

Applications of Machine Learning in Accounting Research

Xikai Chen

Accounting Discipline Group

University of Technology Sydney

Doctor of Philosophy

Year of Submission: 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Xikai Chen, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Accounting Discipline Group at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 2021/August/7

Acknowledgement

It is a genuine pleasure to express my deep sense of thanks and gratitude to my principle supervisor, Distinguished Professor Stephen Taylor. His dedication and keen interest and above all his overwhelming attitude to help his students had been responsible for completing my work. His constant support, meticulous scrutiny, timely advice and scientific approach have helped me to a very great extent to accomplish this thesis.

A debt of gratitude is also owed to my principle supervisor, Professor Yaowen Shan, for his keen interest in me at every stage of my research. His prompt inspirations, timely suggestions with kindness, enthusiasm and dynamism have enabled me to complete my thesis.

I thank profusely all the staff of the Accounting Discipline Group at the UTS Business School for their kind help and co-operation throughout my period of study.

It is my privilege to thank my parents, Mrs. Shaohong Huang and Mr. Jinzhang Chen, for their constant encouragement throughout my research period.

I am extremely thankful to my friends Mr. Feng Xiao and Mr. Claudio Campi. Without their kind emotional support this thesis would not have been possible.

Table of Contents

Chapter 1 INTRODUCTION.....	1
Chapter 2 A RESEARCH GOLD RUSH?	5
2.1 INTRODUCTION.....	5
2.2 MAJOR THREATS	8
2.3 CONCERNS OVER ARBITRARY MODEL SELECTION	11
2.3.1 How is model selection arbitrary?	11
2.3.2 Why is model selection arbitrary?.....	14
2.3.3 The importance of a robust baseline model	16
2.4 APPLICATIONS TO ACCOUNTING RESEARCH.....	19
2.4.1 Audit fees	19
2.4.2 Tax avoidance	20
2.5 A ROLE FOR MACHINE LEARNING TECHNIQUES	22
2.5.1 Why use machine learning?.....	22
2.5.2 Systematic variable review	23
2.5.3 LASSO and random forest.....	26
2.6 CONCLUSION	28
Table 2.1 Number of studies with machine learning related topic on FT50 journals	30
Appendix 2A Summary of empirical audit fee studies published on prestigious journals	31
Chapter 3 MACHINE LEARNING AND AUDIT FEES	35
3.1 INTRODUCTION.....	35
3.2 LITERATURE REVIEW.....	40
3.2.1 Right-hand-side covariates in audit fee model.....	40
3.2.1.1 <i>Client attributes</i>	40
3.2.1.2 <i>Auditor attributes</i>	43
3.2.2 Machine learning techniques.....	45
3.3 METHODOLOGY	47
3.3.1 Data and sample	47
3.3.2 Variable definitions.....	48

3.3.3 Sample summary statistics.....	48
3.3.4 Research method.....	51
3.4 MAIN RESULTS.....	53
3.4.1 Step 1: Weak covariates eliminated based on agreement between AL and RF	53
3.4.2 Step 2: Strongest candidate from each group selected using RF	54
3.4.3 Step 3: Strong variables selected based on agreement between AL and RF	55
3.5 FURTHER ANALYSIS.....	56
3.5.1 Corporate governance variables	56
3.5.2 Fixed-effects test.....	58
3.5.3 Standardized regression analysis.....	59
3.6 REPLICATION OF PRIOR WORKS	60
3.6.1 Dividend payout, earnings persistence and audit fees.....	61
3.6.2 CEO Vega and audit fees.....	62
3.6.3 Aggressive real earnings management and audit fees	63
3.6.4 Short-term debt, credit rating and audit fees	64
3.7 CONCLUSION.....	65
Appendices.....	67
Appendix 3A Number of studies with machine learning related topic in FT50 journals.....	67
Appendix 3B Variable definitions.....	68
Appendix 3C Criteria for variable rating.....	72
Appendix 3D Machine learning algorithm.....	73
Figure 3.1 Flow chart for the three-step analyses	75
Tables	76
Table 3.1 Summary statistics.....	76
Table 3.2 Weak covariates eliminated based on agreement between AL and RF	78
Table 3.3 Strongest candidate from each group selected using RF	80
Table 3.4 Strong variables selected based on agreement between AL and RF	81
Table 3.5 Strong variables selected based on agreement between AL and RF (further analysis)	82
Table 3.6 Fixed-effects tests for the key variables	83

Table 3.7 Fixed-effects tests for the key variables (standardized).....	85
Table 3.8 Dividend payouts, earnings persistence, and audit fees.....	87
Table 3.9 CEO Vega, CEO Delta and audit fees	90
Table 3.10 Aggressive real earning management and audit fees	91
Table 3.11 Ratings, short-term debt and audit fees	92
Chapter 4 MACHINE LEARNING AND TAX AVOIDANCE	93
4.1 INTRODUCTION.....	93
4.2 LITERATURE REVIEW	98
4.2.1 Measures for tax avoidance	98
4.2.2 Determinants of tax avoidance	101
4.2.3 Machine learning techniques.....	107
4.3 DATA AND METHODOLOGY	109
4.3.1 Sample and variables	109
4.3.2 Summary statistics	111
4.3.3 Methodology	113
4.4 MAIN RESULTS	114
4.4.1 Step 1: Strongest variable selected from each group using RF	114
4.4.2 Step 2: Robust variable selected using ALASSO and RF	115
4.5 FURTHER ANALYSIS.....	117
4.5.1 Alternative tax avoidance proxy	117
4.5.2 Median threshold	119
4.5.3 Corporate governance attributes	120
4.5.4 Fixed-effects test.....	122
4.5.5 Replication of prior studies.....	123
4.5.5.1 Internal control weakness and tax avoidance	124
4.5.5.2 Corporate social responsibility and tax avoidance	125
4.5.5.3 Price-cost margin and tax avoidance	125
4.5.6 Comparison between theoretical predictions and empirical outcomes	127
4.6 CONCLUSION	129

Appendix 4A Variable definition.....	131
Figures.....	135
Figure 4.1 Flow chart of the two-round analysis	135
Figure 4.2 Summary matrix of theoretical prediction and empirical outcome	136
Tables	137
Table 4.1 Summary statistics.....	137
Table 4.2 Strongest variable selected from each group for ETR measures	139
Table 4.3 Variable selection for ETR models using fixated threshold	140
Table 4.4 Variable selection for alternative tax avoidance models using fixated threshold.....	142
Table 4.5 Variable selection for ETR models using median threshold	144
Table 4.6 Variable selection for alternative tax avoidance models using median threshold	146
Table 4.7 Variable selection for ETR models using both fixated and median thresholds (further analysis)	148
Table 4.8 Variable selection for alternative tax avoidance models using both fixated and median thresholds (further analysis).....	151
Table 4.9 Fixed-effects tests for the key variables of ETR models.....	154
Table 4.10 Fixed-effects tests for the key variables of alternative tax avoidance models	161
Table 4.11 Tax-related ICW and tax avoidance.....	165
Table 4.12 Corporate Social Responsibility (CSR) and tax avoidance	166
Table 4.13 Price-cost margin (PCM) and tax avoidance	167
Table 4.14 Comparison between theoretical predictions and empirical outcomes	170
Chapter 5 CONCLUSION.....	171
Reference.....	173
Appendices for Chapter 3.....	196
Appendix 3E.....	196
Appendix 3E.1 Adaptive LASSO analysis in Step 1 (main test)	196
Appendix 3E.2 Random Forest analysis in Step 1 (main test)	201
Appendix 3F Random Forest analysis in Step 2 (main test)	202
Appendix 3G	205

Appendix 3G.1 Adaptive LASSO analysis in Step 3 (main test).....	205
Appendix 3G.2 Random Forest analysis in Step 3 (main test).....	207
Appendix 3H	208
Appendix 3H.1 Adaptive LASSO and Random Forest analyses in Step 1 (further analysis).....	208
Appendix 3H.2 Adaptive LASSO analysis in Step 1 (further analysis).....	210
Appendix 3H.3 Random Forest analysis in Step 1 (further analysis).....	215
Appendix 3H.4 Random Forest analysis within each group in Step 2 (further analysis).....	216
Appendix 3H.5 Adaptive LASSO analysis in Step 3 (further analysis).....	219
Appendix 3H.6 Random Forest analysis in Step 3 (further analysis).....	221
Appendices for Chapter 4.....	222
Appendix 4B Random Forest analysis within each group for ETR measures	222
Appendix 4C.....	232
Appendix 4C.1 Adaptive LASSO analysis for annual cash ETR.....	232
Appendix 4C.2 Adaptive LASSO analysis for annual GAAP ETR.....	234
Appendix 4C.3 Adaptive LASSO analysis for long-run cash ETR.....	236
Appendix 4C.4 Adaptive LASSO analysis for long-run GAAP ETR.....	238
Appendix 4D	240
Appendix 4D.1 Random Forest analysis for annual cash ETR	240
Appendix 4D.2 Random Forest analysis for annual GAAP ETR	241
Appendix 4D.3 Random Forest analysis for long-run cash ETR	242
Appendix 4D.4 Random Forest analysis for long-run GAAP ETR	243
Appendix 4E RF analysis within each group for alternative tax avoidance measures.....	244
Appendix 4F	254
Appendix 4F.1 Adaptive LASSO analysis for UTB	254
Appendix 4F.2 Adaptive LASSO analysis for DTAX	256
Appendix 4F.3 Adaptive LASSO analysis for DDBTD.....	258
Appendix 4F.4 Adaptive LASSO analysis for MPBTD.....	260
Appendix 4G	262
Appendix 4G.1 Random Forest analysis for UTB	262

Appendix 4G.2 Random Forest analysis for DTAX	263
Appendix 4G.3 Random Forest analysis for DDBTD	264
Appendix 4G.4 Random Forest analysis for MPBTD	265
Appendix 4H	266
Appendix 4H.1 Adaptive LASSO analysis for annual cash ETR (further analysis)	266
Appendix 4H.2 Adaptive LASSO analysis for annual GAAP ETR (further analysis)	268
Appendix 4H.3 Adaptive LASSO analysis for long-run cash ETR (further analysis)	270
Appendix 4H.4 Adaptive LASSO analysis for long-run GAAP ETR (further analysis)	272
Appendix 4H.5 Adaptive LASSO analysis for UTB (further analysis)	274
Appendix 4H.6 Adaptive LASSO analysis for DTAX (further analysis)	276
Appendix 4H.7 Adaptive LASSO analysis for DDBTD (further analysis)	278
Appendix 4H.8 Adaptive LASSO analysis for MPBTD (further analysis)	280
Appendix 4I	282
Appendix 4I.1 Random Forest analysis for annual cash ETR (further analysis)	282
Appendix 4I.2 Random Forest analysis for annual GAAP ETR (further analysis)	283
Appendix 4I.3 Random Forest analysis for long-run cash ETR (further analysis)	284
Appendix 4I.4 Random Forest analysis for long-run GAAP ETR (further analysis)	285
Appendix 4I.5 Random Forest analysis for UTB (further analysis)	286
Appendix 4I.6 Random Forest analysis for DTAX (further analysis)	287
Appendix 4I.7 Random Forest analysis on DDBTD (further analysis)	288
Appendix 4I.8 Random Forest analysis on MPBTD (further analysis)	289

Abstract

Over the last two decades, accounting research has identified an increasing number of incremental explanatory variables. However, owing to a combination of inadequate review studies and the limitations of conventional tools, researchers have expressed concerns about the growth in incremental explanatory variables, particularly as p-hackers may manipulate test design and data selection to produce statistically significant results. This thesis presents a comprehensive overview of the acute challenges faced by accounting research, including p-hacking, overreliance on *t*-statistics, arbitrary selection of control variables, lack of replication culture, and a shortage of careful review studies. In response to calls for more review studies and more advanced techniques in accounting research, this thesis applies a novel technique—machine learning—to systematically evaluate the vast number of incremental variables in explaining two popular outcomes, namely, audit fees and tax avoidance engagement. Following developments in other areas of economics and business research, this thesis applies two widely used variable-selection-oriented algorithms, LASSO and random forest, to systematically evaluate the large number of explanatory variables that the extant audit and tax literature has increasingly identified.

By focusing on two well-explored research questions (i.e., the determinants of audit fees and tax avoidance), this thesis identifies strong variables that form robust baseline models. These models provide a solid foundation for subsequent audit fee and tax avoidance studies and thereby enhance the comparability and credibility of their results. By replicating a number of prior works and showing the sensitivity of results to these robust baseline models, this thesis also demonstrates the importance of valid arguments, robust baseline models and strong theory, prior to concluding that a novel independent variable is economically and statistically significant. Overall, this thesis provides an example of applying more advanced techniques to tackle problems that are beyond the capability of conventional regression approaches typically relied upon by accounting research.

Chapter 1 INTRODUCTION

Over the last two decades, accounting research has identified an increasing number of incremental variables. However, owing to a combination of inadequate review studies and the limitations of conventional tools, many well-explored research topics, such as studies of the determinants of audit fees and tax avoidance, suffer from the lack of a systematic variable review. In the absence of widely accepted robust baseline models, research in those fields is at risk of so-called p-hacking, which aims to produce statistically significant results by manipulating tests and data (Harvey, 2017). In response to calls for more review studies and more advanced techniques in accounting research (Gow et al., 2016; Johannesson et al., 2020), this thesis applies a novel technique—machine learning—to systematically evaluate the vast number of incremental variables that the literature has accumulated over the years, with a focus on selecting the strongest ones.

Chapter 2 provides a detailed analysis of the research ‘gold rush’, the positive-result-driven trend that has dominated accounting research over the last two decades. The chapter begins by investigating the motivation for such a ‘gold rush’, demonstrating that research subject, journal, and researcher all play a role in encouraging this research culture, which leads to a number of concerns, such as p-hacking and an overreliance on *t*-statistics. Specifically, the arbitrary model selection issue reflects the urgency for a systematic review of variables and the importance of a robust baseline model, which are illustrated through two well-explored topics in accounting research—audit fees and tax avoidance. As those challenges are beyond the capacity of conventional tools such as ordinary least squares regression and meta-analysis, machine learning is introduced as a more advanced technique to systematically evaluate the

large number of variables. In the interest of robustness and popularity, two variable-selection-oriented algorithms, LASSO and Random Forest (RF), perform the analyses in subsequent chapters.

Chapter 3 presents a discussion of audit fee research. It begins with a review of empirical research published in prestigious accounting journals from 2005–2018, which identified more than 300 audit fee covariates from 82 audit fee studies. Based on popularity and data access constraints, 99 variables proxying for a wide range of firm-level, auditor-level, and corporate governance attributes are selected for a three-round test using LASSO and RF. From a large-scale US sample over the period 2000–2019, 12 key variables were identified by both machine learning techniques as enduring. Proven to have robust explanatory power over audit fees by fixed-effects tests and standardized regression analysis, the identified key variables are employed as alternative control variables in the replication of four prior studies, each of which claimed to find an additional incremental variable for the audit fee model. The replication results show that many of those positive findings are sensitive to the inclusion of the key variables. The results in Chapter 3 not only suggest that the identified key variables constitute a robust baseline model for subsequent audit fee research to consider, but also stress the importance of more intuitive arguments and strong theoretical support for empirical studies directed at extending the determinants of audit fees.

Chapter 4 extends the approach in Chapter 3 to tax avoidance research. Eighty-seven tax avoidance covariates and the eight most widely used tax avoidance measures are included in the analysis, derived from more than 300 tax avoidance covariates and more than 20 tax avoidance measures used by 56 empirical studies published in prestigious finance and

accounting journals. Using a large-scale US sample from 1990–2019, LASSO and RF selected numerous robust variables for eight popular tax avoidance measures, whose performances were validated by commonly used fixed effects. The first main finding in Chapter 4 is that widely investigated corporate governance attributes have secondary importance when a comprehensive list of firm attributes is presented. The second main finding exposes the sensitivity of three prior studies to the selection of control variables. The replication analyses using the key variables as controls show mixed results. Finally, a comparison of the theoretical predictions and empirical outcomes of the key variables exposes the inconsistency and complexity inherent in empirical tax avoidance studies. Overall the results in Chapter 4 suggest the need to identify more robust variables to advance our knowledge and understanding of tax avoidance.

This thesis makes a number of contributions to the literature. First, it extends analysis directed at the rapidly increasing numbers of incremental variables (Dyckman and Zeff, 2014; Gow et al., 2016; Harvey, 2017), and provides an extensive analysis of a research ‘gold rush’, in respect of audit fees and tax avoidance. In reviewing multiple influential discussion papers published over the last few years, this thesis presents a comprehensive overview of the acute challenges faced by accounting research, by investigating the main drivers, the major threats, and the alarming consequences of the prevalent positive result hunting.

Second, to the best of my knowledge, this thesis is among the first to introduce machine learning techniques into two important accounting research topics—audit fees and tax avoidance. With its review of empirical studies published in prestigious journals, this thesis provides a comprehensive overview of a large numbers of variables examined in those two research fields. In employing two powerful machine learning techniques, LASSO and RF, this

thesis not only systematically evaluates the performance of a vast number of variables, but also selects a number of enduring variables to constitute robust baseline models, which will serve as a useful starting point for subsequent research.

Third, this thesis provides early evidence in response to numerous expressions of concern regarding p-hacking (Harvey, 2017), the shortage of review studies (Gow et al., 2016), the overreliance on *t*-statistics (Johannesson et al., 2020), the arbitrary selection of control variables (Swanquist and Whited, 2018), and replication failure (Hou et al., 2020). In particular, this thesis addresses these issues in two well-explored research fields, where academics have been calling for a systematic examination (Hay et al., 2006; Hanlon and Heitzman, 2010; Hay, 2013; Wilde and Wilson, 2018).

Finally, this thesis contributes to the stream of research that examines the determinants of audit fees and tax avoidance. Although there are a few qualitative studies using meta-analysis to examine the determinants of audit fees and tax avoidance, to my knowledge, there is no prior study that conducts a systematic quantitative assessment of the large number of explanatory variables used in audit fee or tax avoidance studies. This thesis fills this research gap by providing a comprehensive review of these large numbers of under-investigated variables and proposing robust baseline models for subsequent audit fees and tax avoidance research.

The remainder of the thesis is structured as follows. Chapter 2 presents a discussion of the ‘gold rush’ and the advantages of machine learning techniques. Chapter 3 considers the application of machine learning in audit fee studies. Chapter 4 details the application of machine learning in tax avoidance studies. Chapter 5 concludes.

Chapter 2 A RESEARCH GOLD RUSH?

As the golden news spread beyond California to the outside world, it triggered the most astonishing mass movement of peoples since the Crusades. From all over the planet they came—from Mexico and Peru and Chile and Argentina, from Oregon and Hawaii and Australia and New Zealand and China, from the American North and the American South, from Britain and France and Germany and Italy and Greece and Russia.

H.W. Brands, *The Age of Gold: The California Gold Rush and the New American Dream*

2.1 INTRODUCTION

In his 2017 presidential address for the *Journal of Finance*, Campbell Harvey expressed his concern over increasing tendencies to produce ‘statistically significant’ results in financial economics research. Although more correlations and causal inferences have been discovered, other issues such as unreported results, and direct and indirect p-hacking, lead to Campbell’s concerns about the replication of results. He concludes by suggesting a need for far greater challenging of the validity of extant research findings.

Similar voices have been heard in accounting research. In a widely cited paper, Gow et al. (2016) revisit the accounting literature with a critical lens and express concerns over the trend towards ‘significance-oriented’ research. Given snowballing sample sizes and claimed incremental explanatory variables, contemporary empirical accounting research risks producing misidentified results due to an overreliance on statistical significance as the ‘gold standard’ for justification.

The fields of finance and accounting research are facing challenges from the exponential number of identified factors with conventional statistical significance levels. Thanks to the rapid

development of computational capacity, and improved access to vast types of data at the macro, firm, and even individual levels, there are enormous opportunities in contemporary empirical research in finance and accounting to explore a broad range of topics. Such a situation is, in many respects, analogous to a ‘variable gold rush’, where researchers’ primary objective is to find gold (the new incremental explanatory variable) with statistical significance representing the golden colour, an important yet limited aspect of gold. Through statistical manipulation, dubious hypothesis development and misuse of methodology, many findings can be ‘painted gold’ but they fail to hold up in future studies. Therefore, Harvey (2017) and Gow et al. (2016) call for urgent attention to such a ‘variable gold rush’ in empirical research, asking researchers to take a pause, step back, and reevaluate the extant assessment system for identifying a new incremental variable.

There are three main reasons for exponential growth in the number of identified incremental variables in empirical research. First, it is simply a natural consequence of a research field developing over time from infancy towards maturity. Contemporary archival/empirical accounting research has witnessed an increasing diversity in research topics. Progress in computational capacity and access to expanding data sources have enriched wide-ranging research topics with countless interesting research questions. As more correlations are identified, and resulting causal inferences are made, subsequent studies are able to build on previous works to produce a bigger picture, expanding the literature with new and unique perspectives. Thus, an increasing number of identified incremental variables in archival studies is, to a degree, an inevitable development.

However, the life cycle of research field development fails to fully explain the extant

significance-oriented research culture, since knowing what the incremental factors *are not* is arguably equally important as knowing what they *are* in order to provide a comprehensive research-based understanding of a given observed outcome. Journals are thus the second reason for the exponential increase in the number of identified incremental variables in empirical research. Journals play an important role in publishing mostly positive (significant) results rather than negative (insignificant) results. Because publishing papers with ‘significant’ results attracts more citations (Fanelli, 2012), and the impact factor of a journal is largely determined by the number of citations, it is understandable that journals prefer papers that claim significant results. Such publication bias contributes to data mining by encouraging researchers to search for statistically significant results, leaving readers with only a selected sample of actual conducted research, with most no-finding papers rejected (Harvey, 2017).

Third, researchers themselves also contribute to publication bias. When results are insignificant, most researchers choose to move on instead of publishing them, aware of journals’ preference for papers with positive results. Such a ‘file drawer effect’ (Rosenthal, 1979) results in significantly fewer no-finding papers submitted to journals, since most researchers tend to find such projects unrewarding or a waste of valuable research resources. Some researchers, however, manipulate data and statistical tests in order to produce statistically significant results, a process that is widely known as p-hacking.

It is worth noting that journals in different research fields demonstrate varying levels of tolerance towards publishing no-finding papers. Fanelli (2010) reviews papers published in journals across 20 science and social science fields, and finds that science journals, such as those in the areas of space science and geoscience, are more accepting of papers with results that do

not support the main hypothesis, while economics and business journals have a strong preference for papers that contain results that support the main hypothesis. To sum up, the ‘variable gold rush’ results from the rapid development of finance and accounting literature, but is also driven by researchers and the preferences of journals. This chapter outlines the major concerns that arise from a variable ‘gold rush’ in more detail, and highlights two specific topics in accounting research (audit fees and tax aggressiveness) where these concerns are evident. A brief introduction is also provided to two approaches that are subsequently used in Chapters three and four to establish the extent to which many so-called empirical findings are sensitive to the identification of a robust model from which reliable inferences about novel incremental determinants can be made.

2.2 MAJOR THREATS

There are a number of potential threats underlying any flood of newly identified causal variables. These threats include unreported test results, an overreliance on *t*-statistics and ad-hoc (even opportunistic) hypothesis development.

Unreported test results can also be viewed as a form of p-hacking. As Harvey (2017) points out, p-hacking can take many forms, but the main one is either direct or indirect manipulation of tests and data to produce statistically significant results. A common and very subtle p-hacking approach is selective reporting of results. For example, authors may try a couple of different statistical approaches with various explanatory and control variables, then pick the result with the most appealing statistical significance and submit it for publication. Or authors may study the correlation between variables and test those variables one by one, carefully ‘tuning’ the model so that the variable of interest becomes significant. In reporting the results, only cherry-

picked variables are revealed to readers, with the rest of the test left unreported, or hidden in untabulated results.

Another form of p-hacking is data manipulation. In the absence of convincing arguments, in order to produce positive results, authors may select a specific sample period rather than use the full sample range. Such p-hacking leads to false positives, and the results subsequently fail to hold in future work. Evidence of the extent of this problem can be seen in variables that are identified as important explanatory factors by studies published in leading accounting journals, but are then rarely cited in subsequent attempts to identify additional causal variables for the same outcome variable.

A second threat arises from overreliance on *p*-values or *t*-statistics, which in turn potentially overstates the marginal contribution of the new identified variable. As Johannesson et al. (2020) point out, researchers are inclined to judge whether an incremental explanatory variable is successfully identified based on two standards: (1) if the direction of the coefficient of the interested variable is the same as that predicted by theory; and (2) most importantly, if the coefficient associated with the ‘novel’ variable is statistically significant at conventional levels. The primary consideration of readers is the development of the hypothesis and whether the test rejects the null hypothesis. The coefficient magnitude of the interested variable, however, receives little attention. Such prioritization given to beating the minimum acceptable *t*-statistic does not necessarily give the identified variable significant explanatory power, especially given the rapidly growing sample size in concurrent empirical research. Unfortunately, most researchers’ perceptions of the statistical test fail to take such a dramatic change in sample size into account.

Yet *t*-statistics' sensitivity to sample size is surely well-understood. For a large sample of more than 100,000 observations, a *t*-statistic of 3 is not necessarily impressive, since the sample is not sufficiently small, but when put in the context of 100 observations, such a *t*-statistic level would be very informative. An overreliance on *t*-statistics inevitably introduces Type I errors (i.e., a false null rejection), instead of Type II errors. Many studies have argued statistical significance should not be the only or even the most important bar to reach in order to prove the significant explanatory power of the identified variable (Dyckman and Zeff, 2014; Dyckman, 2016; Kim et al., 2018; Stone, 2018; Dyckman and Zeff, 2019).

Without careful examination of other aspects of the results, incremental variables certified solely by *t*-statistics may be misleading. The adjusted R^2 might show only a very marginal change, and/or the magnitude of the coefficient is too small to demonstrate a significant economic effect. Surveying papers published in accounting journals in 2014, Kim et al. (2018) find that only 40% of empirical studies demonstrate statistical significance under alternative criteria. Using standardized regression to measure the incremental explanatory power of variables, Johannesson et al. (2020) replicate 10 studies published in prestigious accounting journals, and find that in eight of the replications the explanatory contribution of the claimed incremental explanatory variable is only marginally different from zero. Thus, relying too heavily on *t*-statistics to justify the explanatory power of an incremental variable is problematic.

The third concern is HARKing, which stands for 'hypothesis after the results are known' (Kerr, 1998). A conventional and ethical procedure for empirical research starts with a literature review, identification of research gaps and hypothesis development, followed by data collection and testing of results. HARKing reverses this process, and thus the literature review and

hypothesis development are usually less satisfactory, since they have to serve the known results rather than explore an interesting research question. As Nuzzo (2014, p.151) points out, ‘the more implausible the hypothesis—telepathy, aliens, homeopathy—the greater the chance that the exciting finding is a false alarm, no matter what the p-value is’. When the hypothesis development is tenuous and the story is not convincing, the statistically significant findings are open to question. Therefore, a detailed literature review with careful examination of theory and prior empirical works should be a critical component of any paper claiming an incrementally significant causal contribution.

2.3 CONCERNS OVER ARBITRARY MODEL SELECTION

2.3.1 How is model selection arbitrary?

Apart from the threat of p-hacking and an overreliance on *t*-statistics in incremental explanatory variable identification, contemporary empirical accounting research is also subject to the arbitrary selection problem for the control variable set. Where there is no unanimous agreement on what constitutes an elementary set of controls for a robust baseline model, researchers are faced with opportunities to test a wide range of candidate variables and to cherry-pick those that produce ‘significant’ results, and report only those results, with other unsuccessful attempts left undocumented. Although it is difficult to detect, such selected reporting of multiple tests for significant results is a subtle yet destructive form of p-hacking (Harvey, 2017).

Although there are increasing concerns regarding the challenges brought about by the flood of identified incremental variables and explanatory power issues (Dyckman and Zeff, 2014; Harvey et al., 2016; Stone, 2018; Johannesson et al., 2020), there is another underestimated yet equally concerning problem, namely the use of large sets of control variables where the selection

of these control variables is best described as arbitrary. In order to disentangle the effect of the incremental variable on the dependent variable, empirical accounting research designs have conventionally employed a list of control variables in the hope of alleviating the omitted variable problem. Compared to the (usually) singular protagonist of the analysis—the incremental variable—the number of control variables included depends on the research question, thereby allowing researchers flexibility in deciding what control variables to include. Given increasing numbers of research papers using a common dependent variable, it is not surprising that so-called control variables would expand in number. Swanquist and Whited (2018) examine the number of variables on the right-hand side of the model in archival studies published in *The Accounting Review* in 1995, 2005 and 2015. They find a significant increase in the average right-hand side variable number from 7.4 to 17.9, suggesting a fast-growing control variable set in empirical accounting studies over the last two decades.

What does a typical control variable selection process from empirical research look like? Swanquist and Whited (2018) reveal two of the most common approaches for control variable selection: (1) borrowing control variables used by prior studies with similar research topics (i.e., the ‘borrowing’ approach); and (2) an exhaustive approach which includes all variables that are potentially correlated with the dependent variable (i.e., the ‘kitchen sink’ approach). Whichever selection approach is applied, readers would naturally expect the author to provide solid arguments to validate their choice of control variables. Under the borrowing approach, authors should demonstrate similarities between their study and prior research, and argue why the model from earlier research is appropriate in the context of their own research question. Under the ‘kitchen-sink’ approach, authors should elaborate on why the chosen variables determine the

dependent variable and thus are worthy of consideration. However, Swanquist and Whited (2018) show that the vast majority of the surveyed studies allocate very limited space to discussion of control variable selection. Instead, the primary focus is on the measurement of the dependent variable, followed by a reference to prior studies employing the same variable.

Regardless of which approach is used to selecting control variables, the absence of a valid theoretical framework for the choice of control variables has important consequences for empirical accounting studies. First, there is likely to be a lack of agreement on an appropriate baseline model. For example, empirical research investigating the determinants of audit fees is a well-explored research topic and constitutes a rich literature, with more than 80 papers published in prestigious accounting journals from 2005 to 2019.¹ After reviewing those papers, I identified more than 300 variables of different measures or different proxies used on the right-hand side of audit fees models, and found that the majority of studies provide insufficient justification for the control variable selection. Such arbitrary selection not only risks the introduction of ‘bad’ controls, but also raises concerns over data snooping, that is, selecting the ‘right controls’ from a large pool of candidates, in order to imbue the ‘novel’ explanatory variable (i.e., the independent variable which is introduced as a ‘new’ explanatory variable) with statistical significance.

Second, there is a lack of agreement on how even key control variables should be measured. For example, I surveyed archival studies of audit fees, and found more than 10 different measures for firm loss, compared to relatively few but consistent measures for firm size

¹ *Journal of Accounting Research, Journal of Accounting and Economics, The Accounting Review, Contemporary Accounting Research, Review of Accounting Studies, Accounting, Organization and Society.*

(calculated as total assets, total sales or total market value). Recent research has documented such discrepancies and emphasized the importance of a carefully constructed variable that captures the intended proxy (Bloomfield et al., 2016; Jennings et al., 2020). Although some studies have attempted to assess the performance of highly correlated variables for similar proxies using a meta-analysis approach (Hay et al., 2006; Bruhne and Jacob, 2020), accounting academics have recently recognized the need for more advanced quantitative techniques to provide a systematic review of research where a large number of covariates are beyond the capacity of the traditional OLS econometric tool (Bertomeu, 2020; Krupa and Minutti-Meza, 2021).

2.3.2 Why is model selection arbitrary?

Several factors contribute to the reliance on a growing yet unstable control variable set. First, there is a lack of systematic review of variable explanatory power. As mentioned above, journals in finance and accounting are motivated to publish papers with positive findings out of concerns for citations, which encourages researchers and journals to look for novel correlations or causality inferences, instead of engaging in retrospective descriptive work (Gow et al., 2016). Compared to other fields, empirical accounting research is lacking in replication culture. Replication of extant studies using new settings and data is regarded as less interesting and thus few replication studies appear in top journals. The laborious nature of systematic review work exacerbates this problem. A comprehensive systematic review of different variables requires researchers to replicate a significant number of prior studies from scratch, which is very time-consuming. Accessing private data is another barrier, and even when using publicly available data sources, thorough scanning, cleaning and merging of different databases is required. Such

intense yet unrewarding work discourages researchers from examining a large number of identified variables and consolidating a robust baseline model.

Second, every research question is unique. Critiques on the arbitrary selection of control variables may seem tricky since there are no identical research questions, and thus every research question is novel and may require specific controls. There is nothing wrong with altering a model in the context of a specific research question, but that alone should not exclude efforts to present solid arguments for including (or excluding) potential control variables. Such arguments require strong theory and convincing logical deduction (Bertomeu et al., 2016; Swanquist and Whited, 2018). Establishment of a parsimonious robust baseline model is therefore critical if further developments in explaining a particular outcome (i.e., dependent variable) are to be fairly evaluated (Swanquist and Whited, 2018; Bertomeu, 2020).

Third, extant methodology in archival accounting and finance research has limitations. Most contemporary empirical accounting studies apply OLS regression methods (or a derivative thereof) as the main methodology, either out of habit or in order to avoid advocating other more advanced techniques (Murphy and Miller, 2019). OLS regression is indeed a powerful tool for quantitative analysis when used properly. However, its limitations—such as the curse of dimensionality—have long been debated in academia. When the right-hand side of an OLS regression model includes many variables, overfitting occurs. Taking many independent variables into consideration, the regression model is tailored to fit the sample data rather than reflecting the true pattern from the population. Thus, an overfitted OLS regression model can produce misleading coefficients, p -values and R^2 . The curse of dimensionality thus prevents researchers from conducting either a systematic review of a large number of identified

incremental variables, or considering all viable control variables in a realistic and feasible way.

One popular method for explanatory power assessment of a large number of variables is meta-analysis, which is widely used in management studies. Although meta-analysis can provide a rough summary of variable performance, especially on the direction and magnitude of variable coefficients, its qualitative nature bears the risk of non-replicable results and the limitation of a small number of comparisons at any time. It is not surprising then that accounting academics have been called for more advanced quantitative techniques to provide a systematic review of such a huge number of extant covariates that are beyond the capacity of traditional tools (Bertomeu, 2020; Krupa and Minutti-Meza, 2021).

2.3.3 The importance of a robust baseline model

There are at least three reasons why a robust baseline model is critical. First, it is both unrealistic and unfeasible for new research to control all identified incremental variables based on prior work. Second, more controls do *not* equal better controls. There is a popular misunderstanding that more control variables mean better overall control, alleviating readers' concerns regarding the omitted variable issue and misidentification of causality (Bertomeu et al., 2016). A good control variable set could help alleviate the omitted variable problem, while a bad control variable set may risk distorting the causal interpretation from the model or concealing the absence of important variables due to a host of unnecessary variables (Angrist and Pischke, 2009). Specifically, when including variables that are either the potential outcome of the incremental variable or the dependent variable, or that are mechanically related to either, a model's causal interpretation is impaired. Thus, a lengthy but poorly motivated control set is not better than a parsimonious robust one. Unfortunately, the idea that 'more is better' is popular

among accounting researchers, and possibly mitigates reviewers' concerns regarding the omitted variable issue (Bertomeu et al., 2016; Swanquist and Whited, 2018).

Third, a robust baseline model protects the validity of incremental explanatory variables, providing an extra deterrent against possible p-hacking. After scrutinizing the return anomalies literature in finance and accounting, Hou et al. (2020) replicate a vast number of prior studies with 452 anomalies variables, and find that 65% of the replications fail to demonstrate statistical significance after controlling for some robust variables such as microcaps. Such a high failure rate echoes the concerns of Harvey et al. (2016) who argue that up to 53% of identified significant anomalies are likely false discoveries. Hou et al.'s (2020) replication results reveal an alarming threat plaguing extant empirical research—many significant results are sensitive to the inclusion of robust variables, and such vulnerability calls into question the validity of the claimed incremental explanatory variable (i.e., the 'novel' variable).

Using the opposite logic, when we identify a strong baseline model composed of robust variables, such a baseline model is not only a sound basis for fair evaluation of the new incremental variables, it also alleviates the fake significance issue arising from potential cherry-picking of control variables with a view to reporting the 'best result'. For example, Hou et al. (2015) propose a robust baseline stock return q-factor model, which is composed of the market factor, the investment factor, the profitability factor and the size factor, demonstrating a comparably strong performance relative to the Fama-French (1993) three-factor model. Hou et al. (2015) examine 80 anomalies identified in prior works and find that half of the results become insignificant when the robust baseline model is applied.

Freyberger et al. (2020) provide an insightful response to doubts expressed about the validity of a parsimonious baseline model. They posit that asset pricing studies have long been under the guidance of a parsimonious baseline model without realizing it: the Fama and French (1993) three-factor model. This model has been so powerful that contemporary asset pricing research is analogous to a competition that tries to capture the variances that the three-factor model fails to explain. Fama and French (1993) achieved an unprecedented and successful dimensional reduction, providing a solid common ground and a robust benchmark against which subsequent research could fairly evaluate any newly identified factor. To date, few of these new factors have demonstrated significantly better explanatory power over stock returns than Fama and French's (1993) three-factor model.

As a common ground on which new variables can be evaluated, a robust baseline model is open-minded and will evolve once new robust variables are discovered. Therefore, it is not surprising that the Fama-French three-factor model grew into the Fama-French five-factor model (Fama and French, 2015) and the Fama-French six-factor model (Fama and French, 2018). Still the Fama-French three-factor model serves as one of the most well-known and succinct baseline models in asset pricing research. However, a robust baseline model is not necessarily singular, pointing to the difficulty of finding a perfect, flawless model; academics are openminded regarding a number of competing good baseline models. For example, in asset pricing studies, given the popularity of the Fama-French models, the q-model (Hou et al., 2015) and the Carhart 4-factor model (Carhart, 1997) are also widely employed as alternative baseline models.

2.4 APPLICATIONS TO ACCOUNTING RESEARCH

As noted above, an increasing number of documented correlations and causal inferences will be the result in a fast-growing literature, with the potential for a potentially overwhelming number of incremental variables. In finance, the most salient example is asset pricing research, where more than 100 factors have been identified as important determinants for stock returns (Freyberger et al., 2020; Gu et al., 2020). Accounting research faces similar challenges. In this section, I focus on two well-explored research accounting topics—audit fees and tax avoidance—and demonstrate the importance of establishing a robust baseline model.

2.4.1 Audit fees

Simunic (1980) was among the first examples of an empirical estimation of audit fee determinants. Over subsequent years, the audit fee model has grown lengthier, with more and more incremental variables identified as audit fee determinants. Especially since 2000, when US audit fee data became available, the number of audit fees studies published in leading journals has increased exponentially. From 2005 to 2018, prestigious journals in accounting published 82 papers that claim to have found a novel incremental variable for the audit fee model.² A closer examination of those empirical studies reveals more than 300 covariates of different measures included in the right-hand side of audit fees models, ranging from firm characteristics and auditor features to many corporate governance aspects. Such an extensive number of variables not only confuses subsequent researchers as it is not feasible to take all variables into consideration, but also suggests a high risk of p-hacking. Even for studies with a similar focus, such as those investigating the impact of corporate governance on audit fees, the

² Appendix 2A provides a summary of those 82 studies.

models vary with different control variables and different measurements, while there is very limited defence of the choices made. Such a phenomenon is concerning because of potential ‘window shopping’, multiple testing and cherry-picking.

Hay et al. (2006) use meta-analysis to address the wide variety of variables used to explain audit fee variation. They suggest that the audit fee model is usually composed of two parts, comprising firm-level characteristics and auditor features. The most common firm-level characteristics in the audit fee model are size, complexity, inherent risk, profitability, leverage, ownership, industry and corporate governance. Widely cited auditor features include proxies for auditor quality such as Big N and client-industry specialization/leadership, auditor tenure, report lag, busy season, auditor problems and non-audit services.³ However, rapid growth in audit fee studies has seen the range of potential explanatory variables expand far beyond those considered by Hay et al. (2006). Hence, it is appropriate to carefully assess the extent to which many of these allegedly incremental contributions to explaining audit fees are in fact robust to an appropriate baseline model. Chapter 3 provides such an analysis.

2.4.2 Tax avoidance

The tax avoidance literature⁴ faces similar challenges to the audit fee literature, but with some added problems. Taxation is studied by scholars in economics, law, finance and accounting, all of whom use different languages and different methods to examine corporate tax decisions (Hanlan and Heitzman, 2010). Unlike audit fee research where there is one singular dependent

³ A detailed literature review is provided in Chapter 3.

⁴ I follow Hanlan and Heitzman (2010) and define tax avoidance broadly as the reduction of explicit taxes, covering various terms describing all kinds of tax-reduction-oriented strategies (e.g., ‘aggressiveness’, ‘noncompliance’, ‘evasions’, ‘sheltering’, etc.) used in finance and accounting studies.

variable, namely audit fees, tax researchers have struggled to reach agreement on how corporate tax avoidance should be measured. From a review of finance and accounting studies in prestigious journals from 2000 to 2019, I identified 56 empirical studies on tax avoidance and found more than 20 measures for corporate tax avoidance.⁵ Four variables, in particular, dominate. These are annual cash effective tax rate (ETR hereafter), annual GAAP ETR, long-run cash ETR and long-run GAAP ETR. Other popular measures include unrecognized tax benefits (UTB hereafter), Frank et al.'s (2009) discretionary permanent book-tax difference (DTAX hereafter), Desai and Dharmapala's (2006) discretionary book-tax difference, and Manzon and Plesko's (2002) book-tax difference. In addition to the different dependent variables, those 56 empirical studies include more than 300 different variables on the right-hand side of tax avoidance models. The most widely cited firm-level attributes are size, performance, growth opportunities, asset tangibility, capital structure, financial constraints, complexity, information environment, financial report incentives, corporate governance and industry.

Without a systematic assessment, such a complex situation provides an opportunity for p-hackers to find another incremental variable by cherry-picking the most significant results from different tax avoidance measures and combining them with carefully selected control variables. Further there are relatively few review studies on tax avoidance. Exceptions are Hanlon and Heitzman (2010), who provide a systematic review of different measures for tax avoidance, and Bruehne and Jacob (2019), who survey a large number of identified incremental variables with meta-analysis. Thus, a comprehensive reassessment of extant tax avoidance models is needed

⁵ Further details are provided in Chapter 4.

to provide a clearer picture for future tax research.

2.5 A ROLE FOR MACHINE LEARNING TECHNIQUES

2.5.1 Why use machine learning?

The machine learning technique made its debut in social science research in recent years, in response to calls for more advanced methods to help solve problems that are beyond the capacity of traditional econometric tools. Compared to the widely used OLS model where the estimate is based on a linear relation that minimizes the sum of the squared distances between observations and the fitted line, machine learning employs a training-and-learning strategy, dividing the sample into three sub-samples: the training sample, the validation sample and the test sample. The training sample is used to build up multiple candidate models, which are then assessed using the validation sample to decide which one to take, followed by an unbiased evaluation using a holdout sample to establish out-of-sample performance assessment. Designed for big data analysis, machine learning techniques can be successful in identifying highly complex structures, identifying functional relations with better out-of-sample performance than conventional econometric tools (Mullainathan and Spiess, 2017). Another advantage of machine learning is its high-dimensional nature, which helps overcome the curse of dimensionality that characterizes traditional econometric tools. As linear regression is ill-suited to handle the large number of variables that the literature has accumulated over the years, machine learning provides a powerful mechanism with which to test those variables and select the robust ones.

Leading journals in finance and accounting have already published a number of studies targeting conventionally well-explored fields with machine learning. Using the machine learning technique, asset pricing researchers have been able to pinpoint a number of pivotal

variables for stock returns (Feng et al., 2020; Freyberger et al., 2020; Giglio et al., 2020; Gu et al., 2020). In accounting studies, researchers have applied machine learning to help detect accounting fraud and misreporting (Bao et al., 2020; Bertomeu et al., 2020; Brown et al., 2020). Those studies all recognize the high-dimension challenge faced by conventional regression methods, and employ machine learning to help tackle such issues. For example, one of the most fundamental and well scrutinized fields in finance, the empirical asset pricing literature, has identified more than 100 incremental explanatory variables for stock returns in the last three decades, leaving subsequent researchers with a large pool of candidates for ‘appropriate control variables’, under a shortage of systematic large-scale performance assessment. A similar challenge also exists in the accounting literature (Bertomeu et al., 2020), with areas such as audit fees and corporate tax avoidance facing an extensive number of recently identified incremental variables. The intention of applying machine learning in finance and accounting studies is not to replace the conventional OLS method, but to enhance current models by identifying robust controls and teasing out unnecessary weak variables (Bertomeu, 2020).

2.5.2 Systematic variable review

How then can machine learning assist with a systematic review of extant identified variables, especially when compared to existing methods? Researchers have widely employed two approaches to systematically evaluate variables in a research field. The first is the structured narrative review. Once a research question is established, researchers examine the literature and collect related variables from prior studies, before summarizing their findings and concluding on the effect of those variables. Usually the collected variables are classified as (statistically) significant or insignificant factors, then the frequency of variable usage and statistical test results

are recorded and compared across studies. Such a narrative review approach suffers from a number of obvious limitations.

First, the inherent subjectivity colours the review in many aspects. Different researchers have various standards with regard to topic relevance and thus may select different studies to review. Some may attribute more credibility towards papers published in leading journals, while others may give equal weight to all studies. Even with regard to the same study, some scholars may examine it with a high standard and claim that the results are minor, while others may have a lower threshold and declare that the results are significant.

Second, the effectiveness of the narrative review approach also decreases as the literature becomes larger. When a new research field is discovered and the literature is in its infancy, a narrative review is potentially useful and is likely to be comprehensive. But as more studies are published, it becomes increasingly difficult for a narrative review to synthesize all findings. Very often the effect of a given explanatory variable differs across studies, and the sign of the coefficient may even flip.

Owing to the innate limitations of the structural narrative review, many researchers take the second approach, that is, meta-analysis. Meta-analysis statistically combines findings from prior studies and summarizes the overall effect. Meta-analysis is potentially superior to a narrative review in several respects. For example, meta-analysis sets up clear rules before the literature review, so that explicit and transparent standards are applied to determine which studies are relevant, in order to overcome the subjectivity issue of narrative review (Borenstein et al., 2021). Another advantage of meta-analysis resides in its statistical synthesis of studies.

By reporting the mean and deviation of the variable effect across studies, meta-analysis provides useful inferences to understand the general impact of the variable. Therefore, meta-analysis is able to review and synthesize an extensive number of variables, which is beyond the capacity of narrative review.

However, the primary drawback of meta-analysis is its irreproducibility. Unlike replication studies, meta-analysis fails to directly examine the validity of identified variables using a similar or extended sample. Although recognizing a large number of variables, meta-analysis is unable to simultaneously and quantitatively assess the variables and select the best performers. Given the goal is to achieve a robust baseline model, meta-analysis is poorly equipped to provide convincing answers.

In contrast, machine learning analysis is well suited to overcoming the weaknesses of these two popular systematic variable review approaches for two reasons. First, the quantitative nature of the machine learning algorithm enables a direct comparison of explanatory power between variables. Compared to the subjective yet implicit weights assigned by narrative review, and the synthesis derived from meta-analysis, machine learning analysis allows variables to compete against each other with known algorithms and clear goals, which is more intuitive than a synthesized conclusion. Second, the machine learning algorithm is not subject to dimensionality, the limitation that constrains OLS from high-dimensional analysis. Therefore, machine learning analysis can handle extremely large numbers of observations each associated with a large number of variables, explore latent patterns from big data, and select the top performers. Unlike meta-analysis, machine learning analysis achieves dimensionality reduction by taking off the cream, and is thus well suited to the goal of setting up a robust baseline model.

2.5.3 LASSO and random forest

In order to provide a comprehensive overview of machine learning-related studies in economics and business, I reviewed a sample of the top 50 journals (FT50) used by the *Financial Times* for business school ranking from 2005 to 2021. Then, with regard to subject relevancy, I omitted journals with a primary focus on marketing and operation research, and journals that are pure mathematics-driven, technology-oriented or commentary-based. Table 2.1 summarizes these studies, which include economics, management, strategy, business, and finance research. After reviewing the empirical studies that apply machine learning techniques in the targeted journals, I identified a wide range of popular machine learning algorithms, such as least absolute shrinkage and selection operator (LASSO hereafter), decision trees, Random Forest (RF hereafter), support vector machine (SVM hereafter), k-nearest neighbour, artificial neural networks, and Latent Dirichlet Allocation (LDA). I found that two most popular machine learning methods are RF and LASSO, which are used in 30 and 28 studies respectively. The next most popular techniques are neural networks (25 studies), SVM (19 studies) and gradient boosting (19 studies). Eighteen studies apply textual-analysis-based machine learning algorithms, and 46 studies use other types of techniques. However, with a focus on popularity and robustness, I employee both LASSO and RF as the main analysis tools to revisit the extensive covariates used in the audit fees and tax avoidance literature.

[Table 2.1 about here]

LASSO is the abbreviation for ‘least absolute shrinkage and selection operator’—the name itself signals the sparsity of its nature for variable selection. The model structure of LASSO is similar to OLS regression since they both belong to the regression family. As a basic regression

tool, OLS is subject to the overfitting problem when the variable number increases and the model becomes complex—the lack of regularization on the choices of weight given to each variable will tune the model just to fit the sample when too many factors are considered. LASSO, however, introduces a penalty term for the sum of absolute values of the weight given to each variable, so that the complexity of the model is regularized and only important variables are retained. Therefore, the basic idea behind LASSO is to look for a parsimonious but good model rather than a ‘better’ long and complex one. As a penalty-based regression algorithm, LASSO is designed to tackle high-dimensional data and is a powerful approach for dealing with a large number of variables. By shrinking the coefficients of the weak explanatory variables to zero, LASSO realizes its sparsity feature by eliminating less important variables when competing all variables against each other. Compared to other conventional variable selection methods such as ridge regression and subset selection, LASSO’s variable selection process is less sensitive to data change and the results are easily interpretable (Tibshirani, 1996). Thus, LASSO is a favoured machine learning algorithm for accommodating variable selection issues (Zhang and Huang, 2008; Meinshausen and Yu, 2009).

As a classification-based approach, RF differs from LASSO in its unique tree-like method. Unlike a regression-based method that aims to minimize the residual sum of squares, RF is a repeated double bootstrapping of both the subsample from the training data and the subset of variables. As each double bootstrapping results in a random tree, running a complete RF procedure will generate multiple random trees amounting to a ‘forest’. The bootstrapping procedure is then assessed by comparing the out-of-bag (OOB) error for all variables. Since dropping important variables leads to a large variance in the OOB error, RF goes through each

variable and compares the OOB before and after variable permutation, ranking variables based on their importance. Unlike LASSO's clear-cut shrinkage strategy, RF provides an equally straightforward result with a focus on relativity comparison. Owing to its stable 'out-of-the-box' performance, insensitivity to excessive model tuning, tolerance for nonlinearity and interaction term (Varian, 2014; Athey and Imbens, 2017), RF is an appealing machine learning algorithm which has been applied widely in recent economics and business studies.

LASSO and RF both have their advantages and disadvantages, but they are complementary to some degree. For example, LASSO is based on linear regression and 'erases' the weak variables, while RF allows for nonlinearity and shows the relative importance of variables. The different mechanisms behind the two algorithms likely result in different sets of variables being selected. In the interests of objectivity, I believe it is best to use both methods instead of relying on either one, and then compare the agreements between the two: variables that are selected and certified by both methods should have fairly strong explanatory power.

2.6 CONCLUSION

The primary purpose of this chapter is to highlight the urgent need in areas of accounting research to establish relevant baseline models against which claimed incremental explanatory variables can be evaluated. As an increasingly wide range of causal associations are identified, the likelihood that such results reflect the choice (active or passive) or relevant control variables increases. Given large samples and relatively marginal statistical significance, the chances for p-hacking also increase. Two areas that are identified in this chapter that warrant such investigations are audit fee research, and research examining determinants of tax aggressiveness.

To conclude, I would like to take readers back to the ‘variable gold rush’ metaphor introduced at the beginning of this chapter, where robust incremental explanatory variables are the gold, researchers are the gold hunters, and journals are the gold dealers. The p -value is the golden colour, which is an important aspect of the gold. Since payment for gold is handsome, gold hunters are busy digging for gold and show little interest in unearthing other elements such as sand, silver, or rocks that could in fact be diamonds (non-significant results). Most gold hunters use a conventional and effective gold detector, namely OLS, to find gold. Some gold hunters who claim to find gold paint iron gold and try to sell it to gold dealers and other hunters, misbehaviour known as p -hacking. With more and more hunters claiming to find gold, gold dealers and other gold hunters become increasingly sceptical and reexamine the purity of that gold with a more advanced gold detector, the machine learning technique. In the next two chapters, considering robustness, I use two different types of gold detectors, LASSO and RF, and revisit the audit fees and tax avoidance literature.

Table 2.1 Number of studies with machine learning related topic on FT50 journals

This table presents numbers of papers with machine learning related topics published on FT50 journals from 2005-2020.

Journal	ML#	Journal	ML#
Abacus	3	Journal of Empirical Finance	4
Accounting and Business Research	1	Journal of Finance	2
American Economic Review	8	Journal of Financial Economics	4
Accounting Horizons	2	Journal of Financial Intermediation	0
Auditing: A Journal of Practice and Theory	1	Journal of Financial Markets	4
Academy of Management Executive	0	Journal of Financial and Quantitative Analysis	0
Academy of Management Journal	0	Journal of Financial Reporting	0
Academy of Management Perspectives	0	Journal of Financial Stability	5
Academy of Management Review	0	Journal of International Business Studies	0
Accounting, Organization and Society	0	Journal of Law and Economics	0
Administrative Science Quarterly	0	Journal of Management	0
British Accounting Review	1	Journal of Management Accounting Research	0
Contemporary Accounting Research	0	Journal of Management Studies	0
Critical Finance Review	0	Journal of Political Economy	2
California Management Review	0	Journal of Small Business Management	2
European Accounting Review	0	Management Accounting Research	0
Econometrica	5	Management International Review	0
Entrepreneurship Theory and Practice	4	Management science	22
Financial Analysts Journal	2	Organizational Behaviour and Human Decision Processes	2
Financial Management	4	Organization Science	1
Human Relations	0	Organization Studies	0
Human Resource Management	0	Quarterly Journal of Economics	5
Journal of Accounting Auditing and Finance	0	Review of Asset Pricing Studies	0
Journal of Accounting and Economics	2	Review of Accounting Studies	5
Journal of Accounting Literature	1	Review of Corporate Finance Studies	0
Journal of Accounting and Public Policy	1	Review of Economic Studies	3
Journal of Accounting Research	3	Review of Financial Studies	7
Journal of Business Ethics	3	Rand Journal of Economics	0
Journal of Banking & Finance	12	Review of Finance	0
Journal of Business Finance & Accounting	4	Research Policy	1
Journal of Behavioral Finance	3	Strategic Entrepreneurship Journal	0
Journal of Business Venturing	1	Strategic Management Journal	6
Journal of Corporate Finance	2	The Accounting Review	0

Appendix 2A Summary of empirical audit fee studies published on prestigious journals

This table summarizes the surveyed empirical audit fee studies. Those studies are published in six leading accounting journals from 2005-2018.

Year	Paper	Author
2005	The importance of business risk in setting audit fees: Evidence from cases of client misconduct	Francis et al.
2005	The pricing of national and city-Specific reputations for industry expertise in the US audit market	Lyon and Maher
2006	Pricing of initial audit engagements by large and small audit firms	Ghosh and Lustgarten
2006	Auditors' response to political connections and cronyism in Malaysia	Gul
2008	Audit labor usage and fees under business risk auditing	Bell et al.
2008	Auditor specialization, auditor dominance, and audit fees: The role of investment opportunities	Cahan et al.
2008	Audit pricing, legal liability regimes, and Big 4 premiums: Theory and cross-country evidence	Choi et al.
2008	Evidence on the audit risk model: Do auditors increase audit fees in the presence of internal control deficiencies?	Hogan and Wilkins
2008	Litigation risk, audit quality, and audit fees: Evidence from initial public offerings	Venkataraman et al.
2009	Industry specialization by global audit firm networks	Carson
2009	Cross-listing audit fee premiums: Theory and evidence	Choi et al.
2010	The regulation of public company auditing: Evidence from the transition to AS5	Doogar et al.
2010	Short-term debt maturity structures, credit ratings, and the pricing of audit services	Gul and Goodwin
2010	An empirical analysis of auditor independence in the banking industry	Kanagaretnam et al.
2010	Examining the potential benefits of internal control monitoring technology	Masli et al.
2011	Do control effectiveness disclosures require SOX 404 (b) internal control audits? A natural experiment with small US public companies	Kinney and Shepardson
2011	The effect of using the internal audit function as a management training ground on the external auditor's reliance decision	Messier
2012	Audit fee reductions from internal audit-provided assistance: The incremental impact of internal audit characteristics	Abbott et al.
2012	Audited financial reporting and voluntary disclosure as complements A test of the confirmation hypothesis	Ball et al.
2012	The effects of firm-initiated clawback provisions on earnings quality and auditor behavior	Chan et al.
2012	Shareholder voting on auditor selection, audit fees and audit quality	Dao et al.
2012	City-level auditor industry specialization, economies of scale, and audit pricing	Fung et al.
2012	Agency conflicts and auditing in private firms	Hope et al.
2012	The impact of mandatory IFRS adoption on audit fees: Theory and evidence	Kim et al.
2012	The consequences of protecting audit partners' personal assets from the threat of liability	Lennox and Li

2012	An empirical test of spatial competition in the audit market	Numan and Willekens
2012	Audit partner specialization and audit fees: Some evidence from Sweden	Zerni
2013	Does social trust matter in financial reporting? Evidence from audit pricing	Berglund and Kang
2013	Costs and benefits of requiring an management partner signature recent experience in the United Kingdom	Carcello and Li
2013	How much does IFRS cost IFRS adoption and audit fees	George et al.
2013	Does auditor industry specialization improve audit quality?	Minutti-Meza
2014	The association between individual audit partners' risk preferences and the composition of their client portfolios	Amir et al.
2014	Public equity and audit pricing in the United States	Badertscher et al.
2014	Who's really in charge? Audit committee versus CFO power and audit fees	Beck and Mauldin
2014	The effect of audit committee industry expertise on monitoring the financial reporting process	Cohen et al.
2014	Does corporate tax aggressiveness influence audit pricing?	Donohoe and Knechel
2014	Fair value and audit fees	Goncharov et al.
2014	Is the effect of industry expertise on audit pricing an office-level or a partner-level phenomenon?	Goodwin and Wu
2014	The effect of governance on specialist auditor choice and audit fees in US family firms	Srinidhi et al.
2015	Benefits and costs of auditor's assurance: Evidence from the review of quarterly financial statements	Bedard and Courteau
2015	Auditor industry specialization and evidence of cost efficiencies in homogenous industries	Bills et al.
2015	Did the 2007 PCAOB disciplinary order against Deloitte impose actual costs on the firm or improve its audit quality?	Boone et al.
2015	Executive equity risk-taking incentives and audit pricing	Chen et al.
2015	Who did the audit? Audit quality and disclosures of other audit participants in PCAOB filings	Dee et al.
2015	Assessing financial reporting quality of family firms: The auditors' perspective	Ghosh, Tang
2015	Fee discounting and audit quality following audit firm and audit partner changes: Chinese evidence	Huang et al.
2015	Audit fees and social capital	Jha and Chen
2015	The effect of China's weak institutional environment on the quality of big 4 audits	Ke et al.
2015	CEO equity incentives and audit fees	Kim et al.
2016	Audit hours and unit audit price of industry specialist auditors: Evidence from Korea	Bae et al.
2016	Small audit firm membership in associations, networks, and alliances: Implications for audit quality and audit fees	Bills et al.
2016	Audit pricing for strategic alliances: An incomplete contract perspective	Demirkan and Zhou
2016	On the benefits of audit market consolidation: Evidence from merged audit firms	Gong et al.

2016	Do school ties between auditors and client executives influence audit outcomes?	Guan et al.
2016	A potential benefit of increasing book–tax conformity: Evidence from the reduction in audit fees	Kuo and Lee
2016	The earnings quality information content of dividend policies and audit pricing	Lawson and Wang
2016	Audit report restrictions in debt covenants	Menon and Williams
2017	The impact of litigation risk on auditor pricing behavior: Evidence from reverse mergers	Abbott et al.
2017	Debt covenant violations, firm financial distress, and auditor actions	Bhaskar et al.
2017	Do CEO succession and succession planning affect stakeholders' perceptions of financial reporting risk? Evidence from audit fees	Bills et al.
2017	Audit fee differential, audit effort, and litigation risk: An examination of ADR firms	Bronson et al.
2017	Do PCAOB inspections improve the quality of internal control audits?	Defond, Lennox
2017	Auditor choice and its implications for group-affiliated firms	Fang et al.
2017	Audit office reputation shocks from gains and losses of major industry clients	Francis et al.
2017	Further evidence on consequences of debt covenant violations	Gao et al.
2017	The consequences of audit-related earnings revisions	Haislip et al.
2017	Do social ties between external auditors and audit committee members affect audit quality?	He et al.
2017	Third-party consequences of short-selling threats: The case of auditor behavior	Hope et al.
2017	The role of audit verification in debt contracting: Evidence from covenant violations	Jiang and Zhou
2017	Are related party transactions red flags	Kohlbeck and Mayhew
2017	Board gender diversity, auditor fees, and auditor choice	Lai et al.
2017	Estimation risk and auditor conservatism	Lennox and Kausar
2017	Do clients' enterprise systems affect audit quality and efficiency?	Pincus et al.
2017	Auditors' response to assessments of high control risk: Further insights	Seidel
2018	The effects of PCAOB inspections on auditor-client relationships	Acito et al.
2018	Public company audits and city-specific labor characteristics	Beck et al.
2018	Transaction costs and competition among audit firms in local markets	Chu et al.
2018	Awareness of SEC enforcement and auditor reporting decisions	Defond et al.
2018	Consequences of adopting an expanded auditor's report in the United Kingdom	Gutierrez et al.
2018	Measuring accounting reporting complexity with XBRL	Hoitash and Hoitash

2018	Audit personnel salaries and audit quality	Hoopes et al.
2018	Accounting comparability, audit Effort, and audit outcomes	Zhang

Chapter 3 MACHINE LEARNING AND AUDIT FEES

3.1 INTRODUCTION

In the past two decades, with the increasing availability of new data sources and the prevalence of cross-discipline research topics, empirical finance and accounting research has explored a variety of issues (e.g., audit fees) from a range of different perspectives and identified numerous new variables as important covariates for the key variable of interest. While it is claimed that most of these covariates are incrementally important, many researchers have expressed serious concerns regarding the issue of so-called ‘p-hacking’, where the research relies too heavily on statistical significance (e.g., p -value) as the only justification for significant results (Gow et al., 2016; Harvey, 2017). Accordingly, there have been increasing calls for a systematic review of well-explored fields and application of more advanced empirical tools to improve the reliability of and replicability in empirical research (Dyckman and Zeff, 2014; Gow et al., 2016; Harvey et al., 2016; Harvey, 2017).

This study responds to these calls by applying machine learning techniques to systematically identify and assess important audit fee covariates, one of the most well-explored fields in archival accounting and auditing research. The audit fee data in the US have become widely accessible since 2000, and in the period 2005–2019, there were 82 empirical studies on US audit fees published in six leading accounting journals.⁶ In these, over 300 variables, measuring a wide range of firm characteristics and auditor attributes, were identified as

⁶ *Journal of Accounting Research*, *Journal of Accounting and Economics*, *The Accounting Review*, *Contemporary Accounting Review*, *Review of Accounting Studies*, and *Accounting, Organization and Society*.

important audit fee covariates.

Given the extensive number of audit fee covariates, it is almost impossible and extremely costly for follow-up research to take all identified variables into account. More importantly, the lack of a commonly used baseline audit fee model leaves room for ‘window shopping’ and p-hacking whereby the best results are adopted and reported following numerous iterations of tests while concealing all unsuccessful attempts (Harvey, 2017). For example, many studies allocate very limited space to justifying the selection of control variables for the audit fee model. In addition, for a particular variable, different studies might choose different measures.⁷ Although every research question is unique and thus may require specific controls, this factor alone does not reduce the amount of work involved in providing intuitive arguments and valid theoretical supports, nor does it undermine the importance of a robust baseline as the starting point.

To the best of my knowledge, Hay et al. (2006) and Hay (2013) are the only two studies that attempt to identify important determinants of audit fees among numerous variables via meta-analysis. Although meta-analysis is widely used in management research for reviewing variables in qualitative-driven studies, the irreproducibility issue makes meta-analysis a less than ideal method to provide a convincing conclusion from synthesized results for two reasons. Firstly, systematic review and replication is laborious and likely unrewarding, and so there is a tendency to identify new incremental variables rather than produce retrospective studies (Gow et al., 2016). Secondly, the limitations of conventional tools make a systematic variable review of this vast literature particularly challenging. Regarding ordinary least squares (OLS)

⁷ For example, different studies provide different definitions for variable size, such as total assets, total sales, and total market values.

regression, the first issue is ‘the curse of dimensionality’: an extensive number of variables will inevitably lead to overfitting and invalidate the review. The second issue, resulting from the wide use of OLS, is a growing overreliance on *t*-statistics to justify explanatory power (Johannesson et al., 2020), which ignores the marginal contribution of identified incremental variables and the dramatic change in sample sizes over the last two decades. As rapidly growing audit fee studies have out-accommodated Hay et al.’s (2006) discussions (2006) with more novel data sources, larger sample sizes and more identified incremental variables, it is time to revisit this established literature with a critical lens and more advanced techniques.

Given the limitations of conventional econometric tools and meta-analysis in large-scale systematic variable evaluation, I adopt machine learning to develop a baseline audit fee model. With its unique ‘training-and-learning’ algorithm and big-data orientation, machine learning can successfully identify highly complex structures from latent data trends, resulting in better out-of-sample performance than conventional econometric tools (Mullainathan and Spiess, 2017). Thus, the high-dimensional nature of machine learning provides an ideal and powerful tool to evaluate numerous variables documented in the auditing literature over years.⁸

I begin by reviewing 82 audit fee studies published from 2005–2018 in the six leading journals noted above, and identify 99 commonly used variables in audit fee models based on their accessibility and prevalence. Using two different—but also the most prevalent—machine learning approaches (i.e., LASSO and random forest (RF)) in economics and business research,⁹

⁸ In recent years, finance and accounting research has seen a number of studies use machine learning techniques in well-explored fields such as asset pricing (Feng et al., 2020; Freyberger et al., 2020; Giglio et al., 2020; Gu et al., 2020) and accounting misconduct (Bao et al., 2020; Bertomeu et al., 2020; Brown et al., 2020).

⁹ After reviewing papers with machine-learning-related topics published in the *Financial Times* 50 Journals for Business School Ranking (FT50), I find that LASSO and RF are the most popular methods for variable selection. LASSO can be used even under multicollinearity and performs better than other variable selection approaches with

I formulate a robust baseline model for audit fees by identifying 12 important determinants that outperform other variables in machine learning analysis. Further analysis shows that all 12 variables remain important determinants after accounting for unobserved time-variant and firm-specific heterogeneity using firm and year fixed effects.

In response to calls for greater replicability in empirical research (Hou et al., 2015; Harvey et al., 2016; Hou et al., 2020), I demonstrate the importance of the identified baseline audit fee model by replicating four prior empirical studies. While significant results can be successfully replicated by applying the control variables used in all four original studies, I find that: the significant relation with audit fees remains the same for short-term debt and credit ratings (Gul and Goodwin, 2010), becomes weaker for the interaction term of dividend payout and earnings persistence (Lawson and Wang, 2016) and aggressive income-increasing real earnings management (REM) (Greiner et al., 2017), and disappears for CEO compensation portfolio to stock return volatility (CEO Vega hereafter) (Chen et al., 2015). Thus, the replication results further highlight the importance of a baseline audit fee model in auditing research.

Finally, I follow Johannesson et al. (2020) and conduct the analysis using standardized regression to assess the relative importance of various determinants to audit fees. Standard OLS or panel data regression does not allow a direct comparison of estimation coefficients across different variables in the same regression model, because statistical interpretation depends on the scale of both audit fees and independent variables. Instead, all estimated coefficients based

easy interpretation of results (Nazemi and Fabozzi, 2018; Rapach et al., 2013). RF can handle unbalanced data and is robust to outliers (Chen et al., 2004; Hastie et al., 2009). As LASSO is a linear model and RF allows for nonlinearity, each method has strengths as well as weaknesses, and using both methods provides a more comprehensive review of audit fees covariates.

on standardized regressions are presented in comparable units. I find that firm size is the most important determinant of audit fees, followed by business complexity factors and a Big 4 indicator. In addition, the replication results of the four studies present similar results. Credit ratings and short-term debt continue to be economically important determinants of audit fees. The impact of aggressive income-increasing REM and the interaction between dividend payout and earnings persistence is weaker but still significant. In addition, I find no significant relation between CEO Vega and audit fees.

My study makes two important contributions to extant literature. Firstly, to the best of my knowledge, it is among the first to apply machine learning techniques to the audit fee literature and to identify a set of important determinants of audit fees that should be controlled in audit fee models. Therefore, the study differs from those using meta-analysis (Hay et al., 2006, 2011) by comparing the importance of different audit fee covariates. Secondly, the study responds to increasing concerns surrounding the call for more retrospective works (Gow et al., 2016), the importance of a robust baseline model in asset pricing (Harvey et al., 2016; Hou et al., 2020), the overreliance on *t*-statistics (Johannesson et al., 2020) and the arbitrary selection of control variables (Swanquist and Whited, 2018). Applying machine learning techniques, my study addresses these concerns and develops a baseline model with 12 important covariates in one of the most well-researched areas in accounting (i.e., audit fees). My replication evidence further highlights the importance of the baseline model for enhancing the robustness and credibility of empirical audit research.

Overall, my findings suggest that future research should consider both the use of the baseline audit fee model and an intuitive demonstration of how the identified incremental

variable is observed and acted on by auditors before claiming that such a variable is an incremental determinant of audit fees. Without such criteria, even top journals will witness more ‘false positives’ via p-hacking.

3.2 LITERATURE REVIEW

3.2.1 Right-hand-side covariates in audit fee model

Most audit fee studies evaluate the competitiveness of the audit market and examine the issues of contracting and independence during the audit process (Hay et al., 2006) by regressing audit fees on a variety of measures for the attributes that are hypothesized to relate to audit fees. Simunic (1980) provides one of the earliest examinations of the audit process by identifying a number of explanatory audit fee variables. Since then, the number of right-hand-side variables in audit fee studies has grown substantially.

From 2005–2018, the leading accounting journals examined in my study published more than 80 papers on determinants of audit fees. Those papers include more than 300 variables, either as interested variables or control variables, on the right-hand side of the audit fee model. Those variables could be broadly summarized as client attributes and auditor attributes, which empirical studies have further classified into a number of categories. Some variables, such as policy change or country-level variables, are contextual and model-specific, and are included in audit fee models that relate to unique research settings. The following subsections mainly focus on the most widely cited client and auditor attributes.

3.2.1.1 Client attributes

Firm size. Almost all studies consider size as a pivotal variable in the audit fee model, as larger

firms require more efforts to audit (Simunic, 1980; Hay et al., 2006). Size is typically measured as total assets, while some studies use market value equity or revenues instead. To better fit the linear relationship with audit fees, size is usually taken in the natural logarithm form of the raw data. The effect of size on audit fees is overwhelmingly positive and statistically significant across almost all studies, suggesting a strong explanatory power of firm size over audit fees.

Complexity. When a client's business mode is more complex, an auditor's task is more complex and thus their fees are higher (Simunic, 1980; Hackenbrack and Knechel, 1997). There are numerous measures for client complexity, but two measures are more popular than others: (1) number of segments; and (2) foreign operation. Different studies provide different definitions for those two measures. For instance, some studies use the number of geographic segments, while others use the number of business segments, or the sum of business and geographic segments. Some studies measure foreign operation as a ratio of foreign sales to total sales, while others use a dummy variable that indicates foreign operation activity. Regardless of the different measures for complexity, the empirical evidence generally supports a positive relation between complexity and audit fees.

Inherent risk. Inherent risk affects audit fees because certain aspects of auditing carry a higher risk of error and require specialized audit procedures (Simunic, 1980). In empirical research, inventory and receivables are often considered the most difficult elements to audit. The majority of studies measure inherent risk as accounts receivables divided by total assets, inventory divided by total assets, or sum of inventory and receivables divided by total assets. Other measures for inherent risk include current assets divided by total assets, current assets divided by current liabilities, and current assets excluding inventory divided by total assets. Empirical

evidence shows a positive relation between inherent risk and audit fees, while the statistical significance varies from weak to strong.

Profitability. Firms facing low profitability may bear higher financial risk, which leads to higher audit fees (Simunic, 1980). Extant literature has documented a negative relation between profitability and audit fees, with three widely used measures of profitability: (1) return on assets (ROA); (2) cash flow from operations; and (3) occurrence of loss.

Leverage. High leverage increases the risk of firm failure, and thus may lead to higher audit fees (Simunic, 1980). Two measures for leverage are widely used: (1) long-term debt divided by total assets; and (2) sum of long-term debt and short-term debt divided by total assets.

Form of ownership. Form of ownership has been studied as a potential determinant of audit fees because of the varying agency costs inherent in different forms of ownership (Hope et al., 2012; Fang et al., 2017). Research has focused on two types of ownership structure: (1) private firms, as private firms bear higher agency costs and thus pay higher audit fees; and (2) firms with institutional ownership—because a dominant shareholder could either exacerbate or alleviate the agency problem, which leads to conflicting effects on audit fees.

Industry. Auditors and researchers have found that some industries are more difficult to audit than others (Simunic, 1980; Pearson and Trompeter, 1994). Many studies choose to delete observations in financial and utility industries due to their significantly distinctive industry-driven business models (Fama and French, 1992), or apply the industry dummy variable to control such industry heterogeneity.

Governance. Better corporate governance implies a more effective control environment, which

might affect audit fees. Prior studies have explored the relation between corporate governance and audit fees (Beck and Mauldin, 2014; Kim et al., 2014; Bills et al., 2017). Some popular measures include board independence, board size, committee constitution, separation of chief executive officer (CEO) and board chair duties, and committee expertise.

3.2.1.2 Auditor attributes

Auditor quality. Research has found that superior auditors charge higher audit fees (Craswell et al., 1995; Francis et al., 2005; Cahan et al., 2008). There are multiple popular measures for auditor quality. The most common is the dummy variable for Big 4 accounting firms. Auditor specialization is another measure for auditor quality, given specialists have the advantage of better industry knowledge and a larger market share. Auditor specialization is commonly measured as the industry market share of an auditor, in terms of client number or aggregated audit fees.

Auditor tenure. Prior works suggest that auditor tenure affects audit fees (Chan, 1999; Hay et al., 2006; Hay, 2013). For example, as a business strategy, auditors tend to offer lower prices to new clients.¹⁰ Commonly used measures for auditor tenure include (1) a dummy variable that indicates a change of auditor; (2) the length of auditor engagement; and (3) a dummy variable that indicates new auditor engagement within a specific number of years.

Report lag. A longer delay in issuing the audit report is likely to suggest problems or difficulties encountered during the course of auditing (Knechel and Payne, 2001), and thus results in higher audit fees.

¹⁰ This ‘low balling’ phenomenon is revisited and re-examined by Barua et al. (2019), who point out that fee discounting is due to significant measurement bias during the auditor turnover year.

Busy season. In the United States, where 31 December is the fiscal year-end for most firms, auditors are very busy in January and February. Any audit work carried out during this time tends to be more costly due to the need to pay staff overtime, and therefore leads to higher fees.

Audit problems. Problems in completing the audit increase auditors' input and the risk they bear, resulting in higher audit fees (Simunic, 1980). The existence of audit problems is commonly measured as: (1) a dummy variable indicating the issuance of unqualified opinion; or (2) a dummy variable indicating the issuance of going concern.

Non-audit services. Purchase of non-audit services also affects audit fees. Firstly, the extensive change in firm brought by non-audit services could end up extra audit efforts. Secondly, firms purchasing non-audit services tend to be problematic in general (Parkash and Venable, 1993; Firth, 1997; Mitra and Hossain, 2007), which leads to higher risk and thus more auditing.

In addition to the widely cited variables listed above, there are a large number of model-specific variables. Also, many auditor attributes are measured at different levels, such as partner, office and national levels. Following a detailed examination of how these covariates are used in the literature, I identify two prevalent problems related to arbitrariness. Firstly, selection of control variables is often arbitrary, lacking strong arguments or solid theories to support the author's choice. Secondly, there is a lack of agreement regarding a measurement for commonly used variables. For example, I have identified more than 10 different measures for variable loss¹¹ in prior audit fee research. With limited access to data, many papers are difficult to replicate and

¹¹ Loss is widely included as a dummy variable in audit fees models and extant studies present numerous definitions of loss: negative net income, negative net income before extraordinary items, loss reported in previous year, loss reported in the prior three years, net loss reported in either of the last two years, etc.

their identified incremental variables end up as orphan variables, which are rarely cited by subsequent audit fee studies, resulting in a marginal contribution to the literature.

3.2.2 Machine learning techniques

In recent years, accounting research has witnessed the increasing application of machine learning techniques. For example, Bao et al. (2020) apply machine learning to detect accounting fraud in publicly traded US firms, and show that the machine learning algorithm outperforms the conventional logistic regression model. Also, Bertoneu et al. (2020) use machine learning to investigate misstatements, highlighting the benefits of the machine learning model in helping to detect and interpret patterns in ongoing accounting misstatements. Both of these studies demonstrate the advantage of machine learning techniques, especially when dealing with a large sample and high-dimensional analysis. By dividing a sample into training, validation and testing subsamples, machine learning techniques have an advantage over traditional econometric tools by providing robust out-of-sample inferences from the in-sample analysis (Mullainathan and Spiess, 2017). In reviewing a large number of identified incremental variables for audit fees, my study extends machine learning into accounting research with a focus on identifying a robust audit fee baseline model.

In order to provide a comprehensive overview of machine learning application in business and economics research, I collected and reviewed papers with machine learning-related topics published in 50 journals used by *Financial Times* for business school ranking (FT50) during the period 2005–2020.¹² Appendix 3A summarizes the journals and identifies the studies with

¹² I omitted journals whose primary focus is on marketing and operational research, and journals that are pure-mathematics-based, technology-driven and with a commentary style.

machine-learning-related topics. Specifically, 30 studies use RF, 28 studies apply LASSO, followed by neural network (26 studies), support vector machine (19 studies) and gradient boosting (19 studies). Eighteen studies apply textual-analysis-based algorithms and 46 studies use other techniques. With a focus on popularity and robustness, I employ both LASSO and RF as the main analysis tools.

LASSO is a powerful regression-based variable selection algorithm. A shrinkage method that reduces the coefficient of less important variables to zero, LASSO introduces a penalty factor for variable coefficients by eliminating weak variables in order to reduce the number of factors included in the model. LASSO has advantages over other conventional variable selection methods because it is immune to the ‘curse of dimensionality’ and multicollinearity, and its results are easy to interpret (Rapach et al., 2013; Nazemi and Fabozzi, 2018). I follow Reeb and Zhao (2018) and apply a modified form of LASSO, Adaptive LASSO (AL). AL applies different penalty weights to different coefficients in order to achieve the oracle property of the estimators (Zou, 2006), while LASSO applies a common penalty factor to all coefficients.

First introduced by Ho (1995), RF is a tree-based algorithm that combines ‘bagging’ (i.e., random data selection) with random variable selection. With its reliable ‘out-of-the-box’ performance and its need for little model tuning, it is regarded as one of the most popular methods (Athey and Imbens, 2016). RF starts by building many regression trees, whose best node split is determined by a number of randomly chosen variables as predictors. Then the prediction is made by averaging the predictions from all regression trees. This bootstrapping aggregating process across data and variables is evaluated by an estimated error known as out-of-bag (OOB) error, which is recorded for each data point before averaging across all trees. Then

the value of each feature is permuted and the OOB error is computed again based on the perturbed data. Because dropping variables with strong explanatory ability leads to a relatively large variance in OOB error, by averaging the difference between the OOB error before and after permutation, RF computes an importance score for each variable and ranks all variables accordingly.

Since AL presumes a linear relation in the model while RF allows non-linearity, I employ both AL and RF in the hope of providing a more comprehensive picture and to select strong variables for establishing a robust baseline audit fee model. Appendix 3D provides a detailed introduction to LASSO and RF.

3.3 METHODOLOGY

3.3.1 Data and sample

I obtained data for my study from multiple publicly available sources. For accounting and financial information, I accessed the Compustat and Center for Research in Security Prices (CRSP) databases. Thomson Reuters and Institutional Brokers' Estimate System (I/B/E/S) provided information on institutional ownership and analyst following. I also used ExecuComp and BoardEx to capture corporate governance information, such as board size and audit committee structure. I collected audit fees and auditor characteristics from Audit Analytics. As information pertaining to audit fees was not available until 2000, my main sample spans the years 2000–2019 with 106,352 firm-year observations, which cover a wide range of firm attributes and auditor characteristics. Because coverage of corporate governance information is much more limited, in addition to the main sample, I constructed another sample that includes firm variables, auditor features and corporate governance characteristics for further analysis.

3.3.2 Variable definitions

The dependent variable in my analysis is audit fees. I selected 99 variables out of more than 300 covariates, based on popularity and data accessibility in prior audit fee research. For example, extant literature contains more than 10 different measures for variable loss in the audit fee model. Considering data access, I constructed three variables to measure loss, following three measurements that are most widely used by prior studies. Those 99 variables provide a wide coverage of firm, auditor and corporate governance attributes. Firm attributes include size, segments, foreign operation, leverage, performance, loss, age, value, financing, industry, liquidity, tangibility, growth, earnings management, unusual activities, discontinued operation, special items, ownership, stock return, business risk, variance, analyst coverage, employee and bankruptcy likelihood. Auditor attributes include Big 4, fees, internal control, tenure, client importance, busy season, specialist, report lag and competition. Corporate governance attributes include board independence, board size, audit committee, committee expertise and CEO–chair duality. To reduce the impact of outliers, I winsorized continuous variables at the 1% level. Appendix 3B provides detailed variable definitions.

3.3.3 Sample summary statistics

Table 3.1 presents the summary statistics of the main sample. In terms of total assets, the median firm size is roughly \$422.57 million, with a \$372.81 million market value of equity and \$282.47 million in revenues. On average, firms have 2.54 geographic segments, 1.95 business segments and 4.49 segments when considering geographic and business segments together. Thirty-four per cent of firms in the main sample report non-zero foreign pretax income, compared to 28% who report foreign exchange income and 28% who have foreign currency translation

adjustments; the average ratio of foreign sales to total sales is 0.16. The average leverage ratio is 20% for long-term debt. The median return on assets is 2% (using net income), 2% (using net income before extraordinary items) and 8% (using EBITDA). On average, 39% of firms report negative net income, 38% firms report loss in the current year and 40% report negative net income before extraordinary items. Firms on average have 17.5 years and 18.1 years of data listed on CRSP and Compustat, respectively. The average Tobin's Q is 3.05, with a market-to-book ratio of 3.90. Cash flow from operations constitutes on average less than 1% of total assets, while net cash flow from financing activities amount to 10% of total assets.

Approximately 26% of firms issue long-term debt or equity, and 10% of firms are more aggressive. About 7% of firms sell common and preferred stock, 10% of firms issue long-term debt only and 2% of firms issue dividends. About 3% are financial firms and 4% are utility firms, 47% firms belong to the high-tech industry, and 1% of firms are accelerated filers. The average ratio of current assets to total assets is 0.49, compared to the average current ratio of 2.65 and quick ratio of 2.20. Receivables and inventory are roughly equal to 24% of total assets, with receivables and inventory being around 15% and 9%, respectively. Approximately 25% of total assets are property, plant and equipment. Intangible assets amount to 15% of total assets. The average growth rate for total assets is 18%, and for sales it is 24%. The average abnormal accruals, calculated following Jones (1991), Dechow and Dichev (2002) and Kathori et al. (2005), are less than 1% after being scaled by total assets. Around 29% of firms have experienced merger and acquisition, with a 10% rate of restatement and a 23% rate of restructure. Fourteen per cent (10%, if only considering discontinued operation that is above 1%) of firms report discontinued operations. Sixty-one per cent of firms report special items, and 52% of the

shares are owned by institutional owners. Annual stock returns are around 11%, compared to 6% for market-adjusted annual returns. The average standard deviation over five years is 10% for cash flow from operations, 20% for sales and 14% for ROA. The average standard deviation of daily returns is around 3%, with a monthly stock return variance of 15%. Approximately 9% of firms beat the analyst earnings forecast, with an average analyst coverage of 7.3. The median number of employees is around 900. The means of the Altman Z score and the Zmijewski score are 4.20 and 1.03, respectively. There are around 31% firms in the litigation industry.

[Table 3.1 about here]

With regard to auditor attributes, around 71% (71%) of firms are audited by Big 4 (Big 5) accounting firms. On average, total audit fees amount to \$1.99 million and non-audit fees cost about \$0.67 million. Tax service fees amount to about 13% of total audit fees, while audit-related fees cost around \$0.26 million. Approximately 65% of firms receive audit opinions that are not unqualified, 9% firms receive going concern and 8% of firms have inefficient internal controls. About 8% of firms experience a change in auditor; 11% of firms with first-year audit engagement and 29% of firms with audit engagement of less than three years. The average length of audit engagement is 7.82 years (sourced from Compustat) and 4.33 years (sourced from Audit Analytics). Using clients' assets as a measure for auditor market power, I observe a relatively independent client–auditor relation. On an industry-year level, the sum of a client's audit and audit-related fees take on average around 22% of an auditor's audit and audit-related fees. A busy season where the fiscal year-end is 31 December contributes 68% of audit work (83% if the fiscal year-end is between December and March), and the average report lag is 81 days. Approximately 23% of firms are audited by industry specialists. The average Herfindahl

index is 0.16 for the four-digit SICC industry classification and 0.04 for the two-digit SICC industry classification.

In regard to corporate governance attributes, each board has an average of nine members, with 81% of positions occupied by independent directors. The audit committee on average has 3.89 members, and a third of these are accounting or financial experts. Roughly 48% of CEOs also act as chairman of the board.

3.3.4 Research method

I ran three rounds of tests to obtain a parsimonious robust baseline model. In the first round, I applied both AL and RF. Specifically, I used a two-year rolling window approach¹³ in the AL analysis and counted the frequency of each variable selected by AL across 19 rolling windows. I ranked all variables from the most frequently selected to the least frequent, and divided them into four groups: Best (always selected or selected more than 15 times); Good (selected 13 to 15 times); Medium (selected 9 to 12 times); and Bad (selected less than 9 times). With RF I ran a variable importance analysis and normalized the importance ranking into a 0 to 1 score: a higher score represents higher variable explanatory power over audit fees. Based on variable importance I divided the variables into four groups: Best (importance loading between 0.3 and 1); Good (importance loading between 0.2 and 0.3); Medium (importance loading between 0.1 and 0.2); and Bad (importance loading between 0 and 0.1).¹⁴ I then compared the results from

¹³ I followed Reeb and Zhao (2018) and adopted a two-year-window approach. Compared to a one-year-window approach, a two-year rolling window ensures not only a relatively large number of observations for each sub-sample, but also a relatively large number of sub-samples. For example, for 2000–2019, a one-year-window approach gives 20 sub-samples with each sub-sample containing observations for each individual year. A two-year-window approach gives 19 sub-samples with each sub-sample containing observations for two years, which provides a larger training data set for the machine learning algorithm and enhances its performance.

¹⁴ I failed to find unanimous agreement in the literature on the threshold for RF importance loading analysis. In biology and engineering research where RF is applied to select important features, 10% is used as the bottom line threshold, and a loading higher than 30% is viewed as high. I modified these observations into my rating rules.

both approaches for each variable. In order to select the variable favoured by both methods, when comparing the results between the two methods, I adopted a weak-bias approach. For instance, if both approaches determined a variable to be Best, I labelled it 'Best'. If one approach rated a variable Good while another rated it Best, I labelled it 'Good'. If one approach resulted in Medium while the other resulted in Good or Best, I labelled it 'Medium'. Variables in other scenarios were labelled 'Bad' (Appendix 3C details the rating criteria). The purpose of the first-round test was to eliminate weak variables. Thus, variables needed to be Medium or above to qualify for second-round analysis.

In the second-round test, I began by grouping similar variables together. For example, I put *Size_ta* (log of assets), *Size_mv* (log of market value equity) and *Size_sale* (log of sales) together, as they all proxy for the same aspect. I then ran RF analysis within each group to select the strongest performer. I relied solely on RF in the second-round analysis mainly because the high-dimensional nature of AL makes it a less than ideal tool to select the strongest candidate from a small group. RF, however, illustrates the importance of each variable to audit fees by ranking the variables from highly important to less important, making it easier to identify the top performer. From each group I selected the variable with the highest importance score. Some groups contained only one variable, thus they immediately qualified for the third-round analysis.

In the third-round test, I reran AL and RF analyses of the variables that survived the previous round. I repeated the same procedures used in the first-round analysis and ranked variable performance as Best, Good, Medium or Bad. Out of consideration for robustness, only

Although the division standard is to some degree subjective, I believe my division threshold is straightforward. The untabulated analysis shows that the results are not sensitive to division standard because my primary focus is on finding the strong variables that survive cross-checking from both machine learning algorithms.

variables ranked Best or Good after comparing results from both approaches remained. Figure 3.1 is a flow chart of the three-round tests.

[Figure 3.1 about here]

3.4 MAIN RESULTS

3.4.1 Step 1: Weak covariates eliminated based on agreement between AL and RF

Table 3.2 summarizes the results of the AL and RF analyses of all variables. In the AL analysis, from 19 two-year rolling windows during the period 2000–2019, six variables that failed the test more than 10 times were eliminated. According to the RF analysis, 22 variables performed poorly, with an importance loading lower than 10%. When comparing the results of the two approaches, AL and RF had relatively conflicting opinions on only 10 variables, indicating a low disagreement rate. Appendix 3E provides detailed results of the AL and RF analyses.

Sixty-seven variables qualified for Step 2 analysis. Specifically, size, segments, foreign operation, profitability, loss, age and asset tangibility survived the first test. Leverage, Tobin's Q and five measures of liquidity also passed the test. *Bank*, *DRev*, *Abnacc3*, *MA*, *Restruct*, *Discon2*, *SPII*, *Inst*, *Analysts*, *Employee*, *Atmanz* and *Zmjewski* also entered the next round, together with three business risk measures and two variance proxies. Multiple auditor attributes also survived the test, including auditor quality, non-audit services, audit-related fees, specialist and competition. *GOC*, *Relag* and *TSF*, together with four measures of tenure and two measures of client importance, also qualified for second-round analysis.

[Table 3.2 about here]

3.4.2 Step 2: Strongest candidate from each group selected using RF

Following prior empirical studies, I divided the variables selected in Step 1 into groups (in bold) based on the attributes for which they proxy. Specifically, *Size_ta*, *Size_mv* and *Size_sale* form the **Size** measure. I put *NGS*, *NBSU*, *NBS*, *NSU* and *NS* into **Segments**, and *FO_sale*, *FO_fca*, *FO_D* and *FO_pifo* into **Foreign Operation**. *ROA_ib*, *ROA_nibs*, *ROA_ebitda* and *CFO* comprise **Performance**. *Loss_ni*, *Loss_ibc* and *Loss_libc* constitute **Loss**. **Age** includes *Age1* and *Age2*. **Liquidity** contains *INVREC*, *Invt*, *CATA*, *CUR_r* and *Quick_r*. **Tangibility** is made up of *PPEAT*, *PPEINT* and *Intan*. *Stdoc*, *Stdtoa* and *Stdsc* belong to **Business Risk**, while *Stdmm* and *Varmsr* are in **Variance**. **Bankruptcy** contains *Atmanz* and *Zmijewski*. *Big4* and *Big5* were allocated to **Auditor Quality**, while *CII* and *CI2* form **Client Importance**. **Fees** is composed of *Nonaudit1*, *Nonaudit2*, *TSF* and *AuditRe*. **Tenure** has *Tenure1*, *Tenure2*, *New2* and *Change2*, while **Specialist** contains *SPEC1* and *SPEC2*. Finally, *HHI1* and *HHI2* belong to **Competition**. The rest of the variables are relatively independent measures for different aspects.

RF analysis within each group determined the strongest variable over other similar measurements. Table 3.3 summarizes the selection results. *Size_ta* has the highest importance loading in Size. In Segments, *NS* is the most important factor. In regard to Foreign Operation, *FO_pifo* dominates the other variables. *ROA_ebitda* is the preferred measure for Performance. *Loss_libc* is followed by *Loss_ibc* and *Loss_ni* in Loss. *Age2* shows a significant higher loading than *Age1*. *Quick_r* is the best performer in the Liquidity group, as is *Intan* in Tangibility. In Business Risk, *Stdoc* is stronger than *Stdtoa* and *Stdsc*. *Varmsr* dominates *Stdmm* in Variance, and *Zmijewski* is the strongest variable in Bankruptcy. *Big4* is a better measure for auditor quality than *Big5*. Among the fee-related variables, *Nonaudit1* is the most important factor.

Change2 is the strongest variables in Tenure. Finally, *CII*, *SPEC2* and *HHI2* dominate in their groups. Appendix 3F provides detailed results of the RF analysis.

[Table 3.3 about here]

3.4.3 Step 3: Strong variables selected based on agreement between AL and RF

In the final test, I ran AL and RF analyses on the strongest variables, which, with the rest of the independent measures, entered into the third-round analysis. I increased the bar and required those remaining to be rated either Good or Best to secure a position in the baseline model. Table 3.4 shows that 24 variables survived AL analysis, and 15 variables reached the increased threshold under RF analysis. Twelve variables are thus classified as robust factors for audit fees according to both methods: *Size_ta*, *FO_pifo*, *Lev_debt*, *ROA_ebitda*, *Intan*, *Restruct*, *SPI*, *Stddoc*, *Employee*, *Big4*, *Nonaudit1* and *Relag*. Appendix 3G provides detailed results of the AL and RF analyses.

Consistent with extant literature, size and complexity-related factors (number of segments, foreign operation and employee number) are important: larger firms or firms with more complex business requires greater audit efforts, which result in higher audit fees. Leverage and profitability are another two important factors to consider in audit fee studies. The results reflect the inherent risk in the use of leverage and in poor profitability.

The inclusion of standard deviation of cash flow suggests that business risk is considered by auditors, in addition to other firm-level risks. Another selected variable, intangibility, reflects the need for different audit treatments of tangible assets and intangible assets (e.g., tangible assets are easier to audit because of the physical substance). Two measures for unusual activities,

restructure and special items, are also priced into audit fees, reflecting the changes and risks brought by those activities. With regard to auditor attributes, auditor quality is selected, indicating the fee premium charged by superior auditors. The higher likelihood of exposure to problems may explain the selection of report lag, which reflects the need for more auditing and results in higher audit fees. Regarding the inclusion of non-audit fees, the explanations for higher audit fees could reside in the problematic nature of firms purchasing non-audit services or that the extensive changes brought by non-audit services impact the nature and amount of auditing, and therefore influence audit fees.

[Table 3.4 about here]

3.5 FURTHER ANALYSIS

3.5.1 Corporate governance variables

Extant literature has witnessed the increasing use of corporate governance variables in audit fee models. Following the same procedures and standards used in the main test, I repeated the same three-round tests, instead including the five most widely cited corporate governance variables: board independence, board size, audit committee size, expertise of audit committee and CEO duality as chair. Appendix 3A provides detailed definitions of those five variables. Appendix 3H.1 summarizes the Step 1 results of the AL and RF analyses, while Appendices 3H.2 and 3H.3 present the detailed results. Even with a smaller sample, the variables for firm and auditor attributes performed very consistently, as in the main test in the first-round analysis. With regard to corporate governance measures, only *CEOChair* failed the test. In total, 69 variables qualified for second-round analysis.

These were placed into groups based on the attributes for which they proxy: Size, Segments, Foreign Operation, Performance, Loss, Age, Liquidity, Tangibility, Growth, Business Risk, Variance, Bankruptcy, Big 4, Fees, Tenure, Client Importance, Specialist and Competition. RF analysis was used within each group to identify the strongest factor. Appendix 3H.4 shows the detailed results, which are highly consistent with those from the main test. Thirty-six variables qualified for the third-round test.

In the third-round analysis, comparing the outcomes of AL and RF analyses, 14 variables outperformed the others. Table 3.5 provides an overview of the results. Appendices 3H.5 and 3H.6 present detailed results of the AL and RF analyses. With regard to firm and auditor attributes, consistent with the main test, size, foreign operation, leverage, restructure, special items, standard deviation of cash flow, employee, Big 4, non-audit fees and report lag are considered important factors in the audit fee model. Institutional ownership and analyst coverage are another two robust factors, which indicate that the auditor takes large shareholder control and information environment into consideration. Board independence and board size are two important variables for audit fee studies with corporate-governance settings, reflecting the premium charged by auditors for higher agency costs related to poor corporate governance quality. Prior studies have documented mixed evidence of how audit committees affect audit fees (Vafeas and Waagelein, 2007; Krishnan and Visvanathan, 2009; Hay et al.2013), the machine learning analysis result echoes to this empirical evidence, showing that the widely used audit committee factor is not the most influential corporate governance determinant of audit fees.

[Table 3.5 about here]

3.5.2 Fixed-effects test

Including industry fixed effects, firm fixed effects and time fixed effects is commonly used in accounting research to control omitted variables. For example, studies include industry fixed effects for two main reasons. Firstly, audit fee distribution across industries could be non-random; firms in one industry may pay higher audit fees than their counterparts in another industry due to business complexity differences. Industry fixed effects are used to help mitigate such self-selection bias. Secondly, including industry fixed effects helps alleviate concerns regarding omitted industry characteristics. Likewise, firm fixed effects are widely used in empirical accounting research to control firm-level time-invariant omitted variables, and time fixed effects are applied to control omitted variables that are constant across individual firms but which vary over time. In this section, I assess the robustness of the identified key variables with industry, firm and time fixed effects.

[Table 3.6 about here]

Table 3.6, Panel A presents the results of regressing audit fees on the 12 identified key variables with industry fixed effects, firm fixed effects and year fixed effects. In column (1), with no fixed effects, all variables are highly statistically significant. In column (2), only with industry fixed effects, all variables are highly statistically significant. In column (3), with firm fixed effects, *Stddev* loses statistical significance, while the other variables remain highly significant. Column (4) shows the results with year fixed effects only, and all variables demonstrate statistical significance. In column (5), with both industry fixed effects and year fixed effects, all variables remain statistically significant. In column (6), with both firm fixed effects and year fixed effects, all variables but *Nonaudit1* are statistically significant. In columns

(7) and (8), in addition to time fixed effects and entity fixed effects, I include auditor fixed effects to control varying auditor attributes, and find that all identified key firm attributes are highly robust.

Table 3.6, Panel B presents the fixed effects test results of 14 key variables identified in further analysis (section 3.5.1). The key variables that were selected in both the main test and further analysis, such as size and foreign operation, show consistent performance. Regarding the identified variables in the smaller sample, *Inst* is statistically significant only when firm fixed effects and year fixed effects are considered individually, but not at the same time. *Boardsize* is robust only when year fixed effect is included, while *Boardind* remains statistically significant across all columns. When adding auditor fixed effects to control the various auditor attributes, all firm attributes and corporate governance attributes show high statistical significance, except for *Inst* under firm fixed effects.

The overall results suggest that even under the stringent fixed-effect control, the identified key variables appear to be robust. As commonly used fixed effects fail to invalidate the explanatory power of the identified key variables, I argue that those key variables constitute a universally robust control set for audit fee studies.

3.5.3 Standardized regression analysis

Following Johannesson et al. (2020), who express concerns regarding the incremental explanatory power of the interested variable in empirical accounting research, I applied standardized regression to reassess variable relevance focusing on the magnitude of the estimated coefficients. Compared to traditional regression analysis, standardized regression

normalizes all variables to meet unit variance, which emphasizes the magnitudes of the estimated coefficients instead of t -statistics. A larger standard regression variable coefficient indicates a higher relevance to the dependent variable.

Table 3.7 presents the fixed-effects test results of the 12 key variables using standardized regression. Because the t -statistics are identical to those from the regular regression analysis in Table 3.6, the focus is on the magnitude of the variable coefficients. Size is the most important factor in the audit fee model with an overwhelmingly larger coefficient magnitude than any other variables, followed by business complexity factors such as employee number and foreign operation. With regard to auditor attributes, the coefficient magnitude of *Big4* dominates the other two variables, suggesting that auditor quality is the most important auditor attribute to consider in audit fee models. The coefficient magnitudes of firm attributes and auditor attributes are consistent between Table 3.7, Panels A and B. There are obvious variations in the coefficient magnitudes of institutional ownership and board size, if time fixed effects are included, while the coefficient magnitudes of analyst coverage and board independence remain relatively consistent across different fixed effects tests. To conclude, the standardized regression results suggest an overall solid explanatory power of the identified key variables. No variable shows high statistical significance with a trivial coefficient magnitude.

[Table 3.7 about here]

3.6 REPLICATION OF PRIOR WORKS

In response to calls for more replication research (Harvey, 2017; Hou et al., 2020) I selected

four studies¹⁵ published in prestigious accounting journals whose identified incremental variables are rarely cited by subsequent audit fee studies, to evaluate the importance of a robust baseline model. Closely following the sample and variable construction indicated by the original studies, I first replicated the main results of those studies, and then replaced the control variables with the identified key variables. In order to achieve comparable results, I also applied the same fixed effect set and decimal places as in the original studies.

3.6.1 Dividend payout, earnings persistence and audit fees

Lawson and Wang (2016) find that dividend-paying firms represent lower-risk audit engagement, which results in lower audit fees. Using three different measures for earnings persistence, Lawson and Wang (2016) show that such a negative association between audit fees and dividend payouts is more pronounced for firms with more persistent earnings. Table 3.8, Panel A presents the replication results of the main analysis in Lawson and Wang (2016), confirming a significant negative relation between audit fees and dividend payout. Table 3.8, Panel A also documents a moderate negative association between audit fees and the interaction term of dividend payout and earnings persistence, for two out of three earnings persistence measures.

Table 3.8, Panel B presents results of replication using the 12 identified key variables. The explanatory power of the dividend payout variable shows a moderate decrease, and the coefficient magnitude of the dividend dummy variable shrinks by almost 25%. The interaction term of dividend payout and earnings persistence loses statistical significance for two out of three earnings persistence measures, and becomes marginally statistically significant for the

¹⁵ Lawson and Wang (2016), Chen et al. (2015), Greiner et al. (2017) and Gul and Goodwin (2010). Those four studies use public databases.

third earnings persistence measure. Although Lawson and Wang (2016) employee 20 control variables, the R^2 of the 12-variable model is higher than that of Lawson and Wang's (2016) lengthy model, suggesting that the identified baseline model has better explanatory power. Table 3.8, Panel C presents the results of replication using the 14 key variables. The statistical significance and the coefficient magnitude of the dividend payout variable are similar to those in the original study, with a weak explanatory power of the interaction term. Table 3.8, Panels D, E and F present the replication results using standardized regression. The analyses show a good coefficient magnitude of dividend payout across three models, indicating its high relevance to audit fees. However, the coefficient magnitude of the interaction term is small, which is marginally different from zero once identified key variables are controlled.

In sum, using the robust baseline model, although audit fees are negatively correlated with dividend payment, it is not associated with dividend-payment policy complexed by earnings persistence.

[Table 3.8 about here]

3.6.2 CEO Vega and audit fees

Using a US sample from 2000–2010, Chen et al. (2015) document a positive association between audit fees and CEO Vega, suggesting that executive risk-taking incentives are considered by auditors. Column (1) in Table 3.9, Panel A presents the replication results of the main analysis in Chen et al. (2015), and shows a significant positive effect between CEO Vega and audit fees.

Columns (2) and (3) in Table 3.9, Panel A present the results of replication using the

identified key variables. The statistically significant association between CEO Vega and audit fees becomes insignificant once the key variables are considered, indicating the result is sensitive to the inclusion of robust variables. Although Chen et al. (2015) use 20 variables as controls in their audit fee model, their model R^2 is the same as that of the more parsimonious key-variables models, suggesting that the baseline models have strong explanatory power. Table 3.9, Panel B presents the replication results using standardized regression. Column (1) shows that although highly statistically significant, the coefficient magnitude of the interested variable is relatively small. In columns (2) and (3), once the identified key variables are controlled, the coefficient magnitude of the interested variable becomes marginally different from zero.

To conclude, I fail to find support for CEO Vega as an important explanatory variable for audit fees, once robust key variables are considered.

[Table 3.9 about here]

3.6.3 Aggressive real earnings management and audit fees

Using a US sample from 2004–2011, Greiner et al. (2017) identify a positive relation between aggressive income-increasing REM and audit fees. Greiner et al. (2017) argue that higher audit fees for firms with aggressive income-increasing REM could be interpreted as a compensation for additional audit efforts and increased perceived business risk.

Table 3.10, Panel A, column (1) presents the replication results of the main analysis in Greiner et al. (2017), and shows a statistically significant positive correlation between audit fees and aggressive income-increasing REM in both the current and previous periods. The replication result using the 12 identified key variables in column (2) is similar to that in the original study.

Column (3) shows the results of replication using the 14 identified key variables, and the dummy variable for aggressive income-increasing REM loses statistical significance, suggesting that the explanatory power of aggressive income-increasing REM is sensitive to the inclusion of corporate governance factors. In Table 3.10, Panel B, the standardized regression analysis shows that although the coefficient magnitude of the aggressive income-increasing REM variable looks reassuring, such relevance becomes trivial once corporate governance factors are considered.

[Table 3.10 about here]

3.6.4 Short-term debt, credit rating and audit fees

Using US data from 2003–2006, Gul and Goodwin (2010) find that short-term debt is negatively correlated with audit fees for firms rated by Standard and Poor's, and such a negative correlation is more pronounced for low-rated firms. Gul and Goodwin (2010) attribute these findings to tougher monitoring and better corporate governance for firms with more short-term debt.

Table 3.11, Panel A, column (1) presents replication results of the main analysis in Gul and Goodwin (2010), and shows a statistically significant negative correlation between short-term debt and audit fees. In column (2), the coefficients of short-term debt, as well as the interaction term of short-term debt and rating, remain statistically significant after controlling the 12 key variables. In addition, the negative correlation between credit ratings and audit fees also remains robust, supporting Gul and Goodwin's (2010) argument that ratings reflect company liquidity risk and therefore are considered by auditors. The proposed correlation between audit fees, low-rated firms and short-term debt is weakened when corporate governance factors are considered, as shown in column (3). In Table 3.11, Panel B, the coefficients of three interested variables—

rating, short-term debt and the interaction term—demonstrate a robust magnitude across three models, suggesting a high relevance between short-term debt and audit fees for rated firms.

[Table 3.11 about here]

3.7 CONCLUSION

Using US data from 2000–2019, this study reviews extant audit fee research and reassesses the explanatory power of a large number of audit fee covariates using machine learning. From more than 300 right-hand-side variables that are used by audit fee studies published in prestigious accounting journals, I selected 99 variables that are widely cited with public data access. Employing AL and RF analysis for variable selection, I identified 12 robust variables as essential controls for a baseline audit fee model.

This study contributes to extant literature in three ways. Firstly, it provides a practical application which is relevant to the rapidly growing literature on accounting research using machine learning techniques. To the best of my knowledge, this is the first study that introduces machine learning methods into audit fees research. I adopted two widely used variable-selection techniques from extant interdisciplinary machine learning studies, namely AL and RF, to select variables with robust explanatory power over audit fees, from a large pool of variables. Secondly, in response to increasing debate over p-hackings (Harvey, 2017), replication failure (Hou et al., 2020), overreliance on *t*-statistics (Johannesson et al., 2020) and arbitrary control selection (Swanquist and Whited, 2018), this study not only provides a parsimonious set of robust control variables that future audit fee researchers should consider to benchmark the explanatory power of the new incremental variable, it also demonstrates the importance of a robust baseline model

by exposing the sensitivity of results and trivial coefficient magnitude in prior works.

There are a few caveats for readers when interpreting the results. Firstly, the main purpose of this study is *not to find the perfect model*, but to establish *a robust baseline model*. Therefore, I have aimed to select a set of control variables, with relatively robust explanatory power and public data access, for a wide range of settings in audit fee research. Such motives led me to choose variables based on agreement between two widely used variable-selection oriented machine learning techniques. That is not to say that the variables that were not selected for the final key variable list have *no* impact on audit fees. However, given the primary goal was to establish a robust parsimonious baseline model as a starting point for subsequent research, I argue that the 12 key variables are essential controls. In regard to the replication of prior works, I argue that the misidentification of incremental variables is unintended, and is largely attributable to the limited sample issue and the disadvantage of traditional econometric tools when dealing with large numbers of covariates. The absence of a widely agreed robust baseline model confuses prior works with no benchmark to follow and thus leads to some robust controls being excluded.

This study has its limitations. Owing to limited data access, the analysis only covers variables constructed from public data sources. However, variables based on private data are not widely used in the mainstream literature, but rather usually require a specific setting. Another limitation is the relatively small sample size for analysis including corporate governance variables, as a result of the small coverage in public databases.

Appendices

Appendix 3A Number of studies with machine learning related topic in FT50 journals

This table presents the number of papers with machine learning related topics published in FT50 journals from 2005-2020.

Journal	ML#	Journal	ML#	Journal	ML#
Abacus	3	Journal of Accounting Auditing and Finance	0	Journal of Management Accounting Research	0
Accounting and Business Research	1	Journal of Accounting and Economics	2	Journal of Management Studies	0
American Economic Review	8	Journal of Accounting Literature	1	Journal of Political Economy	2
Accounting Horizons	2	Journal of Accounting and Public Policy	1	Journal of Small Business Management	2
Auditing: A Journal of Practice and Theory	1	Journal of Accounting Research	3	Management Accounting Research	0
Academy of Management Executive	0	Journal of Business Ethics	3	Management International Review	0
Academy of Management Journal	0	Journal of Banking & Finance	12	Management science	22
Academy of Management Perspectives	0	Journal of Business Finance & Accounting	4	Organizational Behaviour and Human Decision Processes	2
Academy of Management Review	0	Journal of Behavioral Finance	3	Organization Science	1
Accounting, Organization and Society	0	Journal of Business Venturing	1	Organization Studies	0
Administrative Science Quarterly	0	Journal of Corporate Finance	2	Quarterly Journal of Economics	5
British Accounting Review	1	Journal of Empirical Finance	4	Review of Asset Pricing Studies	0
Contemporary Accounting Research	0	Journal of Finance	2	Review of Accounting Studies	5
Critical Finance Review	0	Journal of Financial Economics	4	Review of Corporate Finance Studies	0
California Management Review	0	Journal of Financial Intermediation	0	Review of Economic Studies	3
European Accounting Review	0	Journal of Financial Markets	4	Review of Financial Studies	7
Econometrica	5	Journal of Financial and Quantitative Analysis	0	Rand Journal of Economics	0
Entrepreneurship Theory and Practice	4	Journal of Financial Reporting	0	Review of Finance	0
Financial Analysts Journal	2	Journal of Financial Stability	5	Research Policy	1
Financial Management	4	Journal of International Business Studies	0	Strategic Entrepreneurship Journal	0
Human Relations	0	Journal of Law and Economics	0	Strategic Management Journal	6
Human Resource Management	0	Journal of Management	0	The Accounting Review	0

Appendix 3B Variable definitions

This table reports variable definitions.

Variable	Definition
Size_at	Log of total assets.
Size_mv	Log of market value equity.
Size_sale	Log of sales.
NGS	Log of one plus number of geographic segments.
NBS	Log of one plus number of business segments.
NBSU	Log of one plus number of unique business segments.
NS	Log of one plus number of business and geographic segments.
NSU	Log of one plus number of unique business segments and geographic segments.
FO_D	Dummy variable, equals to 1 if there is any foreign currency translation adjustment, 0 otherwise.
FO_sale	The ratio of foreign sales divided by total sales.
FO_pifo	Dummy variable, equals to 1 if the firm reports non-zero foreign income, 0 otherwise.
FO_fca	Dummy variable, equals to 1 if firm reports foreign exchange income or loss, 0 otherwise.
Lev_debt	The ratio of long-term debt scaled by total assets.
ROA_nibs	The ratio of net income scaled by total assets.
ROA_ib	The ratio of net income before extraordinary items scaled by total assets.
ROA_ebitda	The ratio of earnings before interest, tax, depreciation and amortization scaled by total assets.
Loss_ni	Dummy variable, equals to 1 if the firm reports negative net income, 0 otherwise.
Loss_ibc	Dummy variable, equals to 1 if the firm report net income before extraordinary items in current year, 0 otherwise.
Loss_libc	Dummy variable, equals to 1 if the firm report a negative net income before extraordinary items in the previous year, 0 otherwise.
Age1	Log of the number of years the firm has CRSP data.
Age2	Log of the number of years the firm has Compustat data.
MB	Market-to-book ratio.
TQ	Tobin's Q ratio.
CFO	The ratio of cash flow from operations divided by total assets
Issue1	Dummy variable, equals to 1 if there is no merger in the current year, and either of the following conditions applies: long-term debt increased by 20 percent or more, or the number of shares outstanding increased by 10 percent or more after controlling for stock, 0 otherwise.
Issue2	Dummy variable, equals to 1 when the client firm issues equity or long-term debt during the year that is more than 5 percent of total assets, 0 otherwise.
Issue3	The ratio of financing activities net cash flow scaled by total assets.

Issue4	Dummy variable, equals to 1 for firms with sale of common and preferred stock in current period but not in previous period, 0 otherwise.
Issue5	Dummy variable, equals to 1 for firms that issue long-term debt in current period but not in previous period, 0 otherwise.
Issue6	Dummy variable, equals to 1 for firms that issue dividends in current period, but not in previous period, 0 otherwise.
Bank	Dummy variable, equals to 1 if the firm is a bank, 0 otherwise.
Utility	Dummy variable, equals to 1 if the firm is a public utility, 0 otherwise.
ACCF	Dummy variable, equals to 1 if the company is an accelerated filer, 0 otherwise.
HighTech	Dummy variable, equals to 1 if the firm belongs to a high-tech industry and otherwise “0”. Classification is based on OECD two-digit SIC code classification (codes 28, 35, 36, 37, 38, 48, 73, and 87, are classified as high-tech.), 0 otherwise.
CATA	The ratio of current assets divided by total assets.
CUR_r	Current ratio, current assets divided by current liabilities.
Quick_r	Quick ratio, calculated as the current assets less inventory and divided by total assets.
Rect	The ratio of accounts receivables divided by total assets.
Invt	The ratio of total inventory divided by total assets.
INVREC	The ratio of sum of auditee inventory and receivables divided by total assets.
Intan	The ratio of intangible assets divided by total assets.
PPEAT	The ratio of property, plant, and equipment divided by total assets.
PPEINT	The ratio of property, plant, and equipment divided by sales.
DAT	Annual asset growth.
DRev	Annual revenue growth.
Abnacc1	The residuals from the modified jones model with intercept following Kothari et al. 2005, divided by total assets.
Abnacc2	The residuals from the modified Dechow and Dichev (2002) model by McNichols (2002), divided by total assets.
Abnacc3	The performance-matched abnormal accruals based on ROA following Kothari et al. 2005, divided by total assets.
Totacc	Total accruals calculated from the statement of cash flow.
MA	Dummy variable, equals to 1 if there is merger and acquisition at current year, 0 otherwise.
Restate	Dummy variable, equals to 1 if there is restatement at current year, 0 otherwise.
Restruct	Dummy variable, equals to 1 if there is a restructure at current year, and 0 otherwise.
Discon1	Dummy variable, equals to 1 if the firm reports discounted operations, 0 otherwise.

Discon2	Dummy variable, equals to 1 if the firm reports discounted operations and the absolute value of discounted operation is not smaller than 1% of total assets, 0 otherwise.
SPI1	Dummy variable, equals to 1 if the firm reports non-zero special items, 0 otherwise.
SPI2	Dummy variable, equals to 1 if the firm reports special items and the absolute value of special items is not smaller than 1% of total assets, 0 otherwise.
Inst	Percentage of shares held by institutional owners.
ReAdj	Market-adjusted annual returns, after considering delisting return.
Return	The client firm's annual stock return for the current fiscal year, after considering delisting return.
StdDoc	The standard deviation of cash flows from operations calculated over five years.
Stdsale	The standard deviation of sales calculated over five years.
StdROA	The standard deviation of ROA over the past five years.
Stdmm	Standard deviation of residuals from the market model, estimated by daily returns during the year.
Varmsr	The variance of the client firm's monthly stock returns for the current fiscal year
Beat	Dummy variable, equals to 1 if firm meets or just beats the consensus analyst earnings forecast by \$0.01, 0 otherwise.
Analysts	Log of one plus number of analysts contributing to the firm's consensus forecast.
Employee	Log of one plus number of employees.
AltmanZ	Altman Z score.
Zmijewski	Zmijewski's probability of bankruptcy score.
Litigation	Dummy variable, equals to 1 if the firm has litigation risk, 0 otherwise.
Big4	Dummy variable, equals to 1 if the auditor is from Deloitte, PricewaterhouseCoopers, Ernst & Young or KPMG, 0 otherwise.
Big5	Dummy variable, equals to 1 if the auditor is from Arthur Andersen, Deloitte & Touche, Ernst & Young, KPMG, or PricewaterhouseCoopers, 0 otherwise.
LnAF	Log of audit fees.
Nonaudit1	Log of one plus non-audit fees.
Nonaudit2	The ratio of nonaudit fees divided by the audit fees.
TSF	The ratio of tax services fees divided by the audit fees.
AuditRe	Log of audit related fees.
Opinion	Dummy variable, equals to 1 if there is qualified or unqualified opinion at current year, 0 otherwise.
GOC	Dummy variable, equals to 1 if there is going on concern at current year, 0 otherwise.
MW	Dummy variable, equals to 1 if there are ineffective internal controls at current year, 0 otherwise.

Change1	Dummy variable, equals to 1 if there is change of auditor at current year, 0 otherwise. Data come from Compustat.
Change2	Dummy variable, equals to 1 if there is change of auditor at current year, 0 otherwise. Data come from Audit Analytics.
Tenure1	The number of years of the audit engagement, calculated based on the information from Compustat.
Tenure2	The length of the audit engagement, calculated based on the information from Audit Analytics.
New1	Dummy variable, equals to 1 if it is auditor's first year engagement, 0 otherwise.
New2	Dummy variable, equals to 1 if it is auditor's first- or second-year engagement, 0 otherwise.
CI1	Client importance to an audit firm, measured as a client's (logged) assets divided by the sum total of the audit firm's clients' (logged) assets.
CI2	Ratio of the company's audit and audit-related fees to their auditor's total audit and audit-related fees in the industry market.
BusyD	Dummy variable, equals to 1 if the fiscal year ending is the last day of December, 0 otherwise
BusyDM	Dummy variable, equals to 1 if the firm's fiscal year-end is between December and March, and 0 otherwise.
Relag	Log of number of days between the firm's fiscal year-end and financial statement filing date.
SPEC1	Dummy variable, equals to 1 if the audit firm has the highest audit fee revenue for that two-digit SIC code for that year, 0 otherwise.
SPEC2	Dummy variable, equals to 1 if the market share for the client's audit firm is greater than or equal to 30 percent of the audit fees for the client's 2-digit SIC code in a given year, 0 otherwise.
HHI1	Herfindahl concentration index per audit market, based on 4-digit SIC code.
HHI2	Herfindahl concentration index per audit market, based on 2-digit SIC code.
Boardind	Percentage of independent directors on the board.
Boardsize	Log of one plus the number of board members.
Auditsize	Log of one plus audit committee member number.
AFE	Percentage of audit committee members who are accounting financial experts.
CEOChair	Dummy variable, equals to 1 if the CEO also chairs the board of directors, 0 otherwise.

Appendix 3C Criteria for variable rating

This table reports the rating standards for variable performance. Adaptive LASSO column represents the frequency of variable selected out of 19 rolling windows. Random Forest column indicates the variable importance loading.

Division	Adaptive LASSO	Random Forest	Agreement between AL and RF
Best	16-19	(0.3, 1]	If the variable is labelled by both AL and RF as Best.
Good	13-15	(0.2, 0.3]	If the variable is labelled by both AL and RF as Good, or is labelled as Good by one technique while as Best by another.
Medium	9-12	(0.1, 0.2]	If the variable is labelled as Medium, by either AL or RF, while labelled as better than Medium (Good or Best) by another.
Bad	0-8	[0, 0.1]	The rest.

Appendix 3D Machine learning algorithm

1. LASSO

LASSO is abbreviation for ‘least absolute shrinkage selection operator’, a shrinkage-based method that reduces the coefficient value of less important variables to zero. Similar to conventional OLS regression, LASSO also belongs to regression family, with an additional regularization term for model complexity. Let y_t represent a dependent variable observation, and x_{it} stand for an observation of the i^{th} explanatory variable, the LASSO coefficient estimates satisfy the following equation:

$$\hat{\beta} = \arg \min_{\beta} \left[\left(\sum_{t=1}^T \left(y_t - \sum_{i=1}^k x_{it} \beta_i \right) \right)^2 + \lambda \sum_{i=1}^k |\beta_i| \right]$$

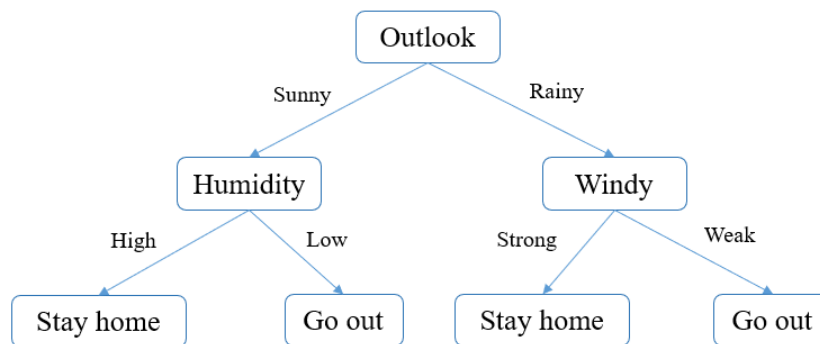
Where β_i is the coefficient for the i^{th} explanatory variable, and λ is the regularization term that controls model complexity. When λ is set to zero, the function of LASSO collapses to OLS regression and produces OLS estimates. If the value of λ is set too large, all coefficients values will shrink to zero. Ten-fold cross-validation is applied to obtain the tuning parameter λ , which minimizes the mean square error in the out-of-sample testing (Hui et al., 2015). The ten-fold cross-validation randomly partitions the training sample into ten subsamples, nine of which are used to fit the model, while the excluded one is for performance evaluation. This process is repeated until every subsample is rotated as the excluded fold, then the value that minimizes the mean square error is selected as λ .

Adaptive LASSO is a modified version of LASSO with the advantage of oracle property (Zou, 2006), which requires an estimator to be consistent in both variable selection and parameter estimation. Therefore, compared to LASSO where all variables are regularized on the same degree, Adaptive LASSO introduces different levels of regularization for each variable and adjusts the penalty factor for each coefficient, so that to achieve the oracle property for the estimators.

2. Random Forest

Random forest is a tree-based method, which is composed of a bunch of individual decision trees (here comes the name “random forest”). A decision tree is a classification method that can be applied to both categorical and continuous variables. Starting with the root node, the decision tree splits into different branches with maximum homogeneity at every step, where the cut-off values and variables that are used to expand on are chosen to minimize the forecast error, with a greedy search algorithm to gain maximum information at each split. Figure 3.2 below illustrates a simple example of decision tree.

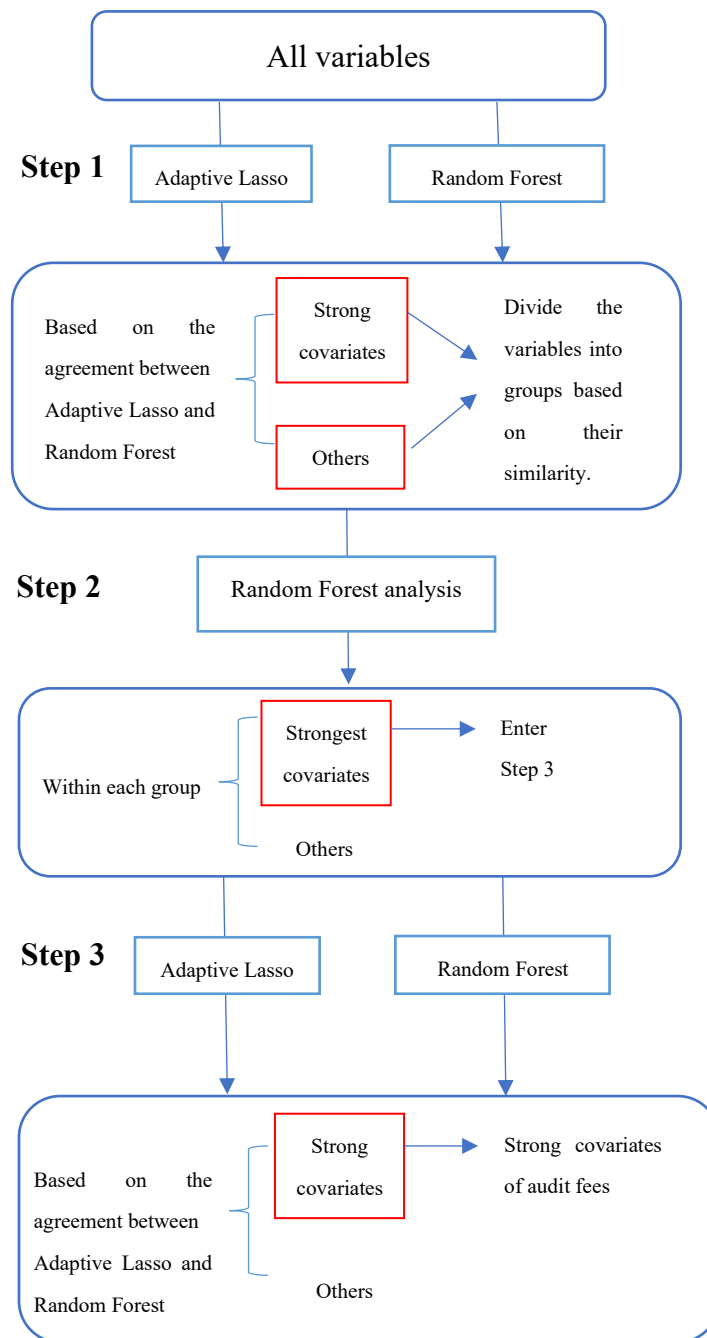
Figure 3.2 Illustration of a simple decision tree



Random forest seeks to avoid overfitting and decreases correlation between trees by randomly selecting subsets of features from a couple of randomly selected trees, a so-called double bootstrapping procedure. Using the subsample, a general relation is estimated between dependent and explanatory variables through a decision tree. Such procedure is repeated until many trees are built, followed by a bootstrapping implemented on an ensemble of trees to estimate the fitted values for the dependent variable. In the end, by comparing the output's sensitivity to the changes on the explanatory variables, Random forest ranks the variable importance accordingly. For random forest, two important parameters are the number of variables to consider for each split and the number of the trees. I choose the square root of the number of the variables for each split, and build up 100 trees for each random forest.

Figure 3.1 Flow chart for the three-step analyses

This figure plots the flow chart for the three-round analyses using Adaptive LASSO and random forest.



Tables

Table 3.1 Summary statistics

This table reports descriptive statistics of the variables. Appendix 3A provides detailed variable definitions.

Panel A: Client attributes

	Mean	SD	Median		Mean	SD	Median
Size_ta	5.95	2.73	6.05	Lev_debt	0.20	0.24	0.13
Size_mv	5.80	2.54	5.92	ROA_nibs	-0.06	0.29	0.02
Size_sale	5.35	2.83	5.64	ROA_ib	-0.06	0.29	0.02
NGS	1.13	0.47	1.10	ROA_ebitda	-0.00	0.29	0.08
NBS	1.00	0.38	0.69	Loss_ni	0.39	0.49	0.00
NBSU	0.86	0.28	0.69	Loss_ibc	0.40	0.49	0.00
NS	1.60	0.43	1.61	CUR_r	2.65	2.88	1.78
NSU	1.53	0.40	1.39	Quick_r	2.20	2.72	1.34
FO_D	0.28	0.45	0.00	Rect	0.15	0.15	0.11
FO_sale	0.16	0.26	0.00	Invt	0.09	0.13	0.03
FO_pifo	0.34	0.47	0.00	INVREC	0.24	0.21	0.19
FO_fca	0.28	0.45	0.00	Intan	0.15	0.20	0.05
PPEAT	0.25	0.26	0.15	Loss_libc	0.38	0.49	0.00
PPEINT	0.80	2.04	0.17	Age1	2.55	0.89	2.64
DAT	0.18	0.63	0.05	Age2	2.62	0.86	2.64
DRevt	0.24	0.90	0.07	MB	3.90	7.17	1.99
Abnacc1	-0.00	0.13	0.01	TQ	3.05	6.54	1.46
Abnacc2	0.00	0.07	0.00	CFO	-0.01	0.26	0.06
Abnacc3	-0.00	0.15	-0.00	Issue1	0.09	0.28	0.00
Totacc	-0.08	0.14	-0.05	Issue2	0.26	0.44	0.00
MA	0.29	0.45	0.00	Issue3	0.10	0.38	0.00
Restate	0.10	0.30	0.00	Issue4	0.07	0.25	0.00
Restruct	0.23	0.42	0.00	Issue5	0.10	0.30	0.00
Discon1	0.14	0.35	0.00	Issue6	0.02	0.16	0.00
Bank	0.03	0.18	0.00	Stddoc	0.10	0.16	0.05
Utility	0.04	0.21	0.00	Stdsale	0.20	0.26	0.11
ACCF	0.01	0.10	0.00	Stdroa	0.14	0.25	0.05
HighTech	0.47	0.50	0.00	Stdmm	0.03	0.02	0.03
CATA	0.49	0.27	0.48	Varmsr	0.15	0.09	0.13
Litigation	0.31	0.46	0.00	Beat	0.09	0.29	0.00
Discon2	0.10	0.30	0.00	Analysts	1.83	0.76	1.79
SPI1	0.62	0.49	1.00	Employee	6.78	2.45	6.88
SPI2	0.32	0.47	0.00	Atmanz	4.20	8.40	2.77
Inst	0.52	0.34	0.57	Zmijewski	1.03	11.66	-1.22
ReAdj	0.06	0.56	-0.03	Return	0.11	0.59	0.05

Panel B: Auditor attributes

	Mean	SD	Median		Mean	SD	Median
LnAF	13.31	1.53	13.38	TSF	0.13	0.23	0.03
Big4	0.71	0.46	1.00	AuditRe	5.95	5.83	8.29
Big5	0.71	0.45	1.00	Tenure2	4.33	4.47	3.00
Nonaudit1	9.99	4.62	11.42	New1	0.11	0.31	0.00
Nonaudit2	0.37	0.61	0.16	New2	0.29	0.46	0.00
CI1	0.00	0.01	0.00	GOC	0.09	0.28	0.00
CI2	0.22	0.35	0.04	MW	0.08	0.27	0.00
BusyD	0.68	0.47	1.00	Change1	0.08	0.26	0.00
BusyDM	0.83	0.38	1.00	Change2	0.08	0.28	0.00
Opinion	0.65	0.48	1.00	Tenure1	7.82	7.92	5.00
Relag	4.35	0.33	4.33	HHI1	0.16	0.18	0.10
SPEC1	0.23	0.42	0.00	HHI2	0.04	0.05	0.02
SPEC2	0.22	0.42	0.00				

Panel C: Corporate governance attributes

	Mean	SD	Median		Mean	SD	Median
Boardind	0.81	0.12	0.83	Boardsize	2.18	0.28	2.20
Auditsize	1.55	0.31	1.61	CEOChair	0.48	0.50	0.00
AFE	0.34	0.25	0.33				

Table 3.2 Weak covariates eliminated based on agreement between AL and RF

This table reports the results of AL and RF analyses in Step 1. Appendix 3A provides variable definitions. AL column shows the frequency of the variable coefficient shrinking to zero in Adaptive LASSO test. RF column shows the variable importance ranking and importance loading in random forest analysis.

	AL	RF	Pass		AL	RF	Pass
Size_ta	0	1(0.3-1)	Yes	Abnacc3	7	71(0.1-0.2)	Yes
Size_mv	0	5(0.3-1)	Yes	Totacc	14	61(0.1-0.2)	No
Size_sale	0	2(0.3-1)	Yes	MA	1	69(0.1-0.2)	Yes
NGS	0	40(0.1-0.2)	Yes	Restate	4	90(0-0.1)	No
NBS	0	17(0.2-0.3)	Yes	Restruct	0	14(0.3-1)	Yes
NBSU	1	56(0.1-0.2)	Yes	Discon1	0	75(0-0.1)	No
NS	4	37(0.1-0.2)	Yes	Discon2	0	30(0.2-0.3)	Yes
NSU	4	39(0.1-0.2)	Yes	SPI1	0	22(0.2-0.3)	Yes
FO_D	7	29(0.2-0.3)	Yes	SPI2	0	87(0-0.1)	No
FO_sale	0	16(0.2-0.3)	Yes	Inst	0	28(0.2-0.3)	Yes
FO_pifo	0	15(0.2-0.3)	Yes	ReAdj	10	83(0-0.1)	No
FO_fca	4	33(0.1-0.2)	Yes	Return	10	76(0-0.1)	No
Lev_debt	2	18(0.2-0.3)	Yes	StdDoc	0	24(0.2-0.3)	Yes
ROA_nibs	1	26(0.2-0.3)	Yes	Stdsale	1	43(0.1-0.2)	Yes
ROA_ib	2	34(0.1-0.2)	Yes	Stdroa	0	23(0.2-0.3)	Yes
ROA_ebitda	2	19(0.2-0.3)	Yes	Stdmm	1	21(0.2-0.3)	Yes
Loss_ni	3	41(0.1-0.2)	Yes	Varmsr	3	42(0.1-0.2)	Yes
Loss_ibc	5	36(0.1-0.2)	Yes	Beat	8	93(0-0.1)	No
Loss_libc	1	50(0.1-0.2)	Yes	Analysts	0	55(0.1-0.2)	Yes
Age1	0	66(0.1-0.2)	Yes	Employee	0	6(0.3-1)	Yes
Age2	2	47(0.1-0.2)	Yes	Atmanz	3	52(0.1-0.2)	Yes
MB	12	72(0.1-0.2)	No	Zmijewski	0	35(0.1-0.2)	Yes
TQ	6	45(0.1-0.2)	Yes	Litigation	2	84(0-0.1)	No
CFO	3	20(0.2-0.3)	Yes	Big4	3	3(0.3-1)	Yes
Issue1	4	79(0-0.1)	No	Big5	0	4(0.3-1)	Yes
Issue2	8	88(0-0.1)	No	Nonaudit1	0	7(0.3-1)	Yes
Issue3	0	44(0.1-0.2)	Yes	Nonaudit2	0	27(0.2-0.3)	Yes
Issue4	9	89(0-0.1)	No	TSF	6	31(0.2-0.3)	Yes
Issue5	13	91(0-0.1)	No	AuditRe	1	8(0.3-1)	Yes
Issue6	10	92(0-0.1)	No	Opinion	2	86(0-0.1)	No
Bank	6	59(0.1-0.2)	Yes	GOC	6	12(0.3-1)	Yes
Utility	1	78(0-0.1)	No	MW	0	82(0-0.1)	No
ACCF	10	94(0-0.1)	No	Change1	4	77(0-0.1)	No
HighTech	1	81(0-0.1)	No	Change2	3	68(0.1-0.2)	Yes
CATA	0	48(0.1-0.2)	Yes	Tenure1	0	49(0.1-0.2)	Yes
CUR_r	2	38(0.1-0.2)	Yes	Tenure2	0	13(0.3-1)	Yes
Quick_r	6	51(0.1-0.2)	Yes	New1	3	74(0-0.1)	No
Rect	15	53(0.1-0.2)	No	New2	2	70(0.1-0.2)	Yes
Invt	2	63(0.1-0.2)	Yes	CI1	0	9(0.3-1)	Yes

INVREC	6	57(0.1-0.2)	Yes	CI2	1	10(0.3-1)	Yes
Intan	0	25(0.2-0.3)	Yes	BusyD	2	85(0-0.1)	No
PPEAT	3	58(0.1-0.2)	Yes	BusyDM	9	80(0-0.1)	No
PPEINT	6	54(0.1-0.2)	Yes	Relag	0	11(0.3-1)	Yes
DAT	11	60(0.1-0.2)	No	SPEC1	3	46(0.1-0.2)	Yes
DRevt	3	65(0.1-0.2)	Yes	SPEC2	0	32(0.2-0.3)	Yes
Abnacc1	14	67(0.1-0.2)	No	HHI1	6	62(0.1-0.2)	Yes
Abnacc2	3	73(0-0.1)	No	HHI2	4	64(0.1-0.2)	Yes

Table 3.3 Strongest candidate from each group selected using RF

This table reports the results of random forest analysis in Step 2. Variables are ranked according to their importance loading. Appendix 3A provides variable definitions.

Group	Size	Segments	Frgn_Op	Performance	Loss
Rank	1.Size_ta	1.NS	1.FO_pifo	1.ROA_ebitda	1.Loss_libc
	2.Size_sale	2.NSU	2.FO_D	2.CFO	2.Loss_ibc
	3.Size_mv	3.NBS	3.FO_fca	3.ROA_nibs	3.Loss_ni
		4.NBSU	4.FO_sale	4.ROA_ib	
		5.NGS			
Group	Age	Liquidity	Tangibility	BizRisk	Variance
Rank	1.Age2	1.Quick_r	1.Intan	1.Stdoc	1.Varmsr
	2.Age1	2.CUR_r	2.PPEINT	2.Stdroa	2.Stdmm
		3.CATA	3.PPEAT	3.Stdsale	
		4.Invt			
		5.INVREC			
Group	Bankruptcy	Big4	Fees	Tenure	Importance
Rank	1.Zmijewski	1.Big4	1.Nonaudit1	1.Change2	1.CI1
	2.Atmanz	2.Big5	2.AuditRe	2.New2	2.CI2
			3.Nonaudit2	3.Tenure2	
			4.TSF	4.Tenure1	
Group	Specialist	Competition			
Rank	1.SPEC2	1.HHI2			
	2.SPEC1	2.HHI1			

Table 3.4 Strong variables selected based on agreement between AL and RF

This table reports the results of AL and RF analyses in Step 3. Appendix 3A provides variable definitions. AL column shows the frequency of the variable coefficient shrinking to zero in Adaptive LASSO analysis. RF column shows the variable importance ranking and importance loading in random forest analysis.

	AL	RF	Pass		AL	RF	Pass
Size_ta	0	1(0.3-1)	Yes	Discon2	0	19(0.1-0.2)	No
NS	0	17(0.1-0.2)	No	SPI1	0	10(0.2-0.3)	Yes
FO_pifo	0	6(0.3-1)	Yes	Inst	4	18(0.1-0.2)	No
Lev_debt	0	15(0.2-0.3)	Yes	StdDoc	0	13(0.2-0.3)	Yes
ROA_ebitda	5	12(0.2-0.3)	Yes	VarmSr	6	22(0.1-0.2)	No
Loss_libc	4	16(0.1-0.2)	No	Analysts	1	26(0.1-0.2)	No
Age2	1	25(0.1-0.2)	No	Employee	0	3(0.3-1)	Yes
TQ	0	24(0.1-0.2)	No	Zmijewski	0	20(0.1-0.2)	No
Issue3	7	21(0.1-0.2)	No	Big4	0	2(0.3-1)	Yes
Bank	8	28(0.1-0.2)	No	Nonaud~1	4	4(0.3-1)	Yes
Quick_r	9	23(0.1-0.2)	No	GOC	10	9(0.3-1)	No
Intan	0	11(0.2-0.3)	Yes	Change2	5	27(0.1-0.2)	No
DRevt	11	31(0.1-0.2)	No	CI1	10	5(0.3-1)	No
Abnacc3	13	32(0-0.1)	No	Relag	0	7(0.3-1)	Yes
MA	5	30(0.1-0.2)	No	SPEC2	8	14(0.2-0.3)	No
Restruct	0	8(0.3-1)	Yes	HHI2	0	29(0.1-0.2)	No

Table 3.5 Strong variables selected based on agreement between AL and RF (further analysis)

This table reports the results of AL and RF analyses in Step 3, including corporate governance attributes. Appendix 3A provides variable definitions. AL column shows the frequency of the variable coefficient shrinking to zero in Adaptive LASSO analysis. RF column shows the variable importance ranking and importance loading in random forest analysis.

	AL	RF	Pass		AL	RF	Pass
Size_ta	0	2(0.3-1)	Yes	Return	8	36(0.1-0.2)	No
NS	0	20(0.1-0.2)	No	StdDoc	0	15(0.2-0.3)	Yes
FO_pifo	0	6(0.3-1)	Yes	VarmSr	3	23(0.1-0.2)	No
Lev_debt	0	14(0.2-0.3)	Yes	Analysts	0	17(0.2-0.3)	Yes
ROA_ebitda	6	19(0.1-0.2)	No	Employee	0	3(0.3-1)	Yes
Loss_ni	6	30(0.1-0.2)	No	Zmijewski	0	25(0.1-0.2)	No
Age2	1	22(0.1-0.2)	No	Big4	1	1(0.3-1)	Yes
TQ	0	28(0.1-0.2)	No	Nonaudit1	3	4(0.3-1)	Yes
Issue3	6	29(0.1-0.2)	No	GOC	12	10(0.2-0.3)	No
Bank	10	32(0.1-0.2)	No	Change2	5	26(0.1-0.2)	No
Quick_r	8	27(0.1-0.2)	No	CI1	12	8(0.3-1)	No
Intan	0	18(0.1-0.2)	No	Relag	0	5(0.3-1)	Yes
DRevt	13	33(0.1-0.2)	No	SPEC2	7	16(0.2-0.3)	No
Abnacc3	14	35(0.1-0.2)	No	HHI2	1	31(0.1-0.2)	No
Restruct	0	7(0.3-1)	Yes	Boardind	4	12(0.2-0.3)	Yes
Discon2	0	24(0.1-0.2)	No	Boardsize	3	9(0.3-1)	Yes
SPI1	1	11(0.2-0.3)	Yes	Auditsize	9	21(0.1-0.2)	No
Inst	5	13(0.2-0.3)	Yes	AFE	14	34(0-0.1)	No

Table 3.6 Fixed-effects tests for the key variables

This table reports the results of regressing audit fees on the identified key variables, with industry fixed effects, firm fixed effects, year fixed effects, and auditor fixed effects. Panel A shows the performance of 12 key variables identified in the main analysis. Panel B shows the performance of 14 key variables identified in the further analysis including corporate governance attributes. Appendix 3A provides variable definitions. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the *t*-statistic is reported in the parentheses.

Panel A: 12 key variables identified in the main analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF
Size_ta	0.414*** (227.71)	0.419*** (177.69)	0.454*** (81.11)	0.351*** (206.90)	0.347*** (160.32)	0.241*** (50.47)	0.337*** (156.14)	0.236*** (49.93)
FO_pifo	0.433*** (83.69)	0.367*** (65.97)	0.354*** (38.50)	0.352*** (76.66)	0.281*** (57.19)	0.116*** (17.10)	0.274*** (56.29)	0.114*** (16.80)
Lev_debt	-0.042*** (-4.00)	0.059*** (5.37)	0.078*** (4.99)	-0.021** (-2.30)	0.079*** (8.11)	0.083*** (6.68)	0.084*** (8.72)	0.086*** (7.01)
ROA_ebitda	-0.603*** (-48.10)	-0.584*** (-46.34)	-0.457*** (-27.16)	-0.495*** (-42.94)	-0.475*** (-41.18)	-0.282*** (-19.88)	-0.491*** (-43.17)	-0.278*** (-19.84)
Intan	0.303*** (25.06)	0.170*** (13.04)	0.180*** (7.90)	0.204*** (18.75)	0.074*** (6.35)	0.046** (2.55)	0.084*** (7.33)	0.050*** (2.83)
Restruct	0.198*** (34.43)	0.175*** (30.26)	0.113*** (20.21)	0.196*** (38.81)	0.169*** (33.67)	0.050*** (12.47)	0.163*** (32.73)	0.048*** (12.02)
SPI1	0.161*** (28.55)	0.161*** (29.04)	0.080*** (16.87)	0.134*** (26.58)	0.137*** (27.82)	0.046*** (12.80)	0.134*** (27.65)	0.045*** (12.66)
Stddoc	0.362*** (15.34)	0.303*** (12.83)	-0.006 (-0.21)	0.357*** (16.33)	0.303*** (13.91)	0.062*** (2.58)	0.350*** (16.27)	0.074*** (3.10)
Employee	0.093*** (49.63)	0.109*** (45.60)	0.074*** (13.20)	0.115*** (68.34)	0.140*** (64.74)	0.165*** (34.89)	0.140*** (65.47)	0.164*** (34.97)
Nonaudit1	-0.001** (-2.16)	-0.004*** (-6.75)	-0.015*** (-21.27)	0.014*** (24.18)	0.012*** (19.96)	-0.000 (-0.69)	0.012*** (21.07)	-0.000 (-0.05)
Big4	0.128*** (20.13)	0.114*** (17.85)	-0.043*** (-4.41)	0.327*** (53.83)	0.324*** (53.25)	0.336*** (40.55)	0.000 (.)	0.000 (.)
Relag	-0.002*** (-15.82)	-0.002*** (-15.23)	-0.001*** (-8.80)	-0.000*** (-2.72)	-0.000** (-1.98)	0.001*** (12.43)	-0.000 (-0.21)	0.001*** (12.98)
Industry FE	Yes				Yes	Yes		
Firm FE			Yes			Yes	Yes	
Year FE			Yes		Yes	Yes	Yes	Yes
Auditor FE							Yes	Yes
N	85574	85574	84372	85574	85574	84372	85572	84370
Adj_R ²	0.807	0.814	0.904	0.849	0.857	0.943	0.860	0.944

Panel B: 14 key variables identified in the further analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF
Size_ta	0.387*** (122.41)	0.392*** (92.99)	0.502*** (54.65)	0.370*** (135.44)	0.361*** (102.27)	0.258*** (35.26)	0.357*** (103.00)	0.255*** (34.84)
FO_pifo	0.460*** (71.55)	0.380*** (52.65)	0.285*** (24.56)	0.402*** (73.13)	0.296*** (49.12)	0.103*** (12.52)	0.292*** (49.15)	0.104*** (12.65)
Lev_debt	0.018 (1.23)	0.101*** (6.34)	0.118*** (5.23)	-0.025* (-1.94)	0.056*** (4.09)	0.115*** (7.04)	0.061*** (4.51)	0.117*** (7.26)
Restruct	0.153*** (22.25)	0.116*** (16.78)	0.091*** (14.26)	0.158*** (26.81)	0.112*** (19.20)	0.039*** (8.82)	0.110*** (18.99)	0.039*** (8.83)
SPI1	0.157*** (20.57)	0.148*** (19.71)	0.070*** (11.38)	0.112*** (17.43)	0.104*** (16.85)	0.041*** (9.81)	0.103*** (16.95)	0.041*** (9.78)
Inst	0.008 (0.81)	0.014 (1.38)	0.199*** (18.61)	-0.046*** (-4.27)	-0.040*** (-3.74)	-0.006 (-0.55)	-0.071*** (-6.79)	-0.008 (-0.78)
Stddev	0.800*** (20.91)	0.711*** (18.19)	-0.027 (-0.54)	0.781*** (23.03)	0.691*** (20.28)	0.113*** (3.16)	0.732*** (22.09)	0.133*** (3.74)
Analysts	-0.074*** (-13.57)	-0.080*** (-13.69)	-0.109*** (-14.45)	-0.029*** (-6.29)	-0.028*** (-5.69)	-0.057*** (-10.19)	-0.026*** (-5.46)	-0.055*** (-9.82)
Employee	0.061*** (24.07)	0.085*** (24.63)	0.056*** (6.28)	0.092*** (42.38)	0.137*** (46.08)	0.176*** (24.47)	0.136*** (46.65)	0.176*** (24.37)
Big4	0.169*** (18.63)	0.160*** (17.65)	0.153*** (8.88)	0.332*** (39.36)	0.330*** (39.67)	0.295*** (20.85)	0.000 (.)	0.000 (.)
Nonaudit1	-0.003*** (-3.36)	-0.007*** (-8.40)	-0.014*** (-15.80)	0.012*** (16.13)	0.008*** (10.84)	0.000 (0.73)	0.009*** (11.80)	0.001 (0.94)
Relag	-0.314*** (-13.21)	-0.316*** (-13.35)	-0.340*** (-12.34)	0.376*** (18.71)	0.414*** (20.72)	0.380*** (18.19)	0.428*** (21.44)	0.383*** (18.34)
Boardind	1.087*** (28.57)	1.110*** (29.64)	1.545*** (34.96)	0.391*** (12.33)	0.350*** (11.38)	0.200*** (6.47)	0.323*** (10.68)	0.192*** (6.23)
Boardsize	-0.008 (-0.48)	-0.000 (-0.01)	0.007 (0.29)	0.080*** (5.64)	0.094*** (6.71)	0.041*** (2.67)	0.092*** (6.67)	0.041*** (2.66)
Industry FE	Yes				Yes		Yes	
Firm FE			Yes			Yes		Yes
Year FE				Yes	Yes	Yes	Yes	Yes
Auditor FE							Yes	Yes
N	44941	44941	44460	44941	44941	44460	44939	44458
Adj_R ²	0.725	0.741	0.871	0.804	0.822	0.936	0.827	0.936

Table 3.7 Fixed-effects tests for the key variables (standardized)

This table reports the results of regressing audit fees on the identified key variables using standardized regression, with industry fixed effects, firm fixed effects, year fixed effects, and auditor fixed effects. Panel A shows the performance of 12 key variables identified in the main analysis. Panel B shows the performance of 14 key variables identified in the further analysis. Appendix 3A provides variable definitions. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the *t*-statistic is reported in the parentheses.

Panel A: 12 key variables identified in the main analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF
Size_ta	0.726*** (227.71)	0.735*** (177.69)	0.797*** (81.11)	0.616*** (206.90)	0.608*** (160.32)	0.423*** (50.47)	0.591*** (156.14)	0.414*** (49.93)
FO_pifo	0.278*** (83.69)	0.236*** (65.97)	0.227*** (38.50)	0.226*** (76.66)	0.180*** (57.19)	0.075*** (17.10)	0.176*** (56.29)	0.073*** (16.80)
Lev_debt	-0.006*** (-4.00)	0.009*** (5.37)	0.012*** (4.99)	-0.003** (-2.30)	0.012*** (8.11)	0.013*** (6.68)	0.013*** (8.72)	0.013*** (7.01)
ROA_ebitda	-0.113*** (-48.10)	-0.110*** (-46.34)	-0.086*** (-27.16)	-0.093*** (-42.94)	-0.089*** (-41.18)	-0.053*** (-19.88)	-0.092*** (-43.17)	-0.052*** (-19.84)
Intan	0.039*** (25.06)	0.022*** (13.04)	0.023*** (7.90)	0.026*** (18.75)	0.009*** (6.35)	0.006** (2.55)	0.011*** (7.33)	0.006*** (2.83)
Restruct	0.127*** (34.43)	0.112*** (30.26)	0.072*** (20.21)	0.126*** (38.81)	0.109*** (33.67)	0.032*** (12.47)	0.105*** (32.73)	0.031*** (12.02)
SPI1	0.103*** (28.55)	0.104*** (29.04)	0.051*** (16.87)	0.086*** (26.58)	0.088*** (27.82)	0.030*** (12.80)	0.086*** (27.65)	0.029*** (12.66)
Stddoc	0.037*** (15.34)	0.031*** (12.83)	-0.001 (-0.21)	0.037*** (16.33)	0.031*** (13.91)	0.006*** (2.58)	0.036*** (16.27)	0.008*** (3.10)
Employee	0.146*** (49.63)	0.172*** (45.60)	0.116*** (13.20)	0.181*** (68.34)	0.220*** (64.74)	0.261*** (34.89)	0.221*** (65.47)	0.257*** (34.97)
Nonaudit1	-0.004** (-2.16)	-0.012*** (-6.75)	-0.043*** (-21.27)	0.042*** (24.18)	0.034*** (19.96)	-0.001 (-0.69)	0.035*** (21.07)	-0.000 (-0.05)
Big4	0.082*** (20.13)	0.073*** (17.85)	-0.027*** (-4.41)	0.210*** (53.83)	0.208*** (53.25)	0.216*** (40.55)	0.000 (.)	0.000 (.)
Relag	-0.045*** (-15.82)	-0.044*** (-15.23)	-0.028*** (-8.80)	-0.007*** (-2.72)	-0.005** (-1.98)	0.032*** (12.43)	-0.001 (-0.21)	0.034*** (12.98)
Industry FE	Yes				Yes		Yes	
Firm FE	Yes				Yes		Yes	
Year FE	Yes				Yes		Yes	
Auditor FE							Yes	
N	85574	85574	84372	85574	85574	84372	85572	84370
Adj_R ²	0.807	0.814	0.904	0.849	0.857	0.943	0.860	0.944

Panel B: 14 key variables identified in the further analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF	lnAF
Size_ta	0.679*** (122.41)	0.688*** (92.99)	0.881*** (54.65)	0.649*** (135.44)	0.633*** (102.27)	0.453*** (35.26)	0.627*** (103.00)	0.448*** (34.84)
FO_pifo	0.295*** (71.55)	0.244*** (52.65)	0.183*** (24.56)	0.258*** (73.13)	0.190*** (49.12)	0.066*** (12.52)	0.188*** (49.15)	0.067*** (12.65)
Lev_debt	0.003 (1.23)	0.015*** (6.34)	0.018*** (5.23)	-0.004* (-1.94)	0.008*** (4.09)	0.017*** (7.04)	0.009*** (4.51)	0.018*** (7.26)
Restruct	0.098*** (22.25)	0.075*** (16.78)	0.059*** (14.26)	0.101*** (26.81)	0.072*** (19.20)	0.025*** (8.82)	0.070*** (18.99)	0.025*** (8.83)
SPI1	0.101*** (20.57)	0.095*** (19.71)	0.045*** (11.38)	0.072*** (17.43)	0.067*** (16.85)	0.026*** (9.81)	0.066*** (16.95)	0.026*** (9.78)
Inst	0.002 (0.81)	0.003 (1.38)	0.043*** (18.61)	-0.010*** (-4.27)	-0.009*** (-3.74)	-0.001 (-0.55)	-0.015*** (-6.79)	-0.002 (-0.78)
Stddev	0.083*** (20.91)	0.074*** (18.19)	-0.003 (-0.54)	0.081*** (23.03)	0.072*** (20.28)	0.012*** (3.16)	0.076*** (22.09)	0.014*** (3.74)
Analysts	-0.036*** (-13.57)	-0.039*** (-13.69)	-0.054*** (-14.45)	-0.014*** (-6.29)	-0.014*** (-5.69)	-0.028*** (-10.19)	-0.013*** (-5.46)	-0.027*** (-9.82)
Employee	0.096*** (24.07)	0.134*** (24.63)	0.089*** (6.28)	0.145*** (42.38)	0.215*** (46.08)	0.277*** (24.47)	0.214*** (46.65)	0.277*** (24.37)
Big4	0.109*** (18.63)	0.103*** (17.65)	0.098*** (8.88)	0.213*** (39.36)	0.212*** (39.67)	0.190*** (20.85)	0.000 (.)	0.000 (.)
Nonaudit1	-0.009*** (-3.36)	-0.021*** (-8.40)	-0.043*** (-15.80)	0.037*** (16.13)	0.024*** (10.84)	0.001 (0.73)	0.026*** (11.80)	0.002 (0.94)
Relag	-0.067*** (-13.21)	-0.068*** (-13.35)	-0.073*** (-12.34)	0.081*** (18.71)	0.089*** (20.72)	0.082*** (18.19)	0.092*** (21.44)	0.082*** (18.34)
Boardind	0.082*** (28.57)	0.083*** (29.64)	0.116*** (34.96)	0.029*** (12.33)	0.026*** (11.38)	0.015*** (6.47)	0.024*** (10.68)	0.014*** (6.23)
Boardsize	-0.001 (-0.48)	-0.000 (-0.01)	0.001 (0.29)	0.014*** (5.64)	0.017*** (6.71)	0.007*** (2.67)	0.016*** (6.67)	0.007*** (2.66)
Industry FE	Yes				Yes		Yes	
Firm FE			Yes			Yes		Yes
Year FE				Yes	Yes	Yes	Yes	Yes
Auditor FE							Yes	Yes
N	44941	44941	44460	44941	44941	44460	44939	44458
Adj_R ²	0.725	0.741	0.871	0.804	0.822	0.936	0.827	0.936

Table 3.8 Dividend payouts, earnings persistence, and audit fees

This table presents the replication results of the main analysis in Lawson and Wang (2016). Panel A reports the replication results of the main analysis in the original paper. Panel B reports the replication results using the 12 key variables. Panel C reports the replication results using the 14 key variables. Panels D, E and F report the replication results using standardized regression. DIVDUM is an indicator variable equal to one if a firm pays a cash dividend to common stockholders in current period. PERSIST is the coefficient for current-period earnings obtained from regression of earnings for period $t + 1$ on earnings for period t . SUSTAIN is the coefficient of variation for earnings before extraordinary items for the five-year period covering period $t - 4$ through period t . OCF_EARN is the coefficient for current-period earnings obtained from a regression of one period ahead operating cash flows on current-period earnings and current-period operating cash flows. Interaction is the interaction term between DIVDUM and PERSIST. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the t -statistic is reported in the parentheses.

Panel A: Replication of the main results in the original study

	(1)	(2)	(3)	(4)	(5)	(6)
	PERSIST	PERSIST	SUSTAIN	SUSTAIN	OCF_EARN	OCF_EARN
DIVDUM	-0.081*** (-3.89)	-0.042* (-1.96)	-0.087*** (-4.21)	-0.096*** (-4.37)	-0.089*** (-4.34)	-0.091*** (-4.18)
PERSIST	-0.071*** (-4.11)	-0.038 (-1.86)	-0.002 (-1.85)	-0.001 (-0.75)	-0.007 (-0.72)	-0.009 (-0.82)
Interaction		-0.097** (-2.99)		-0.004* (-2.24)		0.005 (0.30)
Controls	20 control variables used in the original study					
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
N	17051	17051	17051	17051	17051	17051
Adj_R ²	0.849	0.850	0.849	0.849	0.849	0.849

Panel B: Replication with the 12 key variables

	(1)	(2)	(3)	(4)	(5)	(6)
	PERSIST	PERSIST	SUSTAIN	SUSTAIN	OCF_EARN	OCF_EARN
DIVDUM	-0.058** (-2.90)	-0.032 (-1.46)	-0.061** (-3.05)	-0.063** (-2.96)	-0.064** (-3.21)	-0.063** (-3.11)
PERSIST	-0.069*** (-4.58)	-0.047** (-2.73)	-0.003*** (-3.56)	-0.003** (-3.16)	-0.001 (-0.16)	-0.000 (-0.05)
Interaction		-0.065* (-2.20)		-0.001 (-0.58)		-0.002 (-0.15)
Controls	12 key variables identified in the main analysis					
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
N	19223	19223	19223	19223	19223	19223
Adj_R ²	0.856	0.856	0.855	0.855	0.855	0.855

Panel C: Replication with the 14 key variables

	(1)	(2)	(3)	(4)	(5)	(6)
	PERSIST	PERSIST	SUSTAIN	SUSTAIN	OCF_EARN	OCF_EARN

DIVIDUM	-0.086*** (-4.40)	-0.057** (-2.37)	-0.091*** (-4.64)	-0.092*** (-4.53)	-0.093*** (-4.74)	-0.088*** (-4.29)
PERSIST	-0.072*** (-3.85)	-0.049** (-2.37)	-0.002** (-2.83)	-0.002** (-2.76)	-0.023** (-2.62)	-0.018 (-1.79)
Interaction		-0.070* (-1.97)		-0.000 (-0.16)		-0.012 (-0.68)
Controls	14 key variables identified in the further analysis					
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
N	13226	13226	13226	13226	13226	13226
Adj_R ²	0.816	0.816	0.815	0.815	0.815	0.815

Panel D: Replication of the main results in the original study (Standardized)

	(1)	(2)	(3)	(4)	(5)	(6)
	PERSIST	PERSIST	SUSTAIN	SUSTAIN	OCF_EARN	OCF_EARN
DIVIDUM	-0.052*** (-3.89)	-0.027* (-1.96)	-0.056*** (-4.21)	-0.062*** (-4.37)	-0.057*** (-4.34)	-0.058*** (-4.18)
PERSIST	-0.024*** (-4.11)	-0.013 (-1.86)	-0.006 (-1.85)	-0.002 (-0.75)	-0.004 (-0.72)	-0.005 (-0.82)
Interaction		-0.023** (-2.99)		-0.007* (-2.24)		0.002 (0.30)
Controls	20 control variables used in the original study					
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
N	17051	17051	17051	17051	17051	17051
Adj_R ²	0.849	0.850	0.849	0.849	0.849	0.849

Panel E: Replication with the 12 key variables (standardized)

	(1)	(2)	(3)	(4)	(5)	(6)
	PERSIST	PERSIST	SUSTAIN	SUSTAIN	OCF_EARN	OCF_EARN
DIVIDUM	-0.037** (-2.90)	-0.020 (-1.46)	-0.039** (-3.05)	-0.040** (-2.96)	-0.041** (-3.21)	-0.041** (-3.11)
PERSIST	-0.023*** (-4.58)	-0.016** (-2.73)	-0.010*** (-3.56)	-0.009** (-3.16)	-0.001 (-0.16)	-0.000 (-0.05)
Interaction		-0.015* (-2.20)		-0.002 (-0.58)		-0.001 (-0.15)
Controls	12 key variables identified in the main analysis					
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
N	19223	19223	19223	19223	19223	19223
Adj_R ²	0.856	0.856	0.855	0.855	0.855	0.855

Panel F: Replication with the 14 key variables (standardized)

	(1)	(2)	(3)	(4)	(5)	(6)
	PERSIST	PERSIST	SUSTAIN	SUSTAIN	OCF_EARN	OCF_EARN
DIVIDUM	-0.055*** (-4.40)	-0.036** (-2.37)	-0.059*** (-4.64)	-0.059*** (-4.53)	-0.059*** (-4.74)	-0.057*** (-4.29)

PERSIST	-0.024*** (-3.85)	-0.016** (-2.37)	-0.008** (-2.83)	-0.008** (-2.76)	-0.014** (-2.62)	-0.011 (-1.79)
Interaction		-0.016* (-1.97)		-0.000 (-0.16)		-0.005 (-0.68)
Controls	14 key variables identified in the further analysis					
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
N	13226	13226	13226	13226	13226	13226
Adj_R ²	0.816	0.816	0.815	0.815	0.815	0.815

Table 3.9 CEO Vega, CEO Delta and audit fees

This table presents the replication results of main analysis in Chen et al. (2015). Panel A reports the replication results of main analysis in the original paper. Panel B reports the replication results using standardized regression. LVOLSEN is log volatility sensitivity, defined as the dollar change in the CEO's option holdings in response to 0.01 unit change in stock return volatility. LPRCSEN is log of price sensitivity, defined as the dollar change in the CEO's stock and option holdings with regard to a 1 percent change in stock price. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the *t*-statistic is reported in the parentheses.

Panel A: Replication results of the main analysis

	Replication of original study	Replication with 12 key variables	Replication with 14 key variables
	lnAF	lnAF	LnAF
LVOLSEN	0.01*** (2.86)	0.00 (0.94)	0.00 (1.57)
LPRCSEN	-0.02** (-1.98)	-0.01 (-1.54)	-0.00 (-0.14)
Controls	20 variables	12 key variables	14 key variables
Ind FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	12955	13629	13098
Adj_R ²	0.82	0.82	0.81

Panel B: Replication results of the main analysis (standardized)

	Replication of original study	Replication with 12 key variables	Replication with 14 key variables
	lnAF	lnAF	LnAF
LVOLSEN	0.01*** (2.86)	0.00 (0.94)	0.01 (1.57)
LPRCSEN	-0.02** (-1.98)	-0.01 (-1.54)	-0.00 (-0.14)
Controls	20 variables	12 key variables	14 key variables
Ind FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	12955	13629	13098
Adj_R ²	0.82	0.82	0.81

Table 3.10 Aggressive real earning management and audit fees

This table presents the replication results of main analysis in Greiner et al. (2017). Panel A reports the replication results of main analysis in the original paper. Panel B reports the replication results using standardized regression. Q1REM is an indicator variable equal to one if the current year combined REM measure (abnormal R&D, abnormal gains, and abnormal production costs) is in the top quintile. Q1REM11 is the lagged value of Q1REM by one period. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the *t*-statistic is reported in the parentheses.

Panel A: Replication results of the main analysis

	Replication of original study	Replication with 12 key variables	Replication with 14 key variables
	lnAF	lnAF	LnAF
Q1REM	0.048*** (5.009)	0.050*** (4.937)	0.003 (0.326)
Q1REM11	0.030** (2.556)	0.027* (2.139)	0.026** (2.696)
Controls	21 variables	12 key variables	14 key variables
Ind FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	19596	24562	14440
Adj_R ²	0.859	0.854	0.805

Panel B: Replication results of the main analysis (standardized)

	Replication of original study	Replication with 12 key variables	Replication with 14 key variables
	lnAF	lnAF	LnAF
Q1REM	0.031*** (5.007)	0.032*** (4.937)	0.002 (0.328)
Q1REM11	0.020** (2.556)	0.017* (2.139)	0.017** (2.680)
Controls	21 variables	12 key variables	14 key variables
Ind FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	19596	24562	14440
Adj_R ²	0.859	0.854	0.805

Table 3.11 Ratings, short-term debt and audit fees

This table presents the replication results of main analysis in Gul and Goodwin (2010). Panel A reports the replication results of main analysis in the original paper. Panel B reports the replication results using standardized regression. DEBT3 is the proportion of debt maturing within three years after the fiscal year-end. RATING is a categorical variable equal to 1 if the firm is rated in the lowest S&P rating category through 21 if the firm is rated in the highest S&P rating category. INTER is an interaction term of DEBT3 multiplied by RATING. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the *t*-statistic is reported in the parentheses.

Panel A: Replication results of the main analysis

	Replication of original study	Replication with 12 key variables	Replication with 14 key variables
	lnAF	lnAF	LnAF
RATING	-0.04*** (-6.41)	-0.04*** (-6.74)	-0.04*** (-5.36)
DEBT3	-0.53*** (-3.82)	-0.36*** (-2.95)	-0.36* (-1.96)
INTER	0.03*** (2.67)	0.03** (2.33)	0.03* (1.87)
Controls	14 variables	12 key variables	14 key variables
Ind FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	3063	4207	3133
Adj_R ²	0.79	0.78	0.78

Panel B: Replication results of the main analysis (standardized)

	Replication of original study	Replication with 12 key variables	Replication with 14 key variables
	lnAF	lnAF	LnAF
RATING	-0.10*** (-6.41)	-0.10*** (-6.74)	-0.09*** (-5.36)
DEBT3	-0.08*** (-3.82)	-0.06*** (-2.95)	-0.06* (-1.96)
INTER	0.06*** (2.67)	0.05** (2.33)	0.05* (1.87)
Controls	14 variables	12 key variables	14 key variables
Ind FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	3063	4207	3133
Adj_R ²	0.79	0.78	0.78

Chapter 4 MACHINE LEARNING AND TAX AVOIDANCE

4.1 INTRODUCTION

What is the best way to measure corporate tax avoidance? And what factors explain corporate tax avoidance behaviour? Prior literature in finance and accounting has identified not only a wide range of measures of corporate tax avoidance, but also a large number of tax avoidance covariates (Hanlon and Heitzman, 2010; Wilde and Wilson, 2018; Bruhne and Jacob, 2019). Although more correlations have been identified with novel data and larger samples, a systematic quantitative assessment of this vast literature is wanting, which, due to the lack of agreement on a unique tax avoidance measure and a robust baseline model, raises concerns over p-hacking (Harvey, 2017). In reviewing empirical tax avoidance studies published in leading finance and accounting journals from 2000–2019, this study aims to provide a systematic evaluation of widely used covariates for a number of popular tax avoidance measures, with a view to selecting robust variables for tax avoidance models.

Corporate tax avoidance is a complicated concept to define and measure. Researchers in economics, law, finance, and accounting view tax behaviours through different lenses and therefore use different language to describe them (Hanlon and Heitzman, 2010). Furthermore, unlike other aspects of corporate performance, tax avoidance¹⁶ is by nature hidden, and thus the detection of this intentionally undisclosed behaviour is inevitably challenging and likely imperfect. Therefore, empirical finance and accounting research has witnessed diverse and evolving measurements for tax avoidance. Specifically, the corporate effective tax rate (ETR) using cash taxes paid or income tax expenses is the most widely cited measure for tax avoidance,

¹⁶ For brevity, ‘tax avoidance’ in this study refers to corporate tax avoidance and not personal tax avoidance.

followed by book-tax-difference-based (BTD) measures. Although some are more popular than others, each measure has its strengths and weaknesses.

In addition to the lack of unanimous agreement on a single tax avoidance measure, the growing number of incremental variables perpetuates the challenge for empirical tax avoidance research. On the one hand, more than 300 variables have been used on the right-hand side of different tax avoidance models¹⁷ in empirical studies, making it unrealistic for subsequent research to consider all covariates. On the other hand, owing to the lack of review studies, there is no wide agreement on what factors are essential for which tax avoidance measure, resulting in an arbitrary selection of control variables. Without a systematic review or a widely applied robust baseline model, the diverse measures for tax avoidance and the large number of under-investigated covariates not only confuse subsequent research by inconsistently modelling tax avoidance, but also provide opportunities for p-hacking—after multiple testings with different tax avoidance measures and covariates, p-hackers tune their models with carefully appointed dependent variables and controls to produce statistically significant results (Harvey, 2017).

In response to the deficiencies in tax avoidance research, this study aims to provide a comprehensive evaluation of extant tax avoidance covariates with two popular variable-selection-driven machine learning techniques—LASSO and Random Forest (RF). With great potential for high-dimensional analysis, machine learning techniques have been applied in finance and accounting research in recent years to reinvestigate some well-explored topics (Bao et al., 2020; Freyberger et al., 2020; Gu et al., 2020). The most popular, conventional method for variable performance assessment, ordinary least squares (OLS) regression, leads to

¹⁷ Tax avoidance models refer to different tax avoidance measures; model and measure are used interchangeably throughout this study.

overfitting when the number of variables becomes too large. Another widely used variable review method in management research—meta-analysis—suffers from irreproducibility with its synthesized conclusion. Machine learning techniques have advantages over traditional approaches with their unique quantitative-driven logic. A typical machine learning technique starts with a subsample for algorithm training, and then uses a validation subsample to select the best algorithm, followed by reassessment using another hold-out subsample. This training–learning–evaluating approach provides robust inferences with low out-of-sample errors (Mullainathan and Spiess, 2017). Considering both robustness and popularity, I applied two machine learning techniques with two different approaches—LASSO and RF—for the analysis. Proved to be vigorous methods for variable selection, LASSO and RF are widely used in natural science and recently business economics research to detect patterns from big data and select robust factors (Zhang and Huang, 2008; Meinshausen and Yu, 2009; Varian, 2014; Athey and Imbens, 2017; Reeb and Zhao, 2018; Anand et al., 2019; Amini et al., 2020).

After a review of empirical tax avoidance studies published in 12 prestigious finance and accounting journals¹⁸ from 2000–2019, I identified more than 300 tax avoidance covariates from 56 studies. Considering popularity and data access, I narrowed these down to 87 variables for the analysis. Using machine learning techniques, I tested those 87 variables against each other using the four most widely cited tax avoidance measures, namely annual cash taxes paid (cash ETR hereafter), annual book-tax expense (GAAP ETR hereafter), long-run cash ETR, and long-run GAAP ETR. In addition to these ETR measures, I repeated the analysis using four

¹⁸ *Journal of Accounting Research*, *Journal of Accounting and Economics*, *The Accounting Review*, *Contemporary Accounting Review*, *Review of Accounting Studies*, *Accounting, Organization and Society*, *Journal of Finance*, *Journal of Financial Economics*, *Review of Financial Studies*, *Review of Finance*, *Journal of Quantitative Financial Analysis*, and *Journal of Corporate Finance*.

alternative tax avoidance measures, namely unrecognized tax benefit (UTB), Frank et al.'s (2009) discretionary permanent book-tax difference (DTAX), Desai and Dharmapala's (2006) discretionary book-tax difference (DDBTD hereafter), and Manzon and Plesko's (2002) book-tax difference (MPBTD hereafter). Thirteen out of 26 variables were found by both machine learning techniques to be robust factors for popular tax avoidance measures. Among those 13 key variables, variables that proxy for ROA, new investment, and foreign operation were the strongest performers, and were thus selected by all tax avoidance models. In order to verify the validity of those key variables, I tested them with commonly used fixed effects.

In response to calls for more replication studies (Harvey, 2017; Hou et al., 2020), I revisited three empirical tax avoidance papers (i.e., Kubick et al., 2015; Bauer, 2016; Davis et al., 2016) published in leading journals and replicated their main results. Out of concern for robustness, I used both the original sample period and the extended full sample period where data were available. Although the main results in those studies were able to be successfully replicated, some findings proved sensitive to the control of identified key variables. Once the robust variables were considered, some identified correlations lost statistical significance, and some even flipped the sign, thus contradicting the original findings.

In addition to demonstrating the importance of controlling robust variables, I extend the discussions in Bruhne and Jacob (2019) and provide a comparison between the theoretical predictions and empirical outcomes of the key variables. The comparison shows that theory and empirical evidence are fully consistent in only a number of tax avoidance covariates, reflecting the challenging complexity of empirical tax avoidance research and an insufficient understanding of corporate tax decisions (Hanlon and Heitzman, 2010; Wilde and Wilson, 2018).

This study contributes to the extant literature in three ways. Firstly, to the best of my knowledge, this paper is among the first to introduce machine learning techniques into tax avoidance research. Specializing in detecting latent patterns from big data, machine learning techniques help to identify strong performers from a vast number of tax avoidance covariates. Offering a wide coverage of popular tax avoidance measures and public data access, the identified key variables provide some reliable baseline models as a starting point for future researchers to consider.

Secondly, this study fills the research gap of a systematic quantitative assessment of the large number of tax avoidance covariates which the literature has accumulated over the years. Bruhne and Jacob (2019) provide a comprehensive review of tax avoidance determinants with meta-analysis. Using quantitative-driven machine learning techniques and a large sample, the present study differs from Bruhne and Jacob (2019) by evaluating both the right-hand-side covariates and the left-hand-side tax avoidance measures. Comparing the theoretical predictions and empirical outcomes of the key variables, this study bridges empirical evidence and prior theoretical works.

Lastly, I empirically demonstrate that more robust factors are required to advance our understanding of tax avoidance. The mixed replication results suggest that some identified findings on tax avoidance are likely misidentifications due to the limited methodology and the lack of robust baseline models. The results also suggest that subsequent research should be more cautious about identifying additional incremental variables, which should be supported by relative theories, robust methods, and intuitive arguments, especially with regard to the choice of both tax avoidance measures and controls, that are closely related to the investigated research

context. Only by demonstrating such discretion can future empirical research advance our knowledge of corporate tax avoidance behaviour.

4.2 LITERATURE REVIEW

4.2.1 Measures for tax avoidance

What is tax avoidance? A heated topic, tax avoidance is widely studied by researchers in the fields of law, economics, finance and accounting. However, scholars in different fields investigate tax behaviour with different methods and with a different focus, which makes it difficult to reach a universally accepted definition of, or measurement for, tax avoidance. For example, economists and lawyers are more concerned with the legality of tax structures and make a strong distinction between tax avoidance and tax evasion—the former is classified as legal tax planning while the latter is considered illegal (Weisbach, 2003). Meanwhile, finance and accounting research focuses on the real activities undertaken to reduce taxes and increase induced tax benefits. The legality of tax avoidance behaviour, however, is of secondary interest because most of the studied tax-transaction behaviours are technically legal. For instance, it is inevitable for studies on tax shelters to acknowledge the difficulty of defending the legality (or illegality) of the transaction. Therefore, tax-avoidance transactions usually capture both certain and uncertain tax positions that may or may not be judged illegal (Hanlon and Heitzman, 2010). To provide a wide coverage of tax avoidance behaviours with a focus on the drivers of those behaviours, this study follows Hanlon and Heitzman (2010) and broadly defines tax avoidance as the reduction of explicit taxes, covering various terms that describe all kinds of tax-reduction-oriented strategies (e.g., ‘aggressiveness’, ‘noncompliance’, ‘evasions’, ‘sheltering’, etc.) used in finance and accounting studies.

Just as it is difficult to find a universally accepted definition of tax avoidance across different research fields, there is also a lack of agreement on a singular measurement for tax avoidance. Most tax avoidance measures are constructed based on financial statement data and each of them has its strengths and weaknesses. Extant literature records more than 20 different measurements for tax avoidance, with some more popular than others. The most commonly used measures are ETR-based proxies, followed by book-tax differences and discretionary measures, and UTB.

Constructed as cash taxes paid or income tax expenses divided by pre-tax income, cash ETR and GAAP ETR¹⁹ are the most widely used measures of tax avoidance. Dyreng et al. (2008) propose a long-run cash ETR measure, estimated as the sum of cash taxes paid divided by the sum of pre-tax income (net of special items) over 10 years, to overcome year-to-year volatility in annual ETR and the mismatch between cash taxes and earnings. The long-run ETR measure has also been adopted by other studies that use different time-length scales (Baderscher et al., 2013; Donohoe, 2015; Cen et al., 2017) or GAAP ETR (Brown and Drake, 2014; Law and Mills, 2015; Armstrong et al., 2019). Although ETR measures are straightforward, showing all transactions that affect firms' explicit tax liability, they cannot capture implicit taxes, nor can they reflect so-called conforming tax avoidance (Scholes and Wolfson, 1992; Hanlon and Heitzman, 2010).

Estimated as the difference between book and taxable incomes, BTB is another widely cited measure of tax avoidance (McGuire et al., 2012; Higgins et al., 2015; Bird and Karolyi,

¹⁹ GAAP ETR is defined as total income tax expense divided by pre-tax income with or without subtracting special items. Cash ETR is measured as cash taxes paid divided by pre-tax income with or without subtracting special items.

2017). Compared to ETR, BTD only captures non-conforming tax avoidance, and thus caution must be exercised when interpreting its implications (Shevlin, 2001; Desai, 2003). From a discretionary perspective, Desai and Dharmapala (2006) propose another tax avoidance measure using the residual of regressing total book-tax differences on total accruals. Similarly, Frank et al. (2009) estimate the discretionary portion of the gap between the discretionary and statutory tax rates multiplied by pre-tax income (DTAX hereafter). Those discretionary models attempt to remove determinants unrelated to tax avoidance by only capturing the tax-avoidance-driven part with regression-based partitions. Such a difference-based approach is less straightforward and likely to introduce unwanted noise. The level of or change in UTB is another tax avoidance measure. UTB is the difference between the tax return and the recognized benefits in the financial report, thus larger UTB suggests more uncertainty in the tax position and signals higher likelihood of tax avoidance engagement. The interpretation of UTB, however, also requires caution because the amount of UTB recorded in the financial report is subject to management judgement (Hanlon and Heitzman, 2010). Firms guilty of tax-sheltering transactions are often investigated by studies in order to develop a tax avoidance detection model. For example, Wilson (2009) and Lisowsky (2010) develop tax shelter models by exploring the links between firms charged with tax-sheltering activities and their financial attributes. Identified tax shelter participants provide a unique sample for tax avoidance research, but the tax-sheltering transaction is likely endogenous and fails to capture overall avoidance behaviour with only one singular transaction (Hanlon and Heitzman, 2010). In summary, extant literature has recorded a wide range of measures for tax avoidance, but none provides full coverage of tax avoidance engagement.

4.2.2 Determinants of tax avoidance

In addition to the rich variety of tax avoidance measures, extant literature has also recorded a vast number of tax avoidance covariates. The following section provides a brief review of some of the most widely considered covariates in empirical tax avoidance research.

Firm size. Size is usually measured as (1) the total assets, (2) sales, or (3) market value of the firm. Owing to better trained employees and the ability to implement more advanced techniques, larger firms tend to have a larger tax base, which increases the marginal benefit of tax avoidance (Oi, 1983; Idson and Oi, 1999). Political power theory argues that larger companies have stronger political power and more resources to engage in lobbying activities, therefore they have advantages in negotiating and bargaining for more lenient tax treatments compared to smaller firms, resulting in a higher incentive to participate in tax avoidance (Siegfried, 1972; Hill et al., 2013; Kim and Zhang, 2016). However, larger firms also have more public exposure and bear higher political costs than smaller firms (Zimmerman, 1983; Omer et al., 1993). Large firms are usually under the spotlight from regulators and the media, and this public visibility may constrain them from misbehaving, thereby lowering tax avoidance incentives (Aichian and Kessel, 1962; Watts and Zimmerman, 1978; Wong, 1988). Overall, the theories present mixed arguments about the effect of firm size on tax avoidance.

Performance. The impact of performance on tax avoidance is complex. On the one hand, firms with good performance (e.g., high profitability) may attract attention and scrutiny from investors and regulators (Bozanic et al., 2017), which discourages firms from engaging in tax avoidance. On the other hand, higher profitability leads to tax base expansion, making tax avoidance more appealing to firms. Firms with bad performance, especially those with losses, are less

incentivized to engage in tax avoidance if the tax system allows them to offset the losses with profits from other periods (Bethmann et al., 2018). Empirical studies have applied various measures for firm performance, such as return on assets (ROA), earnings before interest and taxes (EBIT), operating cash flow, net operating loss (NOL), and stock return. As firms with losses tend to have insufficient tax-avoidance motives, many studies choose to exclude loss firms from their sample (Gupta and Mills, 2002; Dyreng et al., 2010; Rego and Wilson, 2012; Hope et al., 2013; McGuire et al., 2014; Chen et al., 2019).

Growth opportunities. Growing firms are more likely to engage in tax avoidance for two main reasons. Firstly, growing firms have a greater flexibility with regard to asset allocation compared to established firms, and thus are able to allocate more investments to tax-favoured assets (Chen et al., 2010). Secondly, fast-growing firms, especially those in the early stages of development, are often financially constrained and need funding for new investments (Arditti and Pinkerton, 1978; Lang et al., 1996). Therefore, those growing firms may engage in tax avoidance to raise funds internally. Widely cited measures for growth opportunities include the market-to-book ratio, Tobin's Q, capital expenditure, and sales growth.

Asset tangibility. Many tax systems limit the deductibility of financing costs, and this rule also restricts the tax deductibility of both tangible and intangible assets (Boadway and Bruce, 1984). As opposed to tangible assets (e.g., equipment and machinery) that are less mobile and have physical substance, intangible assets, with their high mobility, provide more income-shifting opportunities, making tax avoidance less costly (Klassen and Laplante, 2012; DeSimone et al., 2019). Some intangible investments, such as research and development (R&D) expenses, enjoy tax credits in some tax systems, and thus reduce some tax burdens for firms. Therefore, empirical

studies have widely considered asset tangibility as an important control, using measures such as property, plant, and equipment (PPE), inventory, R&D, and intangible assets.

Capital structure. Given the tax deductibility of interest expense, a firm is expected to use more debts when tax rates increase (Modigliani and Miller, 1963; DeAngelo and Masulis, 1980). Prior works document the substitution effect of debt financing on tax avoidance engagement (Graham et al., 1998; Graham and Tucker, 2006), making leverage a popular control in empirical studies. On the equity side, tax avoidance engagement facilitates rent-seeking diversion and managerial opportunism, which result in a value transfer from local government to company shareholders (Desai and Dharmapala, 2009). Therefore, institutional investors may use their power and pursue activities that will benefit themselves at the expense of governance quality and other investors' interests (Desai and Dharmapala, 2006; Cheng et al., 2012).

Financial constraints. Because tax avoidance could be an alternative internal funding tool when firms face higher cost to external funding, empirical studies have documented that financially constrained firms tend to conduct more aggressive tax planning when external funding from the capital market is more expensive (Law and Mills, 2015; Edwards et al., 2016). For multinational firms, when foreign income could be repatriated to relieve domestic financing constraints, some commonly used tax avoidance strategies, such as cross-border income shifting, become less enticing because of the inherent tax costs involved (Foley et al., 2007; Hanlon et al., 2015; Dyreng and Markle, 2016).

Corporate complexity. Many prior works have included measures for foreign operation, number of segments, and equity in earnings of unconsolidated affiliates to control corporate complexity in tax avoidance models (Gupta and Mills, 2002; Armstrong et al., 2012; Hope et al., 2013;

Donohoe, 2015; Chen and Lin, 2017). In terms of business range, internationality, and affiliation, firms with higher complexity may have more channels and opportunities for tax avoidance engagement than firms with lower complexity (Armstrong et al., 2012). More complex firms also tend to develop less transparent information environments (Balakrishnan et al., 2019), which lower the cost of tax avoidance and thus incentivize tax avoidance engagement. Among various corporate complexity factors, foreign operation has been widely investigated by tax avoidance studies. Compared to domestic firms, multinational firms have opportunities to shift their income from high-tax jurisdictions to low-tax ones, and thus have a higher incentive to engage in tax avoidance (Dharmapala and Riedel, 2013; Dharmapala, 2014; Dyreng and Markle, 2016).

Information environment. In a more transparent information environment (e.g., one in which there is high-level analyst coverage), information dissemination breadth is wider and thus lowers monitoring costs for regulators and investors (Jensen and Meckling, 1976). Therefore, a more transparent information environment discourages firms from tax avoidance engagement by increasing the costs of tax avoidance engagement, resulting in a higher chance of exposure and extensive scrutiny from the market and government (Chen et al., 2018).

Corporate governance. In addition to conventional firm attributes, an increasing number of studies are exploring the effect of corporate governance features on tax avoidance (Wilde and Wilson, 2018). Board composition, management compensation, and executive personal characteristics have been widely investigated by empirical studies. With regard to the impact of board characteristics on tax avoidance, mixed results and continuing debates over theory mean that extant literature has yet to find conclusive evidence (Rego and Wilson, 2012; Armstrong et

al., 2015; Chi et al., 2017). Designed to align the interests of managers and shareholders, high-powered compensation incentivizes managers to change their rent-extraction strategies and results in more tax avoidance (Desai and Dharmapala, 2006). Studies have used various compensation proxies, such as stock compensation or cash compensation, as model controls. As the main decision-maker for corporate policy, the chief executive officer (CEO) is found to affect corporate tax activities (Rego and Wilson, 2012; Dyreng et al., 2010). Some widely used measures for executive personal attributes include gender and age. To sum up, although theories and empirical results present mixed arguments, corporate governance attributes provide a novel perspective from which to investigate tax avoidance.

Industry and competition. Firms in different industries have different business environments (e.g., different tax policies, competition levels, and productivity levels), which result in different costs and benefits of tax avoidance engagement (Acemoglu and Zilibotti, 2001; Kubick et al., 2015; Dyreng et al., 2019). For example, firms in new-energy industries are likely to enjoy lenient tax schemes from the government, compared to firms in the alcohol and tobacco industries, which are heavily taxed. Also, firms in high-tech industries tend to use more advanced technologies and face more growth opportunities than their counterparts in traditional industries such as agriculture and manufacture. Therefore, empirical studies have widely applied industry fixed effects to control industry heterogeneity.

After a review of empirical tax avoidance studies published from 2000–2019 in the 12 leading journals noted above, I identified more than 300 covariates that are included on the right-hand side of different tax avoidance models. A detailed examination of those covariates revealed two problems. Firstly, different studies employ different measures for the same proxy in their

models,²⁰ but present few, if any, strong arguments or related theories to support their choice. Secondly, there is a distinct lack of agreement on robust control variables. As some variables are more popular than others, the way most studies select control variables seems arbitrary, with little or no justification for their choices. Although different research questions may require different control variables for specific contexts, the selection of control variables should be supported with intuitive arguments and relative theories. Those two arbitrary-selection problems are perpetuated by multiple measures for tax avoidance, providing p-hackers with opportunities to produce fake-positive results by only reporting the best results from multiple testings of various tax avoidance measures and covariates.

With an increasing number of identified incremental variables, empirical tax avoidance research faces two critical challenges. Firstly, it is unrealistic to include all covariates in the analysis. Compared to other research fields such as asset pricing where some robust baseline models are available (e.g., the Fama–French 3-factor model, the q-model), empirical tax avoidance studies need to find a robust benchmark model in order to deal with p-hacking via arbitrary selection of control variables. Secondly, as opposed to the large number of empirical studies investigating new incremental variables, there is no study that quantitatively evaluates this vast number of covariates.

Aiming to tackle those challenges, this paper provides a comprehensive review of the tax avoidance literature, and introduces two powerful machine learning techniques to select robust variables for a tax avoidance model.

²⁰ For example, as a widely used control for firm performance, NOL is measured in different ways across different studies, including: (1) as a dummy variable equal to one when net operating loss carryforward at the current year is positive; (2) as a dummy variable equal to one when net operating loss carryforward is available; or (3) as the change in net operating loss carryforward divided by total assets. Although those three NOL measures all proxy for firm loss, they each have a different focus with different scales and coverage.

4.2.3 Machine learning techniques

Compared to conventional econometrics tools, machine learning has its roots in big data and provides robust out-of-sample inferences with its unique training-and-learning algorithm (Mullainathan and Spiess, 2017). Therefore, finance and accounting research has recently witnessed the growing popularity of machine learning techniques to help answer questions relating to, for example, accounting misbehaviours (Bao et al., 2020; Bertomeu et al., 2020) and risk premium prediction (Freyberger et al., 2020; Gu et al., 2020), which traditional research methods have not always adequately addressed.

For example, machine learning has advantages over two conventional tools for variable performance review, namely ordinary least squares (OLS) regression and meta-analysis. When a vast number of variables are presented, OLS regression struggles to disentangle, compare, and analyze the explanatory power of each individual covariate, as the model is tuned to fit the sample instead of detecting the underlying patterns. Meta-analysis analyzes the extensive number of variables by combining the results from multiple studies into a pooled estimate. This qualitative-based synthesis analysis suffers from the irreproducibility and subjectivity involved in weight allocation to different studies' results (Wanous et al., 1989; Lipsey and Wilson, 2001). As a data-driven method with clear, pre-set rules, machine learning is therefore able to analyze a huge number of variables and offer replicable results.

After reviewing the application of machine learning in economics and business studies, I employed two popular and powerful variable-selection-based machine learning techniques, namely LASSO and RF, to evaluate the large number of tax avoidance covariates the literature has accumulated over the years.

LASSO stands for ‘least absolute shrinkage and selection operator’, and belongs to the regression family. The rationale behind LASSO is straightforward—it shrinks the coefficients of less important variables to zero by applying a penalty term for the model complexity. Therefore, LASSO prefers a simpler and robust model over a more complex and ‘better’ one. This preference for a small number of robust variables makes LASSO a good tool to analyze the vast number of tax avoidance covariates. Compared to other regression-based variable-selection methods, such as subset selection and ridge regression, LASSO is less sensitive to changes in data and thus results are easier to interpret (Tibshirani, 1996; Zhang and Huang, 2008; Meinshausen and Yu, 2009). I applied a modified LASSO method, adaptive LASSO (ALASSO hereafter). Compared to giving equal penalties for all variables in LASSO analysis, ALASSO uses weighted penalty terms on different variables so as to favour variables with univariate strength and to avoid spurious selection of noise variables. This modification yields consistent estimates of the parameters while retaining the attractive convexity property of LASSO (Zou, 2006; Reeb and Zhao, 2018).

The second machine learning method is RF, a classification-oriented tree-based method, famous for its stable ‘out-of-the-box’ performance and insensitivity to excessive model tuning (Athey and Imbens, 2017; Anand et al., 2019; Amini et al., 2020). RF performs repeated double bootstrapping across the subsample of training data and the subset of variables, which builds up a large number of decision trees. Then this bootstrapping procedure is evaluated according to the out-of-bag (OOB) error for all variables. Because dropping important variables results in a large variance in OOB error, by comparing the difference between OOB errors before and after variable permutation, RF ranks all variables with variable-importance scores, which are

analogous to the variables' explanatory power. In contrast to the linearity in LASSO analysis, RF allows nonlinearities and interactions (Varian, 2014), thus providing an alternative variable selection approach with a different mechanism.

ALASSO and RF have their strengths and weaknesses, and they are complementary to some degree. For example, ALASSO is based on linear regression and eliminates weak variables, while RF allows nonlinearity and shows the relative importance of all variables. The two methods are thus likely to select different sets of robust variables. Out of consideration for robustness, I compared the findings between the two methods and selected variables found by both methods to be robust.

4.3 DATA AND METHODOLOGY

4.3.1 Sample and variables

I collected the data from public databases. Specifically, Compustat and CRSP provided company accounting and financial information. From Audit Analytics I downloaded auditor and related auditing information. I collected corporate governance information from BoardEx and ExecuComp. I captured analyst following and institutional ownership information from I/B/E/S and Thomson Reuters. The main sample contains 148,370 firm-year observations which cover the main tax avoidance measures and firm attributes from 1990–2019. Because the corporate governance data coverage is smaller, I constructed a separate sample including corporate governance attributes in further analysis.

Following Dyreng et al. (2008) and Dyreng et al. (2010), I employed annual ETR measures and their long-term forms for the main analysis, using both cash ETR and GAAP ETR. ETR measures have two main advantages. Firstly, annual and long-term ETR are the most widely

used tax avoidance measures in finance and accounting research. As this study aims to provide a retrospective review of variables used in different tax avoidance models, ETR measures, given their popularity, are the preferred tax avoidance measures. Secondly, considering the large number of right-hand-side variables and the large sample period, ETR measures are easily observable as the ‘salient’ measures of tax avoidance. In addition to ETR measures I also examined four alternative tax avoidance measures—UTB, DTAX, DDBTD, and MPBTD—in further analysis.

From leading finance and accounting journals, I identified 56 US-data-based empirical studies that examine the potential drivers of corporate tax avoidance from 2000–2019. From those studies I collected more than 300 covariates used on the right-hand side of tax avoidance models. I then narrowed the selection down to 87 variables based on popularity and accessibility of data. Those 87 variables cover a wide range of firm characteristics: size, industry competition, profitability, operating cash flow, NOL carryforward, stock return, valuation, sales growth, new investment, PPE, inventory, depreciation, goodwill, R&D expenses, intangible assets, leverage, interest expenses, financing activity, institutional ownership, cash holding, financial constraints, foreign operation, segments, employee, analyst coverage and forecast, auditor, accruals, selling, general, and administrative expenses (SGA), advertisement expenses, special and extraordinary items, operating risk, return volatility, lifecycle, and corporate governance. I winsorized all continuous variables at the 1% level to reduce the impact of outliers. Appendix 4A provides detailed definitions of the variables.

4.3.2 Summary statistics

Table 4.1 presents the summary statistics of the variables. In terms of firm size, the average total assets are around \$482.99 million, and the market value is \$411.58 million with \$330.30 million in revenues. The median Herfindahl-Hirschman Index is 0.05 for the two-digit SICC industry classification and 0.16 for the four-digit SICC industry classification. On average, ROA varies between 8% and 11%, and EBIT account for 20% of net operating assets. Cash flow from operations is on average 3% and 10% of total assets, depending on whether capital expenditure is considered. About 30% of firms report positive NOL carryforward, and the average annual change in tax loss carryforward is negative. Median annual stock returns are around 12%, with a market-to-book ratio of 1.90 and Tobin's Q of 1.38. The median growth rate of sale is 9%, and around 31% of firms have experienced merger and acquisition. Capital expenditures are roughly equal to 7% of total assets. Approximately 21% to 44% of total assets are PPE, whose annual growth rate remains at 1% to 3%. Inventory is on average 13% of total assets, while depreciation and goodwill are 5% and 1%, respectively.

R&D costs are about 2% of total assets, and intangible assets are around 12% of total assets. The median leverage ratio is between 15% and 22%, with 1% interest expenses. The mean Mezzanine ratio, Altman Z-score, and Hadlock and Pierce (2010) index are 2%, 5.32, and -3.35, respectively. About 83% of firms sell either common stocks or long-term debts, and 48% of shares are under institutional ownership. The median amount of cash holding is around 8% of total assets. Foreign income equals approximately 1% of total assets and 13% of pre-tax income, with around 35% of firms reporting foreign operations. The average number of geographic segments and business segments are 1.8 and 1.7, respectively. About 17% of firms report equity

income in earnings, and the mean employee number is 1,510. The average standard deviation of analyst earnings forecasts is 0.07, with 29% of firms indicating a positive earnings surprise. The average analyst coverage is 1.8. Around 80% of firms are audited by a Big 4 auditor, and 75% of firms receive an unqualified audit opinion. The average abnormal accruals calculated following Kathori et al. (2005) and Dechow et al. (1995) are around 3% and less than 1%, respectively. Selling, general, and administrative expenses on average are around 19% of sales or 20% of total assets, compared to 1% for advertisement expenses. Special and extraordinary items scaled by total assets are less than 1%. The average standard deviation is 6% for ROA, 6% for operating cash flow, 17% for sales, 13% for monthly stock returns, and 3% for annual daily returns. Firms on average have 10.7 years of data listed on Compustat.

[Table 4.1 about here]

Turning to the main tax avoidance variables, the average annual cash ETR is 23%, compared to 24% for the three-year long-run cash ETR. The average annual GAAP ETR and the three-year long-run GAAP ETR are 27% and 26%, respectively. As an alternative tax avoidance measure, UTB is around 1% of total assets; DTAX has a mean of 6.31. The average DDBTD and MPBTD are 0.03 and 0.01, respectively.

Regarding corporate governance attributes, each board has an average of 8.4 board members, with 80% of positions assumed by independent directors. Only 3% of CEOs are female, and about half of the CEOs are also the board chair. The average CEO age is 55.8, with an average tenure of around eight years. About 38% of firms report stock compensation, and CEOs on average own about 3% of stocks. Approximate firms spend \$845,000 on cash compensation.

4.3.3 Methodology

Considering the vast number of variables and the similarities between some variables, I ran two rounds of tests to select robust variables. In the first round, I began by dividing the variables into groups based on similarities between them,²¹ and applied RF to select the strongest performer within each group. The sparsity-oriented nature of LASSO makes it a strong tool in high-dimension settings, but a less than ideal one when only a small number of variables is presented. Given the small number of variables within each group, I relied on RF, which illustrates variable importance with a relative-explanatory-power ranking, to select the strongest candidate from each group. Some groups had only one variable and thus immediately qualified for second-round analysis. In the second round, the ‘survivors’ from each group and the other independent measures were tested by both ALASSO and RF.²² The variables rated as important by both ALASSO and RF were the ‘winners’. Figure 4.1 is a flow chart of the second-round analysis.

[Figure 4.1 about here]

In order to compare the results of ALASSO and RF, I used fixated threshold with 50% as the cut-off point. This required: (1) that the variable importance loading exceeded 50% in RF analysis; and (2) that the frequency of the variable selected by ALASSO exceeded 50% of the total rolling windows. I chose a 50% fixated threshold for two main reasons. Firstly, since my study investigates eight tax avoidance measures, each of which has a unique focus, fixated threshold provides a straightforward comparison between models. Secondly, although there is a

²¹ For example, *SPI_at* and *SPI_sale* both proxy for the proportion of special items, with the former scaled by total assets and the latter scaled by sales; thus I combined them and named the group **Special Items**.

²² For ALASSO, I followed the rolling-window approach in Reeb and Zhao (2018) and ran a three-year rolling-window analysis (e.g., 1990–1992 as the first sub-sample, 1991–1993 as the second sub-sample), and counted the frequency of each variable selected by ALASSO from the total number of rolling windows.

risk of missing relevant variables, the 50% threshold provides an intuitive high bar for the large-sample analysis. Taking both machine learning techniques' results into account, variables meeting the 50% threshold should thus have robust explanatory power.²³

4.4 MAIN RESULTS

4.4.1 Step 1: Strongest variable selected from each group using RF

Following extant literature, I divided the variables into groups based on the aspect of firm performance for which they proxy.²⁴

Table 4.2 presents the results of RF analysis within the groups. Across the four ETR measures, *Size_mv* has the highest importance loading in Size. *HHI_sicc* fares better than *HHI_sic* in Competition. *ROA_iblat* dominates in Profitability, as does *DPPEGT* in PPE. *OCFf_lat* is the preferred measure for Operating Cash Flow, as is *NOLI_d* for NOL. *MTB* and *CAPX_lat* are the strongest performers in Valuation and Investment. For RND, long-run ETR measures and annual GAAP ETR prefer *RND_sale*, while annual cash ETR favours *RND_at*. *Intan_sale* has the highest loading in Intangible Assets, as does *DLTT_lat* in Leverage. With regard to Foreign Operation, long-run cash ETR and annual GAAP ETR prefer *PIFOC_at*, while annual cash ETR and long-run GAAP ETR prefer *PIFO_pi*. *NBS2* is the strongest performer in

²³ As Stoppiglia et al. (2003) point out, the choice of variable ranking threshold is problem-dependent, and a trade-off between model parsimony and explanation power is inevitable. When a high bar is chosen, the model is likely to be more parsimonious but there may be a risk of missing some relevant variables. A low bar may provide more comprehensive coverage but is also likely to include variables that are less important.

²⁴ The groupings (in bold) are as follows. **Size:** *Size_at*, *Size_mv* and *Size_sale*. **Competition:** *HHI_sic* and *HHI_sicc*. **Profitability:** *ROA_piat*, *ROA_pilat*, *ROA_pilat2*, *ROA_iblat* and *EBIT*. **Operating Cash Flow:** *OCF_avat* and *OCFf_lat*. **Operating Loss Carry Forward:** *NOLI_d* and *NOL2_d*. **Valuation:** *MTB* and *TBQ*. **Investment:** *CAPX_lat* and *CAPX_at*. **PPE:** *DPPENT*, *DPPEGT*, *PPENT_at*, *PPENT_lat* and *PPEGT_at*. **RND:** *RND_lat*, *RND_at* and *RND_sale*. **Intangible Assets:** *Intan_lat*, *Intan_at* and *Intan_sale*. **Leverage:** *DLTT_at*, *DLTT_lat* and *LEV_at*. **Foreign Operation:** *PIFOc_at*, *PIFO_lat*, *PIFO_at* and *PIFO_pi*. **Segment:** *NGS*, *NBS1* and *NBS2*. **Accruals:** *AbAccr1*, *AbAccr2* and *ToAccr*. **SGNA:** *SGA_sale* and *SGA_at*. **Special Items:** *SPI_at* and *SPI_sale*. **Advertisement Expenses:** *XAD_avat*, *XAD_sale* and *XAD_at*. **Operating Risk:** *StdROA5*, *StdCF5* and *StdSale5*. **Return Volatility:** *StdMRn5* and *StdARn*. The remaining variables are relatively unique individual measures of other aspects or in a different scale, and so immediately qualified for second-round analysis.

Segment. In Accrual, the cash ETR measures prefer *AbAccr1*, while the GAAP ETR measures prefer *AbAccr2*. *SGA_at* has the highest loading in SGNA. For Advertisement Expenses, long-run GAAP ETR favours *XAD_avat*, and the other ETR measures prefer *XAD_sale*. In Special Items, *SPI_sale* is the preferred measure across all ETR measures except for annual GAAP ETR. *StdCF5* and *StdMRn5* dominate in Operating Risk and Return Volatility, respectively. Appendix 4B provides the detailed results of Step 1 analysis.

[Table 4.2 about here]

In summary, except for RND, Foreign Operation, Accruals, Advertisement Expenses and Special Items,²⁵ the four ETR measures show high consistency in selecting the strongest performer within each group, especially for fundamental firm attributes such as size, profitability, operating cash flow, and valuation.

4.4.2 Step 2: Robust variable selected using ALASSO and RF

Next, I took the variables that survived Step 1, together with the other independent measures, into the second-round analysis using both ALASSO and RF. Table 4.3 presents the results of the Step 2 analysis, where Panels A, B, C and D summarize variable performance in the four ETR models. For example, Panel A shows that in the annual cash ETR model, *Size_mv* is selected by ALASSO 17 times out of a total of 28 rolling windows (i.e., higher than a 50% selection rate) with an importance loading between 50% and 60%; therefore *Size_mv* passes the 50% fixated threshold and is selected as a robust variable.

[Table 4.3 about here]

²⁵ In regard to the different outcomes of those five groups, a detailed examination shows that such different is trivial and uninfluential to the following analysis: the importance loading disparity between the winner and the runner-up is very small, with winner and runner-up sharing similar variable construction and distributions.

Regarding agreement between ALASSO and RF, Table 4.3, Panels A and B show that there are 24 variables in the annual cash ETR model, while only three variables met the fixated threshold in the annual GAAP ETR model. In Table 4.3, Panels C and D, there are 26 variables in the long-run cash ETR model, and 14 variables were selected in the long-run GAAP ETR model. Appendices 4D and 4E provide detailed results of ALASSO and RF analyses, respectively.

Readers are advised to interpret the results with caution. Firstly, the results echo concerns voiced by Dyreng et al. (2008) who point out that annual ETR measures suffer from year-to-year volatility and mismatching, while long-run ETR measures help alleviate such issues. In the final selected models, the disparity between the over-parsimonious annual GAAP ETR model and the relatively informative long-run GAAP ETR model signals potential problems in using annual measures as proxies for tax avoidance.

Secondly, the results raise concerns over tax avoidance models in empirical studies where a singular and relatively simple set of control variables is commonly applied across multiple models which have different tax avoidance measures as the dependent variable. My results present four distinctive robust variable sets for four ETR models, confirming that different tax avoidance measures focus on different aspects (Hanlon and Heitzman, 2010), and thus future research should consider either applying different control sets for different tax avoidance measures, or employing a comprehensive control set across different models.

As a relatively high bar, the 50% fixated threshold allows a straightforward comparison across different models, but such a clear-cut method could be tricky when many variables cluster

right below the threshold.²⁶ Therefore, I conducted another two analyses to reassess the variables' performance without compromising the selection of strong variables. Firstly, in addition to extant selected variables for annual GAAP ETR, I added another variable selection by lowering the fixated threshold to 30% in RF analysis only, while keeping the 50% fixated threshold in ALASSO analysis. This reassessment led to an additional list of 23 robust variables for annual GAAP ETR. Secondly (see section 5.2), I employed a median threshold to reassess variable performance across all models.

Although those four ETR models have different preferences, some variables are more popular than others. For example, ROA, new investment, and special items feature in all models. Size, cash flow, stock return volatility, bankruptcy, foreign operation, and SG&A fees are the next popular variables. Some variables are never selected, such as merger and acquisition, goodwill, financing activity, foreign operation indicator, segments, equity income in earnings, analyst coverage, auditor, and extraordinary items.

4.5 FURTHER ANALYSIS

4.5.1 Alternative tax avoidance proxy

Following the same procedures and standards used in the main analysis, I repeated the two-step analyses and tested the variables with four alternative tax avoidance proxies—UTB, DTAX, DDBTD, and MPBTD.²⁷ In the first-round analysis, I divided the variables into the same groups

²⁶ A closer examination of the annual GAAP ETR model reveals that the over parsimonious result is driven by a large number of variables with importance loading between 0.3 and 0.5 in RF analysis.

²⁷ Given limited data availability, the UTB sample is from 2007–2019, while the samples for DTAX, DDBTD, and MPBTD are from 1995–2019.

as in the main test, and ran the RF analysis within each group to select the strongest performer.

Appendix 4E presents the detailed results of Step 1 RF analysis.²⁸

In the second-round analysis, I tested the individual measures and survivors from Step 1, using both ALASSO and RF. Appendices 4F and 4G present the results for the ALASSO and RF analyses, respectively. Table 4.4 presents the results of alternative tax avoidance measures. Based on agreement between ALASSO and RF, 13 variables were selected by the UTB model, 15 by the DTAX model, 17 by the DDBTD model, and 17 by the MPBTD model.

[Table 4.4 about here]

Measures for ROA, new investment, R&D expenses, and foreign operation were chosen by all four models. Some variables were not selected at all, including size,²⁹ industry competition, NOL carryforward indicator, stock return, merger and acquisition, goodwill, financing activity, institutional ownership, SA index, foreign operation indicator, earnings in equity indicator, earning forecast information, auditor, advertisement expenses, and extraordinary items.

Comparing the results of the four alternative tax avoidance models to those in the main test, measures for ROA, investment, and foreign operation were the most popular across all models,

²⁸ *Size_mv* dominates in Size, as does *ROA_iblat* and *HHI_sicc* in Profitability and Competition. *PIFO_pi* and *NGS* are the strongest performers in Foreign Operation and Segment, respectively. The most important variables in Operating Cash Flow, Valuation, and Return Volatility are *OCFf_lat*, *MTB*, and *StdMRn5*, respectively, which echoes the main test. In NOL and RND, UTB and DTAX prefer *NOL1_d* and *RND_sale*, while DDBTD and MPBTD prefer *NOL2_d* and *RND_lat*. For PPE, UTB and DTAX choose *DPPEGT*, while DDBTD and MPBTD choose *PPEGT_at*. In Intangible Asset and Leverage, DTAX chooses *Intan_lat* and *DLTT_lat*, while UTB, DDBTD, and MPBTD choose *Intan_sale* and *LEV_at*. For Accruals, SGNA and Advertisement Expense, the importance loading difference between the variable in first place and that in second is marginal across four tax avoidance indexes. In Special Items and Operating Risk, UTB prefers *SPI_sale* and *StdROA5*, while DTAX, DDBTD, and MPBTD prefer *SPI_at* and *StdCF5*.

²⁹ In my opinion, the results not only reflect the mixed theoretical arguments about the effect of firm size on tax avoidance (Aichian and Kessel, 1962; Siegfried, 1972; Watts and Zimmerman, 1978; Hill et al., 2013), but also show the importance of a cross examination between two machine learning techniques due to their different focus. Therefore, variables marked as significant factors by both techniques are likely to have robust explanatory power.

followed by measures for employee, operating cash flow, inventory and depreciation, intangible assets, bankruptcy risk, special items, and operating risk. Eighteen unpopular variables were selected by none or only one model. These include net operating loss carryforward indicator, merger and acquisition, goodwill, financing activity, institutional ownership, SA index, foreign operation indicator, segments, equity in earnings, analyst coverage, earnings forecast, auditor, advertisement expense, and extraordinary items.

4.5.2 Median threshold

In considering robustness, I reassessed variable performance using an alternative standard—median threshold—which requires variable performance to be in the top 50% among all variables in both ALASSO and RF analysis. Compared to the 50% fixated threshold, median threshold is less sensitive to performance distribution, providing more consistent and stable results in cases where variables cluster right below the cut-off point or where there is a dramatic performance gap between two consecutive-ranking variables. Median threshold also compares any agreement between two techniques from a different perspective—the fewer variables selected based on median threshold, the more disagreements occur between the two methods. Median threshold considers varying performances across different models, and selects the top performers based on agreement between two machine learning techniques. Although more relaxed, this kind of floating threshold is not uncommon in machine learning research, especially in studies using different methods for feature ranking or variable selection (Clemencon et al., 2013; Petkovic et al., 2020).

[Table 4.5 about here]

Table 4.5 presents the reassessment results of ETR models using median threshold. For example, in Table 4.5, Panel A, *Size_mv* is in the top 50% among all variables in both ALASSO and RF analyses, and thus meets the median threshold. In total, median threshold shortlists 14 variables in the annual cash ETR model, 11 variables in the annual GAAP ETR model, 14 variables in the long-run cash ETR model, and 15 variables in the long-run GAAP ETR model. Table 4.6 presents the results of alternative tax avoidance models using median threshold. Median threshold selects 17 variables in the UTB model, 15 variables in the DTAX model, 16 variables in the DDBTD model, and 15 variables in the MPBTD model. The numbers of selected variables across the eight tax avoidance models are relatively consistent, suggesting wide agreement on the top 50% of performers between the two methods.

[Table 4.6 about here]

Consistent with the results from fixated threshold, measures for ROA and new investment were selected by all tax avoidance models. Measures for operating cash flow, bankruptcy risk, and special items were also popular across all models using median threshold. The unpopular variable list is comparable to that in the main analysis, and includes competition, NOL carryforward indicator, stock return, merger and acquisition, goodwill, financing activity, institutional ownership, SA index, foreign operation indicator, number of segments, equity in earnings, analyst coverage, earnings forecast, auditor, advertisement expenses, and extraordinary items.

4.5.3 Corporate governance attributes

There is an increasing trend in extant literature to include corporate governance variables in tax avoidance models. Following the same procedures and standards used in the main test, I

extended my analysis by including a number of popular corporate governance variables with public data access: board size, director independency, compensation, CEO gender, CEO age, CEO tenure, CEO stock ownership, and CEO–chair duality. Given the smaller coverage of corporate governance data, I restricted the sample years to 2007–2019 for the UTB model, and 2001–2019 for the other models. Since Step 1 analysis from the main test had already selected the strongest performer out of each group for the eight tax avoidance measures, I took the Step 1 results from the previous analyses and immediately ran Step 2 analysis with the corporate governance variables. Appendices 4H and 4I present the detailed results of the ALASSO and RF analyses.

[Table 4.7 about here]

Table 4.7 presents the results of the ETR models using both fixated and median thresholds. Measures for ROA are selected by all ETR models. Other popular variables in the main test such as cash flow, foreign operation, bankruptcy risk, SGNA fees, and special items also secure a place when corporate governance attributes are present. With regard to corporate governance variables, none is selected across four models, either using fixated threshold or median threshold.

[Table 4.8 about here]

Table 4.8 presents the results of alternative tax avoidance models using both fixated and median thresholds. Popular variables include measures for ROA, cash flow, bankruptcy risk, SGNA fees, and special items. Owing to weak performance, corporate governance attributes fail to meet either fixated or median threshold. The overall results in Tables 4.7 and 4.8 suggest that, when a comprehensive list of firm attributes is presented, the widely investigated corporate governance variables are not as robust as the literature suggests. The weak performance of

corporate governance variables across the eight different models is informative, as the tougher fixated threshold and the more lenient median threshold both reject all corporate governance attributes. The results therefore indicate that corporate governance attributes are likely second-order factors, while the most important variables for tax avoidance remain firm-level attributes.

4.5.4 Fixed-effects test

Empirical tax avoidance research has widely employed fixed effects to alleviate concerns around omitted variables. For example, the distribution of tax avoidance across different firms, industries, and years could be non-random, and thus using fixed effects not only helps mitigate the self-selection bias, but also controls firm-level, industry-level, and year-level omitted variable issues. Therefore, I tested the identified key variables with firm, industry and year fixed effects, individually and in pairs.

[Table 4.9 about here]

Table 4.9 presents the results of regressing ETR measures on the selected variables with fixed effects. Column (1) reports the regression results with no fixed effects, and columns (2), (3), and (4) apply singular industry, firm, and year fixed effects, respectively. Columns (5) and (6) use dual fixed effects. In Table 4.9, Panel A, most of the variables selected for annual cash ETR are statistically significant except for *DPPEGT*. Table 4.9, Panel B reports the results of regressing annual GAAP ETR on the selected variables with fixed effects. In Table 4.9, Panel B1, the three key variables selected by 50% fixated threshold are statistically significant across all fixed-effect tests. In Table 4.9, Panel B2, the 23 key variables, identified by the alternative threshold, also exhibit robust performance in the fixed-effect tests. In Table 4.9, Panel C, the selected variables for the long-run cash ETR model—except for *MTB*, *RND_sale* and

Intan_sale—perform well with fixed effects. In Table 4.9, Panel D, the selected variables for the long-run GAAP ETR model also show robust performances across the fixed-effect tests.

[Table 4.10 about here]

Table 4.10 presents the results of the alternative tax avoidance models with fixed effects. Table 4.10, Panel A shows that most of the variables, except for *DSale*, perform well in the UTB model. In Table 4.10, Panel B, almost all selected variables for the DTAX model are statistically significant in the fixed-effect tests. In Table 4.10, Panel C, except for *DSale* and *Intan_sale*, the performances of the selected variables for the DDBTD model are relatively consistent across the tests. In Table 4.10, Panel D, all selected variables but *CHE_at* demonstrate robust performance.

Comparing the R^2 between columns (2), (3), and (4) in Tables 4.9 and 4.10, the R^2 with firm fixed effects is significantly higher than that with year or industry fixed effects, suggesting that tax avoidance is largely driven by firm attributes. Given the overall low R^2 across models, the results also indicate that the variables used by prior studies have relatively limited explanatory power over popular tax avoidance measures, and thus more robust firm-level variables are needed to enrich extant models.

4.5.5 Replication of prior studies

In response to calls for more replication studies (Harvey, 2017; Hou et al., 2020), I replicated three tax avoidance studies³⁰ and demonstrate the importance of controlling robust variables. I chose the studies for two main reasons. Firstly, although they are published in leading journals, the incremental variables identified by those studies are rarely cited by subsequent tax avoidance

³⁰ Bauer (2016), Davis et al. (2016), and Kubick et al. (2015).

research. Secondly, those studies use public databases, which facilitate replication. Closely following the sample and variable construction described in the original studies, I began by replicating the main results with the same control variables employed by the authors. Next, I reassessed the incremental variable using the key variables identified in the main analysis as controls. Considering robustness, I replicated the results using two samples: (1) the sample used in the original studies; and (2) the extended sample with maximum coverage of available data. In order to achieve comparable results, I employed the same set of fixed effects and error clustering used in the original studies.

4.5.5.1 Internal control weakness and tax avoidance

Using a US sample from 2004–2009, Bauer (2016) documents a higher three-year cash ETR for firms marked with a tax-related internal control weakness (ICW) than for those without such weakness. Bauer (2016) argues that internal controls proxy for the governance mechanism that helps align the interests of shareholders and managers, and tax-related internal controls impact corporate tax avoidance with significant cash flow effects. Table 4.11, Panel A, column (1) presents the replication results of the main analysis in the original study, confirming a robust correlation between tax-related ICW and annual cash ETR. Replacing the control variables with the key variables, Table 4.11, Panel A, column (2) also records a statistically significant impact of tax-related ICW on annual cash ETR. Table 4.11, Panel B shows the results of analysis using the extended sample, which confirm the positive statistically significant correlation between tax-related ICW and annual cash ETR. When comparing results between Table 4.11, Panels A and B, the performance of tax-related ICW remains consistent across different samples and controls.

To sum up, the replication results support the robust explanatory power of tax-related ICW on tax avoidance, suggesting tax-related ICW could be an underestimated factor.

[Table 4.11 about here]

4.5.5.2 Corporate social responsibility and tax avoidance

Davis et al. (2016) find that firms with high corporate social responsibility (CSR) pay fewer taxes, suggesting that taxes and CSR are substitutes instead of complements. Using a US sample from 2002–2011 as indicated in Davis et al. (2016), Table 4.12, Panel A, column (1) shows a statistically significant negative relation between the long-run CSR index and long-run cash ETR. In column (2), once key variables are controlled, the identified negative correlation between long-run cash and long-run cash ETR loses statistical significance. Table 4.12, Panel B presents the results of analysis using the extended sample. Column (3) replicates the main result with control variables used in Davis et al. (2016), and records a statistically significant negative correlation. However, in column (4), the identified statistically significant result becomes insignificant with the inclusion of key variables.

To conclude, the replication results fail to find support for CSR as an important factor when the other robust variables are considered. These results also echo the mixed evidence on the relation between CSR and tax rates (Hoi et al., 2013; Huseynov and Klamm, 2012; Lanis and Richardson, 2012), suggesting a more complex dynamic between CSR and tax payment.

[Table 4.12 about here]

4.5.5.3 Price-cost margin and tax avoidance

Using industry-adjusted price-cost margin (PCM) as the measure of product market power, Kubick et al. (2015) identify a negative correlation between PCM and four ETR measures,

suggesting that the natural hedge of product market power incentivizes tax avoidance. Table 4.13, Panel A1 presents the replication results of the main analysis in Kubick et al. (2015), confirming a statistically significant negative relation between PCM and ETR measures. Table 4.13, Panel A2 presents the replication results using the key variables. For the annual GAAP ETR model, the results in columns (5) and (6)³¹ show no statistically significant correlation between PCM and annual GAAP ETR. The results in Table 4.13, Panel A2, columns (7) and (8) also fail to support a robust correlation between PCM and another two ETR measures (i.e., annual cash ETR and long-run GAAP ETR). Once the key variables are considered, the variable coefficients of PCM even flip the sign from negative to positive in columns (6) and (8). In column (9), long-run cash ETR shows a statistically significant negative correlation with PCM, but with a moderate decrease in coefficient magnitude.

Table 4.13, Panel B presents the replication results of analyses using the extended sample: Panels B1 and B2 show the results of analyses using the original control variables and the key variables, respectively. The results in Table 4.13, Panel B1 show a robust negative correlation between PCM and four ETR measures, and the variable coefficients of PCM are larger than those using the original sample in Table 4.13, Panel A. However, regarding the key variables, the results in Table 4.13, Panel B2 show no supportive evidence for the identified robust correlation, except for the long-run cash ETR model in column (9). Specifically, the variable coefficients of PCM flip the direction and become positive in both the annual and long-run GAAP ETR models.

[Table 4.13 about here]

³¹ Columns (5) and (6) present the results using the parsimonious three-variable controls and the 23-variable controls, respectively.

Kubick et al. (2015) apply the same control variables for all ETR measures, suggesting that their singular control set takes all factors into consideration for the four ETR models. Therefore, I aggregated the key variables for those four ETR measures into a comprehensive robust control set, and performed an additional replication analysis. Table 4.13, Panel C presents the replication results of analysis using the comprehensive control set: Panels C1 and C2 show the results of analyses using the original sample and the extended sample, respectively. In Table 4.13, Panels C1 and C2, the results in the annual GAAP ETR, annual cash ETR, and long-run cash ETR models show weak correlation between tax avoidance and PCM. The identified negative correlation even flips into a positive sign with statistical significance for the long-run GAAP ETR model in column (3), which suggests a contradictory conclusion to that in the original study.

In conclusion, the replication results show that the identified negative correlation between PCM and ETR is sensitive to the inclusion of the key variables. The results reflect the debate about the effect of product market power on tax avoidance. As firms with greater product market power are likely to be large, instead of taking advantage of their market power, they may refrain from engaging in tax avoidance out of concern for political costs or reputation damage (Watts and Zimmerman, 1978).

4.5.6 Comparison between theoretical predictions and empirical outcomes

Owing to the lack of a systematic review of variable performance in tax avoidance research, an evaluation of how well the tax avoidance performances predicted by theories match the empirical evidence is missing. Therefore, I provide a comparison between theoretical predictions and empirical outcomes of the identified key variables to examine the theory–

empiricism consistency. Table 4.14, Panel A presents the theoretical predictions, summarized by Bruhne and Jacob (2019), of correlations between tax avoidance and its covariates. For example, the benefits of tax avoidance increase with firm size, as larger firms tend to have a larger tax base. Regarding costs, the political cost hypothesis (e.g., larger firms attract more attention from regulators and investors) and the political power hypothesis (e.g., larger firms have more resources for negotiating and lobbying) provide conflicting arguments about the correlation between firm size and tax avoidance. According to the comparison between benefits and costs, theories are unable to derive a clear prediction for the correlation between firm size and tax avoidance. Table 4.14, Panel B presents the aggregated variable performance from the fixed-effect analyses across eight tax avoidance models. For example, *Size* is selected by three ETR models, and shows a generally mixed relation with tax avoidance.

[Table 4.14 about here]

Figure 4.2 plots the comparison between the theoretical predictions and the empirical outcomes in Table 4.14. Theories and empirical evidence are consistent in five variables that show mixed relation with tax avoidance: *Size*, *Market power*, *Profitability*, *Life cycle*, and *Intangible assets*. Consistent with theory, *Financial constraints* and *Foreign operation* show a positive relation with tax avoidance in empirical analyses. The theoretical predictions for *Ownership*, *Growth*, *Complexity*, and *Tangible assets* are not fully supported by empirical evidence. Meanwhile the empirical performance of *Leverage* contradicts theoretical predictions.

[Figure 4.2 about here]

4.6 CONCLUSION

Empirical research in finance and accounting has seen a rapidly growing number of identified incremental variables for a variety of tax avoidance measures. Using a US sample from 1990–2019, this study applies machine learning techniques to investigate this vast number of tax avoidance measures and covariates, and selects numerous robust variables as popular tax avoidance measures.

This study makes several noteworthy contributions to extant literature. Firstly, it is among the first to introduce machine learning techniques into empirical tax avoidance research. Based on opinions from two variable-selection-oriented machine learning techniques, LASSO and RF, the study identifies a number of key variables as popular tax avoidance measures. With public data access, those key variables provide some robust baseline models which future researchers should consider as a starting point. By replicating a number of prior works and showing mixed results with the key variables, this study empirically demonstrates the importance of controlling robust variables, in response to calls for more prudence when another new incremental variable is identified (Dyckman and Zeff, 2014; Gow et al., 2016; Harvey, 2017).

Secondly, to the best of my knowledge, this study provides the first systematic quantitative assessment of tax avoidance covariates. With a comprehensive review of prior empirical studies and implementation of machine learning techniques, this study combines the widely cited tax avoidance covariates, and compares their performance with one another. Extending the discussions in Bruhne and Jacob (2019), this study bridges empirical outcomes and theory by providing the first comparison between the theoretical predictions and empirical evidence of variable performance.

Finally, this study demonstrates the complexity of empirical tax avoidance research, and the need to find more robust variables to enrich extant tax avoidance models. The mixed replication results of prior works, together with the large inconsistencies between theoretical predictions and empirical outcomes, reflect a deep deficiency in our understanding of corporate tax avoidance (Hanlon and Heitzman, 2010; Wilde and Wilson, 2018).

There are a few important caveats for readers when interpreting the results. Firstly, and most importantly, this study aims to provide a systematic quantitative evaluation of the most widely cited tax avoidance covariates, but it has no intention of finding the perfect model. Therefore, I argue that the identified key variables provide a good starting point for subsequent research, and are open to reasonable tailoring in any specific research topic. The variables that are not selected as key variables are not considered unimportant, but are rather second-order factors (e.g., corporate governance attributes). Secondly, I argue that mixed replication results in prior studies are likely unintentional misidentifications due to outdated samples and the lack of a unanimous robust control set in extant literature. Finally, the analyses in this study are based on public data, and thus are unable to take into consideration variables based on private data.

Appendix 4A Variable definition

This table reports the definition of the variables used in machine learning analysis.

Variable	Definition
Cash_ETR1	Cash effective tax rate, calculated as paid income taxes divided by pre-tax income net of special items.
GAAP_ETR1	GAAP effective tax rate, calculated as total income taxes divided by pre-tax income net of special items.
CashETR3	Long-term cash ETR, calculated as the sum of paid income taxes in three years divided by the sum of pre-tax income net of special items.
GAAPETR3	Long-term GAAP ETR, calculated as the sum of total income taxes in three years divided by the sum of pre-tax income net of special items.
UTB	Unrecognized tax benefits, calculated as the amount of unrecognized tax benefits scaled by total assets.
DTAX	Frank et al. (2009) discretionary permanent book-tax difference, calculated as the residual from regressing total book-tax differences less temporary book-tax difference, on intangible assets, income (loss) reported under the equity method, income attributable to minority interest, state income taxes, change in net operating loss carry forwards, and one-year lagged of total BTB less temporary BTB.
MP_BTBD	Manzon and Pleski (2002) book-tax difference, calculated as U.S. domestic financial income minus U.S. domestic taxable income, minus income taxes (state), income taxes (other) and equity in earnings, divided by lagged assets.
DD_BTBD	Desai and Dharmapala (2006) discretionary book-tax difference, calculated as the sum of the average value of residual for firm over the sample period and the deviation of the residual in a given year, from the firm fixed-effect regression of Manzon and Plesko (2002) BTB scaled by lagged total assets, on total accruals scaled by lagged total assets.
Size_at	Natural logarithm of firm total assets.
Size_mv	Natural logarithm of firm market capital.
Size_sale	Natural logarithm of firm sales.
HHI_sic	Herfindahl-Hirschman Index in a two-digit SIC industry in a given year.
HHI_sicc	Herfindahl-Hirschman Index in a four-digit SIC industry in a given year.
ROA_piat	The ratio of pretax income scaled by total assets.
ROA_pilat	The ratio of pretax income scaled by lagged total assets.
ROA_pilat2	The ratio of pretax income net of extraordinary items scaled by lagged total assets.
ROA_iblat	The ratio of income before extraordinary items scaled by lagged total assets.
EBIT	The ratio of earnings before interest and taxes scaled by net operating assets.
OCF_avat	The ratio of operating cash flow scaled by average assets.
OCff_lat	The ratio of operating cash flow minus capital expenditures scaled by lagged total assets.
DNOL	The ratio of annual change of tax loss carry forward scaled by lagged total assets.

NOL1_d	Dummy variable, equal to 1 if the tax loss carry forward is positive.
NOL2_d	Dummy variable, equal to 1 if the lagged tax loss carry forward is positive.
Return	Annual stock returns.
MTB	Ratio of market value to book value.
TBQ	The firm's Tobin's Q value.
DSale	Annual sales growth.
MA_d	Dummy variable, equal to 1 if there is a merger and acquisition at current year.
CAPX_lat	The ratio of capital expenditures scaled by lagged total assets.
CAPX_at	The ratio of capital expenditures scaled by total assets.
DPPENT	The ratio of annual change of net property, plant and equipment scaled by average total assets.
DPPEGT	The ratio of annual change of gross property, plant and equipment scaled by average total assets.
PPENT_at	The ratio of net property, plant and equipment scaled by total assets.
PPENT_lat	The ratio of net property, plant and equipment scaled by lagged total assets.
PPEGT_at	The ratio of gross property, plant and equipment scaled by total assets.
Invt_lat	The ratio of total inventory scaled by lagged total assets.
DP_at	The ratio of depreciation and amortization expense scaled by lagged total assets.
DGW_at	The ratio of annual change of goodwill scaled by total assets.
RND_lat	The ratio of research and development expense scaled by lagged total assets.
RND_at	The ratio of research and development expense scaled by total assets.
RND_sale	The ratio of research and development expense scaled by total sales.
Intan_lat	The ratio of intangible assets scaled by lagged total assets.
Intan_at	The ratio of intangible assets scaled by total assets.
Intan_sale	The ratio of intangible assets scaled by total sales.
DLTT_at	The ratio of long-term debt scaled by total assets.
DLTT_lat	The ratio of long-term debt scaled by lagged total assets.
LEV_at	The sum of long-term debt and debt in current liabilities divided by total assets.
INT_lat	The ratio of interest expense scaled by average total assets.
Mezzanine	The Mezzanine index, calculated as the convertible debt and preferred stock divided by total assets.
Fin_d	Dummy variable, equal to 1 if a firm issues stock or long-term debt in the given year, and 0 otherwise.
IO_ts	The ratio of number of shares held by institutional investors scaled by total shares outstanding.
CHE_at	The ratio of cash and equivalents scaled by lagged total assets.
SA_HP2010	Hadlock and Pierce (2010) index.
AltmanZ	Altman's Z-score.
PIFOc_at	The ratio of after-tax foreign income scaled by total assets
PIFO_d	Dummy variable, equal to 1 if either foreign pre-tax income or foreign income taxes is non-zero, and 0 otherwise.

PIFO_lat	The ratio of foreign pre-tax income divided by lagged total assets.
PIFO_at	The ratio of foreign pre-tax income divided by total assets.
PIFO_pi	The ratio of absolute value of pre-tax foreign income divided by the absolute value of pre-tax total income.
NGS	Natural logarithm of one plus number of geographic segments.
NBS1	Natural logarithm of one plus number of business segments.
NBS2	Natural logarithm of one plus number of unique business segments.
ESUB_lat	The ratio of equity income in earnings scaled by lagged total assets.
ESUB_d	Dummy variable, equal to 1 if equity in earnings is positive, and 0 otherwise.
EMP	Natural logarithm of one plus number of employees.
StdEarnFst	Standard deviation of analyst earnings forecasts.
Num_Analyst	Natural logarithm of one plus number of analysts following the firm.
Goodnews_d	Dummy variable, equal to 1 if earnings surprise is positive, and 0 otherwise. Earnings surprise is measured as actual earnings per share minus the median analyst earnings forecast.
Big4_d	Dummy variable, equal to 1 if the auditor is a big four, and 0 otherwise.
AuditOp_d	Dummy variable, equal to 1 if the firm received an unqualified audit opinion in a given year, and 0 otherwise.
AbAccr1	The amount of abnormal accruals, following Kothari et al. (2005)
AbAccr2	The amount of abnormal accruals, following Dechow et al. (1995)
ToAcrr	The ratio of total accruals using the cash flow approach (Hribar and Collins, 2002) scaled by lagged total assets.
SGA_sale	The ratio of selling, general and administrative expense scaled by sales.
SGA_at	The ratio of selling, general and administrative expense scaled by assets.
XAD_avat	The ratio of advertising expense scaled by average total assets.
XAD_sale	The ratio of advertising expense scaled by sales.
XAD_at	The ratio of advertising expense scaled by total assets.
SPI_at	The ratio of special items scaled by average total assets.
SPI_sale	The ratio of special items scaled by sales.
EI_at	The ratio of extraordinary items scaled by average total assets.
StdROA5	The standard deviation of ROA in the past five fiscal years.
StdCF5	The standard deviation of operating cash flow in the past five fiscal years.
StdSale5	The standard deviation of sales in the past five fiscal years.
StdMRn5	The standard deviation of monthly stock returns in the past five fiscal years.
StdARn	The annual standard deviation of daily raw returns.
Age	Natural logarithm of one plus number of years for the firm listed on Compustat.
BrdSize	Natural logarithm of one plus number of board directors
Ind_dir	Proportion of independent board members.
Female_d	Dummy variable, equal to 1 if the CEO is female, and 0 otherwise.
CEOage	Natural logarithm of one plus the age of CEO.
Tenure_CEO	Number of years the CEO holds the position.
Stk_ceo	Percentage of stock owned by CEO.

StkComp_d	Dummy variable, equal to 1 if the stock compensation is non-zero, and 0 otherwise.
Cash_compen	Natural logarithm of salary and bonus.
CEOChair_d	Dummy variable, equal to 1 if CEO is also the chair of the board, and 0 otherwise.

Figures

Figure 4.1 Flow chart of the two-round analysis

This figure illustrates the variable selection procedures using machine learning.

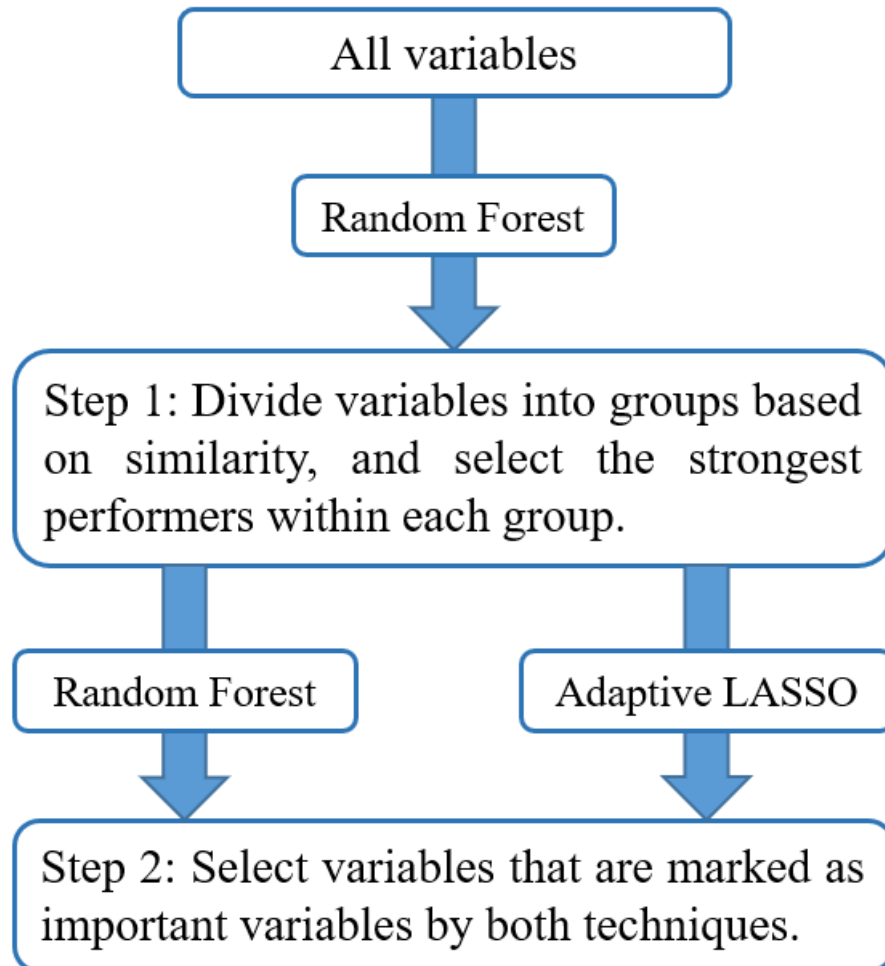


Figure 4.2 Summary matrix of theoretical prediction and empirical outcome

This figure plots a summary matrix comparing the theoretical prediction and empirical outcome. Following Bruhne and Jacob (2019), the comparison matrix aggregates the overall association between variable and corporate avoidance across all tax avoidance models, from both theoretical prediction and empirical outcomes in Table 14.

		Empirical Outcomes		
		Positive relation between construct and tax avoidance	Mixed relation between construct and tax avoidance	Negative relation between construct and tax avoidance
Theoretical Prediction	Positive relation between construct and tax avoidance	<i>Financial constraints</i> <i>Foreign operation</i>	<i>Growth</i> <i>Complexity</i>	
	Mixed relation between construct and tax avoidance	<i>Ownership</i>	<i>Size</i> <i>Market power</i> <i>Profitability</i> <i>Life cycle</i> <i>Intangible assets</i>	<i>Tangible assets</i>
	Negative relation between construct and tax avoidance	<i>Leverage</i>		

Tables

Table 4.1 Summary statistics

This table reports the descriptive statistics of the variables. Appendix 4A provides detailed variable definitions.

Panel A: Main firm attributes

Variables	Mean	SD	Median	Variables	Mean	SD	Median
Size_at	6.18	2.47	6.15	Mezzanine	0.02	0.09	0.00
Size_mv	6.03	2.37	6.08	Fin_d	0.83	0.37	1.00
Size_sale	5.80	2.32	5.84	IO_ts	0.48	0.31	0.49
HHI_sic	0.07	0.07	0.05	CHE_at	0.19	0.32	0.08
HHI_sicc	0.21	0.18	0.16	SA_HP2010	-3.35	0.94	-3.32
ROA_piat	0.09	0.09	0.07	AltmanZ	5.32	8.12	3.39
ROA_pilat	0.11	0.12	0.08	PIFOc_at	0.01	0.02	0.00
ROA_pilat2	0.11	0.12	0.08	PIFO_lat	0.01	0.02	0.00
ROA_iblat	0.08	0.09	0.05	PIFO_at	0.01	0.02	0.00
EBIT	0.20	0.93	0.14	PIFO_pi	0.13	0.35	0.00
OCF_avat	0.10	0.11	0.09	PIFO_d	0.35	0.48	0.00
OCFf_lat	0.03	0.18	0.04	NGS	1.03	0.43	0.69
DNOL	-0.00	0.15	0.00	NBS1	0.99	0.38	0.69
NOL1_d	0.31	0.46	0.00	NBS2	0.88	0.30	0.69
NOL2_d	0.29	0.45	0.00	ESUB_lat	0.00	0.00	0.00
Return	0.23	0.58	0.12	ESUB_d	0.17	0.38	0.00
MTB	3.01	4.33	1.90	EMP	7.32	2.22	7.39
TBQ	1.93	2.07	1.38	StdEarnFst	0.07	0.20	0.02
DSale	0.22	0.67	0.09	Num_Analyst	1.03	1.10	0.69
MA_d	0.31	0.46	0.00	Goodnews_d	0.29	0.45	0.00
CAPX_lat	0.07	0.10	0.04	Big4_d	0.79	0.41	1.00
CAPX_at	0.06	0.07	0.04	AuditOp_d	0.75	0.43	1.00
DPPENT	0.03	0.09	0.01	AbAccr1	0.03	0.10	0.02
DPPEGT	0.06	0.12	0.03	AbAccr2	-0.00	0.13	0.00
PPENT_at	0.30	0.27	0.21	ToAcrr	-0.03	0.13	-0.04
PPENT_lat	0.35	0.33	0.24	SGA_sale	0.19	0.37	0.14
PPEGT_at	0.54	0.45	0.44	SGA_at	0.20	0.26	0.12
Invt_lat	0.13	0.17	0.05	XAD_avat	0.01	0.03	0.00
DP_at	0.05	0.04	0.04	XAD_sale	0.01	0.02	0.00
DGW_at	0.01	0.03	0.00	XAD_at	0.01	0.03	0.00
RND_lat	0.02	0.07	0.00	SPI_at	-0.00	0.02	0.00
RND_at	0.02	0.05	0.00	SPI_sale	-0.01	0.05	0.00
RND_sale	0.03	0.17	0.00	EI_at	0.00	0.00	0.00
Intan_lat	0.12	0.21	0.01	StdROA5	0.06	0.13	0.03
Intan_at	0.11	0.17	0.02	StdCF5	0.06	0.08	0.04

Intan_sale	0.03	0.18	0.00	StdSale5	0.17	0.21	0.10
DLTT_at	0.19	0.20	0.15	StdMRn5	0.13	0.07	0.11
DLTT_lat	0.23	0.24	0.17	StdARn	0.03	0.02	0.03
LEV_at	0.25	0.22	0.22	Age	2.46	0.99	2.56
INT_lat	0.02	0.03	0.01				

Panel B: Corporate governance attributes

Variables	Mean	SD	Median	Variables	Mean	SD	Median
BrdSize	2.24	0.27	2.20	StkCompen_d	0.38	0.48	0.00
Ind_dir	0.80	0.13	0.83	Stk_ceo	0.03	0.06	0.00
Female_d	0.03	0.18	0.00	Cash_compen	6.74	1.04	6.80
CEOage	4.04	0.14	4.04	CEOChair_d	0.52	0.50	1.00
Tenure_CEO	8.01	7.62	5.59				

Panel C: Main tax avoidance variables

Variables	Mean	SD	Median	Variables	Mean	SD	Median
Cash_ETR1	0.23	0.22	0.20	CashETR3	0.24	0.20	0.23
GAAP_ETR1	0.27	0.19	0.30	GAAPETR3	0.26	0.19	0.29

Panel D: Alternative tax avoidance variables

Variables	Mean	SD	Median	Variables	Mean	SD	Median
UTB	0.01	0.01	0.00	MP_BT D	0.01	0.06	0.00
DD_BT D	0.03	0.08	0.03	DTAX	6.31	67.43	0.79

Table 4.2 Strongest variable selected from each group for ETR measures

This table reports the results of Random Forest analysis with each group in Step 1, for annual cash ETR, annual GAAP ETR, long-run cash ETR and long-run GAAP ETR. Variables are ranked according to their importance loading. Appendix 4A provides variable definitions.

	Cash ETR1	GAAP ETR1	Cash ETR3	GAAP ETR3
Size	Size_mv	Size_mv	Size_mv	Size_mv
Competition	HHI_sicc	HHI_sicc	HHI_sicc	HHI_sicc
Profitability	ROA_iblat	ROA_iblat	ROA_iblat	ROA_iblat
Operating Cash Flow	OCFf_lat	OCFf_lat	OCFf_lat	OCFf_lat
NOL	NOL1_d	NOL1_d	NOL1_d	NOL1_d
Valuation	MTB	MTB	MTB	MTB
Investment	CAPX_lat	CAPX_lat	CAPX_lat	CAPX_lat
PPE	DPPEGT	DPPEGT	DPPEGT	DPPEGT
RND	RND_at	RND_sale	RND_sale	RND_sale
Intangible Assets	Intan_sale	Intan_sale	Intan_sale	Intan_sale
Leverage	DLTT_lat	DLTT_lat	DLTT_lat	DLTT_lat
Foreign Operation	PIFO_pi	PIFOc_at	PIFOc_at	PIFO_pi
Segments	NBS2	NBS2	NBS2	NBS2
Accruals	AbAccr1	AbAccr2	AbAccr1	AbAccr2
SGNA	SGA_at	SGA_at	SGA_at	SGA_at
Adv. Expenses	XAD_sale	XAD_sale	XAD_sale	XAD_avat
Special Items	SPI_sale	SPI_at	SPI_sale	SPI_sale
Operating Risk	StdCF5	StdCF5	StdCF5	StdCF5
Return Volatility	StdMRn5	StdMRn5	StdMRn5	StdMRn5

Table 4.3 Variable selection for ETR models using fixated threshold

This table summarizes the variable performance in the models of annual cash ETR, annual GAAP ETR, long-run cash ETR and long-run GAAP ETR using Adaptive LASSO and random forest with fixated threshold. Column AL presents the variable performance in Adaptive LASSO test, calculated as the frequency of selection divided by the total number of rolling windows. Column RF presents the variable importance loading in random forest analysis. Column FT shows if the variable meets the 50% fixated threshold. Appendix 4A provides variable definitions.

Panel A: Annual cash ETR				Panel B: Annual GAAP ETR				Panel C: Long-run cash ETR				Panel D: Long-run GAAP ETR			
Variables	AL	RF	FT	Variables	AL	RF	FT	Variables	AL	RF	FT	Variables	AL	RF	FT
Size_mv	0.61	0.5-0.6	Yes	Size_mv	0.46	0.3-0.4	No	Size_mv	0.54	0.7-0.8	Yes	Size_mv	0.69	0.6-0.7	Yes
EMP	0.79	0.5-0.6	Yes	EMP	0.54	0.3-0.4	No	EMP	0.92	0.8-0.9	Yes	EMP	0.42	0.6-0.7	No
HHI_sicc	0.50	0.7-0.8	Yes	HHI_sicc	0.43	0.5-0.6	No	HHI_sicc	0.54	0.9-1	Yes	HHI_sicc	0.35	0.8-0.9	No
ROA_iblat	0.61	0.9-1	Yes	ROA_iblat	1	0.9-1	Yes	ROA_iblat	0.85	0.9-1	Yes	ROA_iblat	0.73	0.8-0.9	Yes
OCFf_lat	0.93	0.6-0.7	Yes	OCFf_lat	0.96	0.4-0.5	No	OCFf_lat	0.58	0.8-0.9	Yes	OCFf_lat	0.92	0.7-0.8	Yes
DNOL	0.61	0.6-0.7	Yes	DNOL	0.46	0.3-0.4	No	DNOL	0.31	0.7-0.8	No	DNOL	0.27	0.5-0.6	No
NOL1_d	1	0.4-0.5	No	NOL1_d	1	0.2-0.3	No	NOL1_d	1	0.5-0.6	Yes	NOL1_d	0.77	0.4-0.5	No
Return	1	0.5-0.6	Yes	Return	0.79	0.2-0.3	No	Return	0.81	0.6-0.7	Yes	Return	0.12	0.4-0.5	No
MTB	0.50	0.5-0.6	Yes	MTB	0.68	0.3-0.4	No	MTB	0.5	0.7-0.8	Yes	MTB	0.31	0.6-0.7	No
DSale	0.89	0.7-0.8	Yes	DSale	0.5	0.3-0.4	No	DSale	0.35	0.8-0.9	No	DSale	0.5	0.7-0.8	Yes
MA_d	0.75	0.2-0.3	No	MA_d	0.54	0.1-0.2	No	MA_d	0.5	0.2-0.3	No	MA_d	0.31	0.2-0.3	No
CAPX_lat	0.93	0.7-0.8	Yes	CAPX_lat	0.82	0.6-0.7	Yes	CAPX_lat	1	0.9-1	Yes	CAPX_lat	0.54	0.9-1	Yes
DPPEGT	0.57	0.6-0.7	Yes	DPPEGT	0.54	0.4-0.5	No	DPPEGT	0.38	0.8-0.9	No	DPPEGT	0.27	0.7-0.8	No
Invt_lat	0.54	0.6-0.7	Yes	Invt_lat	0.54	0.4-0.5	No	Invt_lat	0.69	0.9-1	Yes	Invt_lat	0.38	0.8-0.9	No
DP_at	0.86	0.6-0.7	Yes	DP_at	0.54	0.4-0.5	No	DP_at	0.69	0.9-1	Yes	DP_at	0.19	0.8-0.9	No
DGW_at	0.54	0.4-0.5	No	DGW_at	0.5	0.2-0.3	No	DGW_at	0.46	0.5-0.6	No	DGW_at	0.27	0.4-0.5	No
RND_at	1.00	0.4-0.5	No	RND_sale	0.75	0.2-0.3	No	RND_sale	0.73	0.6-0.7	Yes	RND_sale	0.65	0.5-0.6	Yes
Intan_sale	0.64	0.4-0.5	No	Intan_sale	0.61	0.3-0.4	No	Intan_sale	0.73	0.6-0.7	Yes	Intan_sale	0.62	0.5-0.6	Yes
DLTT_lat	0.79	0.6-0.7	Yes	DLTT_lat	0.5	0.4-0.5	No	DLTT_lat	0.62	0.8-0.9	Yes	DLTT_lat	0.42	0.7-0.8	No
INT_lat	0.79	0.6-0.7	Yes	INT_lat	0.43	0.3-0.4	No	INT_lat	0.62	0.8-0.9	Yes	INT_lat	0.31	0.7-0.8	No
Mezzanine	0.61	0.2-0.3	No	Mezzanine	0.32	0.1-0.2	No	Mezzanine	0.65	0.3-0.4	No	Mezzanine	0.35	0.3-0.4	No

Fin_d	0.43	0.1-0.2	No	Fin_d	0.43	0.1-0.2	No	Fin_d	0.42	0.2-0.3	No	Fin_d	0.15	0.2-0.3	No
IO_ts	0.50	0.4-0.5	No	IO_ts	0.54	0.2-0.3	No	IO_ts	0.5	0.6-0.7	Yes	IO_ts	0.15	0.4-0.5	No
CHE_at	0.50	0.6-0.7	Yes	CHE_at	0.68	0.4-0.5	No	CHE_at	0.5	0.8-0.9	Yes	CHE_at	0.46	0.7-0.8	No
SA_HP2010	0.57	0.6-0.7	Yes	SA_HP2010	0.54	0.4-0.5	No	SA_HP2010	0.46	0.8-0.9	No	SA_HP2010	0.27	0.7-0.8	No
AltmanZ	0.64	0.5-0.6	Yes	AltmanZ	0.82	0.4-0.5	No	AltmanZ	0.58	0.8-0.9	Yes	AltmanZ	0.65	0.7-0.8	Yes
PIFO_pi	1	0.6-0.7	Yes	PIFOc_at	0.96	0.3-0.4	No	PIFOc_at	0.85	0.6-0.7	Yes	PIFO_pi	0.81	0.5-0.6	Yes
PIFO_d	0.50	0.2-0.3	No	PIFO_d	0.57	0.1-0.2	No	PIFO_d	0.69	0.3-0.4	No	PIFO_d	0.38	0.2-0.3	No
NBS2	0.46	0.2-0.3	No	NBS2	0.5	0.1-0.2	No	NBS2	0.46	0.3-0.4	No	NBS2	0.42	0.3-0.4	No
ESUB_lat	0.54	0.4-0.5	No	ESUB_lat	0.57	0.3-0.4	No	ESUB_lat	0.23	0.5-0.6	No	ESUB_lat	0.19	0.5-0.6	No
ESUB_d	0.43	0.1-0.2	No	ESUB_d	0.5	0.1-0.2	No	ESUB_d	0.31	0.2-0.3	No	ESUB_d	0.15	0.2-0.3	No
StdEarnFst	0.89	0.4-0.5	No	StdEarnFst	0.75	0.2-0.3	No	StdEarnFst	0.58	0.6-0.7	Yes	StdEarnFst	0.35	0.5-0.6	No
Num_Analyst	0.57	0.4-0.5	No	Num_Analyst	0.5	0.3-0.4	No	Num_Analyst	0.46	0.6-0.7	No	Num_Analyst	0.23	0.5-0.6	No
Goodnews_d	0.96	0.2-0.3	No	Goodnews_d	0.54	0.1-0.2	No	Goodnews_d	0.38	0.3-0.4	No	Goodnews_d	0.38	0.3-0.4	No
Big4_d	0.46	0.2-0.3	No	Big4_d	0.21	0.1-0.2	No	Big4_d	0.38	0.2-0.3	No	Big4_d	0.23	0.2-0.3	No
AuditOp_d	0.57	0.2-0.3	No	AuditOp_d	0.61	0.1-0.2	No	AuditOp_d	0.54	0.2-0.3	No	AuditOp_d	0.31	0.2-0.3	No
AbAccr1	0.54	0.5-0.6	Yes	AbAccr2	0.57	0.3-0.4	No	AbAccr1	0.62	0.8-0.9	Yes	AbAccr2	0.23	0.6-0.7	No
Age	0.43	0.5-0.6	No	Age	0.93	0.3-0.4	No	Age	0.19	0.7-0.8	No	Age	0.58	0.6-0.7	Yes
SGA_at	1	0.6-0.7	Yes	SGA_at	0.96	0.4-0.5	No	SGA_at	0.92	0.9-1	Yes	SGA_at	0.62	0.7-0.8	Yes
XAD_sale	0.57	0.3-0.4	No	XAD_sale	0.64	0.2-0.3	No	XAD_sale	0.54	0.5-0.6	Yes	XAD_avat	0.15	0.5-0.6	No
SPI_sale	0.93	0.6-0.7	Yes	SPI_at	1	0.6-0.7	Yes	SPI_sale	0.69	0.8-0.9	Yes	SPI_sale	0.88	0.7-0.8	Yes
EI_at	0.54	0.3-0.4	No	EI_at	0.25	0.2-0.3	No	EI_at	0.38	0.4-0.5	No	EI_at	0.08	0.4-0.5	No
StdCF5	0.50	0.5-0.6	Yes	StdCF5	0.79	0.3-0.4	No	StdCF5	0.65	0.8-0.9	Yes	StdCF5	0.58	0.6-0.7	Yes
StdMRn5	0.86	0.5-0.6	Yes	StdMRn5	0.82	0.3-0.4	No	StdMRn5	0.88	0.7-0.8	Yes	StdMRn5	0.73	0.6-0.7	Yes

Table 4.4 Variable selection for alternative tax avoidance models using fixated threshold

This table summarizes the variable performance in the models of UTB, DTAX, DDBTD and MPBTD using Adaptive LASSO and random forest with fixated threshold. Appendix 4A provides variable definitions. Column AL presents the variable performance in Adaptive LASSO test, calculated as the frequency of selection divided by the total number of rolling windows. Column RF presents the variable importance loading in random forest analysis. Column FT shows if the variable meets the 50% fixated threshold.

Panel A: UTB				Panel B: DTAX				Panel C: DDBTD				Panel D: MPBTD			
Variables	AL	RF	FT	Variables	AL	RF	FT	Variables	AL	RF	FT	Variables	AL	RF	FT
Size_mv	0.18	0.6-0.7	No	Size_mv	0.30	0.9-1	No	Size_mv	0.26	0.6-0.7	No	Size_mv	0.43	0.5-0.6	No
EMP	0.64	0.6-0.7	Yes	EMP	0.65	0.7-0.8	Yes	EMP	0.78	0.7-0.8	Yes	EMP	0.48	0.5-0.6	No
HHI_sicc	0.27	0.6-0.7	No	HHI_sicc	0.35	0.7-0.8	No	HHI_sicc	0.30	0.6-0.7	No	HHI_sicc	0.39	0.5-0.6	No
ROA_iblat	0.64	0.7-0.8	Yes	ROA_iblat	1	0.9-1	Yes	ROA_iblat	0.57	0.8-0.9	Yes	ROA_iblat	0.65	0.9-1	Yes
OCFf_lat	0.55	0.7-0.8	Yes	OCFf_lat	0.70	0.7-0.8	Yes	OCFf_lat	0.48	0.7-0.8	No	OCFf_lat	0.70	0.6-0.7	Yes
DNOL	0.18	0.6-0.7	No	DNOL	0.91	0.7-0.8	Yes	DNOL	0.65	0.5-0.6	Yes	DNOL	0.78	0.5-0.6	Yes
NOL1_d	0.27	0.3-0.4	No	NOL1_d	0.26	0.2-0.3	No	NOL2_d	1	0.3-0.4	No	NOL2_d	1	0.3-0.4	No
Return	0.27	0.4-0.5	No	Return	0.35	0.5-0.6	No	Return	0.39	0.4-0.5	No	Return	0.74	0.3-0.4	No
MTB	0.45	0.5-0.6	No	MTB	0.61	0.6-0.7	Yes	MTB	0.39	0.5-0.6	No	MTB	0.52	0.5-0.6	Yes
DSale	0.64	0.6-0.7	Yes	DSale	0.39	0.7-0.8	No	DSale	0.52	0.6-0.7	Yes	DSale	0.43	0.6-0.7	No
MA_d	0.09	0.2-0.3	No	MA_d	0.26	0.2-0.3	No	MA_d	0.09	0.2-0.3	No	MA_d	0.17	0.2-0.3	No
CAPX_lat	0.64	0.6-0.7	Yes	CAPX_lat	0.65	0.7-0.8	Yes	CAPX_at	0.70	0.6-0.7	Yes	CAPX_at	0.65	0.5-0.6	Yes
DPPEGT	0.73	0.6-0.7	Yes	DPPEGT	0.22	0.6-0.7	No	PPEGT_at	0.87	0.7-0.8	Yes	PPEGT_at	0.78	0.5-0.6	Yes
Invt_lat	0.18	0.5-0.6	No	Invt_lat	0.57	0.6-0.7	Yes	Invt_lat	0.65	0.6-0.7	Yes	Invt_lat	0.91	0.5-0.6	Yes
DP_at	0.64	0.6-0.7	Yes	DP_at	0.48	0.7-0.8	No	DP_at	0.78	0.6-0.7	Yes	DP_at	0.61	0.6-0.7	Yes
DGW_at	0.18	0.4-0.5	No	DGW_at	0.48	0.6-0.7	No	DGW_at	0.22	0.3-0.4	No	DGW_at	0.22	0.3-0.4	No
RND_sale	0.82	0.9-1	Yes	RND_sale	0.83	0.7-0.8	Yes	RND_lat	0.78	0.6-0.7	Yes	RND_lat	0.83	0.5-0.6	Yes
Intan_sale	0.36	0.9-1	No	Intan_lat	0.61	0.6-0.7	Yes	Intan_sale	0.65	0.6-0.7	Yes	Intan_sale	0.65	0.5-0.6	Yes
LEV_at	0.64	0.5-0.6	Yes	DLTT_lat	0.26	0.6-0.7	No	LEV_at	0.43	0.5-0.6	No	LEV_at	0.52	0.4-0.5	No
INT_lat	0.36	0.6-0.7	No	INT_lat	0.48	0.6-0.7	No	INT_lat	0.35	0.5-0.6	No	INT_lat	0.65	0.5-0.6	Yes
Mezzanine	0.55	0.4-0.5	No	Mezzanine	0.26	0.4-0.5	No	Mezzanine	0.35	0.3-0.4	No	Mezzanine	0.48	0.2-0.3	No
Fin_d	0.00	0.2-0.3	No	Fin_d	0.30	0.2-0.3	No	Fin_d	0.35	0.2-0.3	No	Fin_d	0.39	0.2-0.3	No

IO_ts	0.45	0.4-0.5	No	IO_ts	0.22	0.5-0.6	No	IO_ts	0.65	0.4-0.5	No	IO_ts	0.43	0.3-0.4	No
CHE_at	0.36	0.7-0.8	No	CHE_at	0.22	0.7-0.8	No	CHE_at	0.65	0.7-0.8	Yes	CHE_at	0.65	0.6-0.7	Yes
SA_HP2010	0.18	0.6-0.7	No	SA_HP2010	0.30	0.7-0.8	No	SA_HP2010	0.35	0.7-0.8	No	SA_HP2010	0.39	0.5-0.6	No
AltmanZ	0.45	0.6-0.7	No	AltmanZ	0.78	0.7-0.8	Yes	AltmanZ	0.78	0.6-0.7	Yes	AltmanZ	0.78	0.5-0.6	Yes
PIFO_pi	1.00	0.8-0.9	Yes	PIFO_pi	0.61	0.8-0.9	Yes	PIFO_pi	1	0.9-1	Yes	PIFO_pi	0.91	0.8-0.9	Yes
PIFO_d	0.27	0.5-0.6	No	PIFO_d	0.13	0.2-0.3	No	PIFO_d	0.48	0.8-0.9	No	PIFO_d	0.35	0.5-0.6	No
NGS	0.82	0.5-0.6	Yes	NGS	0.52	0.5-0.6	Yes	NGS	0.39	0.4-0.5	No	NGS	0.61	0.3-0.4	No
ESUB_lat	0.36	0.4-0.5	No	ESUB_lat	0.61	0.6-0.7	Yes	ESUB_lat	0.83	0.4-0.5	No	ESUB_lat	0.78	0.3-0.4	No
ESUB_d	0.00	0.2-0.3	No	ESUB_d	0.65	0.3-0.4	No	ESUB_d	0.26	0.2-0.3	No	ESUB_d	0.30	0.1-0.2	No
StdEarnFst	0.18	0.4-0.5	No	StdEarnFst	0.22	0.6-0.7	No	StdEarnFst	0.26	0.4-0.5	No	StdEarnFst	0.30	0.3-0.4	No
Num_Analyst	0.91	0.5-0.6	Yes	Num_Analyst	0.35	0.7-0.8	No	Num_Analyst	0.61	0.5-0.6	Yes	Num_Analyst	0.43	0.4-0.5	No
Goodnews_d	0.00	0.2-0.3	No	Goodnews_d	0.13	0.3-0.4	No	Goodnews_d	0.30	0.2-0.3	No	Goodnews_d	0.22	0.2-0.3	No
Big4_d	0.09	0.3-0.4	No	Big4_d	0.09	0.1-0.2	No	Big4_d	0.43	0.3-0.4	No	Big4_d	0.43	0.2-0.3	No
AuditOp_d	0.45	0.3-0.4	No	AuditOp_d	0.22	0.2-0.3	No	AuditOp_d	0.39	0.2-0.3	No	AuditOp_d	0.57	0.2-0.3	No
AbAccr1	0.36	0.5-0.6	No	AbAccr2	0.57	0.6-0.7	Yes	AbAccr2	0.26	0.6-0.7	No	AbAccr1	0.83	0.7-0.8	Yes
Age	0.27	0.5-0.6	No	Age	0.35	0.6-0.7	No	Age	0.65	0.5-0.6	Yes	Age	0.61	0.4-0.5	No
SGA_sale	0.27	0.8-0.9	No	SGA_sale	0.13	0.6-0.7	No	SGA_at	0.74	0.7-0.8	Yes	SGA_sale	0.30	0.5-0.6	No
XAD_sale	0.45	0.4-0.5	No	XAD_sale	0.17	0.5-0.6	No	XAD_at	0.22	0.4-0.5	No	XAD_at	0.39	0.4-0.5	No
SPI_sale	0.45	0.7-0.8	No	SPI_at	0.39	0.7-0.8	No	SPI_at	0.91	0.7-0.8	Yes	SPI_at	1	0.7-0.8	Yes
EI_at	0.00	0.1-0.2	No	EI_at	0.26	0.2-0.3	No	EI_at	0.13	0.2-0.3	No	EI_at	0.17	0.2-0.3	No
StdROA5	0.91	0.7-0.8	Yes	StdCF5	0.22	0.7-0.8	No	StdCF5	0.43	0.6-0.7	No	StdCF5	0.70	0.5-0.6	Yes
StdMRn5	0.18	0.5-0.6	No	StdMRn5	0.52	0.6-0.7	Yes	StdMRn5	0.35	0.5-0.6	No	StdMRn5	0.61	0.4-0.5	No

Table 4.5 Variable selection for ETR models using median threshold

This table summarizes the variable performance in the models of annual cash ETR, annual GAAP ETR, long-run cash ETR and long-run GAAP ETR using Adaptive LASSO and random forest with median threshold. Column AL shows if the variable performance in Adaptive LASSO test is in the top 50%. Column RF shows if the variable importance loading in random forest analysis is in the top 50%. Column MT shows if the variable meets the median threshold. Appendix 4A provides variable definitions.

Panel A: Annual cash ETR				Panel B: Annual GAAP ETR				Panel C: Long-run cash ETR				Panel D: Long-run GAAP ETR			
Variables	AL	RF	MT	Variables	AL	RF	Median	Variables	AL	RF	MT	Variables	AL	RF	MT
Size_mv	Yes	Yes	Yes	Size_mv	No	Yes	No	Size_mv	No	Yes	No	Size_mv	Yes	Yes	Yes
EMP	Yes	Yes	Yes	EMP	No	Yes	No	EMP	Yes	Yes	Yes	EMP	Yes	Yes	Yes
HHI_sicc	No	Yes	No	HHI_sicc	No	Yes	No	HHI_sicc	No	Yes	No	HHI_sicc	No	Yes	No
ROA_iblat	Yes	Yes	Yes	ROA_iblat	Yes	Yes	Yes	ROA_iblat	Yes	Yes	Yes	ROA_iblat	Yes	Yes	Yes
OCFf_lat	Yes	Yes	Yes	OCFf_lat	Yes	Yes	Yes	OCFf_lat	Yes	Yes	Yes	OCFf_lat	Yes	Yes	Yes
DNOL	Yes	Yes	Yes	DNOL	No	Yes	No	DNOL	No	Yes	No	DNOL	No	No	No
NOL1_d	Yes	No	No	NOL1_d	Yes	No	No	NOL1_d	Yes	No	No	NOL1_d	Yes	No	No
Return	Yes	No	No	Return	Yes	No	No	Return	Yes	No	No	Return	No	No	No
MTB	No	No	No	MTB	Yes	Yes	Yes	MTB	No	No	No	MTB	No	Yes	No
DSale	Yes	Yes	Yes	DSale	No	Yes	No	DSale	No	Yes	No	DSale	Yes	Yes	Yes
MA_d	Yes	No	No	MA_d	No	No	No	MA_d	No	No	No	MA_d	No	No	No
CAPX_lat	Yes	Yes	Yes	CAPX_lat	Yes	Yes	Yes	CAPX_lat	Yes	Yes	Yes	CAPX_lat	Yes	Yes	Yes
DPPEGT	No	Yes	No	DPPEGT	No	Yes	No	DPPEGT	No	Yes	No	DPPEGT	No	Yes	No
Invt_lat	No	Yes	No	Invt_lat	No	Yes	No	Invt_lat	Yes	Yes	Yes	Invt_lat	Yes	Yes	Yes
DP_at	Yes	Yes	Yes	DP_at	No	Yes	No	DP_at	Yes	Yes	Yes	DP_at	No	Yes	No
DGW_at	No	No	No	DGW_at	No	No	No	DGW_at	No	No	No	DGW_at	No	No	No
RND_at	Yes	No	No	RND_sale	Yes	No	No	RND_sale	Yes	No	No	RND_sale	Yes	No	No
Intan_sale	Yes	No	No	Intan_sale	Yes	No	No	Intan_sale	Yes	No	No	Intan_sale	Yes	No	No
DLTT_lat	Yes	Yes	Yes	DLTT_lat	No	Yes	No	DLTT_lat	Yes	Yes	Yes	DLTT_lat	Yes	Yes	Yes
INT_lat	Yes	Yes	Yes	INT_lat	No	Yes	No	INT_lat	Yes	Yes	Yes	INT_lat	No	Yes	No
Mezzanine	Yes	No	No	Mezzanine	No	No	No	Mezzanine	Yes	No	No	Mezzanine	No	No	No
Fin_d	No	No	No	Fin_d	No	No	No	Fin_d	No	No	No	Fin_d	No	No	No

IO_ts	No	No	No	IO_ts	No	No	No	IO_ts	No	No	No	IO_ts	No	No	No
CHE_at	No	Yes	No	CHE_at	Yes	Yes	Yes	CHE_at	No	Yes	No	CHE_at	Yes	Yes	Yes
SA_HP2010	No	Yes	No	SA_HP2010	No	Yes	No	SA_HP2010	No	Yes	No	SA_HP2010	No	Yes	No
AltmanZ	Yes	Yes	Yes	AltmanZ	Yes	Yes	Yes	AltmanZ	Yes	Yes	Yes	AltmanZ	Yes	Yes	Yes
PIFO_pi	Yes	Yes	Yes	PIFOC_at	Yes	No	No	PIFOC_at	Yes	No	No	PIFO_pi	Yes	No	No
PIFO_d	No	No	No	PIFO_d	Yes	No	No	PIFO_d	Yes	No	No	PIFO_d	Yes	No	No
NBS2	No	No	No	NBS2	No	No	No	NBS2	No	No	No	NBS2	Yes	No	No
ESUB_lat	No	No	No	ESUB_lat	Yes	No	No	ESUB_lat	No	No	No	ESUB_lat	No	No	No
ESUB_d	No	No	No	ESUB_d	No	No	No	ESUB_d	No	No	No	ESUB_d	No	No	No
StdEarnFst	Yes	No	No	StdEarnFst	Yes	No	No	StdEarnFst	Yes	No	No	StdEarnFst	No	No	No
Num_Analyst	No	No	No	Num_Analyst	No	No	No	Num_Analyst	No	No	No	Num_Analyst	No	No	No
Goodnews_d	Yes	No	No	Goodnews_d	No	No	No	Goodnews_d	No	No	No	Goodnews_d	Yes	No	No
Big4_d	No	No	No	Big4_d	No	No	No	Big4_d	No	No	No	Big4_d	No	No	No
AuditOp_d	No	No	No	AuditOp_d	Yes	No	No	AuditOp_d	No	No	No	AuditOp_d	No	No	No
AbAccr1	No	Yes	No	AbAccr2	Yes	Yes	Yes	AbAccr1	Yes	Yes	Yes	AbAccr2	No	Yes	No
Age	No	Yes	No	Age	Yes	Yes	Yes	Age	No	Yes	No	Age	Yes	Yes	Yes
SGA_at	Yes	Yes	Yes	SGA_at	Yes	Yes	Yes	SGA_at	Yes	Yes	Yes	SGA_at	Yes	Yes	Yes
XAD_sale	No	No	No	XAD_sale	Yes	No	No	XAD_sale	No	No	No	XAD_avat	No	No	No
SPI_sale	Yes	Yes	Yes	SPI_at	Yes	Yes	Yes	SPI_sale	Yes	Yes	Yes	SPI_sale	Yes	Yes	Yes
EI_at	No	No	No	EI_at	No	No	No	EI_at	No	No	No	EI_at	No	No	No
StdCF5	No	Yes	No	StdCF5	Yes	Yes	Yes	StdCF5	Yes	Yes	Yes	StdCF5	Yes	Yes	Yes
StdMRn5	Yes	No	No	StdMRn5	Yes	No	No	StdMRn5	Yes	Yes	Yes	StdMRn5	Yes	Yes	Yes

Table 4.6 Variable selection for alternative tax avoidance models using median threshold

This table summarizes the variable performance in the models of UTB, DTAX, DDBTD and MPBTD using Adaptive LASSO and random forest with median threshold. Column AL shows if the variable performance in Adaptive LASSO test is in the top 50%. Column RF shows if the variable importance loading in random forest analysis is in the top 50%. Column MT shows if the variable meets the median threshold. Appendix 4A provides variable definitions.

Panel A: UTB				Panel B: DTAX				Panel C: DDBTD				Panel D: MPBTD			
Variables	AL	RF	MT	Variables	AL	RF	MT	Variables	AL	RF	MT	Variables	AL	RF	MT
Size_mv	No	Yes	No	Size_mv	No	Yes	No	Size_mv	No	Yes	No	Size_mv	No	No	No
EMP	Yes	Yes	Yes	EMP	Yes	Yes	Yes	EMP	Yes	Yes	Yes	EMP	No	Yes	No
HHI_sicc	No	Yes	No	HHI_sicc	Yes	Yes	Yes	HHI_sicc	No	Yes	No	HHI_sicc	No	Yes	No
ROA_iblat	Yes	Yes	Yes	ROA_iblat	Yes	Yes	Yes	ROA_iblat	Yes	Yes	Yes	ROA_iblat	Yes	Yes	Yes
OCFf_lat	Yes	Yes	Yes	OCFf_lat	Yes	Yes	Yes	OCFf_lat	Yes	Yes	Yes	OCFf_lat	Yes	Yes	Yes
DNOL	No	Yes	No	DNOL	Yes	Yes	Yes	DNOL	Yes	No	No	DNOL	Yes	Yes	Yes
NOL1_d	No	No	No	NOL1_d	No	No	No	NOL2_d	Yes	No	No	NOL2_d	Yes	No	No
Return	No	No	No	Return	Yes	No	No	Return	No	No	No	Return	Yes	No	No
MTB	Yes	No	No	MTB	Yes	Yes	Yes	MTB	No	Yes	No	MTB	No	Yes	No
DSale	Yes	Yes	Yes	DSale	Yes	Yes	Yes	DSale	Yes	Yes	Yes	DSale	No	Yes	No
MA_d	No	No	No	MA_d	No	No	No	MA_d	No	No	No	MA_d	No	No	No
CAPX_lat	Yes	Yes	Yes	CAPX_lat	Yes	Yes	Yes	CAPX_at	Yes	Yes	Yes	CAPX_at	Yes	Yes	Yes
DPPEGT	Yes	Yes	Yes	DPPEGT	No	Yes	No	PPEGT_at	Yes	Yes	Yes	PPEGT_at	Yes	Yes	Yes
Invt_lat	No	No	No	Invt_lat	Yes	No	No	Invt_lat	Yes	Yes	Yes	Invt_lat	Yes	Yes	Yes
DP_at	Yes	Yes	Yes	DP_at	Yes	Yes	Yes	DP_at	Yes	Yes	Yes	DP_at	Yes	Yes	Yes
DGW_at	No	No	No	DGW_at	Yes	No	No	DGW_at	No	No	No	DGW_at	No	No	No
RND_sale	Yes	Yes	Yes	RND_sale	Yes	Yes	Yes	RND_lat	Yes	Yes	Yes	RND_lat	Yes	Yes	Yes
Intan_sale	Yes	Yes	Yes	Intan_lat	Yes	Yes	Yes	Intan_sale	Yes	Yes	Yes	Intan_sale	Yes	Yes	Yes
LEV_at	Yes	Yes	Yes	DLTT_lat	No	Yes	No	LEV_at	No	No	No	LEV_at	No	No	No
INT_lat	Yes	Yes	Yes	INT_lat	Yes	No	No	INT_lat	No	No	No	INT_lat	Yes	No	No
Mezzanine	Yes	No	No	Mezzanine	No	No	No	Mezzanine	No	No	No	Mezzanine	No	No	No
Fin_d	No	No	No	Fin_d	No	No	No	Fin_d	No	No	No	Fin_d	No	No	No

IO_ts	Yes	No	No	IO_ts	No	No	No	IO_ts	Yes	No	No	IO_ts	No	No	No
CHE_at	Yes	Yes	Yes	CHE_at	No	Yes	No	CHE_at	Yes	Yes	Yes	CHE_at	Yes	Yes	Yes
SA_HP2010	No	Yes	No	SA_HP2010	No	Yes	No	SA_HP2010	No	Yes	No	SA_HP2010	No	Yes	No
AltmanZ	Yes	Yes	Yes	AltmanZ	Yes	Yes	Yes	AltmanZ	Yes	Yes	Yes	AltmanZ	Yes	Yes	Yes
PIFO_pi	Yes	Yes	Yes	PIFO_pi	Yes	Yes	Yes	PIFO_pi	Yes	Yes	Yes	PIFO_pi	Yes	Yes	Yes
PIFO_d	No	No	No	PIFO_d	No	No	No	PIFO_d	Yes	Yes	Yes	PIFO_d	No	Yes	No
NGS	Yes	No	No	NGS	Yes	No	No	NGS	No	No	No	NGS	Yes	No	No
ESUB_lat	Yes	No	No	ESUB_lat	Yes	No	No	ESUB_lat	Yes	No	No	ESUB_lat	Yes	No	No
ESUB_d	No	No	No	ESUB_d	Yes	No	No	ESUB_d	No	No	No	ESUB_d	No	No	No
StdEarnFst	No	No	No	StdEarnFst	No	No	No	StdEarnFst	No	No	No	StdEarnFst	No	No	No
Num_Analyst	Yes	No	No	Num_Analyst	Yes	Yes	Yes	Num_Analyst	Yes	No	No	Num_Analyst	No	No	No
Goodnews_d	No	No	No	Goodnews_d	No	No	No	Goodnews_d	No	No	No	Goodnews_d	No	No	No
Big4_d	No	No	No	Big4_d	No	No	No	Big4_d	No	No	No	Big4_d	No	No	No
AuditOp_d	Yes	No	No	AuditOp_d	No	No	No	AuditOp_d	No	No	No	AuditOp_d	No	No	No
AbAccr1	Yes	Yes	Yes	AbAccr2	Yes	No	No	AbAccr2	No	Yes	No	AbAccr1	Yes	Yes	Yes
Age	No	No	No	Age	Yes	No	No	Age	Yes	No	No	Age	Yes	No	No
SGA_sale	No	Yes	No	SGA_sale	No	Yes	No	SGA_at	Yes	Yes	Yes	SGA_sale	No	Yes	No
XAD_sale	Yes	No	No	XAD_sale	No	No	No	XAD_at	No	No	No	XAD_at	No	No	No
SPI_sale	Yes	Yes	Yes	SPI_at	Yes	Yes	Yes	SPI_at	Yes	Yes	Yes	SPI_at	Yes	Yes	Yes
EI_at	No	No	No	EI_at	No	No	No	EI_at	No	No	No	EI_at	No	No	No
StdROA5	Yes	Yes	Yes	StdCF5	No	Yes	No	StdCF5	No	Yes	No	StdCF5	Yes	Yes	Yes
StdMRn5	No	No	No	StdMRn5	Yes	No	No	StdMRn5	No	No	No	StdMRn5	Yes	No	No

Table 4.7 Variable selection for ETR models using both fixated and median thresholds (further analysis)

This table summarizes the variable performance in the models of annual cash ETR, annual GAAP ETR, long-run cash ETR and long-run GAAP ETR using Adaptive LASSO and random forest with fixated and median thresholds, including corporate governance variables. Appendix 4A provides variable definitions.

Panel A: Annual cash ETR			Panel B: Annual GAAP ETR			Panel C: Long-run cash ETR			Panel D: Long-run GAAP ETR		
Variables	Fixated	Median	Variables	Fixated	Median	Variables	Fixated	Median	Variables	Fixated	Median
Size_mv	No	No	Size_mv	No	No	Size_mv	No	No	Size_mv	No	No
EMP	Yes	Yes	EMP	No	Yes	EMP	Yes	Yes	EMP	No	Yes
HHI_sicc	Yes	Yes	HHI_sicc	No	No	HHI_sicc	Yes	Yes	HHI_sicc	No	No
ROA_iblat	Yes	Yes	ROA_iblat	Yes	Yes	ROA_iblat	Yes	Yes	ROA_iblat	Yes	Yes
OCFf_lat	Yes	Yes	OCFf_lat	No	Yes	OCFf_lat	Yes	Yes	OCFf_lat	Yes	Yes
DNOL	No	Yes	DNOL	No	No	DNOL	No	No	DNOL	No	No
NOL1_d	No	No	NOL1_d	No	No	NOL1_d	No	No	NOL1_d	No	No
Return	No	No	Return	No	No	Return	Yes	No	Return	No	No
MTB	No	Yes	MTB	No	Yes	MTB	No	No	MTB	No	Yes
DSale	Yes	Yes	DSale	No	Yes	DSale	No	No	DSale	No	Yes
MA_d	No	No	MA_d	No	No	MA_d	No	No	MA_d	No	No
CAPX_lat	Yes	Yes	CAPX_lat	No	Yes	CAPX_lat	Yes	Yes	CAPX_lat	Yes	Yes
DPPEGT	No	No	DPPEGT	No	No	DPPEGT	No	Yes	DPPEGT	No	No
Invt_lat	No	No	Invt_lat	No	Yes	Invt_lat	Yes	Yes	Invt_lat	No	Yes
DP_at	Yes	Yes	DP_at	No	Yes	DP_at	Yes	Yes	DP_at	No	No
DGW_at	No	No	DGW_at	No	No	DGW_at	No	No	DGW_at	No	No
RND_at	No	No	RND_sale	No	Yes	RND_sale	Yes	Yes	RND_sale	No	Yes
Intan_sale	No	Yes	Intan_sale	No	Yes	Intan_sale	Yes	Yes	Intan_sale	Yes	Yes
DLTT_lat	Yes	Yes	DLTT_lat	No	Yes	DLTT_lat	Yes	Yes	DLTT_lat	No	Yes
INT_lat	Yes	Yes	INT_lat	No	Yes	INT_lat	Yes	Yes	INT_lat	No	No
Mezzanine	No	No	Mezzanine	No	No	Mezzanine	No	No	Mezzanine	No	No
Fin_d	No	No	Fin_d	No	No	Fin_d	No	No	Fin_d	No	No
IO_ts	No	No	IO_ts	No	No	IO_ts	Yes	No	IO_ts	No	No

CHE_at	No	No	CHE_at	No	Yes	CHE_at	No	No	CHE_at	No	Yes
SA_HP2010	No	No	SA_HP2010	No	No	SA_HP2010	No	No	SA_HP2010	No	No
AltmanZ	Yes	Yes	AltmanZ	No	Yes	AltmanZ	No	Yes	AltmanZ	Yes	Yes
PIFO_pi	Yes	Yes	PIFOc_at	No	Yes	PIFOc_at	Yes	Yes	PIFO_pi	Yes	Yes
PIFO_d	No	No	PIFO_d	No	No	PIFO_d	No	No	PIFO_d	No	No
NBS2	No	No	NBS2	No	No	NBS2	No	No	NBS2	No	No
ESUB_lat	No	No	ESUB_lat	No	No	ESUB_lat	No	No	ESUB_lat	No	No
ESUB_d	No	No	ESUB_d	No	No	ESUB_d	No	No	ESUB_d	No	No
StdEarnFst	No	No	StdEarnFst	No	No	StdEarnFst	Yes	No	StdEarnFst	No	No
Num_Analyst	No	Yes	Num_Analyst	No	No	Num_Analyst	No	No	Num_Analyst	No	No
Goodnews_d	No	No	Goodnews_d	No	No	Goodnews_d	No	No	Goodnews_d	No	No
Big4_d	No	No	Big4_d	No	No	Big4_d	No	No	Big4_d	No	No
AuditOp_d	No	No	AuditOp_d	No	No	AuditOp_d	No	No	AuditOp_d	No	No
AbAccr1	No	No	AbAccr2	No	No	AbAccr1	No	No	AbAccr2	No	Yes
Age	No	No	Age	No	Yes	Age	Yes	Yes	Age	Yes	Yes
SGA_at	Yes	Yes	SGA_at	No	Yes	SGA_at	Yes	Yes	SGA_at	No	Yes
XAD_sale	No	No	XAD_sale	No	No	XAD_sale	Yes	Yes	XAD_avat	No	Yes
SPI_sale	Yes	Yes	SPI_at	No	Yes	SPI_sale	No	Yes	SPI_sale	Yes	Yes
EI_at	No	No	EI_at	No	No	EI_at	No	No	EI_at	No	No
StdCF5	No	No	StdCF5	No	Yes	StdCF5	No	No	StdCF5	No	Yes
StdMRn5	No	Yes	StdMRn5	No	Yes	StdMRn5	Yes	Yes	StdMRn5	No	Yes
BrdSize	No	No	BrdSize	No	No	BrdSize	No	No	BrdSize	No	No
Ind_dir	No	No	Ind_dir	No	No	Ind_dir	No	No	Ind_dir	No	No
Female_d	No	No	Female_d	No	No	Female_d	No	No	Female_d	No	No
CEOage	No	No	CEOage	No	No	CEOage	No	No	CEOage	No	No
Tenure_CEO	No	No	Tenure_CEO	No	No	Tenure_CEO	No	No	Tenure_CEO	No	No
StkCompen_d	No	No	StkCompen_d	No	No	StkCompen_d	No	No	StkCompen_d	No	No
Stk_ceo	No	No	Stk_ceo	No	No	Stk_ceo	No	No	Stk_ceo	No	No

Cash_compen	No	No	Cash_compen	No	No	Cash_compen	No	No	Cash_compen	No	No
CEOChair_d	No	No	CEOChair_d	No	No	CEOChair_d	No	No	CEOChair_d	No	No

Table 4.8 Variable selection for alternative tax avoidance models using both fixated and median thresholds (further analysis)

This table summarizes the variable performance in the models of UTB, DTAX, DDBTD and MPBTD using Adaptive LASSO and random forest with fixated and median thresholds, including corporate governance variables. Appendix 4A provides variable definitions.

Panel A: UTB			Panel B: DTAX			Panel C: DDBTD			Panel D: MPBTD		
Variables	Fixated	Median	Variables	Fixated	Median	Variables	Fixated	Median	Variables	Fixated	Median
Size_mv	No	Yes	Size_mv	No	Yes	Size_mv	No	Yes	Size_mv	No	Yes
EMP	Yes	Yes	EMP	Yes	Yes	EMP	Yes	Yes	EMP	No	No
HHI_sicc	No	No	HHI_sicc	No	No	HHI_sicc	No	No	HHI_sicc	No	No
ROA_iblat	Yes	Yes	ROA_iblat	Yes	Yes	ROA_iblat	No	No	ROA_iblat	No	No
OCFf_lat	No	No	OCFf_lat	Yes	Yes	OCFf_lat	Yes	Yes	OCFf_lat	No	Yes
DNOL	Yes	Yes	DNOL	Yes	Yes	DNOL	No	Yes	DNOL	No	Yes
NOL1_d	No	No	NOL1_d	No	No	NOL2_d	No	No	NOL2_d	No	No
Return	No	No	Return	No	No	Return	No	No	Return	No	No
MTB	Yes	Yes	MTB	Yes	Yes	MTB	No	Yes	MTB	No	Yes
DSale	No	Yes	DSale	Yes	Yes	DSale	No	Yes	DSale	No	Yes
MA_d	No	No	MA_d	No	No	MA_d	No	No	MA_d	No	No
CAPX_lat	Yes	Yes	CAPX_lat	Yes	Yes	CAPX_at	Yes	Yes	CAPX_at	Yes	Yes
DPPEGT	Yes	Yes	DPPEGT	No	Yes	PPEGT_at	No	No	PPEGT_at	Yes	Yes
Invt_lat	Yes	Yes	Invt_lat	No	Yes	Invt_lat	No	Yes	Invt_lat	No	Yes
DP_at	No	No	DP_at	No	Yes	DP_at	Yes	Yes	DP_at	Yes	Yes
DGW_at	No	No	DGW_at	No	No	DGW_at	No	No	DGW_at	No	No
RND_sale	Yes	Yes	RND_sale	Yes	Yes	RND_lat	No	Yes	RND_lat	No	Yes
Intan_sale	Yes	Yes	Intan_lat	Yes	Yes	Intan_sale	Yes	Yes	Intan_sale	No	Yes
LEV_at	Yes	Yes	DLTT_lat	No	Yes	LEV_at	No	No	LEV_at	No	No
INT_lat	No	Yes	INT_lat	No	Yes	INT_lat	No	No	INT_lat	No	No
Mezzanine	No	No	Mezzanine	No	No	Mezzanine	No	No	Mezzanine	No	No
Fin_d	No	No	Fin_d	No	No	Fin_d	No	No	Fin_d	No	No
IO_ts	No	No	IO_ts	No	No	IO_ts	No	No	IO_ts	No	No

CHE_at	No	Yes	CHE_at	No	Yes	CHE_at	Yes	Yes	CHE_at	No	No
SA_HP2010	No	No	SA_HP2010	No	Yes	SA_HP2010	Yes	Yes	SA_HP2010	No	Yes
AltmanZ	Yes	Yes	AltmanZ	Yes	Yes	AltmanZ	No	Yes	AltmanZ	No	Yes
PIFO_pi	Yes	Yes	PIFO_pi	Yes	Yes	PIFO_pi	Yes	Yes	PIFO_pi	Yes	Yes
PIFO_d	No	No	PIFO_d	No	No	PIFO_d	Yes	Yes	PIFO_d	Yes	Yes
NGS	Yes	Yes	NGS	No	No	NGS	No	No	NGS	No	No
ESUB_lat	No	No	ESUB_lat	No	No	ESUB_lat	No	No	ESUB_lat	No	No
ESUB_d	No	No	ESUB_d	No	No	ESUB_d	No	No	ESUB_d	No	No
StdEarnFst	No	No	StdEarnFst	No	No	StdEarnFst	No	No	StdEarnFst	No	No
Num_Analyst	Yes	Yes	Num_Analyst	No	Yes	Num_Analyst	No	No	Num_Analyst	No	No
Goodnews_d	No	No	Goodnews_d	No	No	Goodnews_d	No	No	Goodnews_d	No	No
Big4_d	No	No	Big4_d	No	No	Big4_d	No	No	Big4_d	No	No
AuditOp_d	No	No	AuditOp_d	No	No	AuditOp_d	No	No	AuditOp_d	No	No
AbAccr1	No	No	AbAccr2	No	Yes	AbAccr2	No	No	AbAccr1	Yes	Yes
Age	No	No	Age	No	No	Age	No	Yes	Age	No	Yes
SGA_sale	No	No	SGA_sale	No	No	SGA_at	Yes	Yes	SGA_sale	Yes	Yes
XAD_sale	No	No	XAD_sale	No	No	XAD_at	No	No	XAD_at	No	No
SPI_sale	No	Yes	SPI_at	No	No	SPI_at	Yes	Yes	SPI_at	Yes	Yes
EI_at	No	No	EI_at	No	No	EI_at	No	No	EI_at	No	No
StdROA5	Yes	Yes	StdCF5	No	Yes	StdCF5	No	Yes	StdCF5	Yes	Yes
StdMRn5	No	No	StdMRn5	No	Yes	StdMRn5	No	No	StdMRn5	No	Yes
BrdSize	No	No	BrdSize	No	No	BrdSize	No	No	BrdSize	No	No
Ind_dir	No	No	Ind_dir	No	No	Ind_dir	No	No	Ind_dir	No	No
Female_d	No	No	Female_d	No	No	Female_d	No	No	Female_d	No	No
CEOage	No	No	CEOage	No	No	CEOage	No	No	CEOage	No	No
Tenure_CEO	No	No	Tenure_CEO	No	No	Tenure_CEO	No	No	Tenure_CEO	No	No
StkCompen_d	No	No	StkCompen_d	No	No	StkCompen_d	No	No	StkCompen_d	No	No
Stk_ceo	No	No	Stk_ceo	No	No	Stk_ceo	No	No	Stk_ceo	No	No

Cash_compen	No	No	Cash_compen	No	No	Cash_compen	No	No	Cash_compen	No	No
CEOChair_d	No	No	CEOChair_d	No	No	CEOChair_d	No	No	CEOChair_d	No	No

Table 4.9 Fixed-effects tests for the key variables of ETR models

This table reports the results of regressing ETR measures on the identified key variables meeting 50% fixated threshold, with industry fixed effects, firm fixed effects, year fixed effects, industry-year fixed effects and firm-year fixed effects. Panel A presents the results of annual cash ETR model. Panel B presents the results of annual GAAP ETR model. Panel C presents the results of long-run cash ETR model. Panel D presents the results of long-run GAAP ETR model. Variable definitions are provided in Appendix 4A. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the *t*-statistic is reported in the parentheses.

Panel A: Annual cash ETR model

	(1)	(2)	(3)	(4)	(5)	(6)
	Cash ETR1	Cash ETR1	Cash ETR1	Cash ETR1	Cash ETR1	Cash ETR1
Size_mv	-0.021*** (-21.80)	-0.023*** (-20.98)	0.005** (1.96)	-0.012*** (-12.37)	-0.011*** (-10.08)	0.009*** (3.52)
EMP	0.019*** (20.65)	0.020*** (18.05)	0.016*** (5.45)	0.012*** (12.62)	0.009*** (7.68)	0.001 (0.17)
HHI_sicc	-0.006 (-1.14)	-0.008 (-1.49)	0.002 (0.14)	0.007 (1.38)	0.009* (1.73)	0.019 (1.64)
ROA_iblat	0.090*** (3.01)	0.039 (1.31)	-0.184*** (-5.63)	0.014 (0.50)	-0.040 (-1.44)	-0.186*** (-5.65)
OCFf_lat	-0.233*** (-8.65)	-0.176*** (-6.70)	-0.150*** (-5.83)	-0.175*** (-7.24)	-0.116*** (-4.97)	-0.144*** (-5.53)
DNOL	0.091*** (6.96)	0.088*** (6.79)	0.052*** (4.04)	0.087*** (6.90)	0.084*** (6.72)	0.051*** (4.07)
Return	-0.027*** (-15.40)	-0.027*** (-15.05)	-0.029*** (-15.93)	-0.032*** (-17.53)	-0.032*** (-17.32)	-0.030*** (-16.07)
MTB	0.000 (1.62)	0.001* (1.88)	0.000 (0.32)	0.001* (1.91)	0.000 (1.58)	0.000 (1.49)
DSale	-0.024*** (-6.60)	-0.023*** (-6.49)	-0.028*** (-6.53)	-0.022*** (-6.15)	-0.022*** (-6.11)	-0.026*** (-6.02)
CAPX_lat	-0.264*** (-8.12)	-0.169*** (-5.03)	-0.022 (-0.61)	-0.267*** (-8.72)	-0.183*** (-5.81)	-0.067* (-1.87)
DPPEGT	0.017 (0.91)	0.019 (1.03)	0.009 (0.43)	0.018 (0.97)	0.023 (1.25)	0.006 (0.29)
Invt_lat	0.052*** (7.43)	0.047*** (5.28)	0.032** (1.97)	0.046*** (6.75)	0.039*** (4.46)	0.010 (0.61)
DP_at	0.122*** (2.68)	0.244*** (5.17)	0.429*** (5.74)	0.088** (1.97)	0.223*** (4.79)	0.431*** (5.79)
DLTT_lat	-0.069*** (-9.56)	-0.068*** (-9.28)	-0.028*** (-3.06)	-0.024*** (-3.18)	-0.021*** (-2.80)	-0.012 (-1.24)
INT_lat	0.130 (1.42)	-0.026 (-0.28)	0.091 (0.75)	-0.518*** (-5.23)	-0.666*** (-6.65)	-0.401*** (-3.19)
CHE_at	0.018*** (3.41)	0.020*** (3.79)	0.034*** (4.71)	0.015*** (2.89)	0.014*** (2.71)	0.028*** (3.95)
SA_HP2010	0.015*** (8.61)	0.013*** (7.14)	0.082*** (16.69)	0.005*** (2.64)	-0.000 (-0.24)	-0.053*** (-6.75)
AltmanZ	0.002*** (9.21)	0.002*** (10.11)	0.001*** (5.12)	0.001*** (7.19)	0.002*** (7.65)	0.001*** (3.41)
PIFO_pi	0.031*** (8.60)	0.040*** (10.39)	0.062*** (11.63)	0.040*** (10.87)	0.052*** (13.44)	0.070*** (13.16)
AbAccr1	-0.084*** (-3.35)	-0.006 (-0.23)	0.076*** (3.01)	-0.034 (-1.45)	0.049** (2.03)	0.085*** (3.31)
SGA_at	0.031*** (6.13)	0.030*** (5.18)	0.036*** (2.77)	0.036*** (7.24)	0.035*** (6.14)	0.073*** (5.69)

SPI_sale	0.566*** (13.99)	0.554*** (13.77)	0.492*** (11.17)	0.559*** (14.07)	0.546*** (13.88)	0.465*** (10.85)
StdCF5	-0.042* (-1.90)	-0.060*** (-2.73)	-0.080** (-2.51)	-0.042* (-1.94)	-0.058*** (-2.69)	-0.034 (-1.08)
StdMRn5	-0.495*** (-22.70)	-0.473*** (-21.43)	-0.297*** (-9.88)	-0.404*** (-17.75)	-0.378*** (-16.43)	-0.209*** (-6.51)
N	45162	45162	43926	45162	45162	43926
Adj R ²	0.094	0.107	0.287	0.119	0.133	0.302
Industry FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Panel B1: Annual GAAP ETR model (50% fixated threshold)

	(1)	(2)	(3)	(4)	(5)	(6)
	GAAPETR1	GAAPETR1	GAAPETR1	GAAPETR1	GAAPETR1	GAAPETR1
ROA_iblat	-0.361*** (-57.03)	-0.383*** (-56.08)	-0.444*** (-50.20)	-0.362*** (-57.91)	-0.381*** (-56.37)	-0.463*** (-52.24)
CAPX_lat	0.179*** (33.46)	0.146*** (24.66)	0.172*** (25.33)	0.136*** (25.77)	0.099*** (16.74)	0.124*** (18.18)
SPI_at	1.941*** (45.53)	2.055*** (46.04)	2.385*** (54.79)	1.887*** (44.99)	2.000*** (45.54)	2.371*** (54.92)
N	141967	141957	139130	141967	141957	139130
Adj R ²	0.072	0.166	0.435	0.102	0.190	0.448
Ind FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Panel B2: Annual GAAP ETR model (alternative fixated threshold)

	(1)	(2)	(3)	(4)	(5)	(6)
	GAAPETR1	GAAPETR1	GAAPETR1	GAAPETR1	GAAPETR1	GAAPETR1
EMP	0.003*** (6.59)	0.001** (2.08)	0.005*** (3.52)	0.003*** (5.72)	0.000 (0.30)	-0.002 (-1.18)
ROA_iblat	-0.483*** (-31.98)	-0.482*** (-31.84)	-0.647*** (-37.31)	-0.492*** (-32.75)	-0.487*** (-32.37)	-0.631*** (-36.67)
OCFf_lat	0.109*** (11.40)	0.108*** (11.19)	0.132*** (13.46)	0.125*** (12.89)	0.120*** (12.25)	0.117*** (12.26)
MTB	0.000* (1.82)	0.000 (0.45)	0.001*** (5.36)	0.001*** (4.80)	0.001*** (3.19)	0.002*** (7.38)
DSale	0.007*** (3.51)	0.005*** (2.83)	0.008*** (3.95)	0.005*** (2.60)	0.004** (2.03)	0.008*** (3.94)
CAPX_lat	0.181*** (11.92)	0.204*** (13.26)	0.251*** (15.43)	0.149*** (9.92)	0.163*** (10.70)	0.217*** (13.57)
DPPEGT	0.002 (0.23)	0.009 (0.94)	0.005 (0.56)	0.024** (2.53)	0.033*** (3.37)	0.020** (2.09)
Invt_lat	0.082*** (19.96)	0.102*** (19.32)	0.125*** (13.70)	0.061*** (15.14)	0.076*** (14.65)	0.116*** (12.78)
DP_at	0.054** (2.04)	-0.022 (-0.80)	0.053 (1.22)	-0.012 (-0.46)	-0.096*** (-3.48)	-0.014 (-0.33)
Intan_sale	-0.120*** (-3.81)	-0.098*** (-3.64)	-0.051*** (-2.89)	-0.110*** (-3.85)	-0.087*** (-3.69)	-0.048*** (-2.87)
DLTT_lat	-0.027*** (-7.52)	-0.027*** (-7.07)	-0.043*** (-8.96)	-0.027*** (-7.46)	-0.024*** (-6.50)	-0.043*** (-8.96)
CHE_at	-0.005 (-1.20)	-0.007* (-1.67)	0.021*** (4.64)	0.005 (1.17)	0.001 (0.33)	0.021*** (4.83)
SA_HP2010	-0.008*** (-4.38)	-0.011*** (-5.96)	-0.022*** (-6.69)	-0.022*** (-11.83)	-0.026*** (-13.57)	-0.083*** (-19.27)
AltmanZ	0.002*** (14.30)	0.002*** (15.83)	0.002*** (9.08)	0.002*** (13.11)	0.002*** (14.57)	0.001*** (9.18)
PIFOc_at	-0.593*** (-18.44)	-0.507*** (-15.77)	-0.588*** (-11.52)	-0.474*** (-15.01)	-0.390*** (-12.19)	-0.484*** (-9.43)
ESUB_lat	-0.921*** (-6.71)	-0.840*** (-6.16)	-0.696*** (-3.82)	-0.941*** (-6.91)	-0.893*** (-6.60)	-0.668*** (-3.70)
Num_Analyst	0.002*** (2.68)	0.003*** (3.98)	0.011*** (8.35)	0.001 (0.75)	0.002* (1.77)	0.005*** (4.13)
AbAccr2	-0.048*** (-8.37)	-0.054*** (-9.31)	-0.028*** (-5.27)	-0.038*** (-6.57)	-0.046*** (-7.92)	-0.029*** (-5.58)
Age	-0.016*** (-10.31)	-0.017*** (-10.60)	-0.059*** (-20.77)	-0.019*** (-12.03)	-0.019*** (-11.92)	-0.027*** (-9.08)
SGA_at	0.064***	0.064***	-0.007	0.060***	0.059***	0.005

	(16.45)	(16.30)	(-0.88)	(16.10)	(15.32)	(0.72)
SPI_at	2.661***	2.667***	2.937***	2.643***	2.645***	2.903***
	(48.09)	(47.99)	(54.41)	(48.85)	(48.65)	(54.28)
StdCF5	-0.103***	-0.123***	-0.140***	-0.114***	-0.134***	-0.114***
	(-7.52)	(-8.96)	(-7.58)	(-8.41)	(-9.80)	(-6.25)
StdMRn5	-0.216***	-0.206***	-0.173***	-0.215***	-0.215***	-0.206***
	(-16.94)	(-15.97)	(-10.06)	(-16.22)	(-15.99)	(-11.12)
N	78524	78524	76917	78524	78524	76917
Adj R ²	0.143	0.166	0.402	0.174	0.196	0.416
Industry FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Panel C: long-run cash ETR

	(1)	(2)	(3)	(4)	(5)	(6)
	CashETR3	CashETR3	CashETR3	CashETR3	CashETR3	CashETR3
Size_mv	-0.019*** (-18.19)	-0.018*** (-14.91)	-0.010*** (-4.62)	-0.012*** (-10.81)	-0.010*** (-7.70)	0.013*** (5.31)
EMP	0.017*** (17.12)	0.016*** (14.11)	0.019*** (6.88)	0.013*** (12.19)	0.011*** (8.84)	0.018*** (6.68)
HHI_sicc	0.018*** (3.80)	0.012** (2.27)	-0.009 (-0.87)	0.023*** (4.83)	0.017*** (3.39)	-0.002 (-0.22)
ROA_iblat	0.269*** (10.11)	0.278*** (9.90)	0.188*** (6.21)	0.185*** (6.70)	0.183*** (6.49)	0.066** (2.17)
OCFf_lat	-0.106*** (-5.10)	-0.135*** (-5.74)	-0.097*** (-3.83)	-0.040* (-1.94)	-0.062*** (-2.75)	-0.039 (-1.56)
NOL1_d	-0.029*** (-14.82)	-0.031*** (-16.12)	-0.024*** (-9.51)	-0.020*** (-9.76)	-0.022*** (-10.66)	-0.012*** (-4.50)
Return	-0.012*** (-6.26)	-0.010*** (-5.58)	-0.009*** (-5.01)	-0.015*** (-7.67)	-0.014*** (-7.09)	-0.014*** (-7.67)
MTB	0.000 (0.54)	-0.000 (-0.94)	-0.000 (-0.90)	0.000 (0.39)	-0.000 (-1.27)	-0.000 (-1.32)
CAPX_lat	-0.299*** (-11.88)	-0.297*** (-10.35)	-0.100*** (-3.28)	-0.277*** (-11.11)	-0.274*** (-9.89)	-0.123*** (-4.06)
Invt_lat	0.060*** (8.82)	0.059*** (6.46)	0.034** (2.16)	0.059*** (8.73)	0.053*** (5.82)	-0.011 (-0.72)
DP_at	0.280*** (6.26)	0.280*** (5.82)	0.517*** (6.82)	0.252*** (5.65)	0.274*** (5.75)	0.514*** (6.81)
RND_sale	-0.138 (-0.76)	-0.124 (-0.69)	-0.315 (-1.01)	-0.286 (-1.38)	-0.296 (-1.48)	-0.439 (-1.44)
Intan_sale	0.041 (0.21)	0.039 (0.21)	0.283 (0.95)	0.166 (0.71)	0.192 (0.88)	0.408 (1.39)
DLTT_lat	-0.053*** (-7.15)	-0.049*** (-6.60)	-0.050*** (-5.84)	-0.023*** (-2.97)	-0.019** (-2.44)	-0.016* (-1.82)
INT_lat	-0.205** (-2.15)	-0.191** (-1.97)	0.244** (2.06)	-0.679*** (-6.33)	-0.668*** (-6.22)	-0.266** (-2.20)
IO_ts	-0.043*** (-11.01)	-0.051*** (-12.12)	-0.055*** (-8.78)	-0.017*** (-3.89)	-0.025*** (-5.62)	-0.018*** (-2.78)
CHE_at	0.016*** (2.76)	0.012** (2.12)	0.004 (0.62)	0.010 (1.63)	0.006 (1.06)	0.007 (1.12)
AltmanZ	0.000 (1.14)	0.000 (1.25)	-0.000 (-0.42)	-0.000 (-0.01)	-0.000 (-0.19)	-0.001*** (-4.01)
PIFOc_at	-0.281*** (-6.61)	-0.275*** (-6.12)	-0.449*** (-6.60)	-0.274*** (-6.28)	-0.232*** (-5.16)	-0.299*** (-4.35)
StdEarnFst	-0.009 (-1.12)	-0.011 (-1.38)	-0.003 (-0.39)	-0.014* (-1.69)	-0.015* (-1.76)	-0.003 (-0.36)
AbAccr1	-0.072*** (-3.44)	-0.103*** (-4.29)	-0.060** (-2.46)	-0.012 (-0.58)	-0.028 (-1.18)	0.012 (0.50)
SGA_at	0.032*** (5.73)	0.023*** (3.73)	-0.017 (-1.36)	0.031*** (5.43)	0.023*** (3.70)	-0.004 (-0.35)
XAD_sale	0.080 (0.39)	0.031 (0.16)	-0.466 (-1.53)	-0.063 (-0.26)	-0.167 (-0.72)	-0.683** (-2.30)
SPI_sale	0.270*** (7.20)	0.275*** (7.36)	0.207*** (5.43)	0.266*** (7.12)	0.267*** (7.24)	0.172*** (4.68)
StdCF5	-0.118*** (-4.59)	-0.147*** (-5.69)	-0.173*** (-5.10)	-0.113*** (-4.40)	-0.140*** (-5.44)	-0.163*** (-4.88)
StdMRn5	-0.323***	-0.354***	-0.329***	-0.211***	-0.245***	-0.162***

	(-14.73)	(-15.79)	(-11.52)	(-8.82)	(-10.08)	(-5.17)
N	41186	41186	40211	41186	41186	40211
Adj R ²	0.091	0.106	0.337	0.107	0.122	0.357
Industry FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Panel D: long-run GAAP ETR

	(1)	(2)	(3)	(4)	(5)	(6)
	GAAPETR3	GAAPETR3	GAAPETR3	GAAPETR3	GAAPETR3	GAAPETR3
Size_mv	-0.001*** (-3.05)	-0.004*** (-9.80)	0.002* (1.69)	0.004*** (9.28)	0.000 (0.64)	0.010*** (7.71)
ROA_iblat	-0.142*** (-10.39)	-0.123*** (-9.11)	-0.144*** (-9.13)	-0.149*** (-10.94)	-0.126*** (-9.37)	-0.147*** (-9.28)
OCFf_lat	0.095*** (12.46)	0.088*** (11.71)	0.076*** (9.16)	0.099*** (13.09)	0.089*** (11.88)	0.066*** (7.99)
DSale	-0.005*** (-2.93)	-0.001 (-0.34)	0.002 (1.30)	-0.007*** (-3.70)	-0.002 (-0.98)	0.002 (1.28)
CAPX_lat	0.247*** (24.12)	0.164*** (14.34)	0.133*** (9.47)	0.210*** (20.81)	0.133*** (11.69)	0.113*** (8.09)
RND_sale	-0.125*** (-4.50)	-0.113*** (-4.06)	-0.071 (-1.24)	-0.134*** (-4.92)	-0.115*** (-4.20)	-0.053 (-0.92)
Intan_sale	0.037 (1.47)	0.045* (1.73)	0.040 (0.74)	0.050** (2.01)	0.050** (1.98)	0.021 (0.39)
AltmanZ	0.001*** (7.83)	0.001*** (11.46)	0.001*** (3.86)	0.001*** (7.01)	0.001*** (10.59)	0.001*** (3.27)
PIFO_pi	-0.018*** (-7.43)	-0.021*** (-8.52)	-0.022*** (-6.74)	-0.011*** (-4.70)	-0.015*** (-6.10)	-0.016*** (-4.94)
Age	0.016*** (16.52)	0.006*** (6.48)	-0.038*** (-15.85)	0.020*** (21.22)	0.011*** (11.27)	0.008** (2.54)
SGA_at	0.093*** (26.33)	0.036*** (9.00)	-0.023*** (-2.65)	0.085*** (24.47)	0.032*** (7.96)	-0.024*** (-2.75)
SPI_sale	0.238*** (10.04)	0.208*** (8.76)	0.125*** (4.93)	0.232*** (9.84)	0.204*** (8.60)	0.107*** (4.21)
StdCF5	-0.080*** (-4.99)	-0.125*** (-7.89)	-0.070*** (-3.47)	-0.067*** (-4.25)	-0.112*** (-7.15)	-0.051** (-2.54)
StdMRn5	-0.092*** (-6.48)	-0.222*** (-15.53)	-0.124*** (-6.37)	-0.067*** (-4.50)	-0.212*** (-14.19)	-0.163*** (-7.77)
N	77502	77502	76045	77502	77502	76045
Adj R ²	0.035	0.125	0.355	0.058	0.142	0.362
Industry FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Table 4.10 Fixed-effects tests for the key variables of alternative tax avoidance models

This table reports the results of regressing the alternative tax avoidance measures on the identified key variables meeting 50% fixated threshold, with industry fixed effects, firm fixed effects, year fixed effects, industry-year fixed effects and firm-year fixed effects. Panel A reports the results of UTB model. Panel B reports the results of DTAX model. Panel C reports the results of DDBTD model. Panel D reports the results of MPBTD model. Variable definitions are provided in Appendix 4A. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the *t*-statistic is reported in the parentheses.

Panel A: UTB

	(1)	(2)	(3)	(4)	(5)	(6)
	UTB	UTB	UTB	UTB	UTB	UTB
EMP	0.000** (2.28)	0.000*** (2.77)	-0.002*** (-6.30)	0.000** (2.25)	0.000*** (2.87)	-0.000* (-1.82)
ROA_iblat	0.012*** (4.09)	0.011*** (3.74)	-0.006*** (-3.01)	0.012*** (4.11)	0.011*** (3.76)	-0.006*** (-2.80)
OCFf_lat	0.009*** (3.28)	0.007*** (2.86)	0.002 (1.03)	0.008*** (3.06)	0.007*** (2.65)	0.001 (0.41)
DSale	-0.001 (-1.39)	-0.001 (-1.53)	-0.000 (-0.06)	-0.001 (-1.35)	-0.001 (-1.49)	-0.000 (-0.44)
CAPX_lat	-0.011*** (-2.77)	-0.006* (-1.69)	0.001 (0.36)	-0.012*** (-3.08)	-0.007** (-1.96)	-0.001 (-0.53)
DPPEGT	-0.009*** (-4.31)	-0.008*** (-3.73)	-0.004*** (-3.10)	-0.009*** (-4.14)	-0.008*** (-3.56)	-0.005*** (-3.79)
DP_at	0.008* (1.69)	-0.007 (-1.35)	0.021** (2.23)	0.006 (1.31)	-0.009 (-1.60)	0.016* (1.71)
RND_sale	0.010* (1.70)	0.007 (1.50)	0.001 (0.70)	0.010* (1.71)	0.007 (1.51)	0.001 (0.57)
LEV_at	-0.002*** (-3.10)	-0.001* (-1.79)	-0.003*** (-2.62)	-0.001* (-1.79)	-0.000 (-0.54)	-0.001 (-0.53)
PIFO_pi	0.005*** (15.30)	0.004*** (13.74)	0.000 (0.56)	0.005*** (15.60)	0.004*** (14.03)	0.000 (1.10)
NGS	0.003*** (9.31)	0.002*** (6.19)	-0.001* (-1.72)	0.003*** (9.56)	0.002*** (6.39)	-0.000 (-0.24)
Num_Analyst	0.001*** (9.48)	0.002*** (10.71)	-0.001*** (-3.04)	0.002*** (9.74)	0.002*** (10.90)	-0.000** (-1.99)
StdROA5	0.020*** (8.25)	0.020*** (8.28)	0.010*** (4.99)	0.020*** (8.21)	0.019*** (8.22)	0.009*** (4.45)
N	20324	20324	19779	20323	20323	19778
Adj R ²	0.117	0.139	0.720	0.128	0.150	0.727
Industry FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Panel B: DTAX

	(1)	(2)	(3)	(4)	(5)	(6)
	DTAX	DTAX	DTAX	DTAX	DTAX	DTAX
EMP	1.741*** (8.57)	1.859*** (7.62)	1.630** (2.16)	1.750*** (8.59)	1.915*** (7.82)	0.841 (1.06)
ROA_iblat	160.113*** (23.97)	163.194*** (24.01)	213.810*** (21.94)	164.450*** (24.32)	166.999*** (24.34)	216.457*** (22.05)
OCFf_lat	-16.824*** (-4.56)	-18.279*** (-4.83)	-30.304*** (-6.19)	-19.645*** (-5.21)	-20.888*** (-5.42)	-31.411*** (-6.41)
DNOL	32.455*** (8.38)	32.628*** (8.49)	34.990*** (8.33)	33.279*** (8.46)	33.498*** (8.59)	35.678*** (8.39)
MTB	-0.174 (-1.39)	-0.173 (-1.37)	-0.328* (-1.92)	-0.231* (-1.83)	-0.235* (-1.86)	-0.358** (-2.09)
CAPX_lat	-37.590*** (-7.73)	-44.659*** (-8.69)	-74.756*** (-10.05)	-36.604*** (-7.55)	-42.710*** (-8.36)	-71.371*** (-9.69)
Invt_lat	-15.897*** (-10.04)	-20.470*** (-9.90)	-38.656*** (-8.82)	-14.212*** (-9.12)	-17.619*** (-8.60)	-32.989*** (-7.47)
RND_sale	23.382*** (2.64)	21.592** (2.47)	14.363 (1.17)	23.180*** (2.60)	20.869** (2.43)	14.261 (1.16)
Intan_lat	-10.402*** (-5.56)	-11.834*** (-6.17)	-5.885* (-1.72)	-13.785*** (-7.11)	-15.744*** (-7.91)	-8.422** (-2.43)
AltmanZ	-0.617*** (-10.67)	-0.650*** (-11.11)	-0.865*** (-8.82)	-0.624*** (-10.76)	-0.660*** (-11.24)	-0.865*** (-8.78)
PIFO_pi	4.154*** (3.52)	3.705*** (3.16)	3.608** (2.24)	3.518*** (2.99)	2.987** (2.55)	3.054* (1.89)
NGS	5.498*** (5.14)	5.094*** (4.61)	3.261* (1.65)	3.931*** (3.59)	3.118*** (2.71)	0.492 (0.23)
ESUB_lat	-416.388*** (-4.61)	-419.549*** (-4.65)	-253.073* (-1.89)	-409.987*** (-4.54)	-409.808*** (-4.54)	-254.408* (-1.90)
AbAccr2	13.763*** (4.79)	13.750*** (4.71)	14.794*** (4.48)	12.107*** (4.19)	12.022*** (4.10)	13.762*** (4.16)
StdMRn5	19.781*** (4.10)	18.164*** (3.64)	11.549 (1.24)	24.112*** (4.97)	22.737*** (4.53)	13.903 (1.44)
N	44892	44892	43311	44892	44892	43311
Adj R ²	0.030	0.032	0.061	0.033	0.034	0.063
Industry FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Panel C: DDBTD

	(1)	(2)	(3)	(4)	(5)	(6)
	DD BTD	DD BTB	DD BTB	DD BTB	DD BTB	DD BTB
EMP	-0.005*** (-9.61)	-0.004*** (-7.36)	-0.004*** (-3.45)	-0.004*** (-9.26)	-0.003*** (-6.80)	-0.003*** (-2.65)
ROA_iblat	0.037** (2.06)	0.036** (1.96)	0.084*** (5.39)	0.038** (2.11)	0.034* (1.89)	0.088*** (5.63)
DNOL	-0.061*** (-3.91)	-0.062*** (-3.99)	-0.035*** (-5.00)	-0.060*** (-3.90)	-0.061*** (-3.98)	-0.037*** (-5.21)
DSale	-0.004 (-1.17)	-0.005 (-1.43)	-0.002 (-0.67)	-0.003 (-0.84)	-0.004 (-1.14)	-0.001 (-0.56)
CAPX_at	0.063*** (2.88)	0.033 (1.48)	-0.005 (-0.30)	0.101*** (4.61)	0.076*** (3.41)	0.021 (1.17)
PPEGT_at	0.025*** (8.92)	0.020*** (6.45)	0.017*** (3.54)	0.024*** (8.80)	0.019*** (6.23)	0.015*** (3.26)
Invt_lat	-0.065*** (-10.93)	-0.051*** (-6.42)	-0.071*** (-6.12)	-0.056*** (-9.46)	-0.038*** (-4.84)	-0.063*** (-5.42)
DP_at	-0.135*** (-3.12)	-0.146*** (-3.19)	-0.110** (-2.40)	-0.116*** (-2.70)	-0.124*** (-2.71)	-0.100** (-2.16)
RND_lat	0.065** (2.55)	0.044 (1.57)	-0.022 (-0.61)	0.092*** (3.60)	0.066** (2.40)	-0.017 (-0.48)
Intan_sale	-0.002 (-0.11)	-0.015 (-0.76)	0.046 (1.28)	-0.018 (-1.01)	-0.030 (-1.57)	0.044 (1.21)
CHE_at	-0.012** (-1.97)	-0.008 (-1.28)	-0.008 (-1.10)	-0.016*** (-2.58)	-0.011* (-1.73)	-0.010 (-1.48)
AltmanZ	-0.001*** (-5.40)	-0.001*** (-5.14)	-0.001*** (-3.11)	-0.001*** (-5.24)	-0.001*** (-4.92)	-0.001*** (-2.92)
PIFO_pi	-0.038*** (-24.42)	-0.038*** (-23.61)	-0.023*** (-15.97)	-0.040*** (-25.71)	-0.040*** (-25.08)	-0.023*** (-16.10)
Num_Analyst	-0.004*** (-5.53)	-0.005*** (-6.91)	-0.002* (-1.75)	-0.005*** (-6.29)	-0.006*** (-7.52)	-0.002* (-1.75)
Age	0.006*** (6.76)	0.006*** (6.39)	0.011*** (6.60)	0.004*** (4.96)	0.004*** (4.38)	0.011*** (4.43)
SGA_at	-0.027*** (-5.51)	-0.022*** (-4.02)	0.009 (0.84)	-0.025*** (-5.12)	-0.018*** (-3.27)	0.008 (0.80)
SPI_at	0.394*** (8.08)	0.383*** (7.84)	0.359*** (9.86)	0.397*** (8.21)	0.388*** (8.03)	0.356*** (9.81)
N	17329	17328	16687	17329	17328	16687
Adj R ²	0.140	0.159	0.708	0.154	0.173	0.713
Industry FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Panel D: MPBTD

	(1)	(2)	(3)	(4)	(5)	(6)
	MP BTD	MP BTD	MP BTD	MP BTD	MP BTD	MP BTD
ROA_iblat	-0.001 (-0.05)	-0.010 (-0.43)	0.081*** (3.45)	0.006 (0.27)	0.001 (0.03)	0.078*** (3.15)
OCFf_lat	0.091*** (4.61)	0.101*** (4.69)	0.051*** (2.60)	0.079*** (3.91)	0.086*** (3.80)	0.047** (2.22)
DNOL	-0.050*** (-5.52)	-0.050*** (-5.52)	-0.034*** (-4.33)	-0.050*** (-5.51)	-0.050*** (-5.51)	-0.035*** (-4.46)
MTB	-0.000*** (-3.01)	-0.000** (-2.51)	-0.000*** (-2.76)	-0.001*** (-4.67)	-0.001*** (-4.24)	-0.001*** (-3.46)
CAPX_at	0.087*** (3.02)	0.089*** (2.91)	0.016 (0.55)	0.119*** (4.05)	0.121*** (3.83)	0.067** (2.20)
PPEGT_at	0.018*** (9.48)	0.014*** (6.48)	0.017*** (3.59)	0.018*** (9.44)	0.016*** (6.99)	0.021*** (4.44)
Invt_lat	-0.036*** (-9.15)	-0.029*** (-6.03)	-0.065*** (-6.98)	-0.025*** (-6.32)	-0.013*** (-2.85)	-0.041*** (-4.39)
DP_at	-0.169*** (-5.46)	-0.150*** (-4.56)	-0.110** (-2.35)	-0.155*** (-5.05)	-0.146*** (-4.51)	-0.103** (-2.17)
RND_lat	0.038** (2.14)	0.020 (1.04)	-0.072** (-2.25)	0.063*** (3.61)	0.041** (2.20)	-0.063** (-2.00)
Intan_sale	0.024** (2.06)	0.028** (2.33)	0.049* (1.85)	0.018 (1.55)	0.023* (1.91)	0.047* (1.80)
INT_lat	0.252*** (7.35)	0.232*** (6.54)	0.080* (1.71)	0.390*** (11.24)	0.374*** (10.48)	0.235*** (4.50)
CHE_at	-0.002 (-0.45)	-0.001 (-0.27)	-0.004 (-0.67)	-0.006 (-1.33)	-0.005 (-1.07)	-0.006 (-0.92)
AltmanZ	-0.001*** (-6.69)	-0.001*** (-6.96)	-0.001*** (-3.89)	-0.001*** (-5.49)	-0.001*** (-5.68)	-0.001*** (-2.86)
PIFO_pi	-0.021*** (-23.57)	-0.021*** (-23.35)	-0.018*** (-15.74)	-0.023*** (-26.44)	-0.024*** (-26.22)	-0.020*** (-17.17)
AbAccr1	0.139*** (8.34)	0.146*** (7.49)	0.093*** (5.10)	0.127*** (7.27)	0.127*** (6.11)	0.084*** (4.17)
SPI_at	0.404*** (11.58)	0.396*** (11.32)	0.340*** (10.20)	0.411*** (11.94)	0.406*** (11.79)	0.352*** (10.59)
StdCF5	0.097*** (5.95)	0.102*** (6.23)	0.086*** (4.54)	0.101*** (6.15)	0.106*** (6.42)	0.087*** (4.57)
N	23404	23403	22689	23404	23403	22689
Adj R ²	0.152	0.162	0.414	0.180	0.190	0.427
Industry FE		Yes			Yes	
Firm FE			Yes			Yes
Year FE				Yes	Yes	Yes

Table 4.11 Tax-related ICW and tax avoidance

This table presents the replication results of the main analysis in Bauer (2016). Panel A reports the replication results using sample from 2004–2009. Panel B reports the replication results using extended sample period from 2004–2019. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the *t*-statistic is reported in the parentheses.

	Panel A: Sample from original study (2004–2009)		Panel B: Extended sample (2004–2019)	
	Replication of the original study	Replication with key variables	Replication of the original study	Replication with key variables
	(1) CashETR1	(2) CashETR1	(3) CashETR1	(4) CashETR1
TAX_ICW	0.031** (2.13)	0.040* (1.75)	0.033*** (3.11)	0.033** (2.26)
Other_ICW	0.007 (0.84)	0.021* (1.66)	0.007 (1.24)	0.014 (1.56)
Controls	Controls in the original study	Key Vars for CashETR1	Controls in the original study	Key Vars for CashETR1
N	14240	6491	39534	18380
Adj R ²	0.119	0.123	0.111	0.127
Industry FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Table 4.12 Corporate Social Responsibility (CSR) and tax avoidance

This table presents the replication results of the main analysis in Davis et al. (2016). Panel A reports the replication results using sample from 2002–2011. Panel B reports the replication results using extended sample from 1995–2013. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the p-value is reported in the parentheses.

	Panel A: Sample from original study (2002–2011)		Panel B: Extended sample (1995–2013)	
	Replication of the original study	Replication with key variables	Replication of the original study	Replication with key variables
	(1) CashETR5	(2) CashETR5	(3) CashETR5	(4) CashETR5
CSR	-0.003*	-0.002	-0.003**	-0.002
	(0.061)	(0.111)	(0.047)	(0.140)
Controls	Controls in the original study	Key Vars for CashETR3	Controls in the original study	Key Vars for CashETR3
N	5679	5431	8066	7750
Adj R ²	0.163	0.164	0.165	0.176
Industry FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Table 4.13 Price-cost margin (PCM) and tax avoidance

This table presents the replication results of the main analysis in Kubick et al. (2015). Panel A reports the replication results using sample from 1993–2013. Panel B reports the replication results using extended sample from 1990–2019. Panel C presents the replication results using the aggregated controls composed of key variables for GAAP ETR1, Cash ETR1, GAAP ETR3 and Cash ETR3 models. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, the p-value is reported in the parentheses.

Panel A: Original sample period (1993–2013)

	Panel A1: Replication of original results				Panel A2: Replication with identified key variables				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	GAAPETR1	CashETR1	GAAPETR5	CashETR5	GAAPETR1	GAAPETR1	CashETR1	GAAPETR5	CashETR5
PCM	-0.036***	-0.044***	-0.013	-0.047***	-0.009	0.006	-0.012	0.009	-0.029**
	(0.000)	(0.000)	(0.129)	(0.000)	(0.257)	(0.432)	(0.295)	(0.370)	(0.019)
Controls	Controls in the original study				Key Vars for GAAP ETR1		Key Vars for Cash ETR1	Key Vars for GAAP ETR3	Key Vars for GAAP ETR3
N	24352	24352	24352	24352	24142	21752	20342	22328	14502
Adj R ²	0.074	0.102	0.076	0.119	0.157	0.218	0.113	0.077	0.173
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Extended sample period (1990–2019)

Panel B1: Replication of original results					Panel B2: Replication with identified key variables				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	GAAPETR1	CashETR1	GAAPETR5	CashETR5	GAAPETR1	GAAPETR1	CashETR1	GAAPETR5	CashETR5
PCM	-0.041***	-0.057***	-0.018**	-0.047***	-0.008*	0.000	-0.006	0.005	-0.024**
	(0.000)	(0.000)	(0.011)	(0.000)	(0.077)	(0.940)	(0.510)	(0.538)	(0.018)
Controls	Controls in the original study				Key Vars for GAAP ETR1		Key Vars for Cash ETR1	Key Vars for GAAP ETR3	Key Vars for GAAP ETR3
N	56329	51655	47125	37511	92898	61734	37247	58864	31045
Adj R ²	0.077	0.107	0.106	0.125	0.114	0.178	0.132	0.095	0.151
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel C: Replication using an aggregated key-variable control set

	Panel C1: Sample period from original study (1993–2010)				Panel C2: Extended sample period (1990–2019)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	GAAPETR1	CashETR1	GAAPETR5	CashETR5	GAAPETR1	CashETR1	GAAPETR5	CashETR5
PCM	-0.000	-0.012	0.017*	-0.010	0.004	-0.004	0.018**	-0.008
	(0.975)	(0.272)	(0.082)	(0.387)	(0.586)	(0.666)	(0.040)	(0.433)
Controls	Aggregated key-variables controls				Aggregated key-variables controls			
N	20342	20342	20342	20342	38059	37247	34545	29842
Adj R ²	0.093	0.145	0.108	0.157	0.116	0.169	0.147	0.171
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 4.14 Comparison between theoretical predictions and empirical outcomes

This table presents the comparison between theoretical predictions aggregated by Bruhne and Jacob (2019) and performance of the selected robust variables. Panel A summarizes the association between tax avoidance and different determinants predicted by theories. Panel B reports the aggregated variable performances detailed in Tables 9 and 10. The symbols \uparrow , \downarrow and \uparrow/\downarrow denote a theoretically predicted positive, negative and mixed effect, respectively, of a variable on the benefits or costs from tax avoidance engagement. The symbols +, – and ? stand for an overall prediction or overall outcome of positive, negative and mixed association, respectively, between a variable and tax avoidance. NA means the variable is not selected in the robust baseline model.

	Panel A: Theoretical prediction			Panel B: Aggregated fixed-effect test results of identified key variables							
	Benefits	Costs	Prediction	Cash ETR1	GAAP ETR1	Cash ETR3	GAAP ETR3	UTB	DTAX	DDBTD	MPBTD
Size	\uparrow	\uparrow/\downarrow	?	?	NA	?	– (?)	NA	NA	NA	NA
Market power	\uparrow/\downarrow		?	?	NA	–	NA	NA	NA	NA	NA
Profitability	\uparrow	\uparrow	?	+	?	?	?	?	?	+	+
Growth	\uparrow	\downarrow	+	+	–	+	– (?)	– (?)	–	+	?
Life cycle	\uparrow/\downarrow	\uparrow/\downarrow	?	NA	+	NA	– (?)	NA	NA	+	NA
Tangible assets	\uparrow	\uparrow	?	–	–	–	NA	– (?)	–	?	?
Intangible assets	\uparrow/\downarrow	\downarrow	?	NA	+	?	?	?	?	+	?
Leverage	\downarrow		–	+	+	+	NA	–	NA	NA	+
Financial constraints	\uparrow	(\uparrow)	+	+	+	?	+	NA	+	+	+
Foreign operation		\downarrow	+	–	+	+	+	+	+	–	–
Complexity		\uparrow/\downarrow	?	NA	NA	NA	NA	+	+	NA	NA
Inst. Ownership	\uparrow	\uparrow	?	NA	NA	+	NA	NA	NA	NA	NA

Chapter 5 CONCLUSION

Scholars in financial economics research have expressed concerns regarding the tendency to produce ‘statistically significant’ results (Dyckman and Zeff, 2014; Gow et al., 2016; Harvey, 2017), which has led to an increasing number of incremental variables. A combination of inadequate systematic review and the limitations of conventional tools has meant that many well-explored research fields, such as audit fees and tax avoidance, have accumulated a vast number of under-investigated variables, which confuse subsequent research and provide greater opportunities for p-hacking. In response to these challenges, this thesis has not only provided an extensive investigation into the research ‘gold rush’, but has also employed two powerful machine learning techniques, LASSO and Random Forest, to reassess a large number of right-hand-side variables in audit fee and tax avoidance research.

Drawing on a large US sample of relevant research, this thesis has provided initial evidence of machine learning applications in audit fee and tax avoidance studies. In the interest of subsequent research, the machine learning analyses in this thesis not only selected numerous strong variables as robust baseline models, but also provided an example for applying more advanced techniques to tackle problems that are beyond the capability of conventional tools. Owing to its ability to handle larger sample sizes, novel data sources, and more incremental variables, machine learning is a powerful mechanism that will help advance our knowledge and understanding of financial economics research. Given the prevalence of the research ‘gold rush’, future studies will be able to apply the machine learning analyses used in this thesis in other under-investigated fields, such as corporate diversification and acquisition.

In addition to its contributions, this thesis has a number of limitations. One involves data

access. Although it has identified more than 300 variables from prior works, many of those variables require access to private data sources or time-consuming hand collection, and thus the analyses only cover variables that have been constructed based on public data. With regard to popularity, variables that use public data are still the most widely cited in empirical studies. Another limitation is the smaller coverage of the corporate governance sample because of limited data availability. Therefore, the analysis using corporate governance attributes has been separated from the main analysis. Finally, the review of empirical studies is based on articles in top-tier journals in accounting and finance, which means that second-tier journals or journals in the fields of business and economics have not been investigated. It would therefore be of interest to examine related empirical studies in those additional journals in future research, especially to replicate and cross-check the identified incremental variables with the key variables selected in Chapters 3 and 4.

Reference

- Abbott, L.J., Gunny, K. and Pollard, T., 2017. The impact of litigation risk on auditor pricing behavior: Evidence from reverse mergers. *Contemporary Accounting Research*, 34(2), pp.1103-1127.
- Abbott, L.J., Parker, S. and Peters, G.F., 2012. Audit fee reductions from internal audit-provided assistance: The incremental impact of internal audit characteristics. *Contemporary Accounting Research*, 29(1).
- Acemoglu, D. and Zilibotti, F., 2001. Productivity differences. *The Quarterly Journal of Economics*, 116(2), pp.563-606.
- Acito, A.A., Hogan, C.E. and Mergenthaler, R.D., 2018. The effects of PCAOB inspections on auditor-client relationships. *The Accounting Review*, 93(2), pp.1-35.
- Aichian, A.A. and Kessel, R.A., 1962. Competition, monopoly, and the pursuit of pecuniary gain. In *Aspects of labor economics* (pp. 157-183). Princeton University Press.
- Amini, S., Elmore, R., Öztekin, Ö. and Strauss, J., 2020. Can Machines Learn Capital Structure Dynamics?. *Available at SSRN 3473322*.
- Amir, E., Kallunki, J.P. and Nilsson, H., 2014. The association between individual audit partners' risk preferences and the composition of their client portfolios. *Review of Accounting Studies*, 19(1), pp.103-133.
- Anand, V., Brunner, R., Ikegwu, K. and Sougiannis, T., 2019. Predicting Profitability Using Machine Learning. *Available at SSRN 3466478*.
- Anesa, M., Gillespie, N., Spee, A.P. and Sadiq, K., 2019. The legitimization of corporate tax minimization. *Accounting, Organizations and Society*, 75, pp.17-39.
- Angrist, J.D. and Pischke, J.S., 2009. Instrumental variables in action: sometimes you get what you need. *Mostly harmless econometrics: an empiricist's companion*, pp.113-220.
- Arditti, F.D. and Pinkerton, J.M., 1978. The valuation and cost of capital of the levered firm with growth opportunities. *The journal of finance*, 33(1), pp.65-73.
- Armstrong, C.S., Blouin, J.L., Jagolinzer, A.D. and Larcker, D.F., 2015. Corporate governance, incentives, and tax avoidance. *Journal of Accounting and Economics*, 60(1), pp.1-17.

- Armstrong, C.S., Blouin, J.L. and Larcker, D.F., 2012. The incentives for tax planning. *Journal of accounting and economics*, 53(1-2), pp.391-411.
- Armstrong, C.S., Glaeser, S. and Kepler, J.D., 2019. Strategic reactions in corporate tax planning. *Journal of Accounting and Economics*, 68(1), p.101232.
- Athey, S. and Imbens, G.W., 2017. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), pp.3-32.
- Atwood, T.J., Drake, M.S., Myers, J.N. and Myers, L.A., 2012. Home country tax system characteristics and corporate tax avoidance: International evidence. *The Accounting Review*, 87(6), pp.1831-1860.
- Atwood, T.J. and Lewellen, C., 2019. The complementarity between tax avoidance and manager diversion: Evidence from tax haven firms. *Contemporary Accounting Research*, 36(1), pp.259-294.
- Ayers, B.C., Call, A.C. and Schwab, C.M., 2018. Do Analysts' Cash Flow Forecasts Encourage Managers to Improve the Firm's Cash Flows? Evidence from Tax Planning. *Contemporary Accounting Research*, 35(2), pp.767-793.
- Badertscher, B.A., Katz, S.P. and Rego, S.O., 2013. The separation of ownership and control and corporate tax avoidance. *Journal of accounting and economics*, 56(2-3), pp.228-250.
- Badertscher, B.A., Katz, S.P., Rego, S.O. and Wilson, R.J., 2019. Conforming tax avoidance and capital market pressure. *The Accounting Review*, 94(6), pp.1-30.
- Badertscher, B., Jorgensen, B., Katz, S. and Kinney, W., 2014. Public equity and audit pricing in the United States. *Journal of Accounting Research*, 52(2), pp.303-339.
- Bae, G.S., Choi, S.U. and Rho, J.H., 2016. Audit hours and unit audit price of industry specialist auditors: Evidence from Korea. *Contemporary Accounting Research*, 33(1), pp.314-340.
- Balakrishnan, K., Blouin, J.L. and Guay, W.R., 2019. Tax aggressiveness and corporate transparency. *The Accounting Review*, 94(1), pp.45-69.
- Ball, R., Jayaraman, S. and Shivakumar, L., 2012. Audited financial reporting and voluntary disclosure as complements: A test of the confirmation hypothesis. *Journal of accounting and economics*, 53(1-2), pp.136-166.
- Barua, A., Lennox, C. and Raghunandan, A., 2019. Are audit fees discounted in initial year audit

- engagements?. *Journal of Accounting and Economics*, p.101282.
- Bao, Y., Ke, B., Li, B., Yu, Y.J. and Zhang, J., 2020. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research*, 58(1), pp.199-235.
- Bauer, A.M., 2016. Tax avoidance and the implications of weak internal controls. *Contemporary Accounting Research*, 33(2), pp.449-486.
- Beck, M.J., Francis, J.R. and Gunn, J.L., 2018. Public company audits and city-specific labor characteristics. *Contemporary Accounting Research*, 35(1), pp.394-433.
- Beck, M.J. and Mauldin, E.G., 2014. Who's really in charge? Audit committee versus CFO power and audit fees. *The Accounting Review*, 89(6), pp.2057-2085.
- Bédard, J. and Courteau, L., 2015. Benefits and costs of auditor's assurance: Evidence from the review of quarterly financial statements. *Contemporary Accounting Research*, 32(1), pp.308-335.
- Bell, T.B., Causholli, M. and Knechel, W.R., 2015. Audit firm tenure, non-audit services, and internal assessments of audit quality. *Journal of Accounting Research*, 53(3), pp.461-509.
- Bell, T.B., Doogar, R. and Solomon, I., 2008. Audit labor usage and fees under business risk auditing. *Journal of accounting research*, 46(4), pp.729-760.
- Berglund, N. and Kang, T., 2013. Does Social Trust Matter in Financial Reporting?: Evidence from Audit Pricing. *Journal of Accounting Research*, 12, pp.119-121.
- Bertomeu, J., Beyer, A. and Taylor, D.J., 2016. From casual to causal inference in accounting research: The need for theoretical foundations. *Foundations and Trends in Accounting*, Forthcoming, pp.15-63.
- Bertomeu, J., 2020. Machine learning improves accounting: discussion, implementation and research opportunities. *Review of Accounting Studies*, 25(3), pp.1135-1155.
- Bertomeu, J., Cheynel, E., Floyd, E. and Pan, W., 2020. Using machine learning to detect misstatements. *Review of Accounting Studies*, pp.1-52.
- Bethmann, I., Jacob, M. and Müller, M.A., 2018. Tax loss carrybacks: Investment stimulus versus misallocation. *The Accounting Review*, 93(4), pp.101-125.
- Bhaskar, L.S., Krishnan, G.V. and Yu, W., 2017. Debt covenant violations, firm financial distress,

- and auditor actions. *Contemporary accounting research*, 34(1), pp.186-215.
- Bills, K.L., Cunningham, L.M. and Myers, L.A., 2016. Small audit firm membership in associations, networks, and alliances: Implications for audit quality and audit fees. *The Accounting Review*, 91(3), pp.767-792.
- Bills, K.L., Jeter, D.C. and Stein, S.E., 2015. Auditor industry specialization and evidence of cost efficiencies in homogenous industries. *The Accounting Review*, 90(5), pp.1721-1754.
- Bills, K.L., Lisic, L.L. and Seidel, T.A., 2017. Do CEO succession and succession planning affect stakeholders' perceptions of financial reporting risk? Evidence from audit fees. *The Accounting Review*, 92(4), pp.27-52.
- Bird, A. and Karolyi, S.A., 2017. Governance and Taxes: Evidence from Regression Discontinuity (Retracted). *The Accounting Review*, 92(1), pp.29-50.
- Blay, A.D. and Geiger, M.A., 2013. Auditor fees and auditor independence: Evidence from going concern reporting decisions. *Contemporary Accounting Research*, 30(2), pp.579-606.
- Blaylock, B.S., 2016. Is tax avoidance associated with economically significant rent extraction among US firms?. *Contemporary Accounting Research*, 33(3), pp.1013-1043.
- Bloomfield, R., Nelson, M.W. and Soltes, E., 2016. Gathering data for archival, field, survey, and experimental accounting research. *Journal of Accounting Research*, 54(2), pp.341-395.
- Boadway, R. and Bruce, N., 1984. A general proposition on the design of a neutral business tax. *Journal of Public Economics*, 24(2), pp.231-239.
- Bonsall IV, S.B., Koharki, K. and Watson, L., 2017. Deciphering tax avoidance: Evidence from credit rating disagreements. *Contemporary Accounting Research*, 34(2), pp.818-848.
- Boone, J.P., Khurana, I.K. and Raman, K.K., 2015. Did the 2007 PCAOB disciplinary order against Deloitte impose actual costs on the firm or improve its audit quality?. *The Accounting Review*, 90(2), pp.405-441.
- Borenstein, M., Hedges, L.V., Higgins, J.P. and Rothstein, H.R., 2021. *Introduction to meta-analysis*. John Wiley & Sons.
- Brown, J.L., 2011. The spread of aggressive corporate tax reporting: A detailed examination of the corporate-owned life insurance shelter. *The Accounting Review*, 86(1), pp.23-57.

- Brown, J.L. and Drake, K.D., 2014. Network ties among low-tax firms. *The Accounting Review*, 89(2), pp.483-510.
- Brown, N.C., Crowley, R.M. and Elliott, W.B., 2020. What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, 58(1), pp.237-291.
- Bronson, S.N., Ghosh, A. and Hogan, C.E., 2017. Audit fee differential, audit effort, and litigation risk: An examination of ADR firms. *Contemporary Accounting Research*, 34(1), pp.83-117.
- Bruehne, A. and Jacob, M., 2019. Corporate tax avoidance and the real effects of taxation: A review. *Available at SSRN 3495496*.
- Cahan, S.F., Godfrey, J.M., Hamilton, J. and Jeter, D.C., 2008. Auditor specialization, auditor dominance, and audit fees: The role of investment opportunities. *The Accounting Review*, 83(6), pp.1393-1423.
- Carcello, J.V. and Li, C., 2013. Costs and benefits of requiring an engagement partner signature: Recent experience in the United Kingdom. *The Accounting Review*, 88(5), pp.1511-1546.
- Carhart, M.M., 1997. On persistence in mutual fund performance. *The Journal of finance*, 52(1), pp.57-82.
- Carson, E., 2009. Industry specialization by global audit firm networks. *The Accounting Review*, 84(2), pp.355-382.
- Cen, L., Maydew, E.L., Zhang, L. and Zuo, L., 2017. Customer–supplier relationships and corporate tax avoidance. *Journal of Financial Economics*, 123(2), pp.377-394.
- Chan, D.K., 1999. “Low-balling” and efficiency in a two-period specialization model of auditing competition. *Contemporary Accounting Research*, 16(4), pp.609-642.
- Chan, K.H., Lin, K.Z. and Mo, P.L., 2010. Will a departure from tax-based accounting encourage tax noncompliance? Archival evidence from a transition economy. *Journal of Accounting and Economics*, 50(1), pp.58-73.
- Chan, L.H., Chen, K.C., Chen, T.Y. and Yu, Y., 2012. The effects of firm-initiated clawback provisions on earnings quality and auditor behavior. *Journal of Accounting and Economics*, 54(2-3), pp.180-196.
- Chen, C., Liaw, A. and Breiman, L., 2004. Using random forest to learn imbalanced

- data. *University of California, Berkeley*, 110(1-12), p.24.
- Chen, N.X., Chiu, P.C. and Shevlin, T., 2018. Do analysts matter for corporate tax planning? Evidence from a natural experiment. *Contemporary Accounting Research*, 35(2), pp.794-829.
- Chen, S., Chen, X., Cheng, Q. and Shevlin, T., 2010. Are family firms more tax aggressive than non-family firms?. *Journal of financial economics*, 95(1), pp.41-61.
- Chen, T. and Lin, C., 2017. Does information asymmetry affect corporate tax aggressiveness?. *Journal of Financial and Quantitative Analysis*, 52(5), pp.2053-2081.
- Chen, Y., Ge, R., Louis, H. and Zolotoy, L., 2019. Stock liquidity and corporate tax avoidance. *Review of Accounting Studies*, 24(1), pp.309-340.
- Chen, Y., Gul, F.A., Veeraraghavan, M. and Zolotoy, L., 2015. Executive equity risk-taking incentives and audit pricing. *The Accounting Review*, 90(6), pp.2205-2234.
- Cheng, C.A., Huang, H.H., Li, Y. and Stanfield, J., 2012. The effect of hedge fund activism on corporate tax avoidance. *The Accounting Review*, 87(5), pp.1493-1526.
- Chi, S., Huang, S.X. and Sanchez, J.M., 2017. CEO inside debt incentives and corporate tax sheltering. *Journal of Accounting Research*, 55(4), pp.837-876.
- Choi, J.H., Kim, J.B., Liu, X. and Simunic, D.A., 2008. Audit pricing, legal liability regimes, and Big 4 premiums: Theory and cross-country evidence. *Contemporary Accounting Research*, 25(1), pp.55-99.
- Choi, J.H., Kim, J.B., Liu, X. and Simunic, D.A., 2009. Cross-listing audit fee premiums: Theory and evidence. *The Accounting Review*, 84(5), pp.1429-1463.
- Chu, L., Simunic, D.A., Ye, M. and Zhang, P., 2018. Transaction costs and competition among audit firms in local markets. *Journal of Accounting and Economics*, 65(1), pp.129-147.
- Chung, S.G., Goh, B.W., Lee, J. and Shevlin, T., 2019. Corporate tax aggressiveness and insider trading. *Contemporary Accounting Research*, 36(1), pp.230-258.
- Chyz, J.A., 2013. Personally tax aggressive executives and corporate tax sheltering. *Journal of Accounting and Economics*, 56(2-3), pp.311-328.
- Chyz, J.A., Gaertner, F.B., Kausar, A. and Watson, L., 2019. Overconfidence and corporate tax policy. *Review of Accounting Studies*, 24(3), pp.1114-1145.

- Chyz, J.A., Leung, W.S.C., Li, O.Z. and Rui, O.M., 2013. Labor unions and tax aggressiveness. *Journal of Financial Economics*, 108(3), pp.675-698.
- Cléménçon, S., Depecker, M. and Vayatis, N., 2013. Ranking forests. *Journal of Machine Learning Research*, 14(Jan), pp.39-73.
- Cohen, J.R., Hoitash, U., Krishnamoorthy, G. and Wright, A.M., 2014. The effect of audit committee industry expertise on monitoring the financial reporting process. *The Accounting Review*, 89(1), pp.243-273.
- Core, J.E., Holthausen, R.W. and Larcker, D.F., 1999. Corporate governance, chief executive officer compensation, and firm performance. *Journal of financial economics*, 51(3), pp.371-406.
- Craswell, A.T., Francis, J.R. and Taylor, S.L., 1995. Auditor brand name reputations and industry specializations. *Journal of accounting and economics*, 20(3), pp.297-322.
- Dao, M., Raghunandan, K. and Rama, D.V., 2012. Shareholder voting on auditor selection, audit fees, and audit quality. *The Accounting Review*, 87(1), pp.149-171.
- Davis, A.K., Guenther, D.A., Krull, L.K. and Williams, B.M., 2016. Do socially responsible firms pay more taxes?. *The accounting review*, 91(1), pp.47-68.
- DeAngelo, H. and Masulis, R.W., 1980. Leverage and dividend irrelevancy under corporate and personal taxation. *The Journal of Finance*, 35(2), pp.453-464.
- DeBacker, J., Heim, B.T. and Tran, A., 2015. Importing corruption culture from overseas: Evidence from corporate tax evasion in the United States. *Journal of Financial Economics*, 117(1), pp.122-138.
- Dee, C.C., Lulseged, A. and Zhang, T., 2015. Who did the audit? Audit quality and disclosures of other audit participants in PCAOB filings. *The Accounting Review*, 90(5), pp.1939-1967.
- Dechow, P.M., Sloan, R.G. and Sweeney, A.P., 1995. Detecting earnings management. *Accounting review*, pp.193-225.
- Defond, M.L., Francis, J.R. and Hallman, N.J., 2018. Awareness of SEC enforcement and auditor reporting decisions. *Contemporary Accounting Research*, 35(1), pp.277-313.
- DeFond, M.L. and Lennox, C.S., 2017. Do PCAOB inspections improve the quality of internal control audits?. *Journal of Accounting Research*, 55(3), pp.591-627.

- De George, E.T., Ferguson, C.B. and Spear, N.A., 2013. How much does IFRS cost? IFRS adoption and audit fees. *The Accounting Review*, 88(2), pp.429-462.
- Demirkan, S. and Zhou, N., 2016. Audit pricing for strategic alliances: An incomplete contract perspective. *Contemporary Accounting Research*, 33(4), pp.1625-1647.
- Deng, M., Lu, T., Simunic, D.A. and Ye, M., 2014. Do joint audits improve or impair audit quality?. *Journal of Accounting research*, 52(5), pp.1029-1060.
- Denis, D.K., 2012. Mandatory clawback provisions, information disclosure, and the regulation of securities markets. *Journal of Accounting and Economics*, 54(2-3), pp.197-200.
- Desai, M.A., 2003. The divergence between book income and tax income. *Tax policy and the economy*, 17, pp.169-206.
- Desai, M.A. and Dharmapala, D., 2006. Corporate tax avoidance and high-powered incentives. *Journal of financial Economics*, 79(1), pp.145-179.
- De Simone, L., Mills, L.F. and Stomberg, B., 2019. Using IRS data to identify income shifting to foreign affiliates. *Review of Accounting Studies*, 24(2), pp.694-730.
- Dharmapala, D., 2014. What do we know about base erosion and profit shifting? A review of the empirical literature. *Fiscal Studies*, 35(4), pp.421-448.
- Dharmapala, D. and Riedel, N., 2013. Earnings shocks and tax-motivated income-shifting: Evidence from European multinationals. *Journal of Public Economics*, 97, pp.95-107.
- Donohoe, M.P., 2015. The economic effects of financial derivatives on corporate tax avoidance. *Journal of Accounting and Economics*, 59(1), pp.1-24.
- Donohoe, M.P. and Robert Knechel, W., 2014. Does corporate tax aggressiveness influence audit pricing?. *Contemporary Accounting Research*, 31(1), pp.284-308.
- Doogar, R., Sivadasan, P. and Solomon, I., 2010. The regulation of public company auditing: Evidence from the transition to AS5. *Journal of Accounting Research*, 48(4), pp.795-814.
- Dyckman, T.R. and Zeff, S.A., 2014. Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons*, 28(3), pp.695-712.
- Dyckman, T.R. and Zeff, S.A., 2015. Accounting research: past, present, and future. *Abacus*, 51(4), pp.511-524.
- Dyckman, T.R., 2016. Significance testing: We can do better. *Abacus*, 52(2), pp.319-342.

- Dyckman, T.R. and Zeff, S.A., 2019. Important issues in statistical testing and recommended improvements in accounting research. *Econometrics*, 7(2), p.18.
- Dyreng, S.D., Hanlon, M. and Maydew, E.L., 2008. Long-run corporate tax avoidance. *the accounting review*, 83(1), pp.61-82.
- Dyreng, S.D., Hanlon, M. and Maydew, E.L., 2010. The effects of executives on corporate tax avoidance. *The accounting review*, 85(4), pp.1163-1189.
- Dyreng, S.D., Hoopes, J.L. and Wilde, J.H., 2016. Public pressure and corporate tax behavior. *Journal of Accounting Research*, 54(1), pp.147-186.
- Dyreng, S., Jacob, M., Jiang, X. and Müller, M.A., 2019. Tax incidence and tax avoidance. *Available at SSRN 3070239*.
- Dyreng, S.D., Lindsey, B.P. and Thornock, J.R., 2013. Exploring the role Delaware plays as a domestic tax haven. *Journal of Financial Economics*, 108(3), pp.751-772.
- Dyreng, S.D., Lindsey, B.P., Markle, K.S. and Shackelford, D.A., 2015. The effect of tax and nontax country characteristics on the global equity supply chains of US multinationals. *Journal of Accounting and Economics*, 59(2-3), pp.182-202.
- Dyreng, S.D. and Markle, K.S., 2016. The effect of financial constraints on income shifting by US multinationals. *The Accounting Review*, 91(6), pp.1601-1627.
- Edwards, A., Schwab, C. and Shevlin, T., 2016. Financial constraints and cash tax savings. *The Accounting Review*, 91(3), pp.859-881.
- Engel, E., Hayes, R.M. and Wang, X., 2010. Audit committee compensation and the demand for monitoring of the financial reporting process. *Journal of Accounting and Economics*, 49(1-2), pp.136-154.
- Fama, E.F. and French, K.R., 2015. A five-factor asset pricing model. *Journal of financial economics*, 116(1), pp.1-22.
- Fama, E.F. and French, K.R., 2018. Choosing factors. *Journal of financial economics*, 128(2), pp.234-252.
- Fama, E.F. and French, K.R., 2021. *The cross-section of expected stock returns* (pp. 349-391). University of Chicago Press.
- Fanelli, D., 2010. "Positive" results increase down the hierarchy of the sciences. *PloS one*, 5(4),

p.e10068.

- Fanelli, D., 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), pp.891-904.
- Fang, J., Pittman, J., Zhang, Y. and Zhao, Y., 2017. Auditor choice and its implications for group-affiliated firms. *Contemporary Accounting Research*, 34(1), pp.39-82.
- Feng, G., Giglio, S. and Xiu, D., 2020. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3), pp.1327-1370.
- Feller, A. and Schanz, D., 2017. The three hurdles of tax planning: How business context, aims of tax planning, and tax manager power affect tax expense. *Contemporary Accounting Research*, 34(1), pp.494-524.
- Finley, A.R., 2019. The impact of large tax settlement favorability on firms' subsequent tax avoidance. *Review of Accounting Studies*, 24(1), pp.156-187.
- Firth, M., 1997. The provision of nonaudit services by accounting firms to their audit clients. *Contemporary Accounting Research*, 14(2), pp.1-21.
- Foley, C.F., Hartzell, J.C., Titman, S. and Twite, G., 2007. Why do firms hold so much cash? A tax-based explanation. *Journal of financial economics*, 86(3), pp.579-607.
- Fox, W.F., Luna, L. and Schaur, G., 2014. Destination taxation and evasion: Evidence from US inter-state commodity flows. *Journal of Accounting and Economics*, 57(1), pp.43-57.
- Francis, B.B., Hasan, I., Sun, X. and Wu, Q., 2016. CEO political preference and corporate tax sheltering. *Journal of Corporate Finance*, 38, pp.37-53.
- Francis, J.R., Mehta, M.N. and Zhao, W., 2017. Audit office reputation shocks from gains and losses of major industry clients. *Contemporary Accounting Research*, 34(4), pp.1922-1974.
- Francis, J.R., Reichelt, K. and Wang, D., 2005. The pricing of national and city-specific reputations for industry expertise in the US audit market. *The accounting review*, 80(1), pp.113-136.
- Frank, M.M., Lynch, L.J. and Rego, S.O., 2009. Tax reporting aggressiveness and its relation to aggressive financial reporting. *The Accounting Review*, 84(2), pp.467-496.
- Freyberger, J., Neuhierl, A. and Weber, M., 2020. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5), pp.2326-2377.

- Gaertner, F.B., 2014. CEO After-Tax compensation incentives and corporate tax avoidance. *Contemporary Accounting Research*, 31(4), pp.1077-1102.
- Gallemore, J., Gipper, B. and Maydew, E., 2019. Banks as tax planning intermediaries. *Journal of Accounting Research*, 57(1), pp.169-209.
- Gallemore, J. and Labro, E., 2015. The importance of the internal information environment for tax avoidance. *Journal of Accounting and Economics*, 60(1), pp.149-167.
- Gallemore, J., Maydew, E.L. and Thornock, J.R., 2014. The reputational costs of tax avoidance. *Contemporary Accounting Research*, 31(4), pp.1103-1133.
- Gao, Y., Khan, M. and Tan, L., 2017. Further evidence on consequences of debt covenant violations. *Contemporary Accounting Research*, 34(3), pp.1489-1521.
- Ge, W., Koester, A. and McVay, S., 2017. Benefits and costs of Sarbanes-Oxley Section 404 (b) exemption: Evidence from small firms' internal control disclosures. *Journal of Accounting and Economics*, 63(2-3), pp.358-384.
- Gerakos, J. and Syverson, C., 2015. Competition in the audit market: Policy implications. *Journal of Accounting Research*, 53(4), pp.725-775.
- Ghosh, A. and Lustgarten, S., 2006. Pricing of initial audit engagements by large and small audit firms. *Contemporary Accounting Research*, 23(2), pp.333-368.
- Ghosh, A.A. and Tang, C.Y., 2015. Assessing financial reporting quality of family firms: The auditors' perspective. *Journal of Accounting and Economics*, 60(1), pp.95-116.
- Giglio, S., Liao, Y. and Xiu, D., 2020. Thousands of alpha tests. *Chicago Booth Research Paper*, (18-09), pp.2018-16.
- Glaeser, S. and Guay, W.R., 2017. Identification and generalizability in accounting research: A discussion of Christensen, Floyd, Liu, and Maffett (2017). *Journal of Accounting and Economics*, 64(2-3), pp.305-312.
- Goncharov, I., Riedl, E.J. and Sellhorn, T., 2014. Fair value and audit fees. *Review of Accounting Studies*, 19(1), pp.210-241.
- Gong, Q., Li, O.Z., Lin, Y. and Wu, L., 2016. On the benefits of audit market consolidation: Evidence from merged audit firms. *The Accounting Review*, 91(2), pp.463-488.
- Goodwin, J. and Wu, D., 2014. Is the effect of industry expertise on audit pricing an office-level

- or a partner-level phenomenon?. *Review of Accounting Studies*, 19(4), pp.1532-1578.
- Gow, I.D., Larcker, D.F. and Reiss, P.C., 2016. Causal inference in accounting research. *Journal of Accounting Research*, 54(2), pp.477-523.
- Graham, J.R., Hanlon, M., Shevlin, T. and Shroff, N., 2014. Incentives for tax planning and avoidance: Evidence from the field. *The Accounting Review*, 89(3), pp.991-1023.
- Graham, J.R., Lemmon, M.L. and Schallheim, J.S., 1998. Debt, leases, taxes, and the endogeneity of corporate tax status. *The journal of finance*, 53(1), pp.131-162.
- Graham, J.R. and Tucker, A.L., 2006. Tax shelters and corporate debt policy. *Journal of Financial Economics*, 81(3), pp.563-594.
- Greiner, A., Kohlbeck, M.J. and Smith, T.J., 2017. The relationship between aggressive real earnings management and current and future audit fees. *Auditing: A Journal of Practice & Theory*, 36(1), pp.85-107.
- Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), pp.2223-2273.
- Guan, Y., Su, L.N., Wu, D. and Yang, Z., 2016. Do school ties between auditors and client executives influence audit outcomes?. *Journal of accounting and economics*, 61(2-3), pp.506-525.
- Gupta, S. and Mills, L.F., 2002. Corporate multistate tax planning: Benefits of multiple jurisdictions. *Journal of Accounting and Economics*, 33(1), pp.117-139.
- Gul, F.A., 2006. Auditors' response to political connections and cronyism in Malaysia. *Journal of Accounting Research*, 44(5), pp.931-963.
- Gul, F.A. and Goodwin, J., 2010. Short-term debt maturity structures, credit ratings, and the pricing of audit services. *The Accounting Review*, 85(3), pp.877-909.
- Gul, F.A. and Krishnan, J., 2012. City-level auditor industry specialization, economies of scale, and audit pricing. *The Accounting Review*, 87(4), pp.1281-1307.
- Gutierrez, E., Minutti-Meza, M., Tatum, K.W. and Vulcheva, M., 2018. Consequences of adopting an expanded auditor's report in the United Kingdom. *Review of Accounting Studies*, 23(4), pp.1543-1587.
- Hackenbrack, K. and Knechel, W.R., 1997. Resource allocation decisions in audit

- engagements. *Contemporary Accounting Research*, 14(3), pp.481-499.
- Haislip, J.Z., Myers, L.A., Scholz, S. and Seidel, T.A., 2017. The consequences of audit-related earnings revisions. *Contemporary Accounting Research*, 34(4), pp.1880-1914.
- Hanlon, M. and Heitzman, S., 2010. A review of tax research. *Journal of accounting and Economics*, 50(2-3), pp.127-178.
- Hanlon, M., Lester, R. and Verdi, R., 2015. The effect of repatriation tax costs on US multinational investment. *Journal of Financial Economics*, 116(1), pp.179-196.
- Harvey, C.R., 2017. Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, 72(4), pp.1399-1440.
- Harvey, C.R., Liu, Y. and Zhu, H., 2016. ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1), pp.5-68.
- Hasan, I., HOI, C.K., Wu, Q. and Zhang, H., 2017. Does social capital matter in corporate decisions? Evidence from corporate tax avoidance. *Journal of Accounting Research*, 55(3), pp.629-668.
- Hastie, T., Tibshirani, R. and Friedman, J., 2009. The Elements of Statistical Learning Elements. 233 Spring Street, New York, NY 10013.
- Hay, D., 2013. Further evidence from meta-analysis of audit fee research. *International Journal of Auditing*, 17(2), pp.162-176.
- Hay, D.C., Knechel, W.R. and Wong, N., 2006. Audit fees: A meta-analysis of the effect of supply and demand attributes. *Contemporary accounting research*, 23(1), pp.141-191.
- He, X., Pittman, J.A., Rui, O.M. and Wu, D., 2017. Do social ties between external auditors and audit committee members affect audit quality?. *The Accounting Review*, 92(5), pp.61-87.
- Henry, E. and Sansing, R., 2018. Corporate tax avoidance: data truncation and loss firms. *Review of Accounting Studies*, 23(3), pp.1042-1070.
- Higgins, D., Omer, T.C. and Phillips, J.D., 2015. The influence of a firm's business strategy on its tax aggressiveness. *Contemporary Accounting Research*, 32(2), pp.674-702.
- Hill, M.D., Kubick, T.R., Lockhart, G.B. and Wan, H., 2013. The effectiveness and valuation of political tax minimization. *Journal of Banking & Finance*, 37(8), pp.2836-2849.
- Ho, T.K. and decision Forest, R., 1995, August. Document analysis and recognition, 1995.

- In *Proceedings of the third international conference on* (Vol. 1, pp. 278-282).
- Hogan, C.E. and Wilkins, M.S., 2008. Evidence on the audit risk model: Do auditors increase audit fees in the presence of internal control deficiencies?. *Contemporary Accounting Research*, 25(1), pp.219-242.
- Hoi, C.K., Wu, Q. and Zhang, H., 2013. Is corporate social responsibility (CSR) associated with tax avoidance? Evidence from irresponsible CSR activities. *The Accounting Review*, 88(6), pp.2025-2059.
- Hoitash, R. and Hoitash, U., 2018. Measuring accounting reporting complexity with XBRL. *The Accounting Review*, 93(1), pp.259-287.
- Hoopes, J.L., Merkley, K.J., Pacelli, J. and Schroeder, J.H., 2018. Audit personnel salaries and audit quality. *Review of Accounting Studies*, 23(3), pp.1096-1136.
- Hoopes, J.L., Mescall, D. and Pittman, J.A., 2012. Do IRS audits deter corporate tax avoidance?. *The accounting review*, 87(5), pp.1603-1639.
- Hope, O.K., Hu, D. and Zhao, W., 2017. Third-party consequences of short-selling threats: The case of auditor behavior. *Journal of Accounting and Economics*, 63(2-3), pp.479-498.
- Hope, O.K., Langli, J.C. and Thomas, W.B., 2012. Agency conflicts and auditing in private firms. *Accounting, Organizations and Society*, 37(7), pp.500-517.
- Hope, O.K., Ma, M.S. and Thomas, W.B., 2013. Tax avoidance and geographic earnings disclosure. *Journal of Accounting and Economics*, 56(2-3), pp.170-189.
- Hou, K., Xue, C. and Zhang, L., 2015. Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3), pp.650-705.
- Hou, K., Xue, C. and Zhang, L., 2020. Replicating anomalies. *The Review of Financial Studies*, 33(5), pp.2019-2133.
- Huang, H.W., Raghunandan, K., Huang, T.C. and Chiou, J.R., 2015. Fee discounting and audit quality following audit firm and audit partner changes: Chinese evidence. *The Accounting Review*, 90(4), pp.1517-1546.
- Hui, F.K., Warton, D.I. and Foster, S.D., 2015. Tuning parameter selection for the adaptive lasso using ERIC. *Journal of the American Statistical Association*, 110(509), pp.262-269.

- Huseynov, F. and Klamm, B.K., 2012. Tax avoidance, tax management and corporate social responsibility. *Journal of Corporate Finance*, 18(4), pp.804-827.
- Idson, T.L. and Oi, W.Y., 1999. Workers are more productive in large firms. *American Economic Review*, 89(2), pp.104-108.
- Ittner, C.D., 2014. Strengthening causal inferences in positivist field studies. *Accounting, Organizations and Society*, 39(7), pp.545-549.
- Jennings, J.N., Kim, J.M., Lee, J.A. and Taylor, D.J., 2020. Measurement Error and Bias in Causal Models in Accounting Research. *Available at SSRN 3731197*.
- Jensen, M.C. and Meckling, W.H., 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3(4), pp.305-360.
- Jha, A. and Chen, Y., 2015. Audit fees and social capital. *The Accounting Review*, 90(2), pp.611-639.
- Jiang, L. and Zhou, H., 2017. The role of audit verification in debt contracting: evidence from covenant violations. *Review of accounting studies*, 22(1), pp.469-501.
- Johannesson, E., Ohlson, J.A. and Zhai, W., 2020. The Explanatory Power of Explanatory Variables. *Available at SSRN 3622743*.
- Jones, J.J., 1991. Earnings management during import relief investigations. *Journal of accounting research*, 29(2), pp.193-228.
- Kanagaretnam, K., Krishnan, G.V. and Lobo, G.J., 2010. An empirical analysis of auditor independence in the banking industry. *The Accounting Review*, 85(6), pp.2011-2046.
- Kanagaretnam, K., Lee, J., Lim, C.Y. and Lobo, G., 2018. Societal trust and corporate tax avoidance. *Review of Accounting Studies*, 23(4), pp.1588-1628.
- Ke, B., Lennox, C.S. and Xin, Q., 2015. The effect of China's weak institutional environment on the quality of Big 4 audits. *The Accounting Review*, 90(4), pp.1591-1619.
- Kerr, J.N., 2019. Transparency, information shocks, and tax avoidance. *Contemporary Accounting Research*, 36(2), pp.1146-1183.
- Kerr, N.L., 1998. HARKing: Hypothesizing after the results are known. *Personality and social psychology review*, 2(3), pp.196-217.

- Khan, M., Srinivasan, S. and Tan, L., 2017. Institutional ownership and corporate tax avoidance: New evidence. *The Accounting Review*, 92(2), pp.101-122.
- Kim, C. and Zhang, L., 2016. Corporate political connections and tax aggressiveness. *Contemporary Accounting Research*, 33(1), pp.78-114.
- Kim, J., McGuire, S.T., Savoy, S., Wilson, R. and Caskey, J., 2019. How quickly do firms adjust to optimal levels of tax avoidance?. *Contemporary Accounting Research*, 36(3), pp.1824-1860.
- Kim, J.B., Liu, X. and Zheng, L., 2012. The impact of mandatory IFRS adoption on audit fees: Theory and evidence. *The Accounting Review*, 87(6), pp.2061-2094.
- Kim, J.H., Ahmed, K. and Ji, P.I., 2018. Significance testing in accounting research: a critical evaluation based on evidence. *Abacus*, 54(4), pp.524-546.
- Kim, Y., Li, H. and Li, S., 2014. CEO equity incentives and audit fees. *Contemporary Accounting Research*, Forthcoming.
- Kinney, JR, W.R. and Shepardson, M.L., 2011. Do control effectiveness disclosures require SOX 404 (b) internal control audits? A natural experiment with small US public companies. *Journal of Accounting Research*, 49(2), pp.413-448.
- Klassen, K.J. and Laplante, S.K., 2012. The effect of foreign reinvestment and financial reporting incentives on cross-jurisdictional income shifting. *Contemporary Accounting Research*, 29(3), pp.928-955.
- Klassen, K.J. and Laplante, S.K., 2012. Are US multinational corporations becoming more aggressive income shifters?. *Journal of Accounting Research*, 50(5), pp.1245-1285.
- Knechel, W.R. and Payne, J.L., 2001. Additional evidence on audit report lag. *Auditing: A Journal of Practice & Theory*, 20(1), pp.137-146.
- Kohlbeck, M. and Mayhew, B.W., 2017. Are related party transactions red flags?. *Contemporary Accounting Research*, 34(2), pp.900-928.
- Krishnan, G. and Visvanathan, G., 2009. Do auditors price audit committee's expertise? The case of accounting versus nonaccounting financial experts. *Journal of Accounting, Auditing & Finance*, 24(1), pp.115-144.
- Krupa, J. and Minutti-Meza, M., 2021. Regression and Machine Learning Methods to Predict

Discrete Outcomes in Accounting Research. *Available at SSRN 3801353*.

- Kubick, T.R. and Lockhart, G.B., 2016. Do external labor market incentives motivate CEOs to adopt more aggressive corporate tax reporting preferences?. *Journal of Corporate Finance*, 36, pp.255-277.
- Kubick, T.R., Lockhart, G.B., Mills, L.F. and Robinson, J.R., 2017. IRS and corporate taxpayer effects of geographic proximity. *Journal of Accounting and Economics*, 63(2-3), pp.428-453.
- Kubick, T.R., Lynch, D.P., Mayberry, M.A. and Omer, T.C., 2015. Product market power and tax avoidance: Market leaders, mimicking strategies, and stock returns. *The Accounting Review*, 90(2), pp.675-702.
- Kubick, T.R., Lynch, D.P., Mayberry, M.A. and Omer, T.C., 2016. The effects of regulatory scrutiny on tax avoidance: An examination of SEC comment letters. *The Accounting Review*, 91(6), pp.1751-1780.
- Kuo, N.T. and Lee, C.F., 2016. A potential benefit of increasing book–tax conformity: evidence from the reduction in audit fees. *Review of Accounting Studies*, 21(4), pp.1287-1326.
- Lai, K.M., Srinidhi, B., Gul, F.A. and Tsui, J.S., 2017. Board gender diversity, auditor fees, and auditor choice. *Contemporary Accounting Research*, 34(3), pp.1681-1714.
- Lang, L., Ofek, E. and Stulz, R., 1996. Leverage, investment, and firm growth. *Journal of financial Economics*, 40(1), pp.3-29.
- Lanis, R. and Richardson, G., 2012. Corporate social responsibility and tax aggressiveness: An empirical analysis. *Journal of Accounting and Public Policy*, 31(1), pp.86-108.
- Law, K.K. and Mills, L.F., 2015. Taxes and financial constraints: Evidence from linguistic cues. *Journal of Accounting Research*, 53(4), pp.777-819.
- Law, K.K. and Mills, L.F., 2017. Military experience and corporate tax avoidance. *Review of Accounting Studies*, 22(1), pp.141-184.
- Lawson, B.P. and Wang, D., 2016. The earnings quality information content of dividend policies and audit pricing. *Contemporary Accounting Research*, 33(4), pp.1685-1719.
- Lennox, C. and Li, B., 2012. The consequences of protecting audit partners' personal assets from the threat of liability. *Journal of Accounting and Economics*, 54(2-3), pp.154-173.

- Lennox, C., Lisowsky, P. and Pittman, J., 2013. Tax aggressiveness and accounting fraud. *Journal of Accounting Research*, 51(4), pp.739-778.
- Lennox, C.S., Francis, J.R. and Wang, Z., 2012. Selection models in accounting research. *The accounting review*, 87(2), pp.589-616.
- Lennox, C.S. and Kausar, A., 2017. Estimation risk and auditor conservatism. *Review of Accounting Studies*, 22(1), pp.185-216.
- Liao, P.C. and Radhakrishnan, S., 2016. The effects of the auditor's insurance role on reporting conservatism and audit quality. *The Accounting Review*, 91(2), pp.587-602.
- Lin, K.Z., Mills, L.F., Zhang, F. and Li, Y., 2018. Do political connections weaken tax enforcement effectiveness?. *Contemporary Accounting Research*, 35(4), pp.1941-1972.
- Lipsey, M.W. and Wilson, D.B., 2001. *Practical meta-analysis*. SAGE publications, Inc.
- Lisowsky, P., 2010. Seeking shelter: Empirically modeling tax shelters using financial statement information. *The Accounting Review*, 85(5), pp.1693-1720.
- Lisowsky, P., Robinson, L. and Schmidt, A., 2013. Do publicly disclosed tax reserves tell us about privately disclosed tax shelter activity?. *Journal of Accounting Research*, 51(3), pp.583-629.
- Liu, C. and Wang, T., 2006. Auditor liability and business investment. *Contemporary Accounting Research*, 23(4), pp.1051-1071.
- Lyon, J.D. and Maher, M.W., 2005. The importance of business risk in setting audit fees: Evidence from cases of client misconduct. *Journal of Accounting Research*, 43(1), pp.133-151.
- Magnan, M.L., 2008. Discussion of "Audit pricing, legal liability regimes, and Big 4 premiums: Theory and cross-country evidence". *Contemporary Accounting Research*, 25(1), pp.101-108.
- Manzon Jr, G.B. and Plesko, G.A., 2001. The relation between financial and tax reporting measures of income. *Tax L. Rev.*, 55, p.175.
- Masli, A., Peters, G.F., Richardson, V.J. and Sanchez, J.M., 2010. Examining the potential benefits of internal control monitoring technology. *The Accounting Review*, 85(3), pp.1001-1034.

- McGuire, S.T., Omer, T.C. and Wang, D., 2012. Tax avoidance: Does tax-specific industry expertise make a difference?. *The Accounting Review*, 87(3), pp.975-1003.
- McGuire, S.T., Wang, D. and Wilson, R.J., 2014. Dual class ownership and tax avoidance. *The Accounting Review*, 89(4), pp.1487-1516.
- Meinshausen, N. and Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics*, 37(1), pp.246-270.
- Menon, K. and Williams, D.D., 2016. Audit report restrictions in debt covenants. *Contemporary Accounting Research*, 33(2), pp.682-717.
- Messier Jr, W.F., Reynolds, J.K., Simon, C.A. and Wood, D.A., 2011. The effect of using the internal audit function as a management training ground on the external auditor's reliance decision. *The Accounting Review*, 86(6), pp.2131-2154.
- Miller, M.H., 1977. Debt and taxes. *the Journal of Finance*, 32(2), pp.261-275.
- Mills, L.F., 1998. Book-tax differences and Internal Revenue Service adjustments. *Journal of Accounting research*, 36(2), pp.343-356.
- Minutti-Meza, M., 2013. Does auditor industry specialization improve audit quality?. *Journal of Accounting Research*, 51(4), pp.779-817.
- Minutti-Meza, M., 2014. Issues in examining the effect of auditor litigation on audit fees. *Journal of Accounting Research*, Forthcoming.
- Mitra, S. and Hossain, M., 2007. Ownership composition and non-audit service fees. *Journal of Business Research*, 60(4), pp.348-356.
- Modigliani, F. and Miller, M.H., 1963. Corporate income taxes and the cost of capital: a correction. *The American economic review*, 53(3), pp.433-443.
- Mullainathan, S. and Spiess, J., 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), pp.87-106.
- Murphy, F. and Miller, S., 2018. An Empiricist's Guide to Nonparametric Analysis in Accounting.
- Nazemi, A. and Fabozzi, F.J., 2018. Macroeconomic variable selection for creditor recovery rates. *Journal of Banking & Finance*, 89, pp.14-25.
- Newman, D.P., Patterson, E.R. and Smith, J.R., 2005. The role of auditing in investor

- protection. *The Accounting Review*, 80(1), pp.289-313.
- Numan, W. and Willekens, M., 2012. An empirical test of spatial competition in the audit market. *Journal of Accounting and Economics*, 53(1-2), pp.450-465.
- Nuzzo, R., 2014. Scientific method: statistical errors. *Nature News*, 506(7487), p.150.
- Ohlson, J.A., 2015. Accounting research and common sense. *Abacus*, 51(4), pp.525-535.
- Oi, W.Y., 1983. Heterogeneous firms and the organization of production. *Economic Inquiry*, 21(2), pp.147-171.
- Omer, T.C., Molloy, K.H. and Ziebart, D.A., 1993. An investigation of the firm size—effective tax rate relation in the 1980s. *Journal of Accounting, Auditing & Finance*, 8(2), pp.167-182.
- Parkash, M. and Venable, C.F., 1993. Auditee incentives for auditor independence: The case of nonaudit services. *Accounting Review*, pp.113-133.
- Pearson, T. and Trompeter, G., 1994. Competition in the market for audit services: The effect of supplier concentration on audit fees. *Contemporary accounting research*, 11(1), pp.115-135.
- Petković, M., Kocev, D. and Džeroski, S., 2020. Feature ranking for multi-target regression. *Machine Learning*, 109(6), pp.1179-1204.
- Phillips, J.D., 2003. Corporate tax-planning effectiveness: The role of compensation-based incentives. *The Accounting Review*, 78(3), pp.847-874.
- Pincus, M., Tian, F., Wellmeyer, P. and Xu, S.X., 2017. Do clients' enterprise systems affect audit quality and efficiency?. *Contemporary Accounting Research*, 34(4), pp.1975-2021.
- Powers, K., Robinson, J.R. and Stomberg, B., 2016. How do CEO incentives affect corporate tax planning and financial reporting of income taxes?. *Review of Accounting Studies*, 21(2), pp.672-710.
- Rapach, D.E., Strauss, J.K. and Zhou, G., 2013. International stock return predictability: what is the role of the United States?. *The Journal of Finance*, 68(4), pp.1633-1662.
- Reeb, D.M. and Zhao, W., 2018. Dissecting Innovation. *Available at SSRN 3175055*.
- Rego, S.O., 2003. Tax-avoidance activities of US multinational corporations. *Contemporary Accounting Research*, 20(4), pp.805-833.

- Rego, S.O. and Wilson, R., 2012. Equity risk incentives and corporate tax aggressiveness. *Journal of Accounting Research*, 50(3), pp.775-810.
- Richardson, G. and Lanis, R., 2007. Determinants of the variability in corporate effective tax rates and tax reform: Evidence from Australia. *Journal of accounting and public policy*, 26(6), pp.689-704.
- Rosenthal, R., 1979. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3), p.638.
- Schatzberg, J.W., Sevcik, G.R., Shapiro, B.P., Thorne, L. and Wallace, R.O., 2005. A reexamination of behavior in experimental audit markets: The effects of moral reasoning and economic incentives on auditor reporting and fees. *Contemporary Accounting Research*, 22(1), pp.229-264.
- Seidel, T.A., 2017. Auditors' response to assessments of high control risk: Further insights. *Contemporary Accounting Research*, 34(3), pp.1340-1377.
- Shevlin, T., 2001. Corporate tax shelters and book-tax differences. *Tax L. Rev.*, 55, p.427.
- Siegfried, J.J., 1973. The relationship between economic structure and the effect of political influence: empirical evidence from the Federal Corporation Income Tax Program.
- Simunic, D.A., 1980. The pricing of audit services: Theory and evidence. *Journal of accounting research*, pp.161-190.
- Srinidhi, B.N., He, S. and Firth, M., 2014. The effect of governance on specialist auditor choice and audit fees in US family firms. *The Accounting Review*, 89(6), pp.2297-2329.
- Stone, D.N., 2018. The “new statistics” and nullifying the null: Twelve actions for improving quantitative accounting research quality and integrity. *Accounting Horizons*, 32(1), pp.105-120.
- Stoppiglia, H., Dreyfus, G., Dubois, R. and Oussar, Y., 2003. Ranking a random feature for variable and feature selection. *The Journal of Machine Learning Research*, 3, pp.1399-1414.
- Swanquist, Q.T. and Whited, R.L., 2018. Out of control: The (over) use of controls in accounting research. *Available at SSRN 3209571*.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal*

- Statistical Society: Series B (Methodological)*, 58(1), pp.267-288.
- Vafeas, N. and Waagelein, J.F., 2007. The association between audit committees, compensation incentives, and corporate audit fees. *Review of Quantitative Finance and Accounting*, 28(3), pp.241-255.
- Varian, H.R., 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), pp.3-28.
- Venkataraman, R., Weber, J.P. and Willenborg, M., 2008. Litigation risk, audit quality, and audit fees: Evidence from initial public offerings. *The Accounting Review*, 83(5), pp.1315-1345.
- Walker, P.L. and Casterella, J.R., 2000. The role of auditee profitability in pricing new audit engagements. *Auditing: A Journal of Practice & Theory*, 19(1), pp.157-167.
- Wanous, J.P., Sullivan, S.E. and Malinak, J., 1989. The role of judgment calls in meta-analysis. *Journal of Applied Psychology*, 74(2), p.259.
- Watts, R.L. and Zimmerman, J.L., 1978. Towards a positive theory of the determination of accounting standards. *Accounting review*, pp.112-134.
- Weisbach, D.A., 2003, January. Corporate tax avoidance. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* (Vol. 96, pp. 9-15). National Tax Association.
- Wilde, J.H., 2017. The deterrent effect of employee whistleblowing on firms' financial misreporting and tax aggressiveness. *The Accounting Review*, 92(5), pp.247-280.
- Wilde, J.H. and Wilson, R.J., 2018. Perspectives on corporate tax planning: observations from the past decade. *The Journal of the American Taxation Association*, 40(2), pp.63-81.
- Wilson, R.J., 2009. An examination of corporate tax shelter participants. *The Accounting Review*, 84(3), pp.969-999.
- Wong, J., 1988. Political costs and an intraperiod accounting choice for export tax credits. *Journal of Accounting and Economics*, 10(1), pp.37-51.
- Wu, M.G., 2006. An economic analysis of audit and nonaudit services: The trade-off between competition crossovers and knowledge spillovers. *Contemporary Accounting Research*, 23(2), pp.527-554.
- Wysocki, P., 2010. Corporate compensation policies and audit fees. *Journal of Accounting and*

- Economics*, 49(1-2), pp.155-160.
- Yermack, D., 1996. Higher market valuation of companies with a small board of directors. *Journal of financial economics*, 40(2), pp.185-211.
- Zerni, M., 2012. Audit partner specialization and audit fees: Some evidence from Sweden. *Contemporary Accounting Research*, 29(1), pp.312-340.
- Zhang, C.H. and Huang, J., 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4), pp.1567-1594.
- Zeff, S.A. and Dyckman, T.R., 2018. A historical study of the first 30 years of Accounting Horizons. *Accounting Historians Journal*, 45(1), pp.115-131.
- Zimmerman, J.L., 1983. Taxes and firm size. *Journal of accounting and economics*, 5, pp.119-149.
- Zou, H., 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), pp.1418-1429.

Appendices for Chapter 3

Appendix 3E

Appendix 3E.1 Adaptive LASSO analysis in Step 1 (main test)

This table reports the results of Adaptive LASSO analysis in Step 1. All variables are defined in Appendix 3A. Out# column shows the frequency of variable coefficient shrinking to zero in Adaptive LASSO analysis.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Out#
Size_ta	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Size_mv	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Size_sale	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
NGS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
NBS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
NBSU	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
NS	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	4
NSU	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	4
FO_D	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	No	No	7
FO_sale	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
FO_pifo	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
FO_fca	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	4
Lev_debt	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	2
ROA_nibs	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
ROA_ib	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
ROA_ebitda	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	2
Loss_ni	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	3

Loss_ibc	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	5
Loss_libc	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Age1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Age2	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
MB	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	No	No	No	No	Yes	Yes	No	No	12
TQ	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No	6
CFO	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	3
Issue1	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	4
Issue2	No	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	8
Issue3	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Issue4	Yes	No	Yes	Yes	No	No	Yes	No	Yes	No	No	No	Yes	No	Yes	Yes	Yes	Yes	No	9
Issue5	No	No	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	13
Issue6	Yes	No	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	No	No	10
Bank	Yes	Yes	No	Yes	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	6
Utility	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	1
ACCF	NA ³²	NA	NA	NA	No	Yes	Yes	No	No	No	Yes	No	No	No	No	No	No	No	Yes	10
HighTech	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
CATA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
CUR_r	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	2
Quick_r	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	6
Rect	Yes	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	Yes	15
Invt	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
INVREC	No	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	6
Intan	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0

³² Owing to data unavailability, *ACCF* and *MW* are not available for the analysis from 2000–2003, and 2000–2002, respectively.

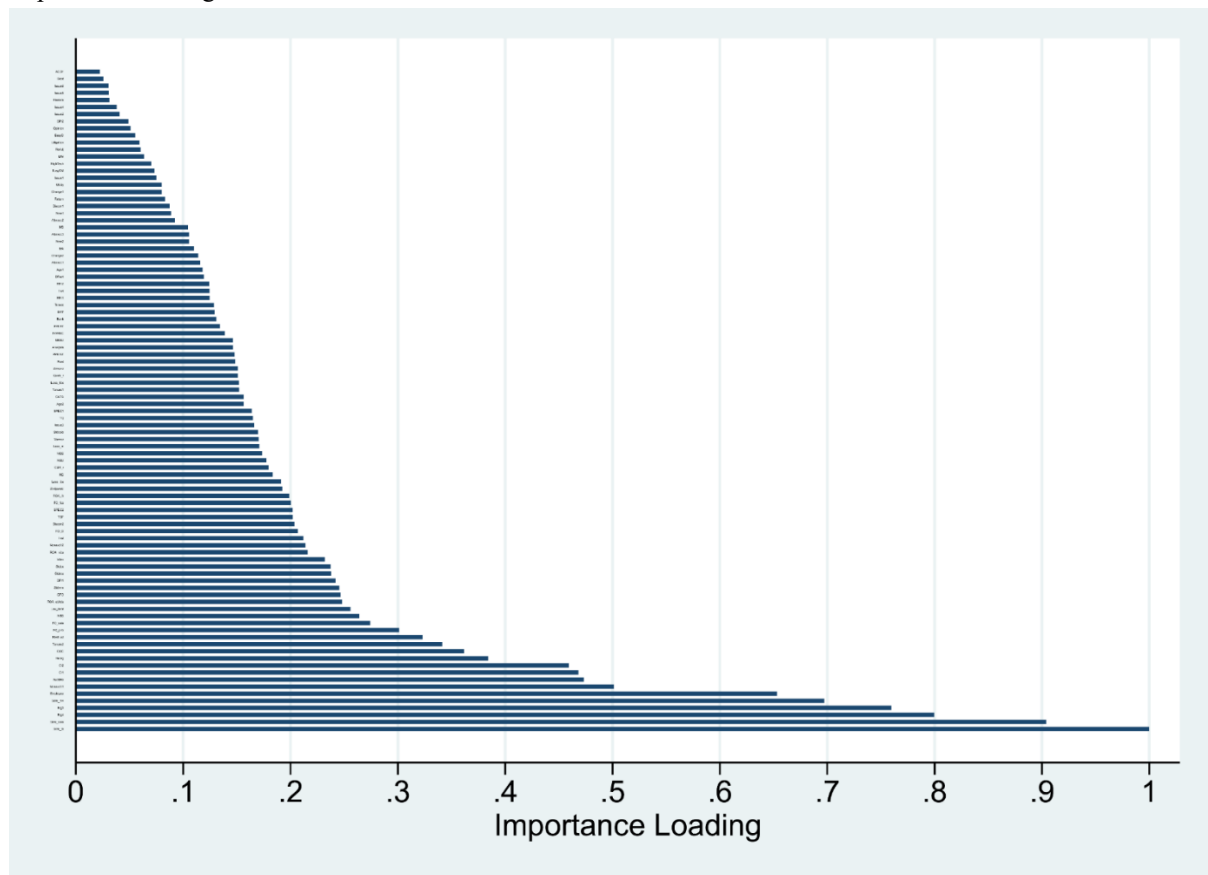
PPEAT	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	3
PPEINT	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	6
DAT	No	Yes	No	No	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	No	11
DRev	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	3
Abnacc1	No	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No	No	No	No	Yes	No	No	No	14
Abnacc2	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	3
Abnacc3	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	No	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	7
Totacc	No	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	Yes	No	Yes	Yes	14
MA	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Restate	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	4
Restruct	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Discon1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Discon2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
SPI1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
SPI2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Inst	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
ReAdj	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	No	No	Yes	Yes	Yes	No	No	10
Return	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No	No	Yes	No	No	Yes	No	10
StdDoc	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Stdsale	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	1
Stdroa	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Stdmm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	1
Varmsr	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	3
Beat	Yes	Yes	Yes	Yes	No	No	No	Yes	No	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes	8
Analysts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0

Employee	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Atmanz	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	3
Zmijewski	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Litigation	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
Big4	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	3
Big5	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Nonaud~1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Nonaud~2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
TSF	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	6
AuditRe	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Opinion	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	2
GOC	Yes	Yes	Yes	No	No	No	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	6
MW	NA	NA	NA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Change1	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	4
Change2	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
Tenure1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Tenure2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
New1	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	3
New2	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
CI1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
CI2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	1
BusyD	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	2
BusyDM	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	9
Relag	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
SPEC1	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	3

SPEC2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
HHI1	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	6
HHI2	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	4

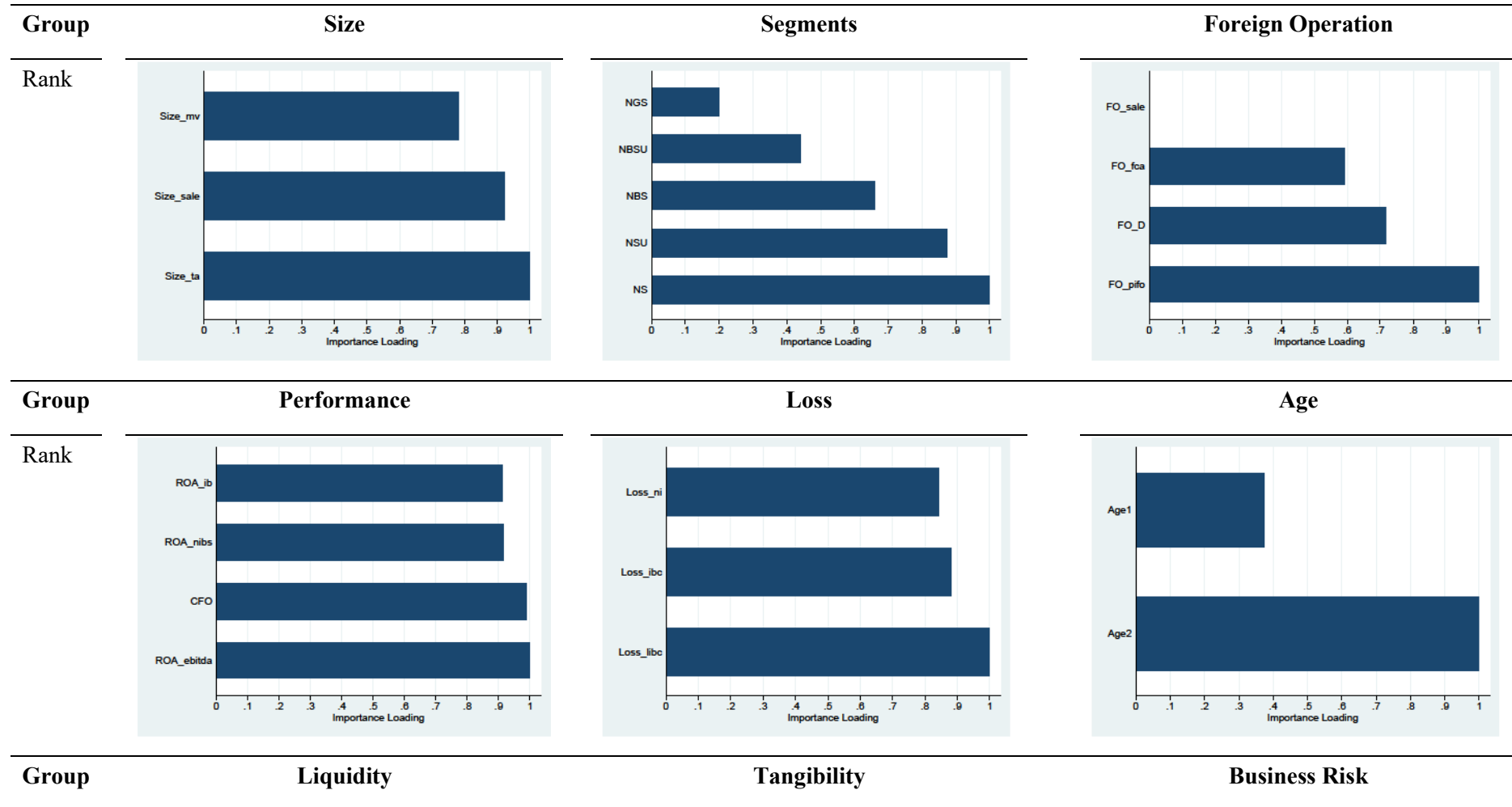
Appendix 3E.2 Random Forest analysis in Step 1 (main test)

This figure plots the results of Random Forest analysis in Step 1. The variables are ranked according to their importance loading.

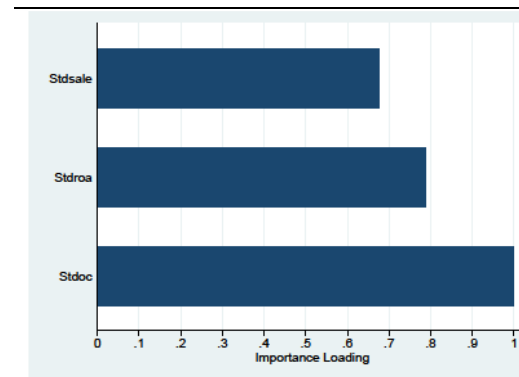
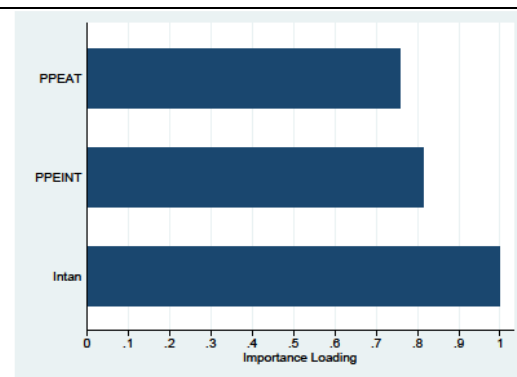
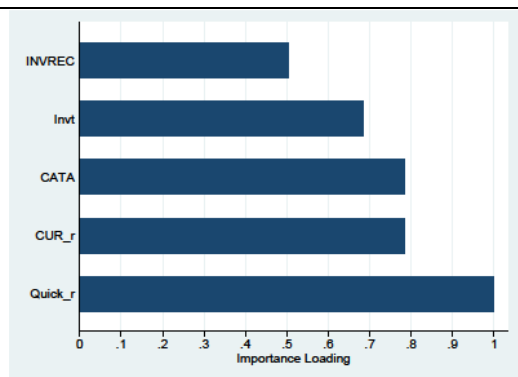


Appendix 3F Random Forest analysis in Step 2 (main test)

This table reports the results of Random Forest analysis within each group in Step 2. Variables are ranked according to their importance loading.



Rank



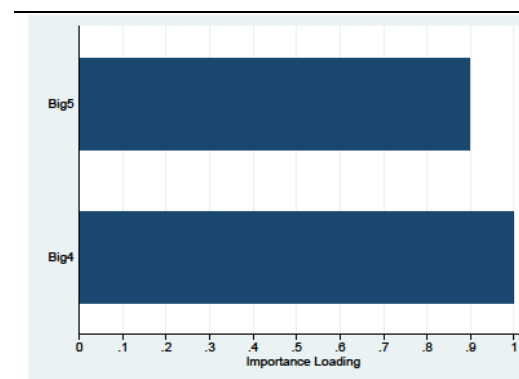
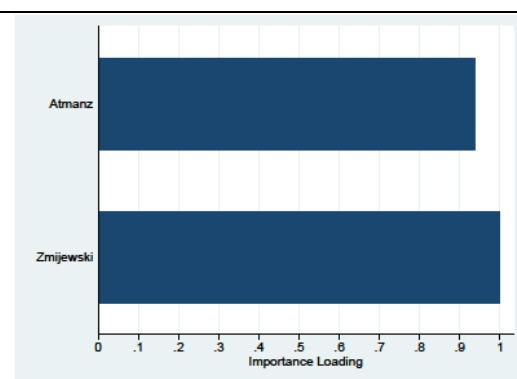
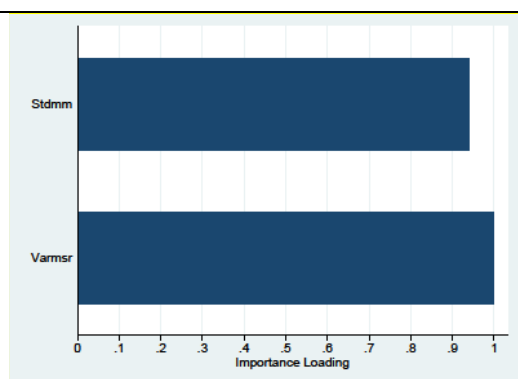
Group

Variance

Bankruptcy

Big 4

Rank



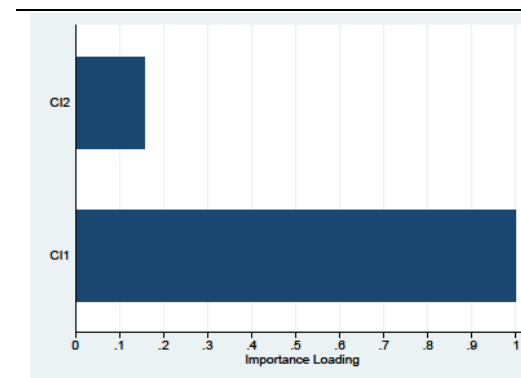
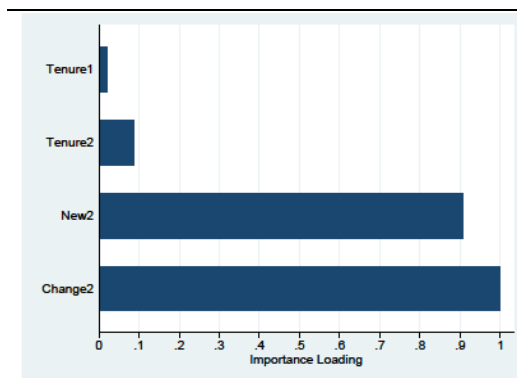
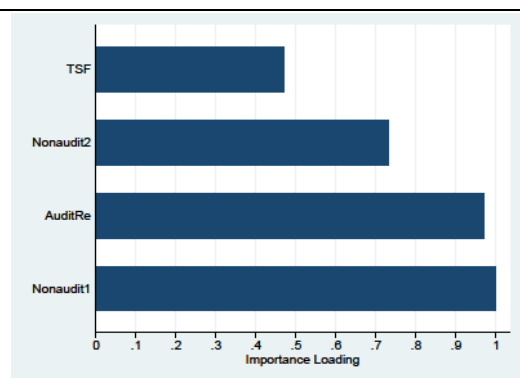
Group

Fees

Tenure

Importance

Rank

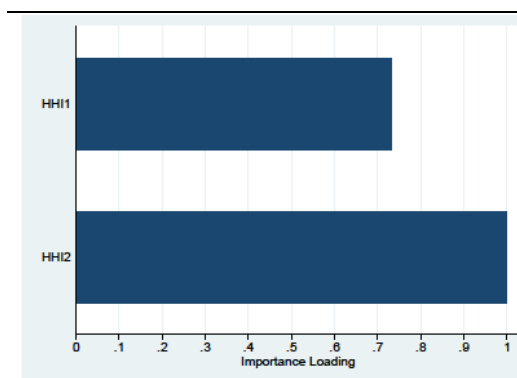
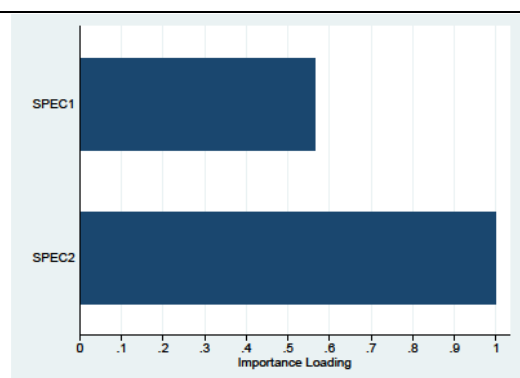


Group

Specialist

Competition

Rank



Appendix 3G

Appendix 3G.1 Adaptive LASSO analysis in Step 3 (main test)

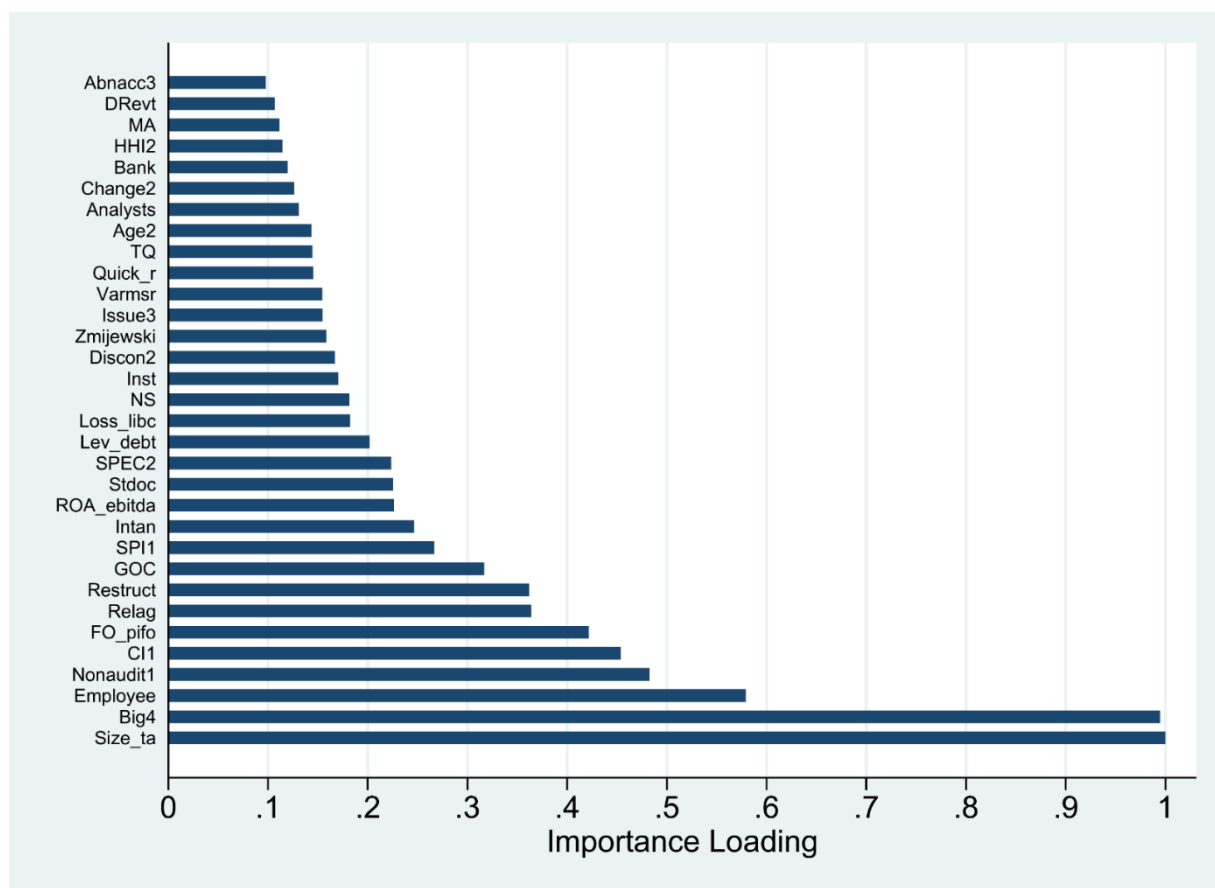
This table reports the results of Adaptive LASSO analysis in Step 3. All variables are defined in Appendix 3A. Out# column shows the frequency of variable coefficient shrinking to zero in Adaptive LASSO analysis.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Out#
Size_ta	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
NS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
FO_pifo	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Lev_debt	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
ROA_ebitda	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	5
Loss_libc	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	4
Age2	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
TQ	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Issue3	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	7
Bank	Yes	Yes	No	No	Yes	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	8
Quick_r	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No	Yes	Yes	Yes	9
Intan	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
DRevt	Yes	Yes	No	No	Yes	No	No	No	No	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No	11
Abnacc3	Yes	No	Yes	No	No	No	No	No	No	No	Yes	No	No	Yes	No	Yes	Yes	No	No	13
MA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	5
Restruct	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Discon2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
SPI1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Inst	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	4

StdDoc	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
VarmSr	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	No	No	Yes	Yes	No	6
Analysts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	1
Employee	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Zmijewski	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Big4	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Nonaud~1	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	4
GOC	Yes	Yes	Yes	Yes	No	No	Yes	No	No	Yes	Yes	No	No	Yes	No	No	Yes	No	No	10
Change2	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	5
CI1	Yes	No	No	No	No	No	No	No	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	10
Relag	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
SPEC2	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	8
HHI2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0

Appendix 3G.2 Random Forest analysis in Step 3 (main test)

This figure plots the results of Random Forest analysis in Step 3. The variables are ranked according to their importance loading.



Appendix 3H

Appendix 3H.1 Adaptive LASSO and Random Forest analyses in Step 1 (further analysis)

This table reports the results of Adaptive LASSO and Random Forest analyses in Step 1, including corporate governance attributes. All variables are defined in Appendix 3A. AL column shows the frequency of the variable coefficient shrinking to zero in Adaptive LASSO analysis. RF column shows the variable importance ranking and the importance loading in Random Forest analysis.

	AL	RF	Pass		AL	RF	Pass
Size_ta	0	2(0.3-1)	Yes	Restate	5	95(0-0.1)	No
Size_mv	0	5(0.3-1)	Yes	Restruct	0	14(0.2-0.3)	Yes
Size_sale	0	4(0.3-1)	Yes	Discon1	2	82(0-0.1)	No
NGS	1	43(0.1-0.2)	Yes	Discon2	1	51(0.1-0.2)	Yes
NBS	0	17(0.2-0.3)	Yes	SPI1	0	21(0.2-0.3)	Yes
NBSU	0	46(0.1-0.2)	Yes	SPI2	0	91(0-0.1)	No
NS	6	31(0.2-0.3)	Yes	Inst	1	23(0.2-0.3)	Yes
NSU	6	37(0.1-0.2)	Yes	ReAdj	9	83(0-0.1)	No
FO_D	10	41(0.1-0.2)	Yes	Return	9	74(0.1-0.2)	Yes
FO_sale	0	19(0.2-0.3)	Yes	StdDoc	1	20(0.2-0.3)	Yes
FO_pifo	0	10(0.3-1)	Yes	Stdsale	3	54(0.1-0.2)	Yes
FO_fca	3	44(0.1-0.2)	Yes	Stdtoa	1	36(0.1-0.2)	Yes
Lev_debt	3	22(0.2-0.3)	Yes	Stdmm	0	18(0.2-0.3)	Yes
ROA_nibs	5	42(0.1-0.2)	Yes	Varmr	1	28(0.2-0.3)	Yes
ROA_ib	6	38(0.1-0.2)	Yes	Beat	10	99(0-0.1)	No
ROA_ebitda	2	26(0.2-0.3)	Yes	Analysts	0	29(0.2-0.3)	Yes
Loss_ni	4	65(0.1-0.2)	Yes	Employee	0	6(0.3-1)	Yes
Loss_ibc	5	61(0.1-0.2)	Yes	Atmanz	3	50(0.1-0.2)	Yes
Loss_libc	1	78(0-0.1)	No	Zmijewski	1	35(0.1-0.2)	Yes
Age1	2	53(0.1-0.2)	Yes	Litigation	3	88(0-0.1)	No
Age2	3	34(0.1-0.2)	Yes	Big4	2	1(0.3-1)	Yes
MB	12	72(0.1-0.2)	No	Big5	5	3(0.3-1)	Yes
TQ	5	52(0.1-0.2)	Yes	Nonaud~1	0	7(0.3-1)	Yes
CFO	5	32(0.2-0.3)	Yes	Nonaud~2	0	33(0.2-0.3)	Yes
Issue1	3	76(0-0.1)	No	TSF	13	39(0.1-0.2)	No
Issue2	7	92(0-0.1)	No	AuditRe	1	8(0.3-1)	Yes
Issue3	1	55(0.1-0.2)	Yes	Opinion	1	90(0-0.1)	No
Issue4	12	93(0-0.1)	No	GOC	5	15(0.2-0.3)	Yes
Issue5	14	96(0-0.1)	No	MW	0	86(0-0.1)	No
Issue6	13	97(0-0.1)	No	Change1	2	81(0-0.1)	No
Bank	9	67(0.1-0.2)	Yes	Change2	3	59(0.1-0.2)	Yes
Utility	2	77(0-0.1)	No	Tenure1	0	47(0.1-0.2)	Yes

ACCF	9	98(0-0.1)	No	Tenure2	1	24(0.2-0.3)	Yes
HighTech	3	87(0-0.1)	No	New1	2	84(0-0.1)	No
CATA	0	49(0.1-0.2)	Yes	New2	3	79(0-0.1)	No
CUR_r	4	45(0.1-0.2)	Yes	CI1	2	13(0.3-1)	Yes
Quick_r	5	48(0.1-0.2)	Yes	CI2	1	12(0.3-1)	Yes
Rect	14	56(0.1-0.2)	No	BusyD	2	89(0-0.1)	No
Invt	1	66(0.1-0.2)	Yes	BusyDM	11	85(0-0.1)	No
INVREC	8	57(0.1-0.2)	Yes	Relag	1	9(0.3-1)	Yes
Intan	0	25(0.2-0.3)	Yes	SPEC1	3	40(0.1-0.2)	Yes
PPEAT	4	60(0.1-0.2)	Yes	SPEC2	3	30(0.2-0.3)	Yes
PPEINT	8	58(0.1-0.2)	Yes	HHI1	5	63(0.1-0.2)	Yes
DAT	7	70(0.1-0.2)	Yes	HHI2	5	62(0.1-0.2)	Yes
DRevt	6	69(0.1-0.2)	Yes	Boardind	0	16(0.2-0.3)	Yes
Abnacc1	14	68(0.1-0.2)	No	Boardsize	0	11(0.3-1)	Yes
Abnacc2	5	80(0-0.1)	No	Auditsize	0	27(0.2-0.3)	Yes
Abnacc3	5	73(0.1-0.2)	Yes	AFE	6	71(0.1-0.2)	Yes
Totacc	11	64(0.1-0.2)	No	CEOChair	5	94(0-0.1)	No
MA	2	75(0-0.1)	No				

Appendix 3H.2 Adaptive LASSO analysis in Step 1 (further analysis)

This table reports the results of Adaptive LASSO analysis in Step 1, including corporate governance attributes. All variables are defined in Appendix 3A. Out# column shows the frequency of variable coefficient shrinking to zero in Adaptive LASSO analysis.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Out #
Size_ta	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Size_mv	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Size_sale	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
NGS	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
NBS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
NBSU	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
NS	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	6
NSU	No	Yes	Yes	Yes	No	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	6
FO_D	No	No	No	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	No	No	10
FO_sale	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
FO_pifo	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
FO_fca	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	3
Lev_debt	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	3
ROA_nibs	No	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	5
ROA_ib	No	Yes	No	No	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	6
ROA_ebitda	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	2
Loss_ni	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	4
Loss_ibc	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	5
Loss_libc	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Age1	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	2

Age2	No	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
MB	No	Yes	Yes	No	No	No	No	Yes	No	No	No	No	No	No	Yes	Yes	Yes	Yes	No	12
TQ	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	5
CFO	No	No	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	5
Issue1	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	3
Issue2	No	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	7
Issue3	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Issue4	No	No	No	No	No	No	Yes	No	Yes	No	No	No	Yes	No	Yes	Yes	Yes	Yes	No	12
Issue5	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	14
Issue6	No	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	No	No	No	No	No	Yes	No	No	13
Bank	No	Yes	No	No	No	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	9
Utility	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	2
ACCF	NA	NA	NA	NA	No	Yes	Yes	No	No	No	Yes	Yes	No	No	No	No	No	Yes	Yes	9
HighTech	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
CATA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
CUR_r	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	4
Quick_r	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	5
Rect	Yes	Yes	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes	14
Invt	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
INVREC	No	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	No	8
Intan	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
PPEAT	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	4
PPEINT	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	8
DAT	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	No	Yes	No	No	7
DRev	No	Yes	No	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	6

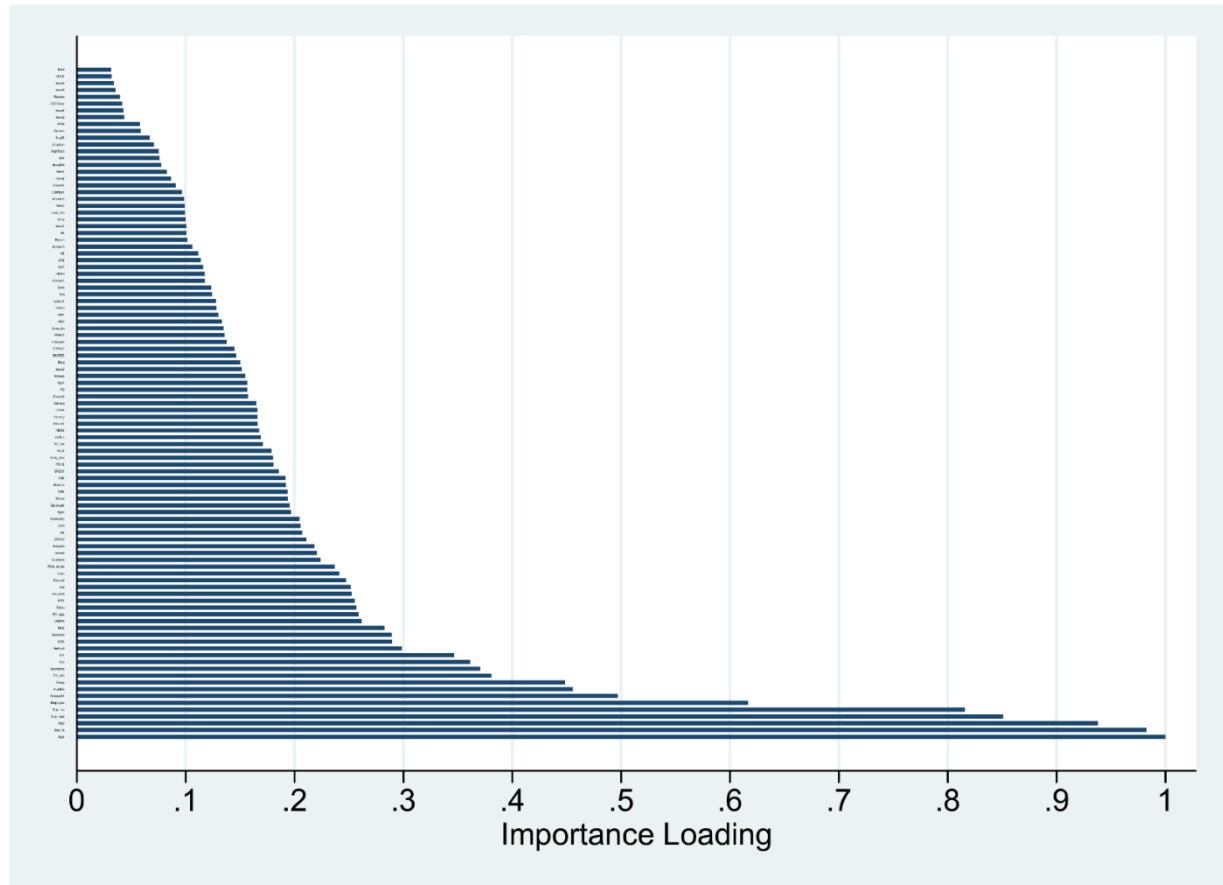
Abnacc1	No	Yes	No	No	No	No	No	No	Yes	Yes	Yes	No	No	No	No	No	No	Yes	No	14
Abnacc2	No	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	5
Abnacc3	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	5
Totacc	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	Yes	No	No	No	No	No	No	Yes	Yes	11
MA	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
Restate	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	5
Restruct	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Discon1	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
Discon2	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
SPI1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
SPI2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Inst	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
ReAdj	No	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	No	No	Yes	Yes	Yes	No	No	9
Return	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	No	Yes	No	No	Yes	Yes	9
Stdoc	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	1
Stdsale	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	3
Stdroa	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Stdmm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Varmsr	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	1
Beat	No	Yes	No	Yes	No	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	No	No	Yes	Yes	10
Analysts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Employee	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Atmanz	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	3
Zmijewski	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Litigation	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3

Big4	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
Big5	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	5
Nonaud~1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Nonaud~2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
TSF	No	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	13
AuditRe	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Opinion	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
GOC	No	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	5
MW	NA	NA	NA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Change1	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
Change2	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
Tenure1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Tenure2	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
New1	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
New2	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	3
CI1	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2
CI2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	1
BusyD	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	2
BusyDM	No	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	No	11
Relag	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
SPEC1	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	3
SPEC2	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
HHI1	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	5
HHI2	Yes	No	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	5
Boardind	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0

Boardsize	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Auditsize	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
AFE	No	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	6
CEOChair	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	5

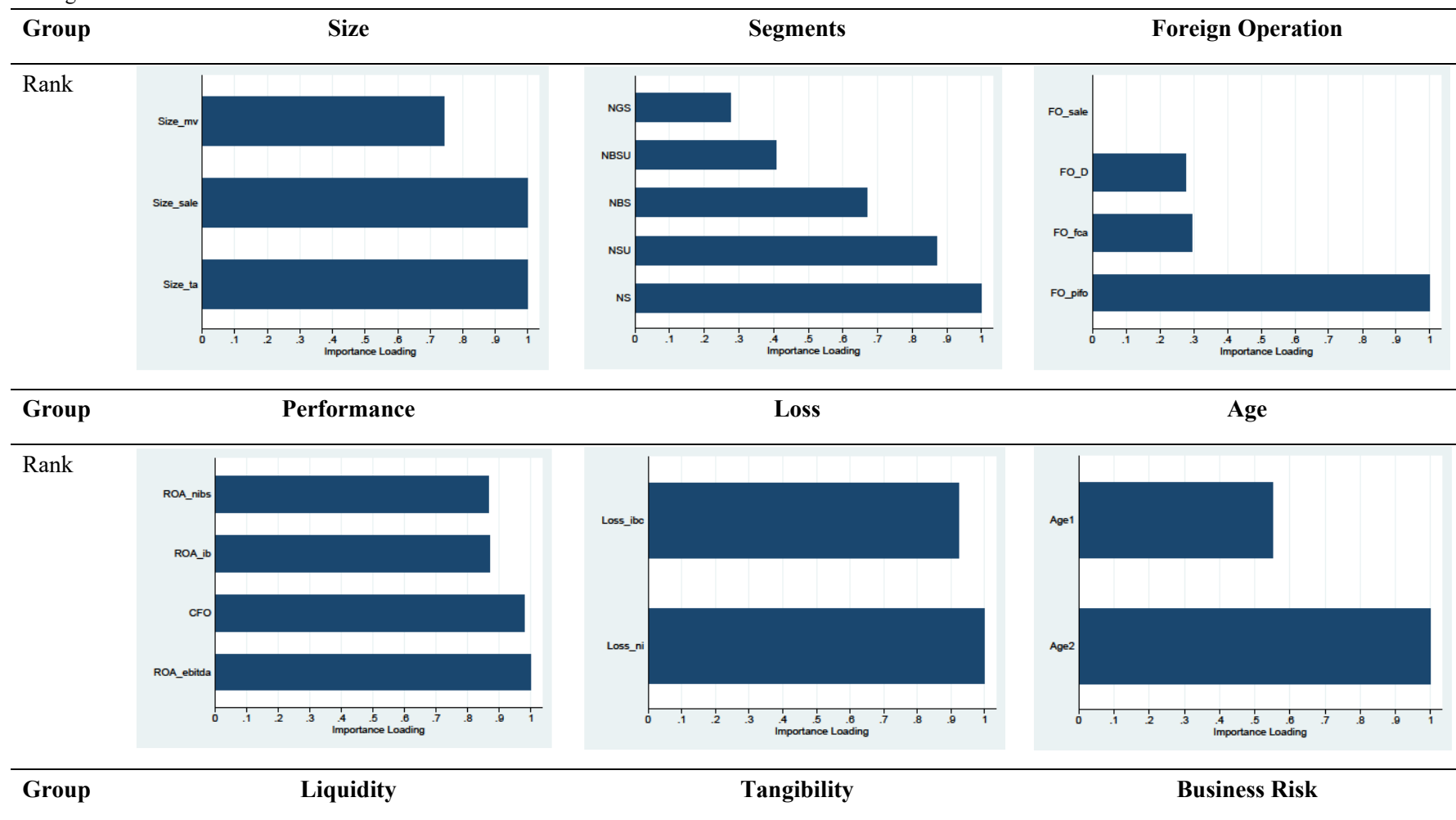
Appendix 3H.3 Random Forest analysis in Step 1 (further analysis)

This figure plots the results of Random Forest analysis in Step 1, including corporate governance attributes. The variables are ranked according to their importance loading.

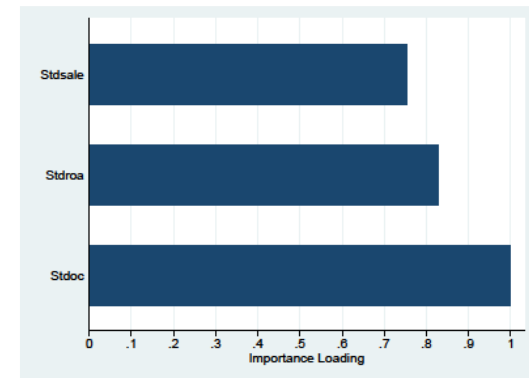
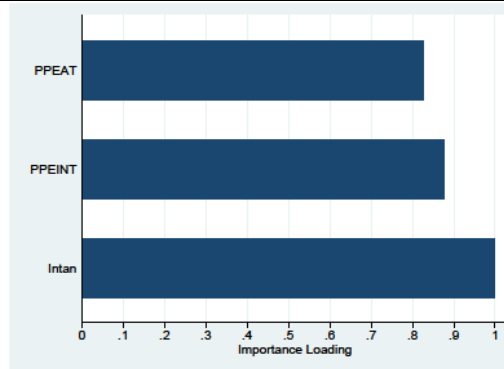
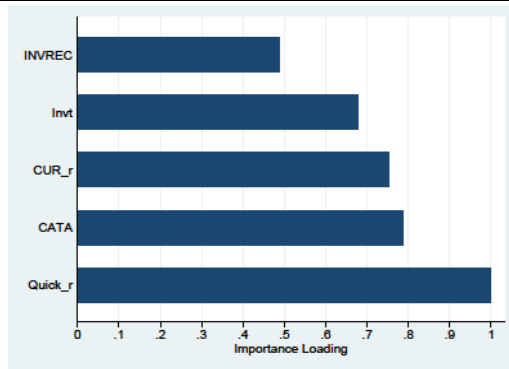


Appendix 3H.4 Random Forest analysis within each group in Step 2 (further analysis)

This table reports the results of Random Forest analysis within group in Step 2, including corporate governance attributes. Variables are ranked according to their importance loading.



Rank



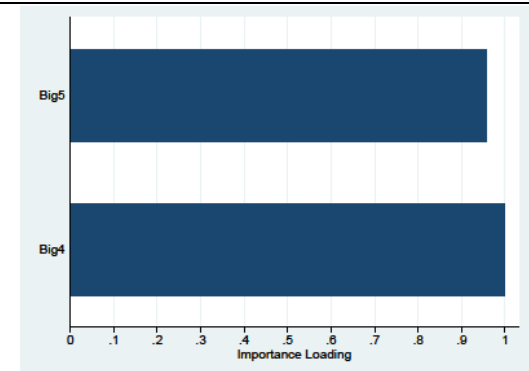
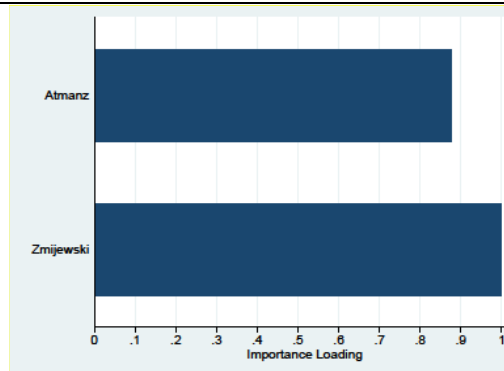
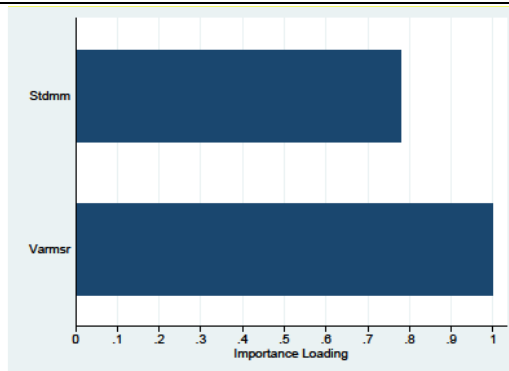
Group

Variance

Bankruptcy

Big 4

Rank



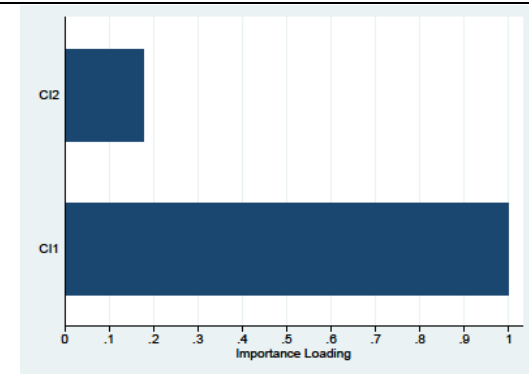
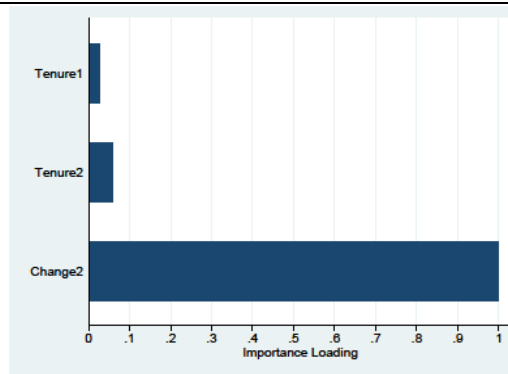
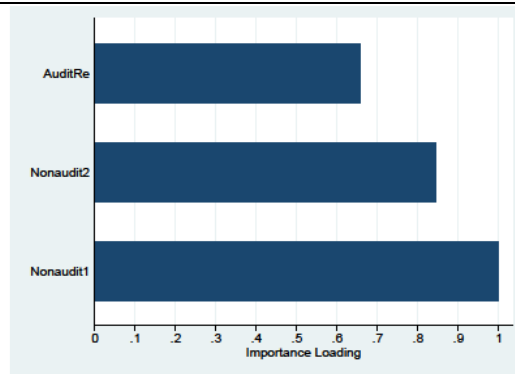
Group

Fees

Tenure

Importance

Rank



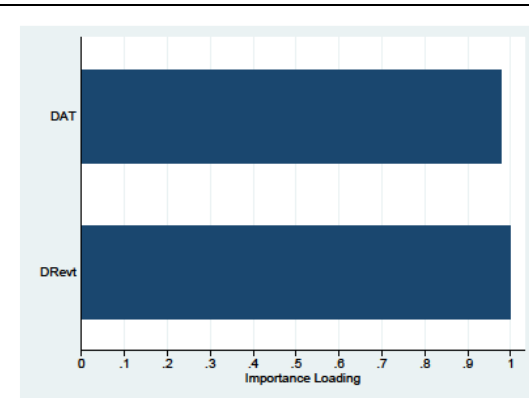
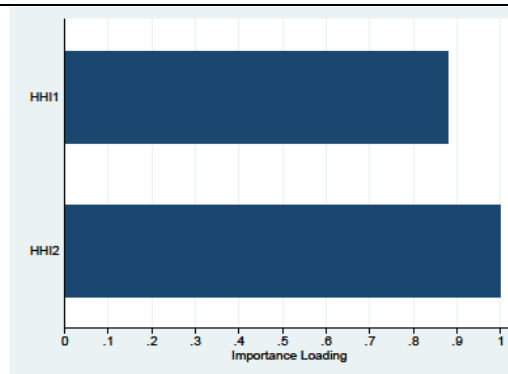
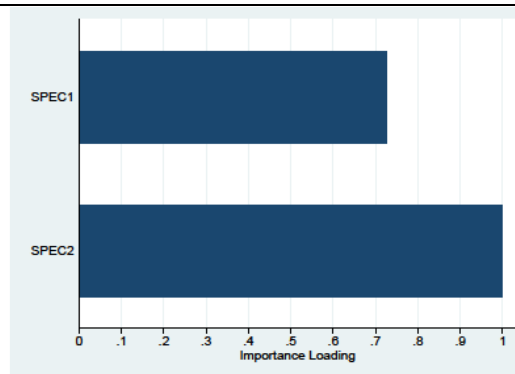
Group

Specialist

Competition

Growth

Rank



Appendix 3H.5 Adaptive LASSO analysis in Step 3 (further analysis)

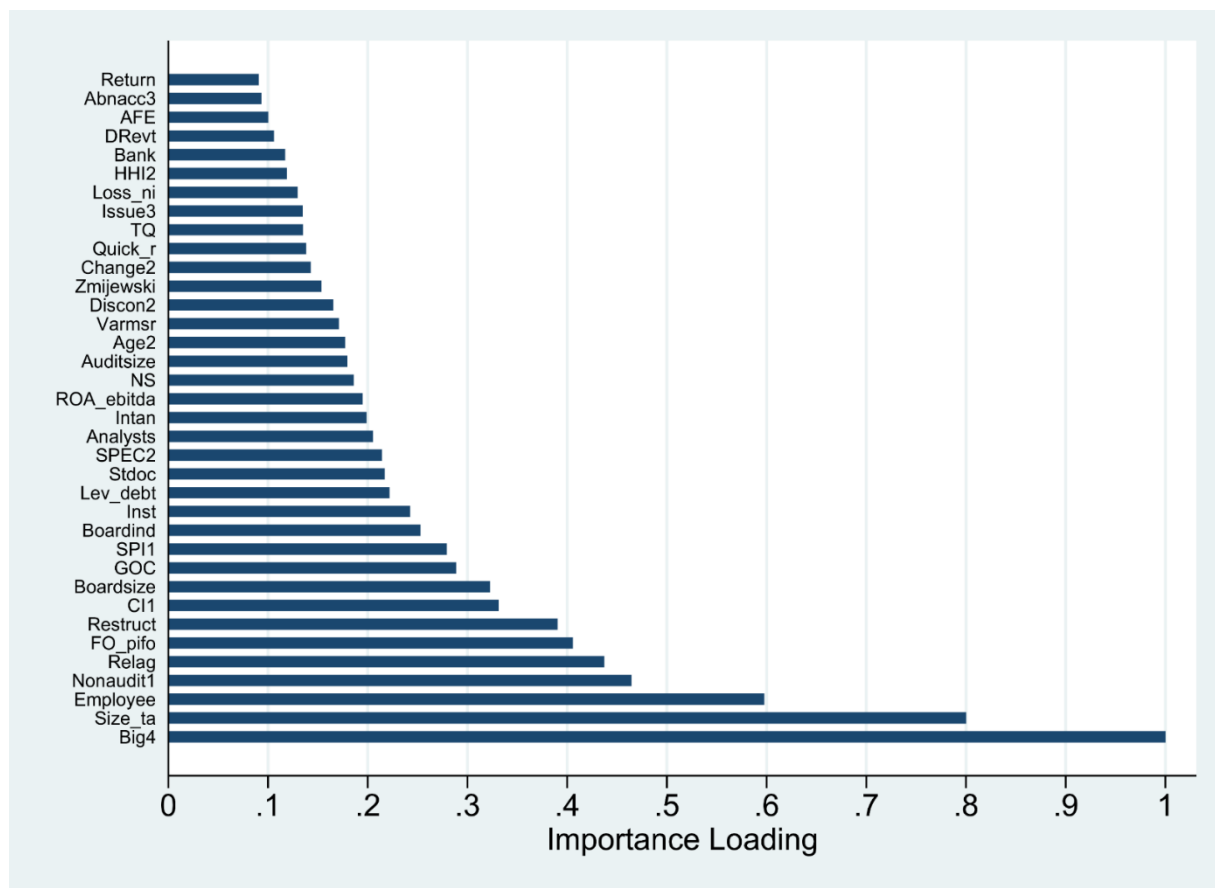
This table reports the results of Adaptive LASSO analysis in Step 3, including corporate governance attributes. All variables are defined in Appendix 3A. Out# column shows the frequency of variable coefficient shrinking to zero in Adaptive LASSO analysis.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Out #
Size_ta	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
NS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
FO_pifo	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Lev_debt	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
ROA_ebitda	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No	6
Loss_ni	No	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	6
Age2	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
TQ	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Issue3	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No	6
Bank	No	No	No	No	Yes	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	10
Quick_r	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	Yes	Yes	Yes	8
Intan	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
DRevt	No	Yes	No	No	Yes	No	No	No	No	Yes	Yes	No	No	Yes	No	No	No	No	Yes	13
Abnacc3	No	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes	No	Yes	Yes	No	No	14
Restruct	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Discon2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
SPI1	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Inst	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	5
Return	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	No	Yes	Yes	No	No	No	Yes	8
Stddev	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0

Varmstr	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	3
Analysts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Employee	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Zmijewski	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
Big4	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Nonaud~1	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
GOC	No	No	No	Yes	No	No	Yes	No	No	Yes	Yes	No	Yes	Yes	No	No	Yes	No	No	12
Change2	No	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	5
CI1	No	No	Yes	No	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	12
Relag	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0
SPEC2	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	7
HHI2	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Boardind	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	4
Boardsize	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
Auditsize	No	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	No	No	No	Yes	9
AFE	No	No	No	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No	14

Appendix 3H.6 Random Forest analysis in Step 3 (further analysis)

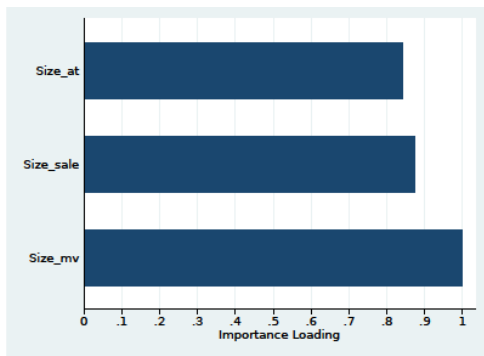
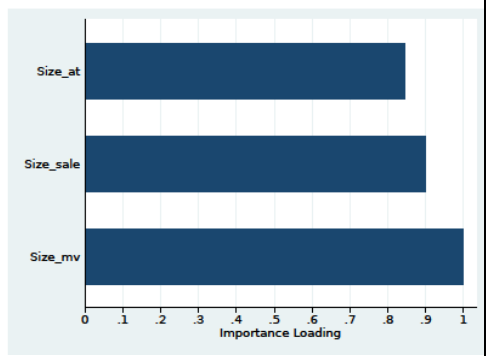
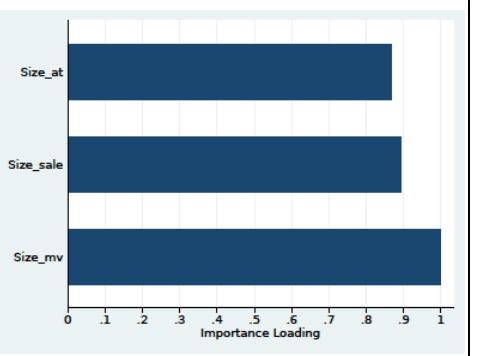
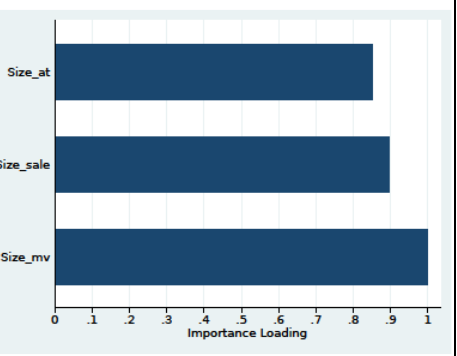
This figure plots the results of Random Forest analysis in Step 3, including corporate governance attributes. The variables are ranked according to their importance loading.

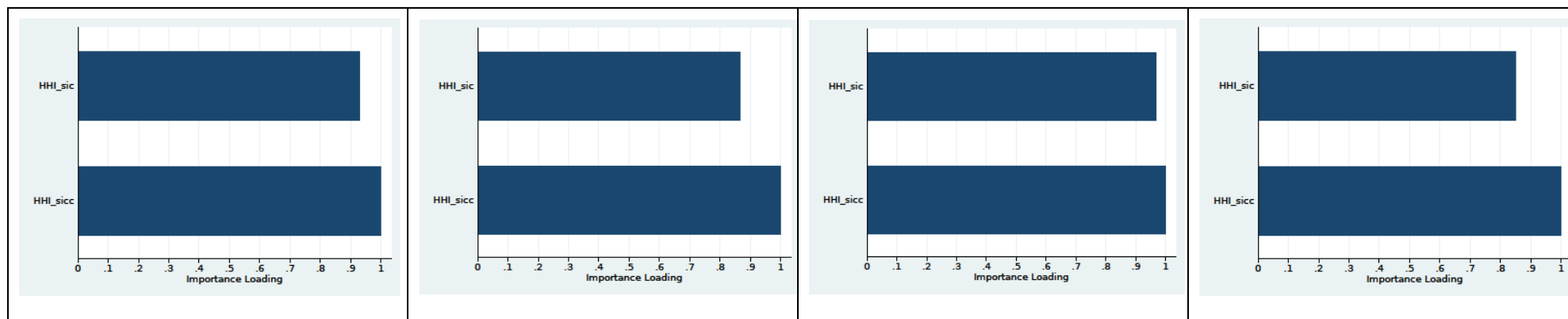


Appendices for Chapter 4

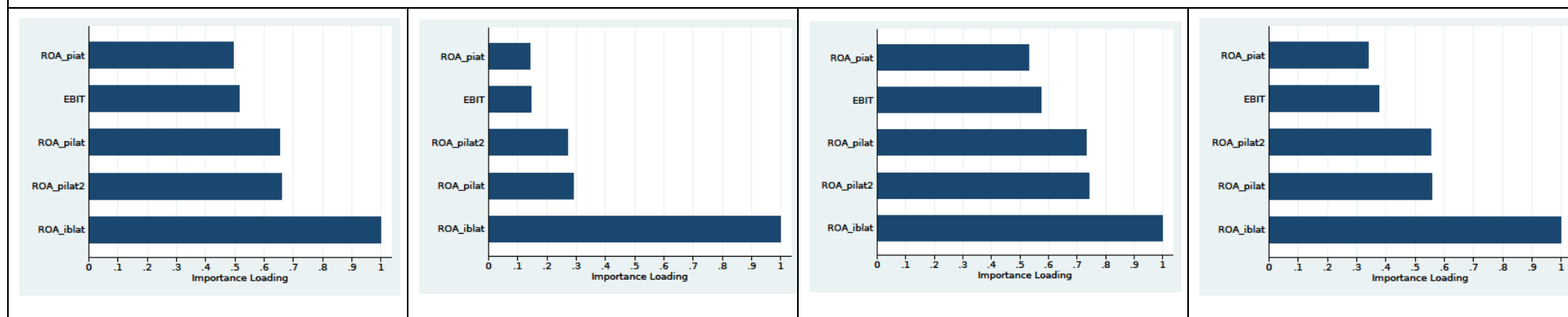
Appendix 4B Random Forest analysis within each group for ETR measures

This table reports the results of Random Forest analysis within each group for annual cash ETR, annual GAAP ETR, long-run cash ETR and long-run GAAP ETR in Step 1. Variables are ranked based on their importance loading. Appendix 4A provides variable definitions.

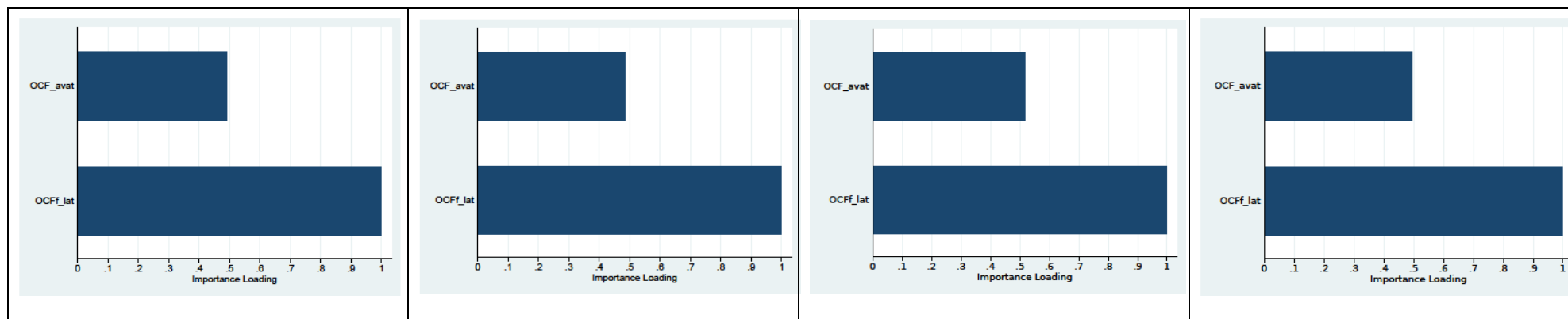
Cash ETR1	GAAP ETR1	Cash ETR3	GAAP ETR3																																
Size																																			
 <table><caption>Importance Loading for Size (Cash ETR1)</caption><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>Size_at</td><td>0.85</td></tr><tr><td>Size_sale</td><td>0.88</td></tr><tr><td>Size_mv</td><td>0.95</td></tr></tbody></table>	Variable	Importance Loading	Size_at	0.85	Size_sale	0.88	Size_mv	0.95	 <table><caption>Importance Loading for Size (GAAP ETR1)</caption><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>Size_at</td><td>0.85</td></tr><tr><td>Size_sale</td><td>0.92</td></tr><tr><td>Size_mv</td><td>0.98</td></tr></tbody></table>	Variable	Importance Loading	Size_at	0.85	Size_sale	0.92	Size_mv	0.98	 <table><caption>Importance Loading for Size (Cash ETR3)</caption><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>Size_at</td><td>0.88</td></tr><tr><td>Size_sale</td><td>0.92</td></tr><tr><td>Size_mv</td><td>0.98</td></tr></tbody></table>	Variable	Importance Loading	Size_at	0.88	Size_sale	0.92	Size_mv	0.98	 <table><caption>Importance Loading for Size (GAAP ETR3)</caption><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>Size_at</td><td>0.88</td></tr><tr><td>Size_sale</td><td>0.92</td></tr><tr><td>Size_mv</td><td>0.98</td></tr></tbody></table>	Variable	Importance Loading	Size_at	0.88	Size_sale	0.92	Size_mv	0.98
Variable	Importance Loading																																		
Size_at	0.85																																		
Size_sale	0.88																																		
Size_mv	0.95																																		
Variable	Importance Loading																																		
Size_at	0.85																																		
Size_sale	0.92																																		
Size_mv	0.98																																		
Variable	Importance Loading																																		
Size_at	0.88																																		
Size_sale	0.92																																		
Size_mv	0.98																																		
Variable	Importance Loading																																		
Size_at	0.88																																		
Size_sale	0.92																																		
Size_mv	0.98																																		
Competition																																			



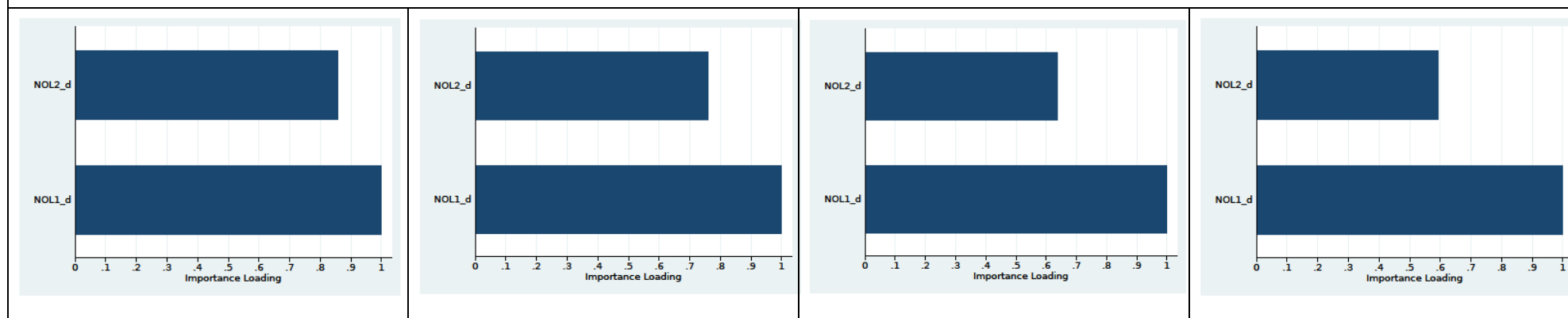
Profitability



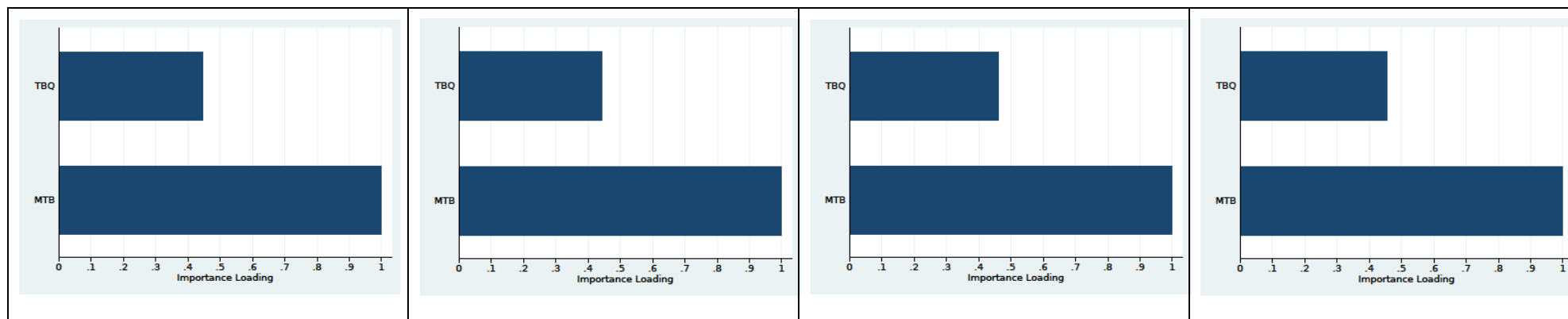
Operating Cash Flow



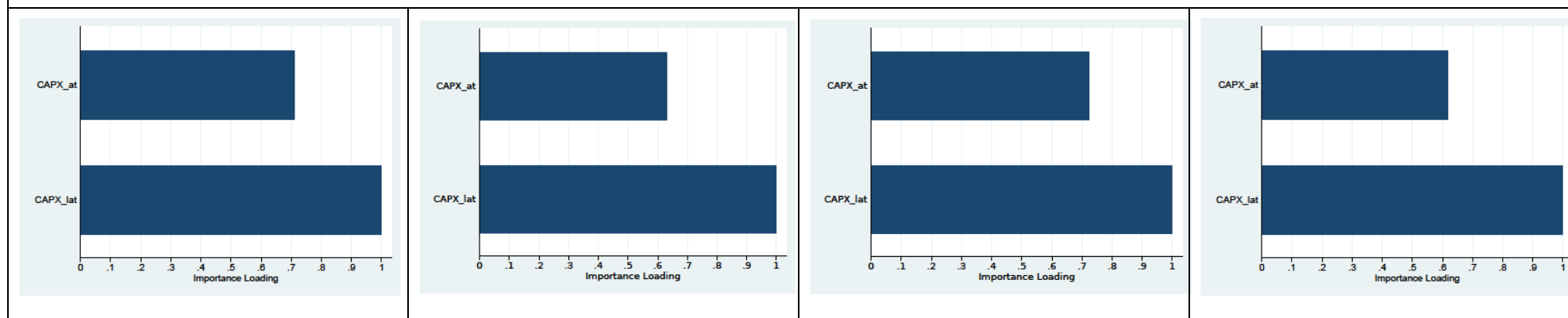
NOL



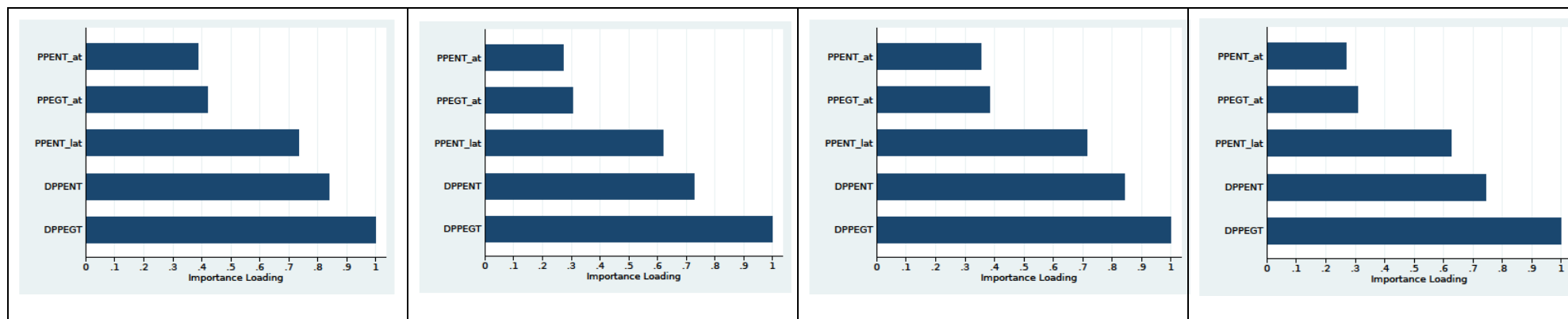
Valuation



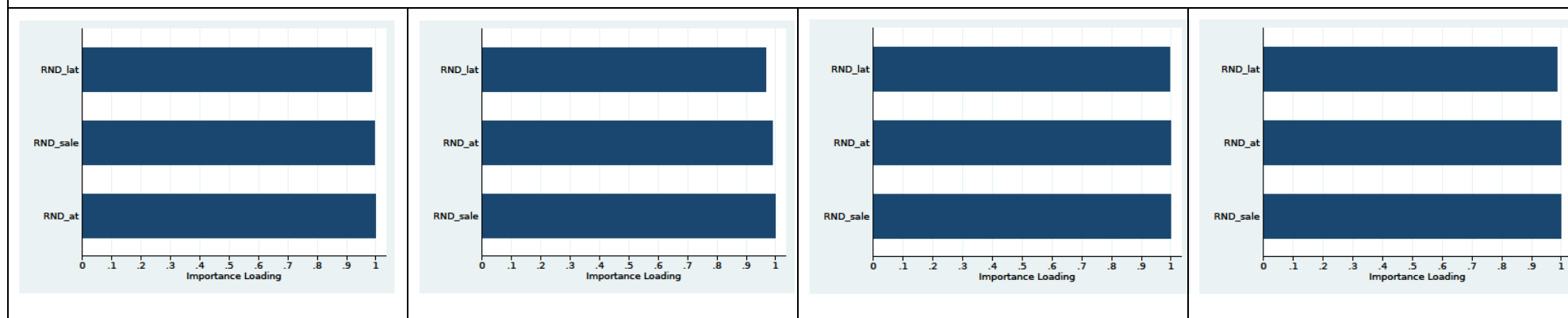
Investment



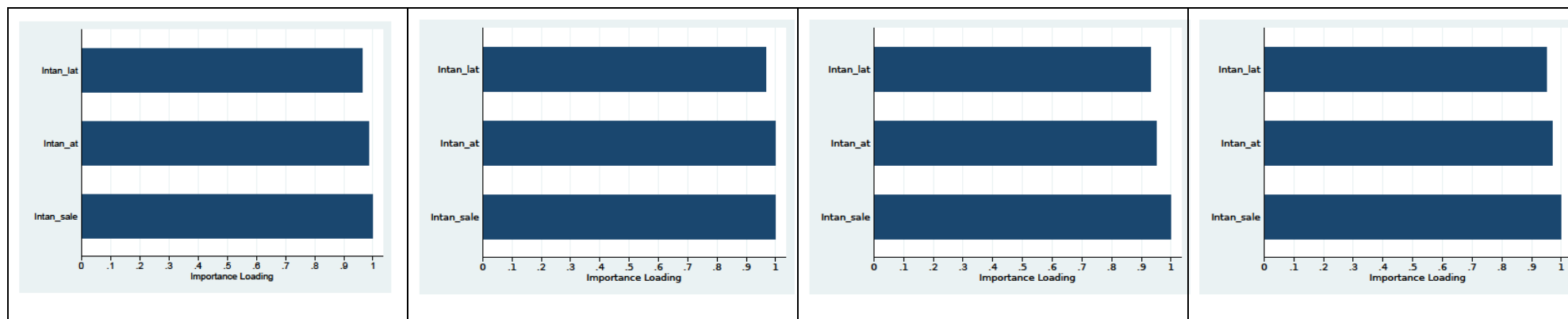
PPE



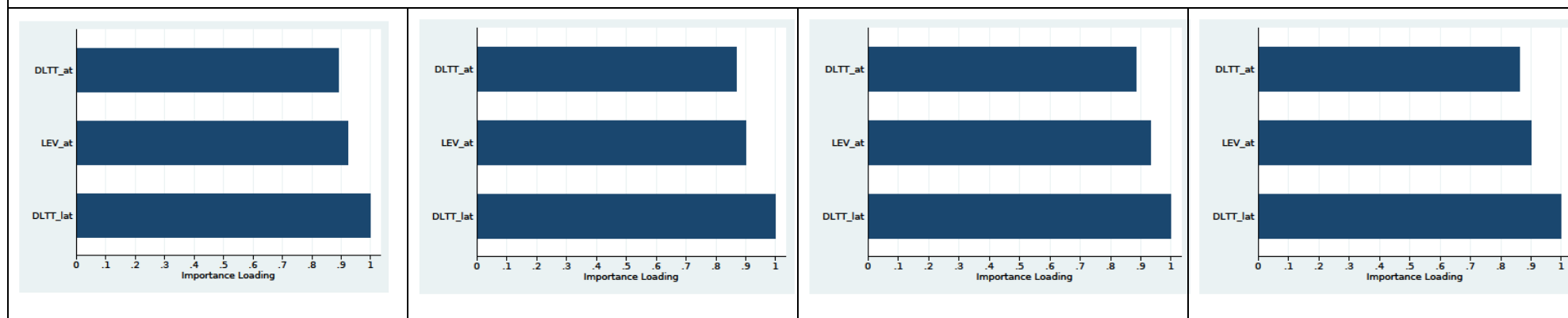
RND



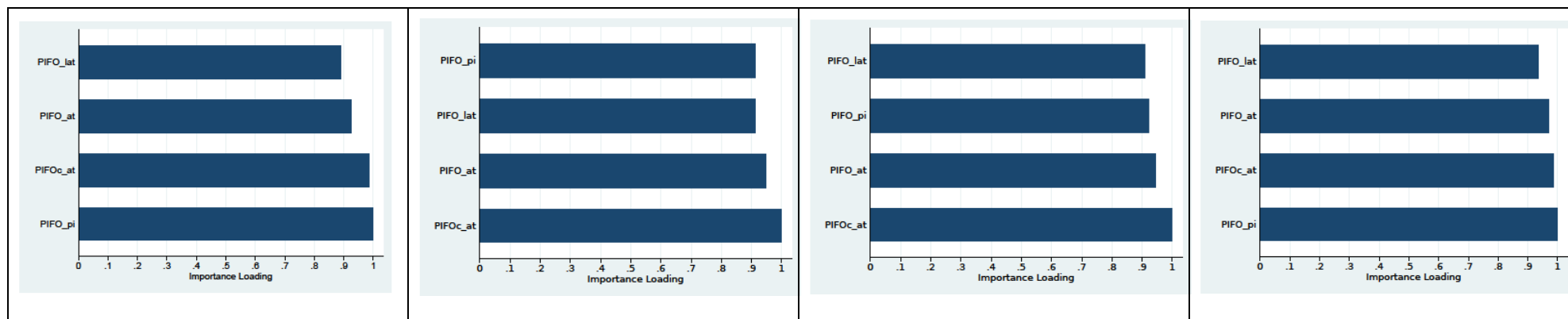
Intangible Assets



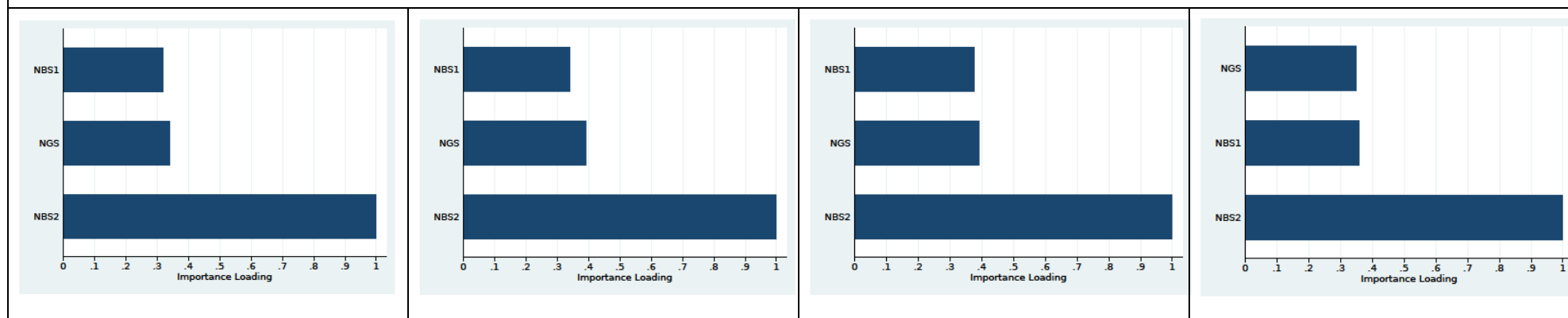
Leverage



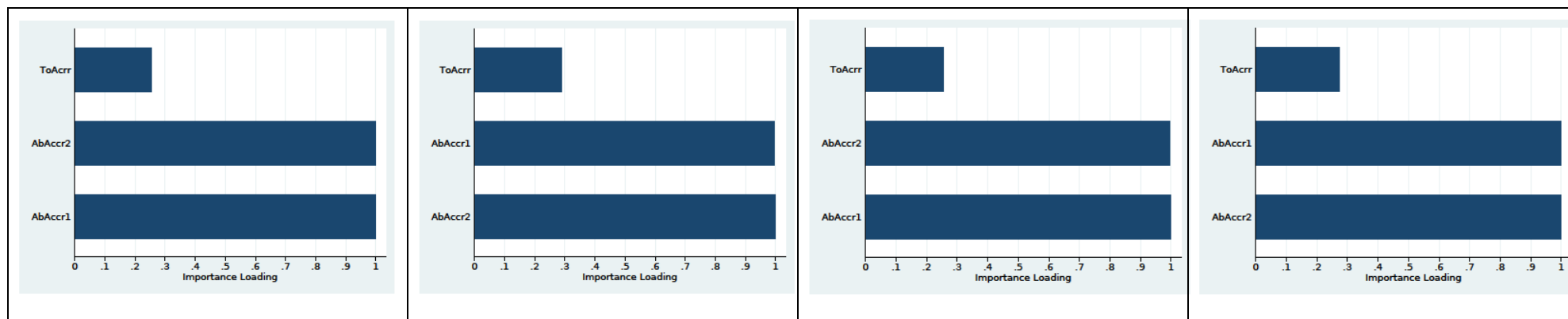
Foreign Operation



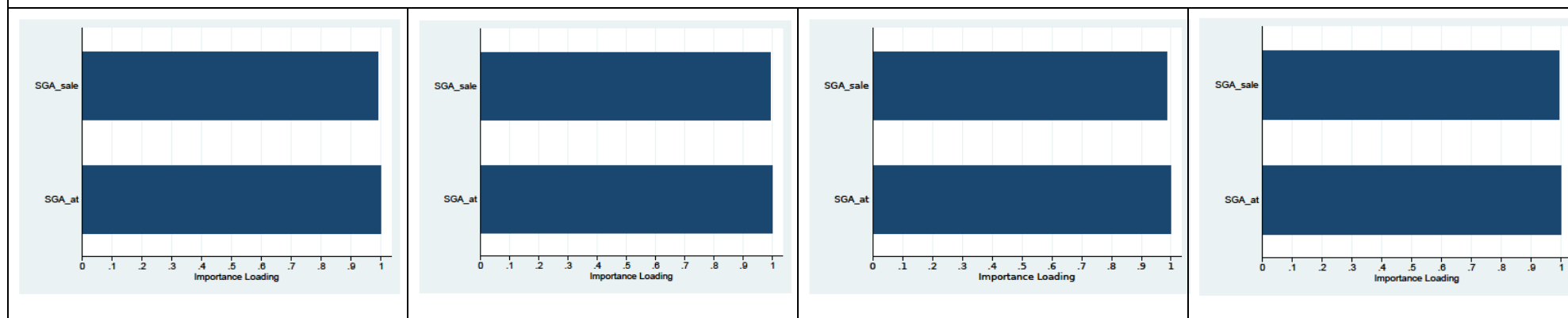
Segments



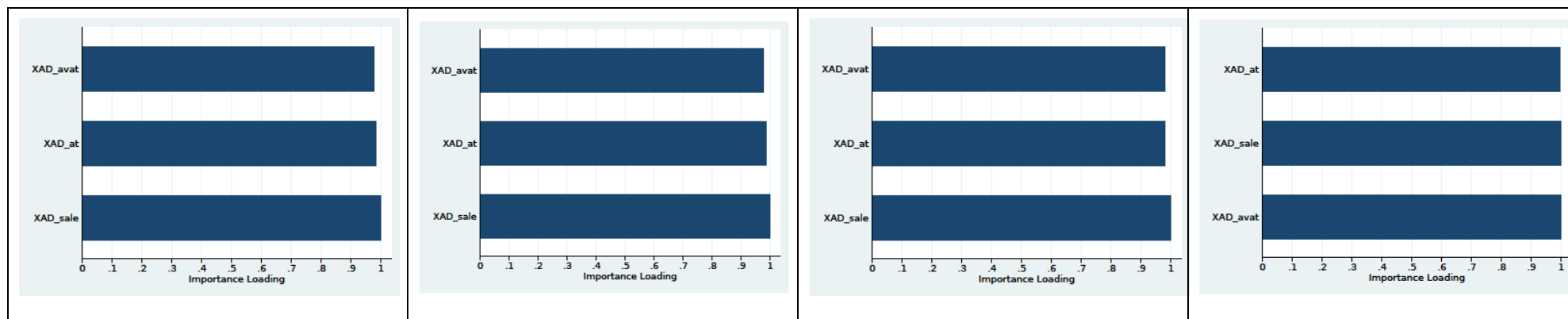
Accruals



SGNA



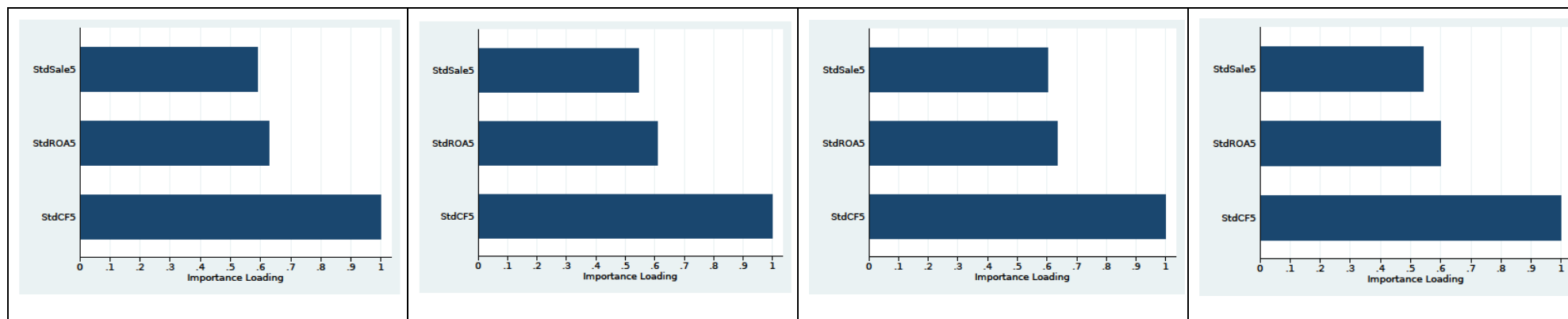
Advertisement Expense



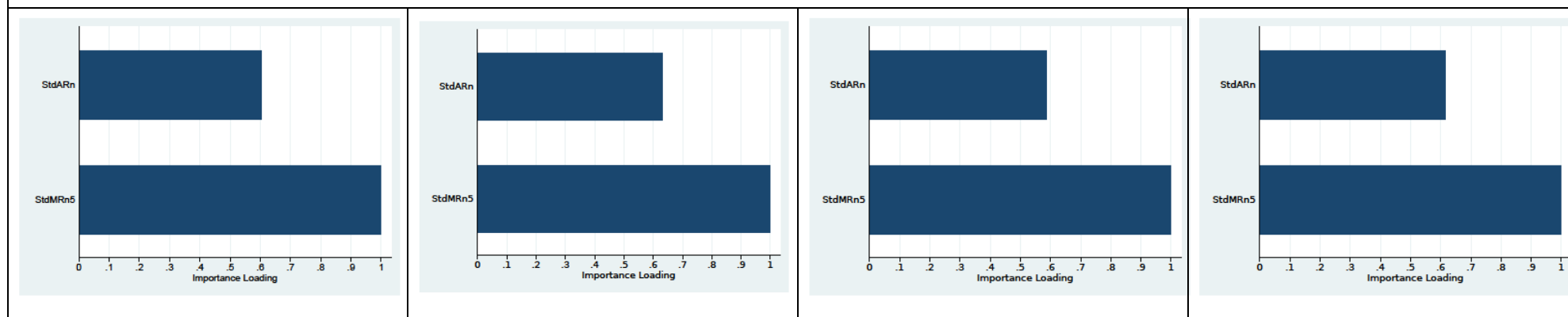
Special Items



Operating Risk



Return Volatility



Appendix 4C

Appendix 4C.1 Adaptive LASSO analysis for annual cash ETR

This table reports the results of Adaptive LASSO analysis for annual cash ETR in Step 2. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv	1	1	1	1		1			1	1	1	1	1	1	1	1				1			1			1	1	
EMP	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1			1			1	1	1	1	1	
HHI_sicc			1	1	1	1								1	1	1			1	1			1	1	1	1	1	
ROA_iblat	1	1	1		1	1				1	1	1	1	1	1	1	1		1	1			1			1		
OCFf_lat	1		1	1	1	1	1		1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	
DNOL	1	1	1	1	1	1	1					1		1	1	1	1	1	1	1			1			1		
NOL1_d	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Return	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
MTB	1		1		1	1							1	1	1	1	1	1				1	1			1	1	
DSale	1		1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
MA_d	1		1	1	1	1		1		1	1	1	1	1	1	1	1	1		1		1	1	1	1	1		
CAPX_lat			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
DPPEGT	1	1	1			1		1			1	1		1	1		1			1	1	1	1			1	1	
Inv_t_lat			1	1	1	1					1	1	1	1	1	1	1	1				1	1			1		
DP_at	1	1	1	1		1	1				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
DGW_at			1			1	1	1	1	1	1	1	1	1	1								1			1	1	
RND_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Intan_sale	1	1	1			1				1	1		1	1	1	1	1	1		1			1	1	1	1		
DLTT_lat	1	1	1	1	1	1		1	1	1	1	1	1	1	1				1		1		1	1	1	1	1	
INT_lat	1	1	1			1	1	1	1		1	1	1	1	1	1	1		1	1	1	1	1	1	1	1		
Mezzanine			1	1	1	1	1				1	1	1	1	1		1	1				1	1	1	1	1		
Fin_d				1	1	1								1	1	1							1	1	1	1	1	
IO_ts						1	1		1	1	1	1	1	1	1								1	1	1	1	1	
CHE_at	1		1	1	1	1					1	1	1	1	1	1				1			1			1		
SA_HP2010			1		1	1	1	1	1				1	1	1		1	1	1	1	1	1		1			1	

AltmanZ	1		1	1							1	1	1	1	1	1	1	1	1	1	1	1		1	1		
PIFO_pi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PIFO_d			1	1	1	1					1	1	1	1				1	1	1	1			1	1		
NBS2			1	1	1	1					1	1	1	1				1	1		1				1	1	
ESUB_lat	1	1	1	1	1	1					1	1	1	1	1					1	1	1			1		
ESUB_d	1		1	1					1	1			1	1	1					1	1				1	1	
StdEarnFst	1	1	1	1		1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1
Num_Analyst			1			1	1	1	1	1		1	1	1	1	1				1					1	1	1
Goodnews_d	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Big4_d			1	1		1					1	1		1	1	1	1	1			1				1	1	
AuditOp_d	1		1	1	1	1			1	1	1	1	1	1	1					1					1	1	1
AbAccr1	1	1	1	1	1	1					1		1	1	1	1		1	1			1			1		
Age			1	1							1	1	1	1	1				1		1	1	1		1		
SGA_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XAD_sale		1	1			1					1	1	1	1	1	1	1			1	1				1	1	1
SPI_sale	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			1	1	1	1	1	1	1
EI_at	1	1	1			1	1	1			1	1	1	1	1	1	1				1						
StdCF5	1	1	1								1		1	1	1	1	1		1		1	1				1	
StdMRn5	1	1	1	1	1	1	1				1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1

Appendix 4C.2 Adaptive LASSO analysis for annual GAAP ETR

This table reports the results of Adaptive LASSO analysis for annual GAAP ETR in Step 2. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv			1				1	1	1			1		1	1		1	1	1	1		1		1				
EMP		1		1	1	1	1	1	1	1	1						1	1	1	1	1	1						
HHI_sicc	1	1	1	1	1	1	1	1	1								1	1	1									
ROA_iblat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OCFf_lat	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DNOL	1	1	1	1	1	1	1	1					1		1					1	1							1
NOL1_d	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Return	1	1	1	1	1		1		1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1		1	
MTB	1	1	1	1	1	1						1	1	1	1		1					1	1	1	1	1	1	1
DSale	1	1	1	1	1		1	1	1		1	1	1	1	1		1											
MA_d			1	1	1	1	1	1	1	1	1	1	1		1	1	1	1										
CAPX_lat		1	1	1		1	1	1				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DPPEGT	1		1		1	1	1		1					1	1		1	1		1	1	1		1			1	
Inv_t_lat	1					1	1					1	1	1	1	1	1			1	1			1		1	1	1
DP_at							1		1	1	1		1	1	1	1	1			1		1		1		1	1	1
DGW_at	1	1					1	1	1	1	1	1		1	1	1	1	1	1									
RND_sale	1		1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1				1				1	1
Intan_sale		1		1			1	1	1		1			1	1	1	1	1		1	1	1	1	1	1			
DLTT_lat	1			1	1	1	1	1	1			1	1		1							1					1	1
INT_lat	1	1	1		1	1	1								1								1	1	1	1	1	
Mezzanine		1	1	1			1							1	1	1	1			1								
Fin_d				1			1							1	1		1	1	1		1	1					1	1
IO_ts	1	1	1	1	1	1	1	1	1						1		1	1									1	1
CHE_at	1				1	1	1					1	1	1	1	1	1	1	1	1	1	1	1			1	1	1
SA_HP2010	1	1	1			1	1		1		1	1	1	1	1	1	1	1				1						
AltmanZ			1	1	1	1	1		1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
PIFOc_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1

PIFO_d	1	1	1	1	1	1					1	1		1	1	1	1	1	1					1			
NBS2	1			1				1	1	1		1	1	1			1			1					1	1	
ESUB_lat	1	1	1	1	1	1	1		1		1			1			1	1	1	1	1	1					
ESUB_d	1				1	1		1	1	1						1							1		1	1	1
StdEarnFst	1	1	1	1	1	1	1		1		1	1	1	1		1	1	1	1	1	1	1		1		1	
Num_Analyst	1			1				1		1	1							1	1	1	1	1	1		1	1	1
Goodnews_d					1			1						1		1	1	1	1	1	1	1	1	1	1	1	1
Big4_d	1							1						1						1	1	1					
AuditOp_d	1	1	1	1	1	1	1	1			1	1	1	1	1					1		1			1	1	
AbAccr2			1	1	1	1	1	1	1		1			1	1	1						1	1	1	1		
Age	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1
SGA_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1
XAD_sale	1	1	1				1	1			1	1	1	1			1	1			1	1	1	1	1		
SPL_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EI_at	1	1	1	1				1			1				1												
StdCF5	1	1	1	1				1	1	1		1	1	1	1	1	1	1					1	1	1	1	1
StdMRn5	1	1	1	1	1	1	1	1				1	1	1	1	1			1	1	1	1	1	1	1	1	1

Appendix 4C.3 Adaptive LASSO analysis for long-run cash ETR

This table reports the results of Adaptive LASSO analysis for long-run cash ETR in Step 2. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Size_mv	1	1	1	1	1	1				1	1		1					1	1					1	1	1
EMP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1	1	1
HHI_sicc			1	1	1	1				1	1	1	1	1				1	1				1	1	1	
ROA_iblat		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1		1	1	1	1	1	
OCff_lat	1	1				1	1			1	1	1	1	1	1	1			1			1	1	1		
DNOL			1	1	1			1			1								1						1	1
NOL1_d	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Return		1		1	1	1	1		1	1	1	1	1	1	1	1		1	1	1	1	1	1		1	1
MTB		1	1	1			1	1	1	1	1						1	1	1					1	1	
DSale	1	1		1						1	1					1	1		1							1
MA_d					1	1	1	1	1	1	1	1	1						1				1	1	1	
CAPX_lat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DPPEGT							1	1		1	1								1	1	1	1		1	1	
Inv_t_lat		1	1	1	1		1		1	1	1	1	1	1	1	1	1	1	1	1					1	
DP_at	1	1		1	1					1	1		1	1	1	1	1		1		1	1	1	1	1	1
DGW_at				1		1	1	1		1	1	1			1	1			1				1		1	
RND_sale	1			1	1	1	1	1	1	1	1			1	1	1	1	1	1		1	1	1			1
Intan_sale		1	1	1	1	1	1	1		1	1	1	1	1	1	1			1	1			1	1	1	
DLTT_lat	1	1	1	1		1			1		1		1	1	1	1	1	1	1				1	1		
INT_lat								1	1	1	1	1	1	1			1		1	1	1	1	1	1	1	1
Mezzanine		1	1	1						1	1	1	1	1	1	1			1	1	1	1	1	1	1	
Fin_d	1	1		1	1					1	1								1	1			1	1	1	
IO_ts				1	1		1	1	1	1	1	1						1	1				1	1	1	
CHE_at		1		1	1					1	1	1	1				1	1	1	1				1	1	
SA_HP2010	1	1		1	1		1	1	1	1	1		1						1	1						
AltmanZ		1		1	1			1	1	1	1			1	1	1		1	1	1		1	1			
PIFOc_at	1	1	1	1	1	1	1	1		1			1	1	1	1	1	1	1	1	1	1	1	1	1	

PIFO_d	1	1	1	1			1	1		1	1	1				1		1	1	1	1	1	1	1	1
NBS2				1	1	1	1	1	1	1	1	1	1	1				1						1	
ESUB_lat		1	1	1							1	1						1							
ESUB_d								1	1	1	1		1	1		1								1	
StdEarnFst		1	1	1				1	1	1	1	1	1	1	1	1	1	1							
Num_Analyst		1	1				1	1	1	1	1	1	1					1						1	1
Goodnews_d							1	1	1	1	1			1	1			1	1						1
Big4_d		1		1	1	1	1	1		1	1							1				1			
AuditOp_d	1	1	1	1	1	1				1	1	1	1	1				1						1	1
AbAccr1	1	1	1		1			1	1	1	1		1	1	1	1		1	1			1		1	
Age		1	1								1		1								1				
SGA_at	1			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XAD_sale	1	1	1		1	1				1	1	1	1					1	1				1	1	1
SPI_sale	1	1	1	1	1	1	1			1	1	1	1	1	1	1		1				1	1	1	
EI_at		1	1	1			1	1	1	1	1		1			1									
StdCF5		1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1							
StdMRn5	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1		1	1	1	1	1	1	1	1

Appendix 4C.4 Adaptive LASSO analysis for long-run GAAP ETR

This table reports the results of Adaptive LASSO analysis for long-run GAAP ETR in Step 2. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

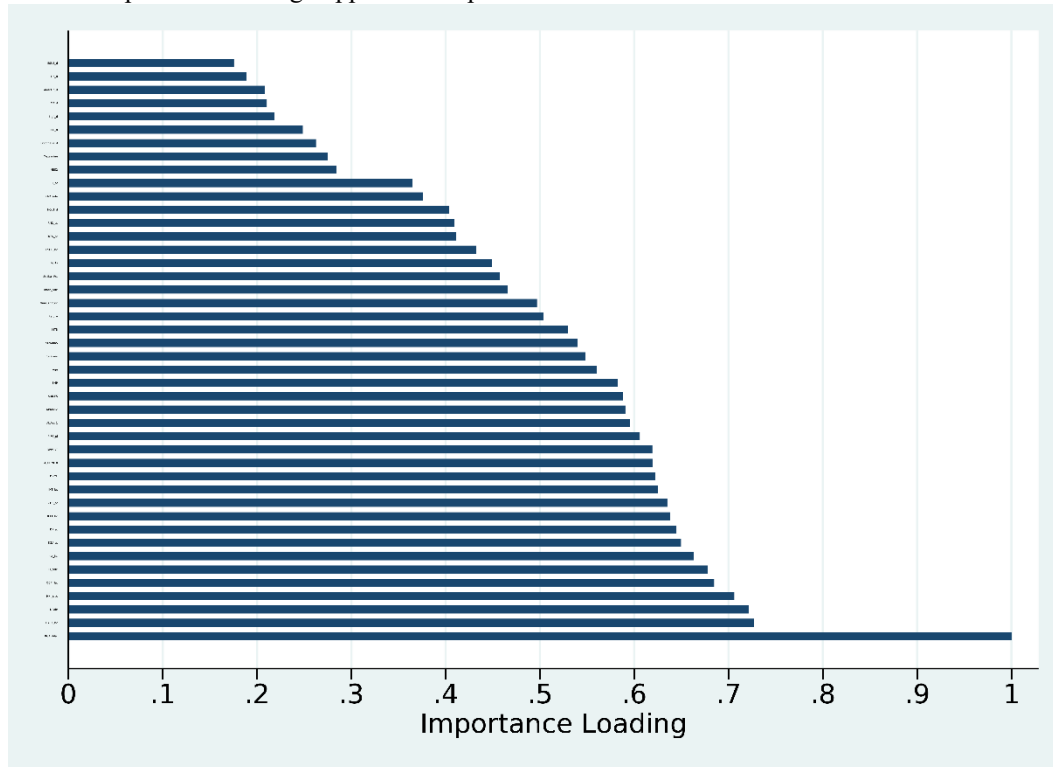
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Size_mv	1	1	1	1	1	1	1	1	1					1	1	1	1	1	1	1					1	1
EMP	1			1						1	1			1	1		1	1	1	1						1
HHI_sicc				1	1	1	1							1			1	1			1			1		
ROA_iblat	1			1				1	1			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OCff_lat	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1
DNOL				1	1	1	1						1												1	1
NOL1_d	1		1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1				
Return							1																		1	1
MTB					1		1	1	1						1						1	1				1
DSale	1					1	1	1	1				1	1			1			1			1	1	1	1
MA_d				1	1	1											1	1	1	1	1					
CAPX_lat									1	1	1	1	1	1	1	1	1		1	1		1	1			1
DPPEGT				1			1								1			1		1	1	1				
Inv_t_lat				1	1	1						1	1	1	1							1			1	1
DP_at															1			1						1	1	1
DGW_at				1	1	1	1				1	1			1											
RND_sale	1	1		1	1	1	1	1	1	1		1			1			1	1	1	1	1	1			
Intan_sale			1								1	1	1	1	1	1	1	1	1		1	1	1	1	1	1
DLTT_lat	1			1		1	1	1	1						1							1		1	1	1
INT_lat				1											1			1	1	1	1	1	1			
Mezzanine	1	1		1	1	1								1	1	1			1							
Fin_d															1			1	1							1
IO_ts	1														1										1	1
CHE_at			1	1	1				1					1	1	1	1				1	1	1			1
SA_HP2010				1	1	1		1						1	1							1				
AltmanZ					1	1	1					1	1	1	1	1	1	1	1	1	1	1	1	1		1
PIFO_pi				1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1		1

[illegible]

Appendix 4D

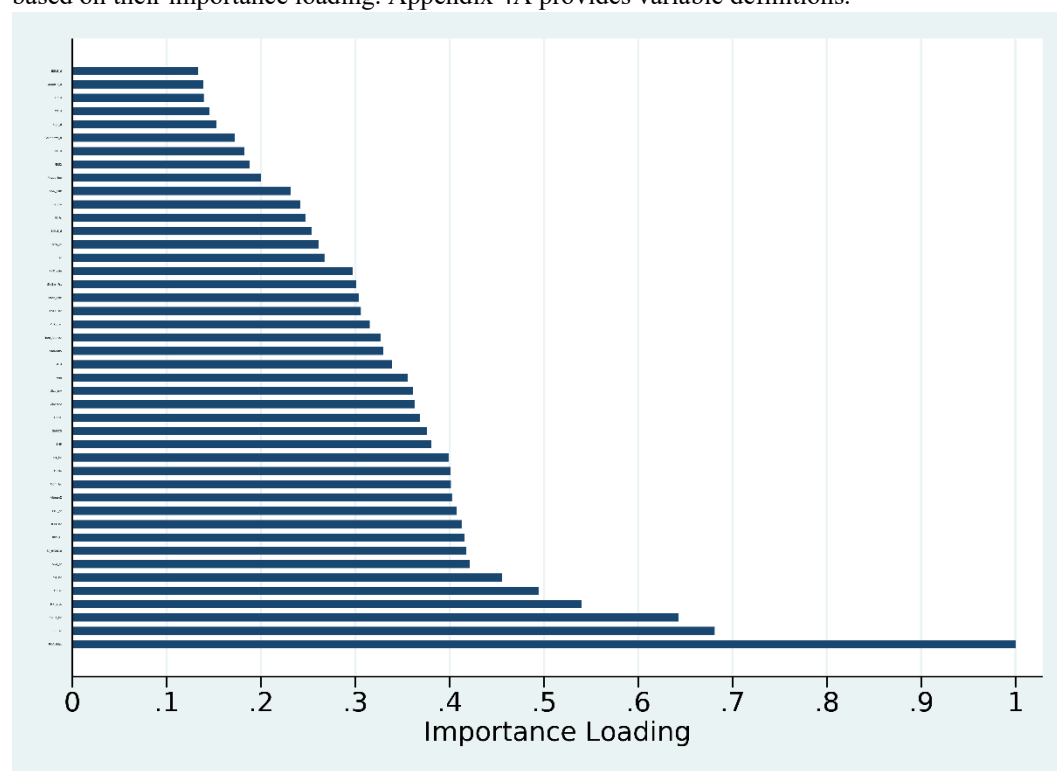
Appendix 4D.1 Random Forest analysis for annual cash ETR

This figure plots the results Random Forest analysis for annual cash ETR in Step 2. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



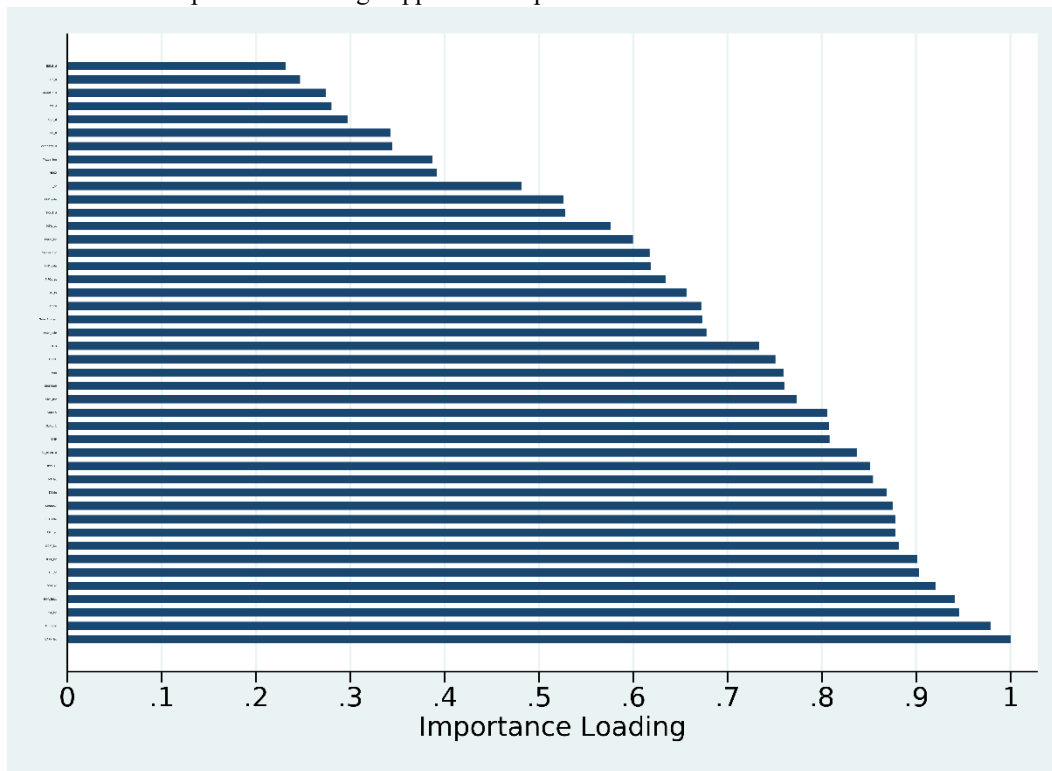
Appendix 4D.2 Random Forest analysis for annual GAAP ETR

This figure plots the results Random Forest analysis for annual GAAP ETR in Step 2. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



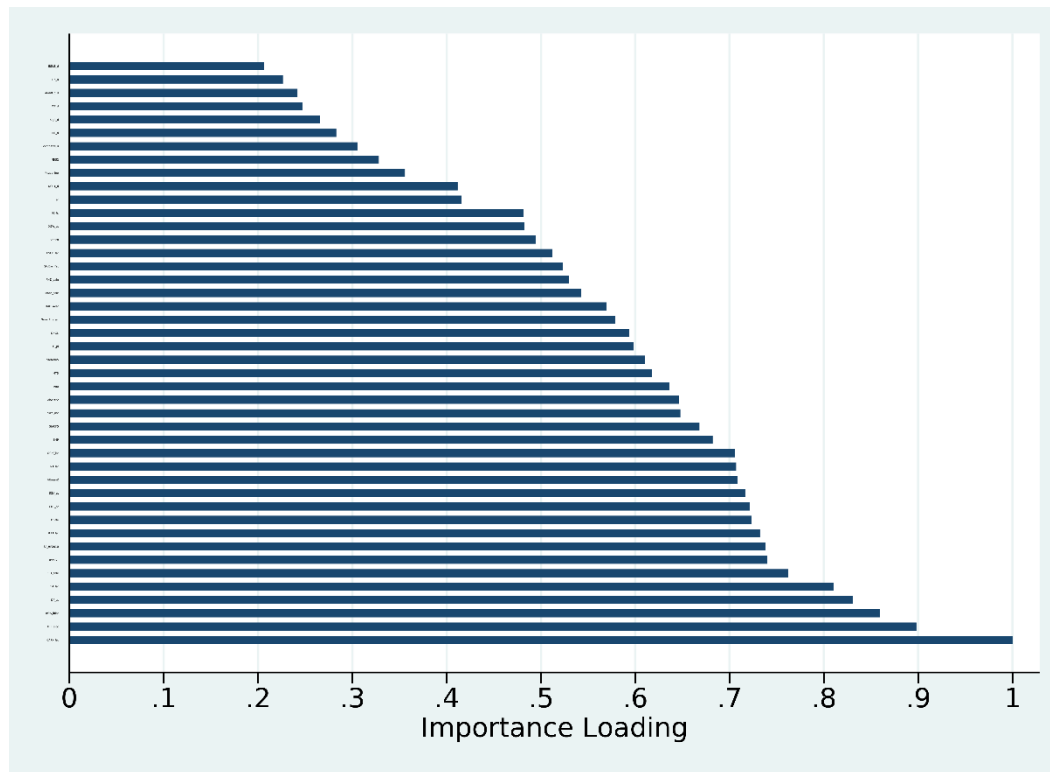
Appendix 4D.3 Random Forest analysis for long-run cash ETR

This figure plots the results Random Forest analysis for long-run cash ETR in Step 2. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



Appendix 4D.4 Random Forest analysis for long-run GAAP ETR

This figure plots the results Random Forest analysis for long-run GAAP ETR in Step 2. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



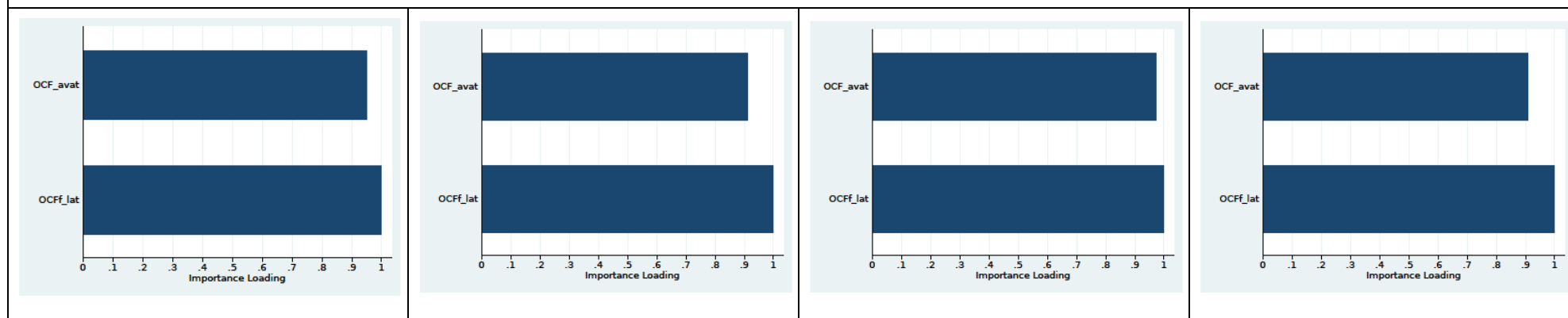
Appendix 4E RF analysis within each group for alternative tax avoidance measures

This table reports the results of Random Forest analyses for UTB, DTAX, DDBTD and MPBTD in Step 1. Variables are ranked based on their importance loading.

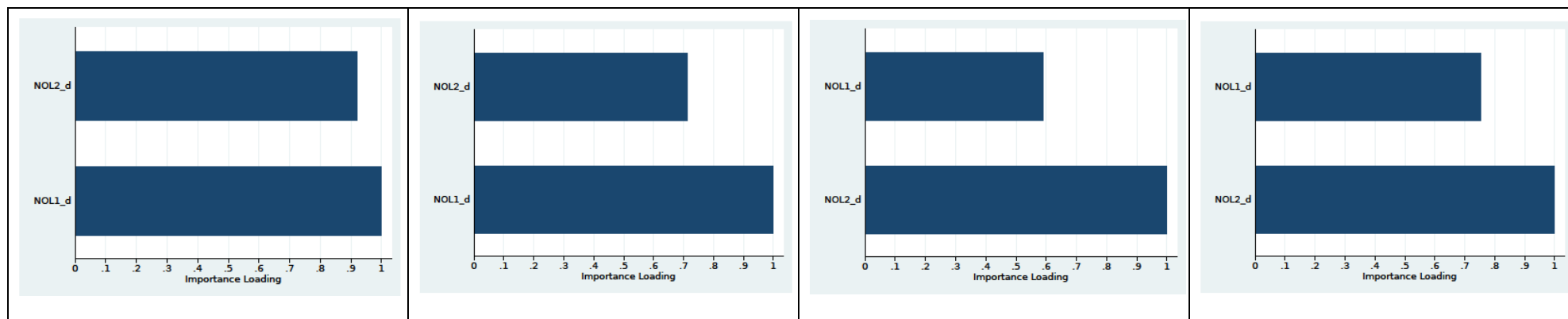
UTB		DTAX		DDBTD		MPBTD																																	
Size																																							
<table><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>Size_mv</td><td>0.95</td></tr><tr><td>Size_at</td><td>0.90</td></tr><tr><td>Size_sale</td><td>0.85</td></tr></tbody></table>		Variable	Importance Loading	Size_mv	0.95	Size_at	0.90	Size_sale	0.85	<table><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>Size_mv</td><td>0.95</td></tr><tr><td>Size_at</td><td>0.90</td></tr><tr><td>Size_sale</td><td>0.85</td></tr></tbody></table>		Variable	Importance Loading	Size_mv	0.95	Size_at	0.90	Size_sale	0.85	<table><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>Size_mv</td><td>0.95</td></tr><tr><td>Size_at</td><td>0.90</td></tr><tr><td>Size_sale</td><td>0.85</td></tr></tbody></table>		Variable	Importance Loading	Size_mv	0.95	Size_at	0.90	Size_sale	0.85	<table><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>Size_mv</td><td>0.95</td></tr><tr><td>Size_at</td><td>0.90</td></tr><tr><td>Size_sale</td><td>0.85</td></tr></tbody></table>		Variable	Importance Loading	Size_mv	0.95	Size_at	0.90	Size_sale	0.85
Variable	Importance Loading																																						
Size_mv	0.95																																						
Size_at	0.90																																						
Size_sale	0.85																																						
Variable	Importance Loading																																						
Size_mv	0.95																																						
Size_at	0.90																																						
Size_sale	0.85																																						
Variable	Importance Loading																																						
Size_mv	0.95																																						
Size_at	0.90																																						
Size_sale	0.85																																						
Variable	Importance Loading																																						
Size_mv	0.95																																						
Size_at	0.90																																						
Size_sale	0.85																																						
Competition																																							
<table><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>HHI_sicc</td><td>0.95</td></tr><tr><td>HHI_sic</td><td>0.90</td></tr></tbody></table>		Variable	Importance Loading	HHI_sicc	0.95	HHI_sic	0.90	<table><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>HHI_sicc</td><td>0.95</td></tr><tr><td>HHI_sic</td><td>0.85</td></tr></tbody></table>		Variable	Importance Loading	HHI_sicc	0.95	HHI_sic	0.85	<table><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>HHI_sicc</td><td>0.95</td></tr><tr><td>HHI_sic</td><td>0.90</td></tr></tbody></table>		Variable	Importance Loading	HHI_sicc	0.95	HHI_sic	0.90	<table><thead><tr><th>Variable</th><th>Importance Loading</th></tr></thead><tbody><tr><td>HHI_sicc</td><td>0.95</td></tr><tr><td>HHI_sic</td><td>0.90</td></tr></tbody></table>		Variable	Importance Loading	HHI_sicc	0.95	HHI_sic	0.90								
Variable	Importance Loading																																						
HHI_sicc	0.95																																						
HHI_sic	0.90																																						
Variable	Importance Loading																																						
HHI_sicc	0.95																																						
HHI_sic	0.85																																						
Variable	Importance Loading																																						
HHI_sicc	0.95																																						
HHI_sic	0.90																																						
Variable	Importance Loading																																						
HHI_sicc	0.95																																						
HHI_sic	0.90																																						
Profitability																																							



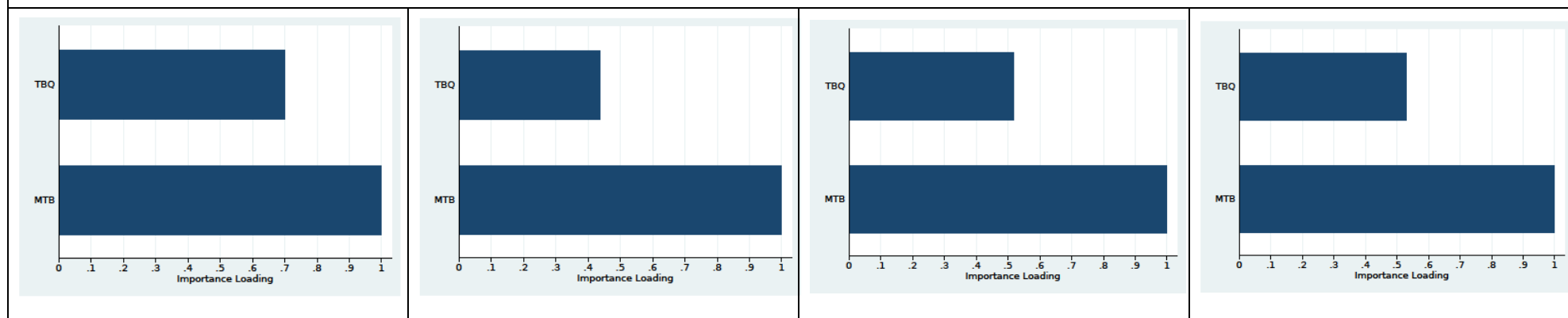
Operating Cash Flow



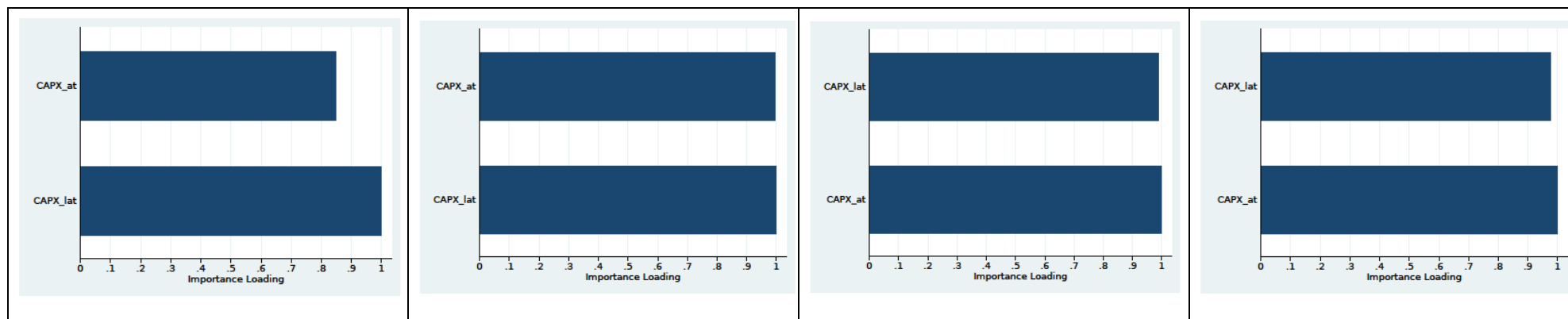
NOL



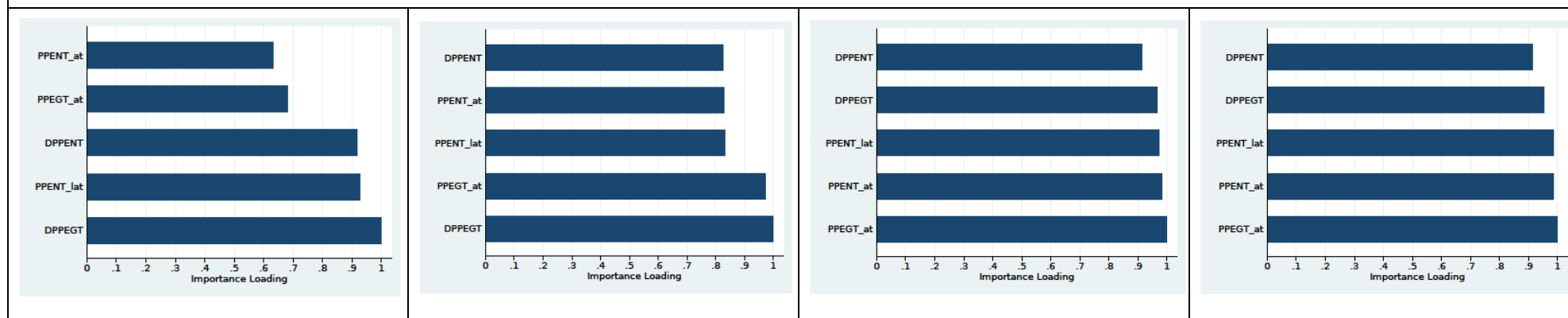
Valuation



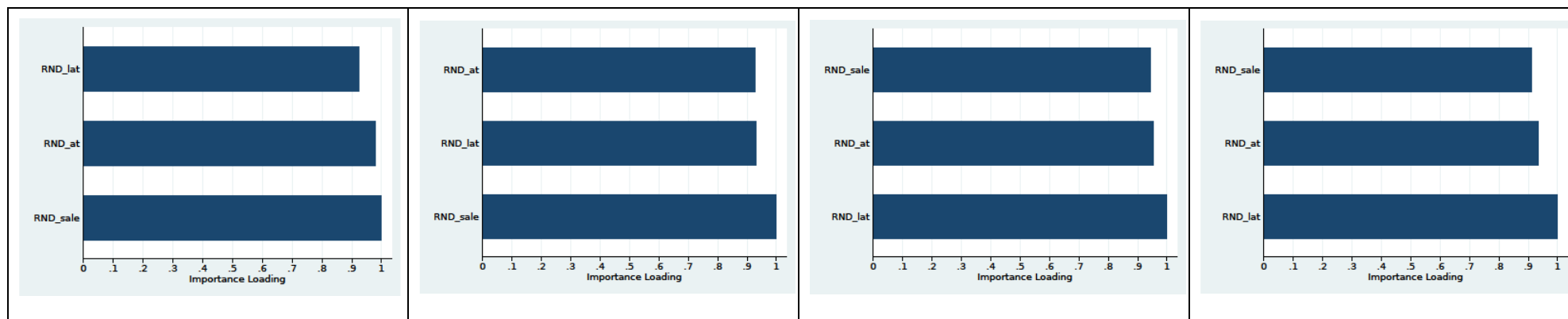
Investment



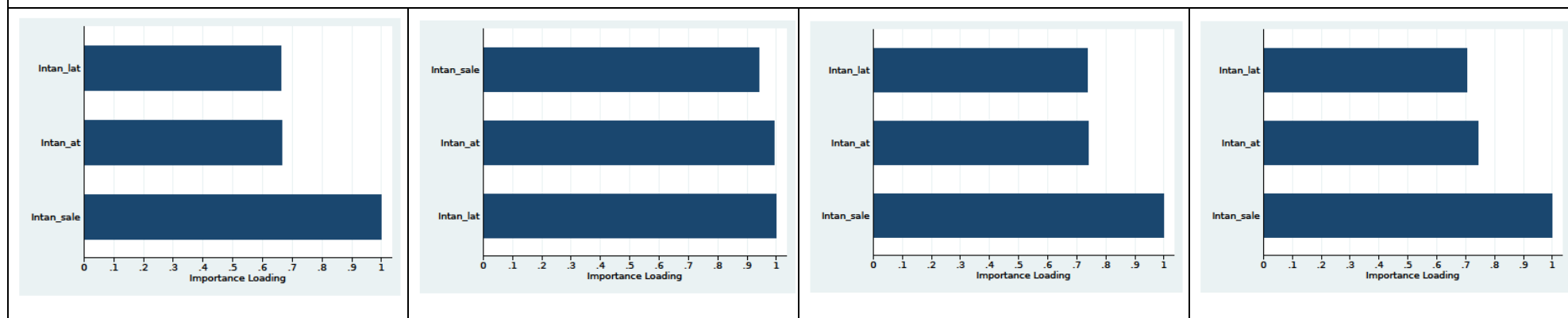
PPE



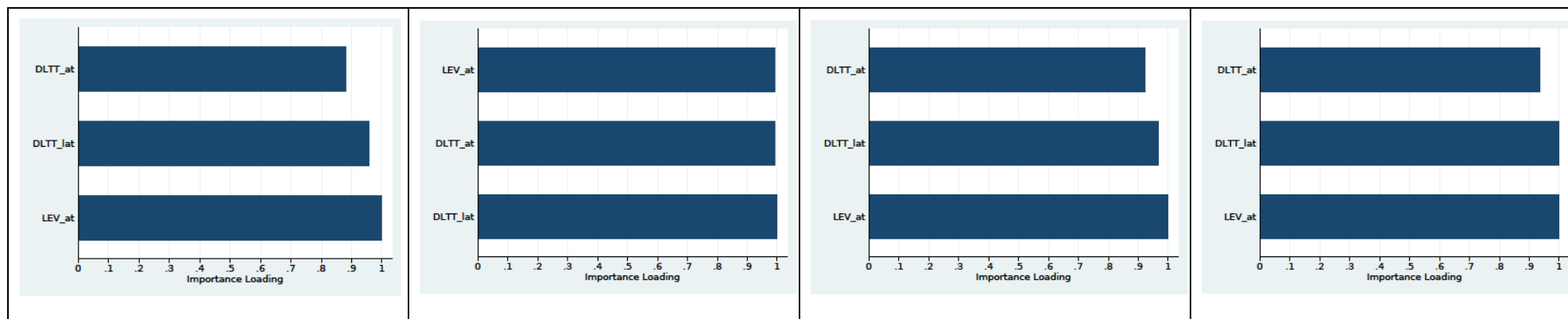
RND



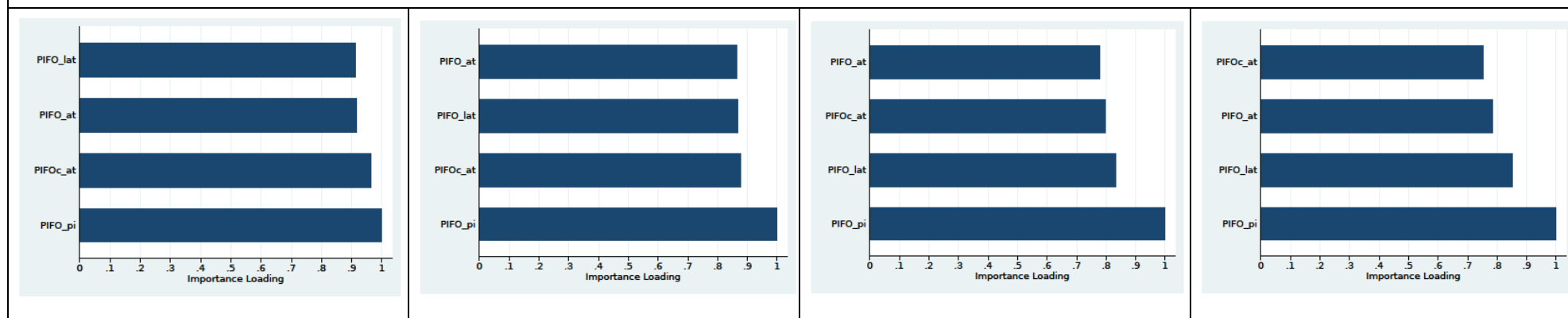
Intangible Assets



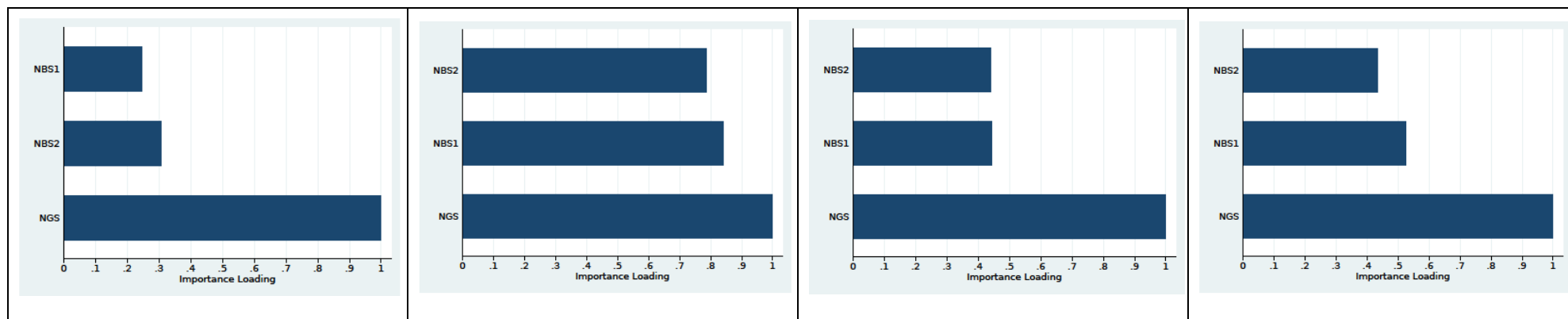
Leverage



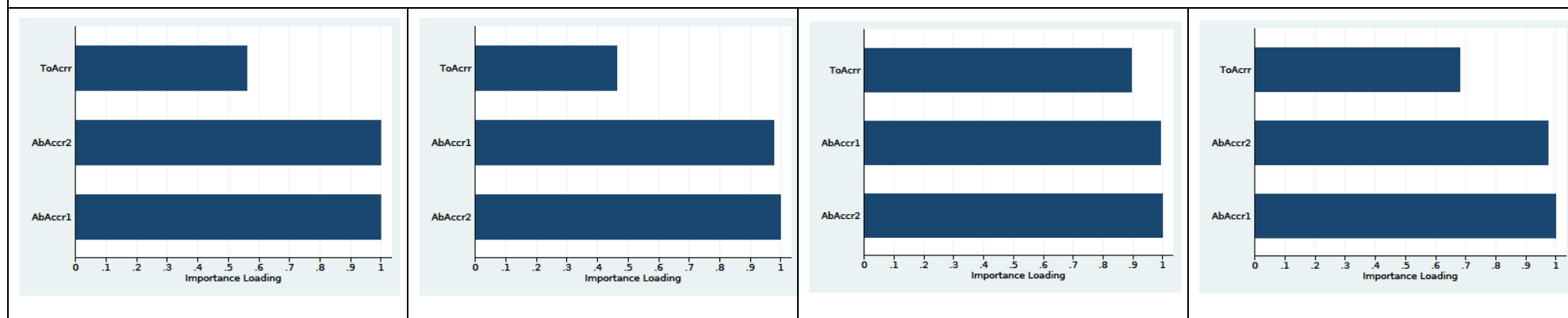
Foreign Operation



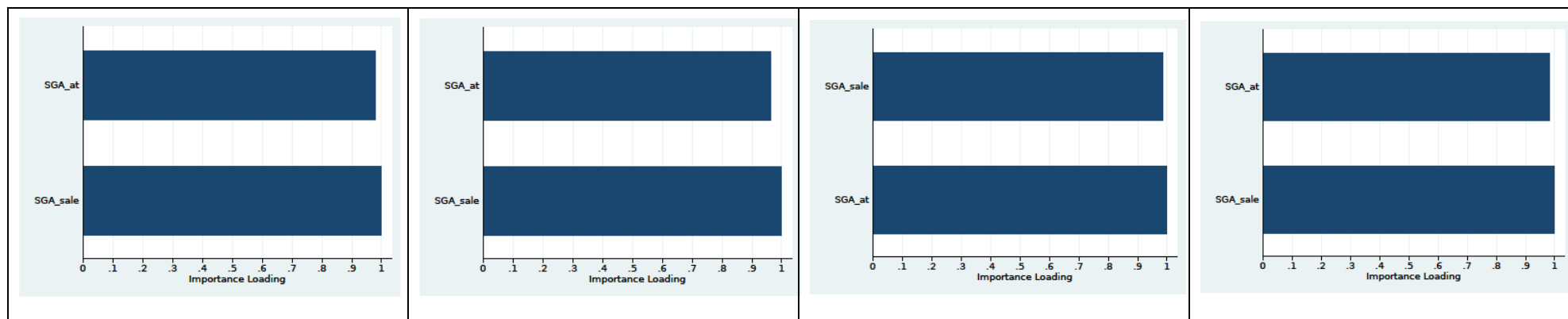
Segments



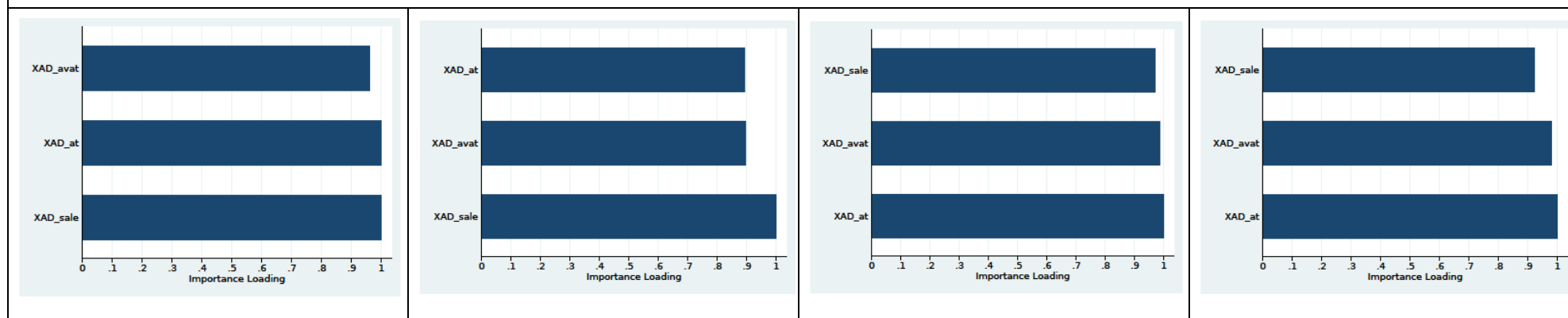
Accruals



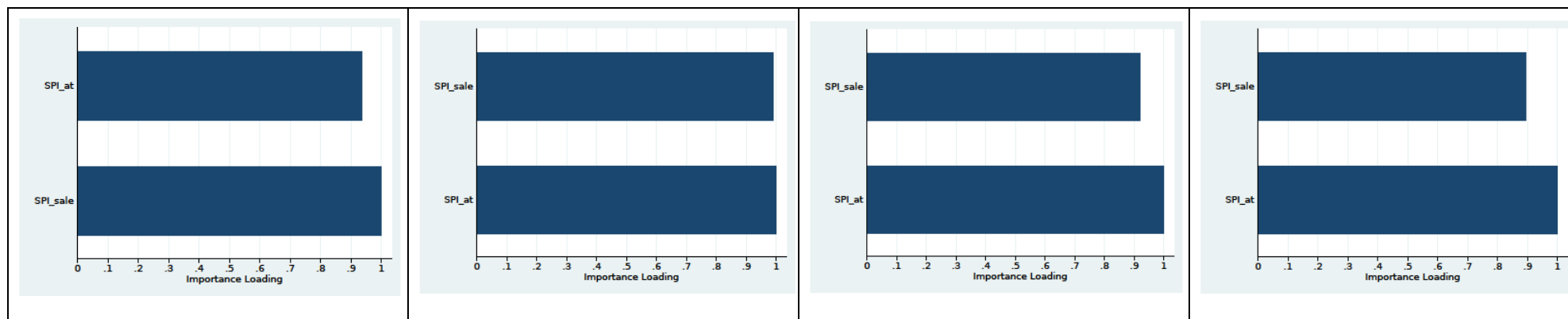
SGNA



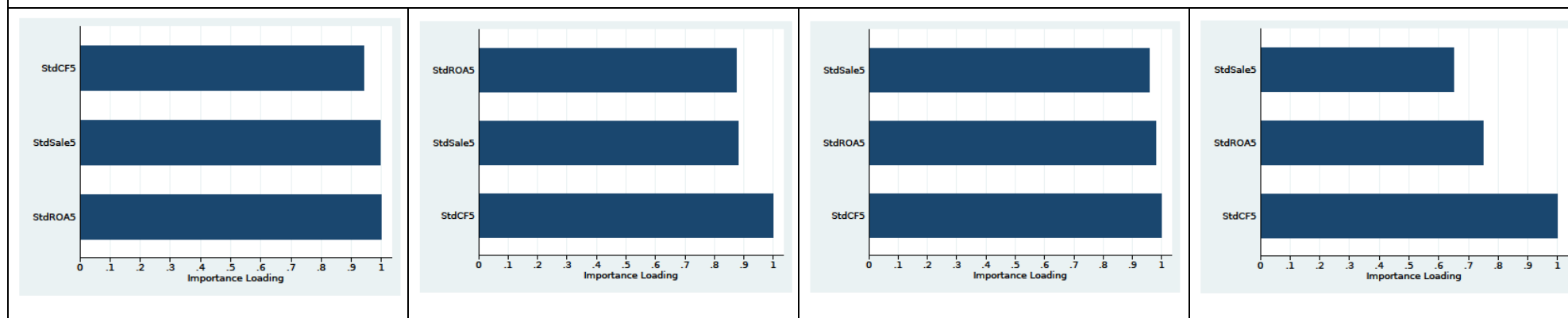
Advertisement Expense



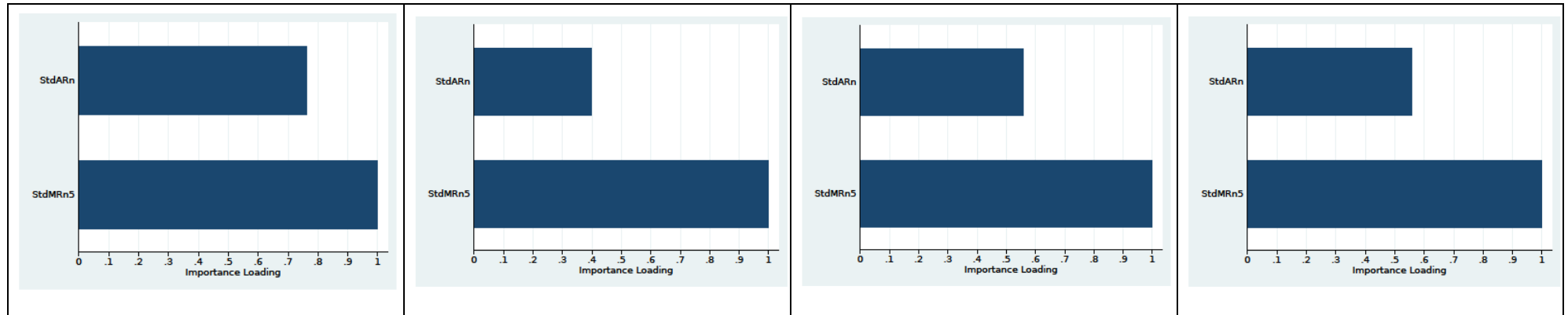
Special Items



Operating Risk



Return Volatility



Appendix 4F

Appendix 4F.1 Adaptive LASSO analysis for UTB

This table reports the results of Adaptive LASSO analysis for UTB in Step 2. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv				1							1
EMP	1	1	1	1	1	1	1				
HHI_sicc		1	1	1							
ROA_iblat		1	1	1	1	1	1	1			
OCFf_lat	1		1	1	1					1	1
DNOL			1			1					
NOL1_d		1	1								1
Return			1	1						1	
MTB	1	1	1	1			1				
DSale	1	1	1	1	1					1	1
MA_d		1									
CAPX_lat	1	1	1	1	1	1	1				
DPPEGT	1			1		1	1	1	1	1	1
Inv_t_lat				1		1					
DP_at	1	1	1	1	1				1	1	
DGW_at	1										1
RND_sale		1	1	1	1		1	1	1	1	1
Intan_sale	1			1	1	1					
LEV_at	1	1	1	1	1	1	1				
INT_lat	1	1	1	1							
Mezzanine	1	1	1	1		1	1				
Fin_d											

IO_ts	1	1	1						1	1	
CHE_at	1	1								1	1
SA_HP2010										1	1
AltmanZ	1	1	1	1					1		
PIFO_pi	1	1	1	1	1	1	1	1	1	1	1
PIFO_d	1	1		1							
NGS			1	1	1	1	1	1	1	1	1
ESUB_lat	1	1	1	1							
ESUB_d											
StdEarnFst	1	1									
Num_Analyst	1		1	1	1	1	1	1	1	1	1
Goodnews_d											
Big4_d		1									
AuditOp_d	1	1	1	1		1					
AbAccr1		1	1	1		1					
Age	1	1	1								
SGA_sale		1		1					1		
XAD_sale		1	1			1				1	1
SPI_sale	1	1	1	1	1						
EI_at											
StdROA5		1	1	1	1	1	1	1	1	1	1
StdMRn5				1							1

Appendix 4F.2 Adaptive LASSO analysis for DTAX

This table reports the results of Adaptive LASSO analysis for DTAX in Step 2. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv				1	1	1	1	1	1												1		
EMP	1			1				1	1	1	1	1	1	1	1	1	1	1	1	1			
HHI_sicc	1		1				1			1	1			1		1			1				
ROA_iblat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OCff_lat							1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1
DNOL	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
NOL1_d							1							1			1	1	1	1			
Return				1	1							1			1	1	1				1	1	
MTB				1	1					1	1	1		1	1	1	1	1	1	1	1	1	
DSale	1					1	1	1	1	1	1			1	1								
MA_d			1											1	1	1	1						
CAPX_lat							1	1	1	1	1			1	1	1	1	1	1	1	1	1	1
DPPEGT		1	1	1	1										1								
Invt_lat			1							1	1	1		1	1	1	1		1	1	1	1	1
DP_at								1	1	1	1			1		1			1	1	1	1	1
DGW_at	1								1	1	1	1	1	1					1	1	1	1	
RND_sale	1		1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
Intan_lat			1	1			1	1	1	1				1	1	1	1	1	1	1	1		
DLTT_lat								1	1	1						1	1		1				
INT_lat	1						1				1			1	1			1	1	1	1	1	1
Mezzanine				1						1	1	1		1	1								
Fin_d									1	1	1		1	1			1				1		
IO_ts							1		1	1	1					1							
CHE_at			1										1	1		1	1						

SA_HP2010	1							1						1				1			1	1
AltmanZ	1							1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PIFO_pi	1							1	1	1	1	1	1	1	1	1	1					
PIFO_d	1									1			1									
NGS	1		1	1	1	1						1	1	1				1	1	1		1
ESUB_lat	1			1	1	1	1	1	1		1	1						1		1	1	1
ESUB_d	1			1	1	1			1	1	1	1	1	1	1	1		1		1		
StdEarnFst									1	1			1	1	1							
Num_Analyst			1							1	1			1	1			1	1		1	
Goodnews_d													1					1	1			
Big4_d													1		1							
AuditOp_d				1				1	1				1				1					
AbAccr2	1			1						1			1	1	1	1	1	1		1	1	1
Age				1							1		1		1	1	1	1		1		
SGA_sale								1					1									1
XAD_sale			1	1														1	1			
SPI_at	1	1	1	1					1		1				1	1					1	
EL_at								1	1	1	1		1	1								
StdCF5																		1	1	1	1	1
StdMRn5	1							1	1	1	1		1	1	1	1			1	1		

Appendix 4F.3 Adaptive LASSO analysis for DDBTD

This table reports the results of Adaptive LASSO analysis for DDBTD in Step 2. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv										1		1							1		1	1	1
EMP	1	1	1			1			1	1	1	1	1	1	1		1	1	1	1	1	1	1
HHI_sicc	1							1	1	1		1	1						1				
ROA_iblat	1	1	1				1	1	1	1		1		1	1				1	1	1		
OCff_lat	1	1				1	1		1	1			1						1	1	1		1
DNOL		1				1	1	1	1	1	1	1	1	1	1				1	1	1		1
NOL2_d	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Return	1				1				1	1				1	1		1	1	1				
MTB				1	1					1	1	1	1			1	1				1		
DSale						1		1	1	1		1	1				1		1	1	1	1	1
MA_d									1	1													
CAPX_at	1	1								1	1	1	1	1	1	1	1	1	1	1	1	1	1
PPEGT_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				1
Invt_lat			1	1	1	1	1	1	1	1	1	1						1	1		1	1	1
DP_at	1	1			1	1			1	1	1	1	1	1	1		1	1	1	1	1	1	1
DGW_at										1				1	1				1		1		
RND_lat	1	1					1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1
Intan_sale					1		1	1	1	1	1	1	1				1	1	1	1	1	1	1
LEV_at					1	1	1	1	1	1	1	1	1						1				
INT_lat									1	1				1			1	1	1		1		1
Mezzanine	1						1	1	1	1				1					1		1		
Fin_d						1			1	1	1			1	1		1				1		
IO_ts						1	1		1	1	1	1	1	1			1	1	1	1	1	1	1
CHE_at	1	1	1	1	1	1	1	1	1	1	1								1		1	1	1
SA_HP2010									1	1				1	1	1	1	1			1		
AltmanZ	1	1	1	1	1	1	1	1	1	1	1	1	1	1				1			1	1	1
PIFO_pi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

PIFO_d	1	1					1			1			1	1	1			1	1		1		1
NGS	1	1							1	1				1	1					1	1		1
ESUB_lat	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1		
ESUB_d							1	1	1	1									1				1
StdEarnFst								1	1	1	1			1					1				
Num_Analyst							1	1	1	1			1	1	1	1	1	1	1	1	1		1
Goodnews_d							1	1	1	1		1		1					1				
Big4_d					1		1	1	1	1	1	1							1		1		1
AuditOp_d								1	1	1			1	1	1				1			1	1
AbAccr2	1	1			1					1									1		1		
Age	1	1	1	1	1	1	1	1	1	1	1	1	1						1				1
SGA_at							1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XAD_at										1		1	1						1		1		
SPI_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
EI_at									1	1		1											
StdCF5	1					1	1	1	1	1		1	1						1				1
StdMRn5	1				1			1	1	1				1					1	1			

Appendix 4F.4 Adaptive LASSO analysis for MPBTD

This table reports the results of Adaptive LASSO analysis for MPBTD in Step 2. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

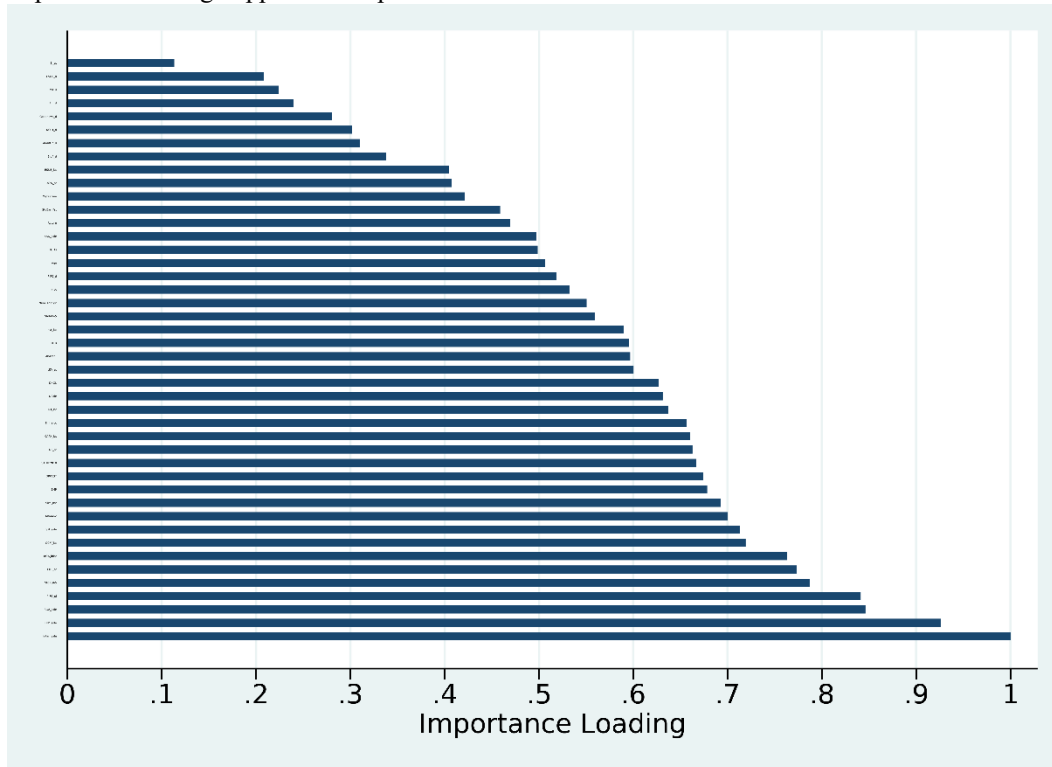
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv											1	1		1	1	1	1		1		1	1	1
EMP		1				1			1	1	1					1	1	1	1	1			
HHI_sicc	1	1								1	1	1	1	1			1				1		
ROA_iblat	1	1				1	1	1	1	1				1	1	1	1		1	1		1	1
OCff_lat	1	1	1			1	1			1	1	1		1	1	1	1	1		1		1	1
DNOL		1				1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1
NOL2_d	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Return		1	1		1		1	1	1	1	1		1	1	1	1	1	1	1	1	1		
MTB		1			1		1			1	1	1	1	1	1	1	1	1					
DSale			1			1	1			1			1				1			1	1	1	1
MA_d										1				1						1	1		
CAPX_at	1	1					1			1	1	1	1	1		1	1	1	1	1	1	1	
PPEGT_at		1	1	1	1	1	1	1	1	1	1		1	1	1		1	1	1			1	1
Inv_t_lat			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DP_at						1				1	1	1	1	1		1	1	1	1	1	1	1	1
DGW_at										1	1			1	1		1						
RND_lat	1	1					1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Intan_sale	1	1	1	1	1	1					1	1	1	1				1		1	1	1	1
LEV_at					1	1	1	1		1	1	1	1	1				1		1			1
INT_lat		1		1					1	1	1		1	1	1	1	1	1	1		1	1	1
Mezzanine							1	1	1	1	1		1	1	1	1				1	1		
Fin_d						1			1	1	1		1	1	1			1			1		
IO_ts		1				1	1						1	1				1	1	1	1	1	
CHE_at	1	1	1	1	1	1			1	1	1		1	1			1				1	1	1
SA_HP2010	1				1				1				1	1	1	1	1				1		
AltmanZ		1				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1
PIFO_pi		1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

PIFO_d													1	1	1	1		1			1	1	1
NGS	1	1								1	1	1	1	1	1	1	1	1	1	1	1	1	1
ESUB_lat	1	1	1			1	1	1	1	1	1	1	1	1	1	1	1	1	1				
ESUB_d									1	1			1	1						1	1	1	
StdEarnFst							1	1	1	1			1				1				1		
Num_Analyst							1			1		1	1	1	1	1	1		1				1
Goodnews_d							1				1	1	1						1				
Big4_d							1		1	1	1		1				1			1	1	1	
AuditOp_d							1	1	1				1	1	1		1	1	1	1	1	1	1
AbAccr1	1	1			1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1		
Age		1			1					1	1	1		1	1	1	1	1	1	1		1	1
SGA_sale															1	1	1	1	1	1	1		
XAD_at										1	1					1	1		1	1	1	1	
SPI_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EI_at									1	1			1	1									
StdCF5						1	1	1	1	1	1		1	1	1		1	1	1	1	1	1	1
StdMRn5		1							1	1	1	1		1		1	1	1	1	1	1	1	1

Appendix 4G

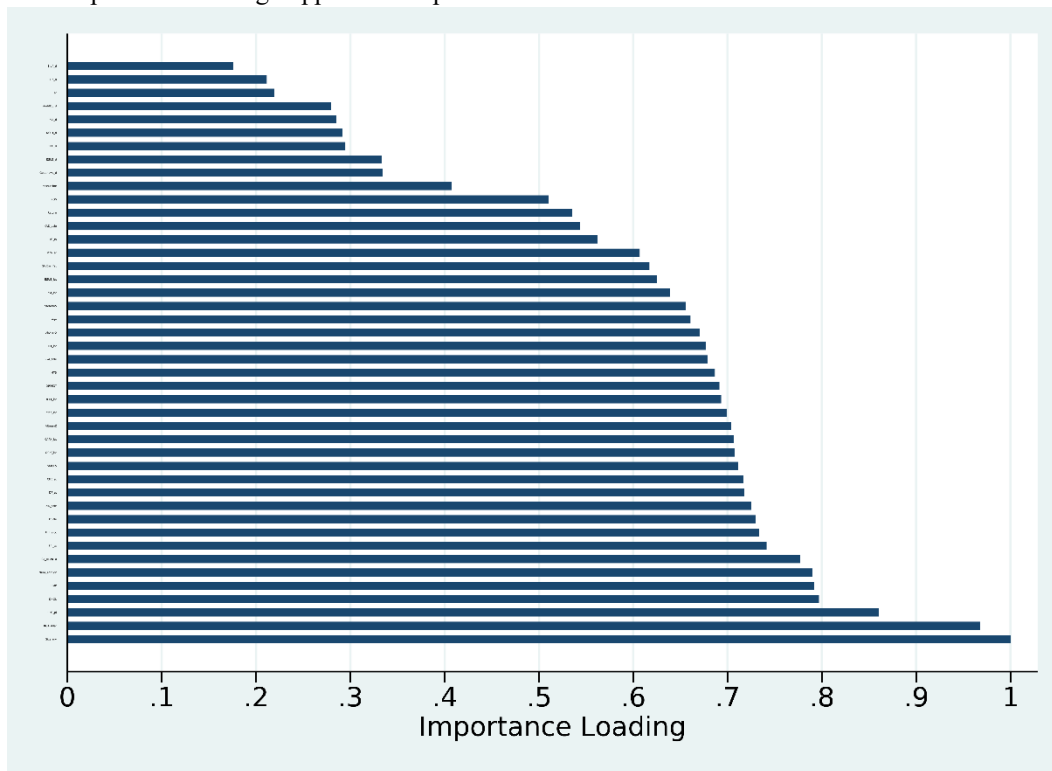
Appendix 4G.1 Random Forest analysis for UTB

This figure plots the results of Random Forest analysis for UTB in Step 2. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



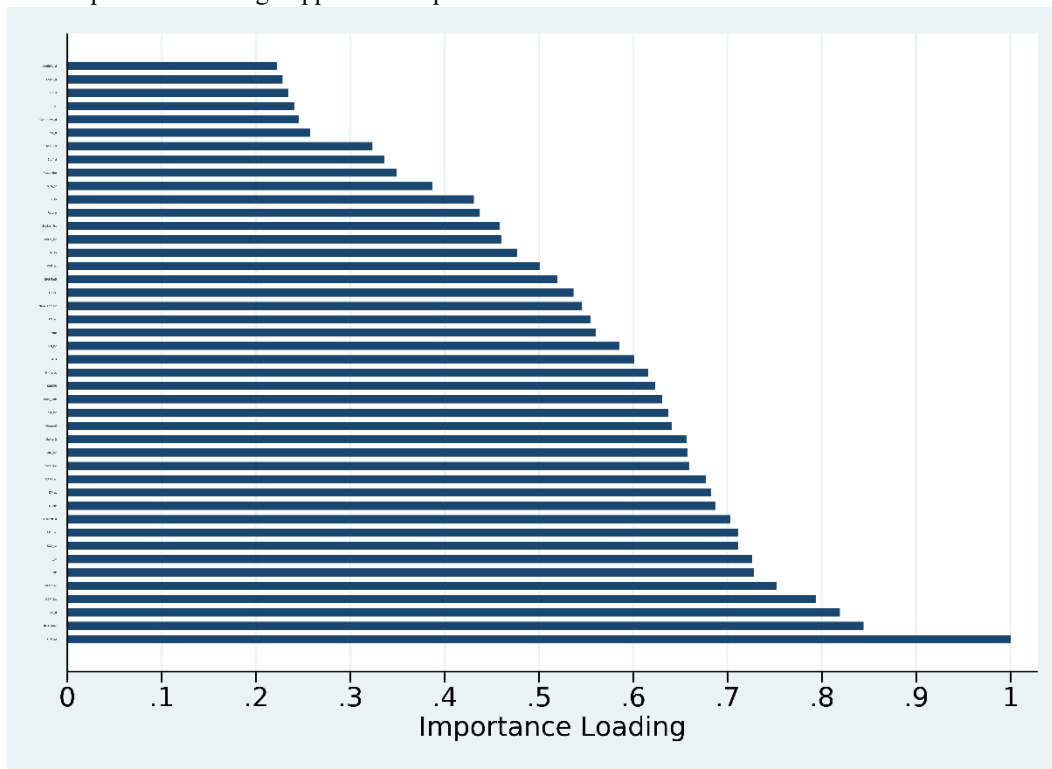
Appendix 4G.2 Random Forest analysis for DTAX

This figure plots the results of Random Forest analysis for DTAX in Step 2. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



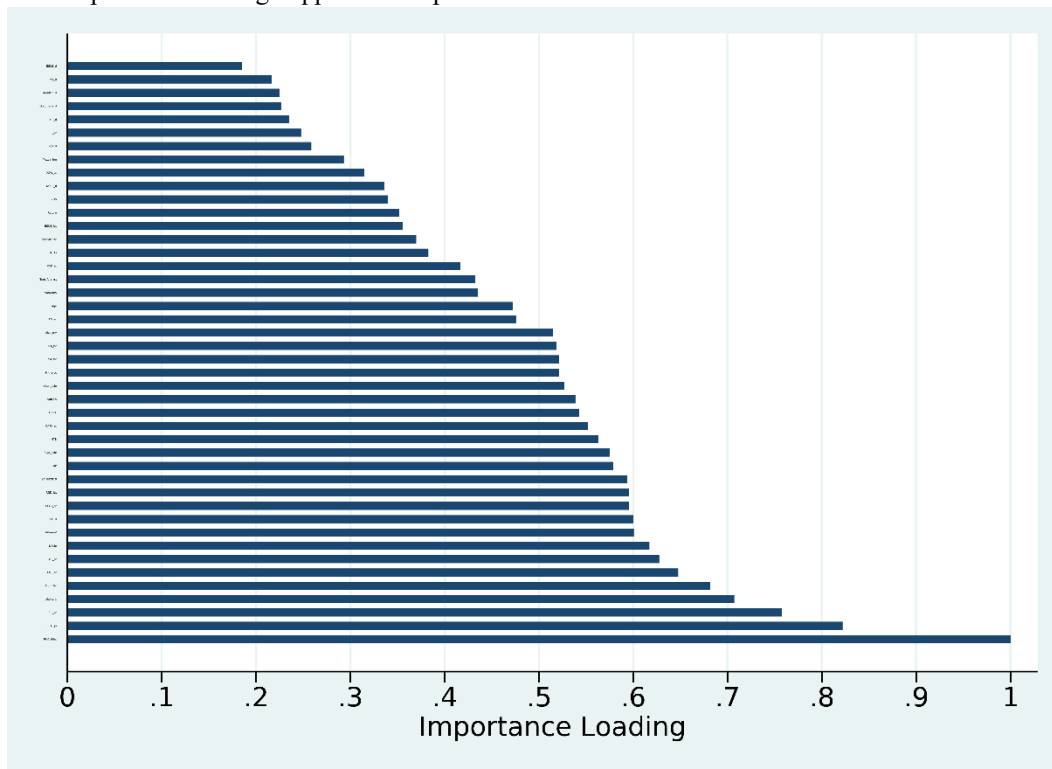
Appendix 4G.3 Random Forest analysis for DDBTD

This figure plots the results of Random Forest analysis for DDBTD in Step 2. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



Appendix 4G.4 Random Forest analysis for MPBTD

This figure plots the results of Random Forest analysis for MPBTD in Step 2. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



Appendix 4H

Appendix 4H.1 Adaptive LASSO analysis for annual cash ETR (further analysis)

This table reports the results of Adaptive LASSO analysis for annual cash ETR in Step 2, including corporate governance covariates. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv			1									1	1		1	1	1
EMP				1				1	1	1		1	1	1	1	1	1
HHI_sicc				1				1	1	1	1	1	1	1	1	1	1
ROA_iblat	1	1		1	1	1	1				1	1	1		1		
OCFf_lat		1	1	1	1	1	1	1		1	1	1	1	1	1	1	1
DNOL		1	1	1	1	1	1	1				1	1	1	1		
NOL1_d	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Return		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MTB			1	1							1	1	1	1	1	1	1
DSale	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MA_d	1										1	1	1	1	1		
CAPX_lat		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DPPEGT			1						1		1	1	1		1	1	1
Invt_lat	1		1	1		1					1	1			1		
DP_at			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DGW_at		1	1				1					1			1		1
RND_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Intan_sale		1				1			1	1	1	1	1	1	1		1
DLTT_lat	1								1	1	1	1	1	1	1		1
INT_lat	1	1	1	1		1	1	1		1	1	1			1	1	
Mezzanine	1	1				1	1	1			1	1	1	1	1		
Fin_d	1	1	1									1		1	1	1	1
IO_ts												1	1	1		1	
CHE_at				1								1	1	1	1	1	1
SA_HP2010			1						1		1	1			1		1

AltmanZ	1	1	1	1	1		1	1	1	1	1			1		1
PIFO_pi	1	1	1				1	1	1	1	1	1	1	1	1	1
PIFO_d			1	1						1	1	1	1	1		
NBS2		1	1	1						1	1			1	1	1
ESUB_lat	1	1	1						1	1	1	1		1		
ESUB_d										1	1			1		
StdEarnFst	1	1	1		1		1	1			1	1		1	1	
Num_Analyst	1	1	1								1	1	1		1	1
Goodnews_d					1	1	1	1	1	1	1	1		1	1	1
Big4_d			1			1	1					1			1	1
AuditOp_d	1	1	1									1			1	1
AbAccr1	1		1					1	1		1	1			1	
Age								1			1	1	1			1
SGA_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XAD_sale											1	1			1	
SPI_sale	1	1	1	1								1	1	1	1	1
EI_at		1	1	1	1							1				
StdCF5		1				1	1			1	1	1			1	
StdMRn5	1	1	1	1	1	1			1	1	1	1	1	1	1	1
BrdSize		1									1	1			1	
Ind_dir			1	1	1	1	1	1	1	1	1	1		1	1	
Female_d			1		1	1	1			1	1	1			1	1
CEOage						1	1					1		1	1	
Tenure_CEO		1	1	1							1	1	1	1	1	1
StkCompen_d			1	1	1							1				1
Stk_ceo		1	1	1				1	1	1	1	1			1	
Cash_compen		1	1			1			1			1	1	1	1	1
CEOChair_d							1		1	1	1	1	1	1	1	1

Appendix 4H.2 Adaptive LASSO analysis for annual GAAP ETR (further analysis)

This table reports the results of Adaptive LASSO analysis for annual GAAP ETR in Step 2, including corporate governance covariates. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv	1				1		1		1	1							
EMP	1	1		1	1	1	1	1	1	1							
HHI_sicc						1	1						1				
ROA_iblat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OCFf_lat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DNOL		1															
NOL1_d		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Return	1	1					1	1		1	1	1			1		
MTB		1							1	1	1	1	1	1	1	1	1
DSale		1		1						1		1		1	1		
MA_d				1	1		1	1	1								
CAPX_lat	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1
DPPEGT	1				1	1	1		1								
Inv_t_lat	1			1	1					1					1	1	1
DP_at	1	1	1	1	1								1	1	1	1	1
DGW_at								1							1		
RND_sale	1		1	1		1	1	1	1		1	1		1	1	1	1
Intan_sale	1	1			1				1	1	1	1	1				
DLTT_lat	1	1								1	1				1	1	1
INT_lat	1	1			1	1						1	1	1			
Mezzanine	1			1	1												
Fin_d										1		1					
IO_ts															1	1	1
CHE_at				1	1	1	1	1	1	1	1	1			1	1	1
SA_HP2010	1				1											1	
AltmanZ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
PIFOc_at	1	1			1	1	1	1	1	1	1	1	1	1		1	1

PIFO_d	1	1		1	1					1		1					
NBS2		1	1	1						1							
ESUB_lat	1	1								1		1					
ESUB_d									1							1	1
StdEarnFst	1	1			1					1	1	1					
Num_Analyst		1										1			1	1	1
Goodnews_d	1	1						1	1	1	1	1	1		1	1	
Big4_d																	
AuditOp_d				1	1					1						1	1
AbAccr2												1	1				
Age	1	1	1	1	1	1	1	1	1	1	1			1	1	1	1
SGA_at	1	1		1	1			1	1	1	1	1	1	1	1	1	
XAD_sale		1	1	1	1			1		1			1		1		
SPI_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EI_at	1			1													
StdCF5	1	1			1					1	1	1	1		1	1	1
StdMRn5	1			1	1			1	1	1	1	1	1	1	1	1	1
BrdSize	1	1	1	1	1			1							1		
Ind_dir				1	1					1							
Female_d	1			1								1	1				
CEOage										1	1	1					1
Tenure_CEO					1					1		1	1	1	1		
StkCompen_d	1	1		1	1			1	1	1					1	1	1
Stk_ceo	1	1						1	1								
Cash_compen	1	1	1	1	1												
CEOChair_d	1											1	1	1	1	1	1

Appendix 4H.3 Adaptive LASSO analysis for long-run cash ETR (further analysis)

This table reports the results of Adaptive LASSO analysis for long-run cash ETR in Step 2, including corporate governance covariates. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Size_mv			1	1										1	1
EMP				1	1			1			1	1	1	1	1
HHI_sicc				1	1	1				1	1	1	1	1	1
ROA_iblat	1	1	1	1	1	1				1	1	1	1	1	1
OCFf_lat		1	1	1		1	1				1		1	1	1
DNOL								1					1	1	
NOL1_d	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Return			1	1	1	1	1	1	1	1	1	1		1	1
MTB		1				1								1	
DSale											1		1	1	1
MA_d	1												1	1	
CAPX_lat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DPPEGT			1	1		1				1	1			1	
Inv_t_lat	1	1	1	1	1	1				1	1	1	1	1	1
DP_at		1	1	1	1	1				1	1	1	1	1	1
DGW_at	1			1									1	1	
RND_sale	1	1	1	1	1		1		1	1	1	1		1	
Intan_sale	1	1		1		1	1	1			1	1	1	1	1
DLTT_lat				1	1	1	1	1	1			1	1	1	
INT_lat	1	1	1	1				1		1	1	1	1	1	1
Mezzanine	1	1	1			1				1	1	1	1	1	
Fin_d								1		1			1	1	
IO_ts		1	1	1	1	1					1	1	1	1	
CHE_at						1							1	1	1
SA_HP2010						1								1	1
AltmanZ			1					1	1	1	1			1	1
PIFOc_at			1	1	1	1	1	1		1	1	1	1	1	

PIFO_d		1	1	1					1	1	1	1	1	1
NBS2				1			1						1	1
ESUB_lat													1	
ESUB_d			1	1	1					1			1	
StdEarnFst	1	1	1	1	1	1			1				1	1
Num_Analyst													1	1
Goodnews_d						1		1						1
Big4_d													1	1
AuditOp_d					1	1							1	1
AbAccr1						1		1	1			1	1	
Age			1	1		1		1	1	1		1	1	1
SGA_at	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XAD_sale			1	1	1	1		1	1	1		1	1	1
SPI_sale	1		1					1				1	1	1
EI_at		1		1	1	1								
StdCF5				1							1		1	
StdMRn5	1	1	1	1	1						1		1	1
BrdSize						1			1	1	1	1	1	
Ind_dir						1		1					1	
Female_d					1	1							1	1
CEOage					1	1						1	1	1
Tenure_CEO			1	1									1	1
StkCompen_d				1		1							1	
Stk_ceo	1	1	1	1	1			1	1	1	1		1	
Cash_compen			1	1	1	1			1	1	1	1	1	
CEOChair_d													1	

Appendix 4H.4 Adaptive LASSO analysis for long-run GAAP ETR (further analysis)

This table reports the results of Adaptive LASSO analysis for long-run GAAP ETR in Step 2, including corporate governance covariates. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Size_mv							1								
EMP					1	1	1	1		1				1	
HHI_sicc											1		1		
ROA_iblat		1	1				1	1		1	1	1	1	1	1
OCFf_lat	1				1	1	1	1		1	1	1	1	1	1
DNOL	1		1	1											
NOL1_d	1		1	1	1		1								1
Return														1	1
MTB				1	1									1	1
DSale			1	1								1	1	1	1
MA_d								1		1					
CAPX_lat	1	1	1					1		1	1	1	1	1	1
DPPEGT							1	1							
Inv_t_lat		1	1							1				1	1
DP_at															
DGW_at							1							1	
RND_sale	1											1			1
Intan_sale		1	1	1	1	1	1	1	1	1	1		1	1	1
DLTT_lat											1	1	1	1	1
INT_lat			1							1					
Mezzanine			1	1	1		1	1							1
Fin_d															
IO_ts														1	1
CHE_at										1				1	1
SA_HP2010															1
AltmanZ	1		1	1	1	1	1	1		1	1	1	1	1	1
PIFO_pi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

[illegible]

Appendix 4H.5 Adaptive LASSO analysis for UTB (further analysis)

This table reports results of Adaptive LASSO analysis for UTB in Step 2, including corporate governance covariates. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv				1					1		1
EMP	1	1	1	1	1	1	1	1	1	1	
HHI_sicc			1	1							
ROA_iblat			1	1	1	1	1	1	1	1	1
OCFf_lat									1		
DNOL			1	1				1	1	1	1
NOL1_d									1	1	
Return					1				1	1	
MTB	1	1	1	1					1	1	1
DSale			1		1				1		
MA_d									1		
CAPX_lat	1		1	1	1				1		1
DPPEGT			1	1	1	1	1	1	1	1	1
Inv_t_lat			1	1	1	1	1	1	1		
DP_at									1		
DGW_at									1		1
RND_sale	1	1	1	1	1	1	1	1	1	1	1
Intan_sale	1	1	1	1	1	1			1		
LEV_at	1		1	1	1	1			1	1	
INT_lat			1	1					1		
Mezzanine	1		1	1			1	1	1		
Fin_d									1		
IO_ts							1	1	1	1	
CHE_at	1			1					1		
SA_HP2010									1		
AltmanZ			1	1	1				1	1	1
PIFO_pi	1	1	1	1	1	1	1	1	1	1	1
PIFO_d											
NGS	1	1	1	1		1	1	1	1	1	1
ESUB_lat	1	1	1	1					1		
ESUB_d											
StdEarnFst			1						1		
Num_Analyst			1	1	1	1	1	1	1	1	1
Goodnews_d									1	1	1
Big4_d									1		
AuditOp_d	1								1		
AbAccr1									1		
Age	1	1	1						1		
SGA_sale									1		
XAD_sale	1	1	1	1	1	1	1	1	1	1	1
SPI_sale		1	1	1	1				1		
EI_at											
StdROA5		1	1	1	1	1	1	1	1	1	1
StdMRn5				1					1		

BrdSize	1		1	1	1	1	1
Ind_dir	1				1	1	1
Female_d	1				1		
CEOage					1		
Tenure_CEO					1		
StkCompen_d	1	1			1		
Stk_ceo					1	1	1
Cash_compen							
CEOChair_d				1	1	1	

Appendix 4H.6 Adaptive LASSO analysis for DTAX (further analysis)

This table reports the results of Adaptive LASSO analysis for DTAX in Step 2, including corporate governance covariates. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv												1	1	1	1		
EMP	1	1	1	1	1		1	1	1	1	1	1	1	1			1
HHI_sicc				1									1				
ROA_iblat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OCFf_lat			1	1	1					1		1	1	1	1	1	1
DNOL	1	1	1	1			1	1	1	1	1	1	1	1			
NOL1_d							1					1	1				
Return			1						1	1	1	1		1	1	1	
MTB			1	1				1	1	1	1	1	1		1		
DSale	1	1	1	1					1			1			1	1	1
MA_d																	1
CAPX_lat			1		1				1	1	1	1	1	1	1	1	1
DPPEGT			1	1					1	1							
Inv_t_lat			1							1		1	1	1	1		1
DP_at			1	1	1					1				1	1		1
DGW_at					1		1	1									
RND_sale	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Intan_lat	1	1							1	1	1	1	1	1	1		
DLTT_lat			1	1					1	1	1	1		1			
INT_lat	1	1						1					1	1	1	1	
Mezzanine																	
Fin_d												1					
IO_ts			1	1					1	1					1		
CHE_at									1	1					1	1	1
SA_HP2010			1	1				1				1	1	1		1	1
AltmanZ			1				1	1	1	1	1	1	1	1	1	1	1
PIFO_pi			1	1	1	1	1	1	1	1	1	1	1	1			

PIFO_d	1	1								1									1
NGS						1							1						
ESUB_lat			1			1		1							1				1
ESUB_d								1	1	1			1						
StdEarnFst												1							
Num_Analyst						1		1			1		1						
Goodnews_d																			
Big4_d								1	1										
AuditOp_d							1		1		1								
AbAccr2						1	1	1	1	1	1	1							1
Age										1	1				1				
SGA_sale			3																
XAD_sale																			1
SPI_at			1	1					1										
EI_at	1	1						1											
StdCF5			1											1	1	1	1	1	
StdMRn5			1	1				1			1			1	1	1	1	1	
BrdSize											1								
Ind_dir											1								
Female_d				1							1								
CEOage	1	1	1	1															
Tenure_CEO								1	1		1	1	1						
StkCompen_d											1	1							
Stk_ceo											1								
Cash_compen						1					1								
CEOChair_d																			

Appendix 4H.7 Adaptive LASSO analysis for DDBTD (further analysis)

This table reports the results of Adaptive LASSO analysis for DDBTD in Step 2, including corporate governance covariates. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv				1			1	1	1				1		1		1
EMP				1					1	1		1	1	1	1	1	1
HHI_sicc							1	1									1
ROA_iblat	1	1										1	1	1			1
OCFf_lat							1	1	1	1		1	1	1		1	1
DNOL	1	1	1	1				1	1	1	1	1	1	1	1	1	1
NOL2_d	1				1	1	1	1	1	1	1	1	1	1	1	1	1
Return	1	1					1	1	1	1		1	1				1
MTB				1	1					1	1			1		1	1
DSale	1						1			1	1			1		1	1
MA_d		1		1				1	1								
CAPX_at	1			1	1	1	1	1	1	1	1	1	1	1	1	1	1
PPEGT_at	1	1	1	1									1				1
Inv_t_lat		1	1	1				1				1	1		1	1	1
DP_at					1	1	1		1	1	1	1	1	1	1		
DGW_at									1	1				1	1		
RND_lat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Intan_sale	1	1	1	1	1				1	1	1	1	1	1	1	1	1
LEV_at																1	1
INT_lat			1	1				1				1				1	1
Mezzanine													1		1		
Fin_d		1					1	1	1								
IO_ts	1													1	1	1	1
CHE_at	1	1	1	1	1	1	1				1	1	1	1	1	1	1
SA_HP2010			1	1	1			1	1	1	1	1					1
AltmanZ					1	1	1	1						1	1	1	1
PIFO_pi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

PIFO_d	1					1	1	1	1	1		1	1		1	1	1
NGS						1	1	1	1						1		1
ESUB_lat	1	1	1	1	1			1	1	1	1	1	1	1	1	1	1
ESUB_d		1						1									1
StdEarnFst		1	1					1				1		1	1		
Num_Analyst				1						1	1	1		1			
Goodnews_d		1															
Big4_d								1	1	1	1	1			1	1	1
AuditOp_d		1	1				1	1	1			1				1	
AbAccr2								1	1								1
Age	1	1				1	1		1	1			1	1		1	1
SGA_at	1	1	1		1	1	1	1	1	1	1	1	1	1	1		1
XAD_at		1	1				1	1				1			1	1	1
SPI_at	1	1	1	1			1	1	1	1	1	1	1	1	1	1	1
EI_at				1			1										
StdCF5				1				1	1	1	1	1				1	1
StdMRn5	1			1			1								1	1	1
BrdSize								1							1	1	1
Ind_dir	1	1	1			1	1	1	1	1		1					
Female_d									1								1
CEOage	1	1		1			1		1	1		1		1			1
Tenure_CEO		1											1	1	1		1
StkCompen_d		1	1	1			1	1	1			1	1	1	1		
Stk_ceo		1		1						1	1	1	1				1
Cash_compen															1		1
CEOChair_d		1						1	1	1				1			1

Appendix 4H.8 Adaptive LASSO analysis for MPBTD (further analysis)

This table reports the results of Adaptive LASSO analysis for MPBTD in Step 2, including corporate governance covariates. Selected variables for the year are marked with 1. Appendix 4A provides variable definitions.

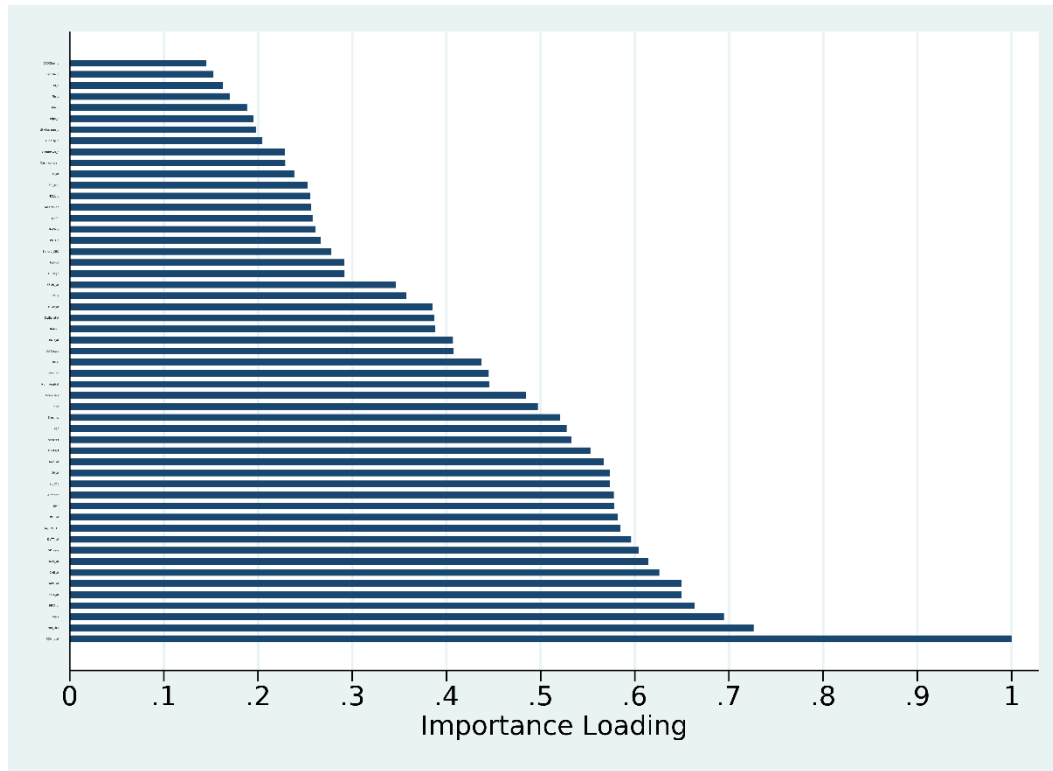
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Size_mv						1	1	1		1		1	1	1	1	1	1
EMP							1			1		1				1	
HHI_sicc						1	1	1		1						1	1
ROA_iblat									1	1	1	1			1	1	1
OCFf_lat	1					1			1	1		1			1	1	1
DNOL	1			1				1	1	1	1	1	1	1	1	1	1
NOL2_d				1	1	1	1	1	1	1	1	1	1	1	1	1	
Return	1	1				1	1	1	1	1		1	1				1
MTB					1	1		1	1	1	1	1				1	1
DSale							1	1		1	1					1	
MA_d				1	1	1		1	1	1		1				1	
CAPX_at				1	1	1	1	1	1	1	1	1		1	1	1	
PPEGT_at	1	1	1	1				1				1	1			1	1
Inv_t_lat				1	1	1		1		1	1	1	1	1	1	1	1
DP_at				1	1	1	1	1		1	1	1	1	1	1	1	1
DGW_at					1	1		1									
RND_lat		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Intan_sale			1	1			1	1				1			1	1	1
LEV_at						1		1				1				1	1
INT_lat				1	1		1	1	1	1		1					
Mezzanine							1	1		1			1	1		1	
Fin_d							1	1	1							1	
IO_ts							1	1	1	1						1	
CHE_at					1			1						1	1	1	1
SA_HP2010					1			1	1	1	1	1				1	1
AltmanZ				1	1	1	1	1	1	1		1					
PIFO_pi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

PIFO_d					1	1	1	1	1	1					1	1
NGS						1	1	1	1	1	1				1	1
ESUB_lat	1	1	1	1				1	1	1	1	1	1	1	1	1
ESUB_d				1				1	1	1					1	1
StdEarnFst					1			1					1		1	
Num_Analyst				1	1			1	1	1	1	1				
Goodnews_d										1						
Big4_d				1	1	1		1		1					1	1
AuditOp_d						1	1	1	1			1			1	1
AbAccr1	1	1	1	1	1	1	1	1		1		1	1	1		1
Age						1	1	1	1	1		1	1	1	1	1
SGA_sale				1		1	1		1	1	1	1	1	1	1	1
XAD_at					1							1				1
SPI_at	1			1	1	1	1	1	1	1	1	1	1	1	1	1
EI_at							1	1		1						
StdCF5		1		1	1	1		1	1	1	1	1	1	1	1	1
StdMRn5			1	1	1	1		1		1		1			1	1
BrdSize				1	1			1	1						1	1
Ind_dir	1	1	1	1	1	1	1	1	1	1		1			1	
Female_d				1				1		1		1			1	1
CEOage				1			1	1	1	1				1	1	1
Tenure_CEO												1	1	1		
StkCompen_d				1	1	1	1	1		1		1	1	1	1	
Stk_ceo		1		1						1		1	1	1		
Cash_compen								1	1	1					1	
CEOChair_d								1	1	1				1	1	1

Appendix 4I

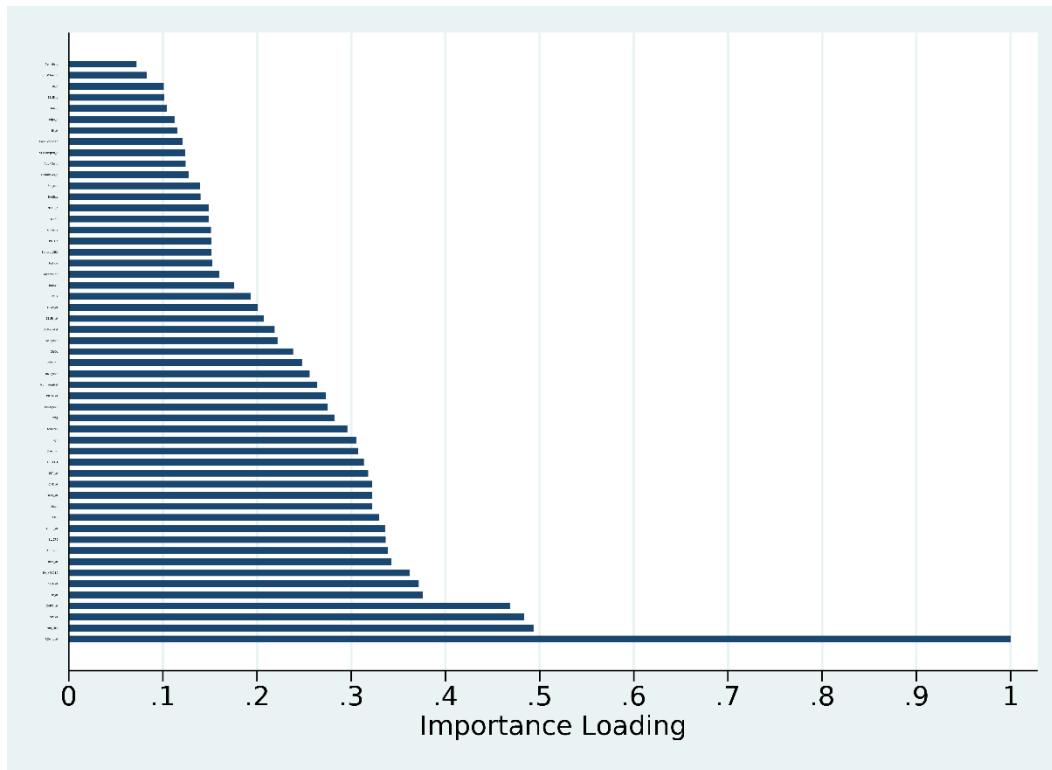
Appendix 4I.1 Random Forest analysis for annual cash ETR (further analysis)

This figure plots the results of Random Forest analysis for annual cash ETR in Step 2, including corporate governance variables. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



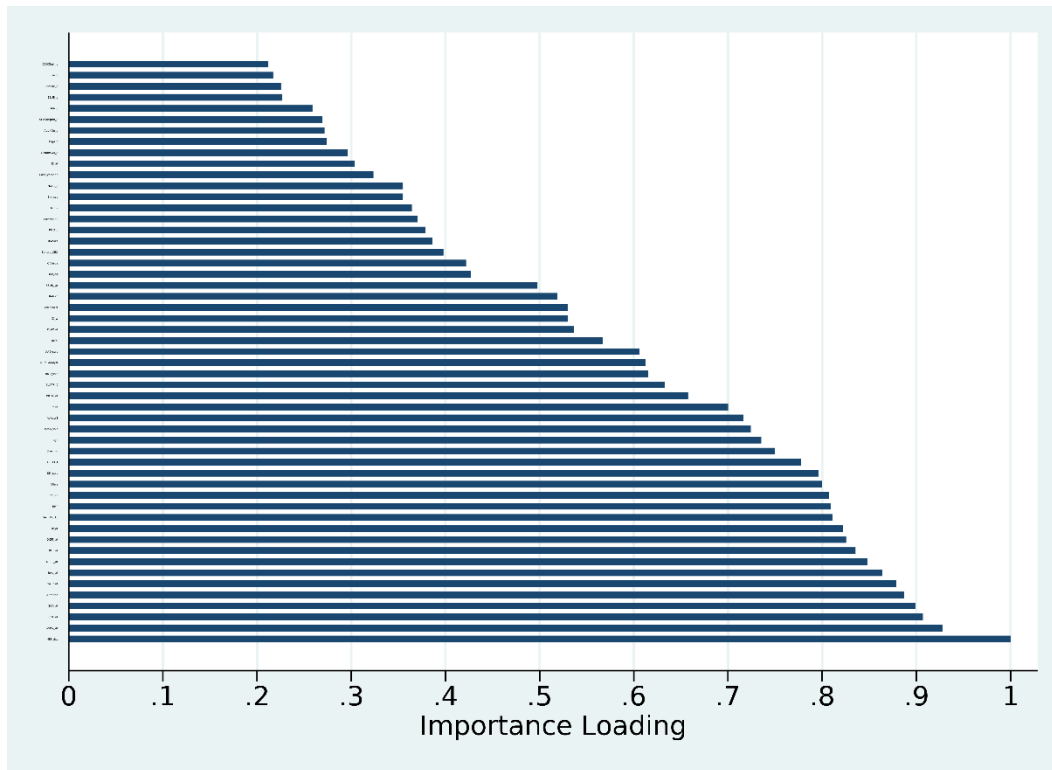
Appendix 4I.2 Random Forest analysis for annual GAAP ETR (further analysis)

This figure plots the results of Random Forest analysis for annual GAAP ETR in Step 2, including corporate governance variables. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



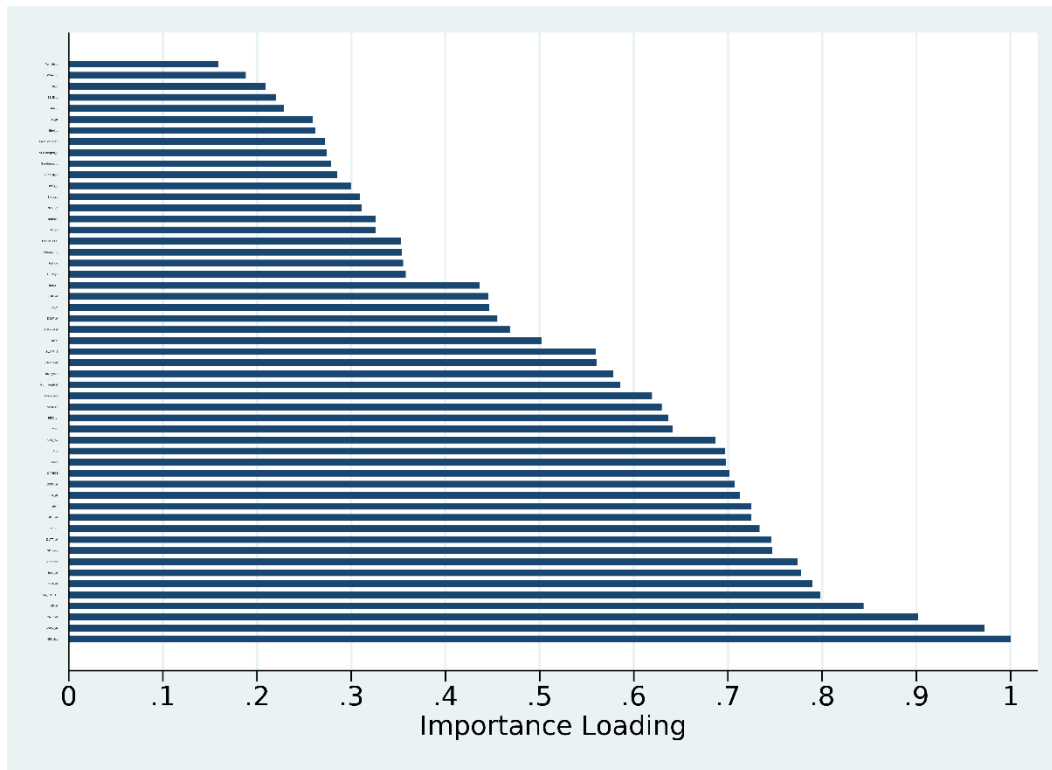
Appendix 4I.3 Random Forest analysis for long-run cash ETR (further analysis)

This figure plots the results of Random Forest analysis for long-run cash ETR in Step 2, including corporate governance variables. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



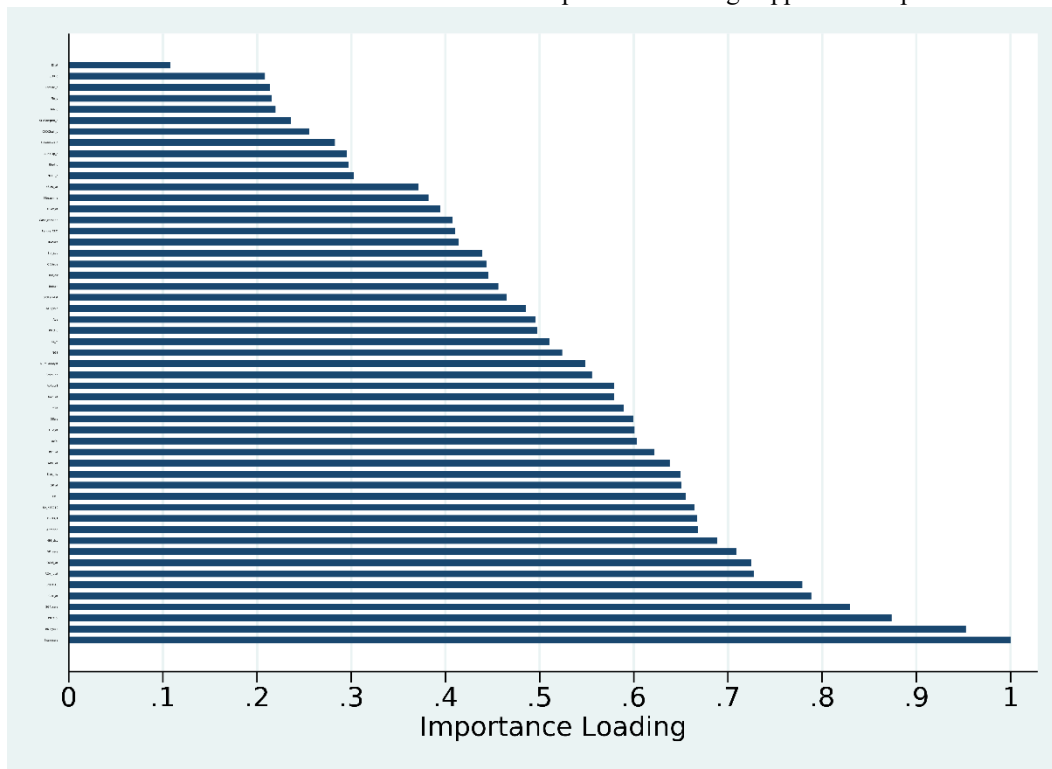
Appendix 4I.4 Random Forest analysis for long-run GAAP ETR (further analysis)

This figure plots the results of Random Forest analysis for long-run GAAP ETR in Step 2, including corporate governance variables. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



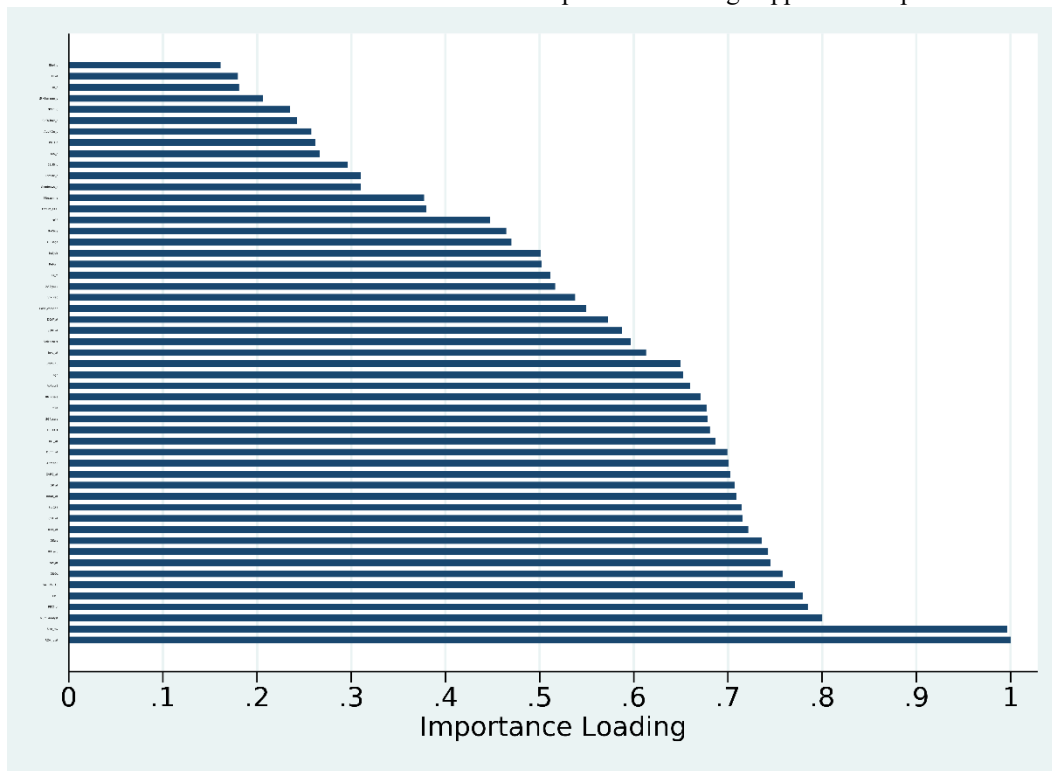
Appendix 4I.5 Random Forest analysis for UTB (further analysis)

This figure plots the results of Random Forest analysis for UTB in Step 2, including corporate governance variables. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



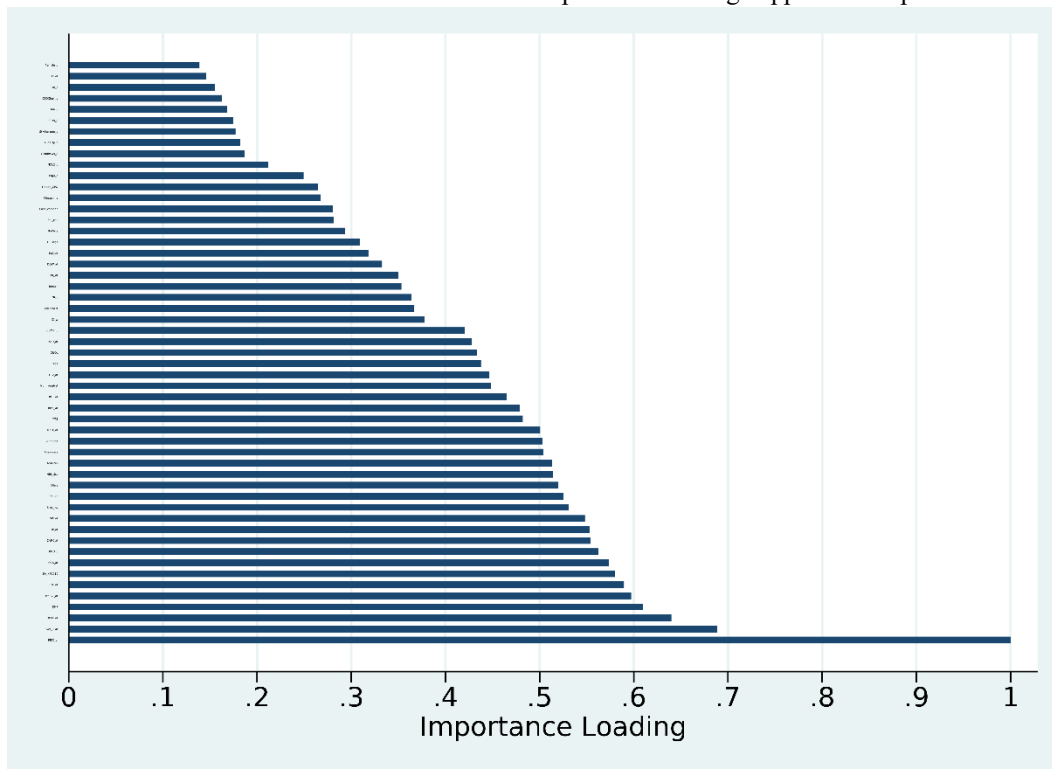
Appendix 4I.6 Random Forest analysis for DTAX (further analysis)

This figure plots the results of Random Forest analysis for DTAX in Step 2, including corporate governance variables. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



Appendix 4I.7 Random Forest analysis on DDBTD (further analysis)

This figure plots the results of Random Forest analysis for DDBTD in Step 2, including corporate governance variables. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.



Appendix 4I.8 Random Forest analysis on MPBTD (further analysis)

This figure plots the results of Random Forest analysis for MPBTD in Step 2, including corporate governance variables. The variables are ranked based on their importance loading. Appendix 4A provides variable definitions.

