

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.