Machines in markets: The impacts of technology on stock valuation and trading

Bao Linh Do

Ph.D. supervisors:

Prof. Tālis Putniņš

Assoc. Prof. Vinay Patel

Prof. David Michayluk

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Finance Discipline Group, UTS Business School

University of Technology Sydney

28th Jun 2023

Certificate of original authorship

I, Bao Linh Do, declare that this thesis, titled "Machines in markets: The impacts of technology on stock valuation and trading", is submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy in the UTS Business School 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:

Bao Linh Do

Date: 28th Jun 2023

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Tālis Putniņš, for his insightful comments and timely suggestions at every stage of the research projects. His immense knowledge and prolific experience in the profession have motivated me all the time in both research and daily life. I am particularly grateful for his unwavering support and belief in a junior researcher like me.

I also want to thank Christina Sklibosios for her helpful advice during my candidature. I would like to extend my thanks to many UTS Finance Discipline academic staff including Vinay Patel, Kenny Phua, Leo Liu, Vitali Alekseev, Marco Navone, Kristoffer Glover, Harry Scheule, David Michayluk, Nihad Aliyev for all their suggestions to improve the projects. I am also thankful to other UTS Finance Discipline staff, including Andrea Greer, Duncan Ford, Chloe Qu, and Stephanie Ough, for their administrative support.

I am also grateful to the Centre for Research in Securities Price, Refinitiv (Thomson Reuters), and Public Access to Court Electronic Records for providing access to the data used in this thesis. The thesis also benefitted greatly from the shared data by Ville Rantala, Andrew Chen, Leonard Kostovetsky, and Rui Dai.

I would like to thank the academics at the conference presentation including Xing Liu, Jingyu He, Georgina Ge, Zhuo Zhong, Carol Alexander, Haiying Yin, Valerie Laturnus, Chanyuan Ge, Hannah Nguyen for many helpful suggestions.

I am also grateful to my honours supervisor, Ivy Zhou, for the understanding and helpful advice for her old student. I cannot forget to thank my fellow Ph.D. students Niki Li, Chung Mai, Anirudh Dhawan, Mai Luong, Huong Nguyen, Atiqur Rasel, Man Nguyen, Prateek Samuel, Khanh Nguyen, Duy Nguyen, Tra Nguyen, Jason Liu, Yifeng Xu, Shu Yang, Xinyi Deng, Justin Hitchen, Sehatpour Mohammadhadi, Madeline Combe, Inji Allahverdiyeva, and other members of “UTS Finance PhD Gang” for the laughs and sweet memories at UTS. My appreciation also goes out to my precious friends Oliver Chi, Vu Trung Nguyen, Huy Ta, Mai Nguyen, Michelle Phan, Trang Nguyen, Minh Nguyen, and my housemates for their thoughtful support and tremendous understanding during the journey.

Lastly, I deeply thank my parents, little brother, and the rest of the family, whom without this would not have been possible.

Preface

Chapters 2 to 4 are structured as working papers and have been presented or have been selected for presentation at various academic conferences. The list of working papers and conference presentations is as follows:

1. Do, L., Putniņš, T., 2023. "What can machine learning teach us about company valuation?" (Chapter 2)
    - Vietnam Symposium in Banking and Finance in 2021
2. Do, L., Putniņš, T., 2023. "Nonlinear market efficiency" (Chapter 3)
    - 12th Financial Markets and Corporate Governance Conference in 2021
    - FMA Annual Meeting 2022
    - 11th FIRN Annual Conference and Ph.D. Symposium in 2022
    - Asian Finance Association Conference in 2023
3. Do, L., Putniņš, T., 2023. "Detecting layering and spoofing in markets" (Chapter 4)
    - 3rd Boca Corporate Finance and Governance Conference in 2022
    - 35th Australian Finance and Banking Conference 2022

Table of contents

Certificate of original authorship i
Acknowledgements ii
Preface iii
Table of contents iv
List of Tables vii
List of Figures ix
List of Abbreviations x
Abstract xi
Chapter 1: Introduction 1
1.1. Company valuation 2
1.2. Evolution of market efficiency definitions 4
1.3. Layering and spoofing in markets 6
1.4. Thesis outline 7
Chapter 2: What can machine learning teach us about company valuation? 8
2.1. Introduction 8
2.2. Overview of valuation theory 11
2.3. Data and methods 14
2.3.1. Data 14
2.3.2. Boosted tree model 18
2.3.3. Parameter tuning and validation test 18
2.3.4. Partial dependence plot 19
2.3.5. Shapley-based R^2 decomposition 19
2.4. Linear regression results 22
2.5. Boosted tree results 26
2.5.1. Improvements in out-of-sample predictions 26
2.5.2. Variable interactions 27
2.5.3. Out-of-sample R^2 decomposition 34
2.6. Comparison with other peer valuation methods 40
2.7. Conclusion 41

Chapter 3: Nonlinear market efficiency 42
3.1. Introduction 42
3.2. Conceptual framework 45
3.3. Data and Methods 47
3.3.1. Data 47
3.3.2. Inverse proxies of market efficiency 48
3.3.3. Linear regression 49
3.3.4. Feed-forward neural network 50
3.3.5. Information sets 52
3.4. Portfolio forecasts and asset pricing tests 54
3.5. Non-linear inefficiency 59
3.6. Drivers of the increase in non-linear efficiency 68
3.7. Conclusion 71
Appendix 3.A: Firm characteristics 72
Appendix 3.B: Accounting Variables 83
Appendix 3.C: Quantitative keywords 84
Appendix 3.D: Time-series tests of portfolios 85
Chapter 4: Detecting layering and spoofing in markets 93
4.1. Introduction 93
4.2. What are layering and spoofing? 98
4.3. Characteristics of layering and spoofing in prosecution cases 100
4.3.1. Overview of the prosecution cases 101
4.3.2. Key features of the prosecution cases 104
4.3.3. Anatomy of a spoofing case 108
4.3.4. Layering and spoofing characteristics 118
4.4. Data and metrics definition 126
4.4.1. Data 126
4.4.2. Metrics 126
4.4.3. Using the Intraday Metrics to Detect Spoofing 134
4.4.4. Using Daily Metrics to Detect Spoofing 136

4.5. Machine learning models for detection and out-of-sample validation.....138
4.6. Conclusion 147
Chapter 5: Conclusion.....148
5.1. What can machine learning models teach about the drivers of company value?.....148
5.2. How does machine learning impact market efficiency? 149
5.3. How to detect layering and spoofing in markets? 149
5.4. Future research direction 150
References.....152

List of Tables

Table 2.1 16
Table 2.2 19
Table 2.3 23
Table 2.4 24
Table 2.5 25
Table 2.6 26
Table 2.7 27
Table 2.8 34
Table 2.9 39
Table 2.10 40
Table 3.1 54
Table 3.2 55
Table 3.3 56
Table 3.4 57
Table 3.5 58
Table 3.6 60
Table 3.7 62
Table 3.8 70
Table 3.A1 72
Table 3.A2 83
Table 3.A3 85
Table 3.A4 86
Table 3.A5 87
Table 3.A6 88
Table 3.A7 89
Table 3.A8 90
Table 3.A9 91
Table 3.A10 92
Table 4.1 102
Table 4.2 124
Table 4.3 132

Table 4.4 133
Table 4.5 134
Table 4.6 136
Table 4.7 138
Table 4.8 145

List of Figures

Figure 2.1 28
Figure 2.2 30
Figure 2.3 31
Figure 2.4 33
Figure 2.5 35
Figure 2.6 36
Figure 2.7 37
Figure 2.8 38
Figure 3.1 46
Figure 3.2 47
Figure 3.3 63
Figure 3.4 65
Figure 3.5 66
Figure 3.6 67
Figure 4.1 110
Figure 4.2 111
Figure 4.3 112
Figure 4.4 114
Figure 4.5 116
Figure 4.6 117
Figure 4.7 140
Figure 4.8 141
Figure 4.9 143
Figure 4.10 144
Figure 4.11 146

List of Abbreviations

AI	Artificial Intelligence
AMEX	American Stock Exchange
CAPM	Capital Asset Pricing Model
CBOE	Chicago Board Options Exchange
CBOT	Chicago Board of Trade
CFTC	Commodity Futures Trading Commission
CME	Chicago Mercantile Exchange
COMEX	Commodity Exchange Inc.
CRSP	Centre for Research in Securities Prices
DOJ	Department of Justice
EDGAR	Electronic Data Gathering, Analysis, and Retrieval system
EMH	Efficient Markets Hypothesis
FCA	Financial Conduct Authority
FINRA	Financial Industry Regulatory Authority
ICE	The Intercontinental Exchange
IOM	Index and Options Market
LASSO	Least Absolute Selection and Shrinkage Operator
LSE	London Stock Exchange
NASDAQ	Nasdaq Stock Exchange
NYMEX	New York Mercantile Exchange
NYSE	New York Stock Exchange
OSC	Ontario Securities Commission
ReLU	Rectified Linear Units
SARD	Sum of Absolute Rank Differences
SEC	Securities and Exchange Commission
SESC	Securities and Exchange Surveillance Commission
TYO	The Tokyo Stock Exchange

Abstract

Technology has brought many changes to financial markets, including automation, substantially faster trading via algorithms, vast amounts of financial data in digital form, and sophisticated data analysis techniques able to learn complex non-linear relations. This thesis examines the implications and applications of technology in finance, specifically in the areas of fundamental valuation of stocks, market efficiency, and market manipulation.

The first chapter discusses the application of machine learning in company valuation. Despite being a core topic in finance/accounting, company valuation receives surprisingly little attention from empiricists. Given the non-linearities between financial variables and company value, machine learning is particularly well-suited to empirically characterizing what drives company value. The chapter uses a tree-based model to not impose a functional form on the relationships while retaining the interpretability of the drivers of company value. The results demonstrate that treating financial variables in isolation and with a linear approach is not a sound valuation practice. Interactions among firm fundamentals play a large role in predicted value – more than 50% of out-of-sample predictability is attributed to variable interactions. Certain interactions, such as the interplay between growth and risk, dividend payout and growth, and reinvestment rate and growth are important in accurately valuing companies and have a sound conceptual basis. Our results also indicate the importance of profitability in company value, as it is the primary driver of price-to-book and enterprise-value-to-invested capital. We find that performance improves after accounting for peer dummies based on analyst coverage in the out-of-sample setting.

The second chapter explores a new dimension of market efficiency. In Eugene Fama's original Efficient Markets Hypothesis (EMH), the different degrees of efficiency are defined by different information sets (e.g., stock prices, public information, private information). We propose an orthogonal dimension – for a given information set (e.g., all public information), how well do stock prices incorporate increasingly complex (e.g., non-linear vs linear) combinations of the information. We test this new concept by utilizing machine learning to generate stock return predictions using non-linear combinations of information and contrast the capability of predicting returns to the ability of linear models. The performance difference between these two is our new measure of "non-linear market efficiency". The idea is that when the difference is large, markets are not doing a good job of reflecting the non-linear relations

between public information, whereas when the difference narrows, it suggests markets are becoming more efficient, not with respect to the *type* of information that they incorporate, but rather, in how they *combine* information. Our findings suggest that overall, using models that capture non-linearities between information substantially improves return predictability, supporting the need to augment the EMH with this second dimension of market efficiency. We show that the predictability difference decays over time. We attribute this increase in non-linear market efficiency to improvements in technology and growth in the numbers of quantitative mutual funds that are likely to be using machine learning models.

The third chapter shows that as the market evolves, new types of market manipulation enabled by algorithms appear. One example is layering and spoofing, which refers to the use of non-bona-fide orders in a market to create a false impression of buying or selling interest, thereby pushing market prices and causing a better execution price on a bona-fide order from the same trader. Such manipulation is enabled by algorithmic trading because it usually requires many orders and cancellations in quick succession to be sufficiently profitable. It sees a rapid increase in the number of prosecution cases in recent years, showing increased prevalence or increased regulatory interest. Using a global sample of hand-collected data from prosecuted cases, we develop empirical metrics to detect layering and spoofing and test their accuracy using out-of-sample cross-validation. Our results suggest that the most important variables to predict intraday spoofing are order imbalance, high quoting activity, and trades occurring on the opposite side of the high quoting activity. Given the complex interactions between the various characteristics of spoofing strategies, we also employ a random forest and boosted tree classification model to predict spoofing at the one-second frequency. Machine learning proves to be a more effective method of prediction for spoofing, thanks to its capability to consider interactions between variables.

Overall, this thesis contributes to the literature by using advances in data science techniques to shed new light on core topics in finance – revealing nonlinearity in company valuation, developing a new market efficiency dimension, and building a detection model for a new, algorithmic type of market manipulation.

Chapter 1: Introduction

Technology is disrupting finance like many other industries. In some instances, for the better, in others, with detrimental effects.

Financial markets in the 17th century were unsophisticated and involved manual processes. Transactions were typically conducted in cash, with buyers exchanging coins or bills for paper certificates. There was little regulation and buyers had to rely on their judgment and experience to determine value. Trading was largely conducted in physical locations, such as trading floors. The trading floor was a chaotic environment, with traders shouting and gesturing to execute trades. The absence of electronic markets also meant that traders had to rely on physical market data, such as newspapers and market reports, to stay informed about market conditions. This information was often slow to arrive and could be unreliable.

Over time, markets integrated many waves of technological innovations. Governments began to regulate markets, imposing rules on trading mechanisms and quality standards. Markets also began to specialize in what types of financial products they list and trade. The development of electronic markets made trading faster, more efficient, and more accessible to a wider range of participants. The more recent replacement of humans with algorithms for many investment decisions and trade execution has changed market dynamics. So too has the growing stream of real-time digital information and the data science methods available to analyze the data.

In electronic markets, trading is conducted electronically, typically through computer networks, eliminating the need for a physical presence on a trading floor. These systems allow trades to execute in milliseconds. Electronic markets enable automation and algorithmic trading—traders program their strategies into computer systems that execute trades automatically based on predefined rules. Digitization of information has enabled markets to leverage advanced data science technologies like machine learning and artificial intelligence to improve investment and trading decisions. The AI algorithms can analyze data, predict market trends, and execute trades automatically, often without human intervention. While electronic markets were a significant step forward, the integration of advanced technologies like AI has taken the financial market to a new level.

This thesis examines both implications and applications of technology in finance, particularly the valuation of companies, market efficiency, and market manipulation.

The main research contributions are discussed in Chapters 2 to 4. In Chapter 2, we deploy the foundation of machine learning techniques to uncover fresh insights about valuation of companies by capturing intricate and non-linear relationships within the data. Chapter 3 explores the mechanism of machine learning on market efficiency and the nature of the information that is reflected in prices. In Chapter 4, the automation and scalability of market manipulation techniques, such as layering and spoofing, facilitated by technology applications in trading, are examined as an undesirable aspect. Additionally, the chapter explores methods for detecting and identifying such manipulative trading practices in financial markets.

The remainder of this chapter provides general background and reviews the literature to which this thesis contributes.

1.1. Company valuation

Valuing a company is one of the core tasks performed by finance practitioners and therefore a central topic in the field of finance. Graham and Dodd (1940) offer one of the initial approaches to fundamental analysis, introducing the term "security analysis" to refer to the examination of available facts with the aim of making conclusions about a company's prospects. According to the neoclassical model of security valuation, the present value of future cash flows (also known as discounted cash flows, DCF) determines a company's value by summing all the future cash flows, with adjustment for risk and the time. The key challenge is that this valuation model requires practitioners to make many assumptions regarding the cash flows, the timing, growth, and risk of the cash flows. These assumptions and the approaches to obtaining these parameters are what make valuation an "art" as well as a science and lead to divergence and inconsistency among practitioners.

A key source of information about company fundamentals is financial statements. Numerous studies examine the extent to which financial statement components facilitate better investment decisions for investors (e.g., Foster, Olsen, and Shevlin, 1984; Ou and Penman, 1989; Lev and Thiagarajan, 1993; Abarbanell and Bushee, 1998; Piotroski, 2000). However, the optimal way to combine the information in financial statements is still an open question that is not resolved in the literature. One strand of literature discusses cash flow discount models, including Ohlson (1995), Burgstahler and Dichev (1997), Penman (1998), Piotroski (2000), and Nissim and Penman (2001). According to Ou and Penman (1989), financial statement elements have the potential to predict forthcoming profits and establishing a lucrative trading plan. This involves making purchasing or selling decisions regarding stocks based on the direction of earnings expansion. Bartram and Grinblatt (2018) investigate a combination of

multiple financial statement variables in a linear fashion. Bartram and Grinblatt (2018) illustrate a profitable trading strategy using the mispricing between predicted equity value and actual equity value. They show the benefit of "kitchen-sink" linear regression in predicting company value.

A body of literature focuses on relative valuation, which aims at finding a company's peers (comparable companies) and then seeks to value one company using the observed value of the peer group. Mukhlynina and Nyborg (2020) review valuation practices and show that the most popular valuation method is "valuation by multiples". Valuation using multiples can be easier to apply than discount cash flow (DCF) model and can avoid having to estimate tricky parameters such as the future growth rate or terminal value. Liu, Nissim, and Thomas (2002) suggest that valuation ratios are beneficial for forecasting company value. Bhojraj and Lee (2002) identify peer firms as those with the closest predicted values of multiples. Similarly, Knudsen, Kold, and Plenborg (2017) identify peers by using the absolute distance between the company fundamentals. Other studies use information in company reports or analyst coverage, such as Hoberg and Phillips (2016) and Kaustia and Rantala (2021). These studies indicate that the amount of information that drives company value can be countless. The challenge is to minimize the mispricing gap between the predicted value and fair value.

Although in theory the key fundamentals that drive value (cash flows, risk, growth, timing) should have non-linear relations with value and interact in complex ways, relatively few empirical studies use models capable of capturing these non-linear functional forms. Exceptions include the following. Burgstahler and Dichev (1997) find that the earnings and book value of equity of a company both have an impact on its equity value, and this relationship exhibits a convex pattern. Additionally, earnings and book value of equity jointly determine the value of equity in a non-linear manner. Collins, Pincus, and Xie (1999) reveal the heterogeneity in the relationship between earnings and price across firms with positive and negative earnings. Including book value of equity, they can explain part of this discrepancy. Both papers use linear regression with interaction terms.

If we include more financial statement variables as in Bartram and Grinblatt (2018), the linear model cannot capture the complicated variable interactions. In the second chapter of this thesis, a tree-based model is used to estimate the relationship between company value and financial statement variables. This approach offers the advantage of capturing more complex interactions among variables beyond simple pairwise relationships, as well as accommodating potential non-linearities. This task is only possible by using a machine learning model such as boosted regression tree. While non-linearities are important theoretically as implied by the DCF

model and are documented in earlier literature, they are not well studied due to the limitation of linear methods until recently. For example, except for the interactions between the book value of equity and earnings, there is no empirical evidence on other financial statement variables while they are also important in company valuation. We use some explainable machine learning techniques such as Shapley values and partial dependence plots to get more insights into the relationship between company value and value drivers. We decompose the unique contribution to R^2 of each value driver. After considering unique variable contributions, the rest of the R^2 is attributed to complex variables interactions. We find that the component of R^2 attributed to variables interactions is large compared to the independent contribution of each value driver. Variable interactions are the second most important component in explaining equity multiple and the most important variables in explaining enterprise multiple. We document some key interactions such as the interaction between growth and risk, growth and dividend, and growth and reinvestment rate. These non-linearities are probably unknown to valuation practitioners.

We also uncover the value of information that is not directly observed by econometricians. We add new variables, which are peer dummies (equal to one if the two firms are peers and zero otherwise). Peers are determined in Kaustia and Rantala (2021) as the firms covered by the same analysts. We find that integrating professionals' views on the economic link between firms improves the prediction of company value.

1.2. Evolution of market efficiency definitions

The concept of market efficiency, developed by Fama (1970), posits that stock prices reflect all available information, making it impossible to consistently outperform the market through investment strategies that rely on that information. He classifies market efficiency into three forms that stock prices may impact upon: weak, semi-strong, and strong, based on the extent to which stock prices reflect stock market information, all public information, and all public and private information, respectively. In other words, Fama defines efficiency with respect to one dimension, being the breadth of the information set that is reflected in prices. A vast body of empirical asset pricing research tests the concept of market efficiency, although the debate is still somewhat unsettled, as illustrated by the 2013 Nobel Prize awarded for conflicting views on the issue of informational efficiency. Prior studies find many asset-pricing anomalies. For example, the number of anomalies detected in empirical studies grow to more than 300 return predictors (Hou, Xue, and Zhang, 2020; Chen and Zimmermann, 2021). But even the anomalies may not be evidence against efficiency as they could simply reflect

deficiencies in the underlying asset pricing models (known as the "joint hypothesis problem") or be the result of data snooping and publication bias.

Stambaugh and Yuan (2017), Green, Hand, and Zhang (2017), and Kozak, Nagel, and Santosh (2020) show that a (linear) combination of signals from a wide range of anomalies achieves less noisy measures for stock mispricing than using any separate return signal. Some other papers are directly connected to the work of return prediction based on asset-pricing anomalies, such as Granger (1992), Lo (2004, 2012), and Daniel and Titman (1999). Campbell and Yogo (2006) demonstrate that detecting predictability becomes more difficult without the diligent application of efficient statistical tests. Gu, Kelly, and Xiu (2020) find that machine learning generates significant profits using a combination of return prediction signals. Barbopoulos et al. (2021) show that increase in information access by cloud computing leads to improvements in market efficiency.

Many papers show that anomalies disappear, possibly for two reasons: they reflect statistical artifacts, or they are arbitraged away. Some papers attribute disappearing anomalies to arbitragers (McLean and Pontiff, 2016; Jons and Pomorski, 2013; Falck, Rej, and Thesmar, 2022, Martin and Nagel, 2022). Indeed, Hou et al. (2020) discover that many anomalies do not persist across various sample periods.

The third chapter builds on this literature about market efficiency and return predictability. It introduces a second dimension to the EMH, being the complexity of the functional forms that link information and stock prices. We argue this new dimension is largely orthogonal to the original dimension of information sets. For a given information set, say for example all public information, the information could be separately and linearly reflected in prices, or non-linear transformations of the information and interactions between the different pieces of information could be reflected in prices. This additional dimension is important in capturing the impacts of advances in data science methods and their use in markets. In general, our findings demonstrate that non-linear machine learning models outperform linear regression models in terms of the Sharpe ratio. But more importantly, we show how the degree of "non-linear market efficiency" (the performance difference between non-linear and linear return prediction models) increases through time with the implementation of machine learning in investment decision-making.

This research also shares the same motivations as Karapandza and Mazin (2014) and Rösch, Subrahmanyam, and Van Dijk (2017), who discuss market efficiency as a relative and time-varying term. We have observed a decline in the superiority of non-linear models compared to linear models in predicting returns over time. This emphasizes how advanced data

modeling techniques are influencing the complication of information embedded in stock prices and the efficiency of the market.

1.3.Layering and spoofing in markets

Market manipulation evolves alongside innovations in financial markets. Putniņš (2020) reviews different types of market manipulation, from techniques that have a long history to the more recent type of manipulation. Certain tactics of market manipulation have been employed for an extended period to artificially impact the price of a security or commodity. Examples include cornering the market and engaging in market squeezes. Jarrow (1992) and Cherian and Jarrow (1995) explore trading strategies used by large traders with market power. Allen, Litov, and Mei (2006) examine corners and squeezes, while Merrick Jr, Naik, and Yadav (2005) model the differences in the settlement between the spot and futures market, leading to favorable conditions for squeezes.

Additional forms of market manipulation encompass various techniques. These include trade-based methods like wash trading and closing price manipulation, information-based strategies such as pump-and-dump, and order-based tactics like quote stuffing and layering. The manipulators may be individuals or groups, sophisticated or unsophisticated backgrounds, in the same or across multiple markets, seeking to profit from their ability to influence the price of a particular security or commodity. Empirical studies on these types of market manipulation are limited. Comerton-Forde and Putniņš (2011, 2014) discuss the prevalence and measures of closing price manipulation. Washing trading is prominent in cryptocurrency exchanges due to limited regulation (Pennec, Fiedler, and Ante, 2021; Cong et al., 2022). Dhawan and Putniņš (2022) examine the impact of pervasive pump-and-dump manipulation schemes on trading volumes and prices within cryptocurrency markets.

The rise of algorithmic trading enables new forms of manipulation and allows some existing forms of manipulation to be substantially scaled up. Those include layering and spoofing. In these tactics of manipulation, a trader executes one or multiple substantial orders on a particular side of the market with the intention of deceiving others into perceiving artificial demand or supply. These false orders are intended to influence other traders and push the market in a particular direction, resulting in the execution of an order at a favorable price. After that, the manipulator cancels the false orders and possibly repeats the pattern in the opposite direction. This tactic is easier to execute using algorithms that can place and cancel orders rapidly, allowing the profitable trade cycle to be repeated many times. Spoofing is examined theoretically by Cartea, Jaimungal, and Wang (2020) and Williams and Skrzypacz (2021), with

Cartea et al. (2020) demonstrating how a trader wanting to sell can achieve a better sale price by using spoofing orders on the buy side of the limit order book. Empirical studies on spoofing, such as Lee, Eom, and Park (2013) and Brogaard, Li, and Yang (2022), examine this type of manipulation based on predefined sets of characteristics and examine the effects on market quality.

The fourth chapter in this thesis extends the research on layering and spoofing by building a comprehensive set of empirical metrics that can be used to detect the presence of layering and spoofing in markets. The approach draws on hand-collected data from prosecuted manipulation cases from around the world. The chapter proposes a detection model for layering and spoofing to help with building surveillance systems and facilitate manipulation prosecution.

1.4. Thesis outline

The remainder of this thesis comprises three studies on the following topics:

i. What does machine learning teach us about company valuation? (Chapter 2);

ii. Non-linear market efficiency (Chapter 3); and

iii. Detection of spoofing and layering in markets (Chapter 4).

Chapter 5 summarizes the findings of the whole thesis and suggests avenues for future research.

Chapter 2: What can machine learning teach us about company valuation?

2.1. Introduction

Company valuation is considered both a science and an art. It is a 'science' because valuation is based on normative theories, such as the dividend discount model or discounted cash flow model. It is an 'art' because it is subject to a lot of assumptions by valuation practitioners. Getting valuations 'right' is crucial for the efficient allocation of resources in an economy. It also has a major bearing on the returns of investment portfolios and risk/uncertainty in markets.

While valuation theories such as discount cash flows are relatively well accepted, they suggest that drivers of value interact in complex ways—many factors jointly determine free cash flows, which in turn interact with growth and risk in non-linear ways. In contrast, there are relatively few empirical studies that validate valuation models and explore their shortcomings. This research aims to address this issue and extend knowledge about company valuation by using machine learning to estimate the relationship between company value and company fundamentals.

Our data-driven approach is guided by valuation theories. We choose the relevant financial statement variables based on valuation theories, estimate the relationship between these variables and company value, and use the estimated model to predict value out-of-sample. First, we look at how relevant company fundamentals explain company value using linear regression—the baseline model used in practice. Then, we use the boosted tree, a machine learning model that has a tree-split structure. The tree-based model allows for complex variable interactions. A boosted tree model is built from many single trees; each tree improves from the previous tree in each step. Unlike linear regression with coefficients which demonstrate the linear relationship between independent and independent variables, boosted tree does not assume any pre-knowledge of the relationship between the value drivers and company value. We let the model tell us what relationships are supported by the data.

We find that the boosted tree model performs significantly better than linear regression models in out-of-sample tests, which proves that the tree model is a preferable approach for valuing companies compared to linear regression. We identify the contribution of each variable to the predicted value by both linear regression and the boosted tree. We decompose the part of R^2 that is explained by unique value driver vs the part that is explained by variable

interactions. We find that variable interactions are the most important component in explaining equity multiple and the third most important component in explaining enterprise multiple.

To further explore which variable interactions the tree suggests, we use partial dependence plots, which show the average change in predicted value with respect to change in one value driver, at every level of the other value driver. These techniques allow us to look inside the 'black box' of machine learning techniques and gain economic insights about the drivers of company value.

The most notable interactions between the drivers and the price-to-book ratio are between the growth rate and risk, and the growth rate and the dividend payout ratio. We find that the price-to-book ratio is positively related to the growth rate. However, at every level of growth rate, price-to-book decreases when the beta increases. We also find that the dividend payout is not always positively related to firm value; the positive effect only begins when the dividend payout is in the top quintile cross-sectionally. The price-to-book is highest when both dividend payout and long-term growth are at their highest level.

We also find that long-term growth is valued more than short-term growth. The effects of growth are different for equity value and enterprise value. While the price-to-book ratio is negatively related to short-term growth only when the short-term growth is high, the price-to-book ratio is positively related to long-term growth at all levels. For both short-term and long-term growth, the relationship between growth rate and enterprise value-to-capital becomes negative when both type of growth rates are at their highest levels. As enterprise value also includes the value of debt, a negative relationship between growth rate and enterprise value ratio indicates that too high growth may destroy the enterprise value of a company.

We also include peer group dummy variables in the boosted tree. The dummies are equal to one if the two firms are peers according to the definition in Kaustia and Rantala (2021), and zero otherwise. Kaustia and Rantala (2021) define peer groups based on analyst coverage. This information allows us to control for information that is not directly observed by econometricians. The performance of ordinary least squared diminishes when a large number of variables are incorporated into the model. Using boosted tree, we incorporate more variables, while allowing for the interactions between financial and other information in company valuation. We find that integrating analysts' views on the economic link between firms improves the accuracy of the empirical valuation models.

Our study mainly focuses on estimating a company's price-to-book ratio and enterprise value-to-capital because estimating these ratios is a more challenging task, having removed scale. Financial ratios are used frequently by valuation practitioners. Mukhlynina and Nyborg (2020) review valuation practices and show that valuation by multiples is the most popular

valuation method. Valuation using multiples appears to be a more straightforward approach because it avoids assumptions about growth rate, discount rate, or terminal value. However, the apparent simplicity of relative valuation masks the fact that multiples are affected in complex ways by a range of company fundamentals.

Many studies examine the value of financial statement variables in helping investors make better investment decisions (Foster, Olsen, and Shevlin, 1984; Ou and Penman, 1989; Lev and Thiagarajan, 1993; Abarbanell and Bushee, 1998; Piotroski, 2000). Careful analysis of past financial statements helps uncover information that is not yet reflected in the price. Extensive studies conducted in the field of fundamental analysis reveal that publicly accessible accounting information is not adequately reflected in security prices (Ball and Brown, 1968; Bernard and Thomas, 1989; Sloan, 1996). Ou and Penman (1989) find that future earnings are predicted by financial statements, which can be used to implement a profitable trading strategy by buying or selling stocks based on predictions of a logistic model about the direction of changes in earnings.

Although equity prices tend to display a higher level of volatility compared to underlying fundamentals, the equity market appears to be efficient at the micro level. Kritzman and Page (2003) suggest that valuation ratios are beneficial forecasting metrics. A large body of literature reports the importance of company information, such as earnings, in determining security values (Ohlson, 1995; Ohlson and Juettner-Nauroth, 2005). Accurate earnings forecasts allow investors to make more informed investment decisions and facilitate efficient capital allocation (Loh and Mian, 2006).

While the key factors that determine value, such as cash flows, risk, growth, and timing, have non-linear relationships theoretically and interact with each other in intricate ways, there are only a limited number of empirical studies that employ models capable of capturing these non-linear functional forms. Some exceptions use linear regression with interaction terms. Burgstahler and Dichev (1997) find that the equity value of a company is influenced by both its earnings and book value of equity and the relationship is complex in nature. Moreover, the equity value is also affected by the ratio of earnings to the book value of equity. Collins, Pincus, and Xie (1999) show that the relationship between earnings and price is not consistent across firms with positive and negative earnings. The inclusion of the book value of equity can help to explain this discrepancy to some extent. Skinner and Sloan (2002) find that value and growth stocks respond asymmetrically to negative and positive earnings. Bartram and Grinblatt (2018) highlight the advantages of using a comprehensive linear regression model, often referred to as a ‘kitchen-sink’ model, for predicting a company's worth.

While non-linearities are important theoretically as implied by the discount cash flow (DCF) model and are documented in earlier literature, they are not well studied due to the predominant use of linear methods until recently. Linear regression is not suitable for the task of integrating all possible interactions among relevant financial statement variables. We show that non-linearities, which are the interaction of financial statement variables, deserve careful consideration in valuation practice.

Our second contribution to the growing literature involves the application of machine learning in finance and accounting research domains. Recent papers use machine learning in asset pricing studies, such as Tobek and Hronec (2021), Gu, Kelly, and Xiu (2020), Bryzgalova, Pelger, and Zhu (2021) and Dong et al. (2022), and in financial statement analysis to forecast earnings or the magnitude of abnormal stock return, such as Cao and You (2021). We contribute to the literature that uses machine learning in finance by answering one of the most critical questions in finance: how is financial information best combined to value a company? We document the benefits of applying a tree-based model in this setting.

2.2. Overview of valuation theory

In this section, we outline classic approaches to company valuation, as normative theory provides guidance on which variables should be included in our model. We choose our inputs based on normative theory models, such as the dividend discount model and the free cash flow discount model.

According to the dividend discount model:

$$MEQ_{i,t} = \sum_{t=1}^{\infty} \frac{D_{i,t}(1 + g_{i,t})}{(1 + k_{i,t}^{e})^{t}} \tag{1}$$

in which, $MEQ_{i,t}$ is the market value of equity of the company i at time t.

$D_{i,t}$ is the dividend payment of the company i at time t.

$k_{i,t}^{e}$ is the cost of equity of the company i at time t.

$g_{i,t}$ is growth rate of the company i at time t.

Ohlson (1995) proposes clean surplus relation, which implies that the market value is equal to the book value plus the present value of future expected abnormal earnings. Abnormal earnings are driven by net income, book value, and dividends. The relationship in (2) demonstrates that the increase in book value is equal to net income minus the dividend. The relationship in (3) implies that abnormal earnings are total earnings minus 'normal' earnings, which is the product of the cost of equity and the previous book value:

$$BEQ_{i,t-1} = BEQ_{i,t} + D_{i,t} - NI_{i,t} \tag{2}$$

$$NI^a_{i,t} = NI_{i,t} - k^e_{i,t} BEQ_{i,t-1} \quad (3)$$

in which, $NI^a_{i,t}$ is the abnormal earnings of the company i at time t.

$NI_{i,t}$ is the earnings of the company i at time t.

$BEQ_{i,t-1}$ is the previous book value of equity of company i at time $t-1$.

$k^e_{i,t}$ is the cost of equity of the company i at time t.

Other information may not be reflected in the financial statements at time t but is relevant to equity market value:

$$NI^a_{i,t+1} = wNI^a_{i,t} + v_{i,t} + \varepsilon_{i,t+1} \quad (4)$$

in which, $NI^a_{i,t+1}$ is the abnormal earnings of the company i at time $t+1$.

$NI^a_{i,t}$ is the abnormal earnings of the company i at time t.

$v_{i,t}$ is the soft information not reflected in the financial statements of the company i at time t.

w is the multiplier of $NI^a_{i,t}$.

$\varepsilon_{i,t+1}$ is the unpredictable shock of company i at time t.

From equations (2) and (3), we can determine the dividend in terms of abnormal earnings and the book value of equity. The market value of equity is equal to:

$$MEQ_{i,t} = BEQ_{i,t} + \sum_{t=1}^{\infty} \frac{NI^a_{i,t}(1+g_{i,t})}{(1+k^e_{i,t})^t} \quad (5)$$

in which, $MEQ_{i,t}$ is the market value of equity of the company i at time t.

$k^e_{i,t}$ is the cost of equity of the company i at time t.

$NI^a_{i,t}$ is the abnormal earnings of the company i at time t.

$BEQ_{i,t}$ is the previous book value of the equity of the company i at time t.

$g_{i,t}$ is growth rate of the company i at time t.

From (3), (4), and (5), we can express $MEQ_{i,t}$ in terms of the book value of equity, earnings, dividend, cost of equity, and growth rate:

$$MEQ_{i,t} = BEQ_{i,t} + \sum_{t=1}^{n} \frac{(1+g_{i,t})(NI_{i,t} - k^e_{i,t}(BEQ_{i,t} + Div_{i,t} - NI_{i,t}))}{(1+k^e_{i,t})^t} \quad (6)$$

The variables' definition for equation (6) is the same as in equations (3), (4), and (5). From equation (6), the drivers of the market value of equity of company i at time t are:

1) Earnings of company i at time t.
2) Cost of equity of company i at time t.
3) Book value of equity of company i at time t.

4) Dividend of company i at time t.
5) The growth rate of company i at time t.
6) Other information of company i at time t that is not reflected in the financial statements.

If we divide both sides by book value, we can identify the drivers of the price-to-book ratio (P/B):

$$P/B_{i,t} = 1 + \sum_{t=1}^{\infty} \frac{(1 + g_{i,t})(roe_{i,t} - k_{i,t}^{e}(1 + dp_{i,t}roe_{i,t} - roe_{i,t}))}{(1 + k_{i,t}^{e})^{t}} \quad (7)$$

in which, $P/B_{i,t}$ is the price-to-book ratio of the company i at time t.

$roe_{i,t}$ is the return on equity of the company i at time t.

$dp_{i,t}$ is the dividend payout of the company i at time t.

$g_{i,t}$ is growth rate of the company i at time t.

The drivers of the equity multiple are:

1) Cost of equity of company i at time t.
2) Return on equity of company i at time t.
3) Dividend payout of company i at time t.
4) The growth rate of company i at time t.
5) Other information of company i at time t that is not reflected in the financial statements.

We also investigate the drivers of a total company valuation. Enterprise value is the total firm value netting cash out:

$$EV_{i,t} = MEQ_{i,t} + TD_{i,t} - C_{i,t} \quad (8)$$

in which, $EV_{i,t}$ is the market firm value of firm i at time t.

$MEQ_{i,t}$ is the market firm value of equity of company i at time t.

$TD_{i,t}$ is the total debt of the company i at time t.

$C_{i,t}$ is the total cash of the company i at time t.

From equations (6) and (8), we can write enterprise value in terms of the book value of equity, earnings, dividend, cost of equity, and growth rate:

$$EV_{i,t} = BEQ_{i,t} + \sum_{t=1}^{\infty} \frac{(1 + g_{i,t})(NI_{i,t} - k_{i,t}^{e}(BEQ_{i,t} + Div_{i,t} - NI_{i,t}))}{(1 + k_{i,t}^{e})^{t}} \quad (9)$$

Net income, dividend, and the book value of equity change at the rate $g_{i,t}$, which is the growth rate of the company i at time t. From equation (9), the drivers of the market enterprise value of company i at time t are:

1) Earnings of company i at time t.
2) Cost of equity of company i at time t.

3) Book value of equity of company i at time t.
4) Dividend of company i at time t.
5) The growth rate of company i at time t.
6) Total debt and cash of company i at time t.
7) Other information of company i at time t that is not reflected in the financial statements.

If we divide both sides by invested capital, we can identify the drivers of multiple EV/IC:

$$EV/IC_{i,t} = \frac{roic_{i,t}}{roe_{i,t}} + \sum_{t=1}^{\infty} \frac{(1+g_{i,t})\left(roic_{i,t} - k_{i,t}^{e}\left(\frac{roic_{i,t}}{roe_{i,t}} + dp_{i,t}roic_{i,t} - roic_{i,t}\right)\right)}{(1+k_{i,t}^{e})^{t}} + dtc_{i,t} - ctc_{i,t} \quad (10)$$

in which, $P/B_{i,t}$ is the price-to-book of company i at time t.

$roe_{i,t}$ is the return on equity of the company i at time t.

$dp_{i,t}$ is the dividend payout of the company i at time t.

$roic_{i,t}$ is the return on capital of the company i at time t.

$dtc_{i,t}$ is the debt-to-capital of the company i at time t.

$ctc_{i,t}$ is the cash-to-capital of the company i at time t.

$g_{i,t}$ is growth rate of the company i at time t

We also control for the reinvestment rate of company i and time t as it is a driver of fundamental growth. The drivers of the enterprise multiple are as follows:

1) Cost of equity of company i at time t.
2) Return on equity of company i at time t.
3) Return on capital of company i at time t.
4) Dividend payout of company i at time t.
5) The growth rate of company i at time t.
6) Debt over the capital of company i at time t.
7) Cash over the capital of company i at time t.
8) Reinvestment rate of company i at time t.
9) Other information of company i at time t that is not reflected in the financial statements.

2.3. Data and methods

2.3.1. Data

Our sample contains quarterly data from January 1990 to December 2020. The stocks are part of the Centre for Research in Securities Prices (CRSP) monthly stock file and have a positive number of common shares outstanding. We exclude stocks for companies with

negative total assets, prices smaller than $5, exchange codes not from 1 to 3, and financial stocks with the SIC codes from 6000 to 6999. Stocks have all non-missing accounting information required as inputs in the models. We extract financial statement information from Compustat Fundamentals Quarterly. A list of variables is provided in Table 2.1.

We use a one-year change in revenue and earnings per share as a proxy for growth rate and the capital asset pricing model (CAPM) beta as a proxy for risk in our model. Beta is calculated using a 36-month rolling window. We cross-sectionally rank all variables (independent and dependent variables) period by period to the range $[-1,1]$.

Our peer group data are based on analyst cross-coverage from 1986 to 2013. Each firm has 10 peers per year. Data are provided by Kaustia and Rantala (2021). The raw data is in the form of firm–peer firm for every year. One firm can have multiple peers in one year. For each year, we create a dummy that is equal to one if the firm is a peer to that peer dummy and repeat this process for every firm and every year in the sample (this process is called one-hot encoding the raw categorical data into dummies). SARD is the sum of absolute rank differences between any two companies in terms of selected fundamental variables. If the potential peer has a low SARD value, this approach suggests that the potential peer and the target company share similarities with respect to the selected variables. The StarMine peers are created through Refinitiv's exclusive algorithm, which merges competitor lists mentioned in official filings, analyst coverage, business classification, and revenue similarity.

Table 2.1
List of variables

This table provides general information about the variables that are constructed from the accounting items from the Compustat database. Variables of company i at quarter t is defined in column Constituents.

Variable name	Variable description	Constituents	Calculation
CA	Current asset	Cash and short-term investments plus other current assets	$actq_{i,t}$
NCA	Non-current asset	Other assets and property, plant and equipment	$atq_{i,t} - actq_{i,t}$
CL	Current liability	Other current liabilities and accounts payable	$dlcq_{i,t}$
NCL	Non-current liability	Other liabilities and long-term debt	$ltq_{i,t} - dltq_{i,t}$
IT	Income tax	Income tax	$txtq_{i,t}$
Dep	Depreciation	Depreciation expense	$dpq_{i,t}$
IE	Interest expense	Interest expense	$xintq_{i,t}$
Div	Dividend	Total amount of cash dividends paid for common/ordinary capital	$dvq_{i,t}$
PS	Preference stock	Prefer stock	$pstkq_{i,t}$
S	Total sale	Total sale	$saleq_{i,t}$
EDO	Extra ordinary items and discontinued operations	Extraordinary items and discontinued operations	$xidoq_{i,t}$
OE	Operating expense	Capital expenditure, general and administrative expense, depreciation expenses and cost of goods sold	$capxq_{i,t} + xsgaq_{i,t} + dpq_{i,t} + cogsq_{i,t}$
NOE	Non-operating expense	Non-operating expense	$nopiq_{i,t}$

$G1S, G1E$	One-year change in revenue and net income	One-year change in revenue and net income	$G1S = (sale_{i,t} - saleq_{i,t-4})/saleq_{i,t-4}$ $G1E = (niq_{i,t} - niq_{i,t-4})/niq_{i,t-4}$
$G5S, G5E$	Five-year change in revenue and net income	Five-year change in revenue and net income	$G5S = (sale_{i,t} - saleq_{i,t-20})/saleq_{i,t-20}$ $G5E = (niq_{i,t} - niq_{i,t-20})/niq_{i,t-20}$
$beta$	Market beta	Market beta	
roe	Return on market		$niq_{i,t}/(teqq - pstkq + txditcq)_{i,t-1}$ Or $niq_{i,t}/(ceqq + pstkq)_{i,t-1}$ Or $niq_{i,t}/(atq - dlttq - dlcq)_{i,t-1}$
$roic$	Return on invested capital		$(piq_{i,t} - nopiq_{i,t}) - txtq_{i,t}/icaptq_{i,t-1}$
dp	Dividend payout ratio		$dvq_{i,t}/niq_{i,t}$
dtc	Debt to capital ratio		$ltq_{i,t}/icaptq_{i,t}$
ctc	Cash to capital ratio		$cheq_{i,t}/icaptq_{i,t}$
$G1roe, G5roe$	One-year change and five-year change in return on equity		$G1roe = (roe_{i,t} - roe_{i,t-4})/roe_{i,t-4}$ $G5roe = (roe_{i,t} - roe_{i,t-20})/roe_{i,t-20}$

For proxy of risk, we use market beta over a 60-month window. We calculate historical growth for the one-year and five-year periods, which are proxies for short- and long-term growth, respectively. We control for growth based on earnings and revenue in regression of enterprise value and equity value on financial statement variables. We control for growth of return on equity in regression of valuation ratio on relevant value drivers. Growth of return on equity is:

$$G1roe_{i,t} = \frac{roe_{i,t} - roe_{i,t-1}}{roe_{t-1}} \tag{11}$$

$$G5roe_{i,t} = \frac{roe_{i,t} - roe_{i,t-5}}{roe_{i,t-5}} \tag{12}$$

in which, $G1roe_{i,t}$ and $G15roe_{i,t}$ is the one-year and five-year growth rate of company i at time t.

$roe_{i,t}$, $roe_{i,t-1}$, $roe_{i,t-5}$ is the return on equity of the company i at time t, $t-1$, and $t-5$.

2.3.2. Boosted tree model

The boosted tree model originates from Friedman (2001) and focuses on a stepwise approach for m from 1 to M (M is the number of trees):

$$F(X) = F_{m-1}(X) + B_m h(X, a_m) \tag{13}$$

in which, $h(X, a_m)$ is a small regression tree.

a_m is the parameters of a regression tree.

X is a vector of the input variables.

B_m is the weight assigned to a tree.

For parameter optimization, we minimize the mean squared error:

$$\{B_m, a_m\}_1^M = \min_{\{B'_m, a'_m\}_1^M} \frac{1}{N} \sum_{i=1}^{N} (y_i - F(X))^2 \tag{14}$$

2.3.3. Parameter tuning and validation test

We use a rolling window to estimate both linear and boosted tree valuation models. Specifically, we run a model on the previous 10 years' data and use the estimated model to predict company value in the next two years. We use three-fold cross-validation for parameter tuning, which means for each window, we further divide training data into three folds and find

the best parameters based on a leave-one-out validation fold. We train the boosted tree based on the following set of parameters in Table 2.2:

Table 2.2
Boosted tree parameters
This table provides the parameters that we consider in tuning the boosted tree model.

Hyper parameters	Values
Learning rate	0.1, 0.01, 0.001
Max depth	3, 4, 5, 6
Number of estimators	100, 200, 300, 400, 500

2.3.4. Partial dependence plot

We use the partial dependence plot to visually investigate the relationship between the value drivers and the company value. A partial dependence plot illustrates the impact of variable x_s on the predicted valuation of a company, showcasing the marginal effect it has in isolation:

$$\hat{y}(x_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_s, x_c) \tag{15}$$

in which, the x_s are variable(s) of interest.

x_c is the subset of all other variables in the model.

For example, we want to examine the relationship between value driver x_1 and predicted firm value $\hat{y}$ if we also have $\{x_2, x_3\} \in x_c$, so we permute x_1 values with values of x_2, x_3 in our dataset. For each x_1 value, we calculate the set of predicted firm values and then average them. We rerun this step for every possible value of x_1 in our dataset to obtain the final average as the predicted marginal relationship of x_1 and firm value implied by our model. We then plot the values of x_1 against the average predicted value of $\hat{y}$. In linear regression, this relationship is a straight line.

The same method is performed if we examine the relationship between two variables on the predicted firm value. For each combination of (x_1, x_2) values, we calculate the set of predicted firm values and then average them. In this way, we can observe how the predicted value changes with the two variables. It also enables us to observe the interaction between the two variables.

2.3.5. Shapley-based R^2 decomposition

We are also interested in how much each value driver contributes to the explained variations of the multiples. We calculate the Shapley values, a metric derived from game theory, to determine the unique contribution of each player in the game (out of the total

contribution generated by the coalition of all players in the game). The Shapley values are applied in machine learning models to explain the contribution of each input variable to the prediction outcome. The Shapley value of variable f at observation i is:

$$\varphi_i^f = \sum \frac{|S|!\,(N-|S|-1)!}{N!}\left(v_i(S \cup \{f\}) - v_i(S)\right) \tag{16}$$

in which, N is the total number of variables in the model including variable f

S is the number of variables in a subset of N

$v_i(S)$ is the prediction at observation i using variables in set S

$v_i(S \cup \{f\})$ is the prediction at observation i using variables in set S without variable f

We translate the Shapley values into the decomposition of R^2 by adopting the method suggested by Redell (2019). This technique is independent of the specific model used, allowing its application to both linear regressions and boosted trees. It serves as a means to evaluate the significance of each input variable in the model in terms of their respective R^2 components. The R^2 component associated with each variable sum to the overall model R^2. Each feature has a single statistic, which shows its contribution to the total R^2.

The in-sample R^2 may be inflated by hyperparameter tuning, overfitting, or simply by adding more variables to the model. Therefore, we use the out-of-sample R^2 as Campbell and Thompson (2008) for both linear and nonlinear models:

$$R^2_{OOS} = 1 - \frac{\frac{1}{N}\sum_{t=1}^{T}(r_t - \hat{r}_t)^2}{\frac{1}{N}\sum_{t=1}^{T}(r_t - \bar{r}_t)^2} \tag{17}$$

One of the more popular approaches involves partial R^2, which is a type of sequential testing in which a feature is added to the model. Any incremental increase in R^2 is a result of the variation explained by the new feature. A limitation of the partial R^2 approach is that when features are correlated, the sequence in which they are included in the model can introduce bias in the assigned variance explained to each feature. The Shapley value decomposition of R^2 is an order-unbiased approach. It assesses the alterations in model fit across every conceivable grouping of the model, considering all possible orders. Shapley values provide insight into how each variable influences a certain prediction. In other words, the prediction at each instance i is:

$$\hat{y}_i = \varphi_0 + \sum_{f=1}^{F} \varphi_i^f \tag{18}$$

in which, $\hat{y}_i$ is the prediction of a given instance i.

φ_0 is the average prediction across instances in a dataset.

φ_i^f is the Shapley value of variable f at observation i.

Equation (18) shows the additive property of the Shapley value. For a given instance i, model prediction is the sum of the average prediction across instances in a dataset and the feature-level attributions. If a feature exhibits non-zero effects, removing that particular feature from the model results in a decrease in model accuracy and an escalation in the variability of the residuals. Features can be prioritized based on the degree to which their exclusion amplifies residual variance, with greater increases in residual variance indicating a higher level of significance for the feature

We first need a Shapley-modified predicted value $\hat{y}_i$ without feature f:

$$\hat{y}_{i(f)} = \hat{y}_i - \varphi_i^{(f)} \tag{19}$$

For each variable f, we can compute the contribution to the total R^2 as follows:

$$R_{i,f}^2 = \frac{R_{baseline}^2 - \min\left(\frac{var_{res_{baseline}}}{var_{res_{shap}}}, 1\right) \times R_{baseline}^2}{\sum_{f=1}^{F} R_{baseline}^2 - \min\left(\frac{var_{res_{baseline}}}{var_{res_{shap}}}, 1\right) \times R_{baseline}^2} \times R_{baseline}^2 \tag{20}$$

in which, $var_{res_{baseline}}$ is the residual variance based on the original model prediction.

$var_{res_{shap}}$ is the residual variances based on the Shapley-modified predictions.

$R_{baseline}^2$ is the total R^2 of the baseline model (the original model that includes all variables).

F is all variables in the model.

The ratio $\frac{var_{res_{baseline}}}{var_{res_{shap}}}$ ranges from 0 to 1. When it is one, it means that removing a variable does not change the model's residual variance. $R_{i,f}^2$ of variable f, in this case, is zero.

We can express equation (20) as follows:

$$R^2_{i,f} = \underbrace{(1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2})}_{\text{Total } R^2} \times \underbrace{(\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_{i(f)}\right)^2 - \sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}\left(y_i - \hat{y}_{i(f)}\right)^2})}_{\text{Marginal explained variance of variable } f \text{ as part of total explain variance (percent)}} \quad (21)$$

2.4. Linear regression results

Our linear regression results are based on the Fama-MacBeth regression every quarter. We first examine the univariate regression of each value driver on the equity value or enterprise value, and then we run the 'kitchen-sink' Fama-MacBeth regression to control for all value drivers.

Table 2.3 reports the results of the univariate and multivariate regression of the market value of equity on relevant accounting information. The relevant variables that drive the equity value of the company is provided in equation (6). Our multivariate regression results show that assets have a positive effect on equity value. While equation (6) suggests that liabilities should have a negative effect on the market value of equity, non-current liability has a positive effect on equity value. According to equation (6), dividends have a negative effect on the market value of equity, conditional on the risk level. Our result shows the opposite. Multivariate regression suggests that when controlling for other variables, dividends have a positive effect on the market value of equity. We break down net income into total revenue and other expenses. Total revenue is shown to positively affect the market value of equity. However, there are mixed results for different types of expenses. Increases in interest expense and operating expense are negatively related to equity value, while the increase in depreciation expense, tax expense, and non-operating expense increases equity value. We observe opposite signs of the coefficients between univariate results and multivariate results for growth variables, expense variables, and non-current liabilities.

Table 2.3
Regression of the market value of equity on all variables

This table shows the results of the Fama-MacBeth regression of the market value of equity on accounting items. Right-hand-side variables are current asset (CA), current liability (CL), non-current asset (NCA), non-current liability (NCL), depreciation expense (Dep), dividend (Div), discontinue operation (EDO), five-year change in net income ($G5E$) and revenue ($G5S$), one-year change in net income ($G1E$) and revenue ($G1S$), income tax (IT), interest expense (IE), non-operating expense (NOE), operating expense (OE), total revenue (S), and beta ($beta$). ***, ** and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Variable	Univariate result		Multivariate result	
	Parameter estimate	t-stat	Parameter estimate	t-stat
Intercept			165.55	$(37.87)^{***}$
CA	2.75	$(107.17)^{***}$	0.87	$(47.51)^{***}$
CL	10.83	$(103.81)^{***}$	-0.39	$(-7.19)^{***}$
NCA	0.74	$(116.79)^{***}$	0.13	$(27.25)^{***}$
NCL	0.93	$(111.46)^{***}$	0.24	$(51.94)^{***}$
Dep	61.67	$(84.32)^{***}$	4.44	$(18.90)^{***}$
Div	166.62	$(91.73)^{***}$	38.67	$(35.16)^{***}$
EDO	995.18	$(13.73)^{***}$	355.76	$(16.54)^{***}$
$G5E$	0.10	$(5.50)^{***}$	0.04	$(2.91)^{***}$
$G5S$	0.13	$(2.84)^{***}$	-0.09	$(-4.15)^{***}$
$G1E$	0.54	$(11.32)^{***}$	-0.07	$(-2.09)^{**}$
$G1S$	-0.13	$(-8.28)^{***}$	-0.44	$(-57.12)^{***}$
IT	141.99	$(124.11)^{***}$	49.89	$(53.50)^{***}$
IE	103.94	$(80.71)^{***}$	-17.25	$(-56.46)^{***}$
NOE	239.64	$(64.98)^{***}$	17.57	$(18.02)^{***}$
OE	4.97	$(102.67)^{***}$	-1.01	$(-13.96)^{***}$
S	4.15	$(106.68)^{***}$	1.13	$(15.61)^{***}$
$beta$	-0.12	-0.48	0.55	$(2.65)^{***}$
R^2			68%	

We also examine how valuation ratios change with the relevant value drivers. Table 2.4 presents regression results of price-to-book (P/B) on profitability (return on equity), investment (dividend payout ratio), growth of profitability, and risk. The drivers of P/B are suggested in equation (7). Our results show that all value drivers positively relate to P/B, except for short-term growth for both univariate and multivariate regression. The linear regression results are consistent with other empirical valuation papers. Burgstahler and Dichev (1997) and Collins, Pincus, and Xia (1999) shows that book value of equity and both positive and negative earnings drive market value of equity. Hand and Landsman (2005) show that dividend is positively priced, especially for firms with low incentives to signal.

Table 2.4
Regression of P/B on value drivers

This table shows the results of the Fama-MacBeth regression of P/B on value drivers according to valuation theory. Value drivers return on equity (roe), dividend payout ratio (dp), short-term and long-term growth rate ($G1roe$ and $G5roe$), and beta ($beta$). T-statistics are reported in parentheses. ***, ** and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Variable	Parameter estimate	Parameter estimate
	Univariate result	Multivariate result
Intercept		23.81
		(100.14)***
roe	24.36	24.46
	(28.93)***	(30.25)***
dp	0.22	0.03
	(7.17)***	1.53
$G1roe$	0.07	0.11
	(9.81)***	(11.78)***
$G5roe$	0.09	0.05
	(14.00)***	(14.66)***
$beta$	0.11	0.14
	(10.45)***	(4.62)***
R^2		17%

Table 2.5 reports the results of the regression of enterprise value on relevant accounting information. From equation (9), we break down the enterprise value into many variables. As enterprise value is the total value of assets in the company (except for cash), current assets and current liabilities, non-current assets, and non-current liabilities have a positive effect on enterprise value. From equation (9), dividends are expected to have a negative effect on enterprise value, conditional on the risk level. However, after controlling for other accounting variables, our linear regression indicates that dividends have a positive effect on enterprise value. Total revenue also significantly and positively affects enterprise value, which is consistent with equation (9). Expenses are expected to have a negative effect on enterprise value. However, our results show that depreciation, income tax, and non-operating expenses are positively related to enterprise value. The beta has a significant positive slope. Most historical growth variables have a negative impact on enterprise value, except for long-term growth in earnings. The direction of coefficients of univariate and multivariate regressions are the same for most variables, except for growth and expense variables. The results imply that controlling for other company fundamentals in the valuation model can affect the relationship between key variables such as growth and the company value.

Table 2.5
Regression of enterprise value on all variables

This table shows the results of the Fama-MacBeth regression of enterprise value on accounting items. Right-hand-side variables are current asset (CA), current liability (CL), non-current asset (NCA), non-current liability (NCL), depreciation expense (Dep), dividend (Div), discontinue operation (EDO), five-year change in net income ($G5E$) and revenue ($G5S$), one-year change in net income ($G1E$) and revenue ($G1S$), income tax (IT), interest expense (IE), non-operating expense (NOE), operating expense (OE), total revenue (S), beta ($beta$) and preference stock (PS). ***, ** and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Variable	Univariate result		Multivariate result	
	Parameter estimate	t-stat	Parameter estimate	t-stat
Intercept			191.05	$(66.55)^{***}$
CA	4.67	$(134.29)^{***}$	0.57	$(33.76)^{***}$
CL	20.64	$(135.57)^{***}$	0.38	$(10.76)^{***}$
NCA	1.38	$(195.27)^{***}$	0.17	$(28.01)^{***}$
NCL	1.82	$(207.78)^{***}$	1.06	$(182.31)^{***}$
Dep	112.99	$(105.41)^{***}$	10.03	$(30.32)^{***}$
Div	282.77	$(102.56)^{***}$	41.75	$(38.70)^{***}$
EDO	863.76	$(6.54)^{***}$	141.76	$(4.23)^{***}$
$G5E$	0.12	$(4.43)^{***}$	0.05	$(3.63)^{***}$
$G5S$	0.28	$(3.39)^{***}$	-0.06	$(-2.68)^{***}$
$G1E$	0.78	$(10.93)^{***}$	-0.10	$(-2.93)^{***}$
$G1S$	0.12	$(3.05)^{***}$	-0.45	$(-51.88)^{***}$
IT	230.68	$(117.89)^{***}$	39.69	$(57.22)^{***}$
IE	205.86	$(100.05)^{***}$	-9.36	$(-18.52)^{***}$
NOE	387.98	$(66.42)^{***}$	6.44	$(7.19)^{***}$
OE	8.73	$(117.80)^{***}$	-0.69	$(-9.37)^{***}$
S	7.28	$(126.26)^{***}$	0.87	$(11.97)^{***}$
$beta$	1.78	$(4.98)^{***}$	1.03	$(66.50)^{***}$
PS	101.61	$(53.07)^{***}$	-16.01	$(-28.49)^{***}$
R^2			81%	

Table 2.6 shows the univariate and multivariate regression results of enterprise value-to-invested capital (EV/IC) on profitability (return on equity and return on capital), investment (dividend payout ratio and reinvestment rate), debt-to-capital, cash-to-capital, growth, and risk. The drivers of EV/IC are suggested in equation (10). We observe a positive relationship between most value drivers and enterprise value-to-invested capital, except for short-term growth, dividend payout, and return on capital. We observe the opposite sign of the coefficient on the dividend payout ratio between univariate and multivariate regression.

Table 2.6
Regression of EV/IC on value drivers

This table shows the results of the Fama-MacBeth regression of EV/IC on value drivers according to valuation theory. Value drivers are return on capital (roc), reinvestment rate (ri), return on equity (roe), dividend payout ratio (dp), short-term and long-term growth rate ($G1roic$ and $G5roic$), beta ($beta$), debt-to-capital (btc) and cash-to-capital (ctc). T-statistics are reported in parentheses. ***, ** and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Variable	Parameter estimate	Parameter estimate
	Univariate result	Multivariate result
Intercept		2.50
		$(24.29)^{***}$
roe	0.03	0.11
	1.05	$(7.59)^{***}$
dp	-0.02	0.07
	-1.18	$(4.71)^{***}$
$G1roic$	-0.01	-0.01
	$(-12.70)^{***}$	$(-10.45)^{***}$
$G5roic$	0.01	0.01
	$(18.54)^{***}$	$(3.05)^{***}$
$beta$	0.01	0.01
	$(1.82)^{*}$	$(2.26)^{***}$
$roic$	-2.59	-0.72
	$(-5.36)^{***}$	$(-4.96)^{***}$
ri	0.01	0.01
	$(47.06)^{***}$	$(51.61)^{***}$
dtc	1.06	0.99
	$(94.39)^{***}$	$(70.55)^{***}$
ctc	11.22	1.23
	$(27.88)^{***}$	$(3.89)^{***}$
R^2		77%

2.5. Boosted tree results

2.5.1. Improvements in out-of-sample predictions

As we observe in equations (6), (7), (9), and (10), there are many complex interactions among value drivers when determining company value. Given its tree-split nature, the boosted tree model provides a better prediction of company value compared to linear regression. We run boosted tree models on the valuation multiples as valuation multiples are more popular in practice than the unscaled company value.

Table 2.7
Predictive performance

The table shows the out-of-sample predictive ability of linear regression, boosted tree on financial information, and boosted tree on both financial information and analyst-based information. Panel A provides out-of-sample performance for the sample period from 1990 to 2013 (as analyst information is only available until 2013). The panel provides out-of-sample performance for the whole sample period from 1990 to 2020. The rows are the mean and median squared error across all rolling windows and the out-of-sample R^2.

Variable	P/B			EV/IC		
	Linear regression on financial information	Boost tree on financial information	Boost tree on financial information and analyst information	Linear regression on financial information	Boost tree on financial information	Boost tree on financial information and analyst information
Panel A: Until 2013						
Mean SE	29%	25%	21%	25%	22%	20%
Median SE	17%	14%	11%	13%	12%	10%
OOS R^2	11%	28%	38%	27%	39%	49%
Panel B: Whole sample						
Mean SE	32%	24%		27%	20%	
Median SE	23%	13%		16%	10%	
OOS R^2	13%	23%		29%	34%	

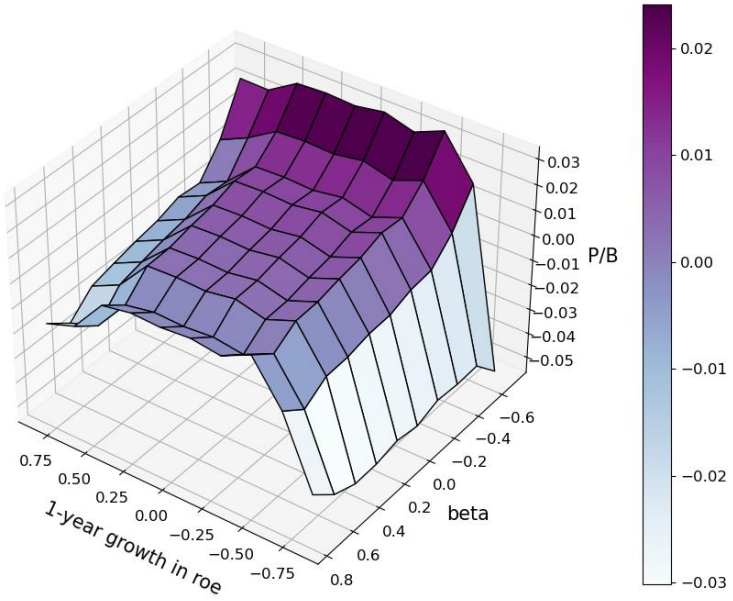
Table 2.7 reports the results of the model's out-of-sample performance. As the analyst-based information is only available until 2013, we run the models that combine the financial statement information and analyst information only until 2013. For the whole sample, the mean and median squared error for the whole period is smaller for the boosted tree, and out-of-sample R^2 is higher for the boosted tree model. The difference in performance between the two models is more pronounced for the P/B than for EV/IC. Controlling for implicit information from analysts, squared error further decreases for both valuation ratios, and out-of-sample R^2 further increases for both valuation ratios. This result shows that implicit information from analysts can also improve the valuation process. The result indicates that machine learning is doing better than linear regression in explaining valuation multiples using financial and other relevant information, by incorporating the non-linearities and interactions among them.

2.5.2. Variable interactions

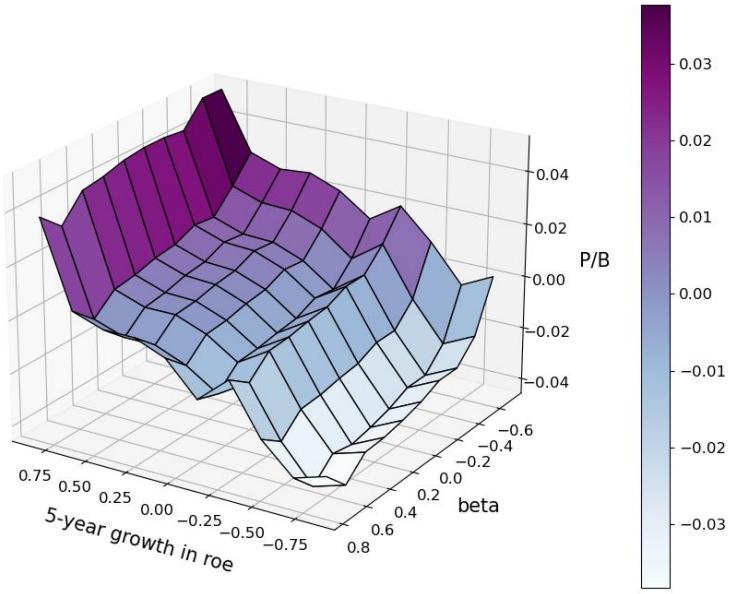
We find some interesting and important variable interactions from the boosted tree using partial dependence plots. Value is created when the short-term growth rate is moderate. For short-term growth, the value increases sharply when growth increase from the smallest level, then it does not change when growth keeps increasing. When the short-term growth rate is at the highest level, it starts destroying value. This is shown in Figure 2.1 Panel A. Unlike short-term growth, long-term growth is positively related to value at all levels, which is shown in Figure 2.1 Panel B. Figure 2.1 also shows the interaction between the one-year and five-year historical growth rate and beta.

Figure 2.1
Joint partial dependence plot between growth and risk on P/B
This figure visualizes the effect of interaction between growth and risk (controlling for other variables) on P/B.

Panel A: Effect of one-year growth and beta on P/B



Panel B: Effect of five-year growth and beta on P/B

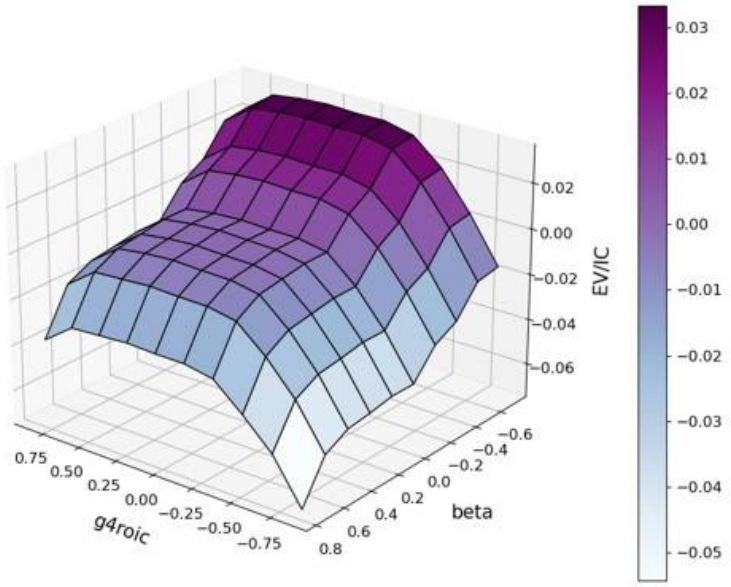


Conditional on beta, a company with the lowest beta and median short-term growth in a cross-section has the highest value. Figure 2.1 Panel B illustrates the same relationship for long-term growth; however, the preferable long-term growth is around the top quintile. High long-term growth seems to be preferable to high short-term growth. Short-term growth strategies aim to achieve immediate gains and generate quick results. The purpose of these strategies is to increase revenue and profit shortly through actions such as launching new products, ramping up marketing efforts, expanding into new markets, or leveraging existing customer relationships for additional sales. Short-term growth strategies prioritize maximizing profits in the short term. On the other hand, long-term growth strategies require patience and a focus on building something that can last for years and create a lasting impact on a business's financial performance. Such strategies often involve developing new products and services, investing in research and development, building infrastructure and personnel, forming strategic partnerships, and emphasizing customer experience. The goal of long-term growth strategies is to establish a sustainable competitive advantage that can help a business remain profitable and successful for years to come. Short-term growth strategies are generally perceived to be lower risk and require a less upfront investment. Therefore, long-term growth strategies may yield greater profits in the long run and help a business create a sustainable competitive advantage over its competitors. Both short-term and long-term growth create the most value when the beta is small.

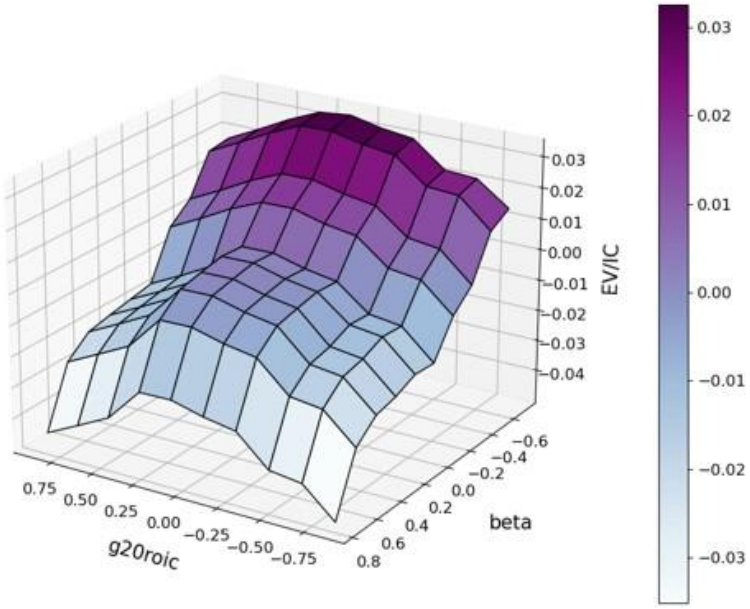
Figure 2.2 Panels A and B illustrate the interaction effect of growth rate and risk on EV/IC. For enterprise value, a company that is not too risky and has the median historical growth rate has the greatest enterprise value. There is no significant difference between short-term growth and long-term growth effect on EV/IC. Companies that experience too high levels of growth tend to possess a greater proportion of intangible assets than those with low growth rates. As a result, it can be challenging for debt holders to identify the heightened risks associated with high-growth firms. This may be the reason for high growth opportunities to be inversely correlated with debt value.

Figure 2.2
Joint partial dependence plot between growth and risk on EV/IC
This figure visualizes the effect of the interaction between growth and risk (controlling for other variables) on EV/IC.

Panel A: Effect of one-year growth and beta on EV/IC



Panel B: Effect of five-year growth and beta on EV/IC



Other notable interactions are the effect of the interactions between the growth rate and dividend payout on P/B and the interaction between the growth rate and reinvestment rate on EV/IC. Figure 2.3 Panel A shows the effect of the interaction between growth rate and dividend

payout ratio on P/B. We observe that the dividend payout ratio does not always have a positive relationship with P/B. For companies with dividend payout ratios less than the cross-sectional median, the value does not change at every level of the growth rate. For companies with a dividend payout ratio in the top quintile, the relationship between dividend payout and P/B is highly positive, especially in the case of long-term growth. If mature companies with significant long-term growth pay small dividends, their value is the smallest in the cross-section, as illustrated by Figure 2.3 Panel B.

Figure 2.3
Joint partial dependence plot between growth and dividend payout on P/B
This figure visualizes the effect of the interaction between growth and dividend payout (controlling for other variables) on P/B.

Panel A: Effect of one-year growth and dividend payout on P/B

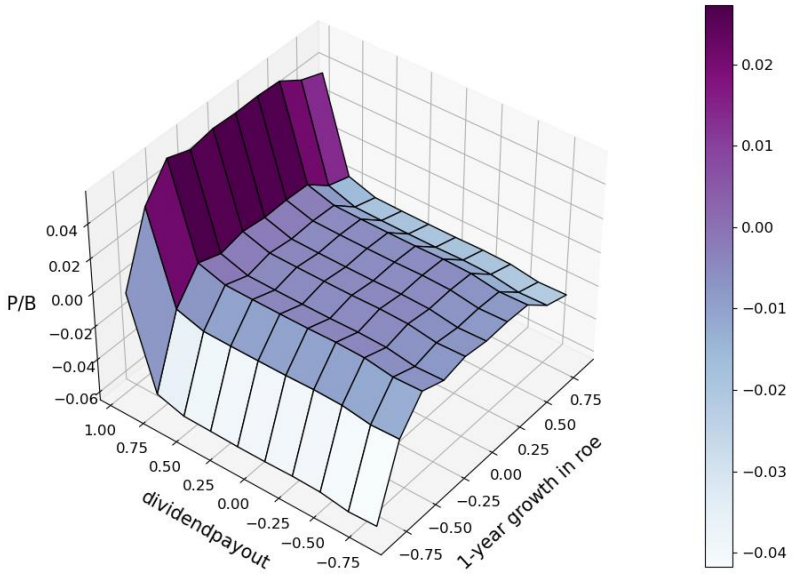


Figure 2.3 (continued)
Joint partial dependence plot between growth and dividend payout on P/B
This figure visualizes the effect of the interaction between growth and dividend payout (controlling for other variables) on P/B.

Panel B: Effect of five-year growth and dividend payout on P/B

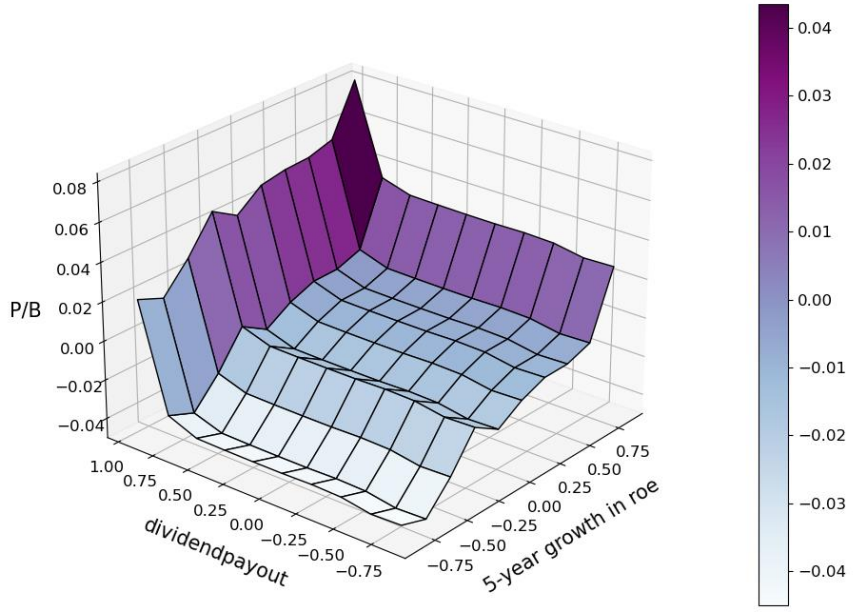
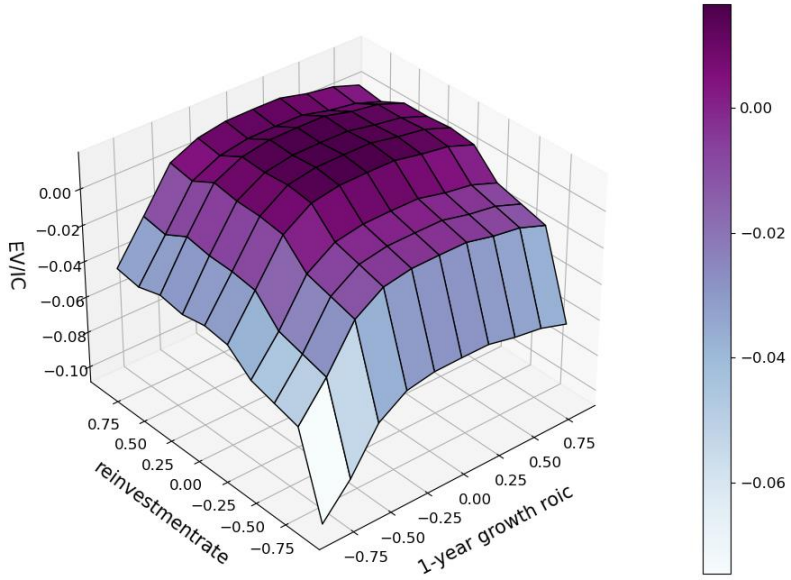


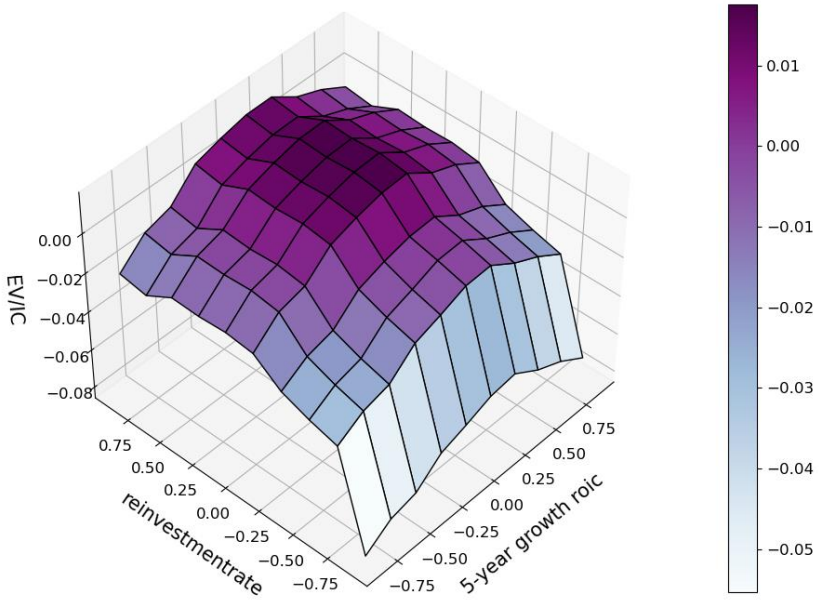
Figure 2.4 Panels A and B show the interaction effects between growth rate and reinvestment rate on EV/IC. Companies with the greatest values are those with median cross-sectional growth and median reinvestment rates. The reinvestment rate measures the amount of capital invested per unit of income.

Figure 2.4
Joint partial dependence plot between growth and reinvestment rate on EV/IC
This figure visualizes the effect of the interaction between growth and reinvestment rate (controlling for other variables) on EV/IC.

Panel A: Effect of one-year growth and reinvestment rate on EV/IC



Panel B: Effect of five-year growth and reinvestment rate on EV/IC



2.5.3. Out-of-sample R^2 decomposition

Table 2.8 shows the out-of-sample R^2 decomposition for P/B prediction for three models: linear regression with financial information, linear regression with financial information and their pair-wise interactions, and a boosted tree with financial information and analyst-based peer group (more than two-way interactions are automatically considered in the boosted tree model).

Table 2.8
Out-of-sample R^2 decomposition of P/B

The table shows the out-of-sample R^2 decompositions based on the Shapley value. The predicted variable is the P/B ratio. Panel A presents results for linear regression on financial statement information, Panel B shows the boosted tree on financial statement information and Panel C presents results for the boosted tree on financial statement information and analyst information. Variables are sorted by their level of R^2 contribution from highest to lowest.

Variable	R^2 contribution (level)	R^2 contribution (%)
Panel A: Linear regression on financial information		
roe	9.65%	89.42%
dp	0.46%	4.31%
$G1roe$	0.44%	4.05%
$G5roe$	0.15%	1.37%
$beta$	0.09%	0.85%
Total	10.80%	100%
Panel B: Linear regression on financial information and interactions		
roe	6.20%	39.14%
$Interactions$	6.12%	38.64%
dp	2.77%	17.51%
$beta$	0.30%	1.92%
$G5roe$	0.27%	1.69%
$G1roe$	0.17%	1.09%
Total	15.84%	100%
Panel C: Boost tree on financial information and analyst-based peer group		
$Interactions$	19.17%	51.13%
roe	15.42%	41.10%
dp	1.09%	2.89%
$beta$	0.82%	2.20%
$G1roe$	0.53%	1.41%
$G5roe$	0.48%	1.28%
Total	37.51%	100%

For the first two regressions, return on equity is the most dominant driver of P/B. In the model with linear regression and pair-wise interactions among value drivers, interaction terms are the second-largest contribution to predictability. Comparing the results from Table 2.8 Panels A and B, after controlling for interaction terms, the importance of return on equity decreases significantly. Other variables contribute smaller proportions to out-of-sample predictability. From Table 2.8 Panel C, the importance of interactions by boosted tree model (after unique variable contribution) and analyst information dominates out-of-sample predictability. Most notably, when controlling for analyst information, not only does total out-of-sample R^2 increase but the contribution of return on equity significantly also decreases. Interactions may occur between return on equity and other information. Therefore, the unique contribution of return on equity decreases from 89.42% in linear regression without interactions to 41.10% in the boosted tree model. Figure 2.5 shows the contribution of each value driver in percentage terms of total predictability (100%) for linear regression with interaction terms.

Figure 2.5
R^2 decomposition of P/B value drivers—linear regression and pairwise interactions
This figure shows the relative contribution of each value driver to P/B prediction out-of-sample. The horizontal axis shows the relative contribution of each value driver to the total out-of-sample R^2 (in percent). The sum of all variable contributions is 100%.

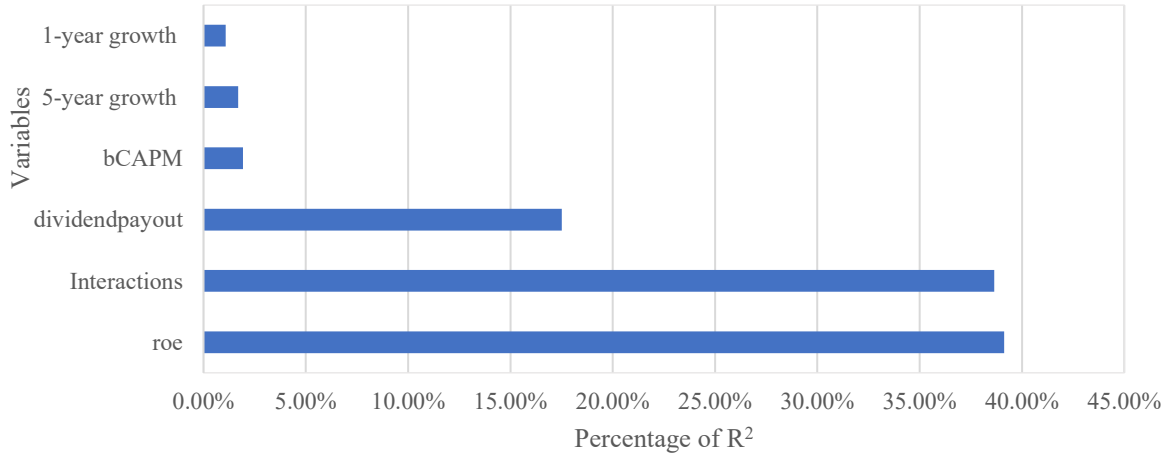


Figure 2.6 provides a visualization of the relative contribution of each value driver to P/B in percentage terms of the total predictability (100%) using the boosted tree.

Figure 2.6
R^2 decomposition of P/B value drivers—boosted tree with analyst information
This figure shows the relative contribution of each value driver to P/B prediction out-of-sample. The horizontal axis shows the relative contribution of each value driver to the total out-of-sample R^2 (in percent). The sum of all variable contributions is 100%.

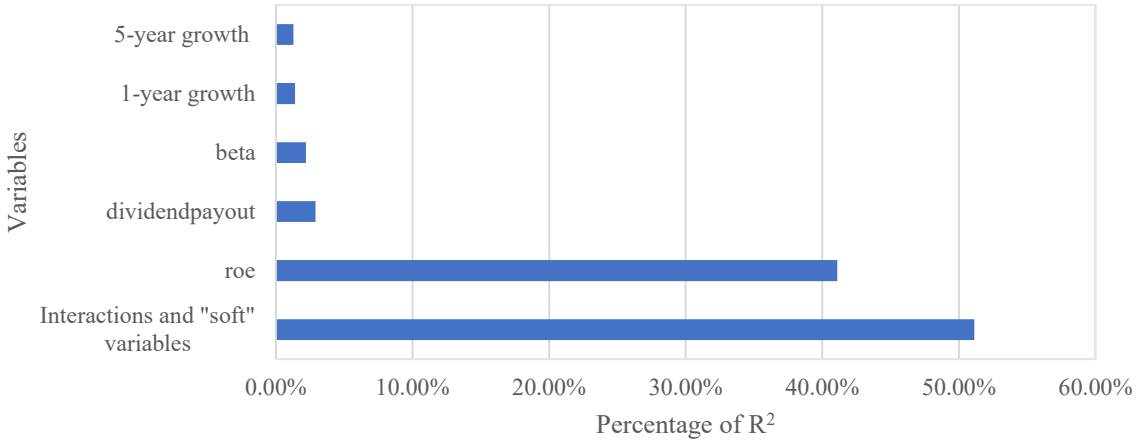


Table 2.9 Panels A and B show the linear regression results for EV/IC. Before controlling for interaction terms, the return on invested capital plays the most important role in predicting EV/IC. Although the contribution is not as high as the return on equity in the case of P/B, the return on invested capital still dominates out-of-sample predictability. After controlling for pairwise interactions in linear regression, total R^2 increases slightly from 26.92% to 28.90%, but each of the components changes significantly. The contribution of return on capital decreases significantly and is replaced by the contribution of interaction terms, which means that most of the previous contributions by return on capital were interactions with other variables. Figure 2.7 shows the relative contribution of each value driver to EV/IC in percentage terms of total predictability for linear regression with interaction terms.

Figure 2.7
R^2 decomposition of EV/IC value drivers—linear regression and pairwise interactions
This figure shows the average contribution of each value driver to EV/IC prediction out-of-sample. The horizontal axis shows the relative contribution of each value driver to the total out-of-sample R^2 (in percent). The sum of all variable contributions is 100%.

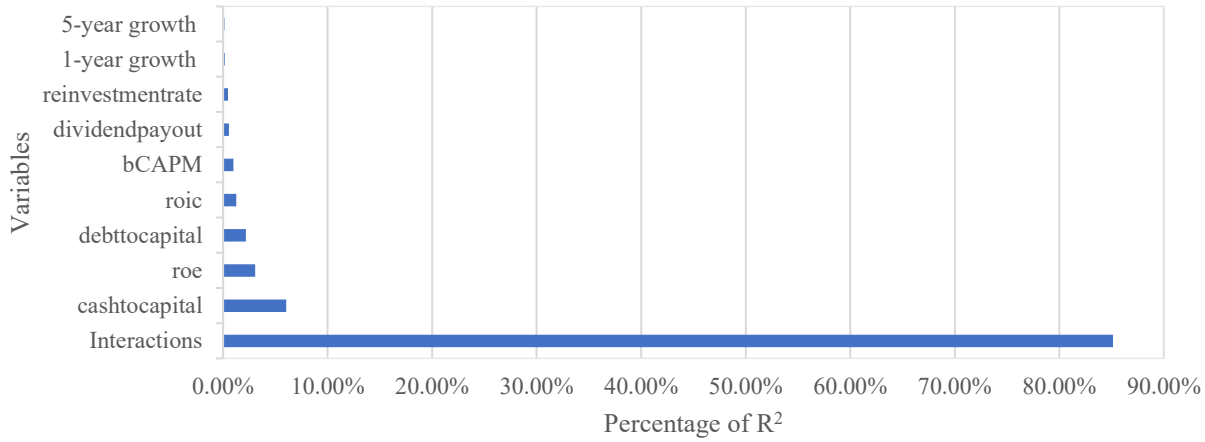


Table 2.9 Panel C shows the boosted tree results. After controlling for analyst information, the interactions account for approximately 23.42% of the total predictability and become the third-largest group. Total R^2 increases by a significant margin, from 28.90% to 49.31%. The unique contribution of return on equity becomes the largest contribution to predictability. The unique contribution of long-term growth is the second-largest contributor. Figure 2.8 shows the relative contribution of each value driver to EV/IC in percentage terms of total predictability for boosted tree with interaction terms.

Figure 2.8
R^2 decomposition of EV/IC value drivers—boosted tree with analyst information

This figure shows the relative contribution of each value driver to the EV/IC prediction out-of-sample. The horizontal axis shows the relative contribution of each value driver to the total out-of-sample R^2 (percent). The sum of all variable contributions was 100%.

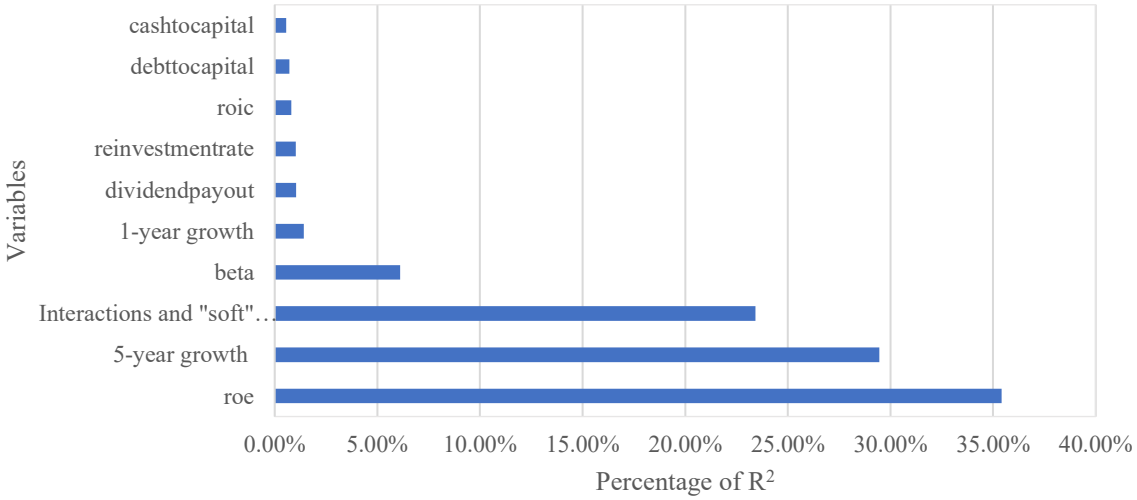


Table 2.9
Out-of-sample R^2 decomposition of EV/IC

The table shows the out-of-sample R^2 decompositions based on the Shapley value. The predicted variable is the EV/IC ratio. Panel A presents results for linear regression on financial statement information, Panel B shows the boosted tree on financial statement information and Panel C presents results for the boosted tree on financial statement information and analyst information. Variables are sorted by their level of R^2 contribution from highest to lowest.

Variable	R^2 contribution (level)	R^2 contribution (%)
Panel A: Linear regression on financial information		
$roic$	11.37%	42.23%
dtc	7.41%	27.52%
ctc	6.06%	22.52%
$beta$	0.97%	3.62%
roe	0.94%	3.49%
$G1roe$	0.11%	0.41%
ri	0.02%	0.09%
$G5roe$	0.02%	0.07%
dp	0.01%	0.04%
Total	26.92%	100%
Panel B: Linear regression on financial information and interactions		
$Interactions$	24.61%	85.14%
ctc	1.75%	6.05%
roe	0.89%	3.08%
dtc	0.63%	2.16%
$roic$	0.36%	1.26%
$beta$	0.29%	1.00%
dp	0.16%	0.56%
ri	0.13%	0.46%
$G1roe$	0.04%	0.15%
$G5roe$	0.04%	0.14%
Total	28.90%	100%
Panel C: Boosted tree on financial information and analyst-based peer group		
roe	17.46%	35.42%
$G5roe$	14.53%	29.46%
$Interactions$	11.55%	23.42%
$beta$	3.01%	6.11%
$G1roe$	0.70%	1.42%
dp	0.52%	1.05%
ri	0.51%	1.04%
$roic$	0.40%	0.81%
dtc	0.35%	0.72%
ctc	0.28%	0.57%
Total	49.31%	100%

2.6. Comparison with other peer valuation methods

We compare our results with other relative valuation methods, such as the sum of absolute rank difference (SARD-10) from Knudsen et al. (2017), the StarMine algorithm to define peer firms of Refinitiv[1], analyst-based peer groups (Kaustia and Rantala, 2021) and product-based competitors (Hoberg and Phillips, 2016). Table 2.10 reports the results of performance between different relative valuation approaches. We restrict the sample period from 2013 to 2020 to compare all methods (as data for analyst-based peer firms is restricted to 2013 only). We find that boosted trees especially boosted trees with a combination of financial information and analyst information, deliver the best out-of-sample prediction of both the P/B and EV/IC.

Table 2.10
Comparison of tree-based grouping and other types of peer groups
The table shows the comparison of different methods for P/B and EV/IC prediction. The methods are linear regression, boosted tree, SARD-10 (Knudsen et al. 2017), StarMine algorithm by Refinitiv, analyst cross-coverage (Kaustia and Rantala, 2021), K-10 product-based competitors (Hoberg and Phillips, 2016) and our boosted tree model.

Squared error	Linear regression on financial information	Boosted tree on financial information	SARD-10	StarMine Refinitiv	Analyst cross-coverage	K10-based competitor	Boosted tree on financial and analyst information
P/B							
Mean	0.29	0.25	0.31	7.84	25.05	39.00	0.21
Median	0.17	0.14	0.18	1.39	0.86	3.64	0.11
EV/IC							
Mean	0.25	0.22	0.33	7.23	12.09	17.38	0.20
Median	0.13	0.12	0.19	1.38	0.46	2.50	0.10

[1] The StarMine peers are created through Refinitiv's exclusive algorithm, which merges competitor lists mentioned in official filings, analyst coverage, business classification, and revenue similarity.

2.7. Conclusion

Valuation practices vary greatly among practitioners, who rely on both the 'science' of valuation, such as the relative valuation of similar companies or cash flow discount models with numerous assumptions, and the 'art' of leveraging additional knowledge to make valuations more precise.

We use a data-driven approach with machine learning methods to provide new empirical insights into valuation. Our results show that there are significant interactions among the fundamentals that influence company value. Treating each value driver in isolation and with a linear approach is not a sound valuation practice.

We find that long-term growth is more valuable than short-term growth. Some noteworthy interactions are between the growth rate and risk, growth rate, and dividend payout, and growth rate and reinvestment rate. Companies with a combination of high growth and low risk generate the highest value cross-sectionally. A high dividend payout creates the most value for companies with high long-term growth. Companies with the greatest value are those with median reinvestment and median growth. All interactions are observed when controlling for the analyst information, proving the usefulness of information not observed in financial statements. Machine learning can take on the task of including such information in the valuation process. The boosted tree model performs better than other relative valuation techniques and linear regression.

Our results provide insights for managers regarding company management strategies and for investors regarding investment decisions. Better valuations of companies can improve market efficiency by reducing mispricing and improving resource allocation.

Chapter 3: Nonlinear market efficiency

3.1. Introduction

Machine learning is increasingly used in investment decision-making and trading in financial markets. Bloomberg projects that machine learning can involve in 90% of investment management by 2040. Compared to other less quantitative approaches or conventional statistics, machine learning is known for its ability to generate superior forecasts. Many mutual funds today use machine learning as part of their investment process, and many other funds start switching to state-of-the-art forecasting techniques.

The increasing use of machine learning in investment decision-making indicates that there is complexity in how information combines asset prices. In this chapter, we introduce a second dimension to the efficient markets hypothesis (EMH), being the complexity of the functional forms that link information and stock prices. This new dimension is largely independent of the original dimension of the EMH, being the breadth of information sets. For example, consider a particular information set, such as all publicly available information. Prices may reflect this information in a linear manner, or non-linear transformations and interactions between different pieces of information may also be reflected in prices. This additional dimension is crucial in capturing the effects of advancements in data science techniques and their implementation in markets.

We use neural networks to predict returns and compare the predictive ability with that of linear regression models. Neural networks learn without being explicitly programmed with prior knowledge about the relationships between predictors and return, as opposed to other parametric models, such as linear regression. Neural networks are effective in combining large amounts of data and trading signals.[2] We refer to the predictive ability difference between the two methods as "nonlinear market inefficiency"—the amount of predictability created by complex interactions and non-linearities among pieces of information.[3] We measure non-linear inefficiency for three main information sets: past return information, accounting information,

[2] As it is most likely that machine learning was not used in the early 1970s (the starting point of our sample), we have to assume that if someone could have used the technology in the past, we want to measure how much predictability they could exploit.

[3] Return prediction signals and firm characteristics are documented in other empirical asset pricing papers compiled by Chen and Zimmermann (2021).

and anomalies. The past return information set includes 39 input variables, which are constructed from historical stock returns during the previous 252 days. The accounting information includes accounting variables from the company's financial statements. The final set includes factors and anomalies constructed based on many asset pricing papers and publications.

On average, we find that the non-linear inefficiency in public information (accounting data and firm characteristics) is higher than that of past return information. This finding implies that most of the complexity reflected in asset price is in the public information set. Adding past return information to public information does not increase non-linear inefficiency significantly. Between the two types of public information (accounting information and anomalies), we observe higher predictability in accounting variables than in factors and anomalies for both linear and non-linear models. However, the predictability difference between linear and non-linear models is greater for anomalies than accounting variables.

The predictive power of linear regression worsens when using the largest set of variables—203 anomalies. Similarly, the forecasting ability of linear regression is worse than using only 30 accounting variables. On average, neural networks have the highest improvement over linear regression when using anomalies as input variables, which implies that the level of complex interactions in the anomalies set is higher than the other two sets of variables.

We find that the increasing use of advanced data modeling reduces non-linear inefficiency. Using the growth of quantitative mutual funds and the quantity of machine learning publications as proxies of technology advances, we find that the technology race makes the market more efficient. There is also an interaction between the two forces. Indeed, there is significant return predictability in the 1960s and 1970s using machine learning, which shows that the market is highly non-linearly inefficient if we measure market inefficiency in the past using today's knowledge and technology. We show that as technology becomes more widespread, non-linear inefficiency disappears.

Fama (1965) emphasizes that the basic linear regression that forms the foundation of the serial correlation model is not sophisticated enough to identify the intricate patterns in stock prices. Jensen (1978) mentions scientific revolutions, which refer to better data and increased econometric sophistication, uncovering inconsistencies that old techniques missed in the past. Shiller (2003) repeats the skeptical view. The challenge for financial economists is not to maintain the EMH in its purest form but to provide a more accurate description of the actual market.

In the past, many asset-pricing anomaly detections indirectly shake the ground of the EMH. Some other papers are directly connected to the work of return prediction based on asset-

pricing anomalies, such as Granger (1992), Lo (2004, 2012), and Daniel and Titman (1999). The number of anomalies detected in empirical studies grow to more than 300 return predictors (Hou, Xue, and Zhang, 2020; Chen and Zimmermann, 2021). Campbell and Yogo (2006) show that it is more challenging to gain predictability without the prudent utilization of econometric models. Indeed, Hou et al. (2020) find that anomalies fail in different sample periods. Anomalies disappear for two reasons: they reflect statistical artifacts, or they are arbitraged away. Some papers attribute disappearing anomalies to arbitragers (McLean and Pontiff, 2016; Falck et al., 2022). Cochrane (2011) mentions the challenge facing us—how to account for enormous independent dimensions of expected returns.

Assuming that some anomalies are subsumed by others, and some provide independent information (Green, Hand, and Zhang, 2017) with large amounts of return signals, how complex can the relationship among return signals be? How much profit we can exploit from the complexity of nonlinear relationships? In the machine learning era, we are empowered with tools to shed light on these questions. Stambaugh and Yuan (2017) and Kozak et al. (2020) show that a combination of signals accommodates a wide range of anomalies and achieves a less noisy measure for stock mispricing than the separate signal in the model. Barbopoulos et al. (2021) show that increase in information access by cloud computing leads to improvements in market efficiency. Analogously, Gu et al. (2020) find that machine learning generates significant profits using an average combination of return prediction signals. Our chapter is related to the empirical approach of Gu et al. (2020) study. However, our study is different from Gu et al. (2020) in that we propose a new dimension of market efficiency based on different modeling methods. We do not aim to show how different machine learning models predict return as in Gu et al. (2020).

Our study complements the literature on efficient markets measured in relative terms. We share the same motivations as Karapandza and Mazin (2014) and Rösch, Subrahmanyam, and Van Dijk (2017), who discuss market efficiency in relative terms. Our chapter is different in that we propose an orthogonal dimension of market efficiency to the traditional dimension of market efficiency. The dimensions we propose are different modeling techniques (linear regression and machine learning) in return prediction.

We also combine a significant and diverse number of return signals to view market efficiency from different perspectives. Since the introduction of EMH, anomaly discovery flourishes, making it hard to believe that the signals are pure statistical artifacts or data snooping. Our chapter provides empirical evidence on why technological developments should not be ignored in testing the EMH. Our chapter is one of the first to empirically study efficient

markets, conditional on the level of technology, to account for the growing utilization of big data and the sophistication of forecasting techniques.

3.2. Conceptual framework

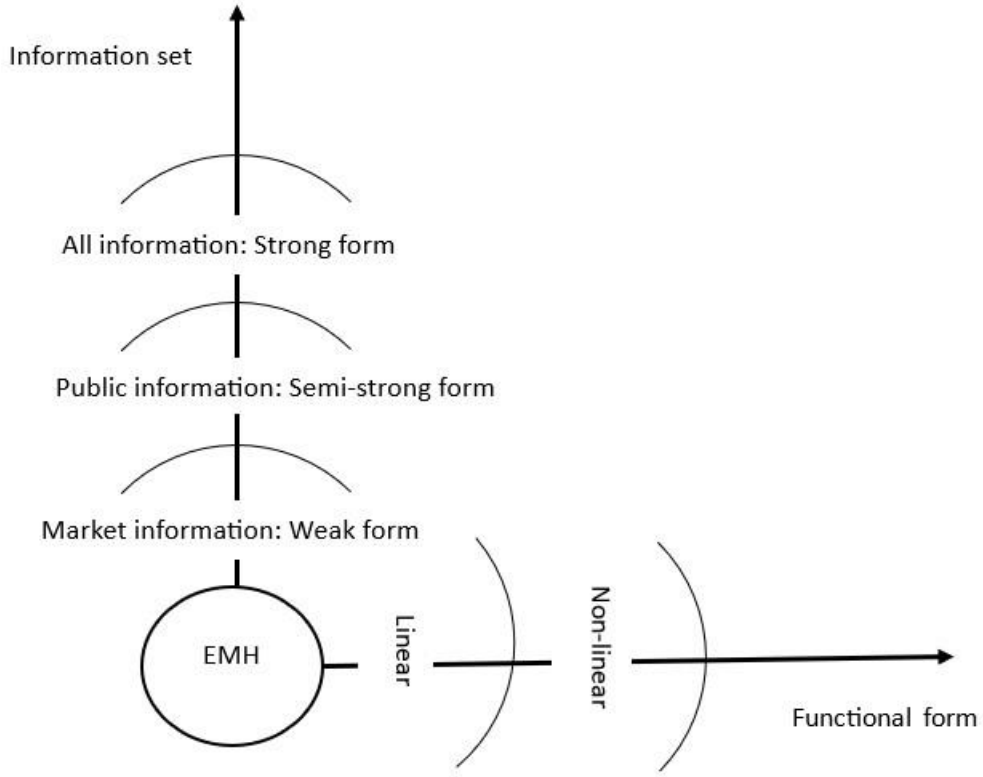
In 1970, Fama introduces the idea of market efficiency, which suggests that stock prices incorporate all available information. Fama (1970) categorizes market efficiency into three categories: weak, semi-strong, and strong, depending on the degree to which stock prices reflect past, public, and all information, respectively.

We explore another (orthogonal) dimension of market efficiency, which is the functional form of the model that investors use to predict returns based on a specific set of information. The two functional forms in our chapter are the linear model and a nonlinear machine learning model. The linear model, or ordinary least squared, is the conventional statistical model used in asset pricing studies. The nonlinear model is a neural network model, which incorporates nonlinearity and interactions among predictors.

Figure 3.1 illustrates these orthogonal dimensions of market efficiency.

Figure 3.1
Orthogonal dimensions of efficiency

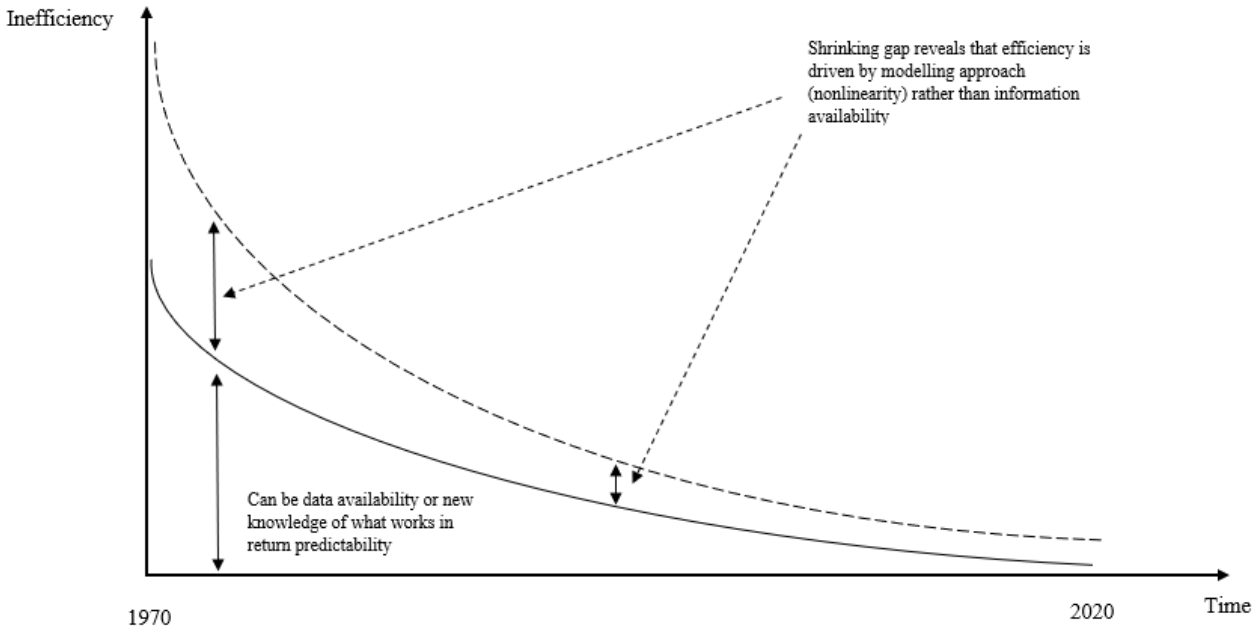
The figure shows the orthogonal dimensions of market efficiency. The y-axis shows the dimensions of the information set. The x-axis shows the dimension of market efficiency by model function form given an information set.



The market can be more efficient with respect to the breadth of information incorporated into prices – from limited information in weak-form efficiency to all available information in strong-form efficiency. But, for a given information set, it can also be more or less efficient in the functional form complexity of how that information gets reflected in prices.

Figure 3.2 illustrates this chapter's core hypothesis, being that as the use of machine learning becomes more prevalent in markets and investment decisions, the difference in how much one can predict future returns using linear models versus using non-linear models is expected to shrink, reflecting increasing non-linear market efficiency. This shrinking gap is expected to be observed in addition to a general tendency for markets to become more efficient through time in the traditional EMH sense.

Figure 3.2
Inefficiency measure
The figure shows the change in inefficiency measures over time with respect to the model's functional forms. The y-axis shows the level of inefficiency (return predictability) and the x-axis shows the time.



3.3. Data and Methods

3.3.1. Data

Our sample period is from January 1965 to December 2019 to minimize the number of missing observations. The number of stocks in our study is almost 30,000, with a monthly average of approximately 6,000 stocks. We filter companies with share codes 10 or 11 (ordinary equity).

We use three sets of return prediction signals, based on predictors identified in other empirical studies. The first set of inputs is past returns, which we download from the Centre for Research in Security Prices (CRSP)'s daily and monthly stock files for all firms listed on New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and NASDAQ Stock Exchange (NASDAQ).

The second set is documented anomalies. We download anomalies from a shared dataset of 203 predictors provided by Chen and Zimmermann (2021).[4] Data on anomalies are at the firm level and updated monthly. The anomalies from Chen and Zimmermann (2021) are collected from published studies since 1970. We fill in missing values in each stock month with the monthly averages of those values. We normalize all variables by transforming all variables for the whole sample period to the range $[-1,1]$, as in Kelly, Pruitt, and Su (2019). We also include six macroeconomic variables downloaded from Goyal and Welch (2008). The anomalies are provided in Appendix 3.A.

The third set of inputs are company fundamentals from quarterly 10-K and 10-Q filings obtained from CRSP-Compustat merged database. The accounting variables are listed in Appendix 3.B.

3.3.2. Inverse proxies of market efficiency

We report three measures of return predictability for different models. Return spread and risk-adjusted excess returns (alphas) are designed to evaluate asset pricing model performance. Cross-sectional R^2 measures the return predictability by measuring the proportion of return variance that it can explain.

First, we measure the return from a long-short trading strategy. We form deciles based on predicted returns in month $t+1$. We form both equal-weighted portfolios and value-weighted portfolios. We calculate the return spread by buying the top decile and selling the bottom decile each month.

$$ReturnSpread_{t+1} = r_{Q10,t+1} - r_{Q1,t+1} \tag{22}$$

in which, $r_{Q10,t}$, $r_{Q1,t}$ is the top and bottom decile portfolio.

Second, we calculate the cross-sectional R^2. We employ portfolio returns to calculate this measure. We use portfolio return instead of stock-level return as R^2 built on portfolio return predictions is less noisy than based on stock return predictions. At the stock level, R^2 can become highly negative due to poor prediction for certain stocks, while prediction at the portfolio level is more consistent and less noisy. Our measure of cross-sectional R^2 is:

[4] Data from Chen and Zimmermann (2020) was downloaded from: https://www.openassetpricing.com/

$$R^2_{t+1} = 1 - \frac{\sum_{p=1}^{10}(r_{p,t+1} - \hat{r}_{p,t+1})^2}{\sum_{p=1}^{10}(r_{p,t+1})^2} \quad (23)$$

in which, $r_{p,t+1}$ is the monthly return of portfolio p at month $t+1$.

$\hat{r}_{p,t+1}$ is predicted return of portfolio p at month $t+1$.

Third, we calculate the risk-adjusted excess return, which is the intercept of equation (24). We estimate rolling monthly regressions using return data from the previous 24 months to obtain the alpha estimates:

$$r_{Q10,t+1} - r_{Q1,t} = \alpha_c + \beta_c^{mkt} MKTRF_t + \beta_c^{smb} SMB_t + \beta_c^{hml} HML_t + \beta_c^{umd} UMD_t + \epsilon_{c,t} \quad (24)$$

in which, $r_{c,Q10,t} - r_{c,q1,t}$ is the long minus short portfolio.

$MKTRF_t, SMB_t, HML_t \text{ and } UMD_t$ are the corresponding Fama-French and momentum factors.

Lastly, we quantify the shared variation among four market efficiency metrics by extracting their components using principal component. We use composite predictability—the first principal component—to provide an overall view of movements in market efficiency.

We define nonlinear inefficiency as the performance differential between linear regression and machine learning:

$$NLIE_t = PL_t - PML_t \quad (25)$$

in which, $NLIE_t$ is the nonlinear inefficiency at time t

PL_t is the composite return predictability by linear regression at time t

PML_t is the composite return predictability by machine learning model at time t

3.3.3. Linear regression

We estimate linear regression models using each of the information sets as the baseline models. Return predictability from linear regression is also used to compute the nonlinear inefficiency in equation (25). The linear regression is:

$$r_{i,t} = \alpha + \sum_{m=1}^{M} \beta_m x_{i,t} \quad (26)$$

in which, $r_{i,t}$ is the return of stock i at time t

$x_{i,t}$ is the return predictor of stock i at time t (in set M, which can be past returns, accounting information, or anomalies)

3.3.4. Feed-forward neural network

The feed-forward neural network is the non-linear method we choose to predict return. It is arguably the most potent model in machine learning. Its flexibility comes from many layers of non-linear predictor interactions.

The best way to describe why our feed-forward neural network brings nonlinearity and interactions to modelling is to start with the simplest single-neuron unit. The input values are multiplied by their weights and summed as follows:

$$v = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m = \sum_{i}^{m} w_i x_i \tag{27}$$

The output is function $y = f(v)$ of the weighted sum in equation (27). This function is called the activation function in the neural network. There are numerous options available for selecting nonlinear activation functions (such as sigmoid, hyperbolic, and SoftMax). Our study uses rectified linear units (ReLU), which promotes sparsity in the active neuron count and enables rapid evaluation of derivatives. ReLU is a nonlinear activation function. In a multiple-layer network, $f(v)$ can be the value of hidden units before the model calculates the prediction of $r_{i,t}$.

The cost function to find the vector of parameters that minimize squared error:

$$L(A, \theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (r_{i,t+1} - g(C_{i,t}))^2 \tag{28}$$

in which, N, T is the number of firms and monthly periods in our estimation.

Function $g(.)$ is the nonlinear function of anomalies if our model is a neural network. Function $g(.)$ does not depend on i, t because we estimate the model using the whole panel instead of rerunning it for every cross-section as a traditional Fama-MacBeth regression. The first layer has 128 neurons and the second has 64 neurons:

$$y_{1,h} = \varphi\left(C_{i,t} w_{0,h} + b_{0,h}\right) \tag{29}$$

$$y_{2,h} = \varphi\left(y_{1,h} w_{1,h} + b_{1,h}\right) \tag{30}$$

$$E(r_{i,t+1}) = \varphi\left(y_{2,h} w_{2,h} + b_{2,h}\right) \tag{31}$$

in which, $w_{0,h}$, $w_{1,h}$, $w_{2,h}$, $b_{0,h}$, $b_{1,h}$, $b_{2,h}$ are the parameters of the two hidden layers. $w_{0,h} \in \mathbb{R}^{211\times128}$, $w_{1,h} \in \mathbb{R}^{128\times64}$, $w_{2,h} \in \mathbb{R}^{64\times1}$, $b_{0,h} \in \mathbb{R}^{1\times128}$, $b_{1,h} \in \mathbb{R}^{1\times64}$, $b_{2,h} \in \mathbb{R}$

The cost function is:

$$L\left(w_{0,h}, w_{1,h}, w_{2,h}, b_{0,h}, b_{1,h}, b_{2,h}\right) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(r_{i,t+1} - E(r_{i,t+1})\right)^2 \tag{32}$$

in which, N, T is the number of firms and monthly periods in our estimation.

We train the neural network by updating the weights. We initially set weights to random values, repeat inputs into the network and compute the output using the activation function until the cost function is minimized. To prevent overfitting (as the neural network has many parameters), we adopt early stopping and batch normalization as regularisation of the neural network. We use the same design and hyperparameters for all information for the sake of easy comparison.

The dataset starts in January 1965 and ends in December 2019. Choosing a sample-splitting scheme is essential in this study. We use rolling window estimation in our estimate. Our results are out-of-sample because in-sample predictions might overestimate the predictability of returns using a combination of signals—investors deciding in real time without the benefit of hindsight. Investors cannot capture the predictability that econometricians observe when we run the in-sample test. They face high-dimensional problems and make estimating errors, which econometricians observe when looking at historical data ex-post. Our approach mimics what can happen in real-time—using historical data to estimate the model, then use the estimated model to predict the next month's return. We employ a rolling estimation scheme. In predicting returns for month $t+1$ using information available up to time t. For example, to predict return from January 1985 to December 1985, we estimate the model using data from January 1965 to December 1984. We choose hyperparameters by comparing estimated model performance over a validation set not used in the training set. We use the best model to make one-month-ahead return predictions from January 1985 to December 1985. We move the training and testing forward by one year. We repeat the whole process to predict returns for out-of-sample periods from January 1985 to December 2019. The model is retrained

every year with the newest data. Therefore, to be on the same footing as investors, we report only out-of-sample predictability.

3.3.5. Information sets

#1 Past returns at different horizons

Similar to Takeuchi and Lee (2013), we compute return predictors for past returns over the horizon $[t-k, t-1]$ where k is the lookback horizon in days and $k = [1, 2, \dots 21, 42, \dots 252]$, that is, k progresses in one-day increments for the first month and one-month increments for the rest of the year.

Our past return variables include some popular signals, such as the 6-month momentum from Jegadeesh and Titman (1993), and the 12-month look back from Asness, Moskowitz, and Pederson (2013), skipping the most recent month. Jegadeesh and Titman (2001) show negative short-term momentum on horizons of up to a month.

In total, we have 39 variables in the past return information set as we also include the idiosyncratic volatility, alphas, and betas by estimating the model in equation (33). The alphas and betas are the estimated value of the intercept and slope of the equation (33). We include them as part of the past information set. We also include idiosyncratic volatility (standard deviation of the error terms in equation (33)):

$$r_{i,\tau} = \alpha_i + \beta_i r_{m,\tau} + \varepsilon_{i,\tau} \tag{33}$$

in which, $\tau \in [t-k, t-1]$ and $k = [164, 252]$ days.

$r_{i,\tau}$ is the return of stock i.

$r_{m,\tau}$ is the market return in which we use the return on CRSP value-weighted index.

#2 Accounting variables

We also examine how fundamental analysis and accounting information provide a trading signal that predicts returns. We combine accounting information in financial statements, which are measured at a quarterly frequency. Each month, we use data from 20 prior fiscal quarters. We use the 30 most frequently observed numerical accounting variables at the firm level. These variables are extracted from the quarterly financial statements (provided in Appendix 3.B.). Many of the other uncommon variables in financial statements are redundant and are represented by other items. We use 16 items from the most recent balance sheets. For variables from income and cash flow statements, we sum the quarterly value from the most recent 10-K or 10-Q.[5] This input set is the same as that of Bartram and Grinblatt (2018).

Using the predicted market value, we determine the mispricing signal by computing the disparity between a stock's projected value and the actual market value:

$$Mispricing_{i,t} = \frac{E(M_{i,t}) - M_{i,t}}{M_{i,t}} \tag{34}$$

in which, $E(M_{i,t})$ is the expected market value at month t of firm i.

$M_{i,t}$ is the actual market value at time t of firm i.

We use the mispricing signal at time t to sort the return and form the trading signal at time $t + 1$ as in Bartram and Grinblatt (2018).

#3 Anomalies and factors

The 203 trading signals include both predictors that are demonstrated to achieve statistical significance and likely predictors (not statistically in-sample in the relevant literature). The list of signals is provided in Appendix 3.A.

[5] The cashflow statement is cumulative. Quarterly value should be calculated by taking the difference between the adjacent fiscal quarters.

3.4. Portfolio forecasts and asset pricing tests

Table 3.1 presents the results of the weak-form test, in which we use only historical return information.

Table 3.1
Performance of portfolios based on past returns

This table reports the performance of prediction-sorted portfolios over the out-of-sample testing period for equal-weighted and value-weighted portfolios based on past returns. All stocks are sorted into deciles based on their predicted returns at time $t+1$. Columns 'Pred', 'Avg', 'SD', and 'SR' provide predicted returns, average monthly return, standard deviation, and Sharpe ratio, respectively.

	Linear regression				Neural network			
Panel A: Equal-weighted portfolios								
	Pred	Avg	SD	SR	Pred	Avg	SD	SR
Low (L)	−9.27	0.96	6.79	0.49	−1.10	−0.02	7.04	-0.01
2	−4.22	1.09	5.77	0.65	−0.06	0.64	5.79	0.38
3	−2.17	1.05	5.51	0.66	0.40	0.90	5.32	0.59
4	−0.70	1.11	5.26	0.73	0.73	1.01	5.04	0.69
5	0.55	1.13	5.21	0.75	1.01	1.18	4.98	0.82
6	1.74	1.20	5.27	0.79	1.28	1.21	4.94	0.85
7	2.97	1.14	5.24	0.75	1.55	1.24	5.06	0.85
8	4.42	1.13	5.44	0.72	1.87	1.28	5.21	0.85
9	6.43	1.01	5.80	0.60	2.30	1.41	5.83	0.84
High (H)	11.40	1.09	6.68	0.57	3.50	2.06	8.73	0.82
H−L	20.67	0.13	2.43	0.18	4.60	2.08	4.66	1.55
Panel B: Value-weighted portfolios								
Low (L)	−7.94	0.91	6.08	0.52	−0.89	0.49	6.38	0.27
2	−4.13	0.99	5.06	0.68	−0.04	0.84	5.10	0.57
3	−2.15	0.92	4.58	0.70	0.40	0.79	4.66	0.59
4	−0.68	0.95	4.45	0.74	0.73	0.94	4.51	0.72
5	0.56	1.03	4.32	0.83	1.01	0.96	4.42	0.76
6	1.73	1.06	4.46	0.83	1.28	1.00	4.34	0.80
7	2.96	1.05	4.57	0.79	1.55	1.09	4.60	0.82
8	4.38	0.93	4.57	0.70	1.86	1.09	4.51	0.84
9	6.33	0.97	5.08	0.66	2.28	1.19	5.06	0.81
High (H)	10.07	0.92	6.15	0.52	3.03	1.20	6.68	0.62
H−L	18.02	0.01	4.48	0.01	3.92	0.71	4.71	0.52

The out-of-sample performance of the portfolio is consistent with the accuracy of machine learning forecast (Gu et al., 2020; Tobek and Hronec, 2021). Realized returns increase monotonically with the forecasted portfolio. The long-short spread return is greater for neural networks than for linear regression. The results hold for annualized Sharpe ratios, which jump from 0.01 to 0.52 for value-weighted portfolios formed by linear regression and machine learning predictions, respectively.

Table 3.2 shows the results of accounting variables. The best 10−1 strategy comes from equal-weighted neural network portfolios based on accounting variables when the return is equally weighted. The accounting-based portfolio has the greatest long-short spread and Sharpe ratio compared to the other two types of information sets. However, the difference between linear regression and neural networks is not as remarkable as firm characteristics and past returns.

Table 3.2
Performance of portfolios based on accounting variables

This table reports the performance of prediction-sorted portfolios over the out-of-sample testing period for equal-weighted and value-weighted portfolios based on accounting variables. All stocks are sorted into deciles based on their predicted returns at time $t+1$. Columns 'Pred', 'Avg', 'SD', and 'SR' provide predicted returns, average monthly return, standard deviation, and Sharpe ratio, respectively.

	Linear regression				Neural network			
Panel A: Equal-weighted portfolios								
	Pred	Avg	SD	SR	Pred	Avg	SD	SR
Low (L)	−0.50	0.39	6.86	0.20	−0.27	0.10	6.47	0.06
2	−0.25	0.59	5.80	0.35	−0.11	0.45	5.33	0.29
3	−0.16	0.66	5.30	0.43	−0.07	0.65	5.00	0.45
4	−0.09	0.80	5.24	0.53	−0.04	0.78	4.92	0.55
5	−0.03	0.78	5.18	0.52	−0.02	0.88	5.09	0.60
6	0.03	0.88	5.32	0.58	0.01	0.97	5.25	0.64
7	0.09	0.98	5.49	0.62	0.04	1.17	5.52	0.73
8	0.16	1.42	5.86	0.84	0.07	1.37	5.90	0.80
9	0.25	2.07	6.21	1.16	0.12	1.93	6.61	1.01
High (H)	0.49	2.32	7.37	1.09	0.26	2.61	8.11	1.12
H−L	0.99	1.92	5.50	1.21	0.53	2.51	4.92	1.77
Panel B: Value-weighted portfolios								
Low (L)	−0.48	0.61	5.90	0.36	−0.24	0.64	5.53	0.40
2	−0.25	0.88	4.94	0.62	−0.11	0.75	4.77	0.54
3	−0.16	0.94	4.44	0.74	−0.07	0.94	4.46	0.73
4	−0.09	1.00	4.45	0.78	−0.04	1.06	4.47	0.82
5	−0.03	0.99	4.26	0.80	−0.02	1.13	4.67	0.84
6	0.03	1.01	4.34	0.81	0.01	1.15	5.13	0.78
7	0.09	1.13	4.73	0.82	0.04	1.14	5.65	0.70
8	0.16	1.38	5.55	0.86	0.07	1.39	6.12	0.79
9	0.24	1.28	5.77	0.77	0.12	1.52	6.65	0.79
High (H)	0.45	1.66	6.47	0.89	0.24	1.71	7.34	0.81
H−L	0.93	1.05	4.50	0.81	0.48	1.07	5.12	0.73

Table 3.3 reports predicted portfolio returns, actual portfolio returns, standard deviations, and Sharpe ratios for equal-weighted and value-weighted decile portfolios based on anomalies and factors.

Table 3.3
Performance of portfolios based on anomalies

This table reports the performance of prediction-sorted portfolios over the out-of-sample testing period for equal-weighted and value-weighted portfolios based on anomalies. All stocks are sorted into deciles based on their predicted returns at time $t+1$. Columns 'Pred', 'Avg', 'SD', and 'SR' provide predicted returns, average monthly return, standard deviation, and Sharpe ratio, respectively.

	Linear regression				Neural network			
	Pred	Avg	SD	SR	Pred	Avg	SD	SR
Panel A: Equal-weighted portfolios								
Low (L)	−9.58	1.00	6.49	0.53	−1.05	0.13	7.56	0.06
2	−4.66	0.96	5.82	0.57	−0.04	0.65	6.31	0.36
3	−2.37	0.97	5.54	0.61	0.42	0.79	5.49	0.50
4	−0.79	0.92	5.22	0.61	0.76	0.86	5.12	0.58
5	0.47	0.95	5.00	0.66	1.04	0.94	4.83	0.67
6	1.65	0.99	4.95	0.70	1.32	1.10	4.75	0.80
7	2.95	0.94	5.03	0.65	1.62	1.12	4.85	0.80
8	4.57	0.98	5.21	0.65	1.97	1.24	4.89	0.88
9	6.84	1.10	5.34	0.72	2.44	1.34	5.14	0.91
High (H)	11.71	1.08	5.75	0.65	3.48	1.71	6.09	0.98
H−L	21.29	0.08	2.06	0.14	4.53	1.59	3.81	1.44
Panel B: Value-weighted portfolios								
Low (L)	−9.07	0.97	5.15	0.66	−0.93	0.58	6.64	0.30
2	−4.64	0.96	4.80	0.69	−0.03	0.80	5.25	0.53
3	−2.38	1.01	4.82	0.73	0.43	0.85	4.96	0.59
4	−0.80	0.92	4.59	0.70	0.76	0.91	4.62	0.68
5	0.46	0.94	4.61	0.70	1.04	1.00	4.68	0.74
6	1.66	1.08	4.55	0.82	1.32	0.98	4.41	0.77
7	2.96	0.93	4.51	0.72	1.62	1.04	4.40	0.82
8	4.58	0.99	4.57	0.75	1.97	1.16	4.33	0.93
9	6.83	0.94	4.56	0.71	2.43	1.16	4.54	0.89
High (H)	11.19	1.03	4.70	0.76	3.25	1.24	4.96	0.87
H−L	20.26	0.05	2.77	0.07	4.18	0.66	4.35	0.53

For the anomalies, the Sharpe ratio on the value-weighted portfolio also increases from 0.07 to 0.53. There is also a monotonic increase in portfolio returns from decile 1 to decile 10. This indicates that linear regression has limited predictive power, and machine learning can yield higher profit compared to linear regression.

Table 3.4 shows the portfolio performance when we use all information (past returns, anomalies, and accounting variables). The monthly return is 0.66%, resulting in a Sharpe ratio of 0.53. Long-short returns are almost the same as portfolios based on anomalies, but not superior to portfolios based on past returns and accounting variables.

Table 3.4
Performance of portfolios based on all variables

This table reports the performance of prediction-sorted portfolios over the out-of-sample testing period for equal-weighted and value-weighted portfolios based on all variables. All stocks are sorted into deciles based on their predicted returns at time $t+1$. Columns 'Pred', 'Avg', 'SD', and 'SR' provide predicted returns, average monthly return, standard deviation, and Sharpe ratio, respectively.

	Linear regression				Neural network			
Panel A: Equal-weighted portfolios								
	Pred	Avg	SD	SR	Pred	Avg	SD	SR
Low (L)	−0.10	1.00	6.49	0.53	−0.01	0.13	7.56	0.06
2	−0.05	0.96	5.82	0.57	0.00	0.65	6.31	0.36
3	−0.02	0.97	5.54	0.61	0.00	0.79	5.49	0.50
4	−0.01	0.91	5.22	0.61	0.01	0.86	5.12	0.58
5	0.00	0.95	5.00	0.66	0.01	0.94	4.83	0.67
6	0.02	0.99	4.95	0.70	0.01	1.10	4.75	0.80
7	0.03	0.94	5.03	0.65	0.02	1.12	4.85	0.80
8	0.05	0.98	5.21	0.65	0.02	1.24	4.89	0.88
9	0.07	1.10	5.34	0.72	0.02	1.34	5.14	0.91
High (H)	0.12	1.08	5.75	0.65	0.03	1.71	6.09	0.98
H−L	0.21	0.09	2.05	0.14	0.05	1.59	3.81	1.44
Panel B: Value-weighted portfolios								
Low (L)	−0.09	0.97	5.15	0.66	−0.01	0.58	6.64	0.30
2	−0.05	0.96	4.80	0.69	0.00	0.80	5.25	0.53
3	−0.02	1.01	4.82	0.73	0.00	0.85	4.96	0.59
4	−0.01	0.92	4.59	0.70	0.01	0.91	4.62	0.68
5	0.00	0.94	4.61	0.71	0.01	1.00	4.68	0.74
6	0.02	1.08	4.55	0.82	0.01	0.98	4.41	0.77
7	0.03	0.93	4.51	0.71	0.02	1.04	4.40	0.82
8	0.05	0.99	4.58	0.75	0.02	1.16	4.33	0.93
9	0.07	0.94	4.56	0.72	0.02	1.11	4.54	0.85
High (H)	0.11	1.03	4.71	0.76	0.03	1.24	4.96	0.87
H−L	0.20	0.05	2.77	0.07	0.04	0.66	4.35	0.53

Table 3.5
Excess return by decile portfolios

The table reports the excess return (alphas) of the time-series regressions of returns on risk factors. The decile portfolios are across columns. The OLS row presents the alpha of the linear model and the NN row presents the alpha of the neural network model. The sample is from January 1985 to December 2020. T-statistics are reported in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Alpha	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P10−P1
Panel A: Past returns											
OLS	0.58^{***}	0.58^{***}	0.50^{***}	0.53^{***}	0.55^{***}	0.66^{***}	0.60^{***}	0.62^{***}	0.52^{***}	0.76^{***}	0.18^{*}
	(3.49)	(4.87)	(5.16)	(5.43)	(5.64)	(5.84)	(4.86)	(4.62)	(3.96)	(4.50)	(1.41)
NN	-0.30^{**}	0.16^{*}	0.36^{**}	0.42^{***}	0.58^{***}	0.60^{***}	0.61^{***}	0.66^{***}	0.87^{***}	1.95^{***}	2.25^{***}
	(−2.24)	(1.68)	(3.83)	(5.32)	(7.4)	(6.59)	(6.29)	(5.98)	(5.68)	(5.13)	(6.68)
Panel B: Accounting variables											
OLS	−0.09	−0.03	0.01	0.13^{*}	0.15^{*}	0.31^{***}	0.50^{***}	1.06^{***}	1.76^{***}	2.09^{***}	2.18^{***}
	(−1.05)	(−0.47)	(0.25)	(1.95)	(1.95)	(2.82)	(3.42)	(5.58)	(7.57)	(8.00)	(8.42)
NN	-0.34^{***}	-0.11^{*}	0.05	0.14^{**}	0.24^{***}	0.39^{***}	0.62^{***}	0.92^{***}	1.58^{***}	2.43^{***}	2.78^{***}
	(−3.63)	(−1.55)	(0.73)	(2.09)	(3.04)	(4.19)	(5.28)	(5.37)	(6.66)	(8.13)	(9.61)
Panel C: Anomalies											
OLS	0.49^{***}	0.44^{***}	0.45^{***}	0.38^{***}	0.44^{***}	0.45^{***}	0.37^{***}	0.39^{***}	0.49^{***}	0.48^{***}	−0.01
	(−3.45)	(−3.89)	(4.33)	(3.6)	(4.01)	(4.73)	(3.78)	(3.67)	(4.79)	(3.91)	(−0.09)
NN	-0.21^{**}	0.20^{*}	0.29^{**}	0.27^{***}	0.35^{***}	0.51^{***}	0.50^{***}	0.61^{***}	0.73^{***}	1.12^{***}	1.33^{***}
	(−2.11)	(1.44)	(2.37)	(2.75)	(3.74)	(5.66)	(6.17)	(6.94)	(7.19)	(7.25)	(7.26)
Panel D: All variables											
OLS	0.49^{***}	0.44^{***}	0.45^{***}	0.38^{***}	0.44^{***}	0.45^{***}	0.37^{***}	0.39^{***}	0.49^{***}	0.48^{***}	−0.01
	(3.45)	(3.89)	(4.33)	(3.59)	(4.03)	(4.71)	(3.79)	(3.66)	(4.8)	(3.9)	(−0.07)
NN	-0.21^{*}	0.20^{*}	0.28^{**}	0.27^{***}	0.35^{***}	0.51^{***}	0.50^{***}	0.61^{***}	0.73^{***}	1.12^{***}	1.33^{***}
	(−1.11)	(1.44)	(2.37)	(2.74)	(3.72)	(5.65)	(6.16)	(6.96)	(7.19)	(7.25)	(7.26)

Our focus is not on introducing the most powerful forecasting tool but on observing the economic value of nonlinearity. The prediction power of neural networks brings significant economic value. The separate information set enables us to examine how to return prediction responses to different types of information. Predictions based on accounting variables carry the largest returns, which beat other information in predicting monthly returns.

The established risk factors cannot explain the return spreads produced by machine learning. The regressors include the factors based on Fama and French (2016) and the momentum factor.[6] Table 3.5 illustrates the excess returns on portfolios, which is sorted based on return prediction for both linear regression and neural networks.

The information used in Table 3.5 Panel A is past returns. The risk-adjusted excess returns are positive and highly significant for neural networks, while the risk-adjusted excess return for linear regression is small and insignificant. The difference between linear regression and neural networks is the same for Table 3.5 Panels B, C, and D, in which we use accounting variables, anomalies, and all variables, respectively. Alphas are greatest for accounting variables, followed by past returns. Our results agree with those of Bartram and Grinblatt (2018), who also use accounting variables to generate mispricing signals. Overall, our results show that neural networks are better at predicting cross-sections of returns than linear regression. Risk-adjusted excess returns are the highest if using past returns and accounting variables.[7] However, improvement in alphas are the smallest for accounting variables set, compared to other information set.

3.5. Non-linear inefficiency

Based on the three measures, we build a composite measure to examine the summary of predictability for each set of information. Table 3.6 shows the correlations of the predictability measures.

[6] The data are from Kenneth French's website: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[7] Our results are robust if we only use past returns of stocks, without other variables constructed by market returns.

Table 3.6
Correlation of predictability measures

The table reports the correlation among predictability measures, which are portfolio return, cross-sectional R^2, pricing error, and the composite measures of the three proxies. Panels A, B, C, and D report average predictability measures for past returns, anomalies, accounting variables, and all variables, respectively. Column 'Linear regression' indicates the measures of linear regression. Column 'Neural network' reports the measures of neural networks.

	Linear regression			Neural network		
	Cross-sectional R^2	Portfolio return	Pricing error	Cross-sectional R^2	Portfolio return	Pricing error
Panel A: Past returns						
Cross-sectional R^2	1.00			1.00		
Portfolio return	0.69	1.00		0.64	1.00	
Pricing error	0.52	0.36	1.00	0.50	0.88	1.00
Composite measure	0.99	0.72	0.56	0.98	0.77	0.66
Panel B: Anomalies						
Cross-sectional R^2	1.00			1.00		
Portfolio return	0.87	1.00		0.74	1.00	
Pricing error	0.44	0.46	1.00	0.64	0.72	1.00
Composite measure	0.99	0.89	0.47	0.99	0.79	0.69
Panel C: Accounting variables						
Cross-sectional R^2	1.00			1.00		
Portfolio return	0.73	1.00		0.68	1.00	
Pricing error	0.61	0.82	1.00	0.48	0.85	1.00
Composite measure	0.99	0.80	0.68	0.99	0.78	0.60
Panel D: All variables						
Cross-sectional R^2	1.00			1.00		
Portfolio return	0.73	1.00		0.74	1.00	
Pricing error	0.61	0.82	1.00	0.64	0.72	1.00
Composite measure	0.99	0.80	0.68	0.99	0.79	0.68

Three measures are correlated, as shown in Table 3.6, which shows the consistency in statistical measures, such as R^2, and economic measures, such as excess return and return spread.

We report the significance of the average difference between neural networks and linear regression (or the non-linear inefficiency measured as defined in equation (25)) in Table 3.7. Although the return predictability captured by the non-linear models is not the same among the three input sets, we observe a highly significant and average positive difference between the neural network and linear regression prediction, which indicates a positive non-linear inefficiency on average.

Table 3.7 provides the monthly average values of the three separate predictability measures and the composite predictability measure. The first and second columns show the average predictability of linear regression and neural networks, respectively. The last column provides the average difference between the two models. The non-linear inefficiency in the all-variables combination is the greatest. On average, we find that the non-linear inefficiency in public information (accounting data and anomalies) is higher than that of past return information. This finding implies that most of the complexity reflected in asset price is in these information set. Between the two types of public information, we observe higher predictability in accounting variables than in factors and anomalies. However, the non-linear inefficiency is greater for anomalies than accounting variables.

Table 3.7
Average predictability measures by group

The table reports the results of the time-series average for three inverse proxies of market efficiency, which are portfolio return, cross-sectional R^2, pricing error, and the composite measures of the three proxies. Panels A, B, C, and D report average predictability measures for past returns, accounting variables, anomalies, and all variables, respectively. Column OLS indicates the measures of linear regression. Column NN reports the measures of neural networks. Column t-stat is the Newey-West adjusted standard error of the mean difference.

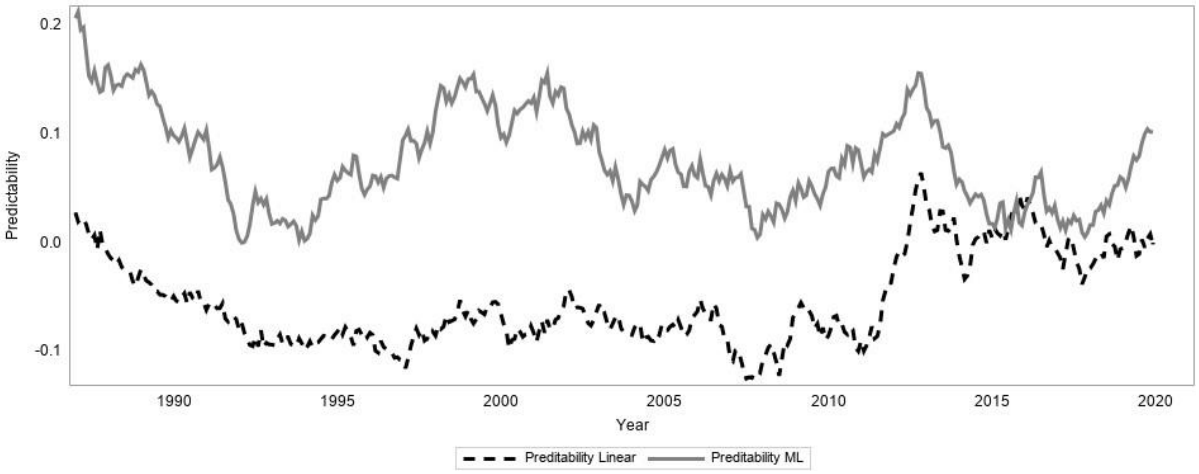
	OLS	NN	Difference	t-stat
	(1)	(2)	(2) − (1)	(2) − (1)
Panel A: Past returns				
Long-short return	0.00	0.02	0.02^{***}	9.64
Cross-sectional R^2	−0.01	0.20	0.22^{***}	12.34
Excess return	0.00	0.02	0.02^{***}	9.63
Composite measure	−0.00	0.08	0.09^{***}	11.84
Panel B: Accounting variables				
Long-short return	0.02	0.03	0.01^{***}	7.16
Cross-sectional R^2	0.23	0.28	0.05^{***}	4.76
Excess return	0.02	0.03	0.01^{***}	6.39
Composite measure	0.10	0.12	0.02^{***}	4.89
Panel C: Anomalies				
Long-short return	0.00	0.02	0.02^{***}	14.74
Cross-sectional R^2	−0.02	0.23	0.25^{***}	14.99
Excess return	0.00	0.01	0.01^{***}	13.79
Composite measure	−0.01	0.10	0.10^{***}	15.32
Panel D: All variables				
Long-short return	0.00	0.02	0.02^{***}	14.76
Cross-sectional R^2	0.02	0.23	0.25^{***}	15.01
Excess return	0.00	0.01	0.01^{***}	13.78
Composite measure	−0.01	0.10	0.10^{***}	15.35

Figure 3.3 Panel A illustrates the change in return predictability using weak-form information, which includes only historical returns of stocks. Most notably, the decreasing trend is noticeable only for neural network predictions, while linear predictions fluctuate around zero for all measures of predictability. Neural networks contribute more to the wedge between nonlinear and linear models. Figure 3.3 Panel B shows the change in the nonlinear gap between the neural network and linear regression. From the linear trend, we can observe a decrease in nonlinear inefficiency.

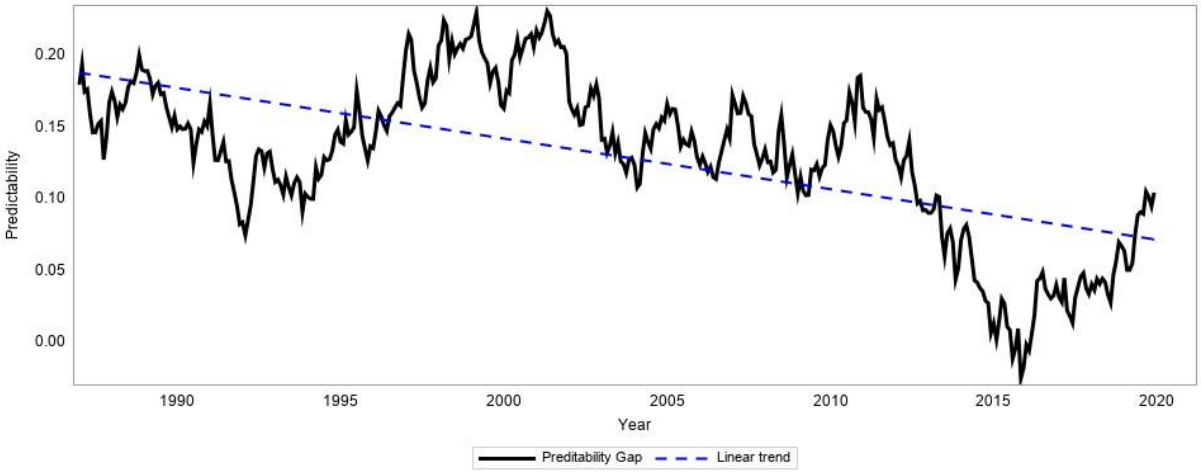
Figure 3.3
Return predictability using past return information

Panel A shows the 12-month moving average of the composite return predictability measures for linear regression and feed-forward neural network. The models are estimated using past returns information. Panel B graph shows the change in non-linear inefficiency over time.

(A) Return predictability



(B) Non-linear inefficiency



The closing gap between the two lines over time means that progressively, increasingly complex relations between information are being reflected in prices—markets are becoming more efficient not just in the types of information they reflect but also in the functional forms they capture. The results provide empirical evidence for the concept of technological market efficiency that was initially proposed by Grabowski (2019). Rösch et al. (2017) show that the efficacy of arbitrage mechanisms, market-making and financial friction govern market efficiency and price convergence toward their linear benchmark. Technological progress is an essential and natural evolution of the financial market that facilitates these mechanisms. Therefore, consistent with Rösch et al. (2017), we show that market efficiency should be treated as a dynamic rather than a static concept. Additionally, there is mounting pressure for asset pricing research to adjust market efficiency by adding nonlinearity to its definition.

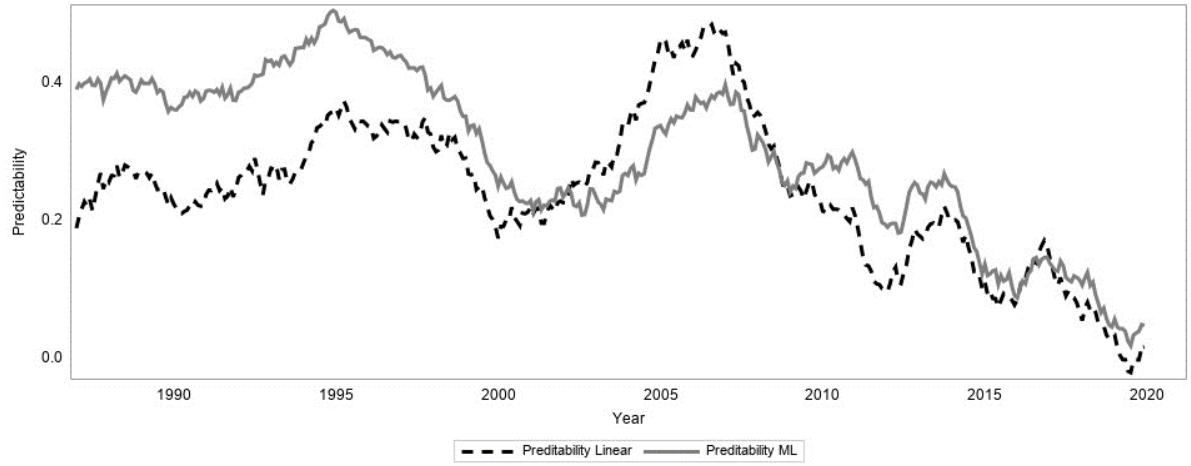
Figure 3.4 Panels A and B show the time series of predictability measures and non-linear inefficiency of accounting variables. The improvement in the performance of neural networks is the smallest in this set of variables. For some periods, non-linearity is even zero.[8] Non-linear inefficiency is also the smallest using this information set. However, we can still observe a decreasing trend in non-linear inefficiency.

[8] When non-linear inefficiency is smaller than 0, we treat it as equal to 0. When linear model performs better, it means that there is no non-linear inefficiency.

Figure 3.4
Return predictability using the accounting variables

Panel A shows the 12-month moving average of the composite return predictability measures for linear regression and feed-forward neural network. The models are estimated using accounting information. Panel B graph shows the change in non-linear inefficiency over time.

(A) Return predictability



(B) Non-linear inefficiency

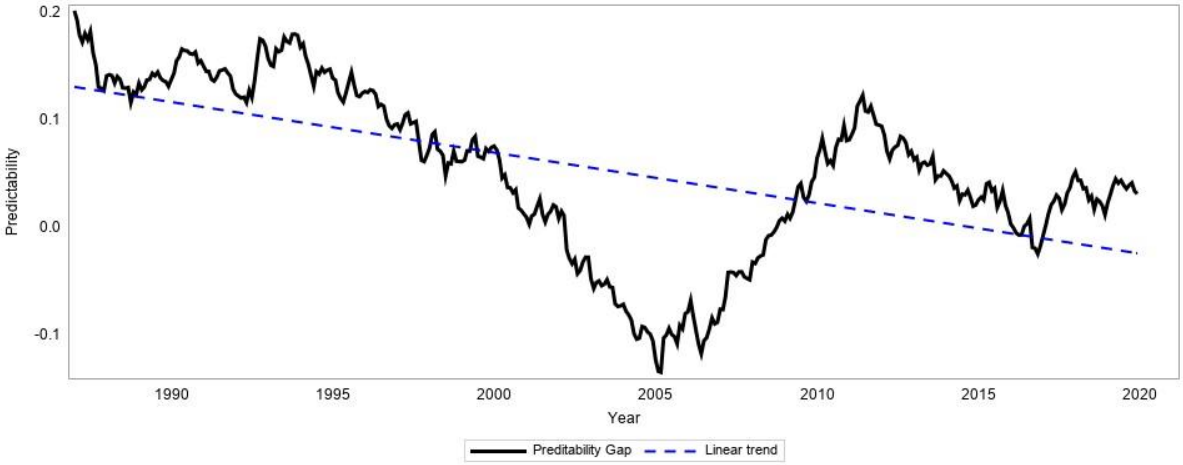
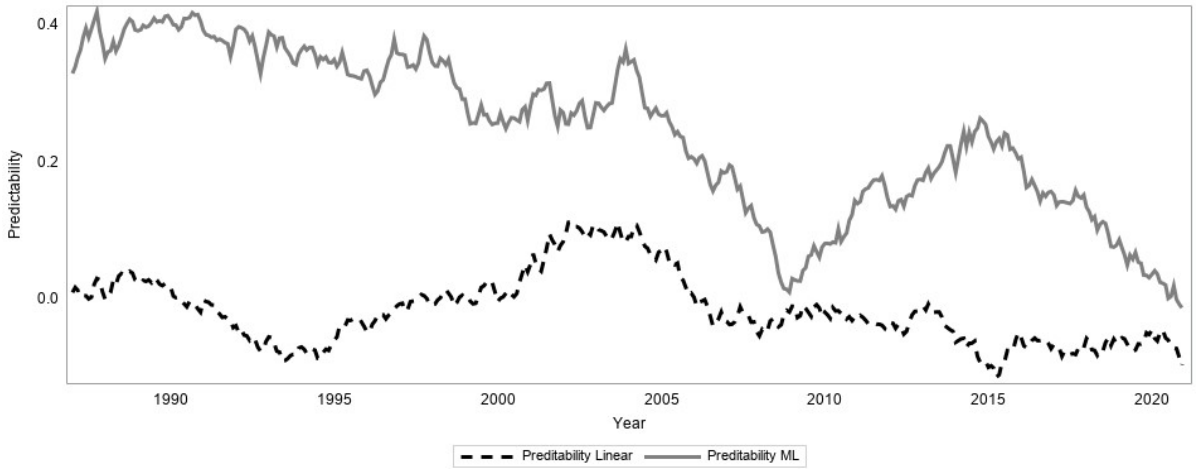


Figure 3.5 shows the time series of composite predictability measures using anomalies. Figure 3.5 Panel A illustrates the time-varying return predictability. The time-series behavior is the same as the weak-form information in Figure 3.3. However, the performance difference between the two models is higher for the anomalies set than the past returns set, as shown in Figure 3.5 Panel B.

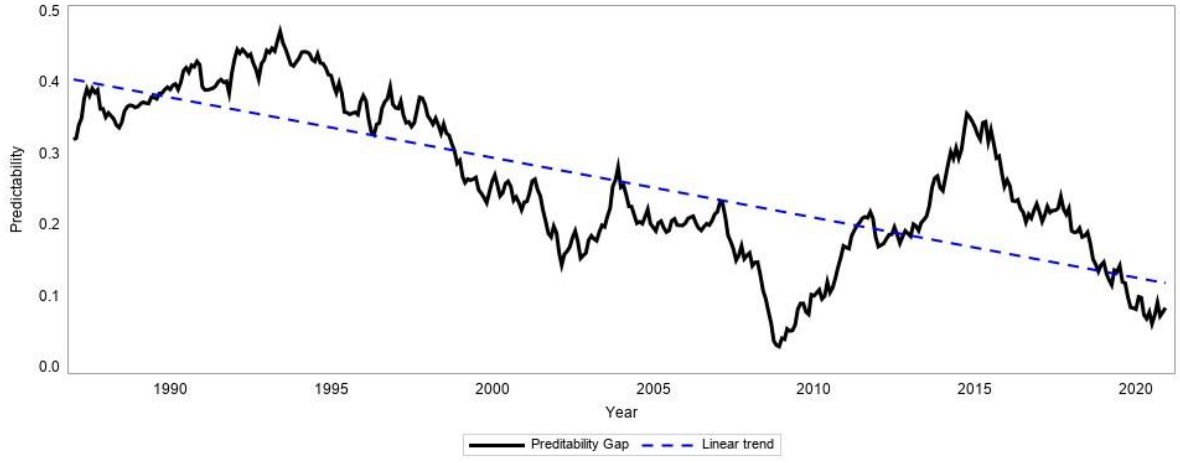
Figure 3.5
Return predictability using anomalies

Panel A shows the 12-month moving average of the composite return predictability measures using linear regression and a feed-forward neural network. The models are estimated using anomalies. Panel B shows the change in non-linear inefficiency over time.

(A) Return predictability



(B) Non-linear inefficiency

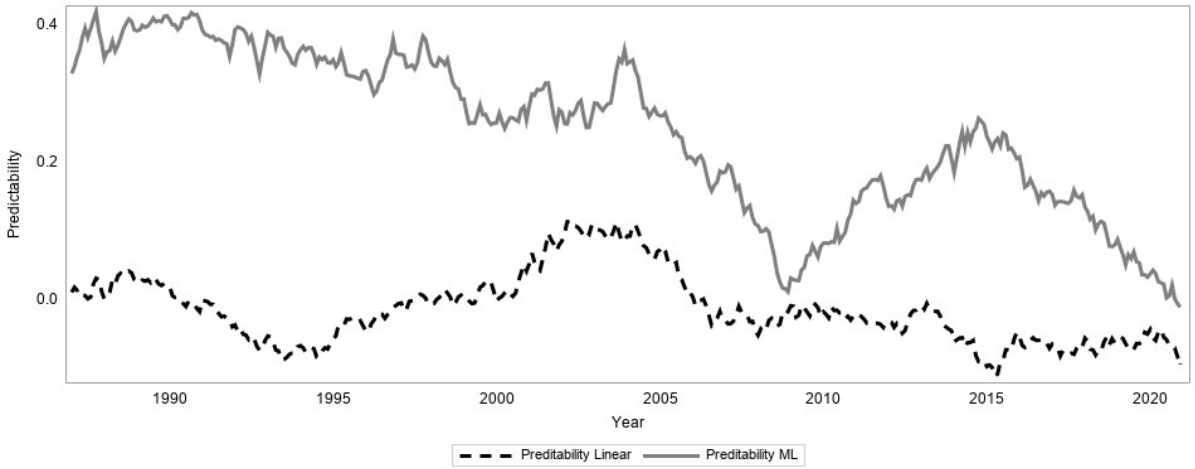


Finally, we combine all variables into our return prediction model. The return predictors used in this set are firm characteristics, accounting variables, and historical returns. Figure 3.6 Panel A illustrates the time-varying return predictability. The time series for the combination of all information is very similar to the pattern using anomalies in Figure 3.5. Figure 3.6 Panel B shows the change in non-linear inefficiency over time. This means that most of the return predictability and non-linear inefficiency are driven by anomalies and adding past returns or accounting information to anomalies does not bring much value to return predictability.

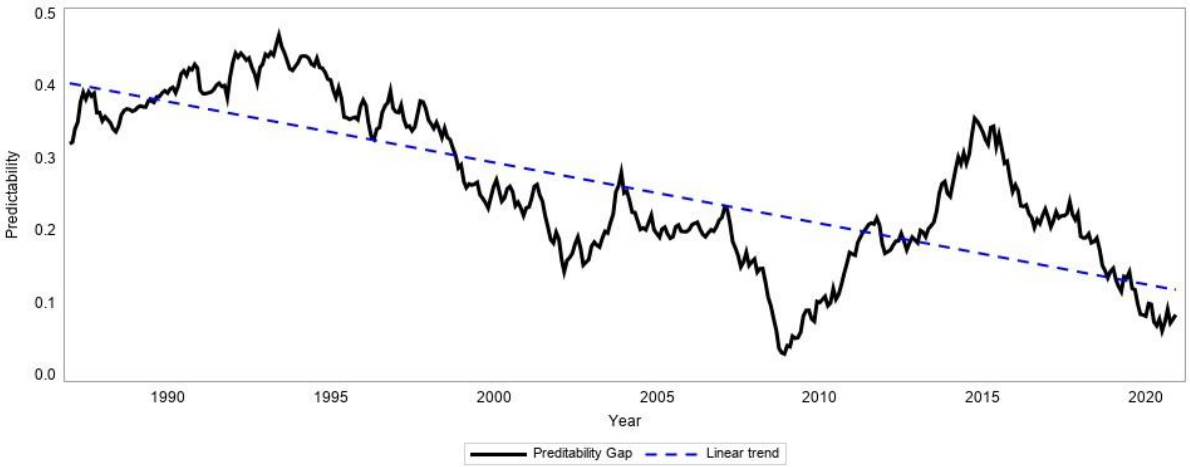
Figure 3.6
Return predictability using all variables

Panel A shows the 12-month moving average of the composite return predictability measures for linear regression and feed-forward neural network. The models are estimated using all information. Panel B graph shows the change in non-linear inefficiency over time.

(A) Return predictability



(B) Non-linear inefficiency



3.6. Drivers of the increase in non-linear efficiency

The three forces that are likely to affect market efficiency are research in finance, research in machine learning, and the number of quantitative funds. Research in finance proxies for more knowledge related to asset pricing and stock returns, which motivates the demand for sophisticated techniques to process more variables. Research in machine learning proxies for the ability to capture mispricing with new nonlinear techniques. Finally, growth of quantitative funds proxies for the actual application of research in both fields to exploit mispricing.

Following Abis (2020) and Beggs, Brogaard, and Hill-Kleespie (2021), we count the number of quantitative funds by extracting quantitative keywords from the mutual fund prospectus. Before 2016, we use prospectus on Morningstar Principia CDs (Kostovesky and Warner, 2020). After 2016, we use the Form 485s reported to the Securities Exchange Commission (SEC). Our sample is from 2000 to 2020. Prospectuses are published at least once every quarter. We only look for quantitative keywords, such as those in Appendix 3.C, which is recommended by Beggs et al. (2021) for the Principal Investment Strategies sections of the mutual fund prospectus after 2016. We exclude other sections in the prospectus, as suggested by Abis (2020). As there can be quantitative keywords in the risk and performance discussion of the prospectus, it is easy to over-identify the number of quantitative funds.

To measure the growth in machine learning, we extract data from peer-reviewed machine learning publications every year from artificial intelligence (AI) index reports (2021).[9] The data is sourced from Elsevier and Scopus, which are subscription-access scientific literature databases. The AI index report (2021) is one of the most comprehensive reports on AI to date. It significantly expands the amount of data available in the report, which is drawn from a broad set of academic, private, and non-profit organizations for calibration. The growth in finance research is measured as the number of finance papers on SSRN as in Dai et al. (2023).

We apply Hamilton's (2017) filter to de-trend the time series. This method is more reliable to control for time trends compared to other detrending approaches. Atanasov, Møller, and Priestley (2020) also use this method for time series regression. The procedure ensures that the component we use in our regression is stationary. We estimate regression with 3-month, 6-month, and 12-month lags of the explanatory variables (quantitative funds, finance papers, and machine learning papers, as it is likely that these variables take time to affect return predictability. We include interaction terms between technology research growth and

[9] Data was downloaded from Stanford AI Index Reports: https://aiindex.stanford.edu/report/

quantitative fund growth as advances in machine learning research enable the discovery of mispricing using new models, while the number of quantitative funds reflects the practical implementation of research to exploit these mispricing. The impact of growth of quantitative fund may be conditional on the growth of research in quantitative models. We estimate the following regression to examine the effect of the growth of finance and machine learning publications and quantitative funds on non-linear inefficiency:

$$NLIE_t = \beta_0 + \sum_{n=3,6,12} \beta_1 FP_{t-n} + \sum_{n=3,6,12} \beta_2 MP_{t-n} + \sum_{n=3,6,12} 3QP_{t-n} + \sum_{n=3,6,12} \beta_4 FP_{t-n} \times QF_{t-n} + \sum_{n=3,6,12} \beta_5 MP_{t-n} \times QF_{t-n} + \varepsilon_t \quad (35).$$

in which, $NLIE_t$ is the non-linear inefficiency measure at month t as defined in equation (25)

FP_{t-n} is the vector of lagged 3-month, 6-month, and 12-month number of finance papers on the SSRN database at time t

MP_{t-n} is the vector of lagged 3-month, 6-month, and 12-month number of machine learning publications at time t

QF_t is the vector of lagged 3-month, 6-month, and 12-month number of quantitative funds at time t

Table 3.8 shows that the decay of non-linearity inefficiency (increase in efficiency) is driven by the technology race of mutual funds and the prevalence of machine learning research over time. The technology race of funds is proxied by quantitative funds over time. The prevalence of machine learning and finance research over time is measured by the number of machine learning and finance publications over time. The significant negative coefficients on the interaction between the application of technology in asset managment and the number of AI and finance research indicate the detrimental effects of technology on non-linear inefficiency, especially when combining research knowledge and applying technology advances into investment practices.

Table 3.8
Drivers of non-linear inefficiency

The table reports the coefficient estimates and t-statistics of the potential driver of non-linear market efficiency. The independent variables are 3-month, 6-month, and 12-month lags in the number of machine learning publications (MP_{t-3}, MP_{t-6} and MP_{t-12}), finance publications (FP_{t-3}, FP_{t-6}, FP_{t-12}), and the number of quantitative mutual funds (QF_{t-3}, QF_{t-6}, QF_{t-12}). The dependent variable is $NLIE_t$, which is non-linear inefficiency (the performance difference between linear and machine learning model). ***, ** and * indicate significant at 1%, 5% and 10% level, respectively.

	$NLIE_t$	
	Parameter Estimate	t-stat
Intercept	0.09	$(9.26)^{***}$
FP_{t-3}	-2.50	$(-1.93)^{**}$
FP_{t-6}	-1.18	(-0.86)
FP_{t-12}	0.42	(0.64)
MP_{t-3}	0.34	(1.57)
MP_{t-6}	-0.01	(-0.04)
MP_{t-12}	-0.48	$(-2.42)^{**}$
QF_{t-3}	22.56	(1.16)
QF_{t-6}	33.35	(1.48)
QF_{t-12}	-14.36	(-1.03)
FP_{t-3}*QF_{t-3}	0.11	(0.92)
FP_{t-6}*QF_{t-6}	0.05	$(1.89)^{*}$
FP_{t-12}*QF_{t-12}	-0.07	$(-0.57)^{**}$
MP_{t-3}*QF_{t-3}	-0.06	(-1.94)
MP_{t-6}*QF_{t-6}	-0.12	(-1.46)
MP_{t-12}*QF_{t-12}	-0.07	$(-3.69)^{**}$
R^2		39%
Observations		179

3.7. Conclusion

Participants in financial markets face an array of data sources and data modelling techniques. Most conventional asset pricing studies implicitly assume that investors use linear models for returns. But in reality, information can have complex non-linear relations with asset prices. How much of the non-linearities and interactions between information are reflected in asset prices is the core issue examined in this chapter.

This chapter examines nonlinear market efficiency, measured by how well machine learning techniques, such as neural networks, can forecast out-of-sample stock returns relative to the performance of linear regression. We find that shallow networks, such as those in our study, have better results for simple information, such as past returns and accounting variables from financial statements, than published anomalies. The greatest performance gap (non-linear inefficiency) is in the combination of all information.

We find that the difference in out-of-sample return predictability using linear regression and neural networks disappears over time, consistent with an increase in the non-linear efficiency of the market. Most of the decrease in the performance differential can be attributed to a diminishing ability of the machine learning models to predict out-of-sample returns, rather than changes in the performance of linear models, consistent with markets through time becoming better at reflecting non-linear combinations of information.

Overall, our study provides empirical evidence of how the static market efficiency definition that is based on different information sets is challenged by the process of investor learning and the adoption of more sophisticated data science models.

Appendix 3.A: Firm characteristics

Table 3.A1
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
ChInvIA	Abarbanell and Bushee	1998	Change in capital inv (ind adj)	AR
ETR	Abarbanell and Bushee	1998	Effective Tax Rate	AR
GrGMToGrSales	Abarbanell and Bushee	1998	Gross margin growth to sales growth	AR
GrSaleToGrInv	Abarbanell and Bushee	1998	Sales growth over inventory growth	AR
GrSaleToGrOverhead	Abarbanell and Bushee	1998	Sales growth over overhead growth	AR
GrSaleToGrReceivables	Abarbanell and Bushee	1998	Change in sales vs change in receiv	AR
LaborforceEfficiency	Abarbanell and Bushee	1998	Laborforce efficiency	AR
pchgm_pchsale	Abarbanell and Bushee	1998	Change in gross margin vs sales	AR
betaCC	Acharya and Pedersen	2005	Illiquidity-illiquidity beta (beta2i)	JFE
betaCR	Acharya and Pedersen	2005	Illiquidity-market return beta (beta4i)	JFE
betaNet	Acharya and Pedersen	2005	Net liquidity beta (betanet,p)	JFE
betaRC	Acharya and Pedersen	2005	Return-market illiquidity beta	JFE
betaRR	Acharya and Pedersen	2005	Return-market return illiquidity beta	JFE
BetaBDLeverage	Adrian, Etula and Muir	2014	Broker-Dealer Leverage Beta	JF
IdioVolAHT	Ali, Hwang, and Trombley	2003	Idiosyncratic risk (AHT)	JFE
EarningsConsistency	Alwathainani	2009	Earnings consistency	BAR
Illiquidity	Amihud	2002	Amihud's illiquidity	JFM
BidAskSpread	Amihud and Mendelsohn	1986	Bid-ask spread	JFE
grcapx	Anderson and Garcia-Feijoo	2006	Change in capex (two years)	JF
grcapx1y	Anderson and Garcia-Feijoo	2006	Investment growth (1 year)	AR
grcapx3y	Anderson and Garcia-Feijoo	2006	Change in capex (three years)	JF
ForecastDispersionLT	Anderson, Ghysels, and Juergens	2005	Long-term forecast dispersion	RFS
betaVIX	Ang et al.	2006	Systematic volatility	JF

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
IdioVol3F	Ang et al.	2006	Idiosyncratic risk (3 factor)	JF
IdioVolCAPM	Ang et al.	2006	Idiosyncratic risk (CAPM)	JF
IdioVolQF	Ang et al.	2006	Idiosyncratic risk (q factor)	JF
CoskewACX	Ang, Chen and Xing	2006	Coskewness using daily returns	RFS
DownsideBeta	Ang, Chen and Xing	2006	Downside beta	RFS
IO_ShortInterest	Asquith Pathak and Ritter	2005	Inst own among high short interest	JFE
Mom6mJunk	Avramov et al	2007	Junk Stock Momentum	JF
OrderBacklogChg	Baik and Ahn	2007	Change in order backlog	Other
ChangeRoA	Balakrishnan, Bartov and Faurel	2010	Change in Return on assets	NA
ChangeRoE	Balakrishnan, Bartov and Faurel	2010	Change in Return on equity	NA
roaq	Balakrishnan, Bartov and Faurel	2010	Return on assets (qtrly)	JAE
MaxRet	Bali, Cakici, and Whitelaw	2010	Maximum return over month	JF
ReturnSkew	Bali, Engle and Murray	2015	Return skewness	Book
ReturnSkew3F	Bali, Engle and Murray	2015	Idiosyncratic skewness (3F model)	Book
ReturnSkewCAPM	Bali, Engle and Murray	2015	Idiosyncratic skewness (CAPM)	Book
ReturnSkewQF	Bali, Engle and Murray	2015	Idiosyncratic skewness (Q model)	Book
CBOperProf	Ball et al.	2016	Cash-based operating profitability	JFE
CBOperProfLagAT	Ball et al.	2016	Cash-based oper prof lagged assets	JFE
CBOperProfLagAT_q	Ball et al.	2016	Cash-based oper prof lagged assets qtrly	JFE
OperProfRD	Ball et al.	2016	Operating profitability R&D adjusted	JFE
OperProfRDLagAT	Ball et al.	2016	Oper prof R&D adj lagged assets	JFE
OperProfRDLagAT_q	Ball et al.	2016	Oper prof R&D adj lagged assets (qtrly)	JFE
Size	Banz	1981	Size	JFE
SP	Barbee, Mukherji and Raines	1996	Sales-to-price	FAJ
SP_q	Barbee, Mukherji and Raines	1996	Sales-to-price quarterly	FAJ
FirmAge	Barry and Brown	1984	Firm age based on CRSP	JFE

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
EP	Basu	1977	Earnings-to-Price Ratio	JF
EPq	Basu	1977	Earnings-to-Price Ratio	JF
hire	Bazdresch, Belo and Lin	2014	Employment growth	JPE
InvGrowth	Belo and Lin	2012	Inventory Growth	RFS
BrandCapital	Belo, Lin and Vitorino	2014	Brand capital to assets	RED
BrandInvest	Belo, Lin and Vitorino	2014	Brand capital investment	RED
Leverage	Bhandari	1988	Market leverage	JFE
Leverage_q	Bhandari	1988	Market leverage quarterly	JFE
ResidualMomentum	Blitz, Huij and Martens	2011	Momentum based on FF3 residuals	JEmpFin
ResidualMomentum6m	Blitz, Huij and Martens	2011	6 month residual momentum	JEmpFin
Price	Blume and Husic	1972	Price	JF
NetPayoutYield	Boudoukh et al.	2007	Net Payout Yield	JF
NetPayoutYield_q	Boudoukh et al.	2007	Net Payout Yield quarterly	JF
PayoutYield	Boudoukh et al.	2007	Payout Yield	JF
PayoutYield_q	Boudoukh et al.	2007	Payout Yield quarterly	JF
NetDebtFinance	Bradshaw, Richardson, Sloan	2006	Net debt financing	JAE
NetEquityFinance	Bradshaw, Richardson, Sloan	2006	Net equity financing	JAE
XFIN	Bradshaw, Richardson, Sloan	2006	Net external financing	JAE
DolVol	Brennan, Chordia, Subra	1998	Past trading volume	JFE
roic	Brown and Rowe	2007	Return on invested capital	WP
DelayAcct	Callen, Khan and Lu	2013	Accounting component of price delay	CAR
DelayNonAcct	Callen, Khan and Lu	2013	Non-accounting component of price delay	CAR
FailureProbability	Campbell, Hilscher and Szilagyi	2008	Failure probability	JF
FailureProbabilityJune	Campbell, Hilscher and Szilagyi	2008	Failure probability	JF
FEPS	Cen, Wei, and Zhang	2006	Analyst earnings per share	WP
AnnouncementReturn	Chan, Jegadeesh and Lakonishok	1996	Earnings announcement return	JF

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
REV6	Chan, Jegadeesh and Lakonishok	1996	Earnings forecast revisions	JF
AdExp	Chan, Lakonishok and Sougiannis	2001	Advertising Expense	JF
RD	Chan, Lakonishok and Sougiannis	2001	R&D over market cap	JF
RD_q	Chan, Lakonishok and Sougiannis	2001	R&D over market cap quarterly	JF
rd_sale	Chan, Lakonishok and Sougiannis	2001	R&D to sales	JF
rd_sale_q	Chan, Lakonishok and Sougiannis	2001	R&D to sales	JF
CashProd	Chandrashekar and Rao	2009	Cash Productivity	WP
invest	Chen and Zhang	2010	Capex and Inventory Change	JF,retracted
DelBreadth	Chen, Hong and Stein	2002	Breadth of ownership	JFE
std_turn	Chordia, Subra, Anshuman	2001	Share turnover volatility	JFE
VolSD	Chordia, Subra, Anshuman	2001	Volume Variance	JFE
CustomerMomentum	Cohen and Frazzini	2008	Customer momentum	JF
retConglomerate	Cohen and Lou	2012	Conglomerate return	JFE
RDAbility	Cohen, Diether and Malloy	2013	R&D ability	RFS
AssetGrowth	Cooper, Gulen and Schill	2008	Asset growth	JF
AssetGrowth_q	Cooper, Gulen and Schill	2008	Asset growth quarterly	JF
Activism1	Cremers and Nair	2005	Takeover vulnerability	JF
Activism2	Cremers and Nair	2005	Active shareholders	JF
EarningsForecastDisparity	Da and Warachka	2011	Long-vs-short EPS forecasts	JFE
CompEquIss	Daniel and Titman	2006	Composite equity issuance	JF
IntanBM	Daniel and Titman	2006	Intangible return using BM	JF
IntanCFP	Daniel and Titman	2006	Intangible return using CFtoP	JF
IntanEP	Daniel and Titman	2006	Intangible return using EP	JF
IntanSP	Daniel and Titman	2006	Intangible return using Sale2P	JF
ShareIss5Y	Daniel and Titman	2006	Share issuance (5 year)	JF
LRreversal	De Bondt and Thaler	1985	Long-run reversal	JF

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
MRreversal	De Bondt and Thaler	1985	Medium-run reversal	JF
ShortInterest	Dechow et al.	2001	Short Interest	JFE
EquityDuration	Dechow, Sloan and Soliman	2004	Equity Duration	RAS
cfp	Desai, Rajgopal, Venkatachalam	2004	Operating Cash flows to price	AR
cfpq	Desai, Rajgopal, Venkatachalam	2004	Operating Cash flows to price quarterly	AR
ZScore	Dichev	1998	Altman Z-Score	JFE
ZScore_q	Dichev	1998	Altman Z-Score quarterly	JFE
ForecastDispersion	Diether, Malloy and Scherbina	2002	EPS Forecast Dispersion	JF
BetaDimson	Dimson	1979	Dimson Beta	JFE
ExclExp	Doyle, Lundholm and Soliman	2003	Excluded Expenses	RAS
ProbInformedTrading	Easley, Hvidkjaer and O'Hara	2002	Probability of Informed Trading	JF
OrgCap	Eisfeldt and Papanikolaou	2013	Organizational capital	JF
OrgCapNoAdj	Eisfeldt and Papanikolaou	2013	Org cap w/o industry adjustment	JF
nanalyst	Elgers, Lo and Pfeiffer	2001	Number of analysts	AR
sfe	Elgers, Lo and Pfeiffer	2001	Earnings Forecast to price	AR
GrLTNOA	Fairfield, Whisenant and Yohn	2003	Growth in long term operating assets	AR
AM	Fama and French	1992	Total assets to market	JF
AMq	Fama and French	1992	Total assets to market (quarterly)	JF
BMdec	Fama and French	1992	Book to market using December ME	JPM
BookLeverage	Fama and French	1992	Book leverage (annual)	JF
BookLeverageQuarterly	Fama and French	1992	Book leverage (quarterly)	JF
OperProf	Fama and French	2006	operating profits / book equity	JFE
OperProfLag	Fama and French	2006	operating profits / book equity	JFE
OperProfLag_q	Fama and French	2006	operating profits / book equity	JFE
Beta	Fama and MacBeth	1973	CAPM beta	JPE
BetaSquared	Fama and MacBeth	1973	CAPM beta squred	JPE

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
EarningsSurprise	Foster, Olsen and Shevlin	1984	Earnings Surprise	AR
AccrualQuality	Francis, LaFond, Olsson, Schipper	2005	Accrual Quality	JAE
AccrualQualityJune	Francis, LaFond, Olsson, Schipper	2005	Accrual Quality in June	JAE
EarningsConservatism	Francis, LaFond, Olsson, Schipper	2004	Earnings conservatism	AR
EarningsPersistence	Francis, LaFond, Olsson, Schipper	2004	Earnings persistence	AR
EarningsPredictability	Francis, LaFond, Olsson, Schipper	2004	Earnings Predictability	AR
EarningsSmoothness	Francis, LaFond, Olsson, Schipper	2004	Earnings Smoothness	AR
EarningsTimeliness	Francis, LaFond, Olsson, Schipper	2004	Earnings timeliness	AR
EarningsValueRelevance	Francis, LaFond, Olsson, Schipper	2004	Value relevance of earnings	AR
roavol	Francis, LaFond, Olsson, Schipper	2004	RoA volatility	AR
AnalystValue	Frankel and Lee	1998	Analyst Value	JAE
AOP	Frankel and Lee	1998	Analyst Optimism	JAE
IntrinsicValue	Frankel and Lee	1998	Intrinsic or historical value	JAE
PredictedFE	Frankel and Lee	1998	Predicted Analyst forecast error	JAE
FR	Franzoni and Marin	2006	Pension Funding Status	JF
FRbook	Franzoni and Marin	2006	Pension Funding Status	JF
BetaFP	Frazzini and Pedersen	2014	Frazzini-Pedersen Beta	JFE
High52	George and Hwang	2004	52 week high	JF
IndMom	Grinblatt and Moskowitz	1999	Industry Momentum	JFE
AbnormalAccrualsPercent	Hafzalla, Lundholm, Van Winkle	2011	Percent Abnormal Accruals	AR
PctAcc	Hafzalla, Lundholm, Van Winkle	2011	Percent Operating Accruals	AR
PctTotAcc	Hafzalla, Lundholm, Van Winkle	2011	Percent Total Accruals	AR
tang	Hahn and Lee	2009	Tangibility	JF
tang_q	Hahn and Lee	2009	Tangibility quarterly	JF
Coskewness	Harvey and Siddique	2000	Coskewness	JF
CapTurnover	Haugen and Baker	1996	Capital turnover	JFE

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
CapTurnover_q	Haugen and Baker	1996	Capital turnover (quarterly)	JFE
RoE	Haugen and Baker	1996	net income / book equity	JFE
VarCF	Haugen and Baker	1996	Cash-flow to price variance	JFE
VolMkt	Haugen and Baker	1996	Volume to market equity	JFE
VolumeTrend	Haugen and Baker	1996	Volume Trend	JFE
AnalystRevision	Hawkins, Chamberlin, Daniel	1984	EPS forecast revision	FAJ
Mom12mOffSeason	Heston and Sadka	2008	Momentum without the seasonal part	JFE
MomOffSeason	Heston and Sadka	2008	Off season long-term reversal	JFE
MomOffSeason06YrPlus	Heston and Sadka	2008	Off season reversal years 6 to 10	JFE
MomOffSeason11YrPlus	Heston and Sadka	2008	Off season reversal years 11 to 15	JFE
MomOffSeason16YrPlus	Heston and Sadka	2008	Off season reversal years 16 to 20	JFE
MomSeason	Heston and Sadka	2008	Return seasonality years 2 to 5	JFE
MomSeason06YrPlus	Heston and Sadka	2008	Return seasonality years 6 to 10	JFE
MomSeason11YrPlus	Heston and Sadka	2008	Return seasonality years 11 to 15	JFE
MomSeason16YrPlus	Heston and Sadka	2008	Return seasonality years 16 to 20	JFE
MomSeasonShort	Heston and Sadka	2008	Return seasonality last year	JFE
NOA	Hirshleifer et al.	2004	Net Operating Assets	JAE
dNoa	Hirshleifer, Hou, Teoh, Zhang	2004	change in net operating assets	JAE
depr	Holthausen and Larcker	1992	Depreciation to PPE	JAE
pchdepr	Holthausen and Larcker	1992	Change in depreciation to PPE	JAE
EarnSupBig	Hou	2007	Earnings surprise of big firms	RFS
IndRetBig	Hou	2007	Industry return of big firms	RFS
BidAskTAQ	Hou and Loh	2016	Bid-ask spread (TAQ)	JFE
PriceDelayRsq	Hou and Moskowitz	2005	Price delay r square	RFS
PriceDelaySlope	Hou and Moskowitz	2005	Price delay coeff	RFS
PriceDelayTstat	Hou and Moskowitz	2005	Price delay SE adjusted	RFS

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
Herf	Hou and Robinson	2006	Industry concentration (sales)	JF
HerfAsset	Hou and Robinson	2006	Industry concentration (assets)	JF
HerfBE	Hou and Robinson	2006	Industry concentration (equity)	JF
STreversal	Jegadeesh	1989	Short term reversal	JF
RevenueSurprise	Jegadeesh and Livnat	2006	Revenue Surprise	JFE
Mom12m	Jegadeesh and Titman	1993	Momentum (12 month)	JF
Mom6m	Jegadeesh and Titman	1993	Momentum (6 month)	JF
ChangeInRecommendation	Jegadeesh et al.	2004	Change in recommendation	JF
OptionVolume1	Johnson and So	2012	Option to stock volume	JFE
OptionVolume2	Johnson and So	2012	Option volume to average	JFE
BetaTailRisk	Kelly and Jiang	2014	Tail risk beta	RFS
fgr5yrLag	La Porta	1996	Long-term EPS forecast	JF
fgr5yrNoLag	La Porta	1996	Long-term EPS forecast (Monthly)	JF
CF	Lakonishok, Shleifer, Vishny	1994	Cash flow to market	JF
CFq	Lakonishok, Shleifer, Vishny	1994	Cash flow to market quarterly	JF
MeanRankRevGrowth	Lakonishok, Shleifer, Vishny	1994	Revenue Growth Rank	JF
sgr	Lakonishok, Shleifer, Vishny	1994	Annual sales growth	JF
sgr_q	Lakonishok, Shleifer, Vishny	1994	Annual sales growth quarterly	JF
KZ	Lamont, Polk and Saa-Requejo	2001	Kaplan Zingales index	RFS
KZ_q	Lamont, Polk and Saa-Requejo	2001	Kaplan Zingales index quarterly	RFS
RDS	Landsman et al.	2011	Real dirty surplus	AR
Tax	Lev and Nissim	2004	Taxable income to income	AR
Tax_q	Lev and Nissim	2004	Taxable income to income (qtrly)	AR
RDcap	Li	2011	R&D capital-to-assets	RFS
zerotrade	Liu	2006	Days with zero trades	JFE
zerotradeAlt1	Liu	2006	Days with zero trades	JFE

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
zerotradeAlt12	Liu	2006	Days with zero trades	JFE
ChEQ	Lockwood and Prombutr	2010	Growth in book equity	JFR
EarningsStreak	Loh and Warachka	2012	Earnings surprise streak	MS
NumEarnIncrease	Loh and Warachka	2012	Earnings streak length	MS
GrAdExp	Lou	2014	Growth in advertising expenses	RFS
EntMult	Loughran and Wellman	2011	Enterprise Multiple	JFQA
EntMult_q	Loughran and Wellman	2011	Enterprise Multiple quarterly	JFQA
CompositeDebtIssuance	Lyandres, Sun and Zhang	2008	Composite debt issuance	RFS
InvestPPEInv	Lyandres, Sun and Zhang	2008	change in ppe and inv/assets	RFS
DivYield	Naranjo, Nimalendran, Ryngaert	1998	Dividend yield for small stocks	JF
DivYieldAnn	Naranjo, Nimalendran, Ryngaert	1998	Last year's dividends over price	NA
Frontier	Nguyen and Swanson	2009	Efficient frontier index	JFQA
GP	Novy-Marx	2013	gross profits / total assets	JFE
GPlag	Novy-Marx	2013	gross profits / total assets	JFE
GPlag_q	Novy-Marx	2013	gross profits / total assets	JFE
IntMom	Novy-Marx	2012	Intermediate Momentum	JFE
OPLeverage	Novy-Marx	2010	Operating leverage	ROF
OPLeverage_q	Novy-Marx	2010	Operating leverage (qtrly)	ROF
AssetLiquidityBook	Ortiz-Molina and Phillips	2014	Asset liquidity over book assets	JFQA
AssetLiquidityBookQuart	Ortiz-Molina and Phillips	2014	Asset liquidity over book (qtrly)	JFQA
AssetLiquidityMarket	Ortiz-Molina and Phillips	2014	Asset liquidity over market	JFQA
AssetLiquidityMarketQuart	Ortiz-Molina and Phillips	2014	Asset liquidity over market (qtrly)	JFQA
cashdebt	Ou and Penman	1989	CF to debt	JAR
currat	Ou and Penman	1989	Current Ratio	JAR
pchcurrat	Ou and Penman	1989	Change in Current Ratio	JAR
pchquick	Ou and Penman	1989	Change in quick ratio	JAR

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
pchsaleinv	Ou and Penman	1989	Change in sales to inventory	JAR
quick	Ou and Penman	1989	Quick ratio	JAR
salecash	Ou and Penman	1989	Sales to cash ratio	JAR
saleinv	Ou and Penman	1989	Sales to inventory	JAR
salerec	Ou and Penman	1989	Sales to receivables	JAR
Cash	Palazzo	2012	Cash to assets	JFE
BetaLiquidityPS	Pastor and Stambaugh	2003	Pastor-Stambaugh liquidity beta	JPE
BPEBM	Penman, Richardson and Tuna	2007	Leverage component of BM	JAR
EBM	Penman, Richardson and Tuna	2007	Enterprise component of BM	JAR
EBM_q	Penman, Richardson and Tuna	2007	Enterprise component of BM	JAR
NetDebtPrice	Penman, Richardson and Tuna	2007	Net debt to price	JAR
NetDebtPrice_q	Penman, Richardson and Tuna	2007	Net debt to price	JAR
PS	Piotroski	2000	Piotroski F-score	AR
PS_q	Piotroski	2000	Piotroski F-score	AR
ShareIss1Y	Pontiff and Woodgate	2008	Share issuance (1 year)	JF
DelDRC	Prakash and Sinha	2012	Deferred Revenue	CAR
OrderBacklog	Rajgopal, Shevlin, Venkatachalam	2003	Order backlog	RAS
DelCOA	Richardson et al.	2005	Change in current operating assets	JAE
DelCOL	Richardson et al.	2005	Change in current operating liabilities	JAE
DelEqu	Richardson et al.	2005	Change in equity to assets	JAE
DelFINL	Richardson et al.	2005	Change in financial liabilities	JAE
DelLTI	Richardson et al.	2005	Change in long-term investment	JAE
DelNetFin	Richardson et al.	2005	Change in net financial assets	JAE
DelSTI	Richardson et al.	2005	Change in short-term investment	JAE
TotalAccruals	Richardson et al.	2005	Total accruals	JAE
AgeIPO	Ritter	1991	IPO and age	JF

Table 3.A1 (continued)
List of firm characteristics
The table summarises the return predictors (firm characteristics) from asset pricing studies that are inputs of our models.

Acronym	Authors	Year	LongDescription	Journal
BM	Rosenberg, Reid, and Lanstein	1985	Book to market using most recent ME	JF
BMq	Rosenberg, Reid, and Lanstein	1985	Book to market (quarterly)	JF
Accruals	Sloan	1996	Accruals	AR
AssetTurnover	Soliman	2008	Asset Turnover	AR
AssetTurnover_q	Soliman	2008	Asset Turnover	AR
ChAssetTurnover	Soliman	2008	Change in Asset Turnover	AR
ChNCOA	Soliman	2008	Change in Noncurrent Operating Assets	AR
ChNCOL	Soliman	2008	Change in Noncurrent Operating Liab	AR
ChNNCOA	Soliman	2008	Change in Net Noncurrent Op Assets	AR
ChNWC	Soliman	2008	Change in Net Working Capital	AR
ChPM	Soliman	2008	Change in Profit Margin	AR
PM	Soliman	2008	Profit Margin	AR
PM_q	Soliman	2008	Profit Margin	AR
RetNOA	Soliman	2008	Return on Net Operating Assets	AR
RetNOA_q	Soliman	2008	Return on Net Operating Assets	AR
ChInv	Thomas and Zhang	2002	Inventory Growth	RAS
ChTax	Thomas and Zhang	2011	Change in Taxes	JAR
Investment	Titman, Wei and Xie	2004	Investment to revenue	JFQA
realestate	Tuzel	2010	Real estate holdings	RFS
secured	Valta	2016	Secured debt	JFQA
WW	Whited and Wu	2006	Whited-Wu index	RFS
WW_Q	Whited and Wu	2006	Whited-Wu index	RFS
AbnormalAccruals	Xie	2001	Abnormal Accruals	AR
skew1	Xing, Zhang and Zhao	2010	Volatility smirk near the money	JFQA
SmileSlope	Yan	2011	Put volatility minus call volatility	JFE
FirmAgeMom	Zhang	2004	Firm Age - Momentum	JF

Appendix 3.B: Accounting Variables

Table 3.A2
List of accounting variables
The table summarises the accounting variables that are inputs to our models.

Variable	Definition
ATQ	Assets - Total - Quarterly
DVPQ	Dividends - Prefered/Preference - Quarterly
SALEQ	Sales/Turnover (Net) - Quarterly
SEQQ	Stockholders Equity - Total - Quarterly
IBQ	Income Before Extraordinary Items - Quarterly
NIQ	Net Income (Loss) - Quarterly
XIDOQ	Extraordinary Items and Discontinued Operations - Quarterly
IBADJQ	Income Before Extraordinary Items - Adjusted for Common Stock Equivalents - Quarterly
IBCOMQ	Income Before Extraordinary Items - Available for Common - Quarterly
ICAPTQ	Invested Capital - Total - Quarterly
TEQQ	Stockholders Equity - Total - Quarterly
PSTKRQ	Preferred/Preference Stock - Redeemable - Quarterly
PPENTQ	Property Plant and Equipment - Total (Net) - Quarterly
CEQQ	Common/Ordinary Equity - Total - Quarterly
PSTKQ	Preferred/Preference Stock (Capital) - Total - Quarterly
DLTTQ	Long-Term Debt - Total - Quarterly
PIQ	Pretax Income - Quarterly
TXTQ	Income Taxes - Total - Quarterly
NOPIQ	Nonoperating Income (Expense) - Quarterly
AOQ	Assets - Other - Total - Quarterly
LTQ	Liabilities - Total - Quarterly
DOQ	Discontinued Operations - Quarterly
LOQ	Liabilities - Other - Total - Quarterly
CHEQ	Cash and Short-Term Investments - Quarterly
ACOQ	Current Assets - Other - Total - Quarterly
DVQ	Cash Dividends (Cash Flow) - Quarterly
LCOQ	Current Liabilities - Other - Total - Quarterly
APQ	Accounts Payable - Quarterly
DPQ	Depreciation - Quarterly
COGSQ	Cost of Goods Sold - Quarterly

Appendix 3.C: Quantitative keywords

quantitative investment, quantitative model, quantitative analysis, quantitative process, quantitative tools, quantitative formula, quantitative computer, statistically driven, statistical methods, quantitative methodology, quantitative management, quantitative method, quantitative models, quantitative analytics, quantitatively-driven, quantitatively-derived, quantitative approach, quantitative value, quantitative statistics, quantitatively investing, quantitative measures, quantitative techniques, quantitative research, quantitative methods, quantitative, factor-based, quantitative three factor, quantitative approaches, quantitative computer valuation, quantitative optimization, quantitatively driven, quantitative studies, quantitative computer valuation, quantitatively assess, quantitative assessment, quantitative research, quantitatively-oriented, multi-factor, multifactor, multi-factor.

Appendix 3.D: Time-series tests of portfolios

Table 3.A3

Time-series tests of portfolios based on firm characteristics by linear regression

The table reports the results of time-series regressions for excess returns on portfolios sorted by predictions of next month's return. The test portfolios are across rows and explanatory variables are across columns. The first column "α" reports excess return in percent; the regressors include five Fama-French factors plus momentum. The last column reports adjusted R^2. The sample is from Jan 1985 to December 2020. T-statistics are reported in parentheses. *, ** and *** indicate significant at 10%, 5% and 1% level, respectively.

Portfolio	α	R_m-R_f	SMB	HML	RMW	CMA	Mom	R^2
P1	0.49	0.97	0.75	−0.02	−0.34	0.02	−0.30	88%
	(3.45)**	(24.31)**	(12.92)**	(-0.30)	(−5.00)**	(0.22)	(−5.20)**	
P2	0.44	0.91	0.67	0.05	−0.23	−0.04	−0.24	91%
	(3.89)**	(29.42)**	(14.34)**	(0.80)	(−4.27)**	(−0.53)	(−5.46)**	
P3	0.45	0.88	0.66	0.06	−0.18	−0.03	−0.22	90%
	(4.33)**	(31.17)**	(14.80)**	(1.17)	(−3.85)**	(−0.41)	(−5.72)**	
P4	0.38	0.85	0.60	0.06	−0.15	−0.02	−0.17	90%
	(3.60)**	(29.47)**	(14.24)**	(1.07)	(−3.26)**	(−0.32)	(−4.15)**	
P5	0.44	0.82	0.56	0.04	−0.14	−0.05	−0.16	89%
	(4.01)**	(26.96)**	(13.33)**	(0.79)	(−2.91)**	(−0.62)	(−3.59)**	
P6	0.45	0.82	0.59	0.02	−0.13	0.05	−0.15	90%
	(4.73)**	(30.82)**	(15.14)**	(0.34)	(−2.61)*	(0.62)	(−4.74)**	
P7	0.37	0.83	0.63	0.06	−0.08	0.03	−0.15	89%
	(3.78)**	(30.08)**	(16.91)**	(0.85)	(−1.73)	(0.37)	(−4.35)**	
P8	0.39	0.86	0.64	0.02	−0.09	0.08	−0.17	90%
	(3.67)**	(29.35)**	(14.44)**	(0.26)	(−1.96)	(1.05)	(−4.00)**	
P9	0.49	0.88	0.66	0.02	−0.10	0.10	−0.16	89%
	(4.79)**	(32.03)**	(14.13)**	(0.25)	(−1.94)	(1.21)	(−4.43)**	
P10	0.48	0.91	0.70	0.03	−0.19	0.13	−0.19	86%
	(3.91)**	(29.59)**	(13.29)**	(0.38)	(−2.81)**	(1.32)	(−4.46)**	
P10−P1	−0.01	−0.06	−0.05	0.05	0.15	0.11	0.11	25%
	(−0.09)	(−2.11)**	(−1.07)	(1.02)	(1.98)*	(1.40)	(2.56)**	

Table 3.A4

Time-series tests of portfolios based on firm characteristics by neural network

The table reports the results of time-series regressions for excess returns on portfolios sorted by predictions of next month's return. The test portfolios are across rows and explanatory variables are across columns. The first column "α" reports excess return in percent; the regressors include five Fama-French factors plus momentum. The last column reports adjusted R^2. The sample is from Jan 1985 to December 2020. T-statistics are reported in parentheses. *, ** and *** indicate significant at 10%, 5% and 1% level, respectively.

Portfolio	α	R_m-R_f	SMB	HML	RMW	CMA	Mom	R^2
P1	−0.21	1.00	0.81	−0.15	−0.45	−0.25	−0.47	88%
	(−1.11)	(21.25)**	(12.17)**	(−2.17)**	(−4.73)**	(−2.36)**	(−5.65)**	
P2	0.20	0.94	0.71	-0.03	−0.25	−0.16	−0.33	90%
	(1.44)	(23.83)**	(15.01)**	(-0.44)	(−3.88)**	(−1.75)	(−5.39)**	
P3	0.29	0.87	0.64	0.00	−0.06	−0.11	−0.26	90%
	(2.37)*	(25.26)**	(15.05)**	(0.01)	(−1.21)	(−1.40)	(−5.42)**	
P4	0.27	0.87	0.60	0.04	−0.02	−0.01	−0.18	91%
	(2.75)**	(29.95)**	(14.74)**	(0.75)	(−0.41)	(−0.20)	(−5.18)**	
P5	0.35	0.83	0.56	0.05	−0.06	0.05	−0.13	91%
	(3.74)**	(29.47)**	(13.87)**	(0.82)	(−1.37)	(0.75)	(−3.83)**	
P6	0.51	0.82	0.57	0.09	−0.05	0.03	−0.12	91%
	(5.66)**	(33.49)**	(15.14)**	(1.38)	(−1.30)	(0.49)	(−3.64)**	
P7	0.50	0.85	0.59	0.11	−0.06	0.08	−0.10	91%
	(6.17)**	(34.95)**	(14.63)**	(2.09)*	(−1.56)	(1.27)	(−3.93)**	
P8	0.61	0.84	0.62	0.08	−0.07	0.14	−0.10	90%
	(6.94)**	(33.28)**	(15.43)**	(1.35)	(−1.57)	(2.03)*	(−3.69)**	
P9	0.73	0.85	0.63	0.14	−0.20	0.13	−0.08	88%
	(7.19)**	(30.21)**	(13.50)**	(2.02)*	(−4.07)**	(1.60)	(−2.52)*	
P10	1.12	0.87	0.75	0.00	−0.40	0.37	−0.12	79%
	(7.25)**	(21.99)**	(10.50)**	(0.00)	(−4.80)**	(2.75)**	(−2.56)*	
P10−P1	1.33	−0.13	−0.05	0.15	0.05	0.62	0.35	38%
	(7.26)**	(−2.55)*	(−0.77)	(1.56)*	(0.42)	(5.05)**	(4.71)**	

Table 3.A5

Time-series tests of portfolios based on past returns by linear regression

The table reports the results of time-series regressions for excess returns on portfolios sorted by predictions of next month's return. The test portfolios are across rows and explanatory variables are across columns. The first column "α" reports excess return in percent; the regressors include five Fama-French factors plus momentum. The last column reports adjusted R^2. The sample is from Jan 1985 to December 2020. T-statistics are reported in parentheses. *, ** and *** indicate significant at 10%, 5% and 1% level, respectively.

Portfolio	α	R_m-R_f	SMB	HML	RMW	CMA	Mom	R^2
P1	0.58	0.95	0.86	−0.07	−0.43	0.11	−0.32	83%
	(3.49)**	(21.67)**	(11.37)**	(−0.68)	(−4.94)**	(0.75)	(−4.98)**	
P2	0.58	0.92	0.79	0.02	−0.21	0.07	−0.21	88%
	(4.87)**	(27.02)**	(14.60)**	(0.31)	(−3.29)**	(0.65)	(−4.65)**	
P3	0.50	0.94	0.72	0.06	−0.14	0.05	−0.19	91%
	(5.16)**	(33.51)**	(15.35)**	(0.97)	(−2.51)*	(0.66)	(−4.76)**	
P4	0.53	0.92	0.70	0.05	−0.10	0.12	−0.17	93%
	(5.43)**	(35.80)**	(18.92)**	(1.03)	(−2.05)*	(1.75)	(−4.36)**	
P5	0.55	0.92	0.69	0.07	−0.08	0.03	−0.16	93%
	(5.64)**	(36.43)**	(19.18)**	(1.39)	(−1.80)	(0.48)	(−3.93)**	
P6	0.66	0.90	0.68	0.07	−0.09	−0.01	−0.19	91%
	(5.84)**	(32.95)**	(14.41)**	(1.35)	(−1.83)	(−0.12)	(−3.87)**	
P7	0.60	0.90	0.68	0.07	−0.09	−0.02	−0.16	90%
	(4.86)**	(31.00)**	(13.45)**	(1.31)	(−1.47)	(−0.30)	(−2.71)**	
P8	0.62	0.89	0.73	0.05	−0.13	−0.06	−0.17	89%
	(4.62)**	(27.39)**	(12.91)**	(0.80)	(−1.70)	(−0.81)	(−2.57)*	
P9	0.52	0.92	0.75	0.02	−0.25	−0.03	−0.17	88%
	(3.96)**	(27.30)**	(11.44)**	(0.29)	(−3.02)**	(−0.42)	(−3.08)**	
P10	0.76	0.91	0.84	0.07	−0.56	−0.14	−0.21	84%
	(4.50)**	(22.26)**	(9.66)**	(0.83)	(−6.10)**	(−1.18)	(−3.04)**	
P10−P1	0.18	−0.04	−0.01	0.14	−0.13	−0.25	0.11	6%
	(1.41)	(−1.15)	(−0.21)	(2.00)*	(−1.29)*	(-2.28)*	(1.40)	

Table 3.A6

Time-series tests of portfolios based on past returns by neural network

The table reports the results of time-series regressions for excess returns on portfolios sorted by predictions of next month's return. The test portfolios are across rows and explanatory variables are across columns. The first column "α" reports excess return in percent; the regressors include five Fama-French factors plus momentum. The last column reports adjusted R^2. The sample is from Jan 1985 to December 2020. T-statistics are reported in parentheses. *, ** and *** indicate significant at 10%, 5% and 1% level, respectively.

Portfolio	α	R_m-R_f	SMB	HML	RMW	CMA	Mom	R^2
P1	−0.30	0.96	0.80	−0.17	−0.60	0.10	−0.39	88%
	(−2.24)*	(26.39)**	(11.15)**	(−2.04)*	(−7.11)**	(0.77)	(−7.87)**	
P2	0.16	0.94	0.73	−0.01	−0.25	0.10	−0.25	92%
	(1.68)	(34.36)**	(14.79)**	(−0.12)	(−4.54)**	(1.11)	(−7.13)**	
P3	0.36	0.90	0.68	0.09	−0.16	0.05	−0.17	91%
	(3.83)**	(35.70)**	(13.05)**	(1.64)	(−3.29)**	(0.71)	(−6.12)**	
P4	0.42	0.89	0.69	0.12	−0.07	0.04	−0.12	92%
	(5.32)*	(39.52)**	(17.59)**	(2.25)**	(−1.75)	(0.61)	(−4.54)**	
P5	0.58	0.88	0.70	0.13	−0.06	0.04	−0.11	92%
	(7.40)**	(41.75)**	(14.30)**	(2.95)**	(−1.34)	(0.71)	(−4.64)**	
P6	0.60	0.88	0.67	0.12	−0.07	0.02	−0.07	91%
	(6.59)**	(35.16)**	(13.87)**	(2.47)*	(−1.47)	(0.32)	(−2.28)*	
P7	0.60	0.91	0.70	0.11	−0.06	0.02	−0.04	91%
	(6.29)**	(41.39)**	(14.85)**	(2.50)*	(−1.14)	(0.30)	(−1.21)	
P8	0.66	0.92	0.72	0.10	−0.07	−0.01	−0.08	90%
	(5.98)**	(34.75)**	(15.55)**	(1.68)*	(−1.11)	(−0.12)	(−1.63)*	
P9	0.87	0.92	0.81	0.04	−0.17	−0.09	−0.12	86%
	(5.68)**	(24.43)**	(14.44)**	(0.51)	(−2.15)*	(−0.83)	(−1.88)	
P10	1.95	0.94	0.94	−0.14	−0.57	−0.17	−0.59	69%
	(5.13)**	(10.62)**	(8.44)**	(−0.80)	(−3.68)**	(−0.67)	(−3.51)**	
P10−P1	2.25	-0.03	0.14	0.03	0.03	−0.27	−0.20	5%
	(6.68)**	(-0.32)	(1.31)	(0.14)	(0.22)	(−1.07)	(−1.31)	

Table 3.A7
Time-series tests of portfolios based on accounting variables by linear regression

The table reports the results of time-series regressions for excess returns on portfolios sorted by predictions of next month's return. The test portfolios are across rows and explanatory variables are across columns. The first column "α" reports excess return in percent; the regressors include five Fama-French factors plus momentum. The last column reports adjusted R^2. The sample is from Jan 1985 to December 2020. T-statistics are reported in parentheses. *, ** and *** indicate significant at 10%, 5% and 1% level, respectively.

Portfolio	α	R_m-R_f	SMB	HML	RMW	CMA	Mom	R^2
P1	−0.09	1.00	0.84	−0.22	−0.68	−0.15	0.09	91%
	(−1.05)	(44.43)**	(20.44)**	(−4.93)**	(−13.91)**	(−2.35)*	(4.38)**	
P2	−0.03	1.01	0.72	−0.09	−0.23	−0.04	−0.02	95%
	(−0.47)	(62.00)**	(17.97)**	(−2.09)*	(−5.34)**	(−0.72)	(−1.41)	
P3	0.01	0.98	0.67	0.00	−0.07	−0.02	−0.07	97%
	(0.25)	(59.83)**	(19.22)**	(0.09)	(−1.65)	(−0.48)	(−3.85)**	
P4	0.13	0.98	0.66	0.06	−0.03	0.00	−0.10	96%
	(1.95)	(58.42)**	(19.51)**	(1.73)	(−0.74)	(0.07)	(−3.97)**	
P5	0.15	0.96	0.64	0.08	−0.04	0.00	−0.13	95%
	(1.95)	(44.29)**	(15.73)**	(1.93)	(−0.74)	(−0.05)	(−4.10)**	
P6	0.31	0.92	0.70	0.06	−0.05	0.03	−0.18	93%
	(2.82)*	(30.98)**	(14.14)**	(1.06)	(−0.96)	(0.35)	(−3.92)**	
P7	0.50	0.86	0.76	0.05	−0.13	0.05	−0.23	89%
	(3.42)**	(21.84)**	(12.91)**	(0.69)	(−1.86)	(0.54)	(−4.01)**	
P8	1.06	0.79	0.80	0.09	−0.22	0.03	−0.31	83%
	(5.58)**	(15.16)**	(10.89)**	(0.81)	(−2.53)**	(0.22)	(−4.21)**	
P9	1.76	0.79	0.79	0.13	−0.28	0.11	−0.39	79%
	(7.57)**	(13.54)**	(8.85)**	(1.26)	(−2.97)**	(0.72)	(−4.02)**	
P10	2.09	0.85	0.86	0.23	−0.37	0.09	−0.61	74%
	(8.00)**	(13.73)**	(8.46)**	(1.66)	(−3.31)**	(0.51)	(−5.65)**	
P10−P1	2.18	−0.15	0.02	0.46	0.31	0.24	−0.69	57%
	(8.42)**	(−2.36)*	(0.15)	(3.10)**	(2.40)*	(1.38)	(−6.54)**	

Table 3.A8
Time-series tests of portfolios based on accounting variables by neural network

The table reports the results of time-series regressions for excess returns on portfolios sorted by predictions of next month's return. The test portfolios are across rows and explanatory variables are across columns. The first column "α" reports excess return in percent; the regressors include five Fama-French factors plus momentum. The last column reports adjusted R^2. The sample is from Jan 1985 to December 2020. T-statistics are reported in parentheses. *, ** and *** indicate significant at 10%, 5% and 1% level, respectively.

Portfolio	α	R_m-R_f	SMB	HML	RMW	CMA	Mom	R^2
P1	−0.34	0.94	0.87	−0.17	−0.59	0.00	−0.02	92%
	(−3.63)**	(39.54)**	(17.83)**	(−3.15)**	(−9.37)**	(−0.04)	(−0.58)	
P2	−0.11	0.94	0.63	−0.07	−0.23	0.02	−0.06	96%
	(−1.55)	(47.70)**	(16.19)**	(−1.59)	(−4.61)	(0.25)	(−2.60)*	
P3	0.05	0.93	0.60	0.02	−0.09	0.00	−0.07	96%
	(0.73)	(57.59)**	(18.39)**	(0.68)	(−2.85)**	(0.06)	(−3.81)**	
P4	0.14	0.93	0.63	0.06	−0.04	0.04	−0.07	96%
	(2.09)*	(53.26)**	(19.52)**	(1.57)	(−1.17)	(0.94)	(−3.30)**	
P5	0.24	0.93	0.66	0.08	−0.01	0.03	−0.12	95%
	(3.04)**	(45.75)**	(16.14)**	1.95	(−0.13)	(0.56)	(−4.18)**	
P6	0.39	0.91	0.71	0.16	−0.06	−0.05	−0.16	94%
	(4.19)**	(33.04)**	(16.47)**	(2.95)**	(−1.15)	(−0.69)	(−4.46)**	
P7	0.62	0.91	0.78	0.12	−0.08	0.02	−0.21	91%
	(5.28)**	(28.71)**	(15.67)**	(2.01)*	(−1.39)	(0.29)	(−4.52)**	
P8	0.92	0.88	0.79	0.10	−0.18	0.01	−0.28	87%
	(5.37)**	(19.62)**	(11.70)**	(1.00)	(−2.29)*	(0.09)	(−3.82)**	
P9	1.58	0.86	0.83	0.05	−0.29	0.01	−0.37	81%
	(6.66)**	(15.20)**	(8.76)**	(0.44)	(−2.62)**	(0.07)	(−3.68)**	
P10	2.43	0.91	0.96	0.03	−0.54	0.01	−0.57	74%
	(8.13)**	(12.96)**	(7.88)**	(0.20)	(−3.86)**	(0.06)	(−4.43)**	
P10−P1	2.78	−0.03	0.09	0.20	0.05	0.01	−0.54	28%
	(9.61)**	(−0.38)	(0.80)	(1.25)	(0.33)	(0.07)	(−4.16)**	

Table 3.A9

Time-series tests of portfolios based on all variables by linear regression

The table reports the results of time-series regressions for excess returns on portfolios sorted by predictions of next month's return. The test portfolios are across rows and explanatory variables are across columns. The first column "α" reports excess return in percent; the regressors include five Fama-French factors plus momentum. The last column reports adjusted R^2. The sample is from Jan 1985 to December 2020. T-statistics are reported in parentheses. *, ** and *** indicate significant at 10%, 5% and 1% level, respectively.

Portfolio	α	R_m-R_f	SMB	HML	RMW	CMA	Mom	R^2
P1	0.49	0.97	0.75	−0.02	−0.34	0.02	−0.30	88%
	(3.45)**	(24.33)**	(12.93)**	(−0.29)	(−4.98)**	(0.21)	(−5.20)**	
P2	0.44	0.91	0.67	0.05	−0.23	−0.04	−0.24	91%
	(3.89)**	(29.37)**	(14.32)**	(0.80)	(−4.27)**	(−0.53)	(−5.45)**	
P3	0.45	0.88	0.66	0.06	−0.18	−0.03	−0.22	90%
	(4.33)**	(31.22)**	(14.79)*	(1.16)	(−3.84)**	(−0.40)	(−5.75)**	
P4	0.38	0.85	0.60	0.06	−0.15	−0.02	−0.17	90%
	(3.59)**	(29.41)**	(14.24)**	(1.06)	(−3.27)**	(−0.32)	(−4.14)**	
P5	0.44	0.82	0.56	0.04	−0.14	−0.05	−0.16	89%
	(4.03)**	(27.09)**	(13.36)**	(0.79)	(−2.91)**	(−0.62)	(−3.62)**	
P6	0.45	0.82	0.59	0.02	−0.13	0.05	−0.15	89%
	(4.71)**	(30.63)**	(15.12)**	(0.33)	(−2.61)**	(0.62)	(−4.69)**	
P7	0.37	0.83	0.63	0.06	−0.08	0.03	−0.15	90%
	(3.79)**	(30.11)**	(16.89)**	(0.85)	(−1.73)	(0.37)	(−4.36)**	
P8	0.39	0.86	0.64	0.02	−0.09	0.08	−0.17	90%
	(3.66)**	(29.28)**	(14.44)**	(0.26)	(−1.96)*	(1.05)	(−3.99)**	
P9	0.49	0.88	0.66	0.02	−0.10	0.10	−0.16	89%
	(4.80)**	(32.08)**	(14.13)**	(0.25)	(−1.94)	(1.21)	(−4.45)**	
P10	0.48	0.91	0.70	0.03	−0.19	0.13	−0.19	86%
	(3.90)**	(29.51)**	(13.26)**	(0.38)	(−2.81)**	(1.31)	(−4.44)**	
P10−P1	−0.01	−0.06	−0.05	0.05	0.15	0.11	0.11	20%
	(−0.07)	(−2.11)*	(-1.07)	(1.01)	(1.97)*	(1.42)	(2.56)*	

Table 3.A10

Time-series tests of portfolios based on all variables by neural network

The table reports the results of time-series regressions for excess returns on portfolios sorted by predictions of next month's return. The test portfolios are across rows and explanatory variables are across columns. The first column "α" reports excess return in percent; the regressors include five Fama-French factors plus momentum. The last column reports adjusted R^2. The sample is from Jan 1985 to December 2020. T-statistics are reported in parentheses. *, ** and *** indicate significant at 10%, 5% and 1% level, respectively.

Portfolio	α	R_m-Rf	SMB	HML	RMW	CMA	Mom	R^2
P1	−0.21	1.00	0.81	−0.15	-0.45	−0.25	−0.47	87%
	(−1.11)	(21.27)**	(12.17)**	(−2.16)*	(-4.73)**	(−2.36)*	(−5.66)**	
P2	0.20	0.94	0.71	−0.03	−0.25	−0.16	−0.33	90%
	(1.44)	(23.76)**	(15.00)**	(−0.44)	(−3.88)**	(−1.76)	(−5.39)**	
P3	0.28	0.87	0.64	0.00	−0.06	−0.11	−0.26	90%
	(2.37)**	(25.34)**	(15.07)**	(0.01)	(−1.22)	(−1.40)	(−5.43)**	
P4	0.27	0.87	0.60	0.04	−0.02	-0.01	−0.18	91%
	(2.74)**	(29.90)**	(14.75)**	(0.75)	(−0.41)	(-0.20)	(−5.17)**	
P5	0.35	0.83	0.56	0.05	−0.06	0.05	−0.13	91%
	(3.72)**	(29.38)**	(13.81)**	(0.82)	(−1.37)	(0.74)	(−3.80)**	
P6	0.51	0.82	0.56	0.09	−0.05	0.03	−0.12	91%
	(5.65)**	(33.38)**	(15.15)**	(1.37)	(−1.30)	(0.48)	(−3.63)**	
P7	0.50	0.85	0.59	0.11	−0.06	0.08	−0.10	91%
	(6.16)**	(34.96)**	(14.66)**	(2.10)*	(−1.56)	(1.28)	(−3.93)**	
P8	0.61	0.84	0.62	0.08	−0.07	0.14	−0.10	90%
	(6.96)**	(33.27)**	(15.41)**	(1.35)	(−1.57)	(2.03)*	(−3.70)**	
P9	0.73	0.85	0.63	0.14	−0.20	0.13	−0.08	88%
	(7.19)**	(30.22)**	(13.51)**	(2.02)*	(−4.07)**	(1.60)	(−2.53)*	
P10	1.12	0.87	0.75	0.00	−0.40	0.37	−0.12	78%
	(7.25)**	(21.98)**	(10.48)**	(0.01)	(−4.81)**	(2.75)**	(−2.56)**	
P10−P1	1.33	−0.13	−0.05	0.15	0.05	0.62	0.35	40%
	(7.26)**	(−2.55)*	(−0.77)	(1.56)	(0.41)	(5.05)**	(4.71)**	

Chapter 4: Detecting layering and spoofing in markets

"Markets today are almost entirely electronic, and algorithms aren't as savvy as their flesh-and-blood counterparts" – Bloomberg (2015)

4.1. Introduction

Throughout the course of financial market history, market manipulation has created vibrant patterns, ranging from traditional tactics like corners and squeezes, to more contemporary methods such as pump-and-dump, benchmark manipulation, and quote stuffing in recent years. New types of market manipulation have evolved and become prevalent with the development of algorithmic and low-latency trading, including spoofing and layering, which are the focus of this chapter. Although there has been a sharp spike in the number of prosecution cases of layering and spoofing in recent years, these forms of manipulation have received relatively little research, particularly the issues of how to detect this form of manipulation and distinguish it from legitimate trading.

Spoofing entails the use of orders with no genuine intention to execute them ("non-bona-fide orders") to influence financial market prices. For example, a trader may place large non-bona-fide buy orders into a market to create a false impression of buying interest, thereby pushing market prices up and causing a better execution price for their sell order. Layering refers to the use of several spoofing orders placed as "layers" in the limit order book to create the impression that several traders have interest in buying or selling. However, some regulators use the two terms interchangeably as both refer to the use of non-bona-fide orders to deceive other traders about supply and demand.

Prosecutions of layering and spoofing have increased substantially in recent years. According to the Financial Times (2018), the number of spoofing cases increased by a factor of five in 2018, including both criminal and civil enforcement actions, compared to 2017. In 2020, JP Morgan was instructed to make a historic settlement payment of $920.2 million due to their involvement in spoofing activities in the metals and Treasury futures markets.

The individuals involved in the prosecuted spoofing cases vary widely, ranging from an individual day-trader conducting trades from their bedroom to some of the biggest institutions on Wall Street. The techniques also range from manual entry of orders to fully automated algorithmic strategies. This variation across real cases prompts the questions: (i)

what are the empirical characteristics of layering and spoofing that are found across cases, i.e., what is the empirical "fingerprint" left in the data by this form of market manipulation, and (ii) how accurately can empirical measures of those characteristics detect layering/spoofing and distinguish it from legitimate trading? We address both questions in this chapter.

We start by hand collecting the most comprehensive database of layering/spoofing. We do so by systematically identifying prosecution cases from markets all around the world during a 10-year period (2010−2020). We extract case information from regulatory and court documents, and where insufficient information is provided, we obtain further details via Freedom of Information (FOI) requests. Using these sources, we compile records of known instances of layering and spoofing with varying degrees of granularity, but in many cases down to the level of individual trades and orders.

We first use the database of layering/spoofing to provide a descriptive anatomy of each case, focusing on the circumstances and motivation, the trading strategy, and any peculiarities or aspects that differ from other cases. A "typical" case of spoofing involves the manipulator placing a relatively minor order at or close to the best available quote on one side of the market (the bona-fide order), followed by the placement of a large order or multiple orders on the opposite side of the market (the illegitimate order(s)). This tactic aims to generate an order imbalance and exert influence over the market. The bona-fide order will then execute if the non-bona-fide orders have a sufficient influence on the market. Finally, the manipulator will typically cancel the non-bona-fide orders following the execution of the bona-fide order. This pattern can be repeated multiple times, alternating between spoofing the buy and sell sides of the market, as each occurrence may only result in a modest profit. However, it is also possible for layering/spoofing to be used in a one-off manipulation to improve a trade execution price.

In analyzing the cases, we find substantial variation in how layering and spoofing is implemented by different traders in different markets. Some use computer algorithms, others place orders manually. Some use multiple layering orders, others prefer to use a single large spoofing order. Some place spoofing orders at the best quotes, others stay back from the best quotes to reduce the chance of the non-bona-fide orders executing. Some repeat the layering/spoofing cycles rapidly switching between the buy and sell sides of the market, while others undertake isolated instances or repeat cycles less frequently.

Despite these differences, we also identify important commonalities across the cases. Non-genuine spoofing orders are typically cancelled before they execute, leading to high

cancellation or order amendment rates. Manipulators create substantial order imbalances in the limit order book. Manipulators tend to execute trades on the opposite side of the limit order book to the non-bona-fide orders and often manipulators repeat the layering/spoofing activities in cyclical patterns, switching between spoofing the buy and sell sides of the market.

Next, we construct daily and intraday empirical metrics to capture the common characteristics of spoofing. We apply logistic regression to test which of the empirical metrics can detect spoofing instances at daily and intraday horizons. In these tests, we use out-of-sample cross-validation to test the accuracy of the metrics. We also estimate machine learning (random forest) prediction models to capture non-linearities between the empirical metrics.

We find that the proposed empirical metrics are able to detect spoofing at both daily and intraday horizons but are more accurate when drawing on intraday data. Of the daily metrics, order book imbalance and the frequency of the imbalance switching sides (from a buy-side imbalance to a sell-side imbalance, and vice versa) are the most effective in predicting spoofing. At intraday horizons, the metrics that are effective in detecting spoofing are order imbalance, abnormal cancellation rates, trades that are on the opposite side to high quoting activity, and cancels that are on the opposite side to a trade.

Spoofing can be regarded as a form of bluffing in markets. It has therefore likely existed for decades if not centuries. However, recent evolution of markets towards increased automation of trading has made spoofing and layering more widespread. These developments have prompted legislative reforms that clarify that the practice is considered illegal, such as the Dodd-Frank Wall Street Reform and Consumer Protection Act.[10] They also explain the rapid rise in the number of prosecution cases, as noted earlier.

The reasons why automation in markets has driven an increase in spoofing are twofold. First, order execution algorithms are often programmed to consider the order book depth (volume of orders in the limit order book) on both sides of the market when deciding whether to place their next order as a passive limit order or to "cross the spread" with a market order and demand liquidity.[11] This programmed behavior of execution algorithms creates a

[10] The Dodd Frank Act (for short) amended the Commodity Exchange Act and adopted an explicit prohibition against spoofing, which it defined as "bidding or offering with the intent to cancel the bid or offer before execution."

[11] Typically, an execution algorithm that is trying to buy a security will be more passive (e.g., use limit orders) when there are more sell orders in the limit order book compared to the volume of buy orders, and more aggressive

predictable response to limit order book conditions that can be profitably exploited through spoofing. Effectively, the success of spoofing algorithms in today's markets is in part due to the presence of other algorithmic traders that can be exploited.

Second, a single instance of spoofing may yield only a small profit. For example, with a bid of \$10.00 and ask of \$10.20, a spoofing cycle that manages to buy 500 shares at the ask and sell them at the bid earns a profit of 500 × \$0.20 = \$100, less fees. Yet if that small profit can be earned hundreds or thousands of times in a day, then the strategy can become highly profitable. Automation enables spoofing strategies to be repeated many times in an efficient manner and has therefore increased the potential profits from spoofing.

This chapter contributes to a broader literature on market manipulation.[12] Most theoretical studies of market manipulation focus on when and how market manipulation is possible and how it affects market quality. For example, Jarrow (1992) and Cherian and Jarrow (1995), investigate the possibility of market manipulation by large traders with market power. Cooper and Glen (1998) and Allen, Litov, and Mei (2006) study traditional market manipulation techniques such as corners and squeezes in which manipulators control prices by obtaining a significant fraction of the supply. Merrick Jr, Naik, and Yadav (2005) model the differences in settlement between the spot and futures markets, creating favorable conditions for squeezes. Eren and Ozsoylev (2006) explore the economic circumstances under which hype-and-dump manipulation is possible. Hanson and Oprea (2009) examine the impacts of manipulators on price accuracy. Similarly, Allen and Gale (1992) use the Glosten and Milgrom (1985) framework to show that trade-based manipulation is possible even without taking actions to alter a firm's value or releasing false information.

Spoofing is examined theoretically by Cartea, Jaimungal, and Wang (2020) and Williams and Skrzypacz (2021). Cartea et al. (2020) show in a model how a trader wanting to sell can achieve a better sale price by using spoofing orders on the buy side of the limit order book to influence the market. They also analyze how fines imposed by regulators affect spoofing revenues, concluding there is a deterrence effect. Williams and Skrzypacz (2021) show that spoofing exists in equilibrium, slows price discovery, raises bid-ask spreads, and raises return volatility, supporting regulatory concerns about this type of trading. They

(e.g., use market orders) when there are more buy orders in the limit order book compared to the volume of sell orders, and vice versa when trying to sell.

[12] For a survey of the literature, see Putniņš (2020).

conclude that spoofing is likely to be most prevalent in markets with a medium level of liquidity.

Using an agent-based simulation model, Wang et al. (2021) find that simple spoofing strategies can mislead traders, distort prices, and reduce welfare. They propose two approaches to mitigating spoofing: (i) mechanism design to disincentivize manipulation; and (ii) changes to trading strategies to improve the robustness of learning from market information.

There is limited empirical research on spoofing or layering. One exception is Lee, Eom, and Park (2013) who examine spoofing in the Korean Stock Exchange (KRX). The spoofing they identify exploits a very specific feature of the KRX, where orders in the limit order book placed well outside of the best quotes with very low execution probability would still get reported in the aggregate limit order book depth. Wang (2019) finds that in the Taiwan Index Futures Market spoofing tends to increase volatility, volume, and influences prices in the direction of the spoofing orders. Debie et al. (2023) examine the JP Morgan spoofing case, proposing that an alternative motivation for the spoofing orders may have been to attract liquidity rather than to influence the price. While their paper provides a useful case study, they only discuss one spoofing case and focus on visual methodology. Our chapter contributes to these empirical studies of spoofing by analyzing a comprehensive global database of spoofing cases and developing/validating metrics to detect spoofing.

Our chapter also contributes to empirical studies of other modern abusive trading techniques. Bernhardt and Davies (2009) demonstrate that mutual fund managers have incentives to influence closing prices at the end of reporting periods. Comerton-Forde and Putniņš (2011, 2014) characterize the effects of closing price manipulation on the US and Canadian stock exchanges and develop a measure of the probability of closing price manipulation. Friederich and Payne (2015), Egginton, Ness, and Ness (2016), and Khomyn and Putniņš (2021) investigate the puzzling high order-to-trade ratios and cancelation rates in markets and whether they may be explained by spoofing or other legitimate trading strategies. Zhai, Cao and Ding (2018) use data mining methods to detect an abnormal pattern in quoting and trading activities of rogue traders. Dhawan and Putniņš (2022) examine pump-and-dump manipulation in cryptocurrency markets.

The findings of this chapter have practical applications in market surveillance and regulation. In enforcing anti-spoofing provisions, proving illegal spoofing has occurred is challenging, and leads to a long time from investigation and prosecution to a final conviction.

The spoofing characteristics that we identify may complement the current statutory and regulatory guidelines for how to identify spoofing.

This chapter proceeds as follows – the next section defines layering and spoofing based on legislation and existing literature. Following that, we characterize the common features of spoofing strategies based on the hand-collected sample of prosecution cases and discuss unique features of the cases to gauge the variability between cases. The next section defines and tests the empirical metrics used to detect spoofing, after which we summarize the conclusions.

4.2. What are layering and spoofing?

For clarity on what it is that we seek to empirically characterize, we first describe the legal definitions of layering and spoofing as well as the way this form of market manipulation is defined in prior literature.

Spoofing is specified as a criminal and civil offence in several jurisdictions. For example, in the US regulatory framework, provisions in the Dodd-Frank Act 2010 (Section 747), Commodity Exchange Act (Section 4c(a)(5)(C)), Securities Exchange Act (Section 10(b) and 9(a)(2)), Security Act 1993 (Section 17(a)), SEC Rule 10b-5, and the FINRA Rule 2020 may be used to enforce spoofing-like behavior.

Section 747 of the Dodd-Frank Wall Street Reform and Consumer Protection Act contains an explicit anti-spoofing provision that prohibits individuals from engaging in any trading, practice, or conduct that is commonly known as "spoofing". It provides a short definition of "spoofing" as "bidding or offering with the intent to cancel the bid or offer before cancellation".

Section 9(a)(2) Prohibition Against Manipulation of Security Prices of the Securities Exchange Act 1934 does not directly provide a definition of spoofing. However, it states that "it is unlawful for any person, directly or indirectly, by the use of the mails or any means or instrumentality of interstate commerce, or of any facility of any national securities exchange, or for any member of a national securities exchange, to effect, alone or with one or more other persons, a series of transactions in any security creating apparent active trading in such security, or raising or depressing the price of such security, to induce the purchase or sale of such security by others".

In 2010, the CFTC publishes a guideline that clarifies some definitions and particularities of spoofing, inviting public comments. In 2011, the CFTC publishes a Proposed Interpretive Order regarding the Dodd-Frank Act to specify the intent requirement of spoofing conduct. According to the proposed guidance, a violator must act with some degree of requisite intent or scienter in that they intended to cancel the bid or ask before execution. Reckless trading or conduct is insufficient. Additionally, cancellation of orders as part of a "legitimate, good faith attempt to consummate trade" is inadequate to meet the requirement of spoofing.

The 2013 CFTC final interpretive guidance adds four specific (non-exclusive) examples of conduct that constitute spoofing: submitting or cancelling bids or offers to overload the quotation system, delaying another person's execution of trades, creating the appearance of false market depth, and making artificial price movements upwards or downwards.

Many other jurisdictions prohibit spoofing through general market manipulation provisions but do not provide an explicit definition of spoofing in legislation. For example, the European legal framework does not explicitly define spoofing apart from more general prohibited manipulative conduct. Spoofing can contravene civil or regulatory provisions in the EU Market Abuse Regulation (596/2014) and can be a criminal offence under the Financial Services Act 2012 and the Fraud Act 2006. In a relevant part of the provisions, Article 12 of the EU Market Abuse Regulation (596/2014) defines a type of market manipulation as behavior that "gives, or is likely to give, false or misleading signals as to the supply of, demand for, or price of, a financial instrument, or a related spot commodity contract." Ultimately, spoofing falls within this description of prohibited conduct.

In a taxonomy of market manipulation types, spoofing is one of the order-based manipulation techniques (Putniņš, 2020). It involves submitting orders to a market to cancel them before they execute (Putniņš, 2020). Orders play a central role in this manipulation, although trades accompany them as part of the strategy. Although spoofing can be conducted manually, this manipulation is often implemented by computer algorithms that submit and cancel orders.

Layering is one type of spoofing strategy that involves submitting multiple orders designed to be cancelled (Putniņš, 2020). Orders are placed in layers across several price levels or on top of one another at a given price step, and cancellations happen on one side of order book.

In placing the spoofing orders, manipulators face a tradeoff – if they are submitted far away from the best prices, they may have limited influences on the market, yet if they are placed at the best quotes, they may inadvertently be executed, contrary to the manipulator's intention. Therefore, in balancing these two considerations, spoofing orders are typically placed close to or at the best price but behind other orders at that price and then dynamically amended or cancelled as the market moves and as the spoofing orders become more likely to be executed.

It is well established in the market microstructure literature that the order book information, such as the limit orders sitting in the book, affects the trading decisions of market participants. For example, many studies find that when there is more depth on the bid (buy) side, buyers respond by increasing the aggressiveness of their orders: placing buy orders at higher prices and/or using market orders to execute their trades immediately rather than being patient and using limit orders.[13] Conversely, when there is more depth on the sell side, sellers respond by increasing the aggressiveness of their orders: placing sell orders at lower prices and/or using market orders to execute their trades immediately rather than being patient and using limit orders. Therefore, layering the bid side is likely to cause other traders to place buy orders with a more assertive approach at higher prices, and vice versa, layering the ask side is likely to cause other participants in the market to submit more aggressive sell orders at lower prices.

4.3. Characteristics of layering and spoofing in prosecution cases

This section analyses a comprehensive set of prosecution cases from around the world. We manually collect spoofing and layering cases prosecuted by market regulators from January 2010 to January 2020. We identify the instances via systematic searches of regulatory/enforcement releases, court filings of prosecutors, news databases (Factiva), and legal databases (LexisNexis). Once a case is identified, we extract as much information as possible from publicly available documents, then supplement that with documents obtained through court repositories, including using the court document service, PACER (Public Access to Court Electronic Records).

[13] For examples see Parlour (1998), Biais, Hillion, and Spatt (1995), Cao, Hansch, and Wang (2008), Griffiths et al. (2000), and Ranaldo (2004).

In cases where information is still missing, we file requests for information with the relevant authority or authorities via Freedom of Information (FOI) requests. Most cases are prosecuted in the US (e.g., by the Financial Industry Regulatory Authority (FINRA), Securities and Exchange Commission (SEC), Commodity Futures Trading Commission (CFTC)) and the UK (e.g., by the Financial Conduct Authority (FCA)).

4.3.1. Overview of the prosecution cases

Table 4.1 below provides a summary of the prosecuted layering and spoofing cases, the manipulation period, the market, and the type of enforcement action. It shows that the cases are prosecuted under both civil and criminal processes. They involve a range of markets, including both equities and futures.

Table 4.1
Summary of Prosecution Cases

This table provides a summary of the spoofing prosecution cases, including the market regulator, accused/prosecuted party, year of manipulation, market, market, and exchange. The cases are from market regulators in the US such as the Financial Industry Regulatory Authority (FINRA), Securities and Exchange Commission (SEC), Commodity Futures Trading Commission (CFTC), in the UK such as the Financial Conduct Authority (FCA), in Canada such as the Ontario Securities Commission (OSC), and Japan such as the Securities and Exchange Surveillance Commission (SESC). The cases are from January 2010 to January 2020. Manipulated stock exchanges and futures exchanges include the New York Stock Exchange (NYSE), Nasdaq Stock Exchange (NASDAQ), New York Stock Exchange Arca (NYSE Arca), London Stock Exchange (LSE), Tokyo Stock Exchange (TYO), Intercontinental Exchange (ICE), Index and Options Market (IOM), Commodity Exchange (COMEX), New York Mercantile Exchange (NYMEX), Chicago Mercantile Exchange (CME), Chicago Board Options Exchange (CBOE), and Chicago Board of Trade (CBOT).

Number	Regulator	Case period	Accused/prosecuted party	Period of manipulation	Market type	Exchange	Action
1	FINRA	2010–2010	Trillium Brokerage Services	1/11/2006–31/1/2007	Equities market	NASDAQ, NYSE	Civil
2	FSA	2011–2014	Swift Trade & Peter Beck	1/1/2007–4/1/2008	Equities market	LSE	Civil
3	SEC	2012–2012	Hold Brothers	1/2009–9/2010	Equities market	NASDAQ, NYSE	Civil
4	SEC	2012–2012	Biremis Corporation	1/2007–6/2010	Equities market	NYSE	Civil
5	SESC	2014–2014	Select Vantage	12/4/2012–24/4/2012	Equities market	TYO	Civil
6	SEC	2014–2014	Visionary Trading LLC	5/2008–11/2011	Equities market	NASDAQ	Civil
7	SEC, FCA	2014–2017	Michael Coscia	8/2011–7/2012	Equities market	ICE	Criminal
8	FCA	2015–2015	Da Vinci Invest	2010–2011	Equities market	CME, CBOT	Civil
9	OSC	2015–2015	Oasis World Trading	9/2013–10/2014	Equities market	Many Canadian equity markets	Civil
10	CFTC	2015–2016	Igor B. Oystacher	12/2011–1/2014	Futures market	COMEX, NYMEX, CME, CBOE	Civil
11	CFTC	2015–2016	Navinder Singh Sarao	6/2009–7/2015	Futures market	CME	Criminal
12	CFTC	2015–2017	Heet Khara and Nasim Salim	2/2015–4/2015	Futures market	COMEX	Civil

Table 4.1 (continued)
Summary of Prosecution Cases

This table provides a summary of the spoofing prosecution cases, including the market regulator, accused/prosecuted party, year of manipulation, market, and exchange. The cases are from market regulators in the US such as the Financial Industry Regulatory Authority (FINRA), Securities and Exchange Commission (SEC), Commodity Futures Trading Commission (CFTC), in the UK such as the Financial Conduct Authority (FCA), in Canada such as the Ontario Securities Commission (OSC), and Japan such as the Securities and Exchange Surveillance Commission (SESC). The cases are from January 2010 to January 2020. Manipulated stock exchanges and futures exchanges include the New York Stock Exchange (NYSE), Nasdaq Stock Exchange (NASDAQ), New York Stock Exchange Arca (NYSE Arca), London Stock Exchange (LSE), Tokyo Stock Exchange (TYO), Intercontinental Exchange (ICE), Index and Options Market (IOM), Commodity Exchange (COMEX), New York Mercantile Exchange (NYMEX), Chicago Mercantile Exchange (CME), Chicago Board Options Exchange (CBOE), and Chicago Board of Trade (CBOT).

Number	Regulator	Case period	Accused/prosecuted party	Period of manipulation	Market type	Exchange	Action
13	SEC	2015−2022	Aleksandr Milrud	1/2013−1/2015	Equities market	NYSE	Criminal
14	SEC	2016−2021	Joseph Taub and Elazar Shmalo	1/2014−12/2015	Equities market	NASDAQ, NYSE	Criminal
15	SEC	2017−2022	Lek Securities	12/2010−9/2016	Equities market	NASDAQ, NYSE	Civil
16	CFTC	2017−2017	Arab Global Commodities	3/2016−8/2016	Futures market	COMEX	Civil
17	CFTC	2017−2017	David Liew	2009−2/2012	Futures market	COMEX	Criminal
18	CFTC	2017−2017	Simon Posen	12/2011−3/2015	Futures market	NYMEX, COMEX	Civil
19	CFTC	2017−2017	Bank of Tokyo-Mitsubishi UFJ	2010−2011	Futures market	CFE, CME	Civil
20	CFTC	2018−2018	HSBC Securities (USA) Inc	7/2011−8/2014	Futures market	COMEX	Civil
21	CFTC	2018−2018	Mizuho Bank, Ltd	5/2016−5/2017	Futures market	CFE, CME	Civil
22	CFTC	2018−2018	Michael D. Franko	5/2013−7/2014	Futures market	COMEX, NYMEX	Civil
23	CFTC	2018−2019	Krishna Mohan	9/2012−03/2014	Futures market	CME, CFE	Criminal
24	CFTC	2018−2019	Jitesh Thakkar	30/1/2013−30/10/2013	Futures market	CME	Acquit
25	CFTC	2018−2023	John Edmonds	2009−2015	Futures market	NYMEX, COMEX	Criminal

4.3.2. Key features of the prosecution cases

1) Michael Coscia

The Michael Coscia spoofing case is important as it was one of the first cases in which an individual was successfully prosecuted for spoofing in accordance with the Dodd-Frank Wall Street Reform Act of 2010, thereby setting a legal precedent. His convictions sent a strong message that regulators will pursue criminal charges against those who engage in spoofing.

Michael Coscia was an experienced trader with more than 20 years of trading experience at the time. He was the principal of Panther Energy Trading, a firm specializing in high-frequency trading. His spoofing strategy was facilitated by algorithms and targeted other high-frequency and algorithmic traders by triggering predictable behaviors of other algorithms.

According to court documents, Coscia engaged in spoofing on at least 36 different occasions between August 2011 and July 2012. He was eventually sentenced to three years in prison. In addition to his prison sentence, Coscia was also ordered to pay a $2.8 million fine and to forfeit $3.9 million in ill-gotten gains.

2) Navinder Singh Sarao

Navinder Singh Sarao is a notable case because he was accused of contributing (through his spoofing strategy) to the 2010 "flash crash" that occurred in the United States stock market. The flash crash, which took place on May 6th, 2010, was characterized by a sudden and sharp drop in stock prices that wiped $1 trillion off the market for a few minutes before prices began to recover. Sarao was indicted on criminal charges in the United States for his alleged role in the flash crash, including fraud and spoofing.

Sarao was a self-taught trader. He traded from his parents' house in a suburb of England. He devised his own trading program after realizing that every time an order was placed, high-frequency traders would make trades milliseconds before these orders were executed and be the first to make money from market changes.

Sarao was arrested in the UK in 2015 and extradited to the United States in 2016 to face trial. In 2017, he admitted guilt for a single charge of spoofing and a separate charge of wire fraud. Consequently, he received a prison sentence of four years and was also mandated to pay a fine amounting to $12.9 million.

3) Swift Trade

In 2013, Swift Trade, a Canadian trading firm, was accused of spoofing. In this case, Swift Trade was accused of placing more than 100,000 spoofing orders over a period of three years, resulting in more than $5 million in profits. Swift Trade operated a network with about 150 trading locations worldwide with hundreds of traders. What is unusual about this case is that the spoofing was not conducted by a single individual or algorithm, but rather, through the network of hundreds of traders in 30 countries. Individual price movements were small but profits were magnified by repeating the pattern many times a day, in many shares across market sectors, and many trading locations. The court imposed a fine of £8 million on Swift Trade in May 2011.

4) Da Vinci Invest

In 2015, Da Vinci Invest, an English company, is accused of engaging in spoofing with three traders from Hungary to manipulate stocks on LSE in 2010. This instance signifies the first occasion in which the Financial Services Authority (FSA, the predecessor of the FCA) employs legal actions to secure permanent restraining orders and monetary sanctions against an individual involved in market manipulation.

It is noteworthy that Da Vinci Invest executes spoofing cycles at high speed. The whole process from submitting orders, canceling orders, and switching sides often takes place within a minute. Da Vinci's algorithm directs the spoofing orders that made “saw-tooth” patterns of trading many times during the day.

5) Joseph Taub and Elazar Shmalo

The Joseph Taub and Elazar Shmalo case involves two individuals who were charged with engaging in layering/spoofing the futures market. What is interesting about this case is that the layering used different trading accounts to submit layered orders and genuine orders to obscure what is otherwise a highly suspicious pattern of trading. Many of the accounts were opened in the names of individuals who neither controlled nor traded the securities held in the accounts.

At least one account was primarily used to place multiple small orders to create upward or downward pressure on the stock price (referred to as a "helper" account). At least one other account (referred to as a "winner" account) was primarily used to buy and sell more substantial quantities of stocks at prices affected by the manipulative orders placed by the helper account. The market manipulation scheme used 36 accounts with at least nine independent brokerage firms. There were no algorithms involved.

6) Alexsandr Milrud

While most strategies in other cases involved computers, Canadian resident Milrud orchestrated spoofing manually. Milrud supposedly engaged in spoofing by having many people, not programs and machines. He led and managed several groups of traders, based primarily in South Korea and China, involved in layering in US securities markets.

To facilitate the spoofing, he worked with a gaming company to develop algorithms with hotkeys for quickly submitting, amending, or cancelling orders. Milrud was eventually convicted of the charges against him and was sentenced to prison for his role in the spoofing scheme.

7) Igor B.Oystacher and 3 Red Trading

Oystacher was accused of engaging in spoofing on numerous occasions between 2010 and 2013 and was eventually charged with six counts of commodities fraud and six counts of spoofing. The most notable characteristic of the Oystacher layering strategies is that he took advantage of a commonly used function of the exchange platform (function called "avoid orders that cross") to facilitate his scheme instead of using a purposefully designed computer program. The traders took advantage of this feature by submitting orders that would automatically cancel the submitted orders on the opposing side, effectively avoiding any potential matches with the new orders. This process occurred swiftly and nearly simultaneously.

Oystacher also used "iceberg" or hidden-quantity orders offered by specifically designated contract markets. The genuine orders were placed as a partially visible iceberg order to maximize the likelihood of execution.

Oystacher was sentenced to three years in jail and a $1.1 million fine. Oystacher also settled $5 million in civil fines to the US CFTC.

8) Jim Zhao

In the Jiongsheng Zhao spoofing case, Zhao was accused of engaging in a scheme to artificially influence prices of futures on the COMEX (Commodity Exchange) and NYMEX (New York Mercantile Exchange). Zhao used both computer programs and manual trading. Zhao placed orders from his home or offices in Sydney, Australia. Interestingly, the daytime trading session in these US markets corresponds with night-time in Australia; therefore, the Zhao focused on carrying out his scheme almost exclusively during the US overnight sessions. Zhao claimed he placed the spoof orders to liquidate the substantial positions prior to mid-day and conclude the trading day.

Zhao was indicted on federal charges of wire fraud, commodities fraud, and spoofing in 2017. He pleaded guilty to the charges in 2018 and was given a sentence of 30 months in jail and ordered to pay a fine of $1.5 million.

9) James Vorley and Cedric Chanu

James Vorley and Cedric Chanu were convicted of spoofing in a US federal court in Chicago in 2018. Vorley and Chanu were employed as traders on the metals desk at Deutsche Bank. Vorley and Chanu joined a scheme with a general pattern of placing numerous spoofing orders for precious metals futures contracts. The manipulators manually placed layering orders. They used functions allowed by trading venues, such as “iceberg” orders, to facilitate their scheme.

10) J.P. Morgan

The J.P. Morgan case against the metals desk team including Michael Nowak, Gregg Smith, and Jeffrey Ruffo is the recent of DOJ’s legal actions against spoofing in our sample. The J.P. Morgan spoofing case involved the use of fraudulent trading techniques by ten individuals within the bank to manipulate markets for financial instruments. The alleged spoofing was so successful that it led to $300 million in loss to other traders in the market from March 2008 to August 2016.

The traders engaged in illicit trading practices and communicated about their activities using electronic means such as chat messages and emails. Most instances of their spoofing lasted less than one minute. The strategies were executed manually and deliberately designed to mislead other market participants who relied on automated trading systems or computer algorithms for their trades. The company settled $920 million in fines.

4.3.3. Anatomy of a spoofing case

This subsection zooms in and analyzes an example of a layering and spoofing case in detail. The example is from the Da Vinci case. The manipulated stock is Admiral Group. The manipulation period was from 14:35 to 15:08 (exchange time) on March 18, 2011.

First, we describe the steps followed by the manipulator in this specific case through one full cycle of spoofing (including buy and sell activity). We define one cycle of spoofing as one buy phase (layer the ask (offer) side to buy at a lower price) and one sell phase (layer the bid side to sell at a higher price).

The buying phase is shown in Figure 4.1. At 14:41:31, the manipulator submits a large (spoofing) sell order of size 14,520 at the best ask (sell) price of 1574. At that time, the total volume at the best ask was only 79. Three seconds later, the manipulator submits another large order (17,500) to sell the stock at 1574. While the two large sell orders remain at the 1574 price level, smaller sell orders are submitted by other market participants at 1573 to get execution priority over the spoofing orders. Therefore, the spoofing orders get pushed back in priority to ask level 2. At 14:41:47, other sell orders are submitted at prices of 1573 and 1572. Therefore, on the sell side at 14:41:47, there are layers of spoofing orders at several price steps behind the best prices.

Between 14:41:31 and 14:41:47, while the spoofing sell orders are placed, the seller-initiated trades and additional sell orders from other market participants that are likely to have been influenced by the spoofing orders drive down the mid-price. Figure 4.2 shows the decrease in mid-quote while the spoofing sell orders are in the market.

Having depressed the market price, at 14:41:52, the manipulator's genuine (intended to execute) buy order executes at a price of 1572. This price is lower than the price that the manipulator would have paid had they bought at 14:41:31 before entering the spoofing sell

orders. Soon after, the manipulator cancels one of the spoofing sell orders of size 12,500 at ask level 2.

The imbalance in the limit order book (appearance of selling pressure) created by the spoofing sell orders is easily observable in Figure 4.3.

Figure 4.1
Example of order placement and cancellation during the buying phase of a layering case

This figure demonstrates the buy phase of a layering case drawn from the prosecution case *FCA v. Da Vinci Invest.* The manipulated security is Admiral Group PLC on the LSE on March 18, 2011. The figure shows the change in the level 1 ask price, level 1 bid price, and the mid-price as a result of order entry, cancellation, and execution. The figure provides time on the horizontal axis and price on the vertical axis. "+" indicates entry of a spoofing order, "×" indicates cancellation of a spoofing order, and "●" indicates a trade.

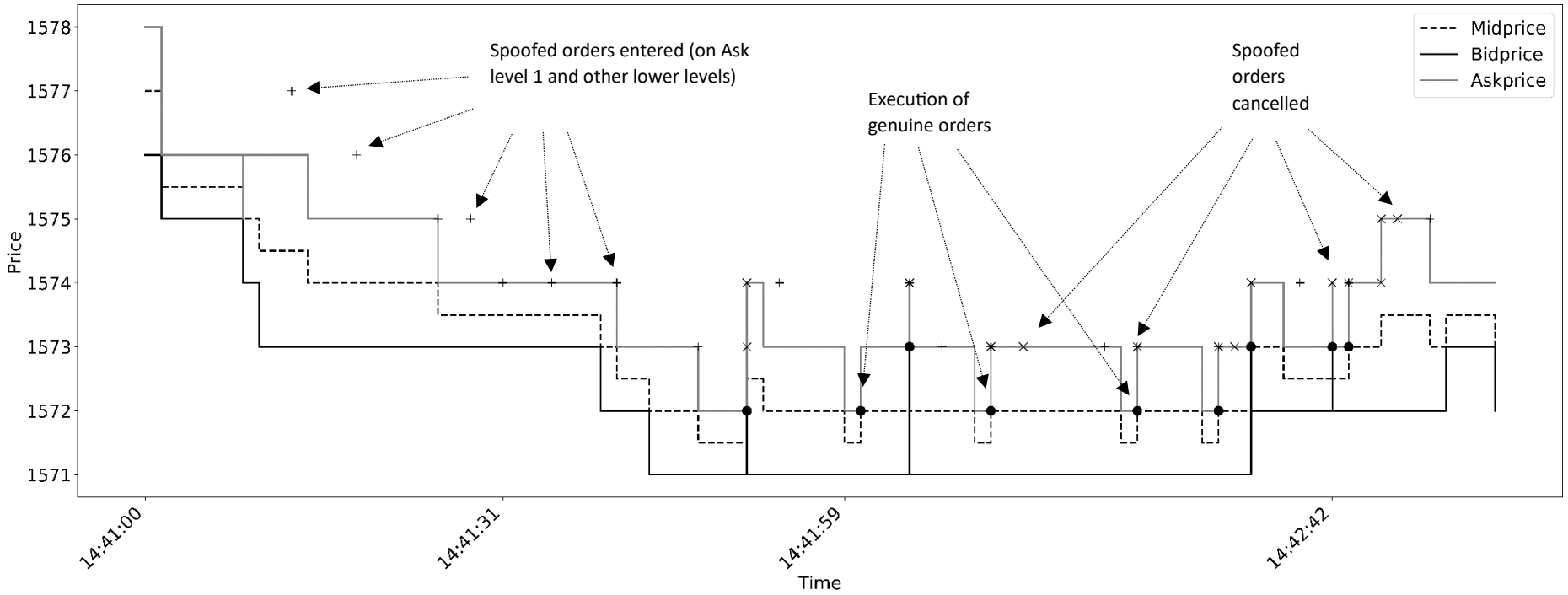


Figure 4.2
Level 1 depth and midquote prices during the buying phase of a layering case

This figure illustrates the level 1 bid and ask depth and midquote prices during the buying phase (layering orders are on the sell (ask) side) of a layering case. The example is from the prosecution case *FCA v. Da Vinci Invest.* The manipulated security is Admiral Group PLC on the LSE on March 18, 2011. For illustrative purposes, we multiply bid volume by −1.

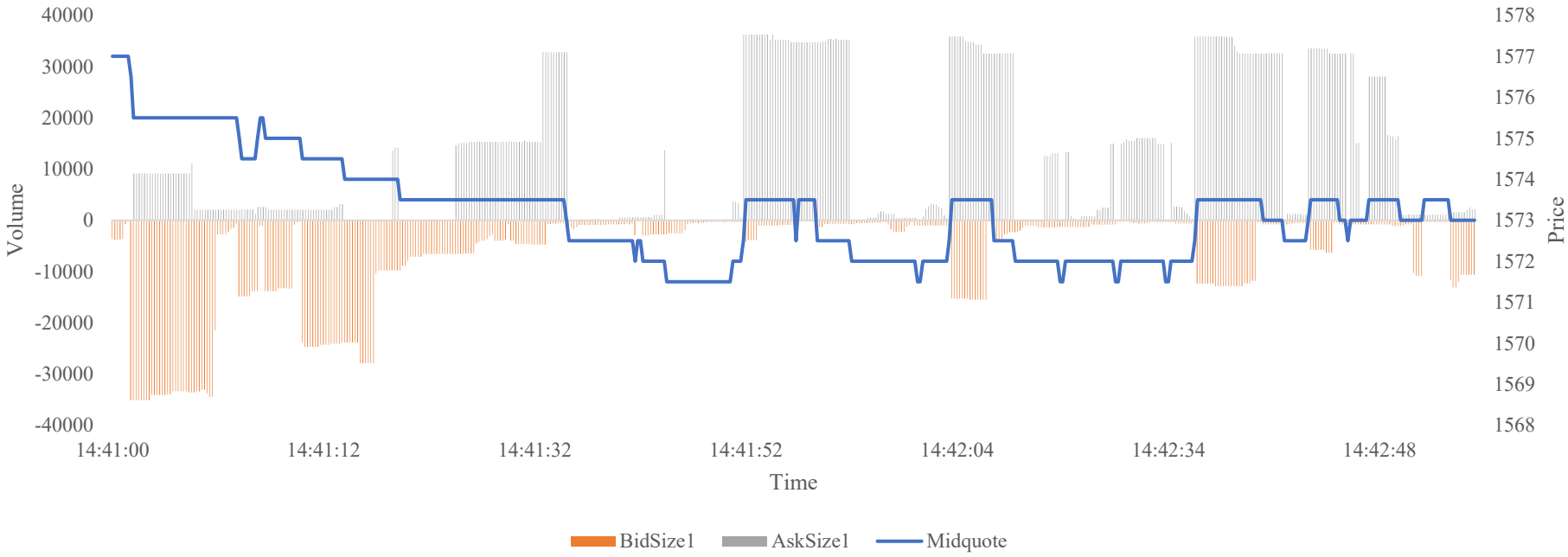
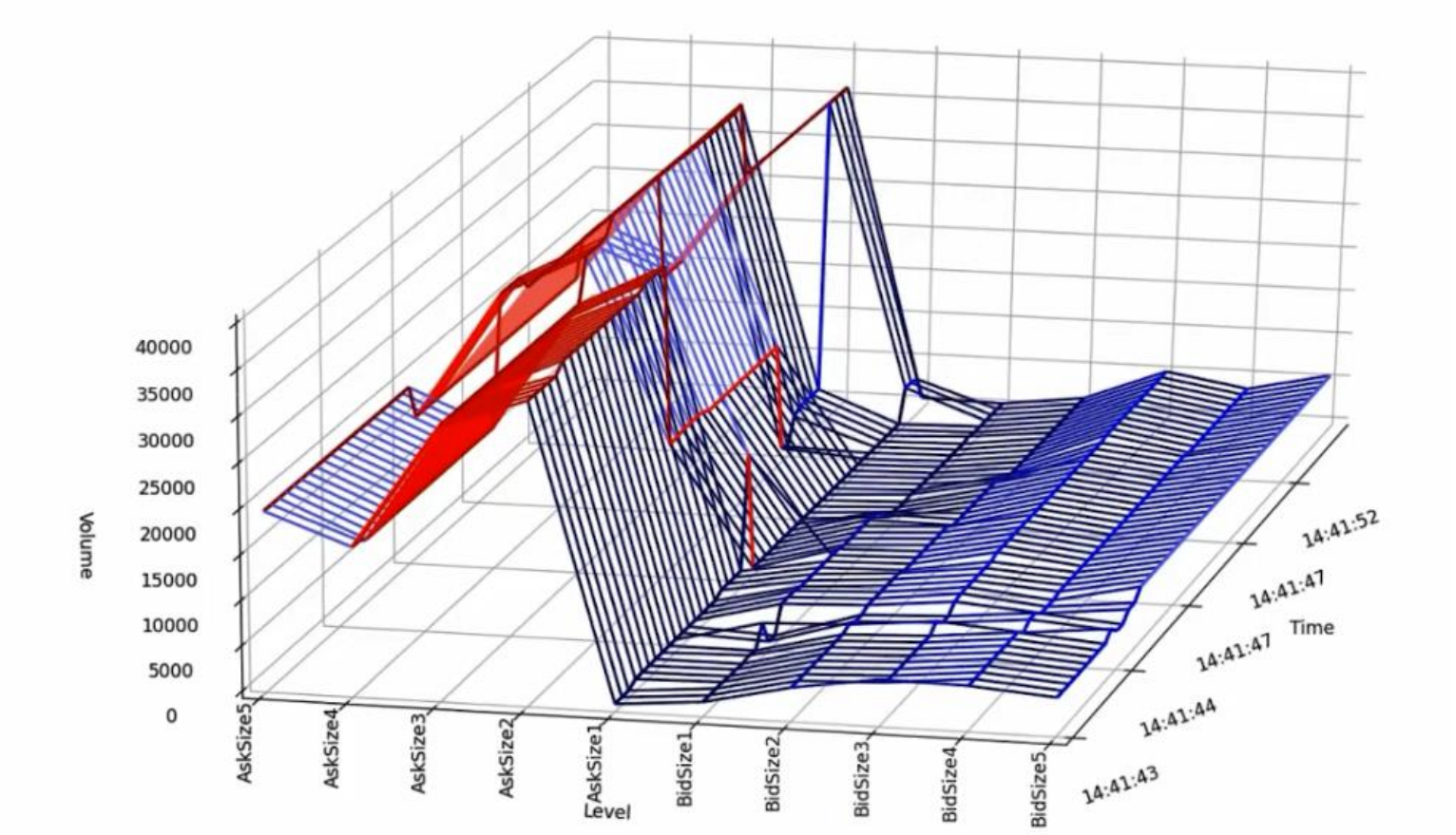


Figure 4.3
Order book imbalance during the buying phase of a layering case

This figure illustrates the first 5 levels of the order book on the bid and ask sides showing the imbalance created by spoofing orders on the sell (ask) side of the market. The example is from the prosecution case *FCA v. Da Vinci Invest.* The manipulated security is Admiral Group PLC on the LSE on March 18, 2011.



For the next few seconds, the manipulator's buy orders continue to execute at low prices against incoming sell orders. At 14:42:01 and 14:42:04, the manipulator executes another two buy orders on the bid side at a price of 1573.

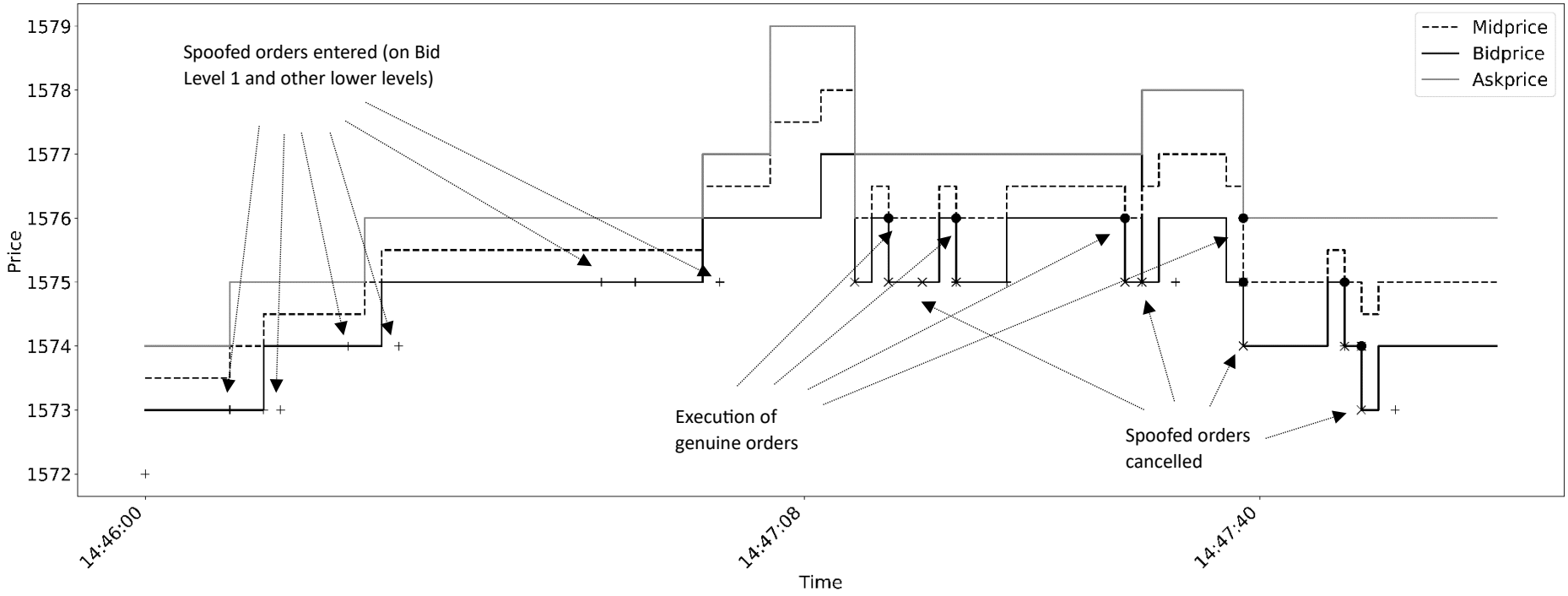
At 14:42:08, the manipulator submits another large spoofing order of size 12,500 at the price of 1573 to support the rest of the spoofing orders on level 2 of the ask side. As soon as this order is submitted, other sellers jump ahead of the spoofing order, bringing the ask price down to 1572. The manipulator takes advantage of the induced selling to buy at 1572. Five seconds later, at 14:42:15, the manipulator cancels the spoofing orders of size 12,500 at the price of 1573 and resubmits them at best ask price, keeps them in the order book for only around 10 seconds and then cancels all non bona-fide orders on the ask side at 14:42:48.

That concludes the "buying phase" of this instance of spoofing.

The manipulator then waits about 5 minutes and then repeats a similar pattern of trading, but on the bid side in what is the "selling phase" – using non-bona-fide orders on the buy side to push the market up and execute ask orders at relatively high prices. That is how the manipulator offloads the inventory of stock accumulated during the buying phase. Figure 4.4. illustrates the selling phase.

Figure 4.4
Example of order placement and cancellation during the selling phase of a layering case

This figure demonstrates the sell phase of a layering case drawn from the prosecution case *FCA v. Da Vinci Invest.* The manipulated security is Admiral Group PLC on the LSE on March 18, 2011. The figure shows the change in the level 1 ask price, level 1 bid price, and the mid-price as a result of order entry, cancellation, and execution. The figure provides time on the horizontal axis and price on the vertical axis. "+" indicates entry of a spoofing order, "×" indicates cancellation of a spoofing order, and "●" indicates a trade.



The manipulator follows the same strategy as before to layer the bid side, starting with a large spoofing buy order of size 12,500 at the best bid price of 1573 at 14:46:07. As soon as this order is entered, a small buy order from another market participant is submitted at a price of 1574 to "leapfrog" the spoofing order and compete with it for execution priority in the order book. Ten seconds later, the manipulator submits another large order of size 14,200 at 1574. As soon as the large spoofing order is entered, buy orders from other market participants are submitted at higher prices and push the spoofing orders back in limit order book priority to price steps behind the best quotes. At this point, both the mid-price and the best ask price have been increased by the spoofing. Figure 4.5 illustrates the change in mid-price and depth at the best quotes.

The manipulator then keeps submitting larger orders at the best buy prices to support the upward pressure on prices. The spoofing buy orders keep attracting other market participants to buy until the best bid price at 14:47:13 is driven up to 1577. The large orders are pushed back to level 2 again due to new buy orders from other participants. The excess buying interest is illustrated in Figure 4.6.

Having inflated the price, the manipulator executes sell orders on the ask side from 14:47:15 to 14:47:39 at a price of 1576, which is higher than what they would have received had they sold at 14:46:07 before placing the spoofing orders. They cancel the spoofing orders between 14:47:46 and 14:48:38, completing the selling phase of the spoofing strategy.

Figure 4.5
Level 1 depth and midquote prices during the selling phase of a layering case

This figure illustrates the level 1 bid and ask depth and midquote prices during the selling phase (layering orders are on the buy (bid) side) of a layering case. The example is from the prosecution case *FCA v. Da Vinci Invest.* The manipulated security is Admiral Group PLC on the LSE on March 18, 2011. For illustrative purposes, we multiply bid volume by −1.

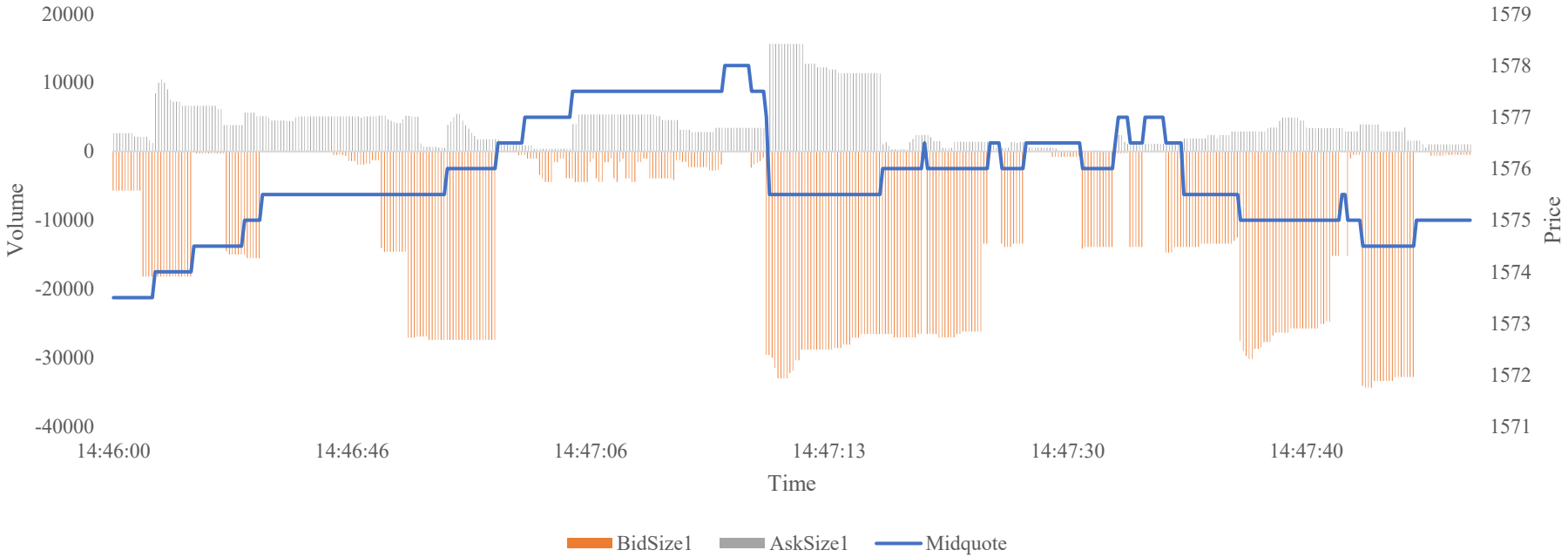
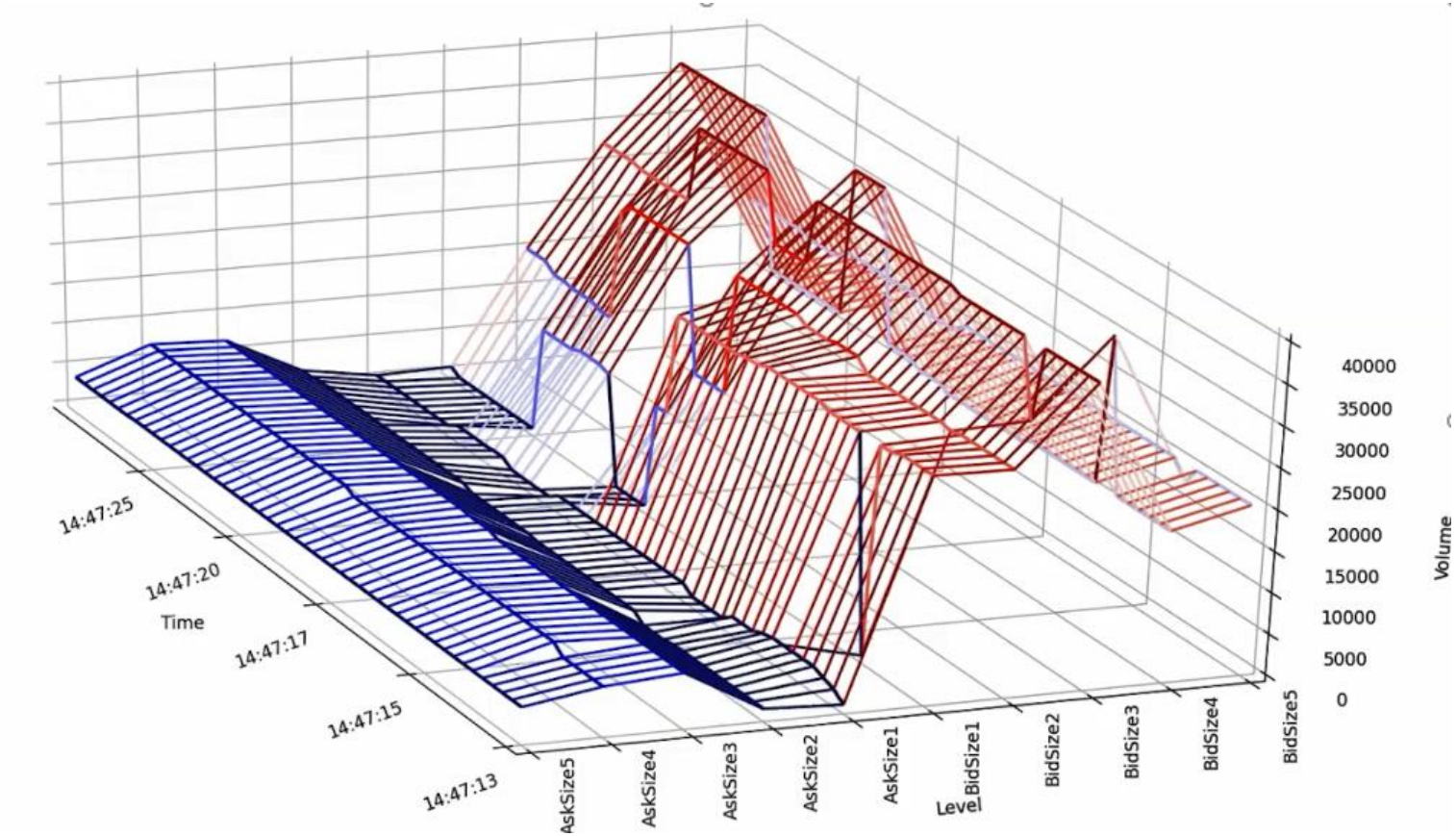


Figure 4.6
Order book imbalance during the selling phase of a layering case

This figure illustrates the first 5 levels of the order book on the bid and ask sides showing the imbalance created by spoofing orders on the buy (bid) side of the market. The example is from the prosecution case *FCA v. Da Vinci Invest.* The manipulated security is Admiral Group PLC on the LSE on March 18, 2011.



4.3.4. Layering and spoofing characteristics

The anatomy of the spoofing case in the previous subsection suggests a pattern of trading that has some distinguishing characteristics. Based on that example as well as a broader review of the other cases in our sample, we develop a set of characteristics that are commonly seen during spoofing – effectively, the empirical "footprint" that is left behind in market data by spoofing strategies. Having defined the set of distinguishing characteristics, we then systematically review each prosecution case in our sample against these characteristics to identify how consistently they present themselves in the cross-section of spoofing cases.

A successful layering and spoofing scheme typically involves placing large orders on a specific side ("spoofing side") of the market in a manner that creates a highly unbalanced order book that conveys the impression of a lot of trading interest on the spoofing side. Orders on the spoofing side are mostly cancelled after obtaining a better execution price for the manipulator on the other side. The genuine interest of the manipulator is in trading on this side ("genuine side").[14] The genuine orders are typically smaller (individually or in aggregate) than the non-bona-fide orders to enable the book imbalance that pushes the market to execute against the manipulator's genuine orders.

In some cases, orders on the genuine side are placed as hidden orders to further amplify the appearance of an imbalance in the visible orders. The manipulator attempts to avoid execution on the spoofing side, although sometimes inadvertently a spoofing order placed by the manipulator will execute. Typically, genuine-side orders are more frequently executed than spoofing side orders, consistent with the intentions of the manipulator.

Cyclical layering is a form of layering in which the layering process is repeated many times in a day by switching from layering the bid side to layering the ask side. Not all spoofing cases have this cyclical property, but when it is present it is a further indication of spoofing and can be used to improve detection accuracy. A typical layering cycle includes: (i) placing small order on the genuine side of the market, (ii) submitting a significant volume of orders on the

[14] While we refer to the orders that the manipulator intends to execute as the genuine orders and the genuine side, that is only intended to convey that the manipulator genuinely wants those orders to execute. From a legal perspective, however, given the trading is conducted for an impermissible purpose, all the manipulator's orders and trades may be regarded as non-genuine.

other side of the market (spoofing side) to move the market towards the genuine order,[15] (iii) parallel to or after the execution of a genuine order, the layering orders are cancelled,[16] and (iv) these steps are then repeated on the opposing side. This series of order placements and cancellations creates a distinctive intraday cyclical pattern. The layering cycles are typically conducted in a short time, within a matter of seconds or minutes. However, the cycles can last for longer. When conducted at high-frequencies (e,g., sub-second layering cycles), they are typically implemented with the use of pre-programmed algorithms, making rapid submission and cancellation possible. When cyclical layering is conducted intraday, there can be hundreds or thousands of instances (cycles) in a day.

Below are the key identifying characteristics of spoofing. No single characteristic is definitive proof of spoofing, but a collection of characteristics seen together can effectively distinguish between spoofing and legitimate trading as per the empirical tests in the next section.

Characteristic 1. High Quoting Activity

The limit orders of a layering strategy are likely to represent a substantial proportion of the depth (resting orders) on one bid or ask side (buying or selling) as the layering strategy is attempting to mislead the market.[17] Using a large volume increases the probability that the layering orders have a substantial effect on the appearance of supply or demand and therefore the desired effect of influencing the price or trading decisions of other market participants.

Characteristic 2. Unbalanced Quoting

The resting (unexecuted) limit orders of a layering strategy at a specific moment are likely to be highly unbalanced (more buy volume than sell volume or vice versa). The

[15] The order of (i) and (ii) can be changed in some cases.

[16] In most cases, spoofing orders are cancelled right after execution of the genuine order(s). However, there are cases where spoofing orders are cancelled before execution.

[17] The resting orders could be large or small, what matters is their aggregate volume being a substantial proportion of the depth. For example, in the alleged layering conducted by Joseph Taub and Elazar Shmalo (see United States of America v Joseph Taub and Elazar Shmalo, Mag. No. 16-8190) the layering orders are small in induvial volume but large in number, while in other cases the layering is conducted with fewer but larger orders (e.g., US CFTC v. Nav Sarao Futures Limited Plc and Navinder Singh Sarao, 2015).

imbalance is used to create a misleading perception regarding the presence of buying and selling interest in order to influence other participants and market prices.

Characteristic 3. Abnormal Cancellation

Layering strategies imply abnormally large cancellation rates (a high ratio of cancellations to trades). This is because the layering orders are not intended to execute so they will typically end in a cancellation. Layering orders may also need to be cancelled and resubmitted as market conditions change (such as when the market moves towards the layering orders) to maintain a low execution probability.

Characteristic 4. Low Execution Probability

Layering orders are placed in the market to deliberately have a low execution probability, either at price steps behind the best quotes or at the back of a long queue of orders at a price level.[18] To the extent that low execution probability is achieved by placing orders away from the best quotes, we see less aggressiveness of the spoofing orders (orders further from the best quotes) on the same side as order imbalance. Depending on the market or trading platform, there are different tactics to lower the probability of execution.

Characteristic 5. Inventory Reversals

Manipulators using intraday layering strategies are likely to accumulate long or short positions intraday, but typically they do not hold positions overnight. In such cases, buying

[18] Low execution probability can be obtained by setting low prices on buy orders (prices falling below the lowest bid price) or high prices on sell orders (prices surpassing the highest ask side). Alternatively, in markets with a lot of depth at the best quotes, layering orders could be submitted at the best quotes relying on time priority to give the orders low execution priority (they would be at the back of a long queue). In doing so, the orders would need to be cancelled and resubmitted or amended in such a way that they lose time priority as they approach the front of the queue.

volume during the day is equal to selling volume, demonstrating inventory reversal characteristics.

Characteristic 6. Trades Oppose Quotes

Manipulators engaging in layering will often have trades carry out on the other side to their resting order imbalance (e.g., if the trader has more buy orders than sell orders, they are likely to execute sells and vice versa). Put differently, manipulators using layering often trade in a direction that is opposite to what their order imbalance would signal as their intended direction of trade because their order imbalance is not a true reflection of their intentions as many or all of those orders are not intended to execute.

Characteristic 7. Cancels Oppose Trades

Manipulators engaging in layering will often cancel orders on specific side after they execute an order on the opposite market side (e.g., after they buy, they cancel sell orders, and vice versa). This occurs because layering orders are not intended to execute and once they have served their purpose of pushing the market towards the price of any bona fide orders resulting in executions, the layering orders can be cancelled.

Characteristic 8. Dark Opposes Lit

When manipulators use "dark" or hidden orders as part of their strategy (not all do so), their "lit" or displayed order imbalance (imbalance between the volume of displayed buy and sell orders) is likely to be opposite in direction to their dark/hidden order imbalance. For example, if a manipulator is trying to sell with dark/hidden orders and they use layering to assist in the execution of those sells, they are likely to have a buy imbalance (more buy volume than sell volume) to create the false impression of buying interest. The opposite imbalances in the lit and dark occur because dark/hidden orders are unlikely to be used to fabricate an inaccurate perception of the supply or demand for a given stock because these orders are not displayed to the market. Dark/hidden orders are therefore more likely to reflect the manipulator's true trading intention at a given point in time.

Characteristic 9. Quoting Opposes Inventory Reversion

Manipulators tend to layer the order book's bid side when they have a long position and vice versa – layer the ask side when they have a short position. Manipulator's order imbalance reflects inventory position. Consequently, the manipulator's order imbalance in the limit order book will tend to mirror their inventory position: a buy-side imbalance (more resting buy orders than sell orders) when they have bought the stock (long position) and a sell-side imbalance (more resting sell orders than buy orders) when they have sold the stock (short position).

Characteristic 10. Quoting Opposes Trading Intention

Manipulators tend to place orders in the limit order book opposite to their trading intention: if they want to sell, they will typically have more buy orders and vice versa. The empirical relevance of this characteristic is that if it is possible to infer a trader's trading intention (to buy or to sell, e.g., if they tend to close the day's trading with zero inventory and happen to be long near the end of the day, then it is likely the manipulator intends to sell), then orders submitted on the opposite side of the trading intention are likely spooking orders.

Characteristic 11. Cyclical Pattern in Depth

During cyclical layering events, large, layered orders create an imbalance of depth o a particular market side, which then switches to the other side, and back again as layering cycles are conducted. This pattern is repeated until the end of the layering instance.

Characteristic 12. Cyclical Pattern in Cancellations

Similarly, cyclical layering results in cycles of high cancellation rates on the bid side, then the sell side, then the buy side again, and so on.

Characteristic 13. Cyclical Pattern in Inventory

Cyclical manipulators layer bid side when they have a long position and layer ask side when they have a short position. Repetition of these cycles creates a cyclical pattern in the net position of the manipulator's inventory during the day.

Characteristic 14. Cyclical Pattern in the Mid-price

Cyclical layering results in cycles of rising then falling mid-prices as the manipulator switches between spoofing both sides of the market.

Table 4.2 illustrates which characteristics are present in each of the prosecution cases in our sample. In explaining the results, it is essential to recognize that only some of the prosecuted cases involve the cyclical form of layering and therefore characteristics 11−14 are not expected to be present in all cases. The results suggest a striking consistency with most of the relevant characteristics being present in most cases.

Given that many of these characteristics are inconsistent with reputable approaches to conducting trades such as market making, arbitrage, and execution algorithms, the consistent presence of these characteristics suggests they should have the ability to empirically detect layering and spoofing and distinguish it from legitimate trading. We test this notion in the next section.

Table 4.2
Characteristics displayed in prosecuted or alleged layering and spoofing cases

This table illustrates which of the characteristics of layering are present in prosecuted or alleged layering cases. Given the limited details of some of the cases and reliance upon documents that are available in the public domain, courts, and via FOI requests, it is not possible to assess the presence of all characteristics in all cases. "N/A" is used to indicate instances where it has not been possible to determine whether a particular characteristic is present or not. "Y" and "N" indicate the characteristic is or is not present in the case, respectively. The characteristics correspond to those described earlier.

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Type	General	General	General	General	General	General	General	General	General	General	Cyclical	Cyclical	Cyclical	Cyclical
Characteristics	Unbalanced Quoting	High Quoting Activity	Abnormal Cancellation	Low Execution Probability	Inventory Reversal	Trades Oppose Quotes	Cancels Oppose Trades	Dark Opposes Lit	Quoting Opposes Inventory Reversion	Quoting Opposes Trading Intention	Cyclical Pattern in Depth	Cyclical Pattern in Cancellations	Cyclical Pattern in Inventory	Cyclical Pattern in Mid-price
Da Vinci Invest	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Michael Coscia	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Peter Beck and Swift Trade	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Biremis Corporation	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Joseph Taub and Elazar Shmalo	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Aleksandr Milrud	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	N/A	N/A	N/A	N/A
Visionary Trading LLC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Lek Secs	Y	Y	Y	N/A	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Hold Brothers	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Trillium Brokerage Services	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Zhen (Steven) Pang	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Igor B. Oystacher	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N/A	N/A	N/A	N/A
Navinder Singh Sarao	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y
Jiong Sheng Zhao	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	N/A	N/A	N/A	N/A

Table 4.2 (continued)
Characteristics displayed in prosecuted or alleged layering and spoofing cases

This table illustrates which of the characteristics of layering are present in prosecuted or alleged layering cases. Given the limited details of some of the cases and reliance upon documents that are available in the public domain, courts, and via FOI requests, it is not possible to assess the presence of all characteristics in all cases. "N/A" is used to indicate instances where it has not been possible to determine whether a particular characteristic is present or not. "Y" and "N" indicate the characteristic is or is not present in the case, respectively. The characteristics correspond to those described earlier.

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Type	General	General	General	General	General	General	General	General	General	General	Cyclical	Cyclical	Cyclical	Cyclical
Characteristics	Unbalanced Quoting	High Quoting Activity	Abnormal Cancellation	Low Execution Probability	Inventory Reversal	Trades Oppose Quotes	Cancels Oppose Trades	Dark Opposes Lit	Quoting Opposes Inventory Reversion	Quoting Opposes Trading Intention	Cyclical Pattern in Depth	Cyclical Pattern in Cancellations	Cyclical Pattern in Inventory	Cyclical Pattern in Mid-price
James Vorley and Cedric Chanu	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	N/A	N/A	N/A	N/A
Krishna Mohan	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Jitesh Thakkar	Y	Y	Y	Y	N/A	Y	Y	N/A	N/A	Y	N/A	N/A	N/A	N/A
Mizuho Bank, LTD	Y	Y	Y	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Michael D. Franko	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N/A	N/A	N/A	N/A
Arab Global Commodities	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	N/A	N/A	N/A	N/A
Citigroup Global Markets	Y	Y	Y	Y	Y	Y	Y	N/A	Y	Y	N/A	N/A	N/A	N/A
David Liew	Y	Y	Y	N/A	Y	Y	Y	N/A	Y	Y	N/A	N/A	N/A	N/A
Simon Posen	Y	Y	Y	N/A	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
The Bank of Tokyo-Mitsubishi	Y	Y	Y	N/A	Y	Y	Y	N/A	Y	Y	N/A	N/A	N/A	N/A
Heet Khara and Nasim Salim	Y	Y	Y	N/A	Y	Y	Y	N	Y	Y	N/A	N/A	N/A	N/A

4.4. Data and metrics definition

4.4.1. Data

We manually extract as many instances of spoofing/layering from the set of prosecution cases. Instances are recorded at two levels of granularity. One level of granularity is security-days (a given security that is manipulated on a given day) and the other is intraday, security-seconds (a given security that is manipulated in a given one-second interval).

The daily sample of prosecuted instances of layering includes 151 manipulated security-days. The intraday sample consists of 1,282 manipulated security-seconds. We obtain detailed trade and quote data (every trade and every quote/order entry/update/cancellation) and limit order book depth data from Thomson Reuters to estimate the daily and intraday metrics. For testing the metrics, we supplement the instances of layering with a matched sample of security-days and security-seconds that have not been manipulated.

4.4.2. Metrics

The metrics are based on the characteristics identified in the previous section as the empirical footprint of spoofing. As we only use publicly available market data in this study, as opposed to regulatory or proprietary data, we can estimate eight of the characteristics: high quoting activity, unbalanced quoting, abnormal cancellations, low execution probability, trades oppose quotes, cancels oppose trades, and cyclical patterns in depth and cancellations[19]:

[19] In a forensic setting, with regulatory or proprietary data, it may be possible to estimate the other characteristics as well. It also may be possible to modify the way these eight metrics are estimated so that they focus in on a particular market participant, rather than being estimated at the market level.

Metric 1. High Quoting Activity

$$HQ_{i,s(d)} = \max_{t \in s(d)} \left(\frac{|EntryAskSize_{i,t} - EntryBidSize_{i,t}|}{AskSize_{i,t} + BidSize_{i,t}} \right) \quad (36)$$

where $EntryAskSize_{i,t}$ is the increase in the aggregate volume of the orders resting on the top 5 ask levels of security i at time t (equal to 0 if there is no increase in the aggregate volume)
$EntryBidSize_{i,t}$ is the increase in the aggregate volume of the orders resting on the top 5 bid levels of security i at time t (equal to 0 if there is no increase in the aggregate volume)
$AskSize_{i,t}$ is the cumulative depth (aggregate order quantity) on the top 5 ask levels of security i at time t
$BidSize_{i,t}$ is cumulative depth on the top 5 bid levels of security i at time t
t indexes time (order book events)
s is a 1-second interval and d is 1-day interval (the metric is calculated for either of these frequencies).

Metric 2. Unbalanced Quoting

$$UQ_{i,s(d)} = \max_{t \in s(d)} \left(\frac{|AskSize_{i,t} - BidSize_{i,t}|}{AskSize_{i,t} + BidSize_{i,t}} \right) \quad (37)$$

where $AskSize_{i,t}$ is the cumulative depth (aggregate order quantity) on the top 5 ask levels of security i at time t
$BidSize_{i,t}$ is the cumulative depth on the top 5 bid levels of security i at time t
t indexes time (order book events)
s is a 1-second interval and d is 1-day interval (the metric is calculated for either of these frequencies).

Metric 3. Abnormal Cancellations

$$AC_{i,s(d)} = \max_{t \in s(d)} \left(\frac{|CancelAskSize_{i,t} - CancelBidSize_{i,t}|}{AskSize_{i,t} + BidSize_{i,t}} \right) \quad (38)$$

where $CancelAskSize_{i,t}$ is the decrease in the aggregate volume of the orders resting on the top 5 ask levels of security i at time t (equal to 0 if there is no decrease in the aggregate volume)
$CancelBidSize_{i,t}$ is the decrease in the aggregate volume of the orders resting on the top 5 bid levels of security i at time t (equal to 0 if there is no decrease in the aggregate volume)
$AskSize_{i,t}$ is the cumulative depth (aggregate order quantity) on the top 5 ask levels of security i at time t
$BidSize_{i,t}$ is the cumulative depth on the top 5 bid levels of security i at time t
t indexes time (order book events)
s is a 1-second interval and d is 1-day interval (the metric is calculated for either of these frequencies).

Metric 4. Low Execution Probability

$$LE_{i,s(d)} = \max_{t \in s(d)} \left(\frac{|AskSizeLevel2to5_{i,t} - BidSizeLevel2to5_{i,t}|}{AskSize_{i,t} + BidSize_{i,t}} \right) \quad (39)$$

where $AskSizeLevel2to5_{i,t}$ is the cumulative depth on ask level 2 to 5 of security i at time t
$BidSizeLevel2to5_{i,t}$ is the cumulative depth on bid level 2 to 5 of security i at time t
$AskSize_{i,t}$ is the cumulative depth (aggregate order quantity) on the top 5 ask levels of security i at time t
$BidSize_{i,t}$ is the cumulative depth on the top 5 bid levels of security i at time t
t indexes time (order book events)

s is a 1-second interval and d is 1-day interval.

Metric 5. Trades Oppose Quotes

For intraday intervals:

$$TOQ_{i,s} = \begin{cases} 1 & \text{if} \quad OIB_{i,s-1}^{Ask} > 10\% \quad \text{and} \quad Trade_{i,s}^{Bid} = 1 \\ 1 & \text{if} \quad OIB_{i,s-1}^{Bid} > 10\% \quad \text{and} \quad Trade_{i,s}^{Ask} = 1 \\ 0 & \text{otherwise} \end{cases} \quad (40)$$

where $Trade_{i,s}^{Ask} = 1$ if there is a trade on the ask side during second s
$Trade_{i,s}^{Bid} = 1$ if there is a trade on the bid side during second s
s is a 1-second interval
The order imbalance variables, $OIB_{i,s-1}^{Ask}$ and $OIB_{i,s-1}^{Bid}$, are defined as:

$$OIB_{i,s-1}^{Ask} = \max_{t \in s-1} \left(\frac{AskSize_{i,t} - BidSize_{i,t}}{AskSize_{i,t} + BidSize_{i,t}} \right) \quad (41)$$

$$OIB_{i,s-1}^{Bid} = \max_{t \in s-1} \left(\frac{BidSize_{i,t} - AskSize_{i,t}}{AskSize_{i,t} + BidSize_{i,t}} \right) \quad (42)$$

where $AskSize_{i,t}$ is the cumulative depth (aggregate order quantity) on the top 5 ask levels of security i at time t
$BidSize_{i,t}$ is the cumulative depth on the top 5 bid levels of security i at time t
t indexes time (order book events)
s is a 1-second interval.

For the daily measure, we sum the number of instances of trades opposing abnormal quoting activity for the whole day:

$$TOQ_{i,d} = \sum_{s \in d} TOQ_{i,s} \tag{43}$$

Metric 6. Cancels Oppose Trades

For intraday intervals:

$$COT_{i,s} = \begin{cases} 1 & \text{if} \quad CIB_{i,s-1}^{Ask} > 10\% \quad \text{and} \quad Trade_{i,s}^{Bid} = 1 \\ 1 & \text{if} \quad CIB_{i,s-1}^{Bid} > 10\% \quad \text{and} \quad Trade_{i,s}^{Ask} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

where $Trade_{i,s}^{Ask} = 1$ if there is a trade on the ask side during second s
$Trade_{i,s}^{Bid} = 1$ if there is a trade on the bid side during second s
s is a 1-second interval
The cancel imbalance variables, $CIB_{i,s-1}^{Ask}$ and $CIB_{i,s-1}^{Bid}$, are defined as:

$$CIB_{i,s-1}^{Ask} = \max_{t \in s-1} \left(\frac{CancelAskSize_{i,t} - CancelBidSize_{i,t}}{AskSize_{i,t} + BidSize_{i,t}} \right) \tag{45}$$

$$CIB_{i,s-1}^{Bid} = \max_{t \in s-1} \left(\frac{CancelBidSize_{i,t} - CancelAskSize_{i,t}}{AskSize_{i,t} + BidSize_{i,t}} \right) \tag{46}$$

where $CancelAskSize_{i,t}$ is the decrease in the aggregate volume of the orders resting on the top 5 ask levels of security i at time t (equal to 0 if there is no decrease in the aggregate volume)
$CancelBidSize_{i,t}$ is the decrease in the aggregate volume of the orders resting on the top 5 bid levels of security i at time t (equal to 0 if there is no decrease in the aggregate volume)
$AskSize_{i,t}$ is the cumulative depth (aggregate order quantity) on the top 5 ask levels of security i at time t
$BidSize_{i,t}$ is the cumulative depth on the top 5 bid levels of security i at time t
t indexes time (order book events)
s is a 1-second interval.

For the daily measure, we sum the number of instances of cancels opposing trades for the whole day:

$$COT_{i,d} = \sum_{s \in d} COT_{i,s} \tag{47}$$

Metric 7. Cyclical Pattern in Depth

The variable $CPD_{i,s(d)}$ counts the number of times that unbalanced quoting, $UQ_{i,t}$, switches sign from greater than 10% to less than -10% or vice versa (in such instances $\mathbb{I}_{\{UQ_{i,t}\}} = 1$) during interval s (a 1-second interval) or interval d (a 1-day interval):

$$CPD_{i,s(d)} = \sum_{t \in s(d)} \mathbb{I}_{\{UQ_{i,t}\}} \tag{48}$$

$$UQ_{i,t} = \frac{AskSize_{i,t} - BidSize_{i,t}}{AskSize_{i,t} + BidSize_{i,t}}$$

where $AskSize_{i,t}$ is the cumulative depth on the top 5 ask levels of security i at time t
$BidSize_{i,t}$ is the cumulative depth on the top 5 bid levels of security i at time t.
t indexes time (order book events)

Metric 8. Cyclical Pattern in Cancellations

The variable $CPC_{i,d(s)}$ counts the number of times that unbalanced cancellations, $UC_{i,t}$, switches sign from greater than 10% to less than -10% or vice versa (in such instances $\mathbb{I}_{\{UC_{i,t}\}} = 1$) during interval s (a 1-second interval) or interval d (a 1-day interval):

$$CPC_{i,s(d)} = \sum_{t \in s(d)} \mathbb{I}_{\{UC_{i,t}\}} \tag{49}$$

$$UC_{i,t} = \frac{CancelAskSize_{i,t} - CancelBidSize_{i,t}}{AskSize_{i,t} + BidSize_{i,t}}$$

where $CancelAskSize_{i,t}$ is the decrease in the aggregate volume of the orders resting on the top 5 ask levels of security i at time t (equal to 0 if there is no decrease in the aggregate volume)
$CancelBidSize_{i,t}$ is the decrease in the aggregate volume of the orders resting on the top 5 bid levels of security i at time t (equal to 0 if there is no decrease in the aggregate volume)
$AskSize_{i,t}$ is the cumulative depth (aggregate order quantity) on the top 5 ask levels of security i at time t
$BidSize_{i,t}$ is the cumulative depth on the top 5 bid levels of security i at time t
t is every point in time

Table 4.3 reports the statistical tests for difference in the daily metrics between security-days that contain manipulation and the matched sample of security-days that do not contain any known manipulation. Most of the metrics show significantly positive differences between

the manipulation and non-manipulation samples. The most significant differences are for manipulation cases executed on the Intercontinental Exchange (ICE) and the New York Mercantile Exchange (NYMEX). NYMEX is the largest energy markets and have long been competitors in terms of liquidity. ICE is the first fully electronic energy market and attracts a high level of liquidity from energy traders.

Table 4.4 shows the same statistics for the intraday granularity (security-seconds). Again, most of the metrics have higher means for the manipulation cases than the non-manipulation observations. The most significant differences are for manipulation cases executed on the Chicago Board Options Exchange (CBOE) and the Index and Option Market (IOM).

Table 4.3
Difference in daily metrics for manipulation vs non-manipulation instances

This table reports averages of the metrics for manipulated security-days compared to non-manipulated security-days. The first column shows the market where manipulation happens (Chicago Board Options Exchange (CBOE), Chicago Board of Trade (CBOT), Commodity Exchange (COMEX), Intercontinental Exchange (ICE), Index and Options Market (IOM), London Stock Exchange (LSE), New York Mercantile Exchange (NYMEX)). The second column (N) shows the number of observations in manipulated and non-manipulated group. In the third column, rows (1) show the mean values for manipulated security-days, rows (2) show the mean values for non-manipulated security-days. Difference (1−2) is the difference between rows (1) and rows (2). The metrics are defined in the text. HQ is *High Quoting Activity*, UQ is *Unbalanced Quoting*, AC is *Abnormal Cancellations*, LE is *Low Execution Probability*, TOQ is *Trades Oppose Quotes*, COT is *Cancels Opposes Trades*, CPD is *Cyclical Pattern in Depth*, CPC is *Cyclical Pattern in Cancellations*. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Exchange	N		HQ	UQ	AC	LE	TOQ	COT	CPD	CPC
CBOE	15	Manipulated securities (1)	1.17	0.76	0.38	0.78	752.40	58.33	121.30	323.30
	388	Non-manipulated securities (2)	0.83	0.57	0.61	0.65	90.26	5.14	18.39	35.34
		Difference (1−2)	0.35	0.19^{***}	−0.22	0.13^{***}	662.10^{***}	53.19^{***}	102.90^{***}	288^{***}
CBOT	10	Manipulated securities (1)	1.41	0.88	1.08	0.89	2,678.10	523.20	932.90	1,238.70
	4,167	Non-manipulated securities (2)	1.61	0.72	1.33	0.76	121.50	30.60	337.10	398.30
		Difference (1−2)	−0.20	0.15^{***}	−0.24	0.13^{*}	$2,556.60^{***}$	492.60^{***}	595.80	840.40
COMEX	17	Manipulated securities (1)	2.96	0.73	1.83	0.91	4,718.90	832.90	2,576.70	6,728.60
	1,581	Non-manipulated securities (2)	3.77	0.91	2.87	0.73	187	66.75	1,955.40	1,735.90
		Difference (1−2)	−0.82	0.17^{***}	−1.04	0.18^{***}	$4,531.80^{***}$	766.20^{***}	621.30	$4,992.70^{***}$
ICE	76	Manipulated securities (1)	4.77	0.92	4.67	0.93	832	134.80	4,639.80	14,511
	5,802	Non-manipulated securities (2)	1.88	0.74	1.69	0.69	29.30	6.57	549.90	933.20
		Difference (1−2)	2.88^{***}	0.18^{***}	2.98^{***}	0.23^{***}	802.70^{***}	128.20^{***}	$4,089.90^{***}$	$13,578.20^{***}$
IOM	17	Manipulated securities (1)	0.96	0.81	0.76	0.91	7,194.40	933.70	976	5,087.10
	587	Non-manipulated securities (2)	1.34	0.72	1.30	0.76	477.40	96.09	1,194	1,087.20
		Difference (1−2)	−0.38	0.09^{*}	−0.55	0.04	$6,717^{***}$	837.60^{***}	−218.30	$3,999.90^{***}$
LSE	23	Manipulated securities (1)	1.38	0.92	0.98	0.92	209	49.91	256.60	357.90
	110,017	Non-manipulated securities (2)	2.27	0.60	1.57	0.63	14.62	3.68	53.80	59.10
		Difference (1−2)	0.89	0.31^{***}	−0.59	0.29^{***}	194.30^{***}	46.23^{***}	202.70	298.70^{***}
NYMEX	9	Manipulated securities (1)	3.16	0.91	3.04	0.93	6,231.60	1,082.30	1,582.20	5,383.10
	6,617	Non-manipulated securities (2)	2.19	0.71	1.47	0.74	105.20	44.50	758	788
		Difference (1−2)	0.95	0.20^{**}	1.57^{***}	0.19^{**}	$6,126.40^{***}$	$1,037.90^{***}$	824.20	$4,595.10^{***}$

Table 4.4

Difference in intraday metrics for manipulation vs non-manipulation instances

This table reports averages of the metrics for manipulated security-seconds compared to non-manipulated security-seconds. The first column shows the market where manipulation happens (Chicago Board Options Exchange (CBOE), Commodity Exchange (COMEX), Index and Options Market (IOM), New York Mercantile Exchange (NYMEX)). The second column (N) shows the number of observations in manipulated and non-manipulated group. In the third column, rows (1) show the mean values for manipulated security-seconds, rows (2) show the mean values for non-manipulated security-seconds. Difference (1−2) is the difference between rows (1) and rows (2). The metrics are defined in the text. HQ is *High Quoting Activity*, UQ is *Unbalanced Quoting*, AC is *Abnormal Cancellations*, LE is *Low Execution Probability*, TOQ is *Trades Oppose Quotes*, COT is *Cancels Opposes Trades*, CPD is *Cyclical Pattern in Depth*, CPC is *Cyclical Pattern in Cancellations*. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Exchange	N		HQ	UQ	AC	LE	TOQ	COT	CPD	CPC
CBOE	95	Manipulated Securities (1)	0.01	0.15	0.01	0.10	0.25	0.03	0.05	0.08
	1,295,054	Non-manipulated securities (2)	0.22	0.08	0.04	0.06	0.01	0.00	0.00	0.00
		Difference (1−2)	0.21^{***}	0.07^{***}	0.03^{***}	0.04^{***}	0.24^{***}	0.03^{***}	0.05^{***}	0.08^{***}
COMEX	388	Manipulated Securities (1)	0.12	0.32	0.05	0.32	0.12	0.02	0.06	0.18
	1,467,509	Non-manipulated securities (2)	0.11	0.31	0.04	0.32	0.05	0.01	0.03	0.08
		Difference (1−2)	0.01	0.01	0.01^{**}	0.00	0.07^{***}	0.01^{**}	0.03^{***}	0.10^{***}
IOM	683	Manipulated Securities (1)	0.17	0.17	0.05	0.14	0.33	0.12	0.20	0.29
	949,068	Non-manipulated securities (2)	0.02	0.14	0.01	0.14	0.07	0.01	0.01	0.05
		Difference (1−2)	0.15^{***}	0.03^{***}	0.04^{***}	0.00	0.26^{***}	0.11^{***}	0.19^{***}	0.23^{***}
NYMEX	793	Manipulated Securities (1)	0.11	0.20	0.04	0.19	0.13	0.03	0.05	0.13
	690,074	Non-manipulated securities (2)	0.09	0.20	0.03	0.19	0.07	0.01	0.02	0.06
		Difference (1−2)	0.02^{***}	0.00^{*}	0.01^{***}	0.00^{*}	0.06^{***}	0.02^{***}	0.03^{***}	0.07^{***}

Table 4.5 shows the correlations among the intraday and daily spoofing metrics. All correlations are positive and range from small values up to 0.95.[20]

Table 4.5
Correlations of spoofing metrics

This table reports the correlations among the spoofing metrics at intraday and daily frequencies. The metrics are defined in the text. HQ is *High Quoting Activity*, UQ is *Unbalanced Quoting*, AC is *Abnormal Cancellations*, LE is *Low Execution Probability*, TOQ is *Trades Oppose Quotes*, COT is *Cancels Opposes Trades*, CPD is *Cyclical Pattern in Depth*, CPC is *Cyclical Pattern in Cancellations*.

Variables	HQ	UQ	AC	LE	TOQ	COT	CPD	CPC
Panel A: Intraday metrics								
HQ	1	0.36	0.28	0.33	0.14	0.08	0.09	0.40
UQ	0.36	1	0.22	0.94	0.15	0.04	0.01	0.05
AC	0.28	0.22	1	0.21	0.17	0.33	0.22	0.40
LE	0.33	0.94	0.21	1	0.12	0.03	0.01	0.05
TOQ	0.14	0.14	0.17	0.12	1	0.22	0.12	0.09
COT	0.08	0.04	0.33	0.03	0.22	1	0.16	0.32
CPD	0.09	0.01	0.22	0.01	0.12	0.16	1	0.19
CPC	0.11	0.05	0.40	0.05	0.09	0.32	0.19	1
Panel B: Daily metrics								
HQ	1	0.06	0.07	0.06	0.01	0.02	0.03	0.03
UQ	0.06	1	0.07	0.95	0.46	0.39	0.23	0.17
AC	0.75	0.07	1	0.07	0.01	0.02	0.04	0.04
LE	0.06	0.95	0.07	1	0.44	0.38	0.21	0.17
TOQ	0.01	0.46	0.04	0.44	1	0.95	0.39	0.39
COT	0.02	0.39	0.02	0.38	0.95	1	0.47	0.47
CPD	0.03	0.22	0.04	0.21	0.47	0.54	1	0.80
CPC	0.03	0.17	0.04	0.17	0.39	0.47	0.80	1

4.4.3. Using the Intraday Metrics to Detect Spoofing

To detect spoofing using the metrics together, we combine them into a probability index where each metric has an optimal weight. To do this, we estimate logit models that use the metrics to predict whether manipulation occurred in a given observation, similar to the approach in Comerton-Forde and Putniņš (2011). For the intraday metrics we estimate the

[20] There is a strong correlation between *Low Execution Probability* and *Unbalanced Quoting* (0.95). Therefore, we exclude variable $LE_{i,t}$ in our subsequent multivariate spoofing detection models to avoid issues of multicollinearity.

following model using the manipulated security-seconds and a non-manipulated sample by taking all non-manipulated security-seconds on the day when there is manipulation:

$$ln\left(\frac{P}{1-P}\right)_{i,s} = \alpha + \beta_1 HQ_{i,s-1} + \beta_2 UQ_{i,s-1} + \beta_3 AC_{i,s} + \beta_4 CPD_{i,s} + \beta_5 CPC_{i,s} + \beta_6 TOQ_{i,s} + \beta_7 COT_{i,s} + \varepsilon_{i,d} \quad (50)$$

where $ln\left(\frac{P}{1-P}\right)$ is the log-odds of the manipulator trading in security i during second s

$HQ_{i,s-1}, UQ_{i,s-1}, AC_{i,s}, CPD_{i,s}, CPC_{i,s}, TOQ_{i,s}, COT_{i,s}$ are the intraday metrics defined earlier

We lag the *High Quoting Activity* and *Order Imbalance* metrics when using the intraday frequency because in most cases, the manipulator submits large spoofing orders and creates a significant limit order book imbalance before executing the genuine order(s). It can take seconds to minutes for other traders to respond to the manipulator's spoofing orders. Cancellation of the spoofing orders typically happens soon after the manipulator has managed to trade so we use contemporaneous cancellation, but even so we may be understating the importance of the cancellation metric as some of the cancellations may occur in subsequent seconds.

These considerations illustrate one of the challenges in measuring spoofing – the characteristics may be displayed sequentially rather than concurrently, and different cases operate at different frequencies. Therefore, our findings can be interpreted as establishing the minimum significance of these characteristics. Applied in a forensic setting to an individual case, one can "tune" the frequency to the particular case. Table 4.6 reports the results of the logit model. The results indicate that the majority of variables significantly predict spoofing at the intraday frequency. The positive signs associated with all variables suggest that an increase in *High Quoting Activity*, *Unbalanced Quoting*, *Abnormal Cancellation*, *Trades Oppose Quotes*, *Cancels Oppose Trades*, *Cyclical Pattern in Depth*, *Cyclical Pattern in Cancellations*, have ability to identify manipulation conducted during that second. The logistic regression suggests that among these variables, the three variables that makes the highest marginal contribution to identifying manipulation are *High Quoting Activity*, *Unbalanced Quoting*, and *Trades Oppose Quotes*.

The AUC reported in Table 4.6 is the Area Under the ROC Curve (AUC), which is a measure of the accuracy of a categorical predictor. The higher the AUC, the better the performance of the model. AUC values above 0.50 indicate predictive accuracy beyond pure

chance. We estimate the AUC as an out-of-sample measure using leave-one-out cross-validation. The AUC score for this model, 0.77, is statistically different from 0.50 at a high confidence level (p-value less than 0.001).

Table 4.6
Logistic regression predicting spoofing at the intraday frequency

This table reports the results of logistic regression models where the dependent variable is log-odds of trades by the manipulator in known spoofing cases and the independent variables are empirical characteristics of spoofing estimated per security-second. HQ is *High Quoting Activity*, UQ is *Unbalanced Quoting*, AC is *Abnormal Cancellations*, LE is *Low Execution Probability*, TOQ is *Trades Oppose Quotes*, COT is *Cancels Opposes Trades*, CPD is *Cyclical Pattern in Depth*, CPC is *Cyclical Pattern in Cancellations*. The table reports coefficient estimates and z-statistics. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Variables	Coefficients	z-statistic
HQ	1.33	8.45^{***}
UQ	1.23	9.36^{***}
AC	0.85	5.40^{***}
TOQ	0.99	12.06^{***}
COT	0.69	4.69^{***}
CPD	0.03	2.20^{***}
CPC	0.10	3.34^{***}
Intercept	-8.52	-227.24^{***}
Observations		4,403,391
Out-of-sample AUC		0.77
Out-of-sample AUC different from AUC=50%, p-value		<0.001

4.4.4. Using Daily Metrics to Detect Spoofing

At the daily frequency, there is a high correlation between the *Cyclical Pattern in Depth* and *Cyclical Pattern in Cancellations* as indicated in Table 4.5 Panel B. Therefore, we combine the *Cyclical Pattern in Depth* and *Cyclical Pattern in Cancellations* into a new variable *Cyclical Pattern in Depth and Cancellations* that captures the number of spoofing cycles in a day:

$$CPDC_{i,d} = CPD_{i,d} + CPC_{i,d} \tag{51}$$

Similarly, we add the *Trades Oppose Quotes* and *Cancels Oppose Trades*, as there is a high correlation between the *Trades Oppose Quotes* and *Cancels Oppose Trades* in Table 4.5 Panel B. The variable *Trades Oppose Cancels and Quotes* is the sum of the two variables, which is defined as:

$$TOCQ_{i,d} = TOQ_{i,d} + COT_{i,d} \tag{52}$$

Following the same approach as for the intraday metrics, we estimate the following logit model using the manipulated and non-manipulated security-days, where the non-manipulated sample is obtained by taking all non-manipulated securities on the same exchange on the days of manipulation:

$$ln\left(\frac{P}{1-P}\right)_{i,d} = \alpha + \beta_1 HQ_{i,d} + \beta_2 UQ_{i,d} + \beta_3 AC_{i,d} + \beta_4 CPDC_{i,d} + \beta_5 TOCQ_{i,d} + \varepsilon_{i,d} \tag{53}$$

where $ln\left(\frac{P}{1-P}\right)_{i,d}$ is the log-odds of the manipulator trading in security i during day d, and

$HQ_{i,d}, UQ_{i,d}, AC_{i,d}, CPDC_{i,d}, TOCQ_{i,d}$, are the metrics at the daily frequency

Table 4.7 reports the result. Among the daily metrics, three stand out as having a strong positive incremental ability to detect spoofing, controlling for other characteristics – *Unbalanced Quoting*, *Trades Oppose Cancels and Quotes*, and *Cyclical Pattern in Depth and Cancellations*. However, *High Quoting Activity* does not contribute to predicting manipulation beyond what is already captured by the other characteristics. At daily level, *High Quoting Activity* may already be captured by the *Unbalanced Quoting* metric.

Abnormal Cancellations also does not predict spoofing at daily level beyond what is captured by the other metrics. As Khomyn and Putniņš (2021) discuss, high-frequency traders can have a high order cancellation rate for legitimate reasons, which might explain why at a daily frequency the *Abnormal Cancellations* metric does not appear statistically significant. Additionally, given that the univariate results did show a positive relation between the *Abnormal Cancellations* and spoofing, it could be that other metrics such as *Unbalanced Quoting*, *Trades Oppose Cancels and Quotes* and the *Cyclical Pattern in Depth and Cancellations* measures capture the high cancellation rates.

Interestingly, at the daily frequency, the model's ability to distinguish between manipulation and non-manipulation is higher, with an AUC of 0.93 in Table 4.7. The increased classification accuracy may be as a result of facing less of a challenge in capturing lead/lag relations between the characteristics, which is an issue when working at the one-second frequency. It may also be due to less noise in the daily metrics, as one-second intervals are more prone to temporary extreme values.

Table 4.7
Logistic regression predicting spoofing at the daily frequency.
This table reports the results of logistic regression models where the dependent variable is log-odds of trades by the manipulator in known spoofing cases and the independent variables are empirical characteristics of spoofing estimated per security-day. HQ is *High Quoting Activity*, UQ is *Unbalanced Quoting*, AC is *Abnormal Cancellations*, LE is *Low Execution Probability*, $TOCQ$ is *Trades Oppose Cancels and Quotes*, $CPDC$ is *Cyclical Pattern in Depth and Cancellations*. The table reports coefficient estimates and z-statistics. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Variables	Coefficients	z-statistic
HQ	-0.02	-0.90
UQ	7.32	7.24^{***}
AC	0.01	0.92
$TOCQ$	3.99	13.19^{***}
$CPDC$	0.39	6.13^{***}
Intercept	-11.61	-12.97^{***}
Observations		129,235
Out-of-sample AUC		0.93
Out-of-sample AUC different from AUC=50%, p-value		<0.001

4.5. Machine learning models for detection and out-of-sample validation

In this section, we examine whether the detection performance of the spoofing metrics can be increased by combining them in machine learning models that allow for interactions and non-linearities. Specifically, we estimate the probability of manipulation using random forest and boosted tree classification models.

Random forest and boosted tree models are well-suited for this task, as they can combine variables into a tree model using split rules, which are particularly useful when there is a significant degree of interaction among independent variables. The spoofing metrics are likely to interact with one another and are closely connected, as shown before. A general spoofing order may involve a manipulator submits spoofing orders, trick other market participants into following the spoofing orders, and then cancel the substantial orders after trading on the genuine side of the market. If any of these steps is unsuccessful, the spoofing may not be easily identified and may be mistaken for other legitimate trading.

We apply *k*-fold cross validation to assess the accuracy of the non-linear models, setting $k = 3$. We randomly divide the sample into three equal subsamples, estimate the model on two samples and test the estimated model on the other one. We then compare the result with logistic regression. To evaluate the precision of the model, we create a graphical representation called the ROC curve. This curve demonstrates the model's effectiveness without considering prior probabilities or classification thresholds. It shows the balance between correctly identified positives (sensitivity) and incorrectly identified negatives (one minus specificity).

Figure 4.7 shows that the detection validity (AUC) of the logit model using intraday metrics is 0.77 in the out-of-sample setting, as discussed previously. In comparison, the AUC of the tree-based and random forest models using the same intraday metrics is 0.81 and 0.82, respectively. The statistical analysis shows non-linear machine learning models outperform simple logistic regression in diagnosing spoofing given the same set of characteristics.

Turning to the daily version of the metrics, Figure 4.8 shows that the out-of-sample detection accuracy (AUC) of the logit model using daily metrics is 0.93, as discussed previously. The random forest and boosted tree models both outperform logistic regression with AUC of 0.96 and 0.97, respectively. Therefore, at both intraday and daily frequencies random forest and boosted tree models outperform logistic regression. These results show that nonlinearity and variable interactions contribute to the ability to detect spoofing.

These are extremely high classification accuracies considering the nature of the problem. For some context, Comerton-Forde and Putniņš (2011) design a model to detect closing price manipulation and obtain an out-of-sample AUC of 0.825 with a logit model. Acknowledge of the finding, our model is better at empirically detecting spoofing using daily metrics compared to previous model that detect closing price manipulation.

Figure 4.7

Out-of-sample classification accuracy of the spoofing detection models using intraday metrics

The figure illustrates the discriminatory power of logistic regression, random forest, and boosted tree models estimated with intraday spoofing metrics. The plot shows the Receiver Operating Characteristics curve (ROC curve) from k-fold cross validation and measures the area under the curve (AUC).

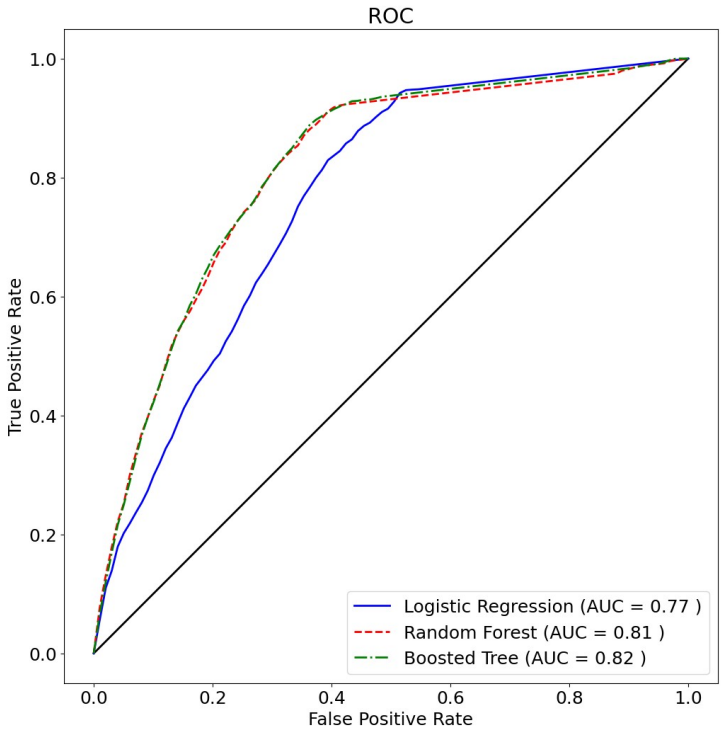
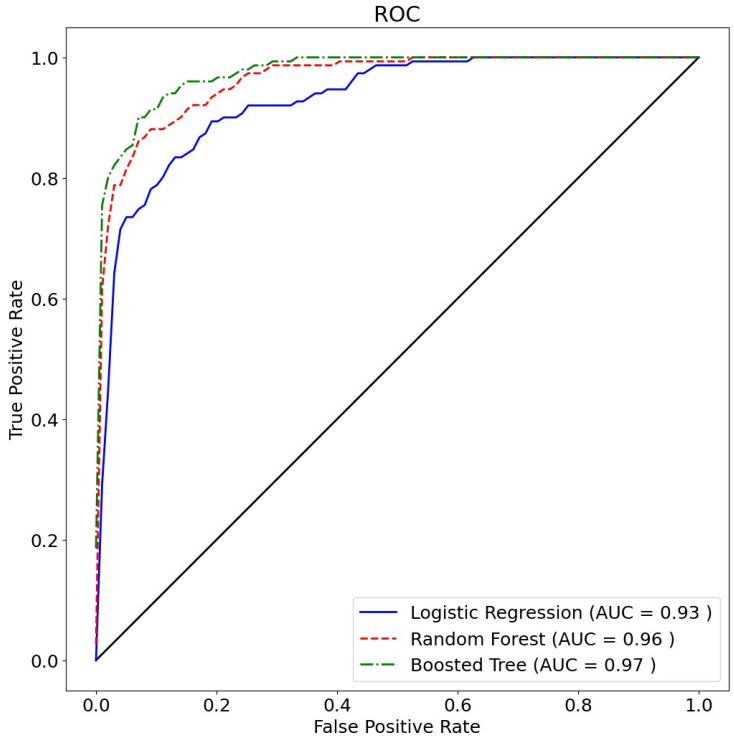


Figure 4.8

Out-of-sample classification accuracy of the spoofing detection models using daily metrics

The figure illustrates the discriminatory power of logistic regression, random forest, and boosted tree models estimated with daily spoofing metrics. The plot shows the Receiver Operating Characteristics curve (ROC curve) from k-fold cross validation and measures the area under the curve (AUC).



While the machine learning models show higher detection accuracy, a potential downside is the ability to easily inspect what is occurring within the model. To obtain some insights about which metrics are deemed the most important by the model, we use mean decrease in prediction error to evaluate the importance of variable i for predicting manipulation by adding up the weighted error decreases for all nodes where variable i is used. Feature importance of a variable i is normalized by the sum of all feature values present in the tree, then standardized by the number of trees T:

$$Importance(M_i) = \frac{1}{T} \times \frac{\sum_j^{node\ j\ splits\ on\ metric\ i} n_{i,j}}{\sum_j^{all\ nodes} n_{i,j}} \tag{54}$$

where $n_{i,j}$ is node j importance of variable i (measured as the decrease in error if including i in that node)

Figure 4.9 shows the results summarizing the importance of each intraday metric in the random forest model. The metrics that make the largest contribution are *High Quoting Activity*, *Unbalanced Quoting*, and *Abnormal Cancellations*. *Cyclical Pattern in Depth* and *Cyclical Pattern in Cancellations* are less important for intraday detection, which indicates that within a second, there might not be enough time for the manipulator to complete a spoofing cycle.

Figure 4.10 presents the relative importance of the daily metrics. In contrast to the intraday measures, the *Cyclical Pattern in Depth and Cancellations*, as well as *Trades Oppose Cancels and Quotes*, are the most important metrics in detecting spoofing at daily level. The result is consistent with the logistic regressions.

Figure 4.9
Importance of each intraday metric in the random forest model

The figure illustrates the level of importance of each intraday spoofing metric in the random forest model. The vertical axis shows the variable importance metric. The horizontal axis shows variable names. HQ is *High Quoting Activity*, UQ is *Unbalanced Quoting*, AC is *Abnormal Cancellations*, LE is *Low Execution Probability*, TOQ is *Trades Oppose Quotes*, COT is *Cancels Opposes Trades*, CPD is *Cyclical Pattern in Depth*, CPC is *Cyclical Pattern in Cancellations.*

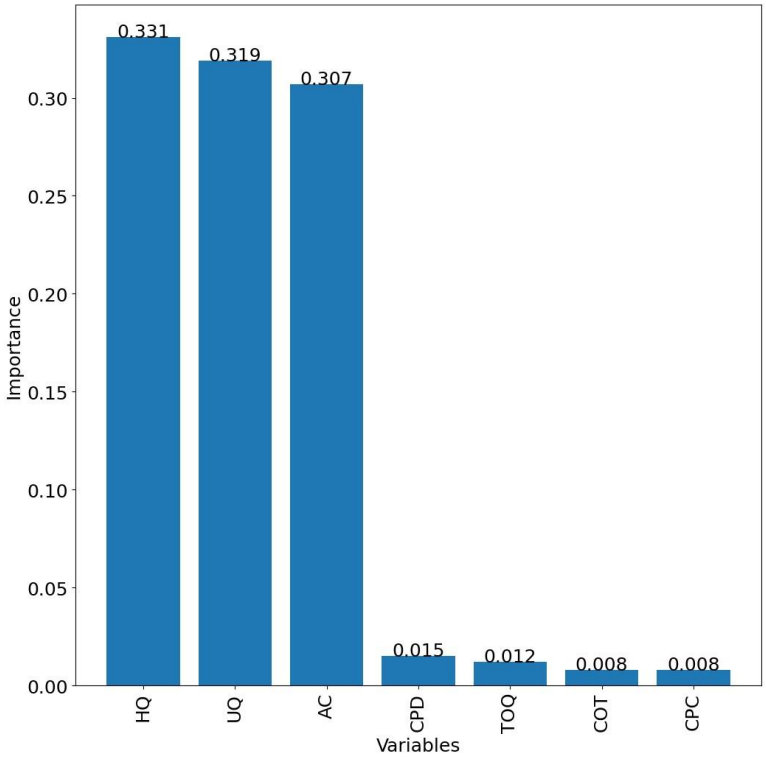
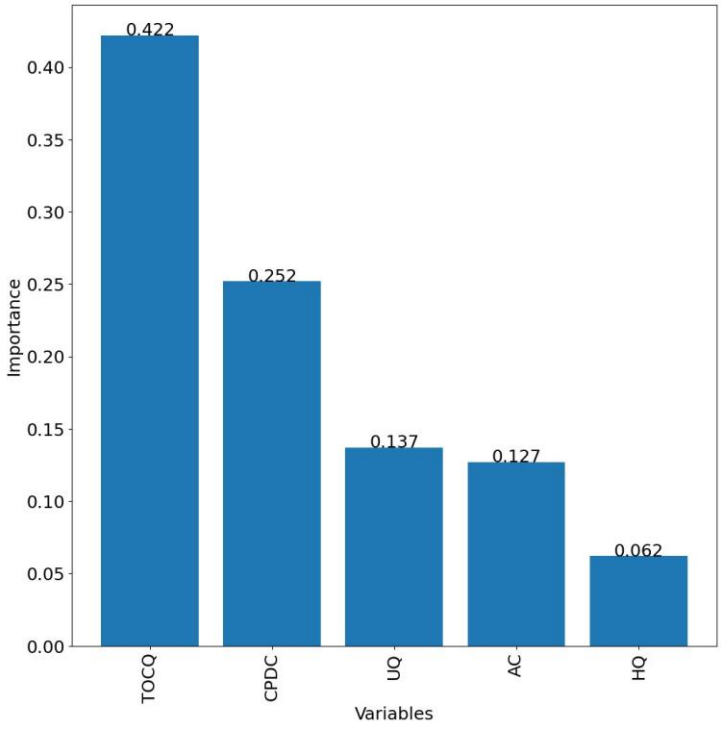


Figure 4.10
Importance of each daily metric in the random forest model

The figure illustrates the level of importance of each daily spoofing metric in the random forest model. The vertical axis shows the variable importance metric. The horizontal axis shows variable names. HQ is *High Quoting Activity*, UQ is *Unbalanced Quoting*, AC is *Abnormal Cancellations*, LE is *Low Execution Probability*, $TOCQ$ is *Trades Oppose Cancels and Quotes*, $CPDC$ is *Cyclical Pattern in Depth and Cancellations*.



In addition to utilizing k-fold cross-validation, a commonly used technique in machine learning literature, we also employ a second validation approach to assess the models' ability to detect spoofing in a specific case that has not been included in the model training process. For this we choose the case of Da Vinci Invest.[21] We observe "saw-tooth" pattern in this case, which takes place over several minutes in a manipulated day, which is representative of layering. We estimate the models without this case included and then test whether the models can detect the spoofing in this case.

Table 4.8 provides results at the classification threshold of 0.5, which means that if the model-implied probability of spoofing is greater than 0.5, we flag the instance as having predicted manipulation.

Table 4.8
Out-of-sample detection of spoofing

This table illustrates the accuracy of the models in out-of-sample detection of spoofing at the intraday level using logistic regression, random forest models, and boosted tree models. The columns labeled "Predicted" provide the number of predicted Non-Manipulation and Manipulation instances using the classification threshold of 0.5. The next three columns provide Accuracy ((True Positives + True Negatives) / Total), Precision (True Positives / (True Positives + False Positives)), and Sensitivity (True Positives / (True Positives + False Negatives)) scores based on the confusion matrix.

At threshold P>= 0.5		Predicted: No	Predicted: Yes	Accuracy	Precision	Sensitivity
Panel A: Logistic Regression	Actual: No	12,545	3			
	Actual: Yes	399	37	0.97	0.93	0.08
Panel B: Random Forest	Actual: No	12,547	1			
	Actual: Yes	407	29	0.97	0.97	0.07
Panel C: Boosted Tree	Actual: No	12,540	8			
	Actual: Yes	239	197	0.98	0.96	0.45

All three models have a strong ability to distinguish spoofing from legitimate trading, as evidenced by their high levels of accuracy and precision. Sensitivity is low, resulting in a conservative classified more prone to false negatives than false positives, but can be increased by choosing a lower classification threshold. The boosted tree model is particularly effective in identifying spoofing in this validation test.

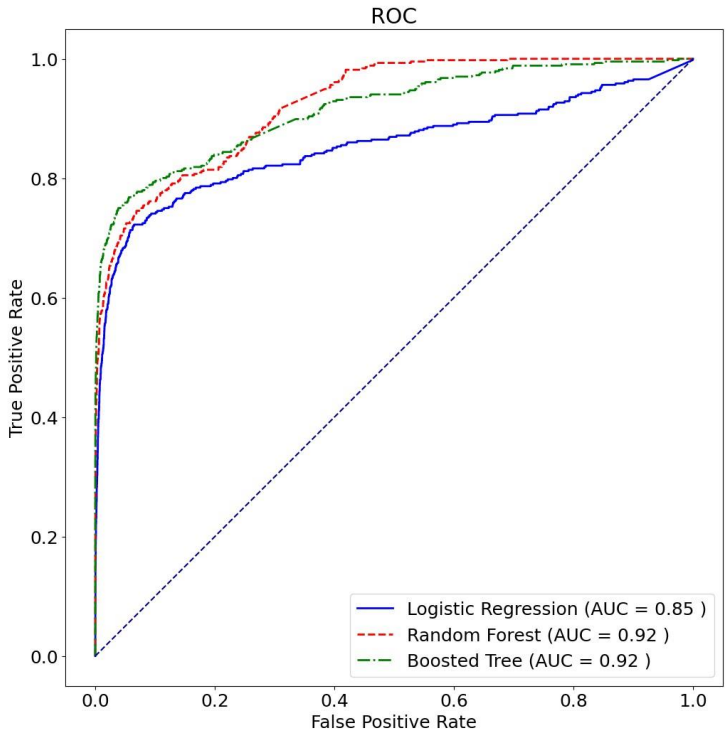
Figure 4.11 shows the ROC curve for this validation exercise and reports the AUC for the three models in this test: 0.85, 0.92, and 0.92 for logistic regression, random forest, and

[21] In this case, we are only provided with the start and end times during which manipulation occurred on a specific day. We assume that all minutes in the manipulated period provided by regulators contain manipulation. This is a conservative assumption because any incorrect labelling of the data is likely to harm the measured classification accuracy.

boosted tree, respectively. The results show that all three models have strong ability to detect spoofing at daily and intraday horizons, using two types of out-of-sample validation tests.

Figure 4.11
Out-of-sample detection of spoofing – Da Vinci case

The figure illustrates the discriminatory power of logistic regression, random forest, and boosted tree models estimated with intraday spoofing metrics. The plot shows the Receiver Operating Characteristics curve (ROC curve) from out-of-sample cross validation on the Da Vinci spoofing case. The plot also reports the area under the curve (AUC).



4.6. Conclusion

In recent years, there has been a proliferation of spoofing and layering in markets, as evidenced by the sharp rise in the number of prosecuted cases. Identification of spoofing poses a significant challenge for market authorities, as it can appear similar to legitimate trading practices. The regulatory framework for spoofing is intentionally vague (to avoid easy loopholes), which can impede the public's understanding of this manipulation category. Our study exhibits an empirical examination of spoofing, using data collected from prosecuted cases.

We find that at the intraday level, metrics such as *Unbalanced Quoting*, *High Quoting Activity*, *and Abnormal Cancellations* are particularly useful in identifying spoofing. At the daily level, the *Cyclical Pattern in Depth and Cancellations* and *Trades Oppose Cancels and Quotes* have the highest ability to detect spoofing.

Given the complex interactions between the various characteristics of spoofing, we also test random forest and boosted tree classification models to detect spoofing at the second-interval level. We find that these machine learning techniques, with their ability to account for interactions between variables, have superior out-of-sample prediction capabilities for spoofing.

Our results suggest that by using machine learning models on a set of spoofing metrics, regulators and market participants can more effectively distinguish between illegitimate and legitimate trading.

Chapter 5: Conclusion

This chapter outlines the key questions and conclusions of this thesis. It also discusses avenues for future research.

5.1. What can machine learning models teach about the drivers of company value?

The boosted tree analysis provides us with new insights into the interactions between different drivers of company value. Specifically, we observe significant interactions between the historical growth rate and risk proxies, and between the historical growth rate and dividend payout ratio.

Our findings also suggest that the growth rate is not the only factor that affects the price-to-book ratio. While the price-to-book ratio generally increases with a higher growth rate, it decreases when the beta on the market increases at the same level of growth rate. Additionally, we find that the dividend payout ratio does not always have a positive relationship with firm value; rather, the positive effect only begins when the dividend payout ratio is in the top quintile cross-sectionally. The positive effect of dividend payout is highest when companies are in the top quintile of long-term growth. Furthermore, we observe that longer-term growth is valued more than short-term growth.

The effects of growth on equity value and total company value differ. Higher growth, for both short-term and long-term, leads to lower enterprise value at all levels of risk, indicating that debt value may be lower for companies with higher growth. Furthermore, our research reveals a concave relationship between the reinvestment rate and the valuation of a company. These interactions are observed even after controlling for peer-analyst groups, which contain implicit information that is not reflected on financial statements.

We use boosted trees because they can handle many input variables, including information from peer firms that are not explicitly part of the financial statements. Linear regression is not suitable for this task because there are too many variables and potential interactions to consider. Boosted trees are better at incorporating soft information and financial information, which results in better predictions of future value.

Including information about a company's peers significantly improves valuation. The peer firms in the dataset reflect similarities among firms that are not fully captured in financial statements. We discover the significance of unobserved information that econometricians may

not directly perceive. Our findings demonstrate that integrating expert opinions on the economic connection between firms leads to a better prediction of the value of a company.

Compared to using hard information alone, our results show that boosted trees outperform linear regression regarding R^2 and squared error for price-to-book and enterprise-value-to-invested-capital. When we use the Shapley value rule to break down the R^2 into variable-level components, our findings indicate that implicit information significantly contributes to explaining the variations in price-to-book and enterprise-value-to-invested-capital ratios.

5.2. How does machine learning impact market efficiency?

Machine learning is better equipped to handle the multi-dimensional and interrelated nature of the information that affects asset prices compared to traditional linear statistics tools including linear regression. We consider that the notable disparity in return predictability between linear and non-linear models as "non-linear inefficiency".

The gap shows that there are complex relations in how information is reflected in prices. We find that applied to past data, machine learning methods show the considerable ability to predict returns. This result holds when the methods are not yet in widespread use, but the predictability gradually declined as more complex models are incorporated into investment decisions. We use the growth of quantitative mutual funds and the success of machine learning publications as indicators of technological advancements and find that they are drivers of increases in non-linear market efficiency.

5.3. How to detect layering and spoofing in markets?

We identify and define 14 key empirical characteristics of spoofing. High quoting activity (entering and cancelling many limit orders on one market side) is a key feature of most of the layering strategies seen in prosecution cases. By employing substantial volumes of fictitious orders, the likelihood of exerting a remarkable influence on the perceived supply or demand increases, thereby influencing the trading decisions of other market participants.

The quoting activities project significant imbalance, which indicate a concentration either on the buy or sell side at a particular point in time. This imbalance results in a false pressure in transaction activities.

Layering orders are not intended to execute, so they are typically canceled and resubmitted as market conditions change to maintain low execution probability. As a result, there is a high order cancellation rate in layering/spoofing strategies. Also, to maintain the low

execution probability, layering orders are often placed in the market at price steps behind the best quotes or the back of a long queue of orders at a price level.

Manipulators using intraday layering strategies are likely to accumulate long or short positions intraday, but typically they do not hold positions overnight. The trade direction of market manipulators using layering often opposes their quotes, e.g., a trader with many buy orders in the market will often actually execute sells, and vice versa. Manipulators rely on the act of order cancellation that is aligned with trade on the other side. The order cancellations is subsequent to execute an order on the opposing bid or ask. For instance, after making a purchase, they cancel the manipulative sell orders.

Some manipulators use dark or hidden orders as part of their strategy while engaging in layering. Manipulators tend to layer the order book's bid side when they have a long position and vice versa and layer the ask side when they have a short position.

The other characteristics are related to cyclical layering, where a manipulator repeats the layering cycle several times on alternating sides of the market. These include cyclical patterns in cancellations, inventory, and mid-prices.

We construct eight empirical metrics that can detect spoofing at both the daily and intraday levels. At the intraday level, metrics such as *High Quoting Activity*, *Unbalanced Quoting*, and *Abnormal Cancellations* are particularly useful in identifying spoofing. Nevertheless, when analyzing spoofing frequency on a daily basis, we observe that *Cyclical Pattern in Depth and Cancellations*, as well as *Trades Oppose Cancels and Quotes* emerge as the most distinguishing characteristics of spoofing.

Given the complex interactions between the various characteristics of spoofing strategies, we also employ random forest and boosted tree classification models to predict spoofing at the minute-interval level. Our findings demonstrate that machine learning models, with their capability to consider the interactions among variables, exhibit higher accuracy in identifying spoofing activities when compared to other approaches in out-of-sample testing.

5.4. Future research direction

Machine learning techniques that help to explain the reason behind the prediction of the machine learning model (as opposed to black-box models) are under-used in finance research. They can be useful in future research in asset pricing, specifically, due to the evidence in this thesis of a degree of inefficiency in how non-linear information and interactions are reflected in prices. Furthermore, given the abundance of alternative data and the availability of big data, future studies can leverage machine learning techniques to explore the predictive

value of these new data types in forecasting returns. For example, data from images, scripts, and audio are not studied widely due to the limitation of traditional approaches.

Another fruitful area of further research is to explore the implications of generative AI such as ChatGPT and similar models. These have only recently become available to the general public and investors but already generated a high level of interest due to their interesting abilities to reason and see links between information. Will these models drive further informational efficiency gains, or will they add noise to investor decisions?

Given the continued digitalization of finance, future research may examine the importance and consequence of new market designs such as decentralized markets and cryptocurrency markets. Due to its high level of transparency, the data on blockchains are more accessible compared to other traditional markets, making it a promising avenue of research in the future.

References

Abarbanell, J.S., and B.J. Bushee, 1998, Abnormal returns to a fundamental analysis strategy, *Accounting Review* 73, 19–45.

Abis, S., 2020, Man vs. Machine: Quantitative and discretionary equity management, *Working Paper*, https://ssrn.com/abstract=3717371.

Allen, F., and D. Gale, 1992, Stock-price manipulation, *Review of Financial Studies* 5, 503–529.

Allen, F., L. Litov, and J. Mei, 2006, Large investors, price manipulation, and limits to arbitrage: An anatomy of market corners, *Review of Finance* 10, 645–693.

Asness, C.S., T.J. Moskowitz, and L. H. Pedersen, 2013, Value and momentum everywhere, *Journal of Finance* 68, 929–985.

Atanasov, V., S.V. Møller, and R. Priestley, 2020, Consumption fluctuations and expected returns, *Journal of Finance* 75, 1677–1713.

Barbopoulos, L.G., R. Dai, T.J. Putniņš, and A. Saunders, 2021, Market efficiency in the age of machine learning, *Working Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3783221.

Bartram, S.M., and M. Grinblatt, 2018, Agnostic fundamental analysis works, *Journal of Financial Economics* 128, 125–147.

Beggs, W., J. Brogaard, and A. Hill-Kleespie, 2021, Quantitative investing and market instability, *Working Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281447.

Bernhardt, D., and R.J. Davies, 2009, Smart fund managers? Stupid money?, *Canadian Journal of Economics* 42, 719–748.

Bhojraj , S., and C.M.C. Lee, 2002, Who is my peer? A valuation-based approach to the selection of comparable firms, *Journal of Accounting Research* 40, 407–439.

Biais, B., P. Hillion, and C. Spatt, 1995, An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* 50, 1655–1689.

Brogaard, J., A. Carrion, T. Moyaert, R. Riordan, A. Shkilko, and K. Sokolov, 2018, High frequency trading and extreme price movements, *Journal of Financial Economics* 128, 253–265.

Brogaard, J., T. Hendershott, and R. Riordan, 2014, High-frequency trading and price discovery, *Review of Financial Studies* 27, 2267–2306.

Brogaard, J., D. Li, and J. Yang, 2022, Does high frequency market manipulation harm market quality?, *Working Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4280120.

Bryzgalova, S., M. Pelger, and J. Zhu, 2021, Forest through the trees: Building cross-sections of stock returns, *Working Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3493458.

Burgstahler, D.C., and I.D. Dichev, 1997, Earnings, adaptation and equity value, *Accounting Review* 72, 187–215.

Campbell, J.Y., and S.B. Thompson, 2008, Predicting excess stock returns out of sample: Can anything beat the historical average?, *Review of Financial Studies* 21, 1509–1531.

Campbell, J.Y., and M. Yogo, 2006, Efficient tests of stock return predictability, *Journal of Financial Economics* 81, 27–60.

Cao, C., O. Hansch, and X. Wang, 2008, The information content of an open limit-order book, *Journal of Futures Markets* 29, 16–41.

Cao, K., and H. You, 2021, Fundamental analysis via machine learning, *Working Paper*, https://repository.hkust.edu.hk/ir/bitstream/1783.1-112730/1/038604_1.pdf.

Cartea, A., S. Jaimungal, and Y. Wang, 2020, Spoofing and price manipulation in order driven markets, *Applied Mathematical Finance* 27, 67–98.

Chen, A.Y., and T. Zimmermann, 2021, Open source cross-sectional asset pricing, *Working Paper*, https://ssrn.com/abstract=3604626.

Cherian, J.A., and R.A. Jarrow, 1995, Market manipulation, In *Handbooks in operations research and management science*, Elsevier Academic Press.

Cochrane, J.H., 2011, Presidential address: Discount rates, *Journal of Finance* 66, 1047–1108.

Collins, D.W., M. Pincus, and H. Xie, 1999, Equity valuation and negative earnings: The role of book value of equity, *Accounting Review* 74, 29–61.

Comerton-Forde, C., and T.J. Putniņš, 2011, Measuring closing price manipulation, *Journal of Financial Intermediation* 20, 135–158.

Comerton-Forde, C., and T.J. Putniņš, 2014, Stock price manipulation: Prevalence and determinants, *Review of Finance* 18, 23–66.

Cong, L.W., X. Li, K. Tang, and Y. Yang, 2022, Cryto wash trading, *Working Paper*, https://www.nber.org/system/files/working_papers/w30783/w30783.pdf.

Cooper, D., and R. Glen, 1998, A strategic analysis of corners and squeezes, *Journal of Financial and Quantitative Analysis* 33, 117–137.

Dai, R., L. Donohue, Q. Drechsler, and W. Jiang, 2023, Dissemination, publication, and impact of finance research: When novelty meets conventionality, *Review of Finance* 27, 79–141.

Daniel, K., and Titman, S., 1999, Market efficiency in an irrational world, *Behavioral Finance* 55, 28–40.

Debie, P., C. Gardebroek, S. Hageboeck, P. Van Leeuwen, L. Moneta, A. Naumann, J.M.E. Pennings, A.A. Trujillo-Barrera, and M.E. Verhulst, 2023, Unravelling the JPMorgan spoofing case using particle physics visualization methods, *European Financial Management* 29, 288–326.

Dhawan, A., and T.J. Putniņš, 2022, A new wolf in town? Pump-and-dump manipulation in cryptocurrency markets, *Review of Finance* 27, 935–975.

Dong, X., Y. Li, D.E. Rapach, and G. Zhou, 2022, Anomalies and the expected market return, *Journal of Finance 77*, 639–681.

Egginton, J.F., B.F. Van Ness, and R.A. Van Ness, 2016, Quote stuffing, *Financial Management* 45, 583–608.

Eren, N., and H.N. Ozsoylev, 2006, Hype and dump manipulation, *Working Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=948814.

Falck, A., A. Rej, and D. Thesmar, 2022, When do systematic strategies decay?, *Quantitative Finance* 22, 1955–1969.

Fama, E. F., 1965, The behavior of stock-market prices, *Journal of Business* 38, 34–105.

Fama, E.F., 1970, Efficient capital markets: A review of theory and empirical work, *Journal of Finance* 25, 383–417.

Fama, E.F., and K.R. French, 2016, Dissecting anomalies with a five-factor model, *Review of Financial Studies* 29, 69–103.

Foster, G., C. Olsen, and T. Shevlin, 1984, Earnings releases, anomalies, and the behavior of security returns, *Accounting Review* 59, 574–603.

Friederich, S., and R. Payne, 2015, Order-to-trade ratios and market liquidity, *Journal of Banking and Finance* 50, 214–223.

Friedman, J.H., 2001, Greedy function approximation: A gradient boosting machine, *Annals of Statistics* 29, 1189–1232.

Graham, B., and D. Dodd, 1940, The scope and limits of security analysis. The concept of intrinsic value, In *Security Analysis: Principles and Techniques*, McGraw Hill.

Granger, C.W.J., 1992, Forecasting stock market prices: Lessons for forecasters, *International Journal of Forecasting* 8, 3–13.

Griffiths, M.D., B.F. Smith, D.A.S. Turnbull, and R.W. White, 2000, The cost and determinants of order aggressiveness, *Journal of Financial Economics* 56, 65–88.

Glosten, L.R., and P.R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.

Goldstein, M., A. Kwan, and R. Philip, 2022, High-frequency trading strategies, *Management Science* Articles in Advance, 1–22.

Grabowski, D., 2019, Technology, adaptation and the efficient market hypothesis, *Working Paper*, https://ssrn.com/abstract=3649446.

Green, J., J.R.M. Hand, and X.F. Zhang, 2017, The characteristics that provide independent information about average U.S. monthly stock returns, *Review of Financial Studies* 30, 4389–4436.

Gu, S., B. Kelly, and D. Xiu, 2020, Empirical asset pricing via machine learning, *Review of Financial Studies* 33, 2223–2273.

Hamilton, J.D., 2017, Why should you never use the Hodrick-Prescott filter, *Working Paper*, http://www.nber.org/papers/w23429.

Hand, J.R.M. and W.R. Landsman, 2005, The pricing of dividends in equity valuation, *Journal of Business Finance and Accounting* 32, 435–469.

Hanson, R., and R. Oprea, 2009, A manipulator can aid prediction market accuracy, *Economica* 76, 304–314.

Hoberg, G., and G. Phillips, 2016, Text-based network industries and endogenous product differentiation, *Journal of Political Economy* 124, 1423–1465.

Hou, K., C. Xue, and L. Zhang, 2020, Replicating Anomalies, *Review of Financial Studies* 33, 2019–2133.

Jarrow, R.A., 1992, Market manipulation, bubbles, corners, and short squeezes, *Journal of Financial and Quantitative Analysis* 27, 311–336.

Jegadeesh, N., and S. Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Jegadeesh, N., and S. Titman, 2001, Profitability of momentum strategies: An evaluation of alternative explanations, *Journal of Finance* 56, 699–720.

Jensen, M.C., 1978, Some anomalous evidence regarding market efficiency, *Journal of Financial Economics* 6, 95–101.

Karapandza, R., and J.M. Marin, 2014, The rate of market efficiency, *Working Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2024552.

Kaustia, M., and V. Rantala, 2021, Common analysts: Method for defining peer firms, *Journal of Financial and Quantitative Analysis* 56, 1505–1536.

Kelly, B.T., S. Pruitt, and Y. Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.

Khomyn, M., and T.J. Putniņš, 2021, Algos gone wild: What drives the extreme order cancellation rates in modern markets?, *Journal of Banking and Finance* 129, 1–16.

Knudsen, J.O., S. Kold, and T. Plenborg, 2017, Stick to the fundamentals and discover your peers, *Financial Analysts Journal* 73, 85–105.

Kostovetsky, L., and J.B. Warner, 2020, Measuring innovation and product differentiation: Evidence from mutual funds. *Journal of Finance* 75, 779–823.

Kozak, S., S. Nagel, and S. Santosh, 2020, Shrinking the cross section, *Journal of Financial Economics* 135, 271–292.

Lee, E.J., K.S. Eom, and K.S. Park, 2013, Microstructure-based manipulation: Strategic behavior and performance of spoofing traders, *Journal of Financial Markets* 16, 227–252.

Lev, B., and S.R. Thiagarajan, 1993, Fundamental information analysis, *Journal of Accounting Research* 31, 190–215.

Liu, J., D. Nissim, and J. Thomas, 2002, Equity valuation using multiples, *Journal of Accounting Research* 40, 135–172.

Lo, A.W, 2004, The adaptive markets hypothesis, *Journal of Portfolio Management* 30, 15–29.

Lo, A.W, 2012, Adaptive markets and the new world order, *Financial Analysts Journal* 68, 18–29.

Loh, R.K., and G.M. Mian, 2006, Do accurate earnings forecasts facilitate superior investment recommendations?, *Journal of Financial Economics* 80, 455–483.

Martin, I.W.R., and S. Nagel, 2022, Market efficiency in the age of big data, *Journal of Financial Economics* 145, 154–177.

Mclean, R.D., and J. Pontiff, 2016, Does academic research destroy stock return predictability?, *Journal of Finance* 71, 5–32.

Merrick, J.J., N.Y. Naik, and P.K. Yadav, 2005, Strategic trading behavior and price distortion in a manipulated market: Anatomy of a squeeze, *Journal of Financial Economics* 77, 171–218.

Nissim, D., and S.H. Penman, 2001, Ratio analysis and equity valuation: From research to practice, *Review of Accounting Studies* 6, 109–154.

Nyborg, K., and L. Mukhlynina, 2020, The choice of valuation techniques in practice: Education versus profession, *Critical Finance Review* 9, 201–265.

O'Hara, M., 2015, High frequency market microstructure, *Journal of Financial Economics* 116, 257–270.

Ohlson, J.A., 1995, Earnings, book values, and dividends in equity valuation, *Contemporary Accounting Research* 11, 661–687.

Ohlson, J.A., and B.E. Juettner-Nauroth, 2005, Expected EPS and EPS growth as determinants of value, *Review of Accounting Studies* 10, 349–365.

Ou, J.A., and S.H. Penman, 1989, Financial statement analysis and the prediction of stock returns, *Journal of Accounting and Economics* 11, 295–329.

Parlour, C.A., 1998, Price dynamics in limit order markets, *Review of Financial Studies* 11, 789–816.

Penman, S.H., 1998, Combining earnings and book value in equity valuation, *Contemporary Accounting Research* 15, 291–324.

Pennec, G.L., I. Fiedler, and L. Ante, 2021, Wash trading at cryptocurrency exchanges, *Finance Research Letters* 43, 1–7.

Piotroski, J.D., 2000, Value investing: The use of historical financial statement information to separate winners from losers, *Journal of Accounting Research* 38, 1–41.

Putniņš, T.J., 2020, An overview of market manipulation. In *Handbook of corruption and fraud in financial markets: Malpractice, misconduct and manipulation*, John Wiley & Sons Inc..

Redell, N., 2019, Shapley decomposition of r-squared in machine learning models, *Working Paper*, http://arxiv.org/abs/1908.09718.

Rösch, D.M., A. Subrahmanyam, and M.A. Van Dijk, 2017, The dynamics of market efficiency, *Review of Financial Studies* 30, 1151–1187.

Shiller, R.J., 2003, From efficient markets theory to behavioral finance, *Journal of Economic Perspectives* 17, 83–104.

Skinner, D.J., and R.G. Sloan, 2002, Earnings surprises, growth expectations, and stock returns or don't let an earnings torpedo sink your portfolio, *Review of Accounting Studies* 7, 289–312.

Stambaugh, R.F., and Y. Yuan, 2017, Mispricing factors, *Review of Financial Studies* 30, 1270–1315.

Takeuchi, L., and Y.-Y. Lee, 2013, Applying deep learning to enhance momentum trading strategies in stocks, *Working Paper*, https://www.semanticscholar.org/paper/Applying-Deep-Learning-to-Enhance-Momentum-Trading-Takeuchi/5991fee5265df4466627ebba62e545a242d9e22d.

Tobek, O., and M. Hronec, 2021, Does it pay to follow anomalies research? Machine learning approach with international evidence, *Journal of Financial Markets* 56, 1–31.

Wang, X., C. Hoang, Y. Vorobeychik, Y., and M.P. Wellman, 2021, Spoofing the limit order book: A strategic agent-based analysis, *Games* 12, 1–43.

Wang, Y.-Y., 2019, Strategic spoofing order trading by different types of ivestors in the futures markets, *Journal of Financial Studies* 27, 65–103.

Welch, I., and A. Goyal, 2008, A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies* 21, 1455–1508.

Williams, B., and A. Skrzypacz, 2021, Spoofing in equilibrium, *Working Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3742327.

Zhai, J., Y. Cao, and X. Ding, 2018, Data analytic approach for manipulation detection in stock market, *Review of Quantitative Finance and Accounting* 50, 897–932.