Tampere University

Joona Viitanen

PREDICTING STOCK PRICES WITH INVESTOR TRADING BEHAVIOUR: A MACHINE LEARNING APPROACH

Master of Science Thesis
Faculty of Management and Business
Examiners: Professor Juho Kanniainen
Dr. Anubha Goel
June 2023

ABSTRACT

Joona Viitanen: Predicting Stock Prices with Investor Trading Behaviour: A Machine Learning Approach
Master of Science Thesis
Tampere University
Master's Degree Programme in Industrial Engineering and Management
June 2023

Stock market prediction has always been considered a difficult task in both academia and industry. Markets have complex dynamics by nature and it is not always clear what drives asset prices. However, the rise of machine learning models has enabled the potential to capture patterns in data that were difficult to uncover using traditional methods. Recently, there has been lots of research about using machine learning methods for different applications in stock markets. However, in behavioural finance and specifically in microstructure literature, machine learning methods remain largely unexplored.

This thesis studies the relationship between investor trading behaviour and stock returns using machine learning methods. The data is from the Finnish stock market between 2000-2009 and the unique dataset was provided by Euroclear Finland Oy. The key questions are 1. How significant is the contemporaneous relation between trading behaviour and stock returns? 2. How significant is the effect of trading behaviour on future returns? and 3. How does the relationship change over different time horizons? The thesis studies 1-day, 5-day and 21-day horizons in both contemporaneous and lead-lag settings using different machine learning models. Logistic regression acted as a benchmark model and a previous period model as a naive model.

The results showed some degree of predictability in stock returns in both contemporary and lead-lag settings. The contemporaneous relationship was stronger as the models were able to beat the naive model by a wide margin. Lead-lag relationship was able to produce results above the naive model, but not by a significant margin. Furthermore, the predictability decreased when the time horizon increased in both settings. Finally, the predictability dropped in the 21-day lead-lag setting, as no model was able to beat the naive model. When it comes to the machine learning models, most of the models were able to beat benchmark logistic regression and a naive model in most configurations, suggesting nonlinear interactions in the system. From Ensemble-based methods, LightGBM, Random Forest and XGBoost performed the best, while AdaBoost struggled to beat logistic regression in other than 1-day horizons.

Keywords: stock market, behavioural finance, investor behaviour, machine learning, return predictability

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.
ChatGPT models have been used to help in structuring the thesis, rephrasing text and the creation of LaTex tables.

TIIVISTELMÄ

Joona Viitanen: Osakkeiden hintojen ennustaminen sijoittajakäyttäytymisen avulla käyttäen koneoppimismalleja
Diplomityö
Tampereen yliopisto
Tuotantotalouden diplomi-insinöörin tutkinto-ohjelma
Kesäkuu 2023

Osakemarkkinoiden ennustamista on aina pidetty vaikeana tehtävänä sekä tiedemaailmassa että teollisuudessa. Markkinoiden dynamiikka on luonteeltaan monimutkainen, eikä aina ole selvää, mikä ohjaa arvopapereiden hintoja. Koneoppimismallien nousu on kuitenkin mahdollistanut potentiaalin löytää datasta kaavoja, joita oli vaikea paljastaa perinteisillä menetelmillä. Viime aikoina on tehty paljon tutkimusta koneoppimismenetelmien käytöstä osakemarkkinoilla erilaisiin käyttötarkoitukseen. Behavioraalisessa rahoituksessa ja erityisesti mikrorakennekirjallisuudessa koneoppimismenetelmiä ei kuitenkaan ole käytetty merkittävästi.

Tämä opinnäytetyö tutkii sijoittajien kaupankäyntikäyttäytymisen ja osakkeiden tuottojen välistä suhdetta koneoppimismenetelmillä. Data on Suomen osakemarkkinoilta vuosilta 2000-2009 ja aineiston on toimittanut Euroclear Finland Oy. Keskeiset kysymykset ovat 1. Kuinka merkittävä kaupankäyntikäyttäytymisen ja osaketuottojen samanaikainen suhde on? 2. Kuinka merkittävä vaikutus kaupankäyntikäyttäytymisellä on tulevaisuuden tuottoon? ja 3. Miten vaikutus muuttuu eri aikahorisonteilla? Diplomityössä tutkitaan 1-, 5- ja 21-päivän horisontteja sekä samanaikaisissa että viive-asetuksissa käyttäen erilaisia koneoppimismalleja. Logistinen regressio toimii vertailumallina ja edellisen ajanjakson malli naiivina mallina.

Tulokset osoittivat jonkin verran ennustettavuutta osakkeiden tuotoissa sekä samanaikaisessa- että viive-asetuksissa. Samanaikainen vaikutus oli vahvempi, kun mallit pystyivät voittamaan naiivin mallin laajalla marginaalilla. Lead-lag -asetus pystyi tuottamaan tuloksia naiivin mallin yläpuolella, mutta ei merkittävällä marginaalilla. Lisäksi ennustettavuus heikkeni, kun aikahorisontti kasvoi molemmissa asetuksissa. Lopulta ennustettavuus putosi 21 päivän viive-asetuksella, koska mikään malli ei pystynyt päihittämään naiivia mallia. Koneoppimismalleista useimmat mallit pystyivät päihittämään logistisen regression ja naiivin mallin useimmissa asetuksissa, mikä viittaa epälineaarisiin vuorovaikutuksiin muuttujien välillä. Ensemble-pohjaisista menetelmistä LightGBM, Random Forest ja XGBoost suoriutuivat parhaiten, kun taas AdaBoost ei voittanut logistista regressiota muissa kuin 1 päivän horisontissa.

Avainsanat: osakemarkkinat, behavioraalinen rahoitus, sijoittajakäyttäytyminen, koneoppiminen, tuottojen ennustettavuus

PREFACE

This thesis is a result of a few months of intense research and a lot more months of not-so-intense research.

I would like to express my sincere thanks to Dr. Kęstutis Baltakys. His guidance was crucial when I was getting started with my research. He provided an interesting topic and offered insightful advice and support.

I'm also grateful to Prof. Juho Kanniainen, who offered valuable assistance as I was wrapping up my thesis. His feedback and encouragement were important during the final stages of my work.

Additionally, thanks to my friends for all the years in university. It has been truly an unforgettable journey.

Finally, I want to thank my family for all these years of support starting from the first grade.

Tampereella, 28th June 2023

Joona Viitanen

CONTENTS

1. INTRODUCTION

1.1 Research background

The financial markets have always been subject to extensive research and analysis in an attempt to understand the complex dynamics that drive asset prices. Investors, analysts, and researchers have sought to uncover patterns and relationships that could provide insights into market behavior and improve investment decision-making. Traditional approaches, such as fundamental analysis and technical indicators, have been widely utilized to assess the value and future performance of financial instruments. However, with the advancements in technology and the availability of vast amounts of data, there has been a growing interest in exploring alternative methods to predict market movements and generate superior returns. (Jiang, 2021)

Machine learning, a subfield of artificial intelligence, has emerged as a promising tool for forecasting financial markets. By leveraging sophisticated algorithms and statistical techniques, machine learning models have the potential to capture intricate patterns and relationships in data that may be difficult to discern through traditional methods. These models can process large datasets quickly and adapt to changing market conditions, offering the possibility of more accurate predictions and enhanced investment strategies. (Henrique et al., 2019; Jiang, 2021)

Financial asset prices exhibit non-linear, dynamic, and chaotic behaviour, making them challenging to predict as they represent financial time series. Among the most recent approaches, machine learning models have garnered significant attention due to their ability to identify intricate patterns across diverse applications. (Henrique et al., 2019)

One of the most sought questions in behavioural finance is how investor behaviour affects asset returns and it has been an important topic in the microstructure literature. Some studies have explored the contemporaneous relationship between trading behaviour and asset returns (Cai & Zheng, 2004; Kumar & Lee, 2006). Others have examined the lead-lag relationships, where trading behaviour is used to forecast future returns based on historical data (Barber et al., 2009a; Kaniel et al., 2008). However, most of the studies have been conducted using traditional statistical methods. Thus, leaving the potential for applying machine learning models in similar research settings.

1.2 The scope of research

This thesis aims to contribute to the existing literature by examining the predictive power of machine learning models in forecasting asset returns using investor trading behaviour. Specifically, the study focuses on the comovement and lead-lag relationships between investor trading and returns across different time horizons. By analyzing a comprehensive dataset and employing a range of machine learning algorithms, this research aims to assess the accuracy and robustness of these models and shed light on their potential practical implications.

The research will utilize a dataset encompassing relevant market data, including asset prices, trading volumes, and investor behaviour indicators. The dataset will cover a representative sample of assets from various asset classes to ensure the generalizability of the findings. Furthermore, the analysis will cover three different time horizons: one-day, five-day, and 21-day, allowing for a comprehensive examination of the models' performance over short and medium-term periods.

The research question is the following: To what extent investor trading behaviour affects stock returns? It can be further divided into three sub-questions:

1. How significant is the contemporaneous relation between trading behaviour and stock returns?
2. How significant is the effect of trading behaviour on future returns?
3. How does the relation change over different time horizons?

1.3 Structure overview

This thesis is structured as follows. Chapter 2 begins with an exploration of investor behaviour and its influence on stock markets. It covers the limitations of efficient markets theory, biases affecting investor decision-making, gaps in traditional finance, investor heterogeneity, the impact of different factors on investor behaviour, investor sentiment and its influence on stock markets. The literature review sets the foundation for the thesis.

Chapter 3 provides an overview of machine learning methods, with a focus on supervised learning algorithms in the context of financial markets. It explains the concept of machine learning and example application in various areas of finance. The chapter also introduces specific machine learning models used in the thesis, discussing their characteristics and applications. These models form the basis for the analysis and predictions in the research.

Chapter 4 focuses on the methodology used in the research. It explains the data processing techniques employed to calculate various metrics related to investor behaviour, such

as net scaled volumes, investor activity, and investor interaction. The chapter also discusses the determination of stock direction states and the selection of features and control variables for the analysis. It further describes the model setup, including the choice of machine learning algorithms and the evaluation of performance metrics. The chapter concludes with an overview of the hyperparameter tuning on model selection.

Chapter 5 presents the results of the study, focusing on the analysis of comovement and lead-lag effects across different time horizons. The main findings and performance of machine learning models are summarized in tables for each section.

Chapter 6 concludes the thesis by summarizing the key findings, discussing the theoretical and practical implications, addressing the limitations and quality assessment of the study, and providing proposals for future research.

2. INVESTOR BEHAVIOUR AND STOCK MARKETS

2.1 Behavioural finance

Thought 1970s and 1980s, the academic finance literature was largely based on efficient markets theory. The theory was based on the assumption that all investors are rational and make decisions based on all available information. However, during the 1980s, the academic discussion started to rise around anomalies that weren't explained by efficient market hypothesis. The most troubling anomaly was excess volatility: research seemed to confirm that stocks had more volatility than the efficient markets hypothesis could explain. The evidence was clear that aggregate market movements are dominated by substantial noise and researchers had to consider other theories. Finally, in the 1990s, the field of behavioural finance started to develop. It became an interdisciplinary field that combines insights from psychology, economics, and finance to study the impact of human behaviour on financial decision-making. (Shiller, 2003)

The fundamental principle that separates behavioural finance from traditional finance is that some agents in the market are not fully rational (Barberis & Thaler, 2002). Barberis and Thaler (2002) provided a comprehensive survey of the behavioural finance literature and identified several biases that affect investor decision-making, including loss aversion, overconfidence, and herding behaviour. Other biases are overextrapolation of past returns, the disposition effect, and limited attention (Barber et al., 2009b). These biases are especially important when studying trading activities of investors.

Subrahmanyam (2008) suggests that traditional finance has limited explanatory power when it comes to understanding certain aspects of investor trading behavior. It fails to address fundamental questions such as why individual investors engage in trading, how they perform in their investments, how they make portfolio choices, and why stock returns vary beyond what can be explained by risk factors. Furthermore, in the field of corporate finance, recent evidence indicates that actions like mergers and acquisitions and decisions regarding capital structure do not align with the rational behavior predicted by traditional theories. This misalignment presents a puzzling phenomenon that requires further explanation and understanding.

Subrahmanyam (2008) divides the literature of behavioural finance literature into three

parts:

1. Research about different patterns in the cross-section of average stock returns,
2. trading activity research and
3. corporate finance research.

This thesis is positioned on the trading activity literature, which will be the main focus of the literature review. Traditional finance primarily focuses on explaining asset prices and often overlooks the significance of trading activity. However, data from the NYSE website reveals a high annual share turnover rate of approximately 99% in 2003, which corresponds to a massive volume of around 350 billion shares. Based on reasonable estimates of per-trade costs, this suggests that the investing public voluntarily pays billions of dollars to financial intermediaries each year. Despite the importance of understanding this substantial trading volume, limited progress has been made by finance scholars in analyzing its origins. (Subrahmanyam, 2008) Fortunately, multiple pieces of research published between 2005-2016 have contributed significantly to our understanding of (individual) investor trading behaviour. The next section addresses how different investors' trading behaviour is affected by different factors.

2.2 Investor heterogeneity

Financial markets are intricate systems characterized by a multitude of factors contributing to their complexity, one of which is the high degree of heterogeneity among investors. Investors vary in numerous aspects, including their risk tolerance, investment size, regulatory restrictions, and access to information. To capture and understand these differences, investor categorization plays a crucial role. Large financial institutions, households, and governmental institutions exhibit distinct characteristics that influence their reactions to external factors, such as news, price changes, or volatility. (Lillo et al., 2015)

Extensive research has been conducted on how different types of investors respond to market information. The first example is reaction on news. Lillo et al. (2015) conducted a study investigating the impact of news on the trading behaviour of different investor categories. Their findings indicated that these categories exhibit varying reactions to factors such as returns, volatility, the number of news articles, and news sentiment.

Another example is the weighting of past returns. Grinblatt and Keloharju (2001) focused on understanding the trading behaviour of sophisticated and less sophisticated investors. They found that sophisticated investors, including non-financial corporations and finance and insurance institutions, tend to assign less importance to past returns compared to less sophisticated investors, such as households, general government, and nonprofit institutions. Furthermore, they observed that domestic investors, especially households, tend to be contrarian in nature, while foreign investors lean towards momentum investing.

Finally even gender may play a role in investor behaviour. Barber and Odean (2001) discovered that men tend to be more overconfident than women, which leads to higher levels of trading activity among men in common stocks. This observation suggests that even gender plays a role in investment decisions.

In conclusion, assigning investors to different categories is crucial when trying to understand trading behaviour. These variations in investor behaviour can be explained by a multitude of factors. The next section describes the dynamics that separate individual and institutional investors from each other.

2.3 Individual investor attributes

Especially before the time behavioural finance developed, academic discussion considered household investors as noise traders that destabilize the market and push prices from fundamentals (Barrot et al., 2016; Kaniel et al., 2008). Noise traders were considered unimportant as the fundamental idea was that arbitrageurs took advantage of the mispricing to drive prices back to their fundamental values (Fama, 1965). Thus, household investors could largely be considered irrelevant. However, recent advances in the field have recognized that not all households are noise traders.

The literature on individual investor trading behaviour has shown consistent evidence of contrarian tendencies among individuals in various markets. Grinblatt and Keloharju (2000, 2001) have documented short-horizon contrarian patterns, where individual investors tend to buy stocks after prices decline and sell after prices rise. On the other hand, there is evidence suggesting that institutions are high-frequency momentum traders in the short term but show contrarian behaviour over longer periods (Campbell et al., 2009). In other words, at least on shorter horizons, individuals and institutions may act completely oppositely. This is also in line with the concept that individuals may act as liquidity providers to institutions, which is discussed later in more detail.

While household investors may be generally less sophisticated than institutions, individuals are not trading under the same constraints (Barrot et al., 2016; Kelley & Tetlock, 2013). First, institutional investors have liquidity constraints (Coval & Stafford, 2007). The need to manage large portfolios and execute trades efficiently requires careful consideration of liquidity conditions and market impact. The second constraint is career concerns (Chevalier & Ellison, 1999). The pressure to achieve consistent performance and attract clients can drive specific investment choices and risk-taking behaviour among institutional investors. Finally, institutional investors may have agency problems. (Lakonishok et al., 1991). These problems stem from conflicts of interest, misaligned incentives, and complex organizational structures, which can potentially influence investment decisions and portfolio management practices. Recognizing these contrasting constraints and challenges emphasizes the importance of understanding and accounting for the heteroge-

neous nature of market participants. Individual investors may be able to take positions in the market that institutional investors are not able to take and vice versa.

Households may also have novel information acquired from the geographic proximity of the firm, relationships with employees, or customer preference insights (Kelley & Tetlock, 2013). Geographic proximity to a firm allows households to have firsthand exposure to local economic conditions and developments that may impact the firm's performance. This proximity can provide valuable insights into the firm's operations, market dynamics, and potential growth opportunities that may not be readily available to institutional investors operating at a distance. For instance, Ivkovic and Weisbenner (2005) found that in the US between 1991 to 1996, households had a strong preference for investing in local companies and they're generating additional returns relative to their nonlocal holdings while doing so. Therefore it is possible that local investors can exploit local knowledge. Additionally, the same logic can be applied to households' relationships with employees. Personal connections as friends or as customers can grant them access to non-public information about the firm's strategies, product pipelines, or upcoming events. Furthermore, their interactions with customers and the broader community enable them to gather customer preference insights, giving them a unique perspective on consumer behaviour and potential shifts in market demand. By leveraging these sources of information, households as retail investors have the potential to contribute distinctive perspectives and insights to the stock market, complementing the perspectives of institutional investors.

In recent years, substantial empirical evidence has emerged, shedding light on the vital role played by individuals in functioning as market makers and fulfilling the liquidity needs of other market participants. Several notable studies, such as those conducted by Kaniel et al. (2012), Kaniel et al. (2008) and Kelley and Tetlock (2013) have contributed to our understanding of this phenomenon. These studies have explored how individuals, particularly household investors, step in to provide liquidity when market conditions become constrained, often due to heightened demand from institutional investors.

The significance of household investors in maintaining market liquidity becomes particularly evident during periods of financial turbulence, such as the infamous 2008-2009 global financial crisis. Notably, retail traders who were actively engaged in the market during this crisis demonstrated an increased inclination towards holding stocks and actively providing liquidity to the French market, as highlighted by Barrot et al. (2016).

Moreover, Kelley and Tetlock (2013) delve into the specific strategies employed by households to facilitate liquidity provision. It is suggested that households predominantly utilize limit orders, a type of order specifying the maximum or minimum price at which they are willing to buy or sell a security, in order to contribute to overall market liquidity. However, when these household investors possess novel and valuable information, they tend to employ market orders, which involve the purchase or sale of a security at the best available

price in the market.

In conclusion, it is clear that household investors have the potential to influence stock prices and market dynamics beyond the notion of being noise traders. Their liquidity provision and ability to leverage unique information sources make them relevant and valuable contributors to the stock market. Their unique set of attributes separates them from institutions and research suggests that their behaviour is an important part of stock market dynamics The next section will delve into the concept of investor sentiment, which further explores how investor behaviour impacts the stock markets.

2.4 Investor sentiment and stock markets

Investor sentiment refers to the expectations about future cash flows and investment risks, which may not necessarily be supported by the existing factual data or evidence (DeLong et al., 1990). Based on efficient markets theory, investor sentiment should not affect realized and future stock returns (Fama, 1965). Later DeLong et al. (1990) discovered that noise traders acting simultaneously could create systematic risk, which is observed in prices. Essentially, this meant that investor sentiment had an unpredictable effect on the deviations from the fundamental values of stocks. If arbitrageurs tried to capitalize on the mispricing, they run a risk that especially in short-run increasing investor sentiment could push the prices even further from fundamental values. Thus, because of the arbitrageurs' risk aversion characteristics, the market couldn't completely eliminate sentiment-driven mispricing to set the security prices to equilibrium. (DeLong et al., 1990)

Investor sentiment has long been recognized as a crucial factor in understanding stock market dynamics. The seminal work of DeLong et al. (1990) presents a comprehensive model that highlights the importance of noise trader sentiment in asset pricing. Their model suggests that both the direction and magnitude of changes in sentiment play a vital role in influencing stock returns.

While prior empirical studies have primarily focused on examining the impact of sentiment on either the mean or variance of asset returns, it is argued that such an approach might be inadequate and incomplete (DeLong et al., 1990). To address this limitation, later research has adopted a GARCH (Generalized Autoregressive Conditional Heteroskedasticity) framework to simultaneously test the four behavioural effects proposed by the DSSW model.

In a study by Lee et al. (2002), a GARCH framework is employed to investigate the joint impact of noise trader sentiment on the formation of conditional volatility and expected returns. The findings indicate a negative correlation between shifts in sentiment and market volatility, implying that volatility increases when investors become more bearish and decreases when they become more bullish. This suggests that conventional measures

of risk fail to capture the influence of sentiment adequately. Moreover, the study reveals that sentiment has a more pronounced impact on the NASDAQ index compared to other indices examined, consistent with the notion that sentiment primarily affects smaller capitalization stocks.

Additionally, the study by Lee et al. (2002) uncovers a significant association between excess returns and shifts in sentiment. Higher excess returns are found to be linked to larger bullish shifts in sentiment and a subsequent decrease in conditional volatility for both small and large capitalization stocks. These results support the market reaction to noise trading as proposed by the DeLong et al. (1990) model, specifically highlighting the Friedman and create-space effects. The influence of noise trading on expected return is found to be through its impact on the market's perception of risk.

Additionally, the study indicates a positive correlation between sentiment changes and excess returns among the analyzed indices. This suggests that the increased risk premium, linked with the hold-more effect, surpasses the adverse influence of the price-pressure effect on anticipated returns. (Lee et al., 2002). These findings underscore that investor sentiment is not limited to individual investors or small capitalization stocks alone; rather, it exhibits relevance and implications across various market segments. Thus, making it important to consider all investor categories when measuring sentiment.

Baker and Wurgler (2006) challenge the classical view in finance theory that investor sentiment does not play a role in the cross-section of stock prices. They provide evidence that investor sentiment has significant cross-sectional effects on stock returns. The authors propose that sentiment-based demands and arbitrage constraints vary across stocks, resulting in cross-sectional patterns influenced by investor sentiment. They find that when sentiment is low, stocks that are attractive to optimists and speculators but unattractive to arbitrageurs, such as small stocks, young stocks, high volatility stocks, unprofitable stocks, non-dividend-paying stocks, extreme growth stocks, and distressed stocks, tend to have higher subsequent returns. Conversely, when sentiment is high, these stocks exhibit lower subsequent returns. (Baker & Wurgler, 2006)

The empirical findings reveal that several firm characteristics that lack unconditional predictive power demonstrate strong conditional patterns when sentiment is considered. The authors conduct various tests to examine the role of systematic risks and find that systematic risk alone does not provide a complete explanation for the observed patterns. These results challenge the classical view of the cross-section of stock prices and emphasize the importance of incorporating investor sentiment into models of prices and expected returns. (Baker & Wurgler, 2006)

Baker and Wurgler (2007) discusses a range of methods has been proposed to measure investor sentiment, covering various aspects of the investor's psychology and behaviours. These methods include surveys to directly capture investor beliefs, mood prox-

ies to assess emotions, analysis of retail investor trades, examination of mutual fund flows, consideration of trading volume as a sentiment indicator, evaluation of premia on dividend-paying stocks, assessment of closed-end fund discounts, analysis of option implied volatility, examination of first-day returns on initial public offerings (IPOs), scrutiny of the volume of IPOs and new equity issues, and exploration of insider trading. The following list by Baker and Wurgler (2007) gives short descriptions for different sentiment proxies.

1. **Investor Surveys:** Surveys conducted among investors to gauge their optimism or pessimism.
2. **Investor Mood:** Connecting stock prices to exogenous changes in human emotions, such as seasonal affective disorder or the mood associated with international soccer results.
3. **Retail Investor Trades:** Analyzing the trading behaviour of individual retail investors, who are more likely to be subject to sentiment.
4. **Mutual Fund Flows:** Examining how investors allocate their investments across different categories, indicating market sentiment.
5. **Trading Volume:** Viewing trading volume and liquidity as indicators of investor sentiment and optimism.
6. **Dividend Premium:** Measuring the difference in market-to-book-value ratios between dividend-paying stocks and non-payers, reflecting sentiment towards "safety."
7. **Closed-End Fund Discount:** Using the average discount on closed-end equity funds, held disproportionately by retail investors, as a sentiment index.
8. **Option Implied Volatility:** Monitoring the implied volatility of options, such as the VIX, as an indicator of investor fear or sentiment.
9. **IPO First-Day Returns:** Examining the remarkable returns on the first trading day of initial public offerings as an indication of investor enthusiasm and sentiment.
10. **IPO Volume:** Observing the fluctuation in the number of initial public offerings, which is often sensitive to changes in investor sentiment.
11. **Equity Issues Over Total New Issues:** Analyzing the equity share of total equity and debt issues by all corporations to gauge sentiment related to firms shifting between equity and debt.
12. **Insider Trading:** Monitoring the personal portfolio decisions of corporate executives, who may reveal their views on mispricing and market sentiment.

In this thesis, the focus is on analyzing (retail) investor trades, because of the extremely detailed data available. Several studies have provided evidence that micro-level trading data, specifically observing retail investors' simultaneous buying or selling actions,

aligns with systematic sentiment patterns (Baker & Wurgler, 2007). For example, a study by Hvidkjaer (2006) investigates trade imbalances to understand the trading behaviour behind momentum. Small-trade imbalances serve as a measure of small-investor sentiment, positively correlating with share turnover among losers. Losers with strong small-trade selling pressures outperform those with buying pressures in subsequent returns, particularly among high-volume stocks. The results indicate that small investors' trading behaviour contributes to momentum and can be used to predict future returns. (Hvidkjaer, 2006) The next section will discuss specifically how investor trading-based sentiment (net investor trading, trading imbalances) affects stock returns.

2.5 Investor trading and stock returns

The consensus regarding the relationship between individual trades and future stock returns isn't completely clear. For example, Odean (1999), Barber and Odean (2000), and Grinblatt and Keloharju (2000) find that stocks purchased by individuals underperform those they sell over longer time horizons, ranging from several months to two years. On the other hand, San (2005) finds positive excess returns in the two years following individual buying.

What seems clear is that individual investors as a group act systematically. Barber et al. (2009b) investigated the systematic trading behaviour of individual investors, also known as noise traders, and its potential impact on asset prices. The study finds strong evidence of systematic trading by individual investors within a month in the buying and selling decisions of stocks. These findings highlight that individual investors, despite their negligible influence as individual traders, have the potential to affect asset prices due to the systematic nature of their trading behaviour. The study suggests that this correlated trading is driven by shared psychological biases, such as the overextrapolation of past returns, the disposition effect, and limited attention. It concludes that individual investors, often considered noise traders, can have a cumulative effect on asset prices due to their systematic trading patterns (Barber et al., 2009b).

Kumar and Lee (2006) had similar findings. By analyzing retail trades from a significant sample of over 1.85 million buy and sell transactions from a large U.S. discount brokerage, they identified a systematic correlation in the trading behaviour of retail investors, indicating the existence of a common directional component. Thus, the existence of a common directional component suggests that retail investors influence stock prices as a group.

The findings of both studies provide support for the predictions of noise trader models and sentiment-based theories of returns comovement. Specifically, the systematic trading activities of retail investors can affect the returns of stocks in which they are concentrated (Barber et al., 2009b; Kumar & Lee, 2006). These results challenge the traditional as-

sumptions that individual irrationalities do not aggregate across a large group of investors and that rational arbitrageurs neutralize the effects of noise trading. It suggests that investor sentiment, particularly among retail investors, plays a significant role in financial markets.

In addition to insight into retail investors' systematic behaviour, Kumar and Lee (2006) provided empirical evidence linking retail investor sentiment to stock return comovements. They measured changes in retail investor sentiment based on these trades and demonstrated its incremental explanatory power for small stocks, value stocks, stocks with low institutional ownership, and lower-priced stocks. (Kumar & Lee, 2006) These results supported the hypothesis that retail investor sentiment seems to be linked to stock return comovements and that they are influencing stock returns beyond traditional risk factors and macroeconomic variables.

Investor trading sentiment effects to stock return comovements are not only limited to retail investors. Cai and Zheng (2004) studied the relationship between institutional trading and stock returns using data from US stock market. The authors address several questions regarding the impact of institutional trading on stock prices, the causality between institutional trading and returns, and the profitability of mimicking institutional trading. The results confirm a strong contemporaneous relation between institutional trading intensity and stock returns. Furthermore, stock returns show a negative relationship with lagged institutional trading, supporting the price-pressure hypothesis. The study also finds that major institutional selling predicts negative excess returns and major institutional buying positive excess returns. (Cai & Zheng, 2004) This adds to Kumar and Lee (2006) results by suggesting that in addition to individuals, institutional trading has contemporaneous relation with stock returns.

Campbell et al. (2009) results extend the work of Cai and Zheng (2004) to the prediction of future returns. Notably, institutional trading was found to predict short-term losses and long-term profits, suggesting liquidity demand. The researchers also discovered that earnings surprises and post-earnings announcement is anticipated by institutions. (Campbell et al., 2009)

In addition to institutions, individual investors also affect future returns. Kaniel et al. (2008) studied the interaction between individual investors and future stock returns. Using a unique dataset provided by the NYSE, the authors analyze the short-horizon individual investor excess selling and buying as well as subsequent returns using regressions. The findings indicate that individuals tend to buy after price decreases and sell after price increases. Additionally, it is observed that the trading actions of individuals can forecast future returns. Specifically, stocks tend to have positive surplus returns following periods of intense purchasing, and negative surplus returns after periods of intense selling by individuals. The authors suggest that individuals' contrarian behaviour positions them as

liquidity suppliers to immediate requiring institutions, which consequently leads to return reversals. The study also controls for factors such as past returns and trading volume, and it demonstrates that individual trading remains a significant predictor of future returns. (Kaniel et al., 2008)

Hvidkjaer (2008) found opposite results using small trade volume as a proxy for retail trading in US stock market. The study finds that intense sell-initiated small-trade volume stocks outperform intense buy-initiated small-trade volume stocks. This return difference persists for up to two years, particularly among small- and medium-sized firms. The results suggest that retail investor-favoured stocks underperform compared to stocks out of favor. The findings imply that retail investor-favoured stocks tend to be overvalued, which is why they underperfom long term. (Hvidkjaer, 2008)

Barber et al. (2009a) confirmed most of the findings above. They present four key findings. First, small trade order imbalances show a strong correlation with order imbalances based on trades from retail brokers, suggesting that small trades serve as reasonable proxies for individual investor trades (Barber et al., 2009a). This finding makes the results between papers mostly comparable regardless if different studies have used small trades as a proxy of individual investor trades. Second, individual investors exhibit herding behaviour, consistently buying or selling the same stocks as each other over consecutive time periods (Barber et al., 2009a). This finding reinforces the idea that individuals trade systematically as a group. Third, annual small trade order imbalances successfully forecast future returns in the following year. Fourth, stocks heavily bought one week earn strong returns the next week, and vice versa. (Barber et al., 2009a). The last two findings are important as they show that the future return predictability is observable over different time horizons.

Similar results about return predictability have also been found in other markets. Yang and Zhou (2015) conducted a study on the Chinese stock market, examining the combined effects of investor trading behaviour and sentiment on stock index returns. Specifically, the authors use transaction data from the Chinese stock market to measure the buy and sell imbalance of the stock index, capturing the aggregate market behaviour and explicitly removing the common dependence on the market factor. They found that both trading behaviour and sentiment have significant impacts on excess returns, with trading behaviour exerting a greater influence. (Yang & Zhou, 2015) The results highlight two important points: Firstly, the effects of trading on stock prices are persistent in other markets than the US market. Secondly, the results may be persistent on stock indexes, indicating that the effects are observable over the total market.

There are further studies revealing similar effects from Korea and Taiwan. Using daily data from the KOSPI and KOSDAQ, the study by Y. Kim and Lee (2022) revealed a significant relationship between investor sentiment and stock returns, with a stronger effect observed

in the KOSDAQ market characterized by high individual investor participation. Lien et al. (2020) analyzed the Taiwan Stock Exchange to understand how different investors' trades influence stock prices. They found that while individuals contributed more to current price movements, institutional investors showed better return predictability, suggesting that their trades were based on more sophisticated information. Large orders from professional institutions, in particular, were found to be closely linked to future stock performance (Lien et al., 2020).

There is also evidence that trading affects returns in intraday horizons. Narayan et al. (2015) conducted a study to examine the predictability of Chinese stock returns using order imbalances as a predictor. The study utilized intraday data from the Shanghai and Shenzhen stock exchanges and considered various trading frequencies (1-minute, 5-minute, 60-minute, and 90-minute). The findings indicated strong evidence of predictability at all trading frequencies, taking into account the persistence and endogeneity of order imbalance and cross-sectional dependence in stock returns. (Narayan et al., 2015) These findings complete the notion that trading affects future returns in all time horizons.

Kelley and Tetlock (2013) take the analysis even further by dividing net trading into (aggressive) market orders and (passive) limit orders. Both aggressive and passive net buying could positively predict a firm's monthly stock returns. Interestingly, only aggressive orders accurately forecasted firm news, such as earnings surprises, implying that these orders convey new cash flow information. Passive net buying, on the other hand, appeared to follow negative returns, suggesting that traders provide liquidity and benefit from the reversal of transient price movements. Limit order return predictability was strongest in stocks that often experienced significant return reversals. (Kelley & Tetlock, 2013)

Kaniel et al. (2012) discovered similar findings regarding the behavior of informed individual traders around announcements and their role in liquidity provision. The researchers observed evidence of informed trading by individual investors, demonstrating that aggregate buying (selling) by individuals predicted significant positive (negative) abnormal returns during and after earnings announcement dates. Kaniel et al. (2012) determined that approximately half of the returns could be attributed to private information and the other half to liquidity provision. This study, along with the work of Kelley and Tetlock (2013), further highlighted the potential information advantage of individuals and their ability to interpret public information. Additionally, some individuals also acted as liquidity providers (Kaniel et al., 2012; Kelley & Tetlock, 2013). To strengthen the findings, Boehmer et al. (2021) successfully replicated the results using more recent data and broader coverage.

Barrot et al. (2016) further studied the liquidy provision and found that retail investors provide liquidity to the stock market, particularly during times of market stress when liquidity is scarce. The authors revealed that a portfolio based on retail investors' trading behaviour generated annualized excess returns of 19% from 2002 to 2010 and up to 40%

during periods of high uncertainty. However, individual retail investors did not benefit from these potential returns due to two main reasons: they often made trades at unfavorable prices on the day of trading and they were too slow in reversing their trades, which caused the liquidity premium to dissipate. Experienced traders were less susceptible to these pitfalls and managed to capture a larger share of the returns. (Barrot et al., 2016) The study emphasizes the notable impact that retail investors' trading behaviour can have on stock returns, especially during periods of low market liquidity,

The study conducted by Z. Chen et al. (2014) in Taiwan's emerging market aligns with the findings of Kaniel et al. (2012) in the United States, indicating that pre-event individual trading positively predicts and post-event returns. However, there is a notable distinction between the two markets regarding the drivers of this relationship. While Kaniel et al. (2012) suggest that the positive relationship in the US can be attributed to both the information advantage and liquidity provisions by individual investors, the study by Z. Chen et al. (2014) reveals that in Taiwan, the positive relationship is primarily explained by liquidity provisions, without a significant presence of an information advantage possessed by individual investors. These results highlight the fact that different markets may have their market-specific dynamics.

Stoffman (2014) examined trading data in Finland over 15 years to explore the relationship between price changes and trading activities of individuals and financial institutions. The author found three main outcomes. Firstly, the study revealed that price movements were particularly associated with the trading demands of institutions and not of households. Specifically, prices tended to rise when institutions bought from households and fell when institutions sold to households. This indicates that institutions largely drive price changes, with households mainly supplying liquidity. Secondly, there was no consistent pattern in price changes when trading occurred solely between individuals or solely between institutions. Lastly, the author found that when prices did change as a result of trading among households, these changes were typically short-lived and followed by reversals. In line with other studies (eg. Kelley and Tetloc, 2013), this suggests that while individual traders may influence prices in the short term, they do not create substantial price distortions in the long term, with institutions playing a more prominent role in driving prices toward fundamental values (Stoffman, 2014). The study highlights the fact that in addition to trading imbalances of different investor groups, trading between groups may also be important driver of future returns.

Ülkü and Weber (2013) discovered significant bi-directional interactions between trading flows and stock returns within the same trading day. This indicates that not only do trading flows influence returns, but returns can also impact trading flows. For instance, individual investors were found to engage in intraday positive feedback trading, buying stocks as prices rise and selling as they fall, potentially amplifying price movements. Institutional investors, including banks and pension funds, adopt contrarian strategies to

minimize their price impact, contributing to market efficiency and stability. (Ülkü & Weber, 2013) The study underlines the problem especially when analyzing comovement of trading and returns. As trading flows and returns have bi-directional interactions, it is difficult to determine which one is driving the other.

To conclude, the research conducted by Qian (2014) provides an insightful perspective to the understanding of the dynamics between retail investors and stock market behaviors. The study distinguishes the processes proposed by earlier research and posits a theory suggesting that sentimental retail investors may not drive prices away from fundamental values through their trades. Instead, they might interfere with the price discovery mechanism for mispriced stocks. This delayed reaction to crucial market information and the associated sluggish price discovery may be factors contributing to the negative correlation between small trade imbalances and future returns. (Qian, 2014)

A key finding of the study by Qian (2014) is that small trade imbalances are negatively correlated with future stock returns only when a substantial difference of opinion exists among investors. This suggests that initial mispricing triggers this correlation. However, for stocks where systematic mispricing is absent, the correlation between small trade imbalances and future returns becomes insignificant. In other words, stocks characterized by low opinion divergence and high small trade imbalances do not display significant overvaluation. Qian (2014).

Similar dynamics are observed in the context of earnings announcements. The study further analyzes the cumulative abnormal returns (CARs) during these periods and discovers that stocks exhibiting high opinion divergence and high small trade imbalances show negative CARs. This suggests that such stocks could be persistently overvalued due to buying pressure from retail investors and their sentiments. (Qian, 2014) The study highlights the importance of opinion divergence in addition to trading imbalances. Therefore, it becomes apparent that a difference of opinion is not just another factor but an essential variable that needs careful consideration when predicting stock returns.

2.6 Synthesis of the theory

There are a few main takes in the literature, which are listed below.

1. Individual investors trade systematically as a group, which has an effect on asset prices
2. Institutional investors' trades affect asset prices
3. Trading predicts contemporary and future stock returns over different time horizons
4. Some individual investors may trade on novel information, while some act as liquidity providers.

5. In addition to order imbalances, trading between investor groups and opinion divergence inside investor groups are important drivers of stock returns.

Main points of the most important research are outlined in table 2.1.

Table 2.1. *The most relevant research*

Article	**Country**	**Main Finding**
Barber et al. (2009b)	US	Individual investors exhibit systematic trading patterns driven by shared psychological biases and these systematic trades cumulatively have the potential to affect asset prices.
Kumar & Lee (2006)	US	Retail investors' trading behaviour has a common directional component, indicating that they influence stock prices as a group.
Cai & Zheng (2004)	US	Strong contemporaneous relation between institutional trading intensity and stock returns.
Campbell et al. (2009)	US	Institutional trading predict short-term losses and long-term profits, suggesting liquidity demand.
Kaniel et al. (2008)	US	Retail order imbalances positively predict the following month's returns on stocks.
Hvidkjaer (2008)	US	Several months lasting order imbalances predict returns at horizons of 1 to 24 months.
Barber et al. (2009a)	US	Trading imbalances predict returns in weekly and yearly horizons.
Yang & Zhou (2015)	China	Both trading behaviour and sentiment significantly impact stock index returns, with trading behaviour exerting a greater influence.
Kim & Lee (2022)	Korea	The study revealed a significant relationship between investor sentiment and stock returns in the Korean stock markets.
Lien et al. (2020)	Taiwan	Individuals contributed more to current price movements, while institutional investors showed better return predictability.
Narayan et al. (2015)	China	Order imbalances can predict stock returns at various intraday trading frequencies.

Table 2.1. *The most relevant research (Continued)*

Article	**Country**	**Main Finding**
Kelley & Tetloc (2013)	US	Aggressive orders forecast firm news while passive net buying tends to follow negative returns, indicating liquidity provision.
Kaniel et al. (2012)	US	Individual investors' aggregate buying and selling predicted large positive and negative abnormal returns around earnings announcements. About half of the returns were attributed to private information.
Barrot et al. (2016)	US	Retail investors provide liquidity, especially during market stress.
Z. Chen et al. (2014)	Taiwan	Pre-event individual trading predicts post-event returns, which is primarily explained by liquidity provision.
Stoffman (2014)	Finland	Institutions largely drive price changes, with households mainly providing liquidity.
Ülkü & Weber (2013)	Korea	There are significant bi-directional interactions between trading flows and stock returns within the same trading day.
Qian (2014)	US	Future returns are negatively related to small trade imbalances when there is a large difference of opinion among investors.

Based on the above-mentioned findings, there is a clear indication that investor trading behaviour has an effect on stock prices. However, there are few methodological gaps that could be filled. First, most of the previous research have focused on the explanatory power of a single group of investors. As made evident in the literature, different groups have different behavioural characteristics and therefore affect differently on stock prices. Thus, introducing multiple different groups to analysis should have more explanatory power compared to, for example, simply using small trades.

Second, different factors and time horizons could be combined into the same study. Previous research has mostly studied certain time horizons or factors, which allows focusing on some effects but possibly fails to capture the aggregate explanatory power of multiple trading behaviour-proxying variables. In addition to trading imbalances used in most of the most relevant studies, findings about the relevance of trading between investors (Stoffman, 2014) or disagreement inside investor groups (Qian, 2014) could be included in the same models to reflect trading behaviour as accurately as possible.

Finally, now that we have a large set of possible trading behaviour variables for different

investor groups over different time horizons, it would be plausible to try more complex models. Most of the previous research about trading behaviour and asset returns has been conducted using different statistical econometrics models. While the models are good for explaining the interactions between different variables, they may lack outright predictive power especially when there are features with complex nonlinear interactions (Mullainathan & Spiess, 2017). Therefore, by using a broad set of trading behaviour explaining features over different time horizons, a machine learning classification-based approach could be the key to beating traditional models in predictability.

3. MACHINE LEARNING METHODS

3.1 Machine learning overview

Machine learning is a subset of artificial intelligence that involves the use of statistical algorithms to enable machines to improve their performance on a specific task with experience (Jordan & Mitchell, 2015) It is a data-driven approach that focuses on building models from data, and these models can be used to make predictions or decisions about new data. There are three main machine learning algorithms types: supervised, unsupervised, and reinforcement. In this thesis, the focus is on supervised learning.

Supervised learning involves training a model on labelled data, where the desired output is known, in order to make predictions on new, unseen data (Goodfellow et al., 2016). The output of the model can be continuous value or labelled value. For example, in binary classification, the output varies between two labels, whereas in multiclass classification, there are multiple possible outcomes (Jordan & Mitchell, 2015). Regressive algorithms are type of supervised learning, where the output takes continuous values (Shalev-Shwartz & Ben-David, 2014). In this thesis, the modelled output is three-class classification.

In the majority of machine learning applications, the data at hand is divided into three distinct sets: the training set, the validation set, and the test set. The training set is utilized to adjust the algorithm's parameters, the validation set is employed to assess the model's performance by utilizing an external dataset while fine-tuning the hyperparameters. Finally, the test set ensures an unbiased evaluation of the actual performance of the final model using unseen data. (Shalev-Shwartz & Ben-David, 2014)

3.2 Supervised learning in financial markets

In recent years, machine learning (ML) has gained significant attention and popularity in various industries, including finance. ML algorithms are capable of extracting patterns from large datasets and providing predictions with high accuracy. The use of ML in financial markets has been gaining momentum due to the increasing amount of data available, the need for automation and speed, and the desire for improved decision-making. (Jiang, 2021)

Supervised learning is a powerful machine learning technique that has been applied to various areas of finance, including stock market prediction (Henrique et al., 2019), credit risk assessment (Shi et al., 2022) and fraud detection (Hilal et al., 2022). This thesis draws inpiration from the stock market prediction literature.

One popular application of supervised learning in financial markets is predicting stock prices. Researchers have used various supervised learning techniques to forecast stock prices, including regression models (Pai & Lin, 2005), decision trees (Krauss et al., 2017), and neural networks (K.-j. Kim & Han, 2000). These models have been shown to outperform traditional time-series models. The advantage of supervised learning in all fields is that it allows for the development of highly accurate models that can be used to make data-driven decisions (Henrique et al., 2019). However, as Arrieta et al. (2020) note, these models can be complex and may be difficult to interpret, which can make it challenging to understand how the model is making its predictions.

In conclusion, there are several supervised learning algorithms used in financial markets literature, including regression models, decision trees, random forests, and neural networks. Each of these algorithms has its strengths and weaknesses, and the choice of algorithm depends on the specific problem at hand. The next sections describe the machine learning models used in this thesis.

3.3 Logistic regression

Logistic regression is one of the most widely used traditional statistical algorithms applied for prediction and interference. The model is especially effective for solving relatively less complex problems. In the context of logistic regression models, the dependent variable is consistently expressed in categorical form, comprising two or more levels. Independent variables, on the other hand, can exist in either numerical or categorical form. Let p denote the conditional probability associated with the first class. The relationship between the probability p of the dichotomous outcome event and a set of explanatory variables $\mathbf{x}$ can be represented as follows (David W. Hosmer, 2000):

$$Logit(p) = \ln\left(\frac{p}{1-p}\right) = \mathbf{b}^T\mathbf{x} \tag{3.1}$$

where $\mathbf{b} = (b_0, b_1, b_2, ..., b_k)$ is the vector of the coefficients of the model and $\mathbf{b}^T$ the transpose vector. We refer to $\frac{p}{1-p}$ as the odds ratio and to the expression (1) as the log-odds or logit transformation.

Consider a training dataset denoted as $D = (x_l, y_l) : l = 1, 2, ..., n_T$, where n_T represents the number of samples. The model assumes that the training sample is a realization of a set of independent and identically distributed random variables. The regression

coefficients b_i, which are unknown and need to be estimated from the data, hold a direct interpretation as log-odds ratios or, in terms of $\exp(b_i)$, as odds ratios. The log-likelihood for n_T observations can be expressed as follows:

$$l(b) = \sum_{l=1}^{nT} [y_l * b^T * x_l - log(1 + e^{b^T * x_l})] \tag{3.2}$$

The log-likelihood function is used for estimating regression coefficients b_i in the model. Coefficients are obtained by iterative methods. The exponential value of regression coefficients ($\exp(\mathbf{b}^T)$) gives the odds ratio, and this value reflects the effect of the risk factor in the disease, and the interpreted values are odds ratios. In addition, after the model is obtained, a class can be predicted based on the estimated coefficients and new observations.

In this thesis, the goal of the logistic regression model is twofold. First, it provides a simple statistical method to study how different input features affect the outcome of the model. Second, it provides an intelligent baseline for more complex machine learning algorithms, which will be discussed in the next sections.

3.4 Random forest

Random Forest is another widely used algorithm for prediction and classification tasks. It is an ensemble learning method that combines multiple decision trees to make predictions. Random Forest aggregates the predictions of several decision trees to improve the accuracy and control overfitting. It was introduced by Breiman (2001) and is particularly effective for datasets with a large number of features or complex interactions.

A Random Forest consists of a set of decision trees. Each tree $T_k, k \in \{1, \ldots, K\}$ is constructed using a bootstrapped sample of the training dataset. Furthermore, at each node, a random subset of features is considered for splitting. Formally, given a training dataset $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where X_i is a feature vector and Y_i is the target variable, a decision tree maps feature vectors to predictions through recursive binary splits.

The quality of a split at a node is often measured by a criterion like Gini impurity or information gain, defined as:

$$\text{Information Gain} = H(D_p) - \sum_{i \in \{\text{left,right}\}} \frac{|D_i|}{|D_p|} H(D_i), \tag{3.3}$$

where H is the entropy function, D_p is the dataset at the parent node, and D_{left} and D_{right} are the datasets for the left and right children after the split.

Once the decision trees are built, predictions are made by inputting a feature vector through each of the trees and aggregating their outputs. For classification tasks, the most common aggregation method is majority voting. Given an input X, the predicted class $\hat{Y}$ is:

$$\hat{Y} = \arg\max_{c \in C} \sum_{k=1}^{K} I(T_k(X) = c), \tag{3.4}$$

where C is the set of classes, $T_k(X)$ is the class predicted by the k-th tree for input X, and $I(\cdot)$ is the indicator function.

The performance of Random Forest depends on the number of trees in the forest and the number of features considered at each split. Increasing the number of trees can improve accuracy, but at the cost of increased computational complexity. Similarly, considering more features at each split can increase diversity but may also lead to higher computational requirements. (Breiman, 2001)

Random Forest offers several advantages. It can handle large datasets with high dimensionality and a mix of categorical and numerical features. It is less prone to overfitting compared to individual decision trees. Moreover, it can provide estimates of feature importance, indicating which features are more influential in making predictions. (Breiman, 2001) Finally, it has been successfully implemented in stock price direction prediction (Ballings et al., 2015)

3.5 Adaboost

A classifier system constructs a model that can predict the category of a new observation given a dataset. The effectiveness of this classification depends on the quality of the method utilized, as well as the complexity of the specific task at hand. If the classification system exhibits higher accuracy compared to a baseline, it suggests that the method has successfully extracted relevant patterns from the data. (Alfaro et al., 2008)

One approach to enhancing the accuracy of classification systems is AdaBoost, introduced by Freund and Schapire. This method iteratively applies a classification system to the training data, wherein each iteration focuses on different examples by adjusting the weights assigned to them. Unlike other ensemble methods such as Bagging, AdaBoost dynamically updates these weights. At the end of training, the individual classifiers are aggregated into a composite classifier, which typically demonstrates superior accuracy on the test data. (Alfaro et al., 2008)

Though multiple variations of the AdaBoost algorithm exist, the most prevalent version is AdaBoost by Freund and Schapire. For simplicity, let's consider a binary classification problem with a training dataset, $T_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, where $Y_i \in \{-1, 1\}$. Each observation X_i is assigned a weight $\omega_b(i)$, initialized to $\frac{1}{n}$. This weight is

updated after each iteration.

A base classifier, denoted as $C_b(X_i)$, is constructed on the training set T_b, and is applied to each training sample. The error of this classifier, ε_b, is given by:

$$\varepsilon_b = \sum_{i=1}^{n} \omega_b(i) n_b(i), \quad \text{where} \quad n_b(i) = \begin{cases} 0, & \text{if } C_b(X_i) = Y_i, \\ 1, & \text{if } C_b(X_i) \neq Y_i. \end{cases} \tag{3.5}$$

The weight for the next iteration, $\omega_{b+1}(i)$, is updated as follows:

$$\omega_{b+1}(i) = \omega_b(i) \exp(\alpha_b n_b(i)), \tag{3.6}$$

where α_b is a constant derived from the error:

$$\alpha_b = \ln\left(\frac{1-\varepsilon_b}{\varepsilon_b}\right). \tag{3.7}$$

The weights are normalized so that they sum to one. As such, $\varepsilon_b = 0.5 - \gamma_b$, where γ_b represents the margin by which the base classifier surpasses the baseline. Consequently, the weights of misclassified observations are increased, whereas the weights of correctly classified instances are reduced. The variable α acts as a learning rate, controlling the influence of each classifier based on its error.

This process is repeated for $b = 1, 2, 3, \ldots, B$. The final ensemble classifier, $C(x)$, is constructed as a linear combination of the base classifiers, weighted by the corresponding α_b:

$$C(x) = \text{sign}\left(\sum_{b=1}^{B} \alpha_b C_b(x)\right). \tag{3.8}$$

By iteratively adjusting the weights of the training instances, AdaBoost focuses on difficult examples and combines the weak classifiers to create a strong ensemble classifier that achieves high accuracy in classifying new observations. (Alfaro et al., 2008)

3.6 XGBoost

XGBoost, which stands for eXtreme Gradient Boosting, is a highly efficient and scalable machine learning algorithm particularly well-suited for tree boosting (T. Chen & Guestrin, 2016). The algorithm employs an ensemble of decision trees and optimizes a regularized objective function.

XGBoost employs an ensemble model consisting of K additive functions. Given a dataset with n examples and m features, denoted as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, where $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$,

the prediction $\hat{y}_i$ for each example is the sum of the outputs of the K functions:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \tag{3.9}$$

where $f_k(\mathbf{x})$ belongs to the function space $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}$, where $q : \mathbb{R}^m \to T$ maps features to the corresponding leaf index and $w \in \mathbb{R}^T$ are the leaf weights. Each function f_k corresponds to a decision tree with structure q and leaf weights w.

The goal of XGBoost is to learn the set of functions that minimize the regularized objective:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3.10}$$

where l is a differentiable loss function that measures the difference between the predicted and actual target, and Ω is a regularization term defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{3.11}$$

where γ and λ are regularization constants.

XGBoost employs an iterative algorithm. At the t-th iteration, a function f_t is added to minimize the objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t). \tag{3.12}$$

Using a second-order approximation to the loss function and removing constants, the optimal weight w_j^* of leaf j for a fixed structure $q(\mathbf{x})$ is:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{3.13}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$. The simplified objective becomes:

$$\mathcal{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \left(\sum_{i \in I_j} g_i \right)^2 \frac{1}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{3.14}$$

Since it is computationally infeasible to enumerate all possible tree structures, XGBoost utilizes a greedy algorithm that iteratively adds branches. The algorithm selects splits by either enumerating all possible splits using an exact greedy algorithm or employing a more computationally efficient approximate algorithm that considers only a subset of the possible splits (T. Chen & Guestrin, 2016).

3.7 LightGBM

LightGBM is an efficient gradient boosting algorithm that is designed for speed and performance. Developed by Microsoft, LightGBM is especially effective when dealing with large datasets. Unlike other tree-based algorithms, LightGBM grows trees leaf-wise which can result in a more accurate model. (Ke et al., 2017)

Traditional gradient boosting algorithms grow trees level-wise, where all nodes of a particular depth are expanded before any nodes of the next depth are considered. In contrast, LightGBM employs a leaf-wise algorithm. At each split, it selects the leaf with the maximum delta loss and splits it, rather than splitting all leaves on a level. This tends to grow more asymmetric trees, but enables the algorithm to achieve lower loss more quickly and often requires fewer boosting rounds.

To speed up the training process, LightGBM incorporates a sampling technique called Gradient-based One-Side Sampling (GOSS). GOSS keeps all instances with large gradients (i.e., those that are under-estimated by the current model) and performs random sampling on the instances with small gradients. This allows LightGBM to focus on the difficult examples, without ignoring a small subset of easy examples, which makes it more efficient without significant loss in accuracy.

LightGBM is designed to handle high-dimensional data with many sparse features efficiently. Through Exclusive Feature Bundling (EFB), it bundles mutually exclusive features (i.e., features that are never non-zero together) into a single feature. This reduces the number of features considerably, especially in sparse datasets, without losing any information.

LightGBM aims to minimize the following objective function over n training instances:

$$\mathcal{L}(\mathbf{p}) = \sum_{i=1}^{n} l(y_i, p_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3.15}$$

where l is the loss function measuring the difference between the predicted and actual target, y_i and p_i are the actual and predicted values respectively, f_k are the boosting functions, and Ω is a regularization term.

Like other gradient boosting algorithms, LightGBM constructs an ensemble of decision trees, by iteratively constructing decision trees, where each tree is trained to correct the errors of the preceding trees.

LightGBM uses the gradients and Hessian information to find the best split points during the training. It employs histogram-based techniques for faster training. Instead of finding the best split points for each feature, LightGBM buckets continuous feature values into discrete bins, which significantly speeds up the training process. (Ke et al., 2017)

4. RESEARCH METHODOLOGY

4.1 Research Design

The objective of this research is to examine how the trading activity of different investor groups can predict movements in individual stocks. The study utilizes various machine learning algorithms to make predictions and compares their performance using multiple evaluation metrics. By combining existing literature on investor sentiment and machine learning-based price predictions, this research aims to contribute to the understanding of stock market dynamics.

The research process is structured as follows:

1. Presentation of the data used in the research.
2. Description of the data processing methods and selected features.
3. Description of the training process of the algorithms and hyperparameter search
4. Presentation of the results of different models based on the chosen evaluation metrics.

4.2 Data Description

The data for this study is provided by Euroclear Finland Oy (previously Finnish Central Security Depository) and consists of all trades conducted on the Finnish stock market from 1995 to 2009. It includes transactional data of individual traders, but the exact timestamps of the trades are not available as the trades are settled at the end of the day. The data contains information about the trader (anonymized owner ID, year of birth, gender, and sector code) and the stock being traded (stock ID, volume traded, transaction type, trading price, trading date, and registration date). Table 4.1 provides a sample of the data format.

The selected time frame for the study spans from 2000 to late 2009, covering two stock market crashes and the corresponding post-crash and pre-crash periods. Further, the testing training split is done at the start of 2007, leaving the financial crisis period for testing. The chosen stocks are mid- to large-cap stocks that were actively traded during this period. The specific stocks chosen for the analysis are listed in Table 4.2.

Table 4.1. *Data Format*

Owner ID	Sector Code	Holding Type	ISIN	Volume	Transaction Type	Trading Date	Price
1001	100	1	FI0009502101	500	10	10.14.2006	8.60
1001	100	1	FI0009502103	350	20	2.4.2001	9.50
5001	200	1	FI0009501234	300	10	7.7.2004	1.30

Table 4.2. *Chosen Stocks*

Stock	ISIN	Trading Days	Days of No Trading	Days of No Returns	Days of Positive Returns	Days of Negative Returns	Mean Monthly Volatility
Nokia	FI0009000681	2452	3	33	1195	1224	0.449
Fortum	FI0009007132	2452	0	114	1217	1121	0.291
UPM	FI0009005987	2452	3	58	1189	1205	0.327
Outokumpu	FI0009002422	2451	2	105	1188	1158	0.384
Sampo	FI0009003305	2452	2	77	1220	1155	0.315
Elisa	FI0009007884	2452	0	68	1206	1178	0.419
NokianRenkaat	FI0009005318	2440	6	148	1175	1117	0.381
Konecranes	FI0009005870	2450	4	143	1177	1130	0.369
StoraEnso	FI0009005961	2406	30	143	1107	1156	0.403
YIT	FI0009800643	2407	2	192	1141	1074	0.327

The data exhibits heterogeneity due to various factors. The investors themselves differ in terms of investor status (e.g., institutional vs. household), investment volumes, investment time horizons, trading strategies, and information sources. The traded stocks also introduce heterogeneity with variations in trading volumes and market environments across different sectors. Consequently, significant data processing is required to normalize these heterogeneous factors.

4.3 Data Processing

The primary objectives of data processing are to provide proxies for investor behavior and to normalize the data by reducing heterogeneity among different stocks and investors. The data processing steps are as follows:

1. Grouping investors based on their status.
2. Scaling the trading volumes to calculate net scaled volumes, which serve as proxies for investor group sentiment.
3. Calculating the proportion of active investors and the share of total volume traded for each group to represent investor group activity.
4. Estimating the trading occurring between investor groups and households as a proxy for investor interaction.
5. Defining the output feature

4.3.1 Investor Grouping

Investor status is the largest source of heterogeneity, encompassing various behavioral differences. In this study, investors are categorized into groups based on the available data, following similar approaches found in prior research (e.g., Lillo et al. (2015), Grinblatt and Keloharju (2001), Tumminello et al. (2012)). The resulting investor categories are shown in Table 4.3.

Table 4.3. *Investor categories*

Investor Category	Total Trades	Mean Trade Value (€)	Median Trade Value (€)	% of Unique Investors
Government Institutions	204,635	491,332	27,575	241
Non-Financial Corporations	3,265,937	54,519	13,584	23,876
Households	8,535,866	11,499	4,478	412,851
Financial-Insurance Corporations	10,634,091	77,945	19,095	671
Non-Profit Institutions	164,962	99,680	10,147	2,717
Foreign	2,111,706	95,193	12,307	3,506

Although alternative approaches, such as semi- or unsupervised clustering algorithms, could be used for investor categorization based on trading behaviour (Challet et al. (2016)), this study opts for interpretable results by employing predefined categories based on available data. Moreover, this choice ensures consistency over time, as dynamically determined categories would introduce further complexity.

4.3.2 Investor Sentiment

To address the heterogeneity arising from investment volumes, the trading volume of each individual investor is scaled to a net scaled volume, which serves as a proxy for investor sentiment. Two methods are used to calculate net scaled volumes.

The first method is the aggregate net scaled volume, defined as:

$$r(g,t) = \frac{V_b(g,t) - V_s(g,t)}{V_b(g,t) + V_s(g,t)}, \tag{4.1}$$

where $V_b(g,t)$ represents the total volume of stocks bought by group g on day t, and $V_s(g,t)$ represents the total volume of stocks sold by group g on day t. The net scaled volume measures the difference between buying and selling volumes and normalizes it by the total volume traded. This scaling ignores the absolute volume traded, effectively reducing the heterogeneity resulting from investment volumes. The net scaled volumes are aggregated at the group level, giving more weight to traders within the group who have traded more.

The second method is the averaged net scaled volume, which calculates net scaled volumes for individual traders as follows:

$$r(i,t) = \frac{V_b(i,t) - V_s(i,t)}{V_b(i,t) + V_s(i,t)}, \tag{4.2}$$

where $V_b(i,t)$ and $V_s(i,t)$ represent the volumes of stocks bought and sold by investor i on day t, respectively. The net scaled volume is not calculated for investors who are not active on a given day. After scaling the daily activities of each investor, the net scaled volumes are averaged daily within investor groups:

$$r(g,t) = \frac{1}{N} \sum_{i \in g} r(i,t), \tag{4.3}$$

where N is the total number of investors in group g who traded on day t. This averaging approach treats all investors equally within a group, regardless of the total volume traded by each investor. By aggregating net scaled volumes in this manner, the scaling emphasizes the activity direction of each individual trader.

Additionally, standard deviation and variance are calculated for the net scaled volumes of each group's investors using the following formulas:

$$\text{std}(r(i,t,g)) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (r(i,t,g) - \bar{r}(t,g))^2}, \tag{4.4}$$

$$\text{var}(r(i,t,g)) = \frac{1}{N} \sum_{i=1}^{N} (r(i,t,g) - \bar{r}(t,g))^2, \tag{4.5}$$

where $\bar{r}(t,g)$ represents the mean net scaled volume of group g investors at time t. These calculations provide a proxy for differences in group opinion, a factor found to be significant in related studies by Qian (2014).

4.3.3 Investor Activity

Two variables are defined to describe group activity:

Buying/Selling Volume Share: This metric represents the percentage of a group's buying or selling volume relative to the total buying or selling volume of the stock on a given day. It is calculated as:

$$\text{VS}(g,t) = \frac{V(g,t)}{V(t)}, \tag{4.6}$$

where $V(g,t)$ is the total buying or selling volume of the stock by group g on day t, and $V(t)$ is the total buying or selling volume of the stock on day t. This metric indicates how actively a group of buyers or sellers participated in trading on a specific day.

Investor Activity: This variable represents the proportion of active investors within a group who traded a particular stock on a given day. It is calculated as:

$$A(g,t) = \frac{N(g,t)}{N(g)}, \tag{4.7}$$

where $N(g,t)$ is the number of investors in group g who traded the stock on day t, and $N(g)$ is the total number of investors in group g. The total number of investors in a group is determined by considering traders who have traded the underlying stock at least once between 2000 and 2009.

4.3.4 Investor Interaction

The research aims to analyze the trading between and within groups of investors to model the provision of liquidity among groups. However, due to the absence of direct matches between the buy-side and sell-side of transactions, a method is required to estimate the amount of trading occurring between groups.

The study adopts a method proposed by Stoffman (2014), which uses the same dataset. This method assumes that trade occurs in proportion to the amount of buying or selling accounted for by each group. It estimates the trading within and between groups by calculating the proportion of shares bought or sold by each group from other members of the same group or from other groups.

The volume sold to households is estimated using the formula:

$$\mathsf{VH}_s(g,t) = \sum_{p \in \mathsf{prices}} \left(V_s(g,t,p) \cdot \frac{V_{b_h h}(t,p)}{V_b(g,t,p)} \right), \tag{4.8}$$

where $V_s(g,t,p)$ represents the volume sold by group g on day t at price p, and $V_{b_h h}(t,p)$ represents the volume bought by households on day t at price p. Similarly, the volume bought from households is estimated using the formula:

$$\mathsf{VH}_b(g,t) = \sum_{p \in \mathsf{prices}} \left(V_b(g,t,p) \cdot \frac{V_{s_h h}(t,p)}{V_s(g,t,p)} \right), \tag{4.9}$$

where $V_b(g,t,p)$ represents the volume bought by group g on day t at price p, and $V_{s_h h}(t,p)$ represents the volume sold by households on day t at price p.

By assuming that trades occur proportionally based on buying and selling activity, this method provides estimates of trading between and within investor groups. It enables the identification of the exact amount of trading between groups in cases where a group has only bought or sold shares at a particular price. The research by Stoffman (2014)

demonstrates the effectiveness of this method in estimating trading interactions.

4.3.5 Model Output - Stock Direction

The study assigns price direction states to each stock, which serve as the model's output. The price direction is determined based on the net scaled volume of a stock:

$$D(s,t,\theta) = \begin{cases} d, & \text{if } r(s,t) < -\theta, \\ f, & \text{if } -\theta \leq r(s,t) \leq \theta \\ u, & \text{if } r(s,t) > \theta. \end{cases} \tag{4.10}$$

where $r(s,t)$ represents the net scaled volume of stock s on day t, θ is the categorization threshold, d represents the down movement state, u represents the up movement state, and f represents the flat state. The threshold is set separately for the training and testing data, ensuring that approximately 33% of the sample falls within the flat movement state. The specific thresholds used for each horizon are presented in Table 4.4.

Table 4.4. *Threshold Values for Different Horizons*

	1-day Horizon	**5-day Horizon**	**21-day Horizon**
θ (Train)	0.0065	0.0155	0.036
θ (Test)	0.0096	0.022	0.049

The set thresholds results in following class balances for testing and training sets 4.5:

Table 4.5. *Class Balanes*

	1-day Horizon	**5-day Horizon**	**21-day Horizon**
Train	[32.3% 33.5% 34.2%]	[30.5% 33.1% 36.4%]	[26.9% 33.2% 39.9%]
Test	[34.0% 33.3% 32.7%]	[33.9% 33.2% 32.8%]	[35.9% 33.6% 30.5%]

The classes are fairly balanced in 1-day and 5-day horizons for testing and training sets. However, In 21-day horizon there is a notable overrepresentation of class 2 in training set and conversely a slight underrepresentation in testing set. This may lead to problems for the 21-day horizon setups.

4.4 Feature Selection

The features used in the analysis are derived from the dataset using the aforementioned data processing methods. The selected features are presented in Table 4.6.

These selected features aim to provide proxies for investor behavior. Scaling the features

Table 4.6. *Features*

Feature	Description	Range
activity	Percentage of groups investors that traded during a day	[0, 1]
buy_household_share	Percentage of groups buying volume that was bought from households	[0, 1]
buy_volume_share	Groups buying volume in proportion to stocks total buying volume	[0, 1]
net_scaled_volume1	Aggregate net scaled volume	[-1, 1]
net_scaled_volume2	Average net scaled volume. All investors under group weighted equally.	[-1, 1]
nsv_std	Standard deviation of groups net scaled volume	[0, 1]
nsv_var	Variance of groups net scaled volume	[0, 1]
sell_household_share	Percentage of groups selling volume that was sold to households	[0, 1]
sell_volume_share	Groups selling volume in proportion to stocks total selling volume	[0, 1]

effectively addresses data heterogeneity and facilitates faster convergence of machine learning algorithms.

In addition to the selected features, there are control variables that are taken into account, as shown in Table 4.7.

Table 4.7. *Control Features*

Variable	Description	Range
Yearly Dummy	Indicator variable for yearly time periods	0 or 1
Monthly Dummy	Indicator variable for monthly time periods	0 or 1
Stock Dummy	Indicator variable for specific stock	0 or 1
Stock Returns (1-day Horizon)	Returns of the specific stock (1-day horizon)	Numerical
Stock Returns (5-day Horizon)	Returns of the specific stock (5-day horizon)	Numerical
Stock Returns (21-day Horizon)	Returns of the specific stock (21-day horizon)	Numerical
Stock Returns (64-day Horizon)	Returns of the specific stock (64-day horizon)	Numerical
Stock Returns (128-day Horizon)	Returns of the specific stock (128-day horizon)	Numerical
Normalized Trading Volumes (1-day Horizon)	Trading volumes normalized for the specific stock (1-day horizon)	Numerical
Normalized Trading Volumes (5-day Horizon)	Trading volumes normalized for the specific stock (5-day horizon)	Numerical
Normalized Trading Volumes (21-day Horizon)	Trading volumes normalized for the specific stock (21-day horizon)	Numerical
Normalized Trading Volumes (64-day Horizon)	Trading volumes normalized for the specific stock (64-day horizon)	Numerical
Normalized Trading Volumes (128-day Horizon)	Trading volumes normalized for the specific stock (128-day horizon)	Numerical
Volatility (5-day Horizon)	Volatility measure for the specific stock (5-day horizon)	Numerical
Volatility (21-day Horizon)	Volatility measure for the specific stock (21-day horizon)	Numerical
Volatility (64-day Horizon)	Volatility measure for the specific stock (64-day horizon)	Numerical
Volatility (128-day Horizon)	Volatility measure for the specific stock (128-day horizon)	Numerical

The control variables provide additional information and context for the analysis. The combined set of features and control variables results in a total of 97 features when considering different dummies and independent features for each investor group. These features are inputted as a vector into the machine learning models, with the price direction state as the target variable.

4.5 Model setup

The modelling was performed in the Pycharm environment, which is a popular integrated development environment (IDE) for Python programming. The Pycharm environment offers a wide range of tools and features that are helpful for data analysis and machine learning tasks, including code completion, debugging, and project management.

To support the modelling process, several libraries were employed, including Pandas, NumPy, SciPy, and Scikit-learn. Pandas is a popular library for data manipulation and analysis, providing tools for working with structured data, including reading and writing data, indexing, filtering, grouping, and joining. NumPy is a numerical computing library that provides support for working with large arrays and matrices, including mathematical functions and algorithms. SciPy is a scientific computing library that provides tools for signal processing, optimization, and statistical analysis. Finally, Scikit-learn is a machine learning library that provides a range of algorithms and tools for data preprocessing, model selection, and evaluation.

To achieve the research objectives, different machine learning algorithms are applied, including logistic regression, random forest, XGBoost, lightGBM and adaboost. Notably, logistic regression functions as a baseline model, which the other models are compared to. In addition to the machine learning algorithms, a naive model is provided as a point of reference for performance evaluation. The naive model is a previous period model, which always predicts the same outcome what happened in the previous period.

The research is conducted over a specific time period, with training data collected between 2000 and 2007, and testing data collected between 2007 and 2009. The time periods are chosen to test if models trained with earlier period work with a future period. Furthermore, both of the periods include a stock market crash and data from pre and post crash periods.

To explore the potential of trading activity data as a predictor over different dimensions, the study employs two different setups, resulting in six different configurations. The first setup focuses on predicting comovements of trading activity and stock returns over different time horizons. The contemporary setup acts as a proof of concept on the relation of stock returns and trading behaviour. The lead-lag setup is the more important one and studies how well trading behaviour can predict future stock returns. The different setups and their description are presented in table 4.8

These different configurations result in the following (table 4.9) training and testing sets for lead-lag setup (samples, features, time-steps). The comovement configurations have n_stocks*n_timesteps more samples as the lead-lag requires shifting the data to achieve the desired lag. Training set is used for for cross-validation for hyperparameter tuning and testing set is used for the final results.

Table 4.8. *Model Configurations*

Time Horizon (days)	Comovement	Lead-lag
1	1-day features to predict stock direction on the same day	1-day features to predict stock direction on the next day
5	5-day features to predict stock direction on the concurrent 5 days	5-day features to predict stock direction on the next 5 days
21	21-day features to predict stock direction on the concurrent 21 days	21-day features to predict stock direction on the next 21 days

Table 4.9. *Training and testing sets*

	1-day Horizon	5-day Horizon	21-day Horizon
Train	(17444, 97, 1)	(17364, 97, 5)	(17044, 97, 21)
Test	(6950, 97, 1)	(6870, 97, 5)	(6550, 97, 21)

Finally, the performance of the different models and setups is evaluated using several performance metrics, including accuracy, precision (macro), recall (macro), F1 score (macro), and ROC curve analysis. These metrics provide insights into the efficacy of the different models and configurations in predicting stock movements.

4.6 Hyperparameter tuning

Because of the time series nature of the data, the cross-validation strategy used here is successive training sets. In this approach, the data is split into training and testing sets in such a way that the training set always precedes the testing set in time, to preserve the temporal order of the data. This method is particularly useful in time-series data where temporal dynamics are important and should not be disrupted. Furthermore, it has been used for machine learning in stock market prediction (Jiang, 2021). The splits are described in table 4.10. The amount of splits could be higher, but it would result in the first folds training sample probably being too small.

Table 4.10. *Cross-Validation Split*

Fold	2000	2001	2002	2003	2004	2005	2006
1	Train	Train	Train	Valid	-	-	-
2	Train	Train	Train	Train	Valid	-	-
3	Train	Train	Train	Train	Train	Valid	-
4	Train	Train	Train	Train	Train	Train	Valid

The hyperparameter search algorithm used is Bayesian optimization. It is an iterative algorithm widely utilized for hyperparameter tuning in machine learning models. Unlike Grid Search and Random Search, Bayesian optimization determines the next evaluation points based on the previously obtained results. It employs two key components: a surrogate model and an acquisition function. The surrogate model aims to fit all the observed points into the objective function, while the acquisition function balances exploration and exploitation by determining which points to sample next. The exploration aspect involves sampling instances in unexplored areas, ensuring comprehensive coverage of the hyperparameter space. Exploitation, on the other hand, focuses on sampling in the regions that exhibit promising performance, as indicated by the posterior distribution. By striking a balance between exploration and exploitation, Bayesian optimization effectively identifies the most likely optimal regions and avoids missing better configurations in unexplored areas. Most importantly, it typically achieves near-optimal hyperparameter combinations within a few iterations. (Snoek et al., 2012)

The core procedures of Bayesian optimization can be summarized as follows (Snoek et al., 2012):

1. Construct a probabilistic surrogate model representing the objective function.
2. Identify the hyperparameter values that optimize the surrogate model.
3. Evaluate these hyperparameter values on the actual objective function.
4. Update the surrogate model with the new results.
5. Repeat steps 2–4 until reaching the maximum number of iterations.

This iterative process enables the algorithm to efficiently explore the hyperparameter space.

For the optimization process, accuracy is used as a scoring method for finding the best hyperparameter combinations. All the searches ran for 100 iterations, except Adaboost and XGBoost both of which run 50 iterations at 21-day horizon setup. Furthermore, the sampling parameters of XGBoost (colsample_bytree, colsample_bylevel, colsample_bynode, subsample) and LightGBM (feature_fraction, feature_fraction_bynode, bagging_fraction) are set to 0.7. The hyperparameter spaces used for AdaBoost, random forest, XGBoost and LightGBM are presented in tables 4.11, 4.12, 4.13 and 4.14 respectively.

Table 4.11. *Hyperparameter Space for AdaBoost (AB)*

Hyperparameter	**Range**
learning_rate	0.001 to 1.0
n_estimators	10 to 200

Table 4.12. *Hyperparameter Space for Random Forest (RF)*

Hyperparameter	**Range**
n_estimators	20 to 200
max_depth	3 to 15
min_samples_split	2 to 10
min_samples_leaf	1 to 10
max_features	sqrt, log2

Table 4.13. *Hyperparameter Space for XGBoost (XGB)*

Hyperparameter	**Range**
n_estimators	20 to 200
max_depth	3 to 15
learning_rate	0.001 to 1.0
gamma	0.0 to 5.0
min_child_weight	1 to 10
reg_alpha	0.0 to 1.0
reg_lambda	0.0 to 1.0

Table 4.14. *Hyperparameter Space for LightGBM (LGBM)*

Hyperparameter	**Range**
n_estimators	20 to 200
num_leaves	2 to 200
max_depth	3 to 15
learning_rate	0.001 to 1.0
min_child_weight	1 to 10
reg_alpha	0.0 to 1.0
reg_lambda	0.0 to 1.0

After running the Bayes search on each model's hyperparameter spaces, the chosen hyperparameters are presented on tables 4.15, 4.16 4.17 and 4.18.

Table 4.15. *Random Forest (RF) Hyperparameters*

Hyperparameters	1-day Horizon	5-day Horizon	21-day Horizon
n_estimators	148	42	84
max_depth	9	8	14
min_samples_split	8	8	8
min_samples_leaf	8	4	1
max_features	sqrt	log2	log2

Table 4.16. *XGBoost (XGB) Hyperparameters*

Hyperparameters	1-day Horizon	5-day Horizon	21-day Horizon
n_estimators	168	20	150
max_depth	6	3	3
learning_rate	0.01	0.0368	0.2148
gamma	5.0	4.3340	1.9108
min_child_weight	10	10	10
reg_alpha	0.0	1.0	0.9899
reg_lambda	0.0	0.0	0.2491

Table 4.17. *AdaBoost (AB) Hyperparameters*

Hyperparameters	1-day Horizon	5-day Horizon	21-day Horizon
learning_rate	0.1565	0.2360	0.0018
n_estimators	55	10	25

Table 4.18. *LightGBM (LGBM) Hyperparameters*

Hyperparameters	1-day Horizon	5-day Horizon	21-day Horizon
n_estimators	164	191	78
num_leaves	20	198	50
max_depth	12	3	3
learning_rate	0.01	0.01	0.3326
min_child_weight	2	1	10
reg_alpha	0.0	0.6614	1.0
reg_lambda	1.0	0.1181	0.0573

The abovementioned hyperparameters were used to train on the training data and testing the results on the testing data. The final results are in the next chapter.

5. RESULTS

5.1 Comovement

For the one-day horizon, all the machine learning models are better than the naive model and baseline logistic regression model. Random forest performed the best across all metrics, although the differences are small to gradient boosting models. All of the models are well above 50% accuracy as shown in table 5.1.

Table 5.1. *Comovement Results - 1-day horizon*

	LogReg	RF	LGBM	XGB	AB	Previous period
accuracy	0.540948	**0.602011**	0.601293	0.601580	0.576724	0.357471
f_score	0.495527	**0.579683**	0.579459	0.578212	0.540647	0.357470
recall	0.540948	**0.601168**	0.600408	0.600783	0.575892	0.357471
precision	0.560957	0.609654	**0.610804**	0.609491	0.584395	0.357470
roc_auc	0.736413	0.784100	0.784113	**0.786831**	0.767022	0.518025

When it comes to 5-day horizon, there is a clear step down in accuracy as the table 5.2 indicates. Furthermore, the differences between models are larger than before. Adaboost is the first model that fails to beat the logistic regression. Random forest and gradient boosting algorithms are still the top performers, but lightGBM performs best this time.

Table 5.2. *Comovement Results - 5-day horizon*

	LogReg	RF	LGBM	XGB	AB	Previous period
accuracy	0.516763	0.543208	**0.558960**	0.540029	0.513006	0.333526
f_score	0.490718	0.526371	**0.541080**	0.512123	0.477645	0.333517
recall	0.516763	0.542638	**0.558003**	0.539747	0.512459	0.333526
precision	0.514434	0.541154	**0.561168**	0.543420	0.502709	0.333511
roc_auc	0.701682	0.730311	**0.752839**	0.738259	0.705793	0.500089

Finally, 21-day horizon continues a similar trend. Shown in table 5.3, overall results drop even more and now in addition to adaboost, also random forest fails to beat logistic regression. Gradient boosting algorithms' results aren't dropping as rapidly as others and lightGBM is again the top performer.

Table 5.3. *Comovement Results - 21-day horizon*

	LogReg	**RF**	**LGBM**	**XGB**	**AB**	**Previous period**
accuracy	0.454586	0.453550	0.494675	**0.503402**	0.427811	0.387130
f_score	0.450490	0.447626	0.492671	**0.501933**	0.338570	0.387131
recall	0.454586	0.459447	0.494758	**0.505507**	0.428596	0.387130
precision	0.452508	0.465460	0.492653	**0.501438**	0.282366	0.387157
roc_auc	0.626274	0.659234	0.677600	**0.682361**	0.584526	0.539629

The results are in line with previous research from Kumar and Lee (2006) and Cai and Zheng (2004) suggesting that investor behaviour affects returns at least contemporaneously across different time horizons. The next section will cover the same results in lead-lag setting.

5.2 Lead-lag

In lead-lag setting, the predictive power is a lot lower than in the contemporaneous setting. The results are presented in table 5.4. However, models still beat the baseline model and naive model by a small but clear margin. Due to computational limitations, only these most important results are trained using 100 different random states. The results are presented in the form [min, median, max] for applicable algorithms. LightGBM is the best predictor by a tiny margin, but similarly to the comovement setting, the results of other RF and XGB are very close. Furthermore, there doesn't seem to be too high sensitivity in the results.

Table 5.4. *Lead-lag results - 1-day horizon (Part 1)*

	LogReg	**AB**	**Previous period**
accuracy	0.3847	0.3997	0.3576
f_score	0.3767	0.4007	0.3576
recall	0.3847	0.3997	0.3576
precision	0.4108	0.4140	0.3576
roc_auc	0.5790	0.5974	0.5181
	RF	**LGBM**	**XGB**
accuracy	[0.3947, 0.4016, 0.4081]	[0.3970, **0.4027**, 0.4069]	[0.3896, 0.3974, 0.4043]
f_score	[0.3906, 0.4011, 0.4086]	[0.3979, **0.4031**, 0.4079]	[0.3690, 0.3853, 0.3991]
recall	[0.3951, 0.4024, 0.4086]	[0.3974, **0.4029**, 0.4075]	[0.3924, 0.4000, 0.4065]
precision	[0.4066, 0.4160, 0.4216]	[0.4120, **0.4170**, 0.4209]	[0.4078, 0.4168, 0.4254]
roc_auc	[0.5916, 0.5975, 0.6008]	[0.5961, 0.5977, 0.5999]	[0.5977, **0.5996**, 0.6022]

In the 5-day horizon, contrary to comovement results, the results presented in table 5.5, don't drop as distinctively in the lead-lag setting. The margins to baseline and naive models are smaller, but lightGBM manages to retain almost the same performance as in the 1-day horizon. Similarly to comovement results, AdaBoost struggles to beat logistic regression in this horizon, but random forest and both gradient boosting algorithms are still performing better.

Table 5.5. *Lead-lag results - 5-day horizon*

	LogReg	**RF**	**LGBM**	**XGB**	**AB**	**Previous period**
accuracy	0.372052	0.397234	**0.399272**	0.377875	0.373362	0.333624
f_score	0.359258	**0.398179**	0.385462	0.371756	0.370629	0.333659
recall	0.372052	0.396971	**0.397155**	0.379262	0.374182	0.333624
precision	0.390427	0.402284	**0.407572**	0.402410	0.398782	0.333695
roc_auc	0.556441	0.577682	0.582086	0.583513	**0.584941**	0.500162

Continuing the analysis for the 21-day horizon in the lead-lag setting, table 5.6 showcases a noticeable decline in the performance of the machine learning models when compared to the shorter horizons. Interestingly, for the 21-day horizon, all machine learning models

fail to outperform the naive model based on the previous period. This is quite contrary to the results seen in the comovement setting as the accuracy of all models is close to random guessing.

Table 5.6. *Lead-lag results - 21-day horizon*

	LogReg	**RF**	**LGBM**	**XGB**	**AB**	**Previous period**
accuracy	0.326260	0.346260	0.332977	0.351908	0.295878	**0.386565**
f_score	0.321562	0.325459	0.327357	0.337390	0.238879	**0.386704**
recall	0.326260	0.360145	0.340832	0.364070	0.296770	**0.386565**
precision	0.339361	0.375852	0.336133	0.374118	0.204089	**0.386861**
roc_auc	0.497182	0.526507	0.500032	0.535033	0.509410	**0.538870**

The results are mixed with respect to previous research. Daily and weekly predictability for example by Campbell et al. (2009) and Barber et al. (2009a) seem to be reproducible but monthly or longer horizons, for example observed by Kaniel et al. (2008) and Hvidkjaer (2008), do not look promising.

In conclusion, the results show a degree of return predictability across different settings and time horizons. Furthermore, most ensemble methods used seem to beat logistic regression as a baseline model.

6. CONCLUSION

6.1 Key findings

This study aimed to examine the predictive power of machine learning models in forecasting asset returns, focusing on the comovement and lead-lag relationships between returns and investor trading behaviour across different time horizons. By analyzing a comprehensive dataset and employing a range of machine learning algorithms, the study sought to assess the accuracy and robustness of these models and shed light on their potential practical implications.

The first finding was that machine learning algorithms were able to provide better results compared to traditional logistic regression. The results are especially clear in comovement settings but also prevalent in most of the lead-lag settings, possibly implying nonlinear behaviour between investor trading behaviour and returns. This is in line with the growing literature where advanced machine learning algorithms are used in different stock market-related prediction tasks (Henrique et al., 2019).

The second finding was that investor trading behaviour affected stock returns in a contemporaneous setting. The results are in line with previous research from Kumar and Lee (2006) and Cai and Zheng (2004). These findings suggest that contemporaneous effects may be prevalent in different markets other than US markets.

The third finding was that investor trading behaviour affected stock returns in a lead-lag setting, although the results weren't as distinct as in a contemporaneous setting. These results are in line with previous literature (Barber et al., 2009a; Campbell et al., 2009). However, the observed effects are probably not economically large enough to be used in creation of trading strategies.

The fourth finding was that prediction power got worse the longer the time horizon was. This was especially evident for the 21-day horizon lead-lag setting, as all the models failed to outperform the naive model. These findings are somewhat contradicting to the results of Kaniel et al. (2008) and Hvidkjaer (2008), as the models failed to provide predictions over longer time periods. These differences could result from the differences between the Finnish and US stock markets. It could also imply that other variables, like news, could affect the prices more than trading behaviour on longer horizons. Furthermore,

the effects could also be because of sample imbalances in training and testing sets. For shorter horizons, the imbalances were not as large, but for 21-day horizon there was a noticeable imbalance.

6.2 Limitations and quality assessment of the study

The research presents informative findings, but it is important to scrutinize the limitations and critically assess the quality of the study. The sample bias is one of the primary limitations. As the dataset utilized was sourced from the relatively small Finnish stock market, it raises concerns about the generalizability of these findings to other markets. Stock markets across the globe are often influenced by a complex interplay of country-specific economic, political, and cultural factors. The absence of a more diverse dataset limits the universal applicability of the results.

In terms of predictive power, the study indicates a decrease in the efficacy of machine learning models with increasing time horizons. This is a noteworthy observation, but it also poses questions regarding the breadth of time horizons that these models can reliably cater to. Future studies may need to explore whether adjustments to model parameters with more available computational resources could help to lessen the effect.

Another limitation lies in the scope of variables considered. While the study explored the relationships between investor trading behavior and stock returns, it did not consider other influential variables such as macroeconomic indicators (Kumar & Lee, 2006) or market news (Barrot et al., 2016; Kaniel et al., 2012). Inclusion of these variables could yield a more holistic understanding and possibly enhance the precision of predictive models.

Additionally, while machine learning models have shown superior predictive power compared to traditional logistic regression, their intrinsic complexity presents a challenge. Often labeled as the "black box" problem, the complexity makes it difficult for practitioners to interpret the underlying mechanics of these models (Arrieta et al., 2020). This could potentially hinder their adoption in practical settings where understanding the rationale behind predictions is vital.

Finally, the assumption of bi-directional interaction between trading flows and returns, though grounded in literature (Ülkü & Weber, 2013), introduces cautions in the interpretation of results. It is challenging to delineate clear causal relationships with certainty due to the interactive nature of these variables. This emphasizes the need for methodologies that can better unravel the cause-effect dynamics at play.

In conclusion, while the study is rigorously executed and contributes to the existing body of knowledge, these limitations need to be recognized for a balanced evaluation of its findings. The research lays the groundwork for further exploration and development of more robust and interpretable machine learning models for asset return predictions.

6.3 Proposals for future research

In light of the findings and limitations, this study opens several possibilities for future research:

The first proposal is the investigation of additional variables: This study focused on the contemporaneous and lead-lag relationships between returns and investor trading behaviour. Firstly, it would be important to replicate previous research results using the same variables and models to be able to precisely compare the results between markets. Next, machine learning models could be tested if they're able to beat the models of previous research using the same set of variables. Finally, future research could consider incorporating additional variables that may not have been exhaustively used such as market news and macroeconomic indicators, to further test if the predictability increases. Most of the price prediction literature have included before-mentioned features in their models (Henrique et al., 2019).

The second proposal is the utilization of deep learning: As machine learning algorithms provided better results than traditional logistic regression, it may be worth exploring deep learning techniques. Neural networks with multiple layers could capture complex patterns and relationships, potentially enhancing the prediction accuracy for asset returns. For example, temporal attention augmented bilinear networks have been used in high-frequency trading prediction (Tran et al., 2017).

The third proposal is cross-market analysis: This study observed differences in the results when compared to previous research, suggesting that the dynamics in the Finnish market might differ from the US markets. Future research can perform a cross-market analysis, studying multiple stock markets simultaneously to assess the generalizability of the machine learning models. This would be a particularly interesting but challenging task, as it could be difficult to combine datasets from different countries.

The final proposal is the exploration of different time horizons. As the predictive power was observed to decline over longer time horizons, investigating a wider range of time horizons, including very short-term (intraday) and very long-term (monthly, yearly) predictions, could provide insights into the optimal horizons for employing machine learning models. Especially intraday research could be interesting, as it has not been in a focus on the previous literature.

REFERENCES

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of adaboost and neural networks [Data Warehousing and OLAP]. *Decision Support Systems*, *45*(1), 110–122. https://doi.org/https://doi.org/10.1016/j.dss.2007.12.002

Arrieta, B. A., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, *61*(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perpectives*, *21*(2), 129–151. https://doi.org/10.1257/jep.21.2.129

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, *42*(20), 7046–7056. https://doi.org/https://doi.org/10.1016/j.eswa.2015.05.013

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, *116*(1), 261–292. https://doi.org/10.1162/003355301556400

Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The Journal of Finance*, *55*(2), 773–806. https://doi.org/https://doi.org/10.1111/0022-1082.00226

Barber, B. M., Odean, T., & Zhu, N. (2009a). Do retail trades move markets? *The Review of Financial Studies*, *22*(1), 151–186. Retrieved June 14, 2023, from http://www.jstor.org/stable/40056908

Barber, B. M., Odean, T., & Zhu, N. (2009b). Systematic noise. *Journal of Financial Markets*, *12*(4), 547–569. https://doi.org/https://doi.org/10.1016/j.finmar.2009.03.003

Barberis, N., & Thaler, R. (2002). *A survey of behavioral finance* (Working Paper No. 9222). National Bureau of Economic Research. https://doi.org/10.3386/w9222

Barrot, J.-N., Kaniel, R., & Sraer, D. (2016). Are retail traders compensated for providing liquidity? *Journal of Financial Economics*, *120*(1), 146–168. https://doi.org/https://doi.org/10.1016/j.jfineco.2016.01.005

Boehmer, E., Jones, C. M., Zhang, X., & Zhang, X. (2021). Tracking retail investor activity. *Journal of Finance*, *76*(5), 2249–2305. https://doi.org/10.1111/jofi.13033

Breiman, L. (2001). Random forests. *MACHINE LEARNING*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cai, F., & Zheng, L. (2004). Institutional trading and stock returns. *Finance Research Letters*, *1*(3), 178–189. https://doi.org/https://doi.org/10.1016/j.frl.2004.06.003

Campbell, J. Y., Ramadorai, T., & Schwartz, A. (2009). Caught on tape: Institutional trading, stock returns, and earnings announcements. *Journal of Financial Economics*, *92*(1), 66–91. https://doi.org/https://doi.org/10.1016/j.jfineco.2008.03.006

Challet, D., Chicheportiche, R., Lallouache, M., & Kassibrakis, S. (2016). Statistically validated lead-lag networks and inventory prediction in the foreign exchange market. https://doi.org/10.48550/ARXIV.1609.04640

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system [22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, AUG 13-17, 2016]. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Lin, W. T., Ma, C., & Tsai, S.-C. (2014). Liquidity provisions by individual investor trading prior to dividend announcements: Evidence from taiwan. *The North American Journal of Economics and Finance*, *28*, 358–374. https://doi.org/https://doi.org/10.1016/j.najef.2014.03.006

Chevalier, J., & Ellison, G. (1999). Career concerns of mutual fund managers. *The Quarterly Journal of Economics*, *114*(2), 389–432. Retrieved April 14, 2023, from http://www.jstor.org/stable/2587013

Coval, J., & Stafford, E. (2007). Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics*, *86*(2), 479–512. https://doi.org/https://doi.org/10.1016/j.jfineco.2006.09.007

David W. Hosmer, S. L. (2000). Introduction to the logistic regression model. In *Applied logistic regression* (pp. 1–30). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/0471722146.ch1

DeLong, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703–738. Retrieved April 14, 2023, from http://www.jstor.org/stable/2937765

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, *38*(1), 34–105. Retrieved April 14, 2023, from http://www.jstor.org/stable/2350752

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Grinblatt, M., & Keloharju, M. (2000). The investment behavior and performance of various investor types: A study of finland's unique data set. *Journal of Financial Eco-*

nomics, *55*(1), 43–67. https://doi.org/https://doi.org/10.1016/S0304-405X(99)00044-6

Grinblatt, M., & Keloharju, M. (2001). What makes investors trade? *The Journal of Finance*, *56*(2), 589–616. https://doi.org/https://doi.org/10.1111/0022-1082.00338

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226–251. https://doi.org/https://doi.org/10.1016/j.eswa.2019.01.012

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, *193*. https://doi.org/10.1016/j.eswa.2021.116429

Hvidkjaer, S. (2006). A trade-based analysis of momentum. *The Review of Financial Studies*, *19*(2), 457–491. Retrieved June 21, 2023, from http://www.jstor.org/stable/3844003

Hvidkjaer, S. (2008). Small Trades and the Cross-Section of Stock Returns. *The Review of Financial Studies*, *21*(3), 1123–1151. https://doi.org/10.1093/rfs/hhn049

Ivkovic, Z., & Weisbenner, S. (2005). Local does as local is: Information content of the geography of individual investors' common stock investments. *Journal of Finance*, *60*(1), 267–306. https://doi.org/10.1111/j.1540-6261.2005.00730.x

Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, *184*, 115537. https://doi.org/https://doi.org/10.1016/j.eswa.2021.115537

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kaniel, R., Liu, S., Saar, G., & Titman, S. (2012). Individual investor trading and return patterns around earnings announcements. *The Journal of Finance*, *67*(2), 639–680. https://doi.org/https://doi.org/10.1111/j.1540-6261.2012.01727.x

Kaniel, R., Saar, G., & Titman, S. (2008). Individual investor trading and stock returns. *The Journal of Finance*, *63*(1), 273–310. https://doi.org/https://doi.org/10.1111/j.1540-6261.2008.01316.x

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree [31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, DEC 04-09, 2017]. In I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30 (nips 2017)*.

Kelley, E. K., & Tetlock, P. C. (2013). How wise are crowds? insights from retail orders and stock returns. *The Journal of Finance*, *68*(3), 1229–1265. https://doi.org/https://doi.org/10.1111/jofi.12028

Kim, K.-j., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with*

Applications, *19*(2), 125–132. https://doi.org/https://doi.org/10.1016/S0957-4174(00)00027-0

Kim, Y., & Lee, K. Y. (2022). Impact of investor sentiment on stock returns. *Asia-Pacific Journal of Financial Studies*, *51*(1), 132–162. https://doi.org/https://doi.org/10.1111/ajfs.12362

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, *259*(2), 689–702. https://doi.org/https://doi.org/10.1016/j.ejor.2016.10.031

Kumar, A., & Lee, C. M. (2006). Retail investor sentiment and return comovements. *The Journal of Finance*, *61*(5), 2451–2486. https://doi.org/https://doi.org/10.1111/j.1540-6261.2006.01063.x

Lakonishok, J., Shleifer, A., Thaler, R., & Vishny, R. (1991). Window dressing by pension fund managers. *The American Economic Review*, *81*(2), 227–231. Retrieved April 14, 2023, from http://www.jstor.org/stable/2006859

Lee, W., Jiang, C., & Indro, D. (2002). Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking & Finance*, *26*(12), 2277–2299. https://doi.org/10.1016/S0378-4266(01)00202-3

Lien, D., Hung, P.-H., & Lin, Z.-W. (2020). Whose trades move stock prices? evidence from the taiwan stock exchange. *International Review of Economics & Finance*, *66*, 25–50. https://doi.org/https://doi.org/10.1016/j.iref.2019.10.011

Lillo, F., Miccichè, S., Tumminello, M., Piilo, J., & Mantegna, R. N. (2015). How news affects the trading behaviour of different categories of investors in a financial market. *Quantitative Finance*, *15*(2), 213–229. https://doi.org/10.1080/14697688.2014.931593

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Narayan, P. K., Narayan, S., & Westerlund, J. (2015). Do order imbalances predict chinese stock returns? new evidence from intraday data. *Pacific-Basin Finance Journal*, *34*, 136–151. https://doi.org/https://doi.org/10.1016/j.pacfin.2015.07.003

Odean, T. (1999). Do investors trade too much? *American Economic Review*, *89*(5), 1279–1298. https://doi.org/10.1257/aer.89.5.1279

Pai, P.-F., & Lin, C.-S. (2005). A hybrid arima and support vector machines model in stock price forecasting. *Omega*, *33*(6), 497–505. https://doi.org/https://doi.org/10.1016/j.omega.2004.07.024

Qian, X. (2014). Small investor sentiment, differences of opinion and stock overvaluation. *Journal of Financial Markets*, *19*, 219–246. https://doi.org/https://doi.org/10.1016/j.finmar.2014.03.005

San, G. (2005). Who gains more by trading – institutions or individuals? *Capital Markets: Market Efficiency eJournal*. https://doi.org/10.2139/ssrn.687415

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Comput. Appl.*, *34*(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, *17*(1), 83–104. https://doi.org/10.1257/089533003321164967

Snoek, J., Larochelle, H., & Adams, R. P. Practical bayesian optimization of machine learning algorithms [Cited by: 3897]. In: *4*. Cited by: 3897. 2012, 2951–2959. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84869201485&partnerID=40&md5=14aa83df115308a2cf91468d634a36aa

Stoffman, N. (2014). Who trades with whom? individuals, institutions, and returns. *Journal of Financial Markets*, *21*, 50–75. https://doi.org/https://doi.org/10.1016/j.finmar.2014.08.002

Subrahmanyam, A. (2008). Behavioural finance: A review and synthesis. *European Financial Management*, *14*(1), 12–29. https://doi.org/https://doi.org/10.1111/j.1468-036X.2007.00415.x

Tran, D. T., Iosifidis, A., Kanniainen, J., & Gabbouj, M. (2017). Temporal attention augmented bilinear network for financial time-series data analysis. *CoRR*, *abs/1712.00975*. http://arxiv.org/abs/1712.00975

Tumminello, M., Lillo, F., Piilo, J., & Mantegna, R. N. (2012). Identification of clusters of investors from their real trading activity in a financial market. *New Journal of Physics*, *14*(1), 013041. https://doi.org/10.1088/1367-2630/14/1/013041

Ülkü, N., & Weber, E. (2013). Identifying the interaction between stock market returns and trading flows of investor types: Looking into the day using daily data. *Journal of Banking & Finance*, *37*(8), 2733–2749. https://doi.org/https://doi.org/10.1016/j.jbankfin.2013.03.021

Yang, C., & Zhou, L. (2015). Investor trading behavior, investor sentiment and asset prices. *The North American Journal of Economics and Finance*, *34*, 42–62. https://doi.org/https://doi.org/10.1016/j.najef.2015.08.003