University of Warsaw
Faculty of Economic Sciences

Turgud Valiyev
Album N°: 466760

ANALYZING STOCK MARKET BEHAVIOR AND PRICE PREDICTIONS ACROSS GLOBAL MARKETS

Magister (master) degree thesis
Field of the study: Data Science and Business Analytics

The thesis written under the supervision of
dr hab. Piotr Wojcik, prof. ucz.
from Department of Data Science
WNE UW

Warsaw, June 2025

Declaration of the supervisor

I declare that the following thesis was written under my supervision and I state that it meets all criterias to be submitted for the procedure of academic degree award.

~~*I declare that my participation in the scientific article, which is a part of this thesis is ....%, while the supplement to the thesis was written independently by the graduate (s).~~

* *cross out if not applicable*

Date

26.05.2025

Signature of the Supervisor

Piotr Woj

*Declaration of the author of the thesis***

Aware of the legal responsibility, I declare that I am the sole author of the following thesis and that the thesis is free from any content that constitutes copyright infringement or has been acquired contrary to applicable laws and regulations.

I also declare that the thesis has never been a subject of degree-awarding procedures in any higher education institution.

Moreover I declare that the attached version of the thesis is identical with the enclosed electronic version.

~~* I declare that my*** participation in the scientific article, which is part of this thesis is ....% (not less than 60%), while the supplement to the thesis was written by me***.~~

* *cross out if not applicable*

Date

26.05.2025

Signature of the Author

** *each of the co-authors of the thesis submit the statement separately.*
*** *in the case of co-author of thesis, substantive and percentage contribution should be declared.*

Summary

This thesis compares traditional econometric and machine learning models for forecasting stock returns under diverse market conditions. A total of 69 U.S. stocks are selected by 6 key sectors, leadership status, and volatility level into 24 sections. The dataset covers 25 years of daily data from 2000 to 2025. Results show that non-linear models are dominating (best model for 73.9% of stocks), with LightGBM emerging as the best model for 53.6% of stocks, particularly for market leaders and stable stocks and 18 out of 24 sections. Econometric models perform best in 6 sections, especially within the energy sector. Key predictive features include Close to Open, 5-day Rolling Return, and rolling mean and standard deviation. The findings offer practical value for investors and analysts.

Key words

stock market, stock price prediction, machine learning, financial investment

Field of the thesis (codes according to the Erasmus program)

Economics (14300)

***The title of th*e thesis in Polish**

Analiza Zachowania Rynku Giełdowego i Przewidywania Cen
Na Rynkach Globalnych

TABLE OF CONTENTS

INTRODUCTION

Predicting stock prices is an interesting topic for both academics and investors due to its theoretical and practical assumptions. In terms of academics, stock price prediction is a broad field for testing financial theories, evaluating market efficiency, and developing advanced modeling techniques that integrate several fields such as economics, statistics, and machine learning (Sonkavde et al., 2023). For investors, accurate forecasts might impact investment strategies, risk management, and portfolio optimization. It is leading to improved returns and competitive advantage. The ability to forecast market movements in advance in a data-driven financial landscape can provide significant value which makes stock price prediction a critical and popular area of research and development, where new ideas and methods are constantly being explored (Mehtab & Sen, 2019).

The prediction of stock prices has always been one of the most challenging tasks in finance. This challenge is occurring due to the complexity and non-linear nature of market behavior. In recent years, the field has encountered a shift with the integration of advanced machine learning (ML) and deep learning (DL) techniques, which demonstrates high potential in financial forecasting (Phuoc et al., 2024; Vijh et al., 2020). These tools allow us to detect patterns in large, noisy, and high-dimensional datasets. The capabilities of these tools outperform many traditional financial models. Their growing application has been observed in different stock markets, including the New York Stock Exchange (Leigh et al., 2005), European Union (Ketsetsis et al., 2020), Turkish (Egeli et al., 2003), Saudi Arabian (Alotaibi et al., 2018), Chinese (Long et al., 2020), and Indian markets (Nayak et al., 2016), as well as globally (Lee et al., 2019). This popularity of applications reinforces the relevance of machine learning and deep learning approaches for global stock prediction.

With the recent popularity of machine learning, deep learning, and time-series analysis, the prediction of stock prices has entered the era of machine learning. These technologies offer advanced instruments capable of revealing patterns in massive datasets, by providing predictions with better accuracy rather than traditional financial models. The better prediction for stocks gives the greater power to make more informed decisions which leads in turn to superior risk management and improved profitability for traders, investors and policymakers (Mehtab & Sen, 2020).

Machine learning models such as decision trees, random forests, and gradient boosting machines have been widely explored for stock return forecasting. The application still faces limitations in interpretability and generalizability within different stock types (Sonkavde et al., 2023; Sheth & Shah, 2023). On the other hand, deep learning models such as LSTM, GRU, and CNN have proven superior performance in modeling temporal dependencies and non-linearities in stock prices (Huang et al., 2020; Yu & Yan, 2020). Despite these advantages, there remains a gap in understanding how these models perform when applied to individual stocks rather than indices. Most prior studies continue to focus on aggregated benchmarks such as the S&P 500, Dow Jones, or Nifty 50 (Mehtab & Sen, 2020; Ketsetsis et al., 2020; Vijh et al., 2020), which can mask firm-specific signals under market-wide trends.

In this study, the focus is specifically on machine learning (ML) and regularized econometric models rather than traditional time series models such as ARIMA or GARCH. Although these models have strong theoretical foundations and are commonly used for financial time series forecasting, they highly depend on assumptions such as stationarity, linearity, and pre-specified lag structures (Gourieroux & Jasiak, 2018; Brainard & Tobin, 1968). This dependence might not hold consistently through a large set of diverse stocks. Given the scope of this dissertation, we cover 69 companies across multiple sectors with different market conditions and volatility levels. The traditional time-series models would require individual tuning and diagnostics for each stock and it is making them computationally intensive and less scalable. In contrast, ML models like Random Forest and LightGBM offer flexibility in handling non-linearity, high-dimensional feature spaces, and heterogeneous behavior without strong distributional assumptions (Krennmair & Schmid, 2022; Ke et al., 2017). Similarly, deep learning models such as LSTM have shown promise in financial forecasting, where they demand extensive data preprocessing, careful architecture design, and longer training times. Due to the limited computational resources and the study's objective to balance performance with interpretability and efficiency, deep learning models are not included (He et al., 2023; Lara-Benítez et al., 2021). Instead, the selected ML and econometric models provide a practical and robust framework for comparative analysis across a wide range of stocks. As markets develop rapidly and become more sensitive to global events, investor sentiment, and firm-level innovations, then the ability to model price movements using flexible, adaptive algorithms becomes critical (Lo, 2004). This challenge has accelerated a change toward more computational and data-driven approaches. This shift is making ML not just an alternative but also an essential tool in modern financial modeling.

In this context, the primary objective of this dissertation is to analyze and compare the effectiveness of traditional econometric models against non-linear or advanced machine learning techniques in predicting stock prices (log return) by considering various company classifications such as sector, market leadership status, and volatility status. Unlike previous studies which primarily focused on enhancing specific model architectures or exploring sector-specific dynamics, this research aims to provide a comprehensive comparison of model performances across different financial sectors and market conditions. With this approach, we not only build forecasting models but also identify predictive capabilities in various market conditions. We apply a multitude of models to identify differences and implement comparisons to bridge the gap between simple and advanced computational models to better understand patterns in predicting stock prices. It explores how machine learning models perform in various market conditions, and emphasizes the respective strengths and limitations.

Research Questions:

Guided by the gaps and opportunities identified in the literature, this research is structured around three critical questions:

- *Which type of predictive model either linear or non-linear is most effective for forecasting individual stock returns across global sectors?*
- *How does predictive performance vary across different financial sectors (e.g., technology, energy, healthcare)?*
- *What is the impact of market capitalization (market leadership) and stock volatility on the accuracy of stock price predictions?*

These research questions display the primary aim of the paper which is to address both methodological and applied dimensions of financial modeling. The first question compares linear and non-linear machine learning models to identify which type of model is better suited to the complex behavior of stock returns. It is expected that non-linear models perform better than linear models because of the complex structure of stocks (Kanas & Yannopoulos, 2001). The second question emphasizes sectoral dynamics, recognizing that industry-specific trends, macroeconomic sensitivities, and regulatory environments might affect predictability. The third question dives deeper into structural firm-level characteristics. The company size (leadership) and volatility are used to evaluate whether large, and stable companies are easier to predict rather than small, and high-risk stocks. In summary, these questions guide a comprehensive investigation that blends data science, financial theory, and empirical market behavior.

Despite the popularity of AI models increasing in finance applications, only a limited number of studies have emphasized individual company-level modeling over broad market indices. Several researchers have recently considered this direction, investigating firm-specific behaviors and achieving promising results in stocks like Infosys, TCS, and ICICI (Chatterjee et al., 2021; Thakur & Kaur, 2023). Modeling at the individual stock level allows for a deeper understanding of volatility, market sentiment, and company-specific shocks. These factors are frequently lost when we work with aggregated index data. This thesis focuses on this idea and introduces a more granular, bottom-up modeling approach by selecting and analyzing 69 stocks from six key global sectors. In addition to sectoral categorization, the selection is also categorized by volatility and market leadership status to create 24 distinct company sections. This study focuses on the current leadership status and current volatility status of the companies. From an initial fetch of over 3,000 global stocks, 69 companies are selected using specific filtering and reliable random sampling. The companies selected have a market presence of at least 25 years. Daily stock data is used in modeling which includes period intervals from January 2000 to May 2025. The structured selection is used for not only the identification of the models that perform best overall but also which are most appropriate for specific types of stocks such as volatile non-leaders in the energy sector or stable market leaders in consumer staples. The selection of stocks is implemented based on 6 key sectors which are technology, energy, financials, industrials, consumer goods, and healthcare. In addition, stock selection is implemented according to volatility status and market leadership status. In this thesis, 2 traditional econometric models and 3 machine learning models are applied to define the best model in various market conditions. In terms of econometric modeling, linear and regularized linear models are used. For machine learning modeling, decision tree, random forest and light gradient boosting models are applied. The thesis seeks to provide both technical and practical contributions to financial forecasting and investment decision-making by linking modeling results with the economic characteristics of firms.

The thesis is structured in the following way – the next chapter gives a summary of the previous works in the form of a literature review. In addition, the gaps from existing research are presented which are addressed in this dissertation. As a next step, the methodology is implemented where the dataset is described in detail, along with information about the applied machine learning models. Besides, a list of evaluation metrics is provided with the characteristics. The evaluation metrics are provided to compare model performances. Subsequently, exploratory data analytics is conducted to gain a deep understanding of the stock market. Then, feature engineering is implemented to overcome potential issues encountered by simple machine learning models, which advanced machine learning models typically do not face. In the empirical results section, the results from all models are compared and connected to economic outcomes. Finally, in the conclusion and future research section, the summary of findings, beneficial information based on analysis, challenges and future research possibilities are described

The outcomes of this research are anticipated to significantly contribute to the fields of financial analysis and predictive modeling. The thesis's focus is to identify the most effective techniques for different market scenarios by providing a detailed comparison of various modeling approaches. The thesis aims to assist investors and financial analysts make better decisions. This paper is answering the gaps that are critical for advancing the understanding of complex market dynamics and developing the predictive performance of stock price models in real-world scenarios.

1. LITERATURE REVIEW

1.1. Introduction to Stock Market Analysis

The word *stock* in North American usage means ownership or equity in a corporation. Stock is issued in the form of shares of the companies which makes this financial instrument attractive in the eyes of individual investors. The sale of stocks has been implemented in primary and secondary markets as same as other securities. The initial sale of securities from the issuing organization to investors is called primary distribution. The secondary transaction happens when the investor obtains the shares from another investor or entity rather than an issued organization. For instance, when an investor purchases 100 shares of General Motors (GM) on the New York Stock Exchange (NYSE), the proceeds of the sale do not go to but rather to the investor who sold the shares (Teweles & Bradley, 1998).

The stock market is an essential part of the global financial system. The stock market acts as a public forum where company shares and derivatives are traded at agreed prices. It is frequently equated with economic prosperity and is viewed as a measure of national economic health. With a regulated framework of issuance and securities trading, the stock market creates conditions in which both public and private companies can raise equity capital without taking on substantial debt. This can improve liquidity and provide income, growth, and return of capital benefits to investors. To understand how the stock market works, we first need to understand how it influences the economy and individual wealth. For investors, the world of the stock market offers the potential for bearing substantial returns on investment via two broad streams, namely, dividends and capital gain. From an economic perspective, it aids in capital formation and economic growth. As an aggregate of investor sentiment, the market offers one of the most immediate measures of a nation's health and faith in its financial stability. Nonetheless, the stock market parentheses, which is a collection of stock prices, may mirror economic shocks and as such predicting decimals is a worthwhile but complex task. The significance of advanced data analytical approaches and techniques to financial applications like predicting stock prices and their associated effects is explored in this research, particularly as predictive accuracy can result in optimized investment strategies and improved market stability (Bosworth, Hymans, & Modigliani, 1975).

Lorie, & Hamilton (1985) argue that the stock market is not only a barometer of the economic climate but also a key staging ground for capital allocation, with implications for aggregate economic outcomes and individual wealth as well. They describe the inherent volatility of stock prices and their vulnerability to a multitude of factors including economic indicators, corporate earnings reports, and investor sentiment. Such vulnerability makes the stock market a complex but important factor within financial infrastructure, requiring powerful analytical tools to enable accurate predictions. Their economic commentary also serves to remind readers and investors alike, that the stock market is a two-edged sword whereby accelerating capital formation we unlock the potential for economic growth, however as investors, we also encounter risks and opportunities that can both reward and punish in the short to medium run. This is why their work emphasizes the need for sound statistical analyses and different forecasting models to predict the stock market.

Continuing the observations offered in the posts above, including those by Lorie and Hamilton, Werner F. De Bondt, a long-time observer of market prices and theories, turns to economists themselves and then to their theories about the stock market relative to what happens. De Bondt's examination of economic models in his 1991 paper challenges the assumptions behind economic predictions and market behavior, suggesting that traditional assumptions may falter when applied to stock market phenomena. He argues that economics, as a systematic study, sets the groundwork, but comprehending the prices of market transactions requires recognizing the unmapped irrationalities and speculative processes that underlie transactions. In its broader implications for economic theory and practice, the work of De Bondt advocates the need for economic theories to account for the complexities of human behavior, and to apply empirical methods that take into consideration the rarely contemplated, yet undeniably present, errors in the assumptions of rationality, information accessibility, and rational decision-making that lie at the core of many economic models. This also has very important consequences in the construction of financial models on stock trading that can fit the observed facts of financial markets (De Bondt, 1991).

The stock market is an essential sector of the global economy, providing companies with access to capital and offering investors the chance of a fair return. Given their complexities and multiple influential components, predicting stock price movements has long been a task for economists, investors, and researchers alike. Recent machine learning, deep learning, and time-series advancements provide methods to tackle these complexities in such a way that predictions of the stock price become increasingly accurate. However, much of the current research is centered on predicting individual or sector-based stocks, which reduces the generalization of these models across diverse sectors and market conditions (Kocaoğlu et al., 2022; Rouf et al., 2021; Sheth & Shah, 2023). Specifically, we have not studied the price (and market) behavior of market leaders, who are those key stocks that determine the directional trends of their respective sectors and how those price actions compare in volatile as opposed to stable sectors. In this thesis, we attempt to reduce this gap by forecasting the price of the stocks by using machine learning models. Analyzing how these models perform in different 6 sectors, considering also market leaders and non-market leaders, volatile and stable stocks.

The stocks are classified into different sectors and several sectors are grouped based on the company's service and working mechanism. The majority of the stocks in the stock market are from sectors such as technology, energy, financials, industrials, consumer goods, and healthcare. As an example of popular companies, Apple is a technology sector stock, where Schneider Electric is an industrial sector stock, and Coca-Cola is a consumer goods sector stock.

1.2. Classification of Stocks

Volatile vs. Stable Stocks: Volatile stocks show significant price fluctuations over short periods (Farmer et al., 2004; Plerou et al., 1999). Stock prices are often influenced by market sentiment, earnings surprises, macroeconomic events etc. On the other hand, stable stocks illustrate more consistent and predictable price movements. Stable stocks are commonly mature companies and big sectors. Volatility is normally measured using metrics such as standard deviation, beta, or average true range.

Market Leaders vs. Non-Leaders: Market leaders are mostly dominant companies within their industry. Marker leader companies are known for their strong financial performance, brand recognition, and high trading volume. Market leadership status in the stock market is typically associated with market capitalization. To be a market leader in the sector, it needs to have a few billion in market capitalization (Smith & Ocampo, 2025; Kollnig & Li, 2023). Non-leader companies include the majority of the companies which are small or medium, less established firms. This may cause higher growth potential but also more risk. Leadership status can influence investor decisions, risk tolerance, and overall stock behavior.

Sectors: The companies are classified into different sectors and several sectors are grouped based on the company's service and working mechanism. The majority of the stocks in the stock market are from sectors such as technology, energy, financials, industrials, consumer goods, and healthcare (Weerawarana et al., 2019; Kollnig & Li, 2023).

Unlike most comparative studies, this dissertation introduces a three-dimensional classification framework by combining volatility, leadership, and sector. By analyzing individual stocks within these categories, the study reveals detailed stock behavior by machine learning modeling.

1.3. Sectoral Differences

Sector Classification: This study considers six key sectors which are Consumer Staples, Energy, Financials, Healthcare, Industrials, and Technology. These sectors are covering a wide scope of economic activity. Each sector has unique characteristics, regulatory environments, and sensitivity to macroeconomic variables. These sectors are commonly recognized in standard classification systems such as the Global Industry Classification Standard (GICS) and Industry Classification Benchmark (ICB) (MSCI & S&P Dow Jones Indices, 2023; FTSE Russell, 2024).

Sector Performance: Sector performance depends on market cycles. For instance, Consumer Staples tend to be defensive during downturns, while Technology companies are acting as defensive during expansion (Bernstein, 2023). Energy is sensitive to geopolitical events and commodity prices (Bariviera et al., 2017), whereas Financials sector stocks correlate with interest rates and credit cycles (Palomino et al., 2019). The sector-based analysis provides detailed insights into stock performance but the broad indices cannot capture.

Stock indices are collections of defined stocks that display the overall performance of a specific part of the market or the entire market. Unlike individual companies, stock indices group multiple companies into a single value to control and track how that group is performing over time. For instance, one of the most famous indices the S&P 500 consists of 500 largest United States companies and these companies are mostly used as a benchmark for the American stock market. The indices assist investors in understanding general market trends, comparing investment returns, and making decisions without analyzing every single stock. However, an individual company's stock shows the unique performance, unique risks, and unique opportunities of the business. Individual stock analysis may provide deeper insights and allow for more precise modeling rather than indices. Single stock is better especially when we aim to investigate factors such as volatility, sector influence, or market leadership (Brealey, 2000).

Why Individual Companies Over Indices?

Indices (e.g., S&P 500, NASDAQ-100) are weighted aggregations of many stocks and the indices provide a high-level market overview (McKillop et al., 2020).

Individual Stocks have more variance, granularity, and modeling challenge which is essential for deep learning and Machine learning applications (Tsantekidis et al., 2017).

As previous studies show, individual stock prediction captures firm-level dynamics better, specifically in volatile environments (Fischer & Krauss, 2018). This research targets micro-level signals by analyzing individual stocks and avoids the smoothing effect of index-level aggregation.

1.4. Machine Learning (ML) and Deep Learning (DL) Models in Stock Market Understanding

Introduction to ML and DL: Machine learning and deep learning are powerful tools to identify patterns in financial data. Machine learning algorithms such as Random Forest and XGBoost are popular for their robustness in handling non-linearities, while deep learning architectures such as Long Short-Term Memory (LSTM) and convolutional neural network (CNN) may capture sequential and spatial dependencies in time series data (Alsharef et al., 2022).

Stock price forecasting has relied on classical statistical and econometric models such as linear regression, autoregressive (AR) models, and ARIMA. These methods use historical price data to estimate future values and are valued for their simplicity and interpretability. However, they are based on assumptions of linearity and stationarity. These conditions are often violated in real-world financial markets. Despite these limitations, these models have provided the foundation for more advanced techniques and continue to serve as useful benchmarks in predictive research (Song et al., 2014).

While traditional models offer analytical clarity, they often fail to capture key characteristics of financial data such as non-linear relationships, volatility, and sudden market shifts. Due to these challenges, researchers have increasingly applied machine learning methods capable of modeling more complex and dynamic patterns. Techniques such as support vector machines, random forests, and recurrent neural networks have demonstrated improved performance, especially in datasets with high noise and dimensionality (Feng et al., 2018). These methods also offer the flexibility to incorporate a wide range of structured and unstructured financial variables.

In addition to increasing predictive power, recent studies have explored the application of deep learning models such as deep neural networks (DNNs), which are well-suited for extracting hierarchical features from time series data. Nikou et al. (2019) compared deep learning algorithms with classical machine learning methods and found that deep models delivered higher accuracy and better generalization, particularly in volatile market environments. Their research demonstrated that deep learning not only improves forecast precision but also provides more robust performance across varying financial conditions, making it a compelling alternative for modern stock price prediction tasks.

Among deep learning models, Recurrent Neural Networks (RNNs) have become one of the most widely used architectures for stock price prediction due to their ability to process and retain sequential information over time. Unlike feedforward networks, RNNs maintain internal memory through feedback loops, making them well-suited for time series models such as daily stock returns or market indices. Their enhanced temporal structure allows them to detect complex patterns and dependencies in historical price movements. Studies such as those by Fischer and Krauss (2018) and Karim et al. (2020) have shown that RNN-based models especially when combined with techniques like LSTM and GRU consistently outperform traditional models in forecasting tasks. These findings highlight the practical strength of RNNs in capturing long-range dependencies and adapting to the dynamic nature of financial markets.

In the field of financial predictions, several innovative approaches also have been explored to enhance the accuracy and efficiency of forecasting models. Li and Shi (2025) introduced a hybrid preprocessing method that combines Empirical Wavelet Transform (EWT) with Dynamic Time Warping (DTW) to refine neural network outputs, offering significant improvements in predictive accuracy by effectively handling different frequency components of stock price data. Similarly, Chang et al. (2025) developed a novel market-embedding model utilizing Mamba (MEM) architecture, which integrates dynamic stock correlations into traditional models, thereby enriching the feature set and enhancing prediction capabilities.

Expanding on neural network innovations, Li et al. (2025) applied a GAN-LSAT-Attention model that exploits the strengths of generative adversarial networks, LSTM, and attention mechanisms to provide a robust predictive framework for U.S. stocks. Kristianti et al. (2024) focused on the transportation industry and utilized LSTM models to explore the impacts of macroeconomic factors on stock prices, which emphasize the role of external economic variables. In addition, Ghallabi et al. (2025) assessed the influence of Environmental, Social, and Governance (ESG) factors on clean energy stocks using advanced machine learning techniques, which highlights the increasing relevance of geopolitical and environmental factors in financial modeling. Unlike, current literature, this dissertation does not focus on specific indices or specific markets. The thesis's core focus is on the general stock market and addressing key research questions:

- Is there a single type of machine learning algorithm that performs best for all dimensions of the analysis?
- How predictive accuracy varies across sectors?
- How predictive performance is influenced by stock volatility and market leadership?

1.5. Previous Studies and Findings

If we review of the previous researches, we observe some important points which are described in a short and detailed way:

- Fischer & Krauss (2018) used LSTM (deep learning model) to outperform machine learning models on S&P 500 stock data.
- Patel et al. (2015) compared distinctive machine learning models such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Raandom Forest (RF), and Logistic Regression on Indian stock market. They improved prediction accuracy in their analysis.
- Tsantekidis et al. (2017) applied deep learning models on limit order book data and they proved the effectiveness of CNN model in intraday prediction.

1.6. Investor Behavior and Study Motivation

Investors are people who deploy their funds to obtain returns. The investors have a multitude of options where to put their money. They can invest in bonds, commodities, and physical assets like real estate, private companies, and start-ups or they can invest in the stock market. There are some factors affecting the investor while the investment process such as the risk profile of the investor, expectations of return, and they're also macroeconomic factors affecting such as political accidents (Shanaev & Ghimire, 2019), environmental problems (Guo, Kuai, & Liu, 2020), sport events (Harjito, Alam, & Dewi, 2021), news (Baker, Bloom, Davis, & Kost, 2019), disasters (Barro & Liao, 2021), pandemics (Wang, Yang, & Chen, 2013), and other types of uncertainty (Al-Awadhi et al., 2020). For instance, the COVID-19 pandemic influenced the equity market a lot since 2020 (Zhang, Hu, & Ji, 2020). The pandemic has caused a significant drop in the confidence of investors in the stock market because of increased uncertainty (Liu et al., 2020). The strongest shock in the stock market was observed in the first stage of the pandemic (Hassan & Riveros Gavilanes, 2021).

The majority of the investors choose to invest their funds in physical assets. Every corporation desires to get a huge amount of investment. The businesses usually benefit from capital infusion to fund their operations, and expansion plans, and to enable CAPEX (capital expenses). These companies either can go to the bank to request money or issue bonds, however, banks usually charge interest and there is a limit to the level of leverage that a company has on its balance sheet. Current time although the interest rates are low, there is still a risk for companies to be overleveraged and therefore be at risk. The other option is infusing the equity for public companies listed on the stock market. This means reaching out to the public for investments in exchange for a part of the ownership of the company. As an example, Schneider Electric is a publicly listed company, and it is traded on the French stock market. Schneider funds are part of the CAC40 which is a capitalization-weighted measure of the 40 most significant stocks in France. Presently, Schneider Electric ranks 12 among the most valuable 40 French companies. In 2003 Market cap of the Schneider was just $10 billion, in 2008 it became $16 billion, and in 2020 it became $58 billion. This shows massive growth in the wealth of the company. This proves that finding small companies with the opportunity to grow might lead to high earnings.

The investment strategy is not only associated with the understanding stock market but also the investor thinking. There are different types of investors based on their risk preferences but they are divided into 3 core groups:

- ◆ **Risk-Averse**: They prefer low-volatility, and stable returns. They frequently opt consumer staples and financial sector stocks (Markowitz, 1952).
- ◆ **Risk-Neutral**: They focus on expected returns without considering risk. They can use algorithmic models to diversify exposure.
- ◆ **Risk-Loving**: They love high volatility and significant gains. They often target small company stocks which has more potential to grow faster and emerging tech (Barberis et al., 2001).

This thesis helps investors to understand how different models perform under various stock conditions. It shows which models perform better in volatile stocks and small companies. So, it assists investors in determining strategy with risk preferences.

1.7. Key Factors Influencing Stock Prices

Stock prices change because of a variety of factors, however, they are primarily influenced by supply and demand in the stock market. In Table 1, the key factors are provided that influence stock price and cause price changes.

Table 1. Main factors influencing stock prices

Factors	Details
Company Performance	A company's stock price depends on its financial health, earnings, and future growth. Strong earnings, innovative products, or positive announcements can cause an increase in stock prices, while poor performance, missed expectations, or bad news can cause the price to drop directly (Avdalovic & Milenković, 2017).
Market Sentiment	Investors' feelings regarding the general market, specific sectors, or specific classes may significantly impact stock prices. Positive sentiment can drive prices up, while factors such as fear, uncertainty, or negative news might cause prices to decrease (Allen, McAleer, & Singh, 2019).

Economic Indicators	Economic indicators such as interest rates, inflation, GDP growth, and unemployment rates, might influence stock prices. For example, increasing interest rates often lead to lower stock prices because borrowing costs enhance and consumer spending may decrease (Aylward & Glen, 2000).
Supply and Demand	Stock prices change quickly and frequently based on the number of people wanting to buy (demand) versus the number wanting to sell (supply). If more people want to buy a stock, then the price rises. İn contrast, if more people want to sell, the price falls (Janor et al., 2010).
Market Speculation	Other drivers such as news, rumors, and investor speculation can also cause stock prices to move up or down. Although, if there is no immediate notable change in the company's situation, these shocks may affect later. For example, rumors of an acquisition or new product launch can send stock prices up, while fears of regulatory changes or legal troubles can cause prices to fall (Andreasson et al., 2016).
Global Events	Political instability, natural disasters, and geopolitical tensions can also affect stock prices. For instance, trade wars, changes in government policies, or crises like pandemics can create uncertainty and cause volatility in the stock market (Niederhoffer, 1971).
Industry Trends	Certain industries can experience fluctuations based on trends in the market. For instance, if the price of oil rises, energy stocks might go up, or if a technology company announces a breakthrough innovation, its stock price could surge (King, 1966).
Investor Behavior	Individual and institutional investors buying or selling stocks can create price movements. Large-scale buying can drive a stock's price up, while mass selling can have the opposite effect (Shiller, 1990).

Source: Own elaboration.

Eight main shocks might affect the stock market and influence the stock prices directly (Tabash et al., 2024; Cieslak & Pang, 2021). These shocks consist of company performance, market sentiment, economic indicators, global events etc. In this dissertation, the shock periods such as the 2008 Global Financial Crisis, the 2020 COVID-19 pandemic and other shocks are included in the dataset which makes models more resilient and reliable regarding these shocks.

It aims to apply both linear and non-linear machine learning models to investigate which types of models perform better through different groups of company stocks. Unlike previous researches, which are mostly focused on a predefined set of companies or major stock indices and identify the best-performing model based on those indices (Teixeira & Barbosa, 2024; Phuoc et al., 2024). Nevertheless, this thesis introduces a more detailed and structured approach where the companies are carefully chosen for the analysis which show a wide variety of market characteristics. Especially, the selection of stocks is based on three important classification criteria: sector, leadership status, and volatility level of each company. In the dissertation, these classifications are defined based on the current situation (2025 April) of companies which also makes this paper worthy and novel.

This detailed classification of stocks allows for a more detailed evaluation of model performance under different market conditions. The study not only identifies the most suitable models for each category but also provides valuable insights into stock behavior by examining how model effectiveness changes across these defined groups. This approach makes the thesis unique and different than the previous studies. It helps to fill a gap in the existing literature. The findings of this research might support more informed decision-making for both investors and policymakers by providing a deeper and detailed understanding of the stock market's complexity and the changing dynamics across company types.

2. DATA AND METHODOLOGY

2.1. Data

Data is collected, selected, and divided into groups according to the sectors, market leadership, and volatility level. After proper division then prepared for the modeling.

2.1.1. Data Collection

For this research, the data is carefully gathered from two main sources, each chosen for their specific strengths in providing essential financial information. The first source, stockanalysis.com, is selected for its comprehensive and easily navigable inventory of stocks categorized by industry. This platform provided not only the list of stocks but also the latest market capitalization. The number of retrieved stocks is 3225.

The 3225 stocks are U.S.-listed stocks which are retrieved from StockAnalysis.com. This source is a widely used financial platform that tracks U.S. stocks traded on major exchanges such as NYSE, NASDAQ, and NYSE American. The stocks retrieved from this source belong to the Consumer Staples, Energy, Financials, Healthcare, Industrials, and Technology sectors. These stocks were chosen based on their market capitalization as of January 26, 2025. For each company, daily historical stock price data is used, which allows for detailed time-series modeling and prediction. The use of daily frequency enables the models to capture short-term fluctuations, trends, and volatility dynamics more effectively. However, it is critical to note that StockAnalysis.com does not list all publicly traded U.S. companies. While the full U.S. market includes over 5,000 listed stocks, StockAnalysis only displays a portion of that. The stocks in this source are mostly mid to large-cap companies and do not include many micro-cap or over-the-counter (OTC) firms. So, the initial dataset excludes some sectors entirely, such as Consumer Discretionary, Utilities, Real Estate, Communication Services, and Materials, meaning the coverage is large but not complete.

A list of stocks from the source StockAnalysis represents the majority of stocks from popular indices such as the S&P 500 and Russell 3000. The S&P 500 displays 500 the largest, most stable U.S. companies and is commonly used to reflect the overall U.S. market's performance. The Russell 3000, on the other hand, includes 3,000 of the largest U.S. companies by market capitalization and is designed to cover approximately 98% of the investable U.S. equity market. While StockAnalysis.com includes all stocks from the S&P 500 and the majority of the Russell 3000, it does not include the full list of smaller-cap or newer firms. Therefore, while 3,225-stock selection includes a significant part of these indices, especially in large and mid-cap segments, it does not fully reflect the entire market universe, particularly in terms of sector balance and company size.

The dataset used in this research is large, diverse, and highly useful for model evaluation, especially because it includes key sectors and companies with varying characteristics. However, due to the exclusion of several sectors and smaller firms, the findings should be understood within this context. The study's results are relevant and valid for the selected stocks and sectors, but they are not fully generalizable to all U.S. publicly traded companies. It serves as a practical sample for understanding stock behavior across leadership, volatility, and sector dimensions using machine learning techniques because the dataset captures a large portion of actively traded and economically significant companies.

Table 2. List of initially collected stocks

No	Sector	The number of stocks
1	Consumer Staple	245
2	Energy	250
3	Financial	603
4	Healthcare	1164
5	Industrials	453
6	Technology	510

Source: Own elaboration.

In Table 2, we observe the selected stocks by defined 6 sectors and there is an imbalance in the initial selection of stocks where the healthcare sector is represented by 1164 stocks however, consumer staples are only 245 stocks. The imbalance in the initial selection of stocks does not have any impact on the results and analysis.

The website stockanalysis.com not only offers an organized overview of stocks segmented by sector but also the market capitalization value of each company which is not available on other platforms like Yahoo Finance, therefore, this source is considered in the analysis. It facilitates targeted data retrieval and it is particularly beneficial for sector-specific analysis.

Moreover, the research uses Yahoo Finance to access a detailed dataset that includes daily trading volumes, adjusted closing prices, and other financial indicators for the last 25 years.

This extended timeframe is chosen to capture multiple economic cycles and ensure a comprehensive view of long-term trends and behaviors in the stock market. The daily data frequency allows the analysis to reflect how markets respond in real time to global events or economic shocks, which is essential for identifying investment opportunities and potential risks. The companies selected from Yahoo Finance are the exact stocks previously identified through StockAnalysis.com. The modeling and analysis are conducted only for the 3,225 U.S.-listed companies gathered from StockAnalysis.com. This list covers nearly the entire S&P 500 index (100%) and the majority of the Russell 3000 index (85–90%), particularly mid- and large-cap companies. While the dataset excludes some micro-cap and sector-specific stocks, it remains broadly representative of the U.S. market. This level of coverage strengthens the reliability of the analysis and ensures that the findings reflect real market dynamics across a wide range of industries and company sizes. It serves as a consistent reference point throughout the study.

2.1.2. Strategic Sector Selection

Investor normally groups the stocks based on sectors and there are 11 defined sectors (described in stockanalysis.com) in the market. Each of the 11 sectors also contains subgroups that provide more specific insight into a company's operations, services, or industry focus.

Table 3. Classification of sectors in stock market

No	Sector	The number of subcategories/industries
1	Communication on services	7
2	Consumer discretionary	23
3	Consumer staples	12
4	Energy	8
5	Financials	15
6	Healthcare	11
7	Information technology	12
8	Materials	14
9	Real estate	12
10	Utilities	6
11	Industrials	25

Source: Own elaboration.

Table 3 shows the number of subcategories (industries) in each of the 11 sectors. There are a total of 145 industries for 11 sectors. Discussing and analyzing the industries is beyond the scope of this dissertation. Even though, companies are in different industries, if they are inside the same sector they are competitors. They compete with each other to obtain more investment and attract investors.

It is important to observe the sectors and other classifications of the selected companies based on their volatility level and market capitalization. The six major impacted sectors are selected. The selected sectors are Technology, Healthcare, Energy, Consumer Goods, Finance, and Industrials. These sectors are selected because they might affect the economy and behave differently in reaction to certain market conditions. These sectors provide sample data to understand stock price movements. In other words, selecting these specific sectors allows for a focused analysis of areas with high economic influence and distinct market dynamics, rather than making weaker findings across all sectors where some may have minimal impact or lack distinct behavioral patterns for meaningful analysis. This targeted approach enhances the efficiency and relevance of the research outcomes (Wielechowski & Czech, 2022).

2.1.3. Market Capitalization and Market Leadership

Market capitalization serves as a primary criterion for categorizing companies and understanding their market influence. Two factors usually applied for the estimation of a company's market capitalization are the share price and the number of outstanding shares. The number of shares outstanding means the total holdings of a shareholder construct which includes shares owned by institutional investors and restricted stock held by company insiders, while the current share price refers to the last trading price on the stock market (Dias, 2013).

$$\text{Market Cap} = \text{Number of Shares Outsdanding} * \text{Current Share Price} \qquad (1)$$

In Equation 1, the formula of market capitalization is provided which is the multiplication of share price and number of shares. The ability to access updated market capitalization data is vital for accurately classifying companies such as Mega Cap, Large Cap, Mid Cap, and Small Cap (Suarez, 2016). For this analysis, the companies are also segmented into four categories based on their market capitalization.

Table 4. Company segmentation by market capitalization

Capitalization level	**Market value**
Mega Cap:	Over $200 billion
Large Cap:	$10 billion to $200 billion
Mid Cap:	$2 billion to $10 billion
Small Cap:	Below $2 billion

Source: (Suarez, 2016)

Table 4 shows the groups of companies according to their capitalization levels. This segmentation helps to differentiate the scale of companies' operations and their potential market influence, which is expected to correlate with different stock price volatility and predictive model performance. Market leaders are identified as companies falling into the mega Cap category. These firms are typically industry giants with substantial market share, stability, and influence over market trends. Identifying market leaders allows us to focus on stocks that not only drive sector performance but also provide a benchmark for comparing the efficacy of predictive models in high-stakes environments.

As it is determined, the market capitalization data used in this study is retrieved on January 26, 2025, from StockAnalysis.com. Given the relatively stable nature of market capitalizations, especially among larger firms, the minor fluctuations in company market caps over a few months are unlikely to significantly impact the validity of stock data analyses extending through April. This ensures that the initial categorization of companies into Mega, Large, Mid, and Small capitalization segments remains accurate and relevant for the duration of the study.

Table 5. Companies classified as market leaders (Market cap ≥ $200 Billion)

No	**Sector**	**The number of Market Leaders**	**The partition of market leaders**
1	Consumer Staple	5	2.04%
2	Energy	3	1.20%
3	Financial	8	1.33%
4	Healthcare	12	1.03%
5	Industrials	0	0.00%
6	Technology	14	2.75%

Source: Own elaboration.

In Table 5, the number of market leaders is listed when we set the threshold as $200 billion which includes companies in Mega cap. If we set a threshold of 200 billion USD to define market leaders, the threshold is insufficient for our analysis. This limitation is revealed because of certain sectors such as industrials, energy, consumer staples and so on…. It is obvious from Table 5 that the majority of sectors lack market leaders in the defined threshold. The market leaders in all 6 sectors are the interval of 0%-2.75%. The mega-cap has a very high threshold level hence, it is hard for companies to surpass this border. It is better to compare market leaders and non-leaders to take a large scale and redefine the border between non-market leaders and market leaders. Although a large cap has a 20 times lower threshold rather than a mega-cap which is 10 Billion USD, still it is changeable to be in this group.

Table 6. Companies classified as market leaders (Market cap ≥ $10 Billion)

No	Sector	The number of Market Leaders	The partition of market leaders
1	Consumer Staple	53	21.63%
2	Energy	50	20.00%
3	Financial	134	22.22%
4	Healthcare	85	7.30%
5	Industrials	114	25.17%
6	Technology	150	29.41%

Source: Own elaboration.

In Table 6, the number of market leaders is described in the $10 billion threshold. The redefined threshold of market leadership status as 10 billion USD which includes stocks classified as large-cap, and mega-cap causes a significant increase in the number of market leaders from around 2% to over 20%. This adjustment improves our scope of analysis by providing a more comprehensive dataset, which allows for detailed and deeper market insights and a more balanced sectoral comparison.

2.1.4. Volatility Analysis

Volatility is a famous issue in economic and financial research which is one of the most important characteristics of financial markets. Volatility is related to market uncertainty directly and influences the enterprises' investment behavior and individuals. Stock price volatility is a high degree of fluctuation in a stock's price in a specific period. High volatility displays significant increases and decreases in stock prices. This might result from different factors such as market sentiment, company performance, economic news, or external events like geopolitical tensions. On the other hand, stock price stability shows the consistency of a stock's price. In stable stocks, smaller fluctuations and less dramatic changes are typically recorded. Stable stocks are well-established companies or sectors with steady earnings, low risk, and less sensitivity to market changes. Investors may prefer stability in uncertain economic environments, while those seeking higher returns may be more attracted to volatile stocks (Bhowmik & Wang, 2020).

Understanding the balance between volatility and stability is crucial for making informed investment decisions. Another important problem in modern financial research is the volatility of financial asset returns and this volatility is frequently described and measured by the variance of the rate of return. However, forecasting perfect market volatility is difficult work and despite the availability of various models and techniques, not all of them work equally for all stock markets (Liu & Gupta, 2022). It is for this reason that researchers and financial analysts face such complexity in market returns and volatility forecasting. In this dissertation, volatility serves as a vital measure, which captures the degree of fluctuation in stock prices over time. Annualized volatility is specifically employed to calculate the current risk level of the stock. This methodology prioritizes recent performance to align with current market conditions, thereby ensuring that the volatility assessments accurately reflect the most recent economic dynamics critical for predictive modeling.

$$Annualized\,Volatility = \sigma * \sqrt{252} \qquad (2)$$

In Equation 2, sigma shows the standard deviation of the daily returns which shows the volatility. After defining the volatility, daily volatility is annualized by scaling it by the square root of the number of trading days in a year (usually around 252). The mean of annualized volatility of all stocks in the last 1 year is 0.6569 for all stocks (more than 3000) and 0.4166 for 967 stocks that are in the stock market at least in 2000, indicating moderate fluctuations in stock prices. This volatility score shows there are some fluctuations on average in stock prices but not extreme changes over the last 12 months. Also, it is observed that the companies which are existed in the stock market for more than 25 years are more stable than the newly entered companies. Investors should consider this volatility level when deciding to invest in the stock market. Hence, year-annualized volatility is opted to exhibit the current behavior of stock price movements. The division for market leadership is also based on current period market capitalization, so using recent volatility ensures alignment with today's market conditions.

For data categorization based on volatility levels, stocks are first, divided into 3 groups. First-group stocks were labeled as volatile and had annualized volatility scores above 0.5. Companies with a volatility score below 0.2 are considered stable, and those with volatility levels between 0.2 and 0.5 are classified as moderate. However, the results of this categorization were not sufficient for the aim of the paper. There was no Industrial non-leader stable stock with these thresholds, and only one stock was in each of the following groups: the Technology Sector (non-leader, stable), the Energy Sector (leader, volatile), and the Energy Sector (non-leader, stable). The categories are simplified into two groups: stable and volatile to improve the clarity and distribution of the analysis. The threshold is finally adjusted to define stable stocks as volatility of 0.25 and below, and volatile stocks as above 0.25 to ensure at least two stocks in each category.

Table 7. Overview of stocks after classification by market leadership and volatility

Sector	Market leadership status	Volatility class	Stock count
Consumer Staple	No	Stable	3
Consumer Staple	No	Volatile	45
Consumer Staple	Yes	Stable	24
Consumer Staple	Yes	Volatile	14
Energy	No	Stable	2
Energy	No	Volatile	52
Energy	Yes	Stable	7
Energy	Yes	Volatile	23
Financial	No	Stable	11
Financial	No	Volatile	161
Financial	Yes	Stable	22
Financial	Yes	Volatile	47
Healthcare	No	Stable	11
Healthcare	No	Volatile	128
Healthcare	Yes	Stable	10
Healthcare	Yes	Volatile	20
Industrials	No	Stable	6
Industrials	No	Volatile	139
Industrials	Yes	Stable	20
Industrials	Yes	Volatile	56

Technology	No	Stable	1
Technology	No	Volatile	97
Technology	Yes	Stable	13
Technology	Yes	Volatile	55

Source: Own elaboration.

Table 7 is provided to show the number of stocks after defining the groups. This classification system is crucial not only for model differentiation but also for validating the selection criteria used in the study. These thresholds as effective for the analytical framework. This involves comparing the predictive performance of models across different categories to observe if and how they influence prediction accuracy. This validation provides that the classification scheme is not arbitrary but is grounded in demonstrable and reproducible results that support the methodological choices made in this research.

2.1.5. Methodological Justifications

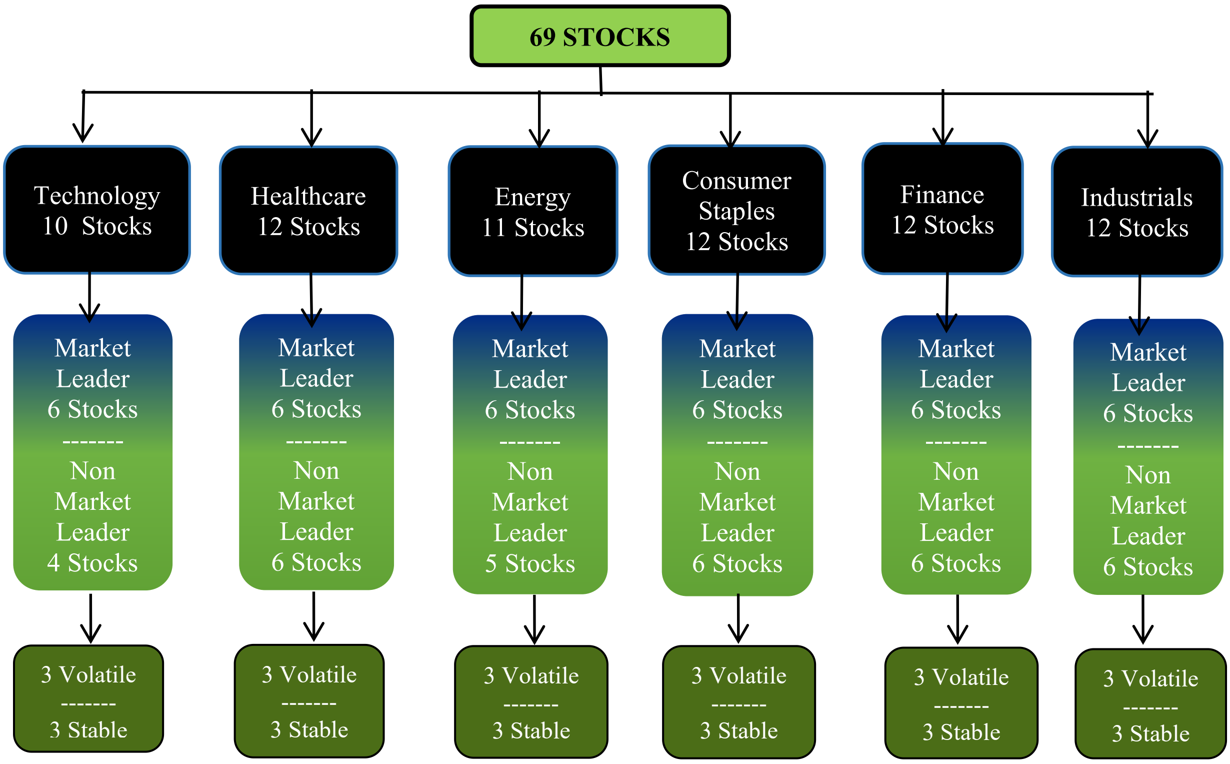
The selection of 25 years of daily data within the interval 2000 to 2025 is justified by the need to cover multiple business cycles, providing a comprehensive view of long-term trends and short-term fluctuations. This extensive dataset allows for robust statistical analysis and enhances the predictive accuracy of the models by training them on diverse market scenarios. Daily data is crucial for capturing the full spectrum of market volatility and immediate reactions to economic events, which are essential for developing responsive predictive models.

Moreover, focusing on sector-specific data ensures that the models are tuned to the unique characteristics of each industry, potentially enhancing their predictive power. By comparing model performances across different sectors, we can identify sector-specific factors that influence stock price movements and adjust the modeling approaches accordingly. Additionally, by differentiating between market leaders and non-market leaders, as well as volatile and stable stocks within these sectors, we delve deeper into the market dynamics. This detailed approach helps us understand the detailed behaviors of stocks more deeply, allowing for more targeted and effective predictive modeling. This wide analysis not only clarifies our understanding of sector impacts but also the distinct influences of market position and stability on stock performance. This section generally outlines the systematic approach to selecting sectors, categorizing companies by market capitalization, and by volatility level, all of which underpin the methodological framework of the study. These elements are essential for ensuring that the research is grounded in practical, real-world market dynamics and is capable of generating actionable insights into stock price predictions.

2.1.6. Final data selection

In this study, a diverse and balanced sample of stocks is selected to provide a deep and structured analysis of different market behaviors. The selection process began by categorizing companies based on sector, market leadership status, and volatility level. As a first step, only companies are opted which have been present in the stock market for at least 25 years. After shortlisting the stocks, the number of stocks decreased from over 3000 to 967. In the following, a pool of 967 U.S.-listed stocks, each of the six sectors is represented by 12 carefully chosen companies: 6 market leaders and 6 non-leaders. Within each of these groups, companies were further classified as stable or volatile, resulting in 3 stocks per subgroup. This structure is intentionally designed to allow for consistent and meaningful comparison across all classification dimensions. Selecting at least three stocks per subgroup enables the application of techniques like majority voting and ensures a more robust assessment of whether linear or non-linear models perform better in various market contexts.

Figure 1. Stock selection framework


69 STOCKS
Technology 10 Stocks
Healthcare 12 Stocks
Energy 11 Stocks
Consumer Staples 12 Stocks
Finance 12 Stocks
Industrials 12 Stocks
Market Leader 6 Stocks ------- Non Market Leader 4 Stocks
Market Leader 6 Stocks ------- Non Market Leader 6 Stocks
Market Leader 6 Stocks ------- Non Market Leader 5 Stocks
Market Leader 6 Stocks ------- Non Market Leader 6 Stocks
Market Leader 6 Stocks ------- Non Market Leader 6 Stocks
Market Leader 6 Stocks ------- Non Market Leader 6 Stocks
3 Volatile ------- 3 Stable
3 Volatile ------- 3 Stable
3 Volatile ------- 3 Stable
3 Volatile ------- 3 Stable
3 Volatile ------- 3 Stable
3 Volatile ------- 3 Stable

Source: Own elaboration.

In Figure 1, the schema illustrates the structured selection of stocks, clarifying the distribution process across different sectors and categories. This visualization aids in understanding how the 69 stocks are systematically categorized based on sector, market leadership, and volatility. If we see 12 stocks from each 6 sectors makes a total of 72 stocks, however, some sections do not include 3 stocks. Thus, 10 stocks are selected in the technology sector and 11 stocks are selected for the energy sector. The problems occurred in (2 stocks in technology-non-leader-stable) and (1 stock in energy-non-leader-stable) 3 stocks which caused to selection of 69 stocks instead of 72 stocks. In total, there are 24 distinct combinations based on sector, market leadership status, and volatility level. Three companies are selected from each group using a random sampling method with a fixed seed to ensure reproducibility. This process follows the principle of stratified random sampling, a widely accepted statistical technique that ensures balanced representation across different categories. This selection process is conducted randomly to avoid any biases, to provide a balanced representation across different market conditions, and to ensure a fair and representative sample for further modeling. The rationale behind this approach is to capture a broad spectrum of market dynamics and to understand how different types of stocks behave under various economic scenarios.

Within each of the 24 sections, 3 stocks are randomly selected by using a fixed random seed to ensure reproducibility. The selection process followed a uniform probability distribution across all eligible stocks in each group.

This means while the selection is random, the seed made it deterministic and reproducible (Toomet & Henningsen, 2008). Three companies were selected for all sections without any problems because all of the sections are made up of more than 3 stocks. This approach ensures avoiding bias by not favoring any specific firm, size, or market position. It is allowed to maintain balance across the full impact of the dataset, which is important for robust modeling and generalizable conclusions.

Since the selections are made from already meaningful subgroups, and the randomization is controlled, the resulting sample avoids both overrepresentation and underrepresentation. It also keeps the diversity of the market, sectors, and volatility patterns, which is critical for any machine learning or econometric modeling that aims to uncover general trends.

Table 8. Final list of selected 69 companies

No	Sector	Leadership status	Volatility status	Company Name	Symbol
1	Consumer Staples	Non-Leader	Stable	Ingredion Inc.	INGR
2	Consumer Staples	Non-Leader	Stable	Tootsie Roll Industries	TR
3	Consumer Staples	Non-Leader	Stable	Flowers Foods Inc.	FLO
4	Consumer Staples	Non-Leader	Volatile	SunOpta Inc.	STKL
5	Consumer Staples	Non-Leader	Volatile	J&J Snack Foods Corp.	JJSF
6	Consumer Staples	Non-Leader	Volatile	G. Willi-Food International Ltd.	WILC
7	Consumer Staples	Leader	Stable	Tyson Foods, Inc.	TSN
8	Consumer Staples	Leader	Stable	Church & Dwight Co., Inc.	CHD
9	Consumer Staples	Leader	Stable	General Mills, Inc.	GIS
10	Consumer Staples	Leader	Volatile	Archer-Daniels-Midland Co.	ADM
11	Consumer Staples	Leader	Volatile	The Estée Lauder Companies Inc.	EL
12	Consumer Staples	Leader	Volatile	The Hershey Company	HSY
13	Energy	Non-Leader	Stable	National Fuel Gas Company	NFG
14	Energy	Non-Leader	Stable	Sabine Royalty Trust	SBR
15	Energy	Non-Leader	Volatile	North European Oil Royalty Trust	NRT
16	Energy	Non-Leader	Volatile	Centrus Energy Corp.	LEU
17	Energy	Non-Leader	Volatile	Riley Exploration Permian, Inc.	REPX
18	Energy	Leader	Stable	Shell plc	SHEL
19	Energy	Leader	Stable	TotalEnergies SE	TTE
20	Energy	Leader	Stable	Exxon Mobil Corporation	XOM
21	Energy	Leader	Volatile	EOG Resources, Inc.	EOG
22	Energy	Leader	Volatile	The Williams Companies, Inc.	WMB
23	Energy	Leader	Volatile	Valero Energy Corporation	VLO
24	Financial	Non-Leader	Stable	Enstar Group Limited	ESGR
25	Financial	Non-Leader	Stable	White Mountains Insurance Group, Ltd.	WTM
26	Financial	Non-Leader	Stable	RLI Corp.	RLI
27	Financial	Non-Leader	Volatile	First Busey Corporation	BUSE
28	Financial	Non-Leader	Volatile	ConnectOne Bancorp, Inc.	CNOB
29	Financial	Non-Leader	Volatile	Globe Life Inc.	GL
30	Financial	Leader	Stable	Royal Bank of Canada	RY
31	Financial	Leader	Stable	Aflac Incorporated	AFL
32	Financial	Leader	Stable	The Bank of New York Mellon Corporation	BK
33	Financial	Leader	Volatile	U.S. Bancorp	USB
34	Financial	Leader	Volatile	ING Groep N.V.	ING

35	Financial	Leader	Volatile	BlackRock, Inc.	BLK
36	Healthcare	Non-Leader	Stable	Quest Diagnostics Incorporated	DGX
37	Healthcare	Non-Leader	Stable	Utah Medical Products, Inc.	UTMD
38	Healthcare	Non-Leader	Stable	AptarGroup, Inc.	ATR
39	Healthcare	Non-Leader	Volatile	Eterna Therapeutics Inc.	ERNA
40	Healthcare	Non-Leader	Volatile	PetMed Express, Inc.	PETS
41	Healthcare	Non-Leader	Volatile	ResMed Inc.	RMD
42	Healthcare	Leader	Stable	Abbott Laboratories	ABT
43	Healthcare	Leader	Stable	Medtronic plc	MDT
44	Healthcare	Leader	Stable	Cencora, Inc.	COR
45	Healthcare	Leader	Volatile	Novo Nordisk A/S	NVO
46	Healthcare	Leader	Volatile	Bio-Rad Laboratories, Inc.	BIO
47	Healthcare	Leader	Volatile	RadNet, Inc.	RDNT
48	Industrials	Non-Leader	Stable	Barrett Business Services, Inc.	BBSI
49	Industrials	Non-Leader	Stable	Casella Waste Systems, Inc.	CWST
50	Industrials	Non-Leader	Stable	Maximus, Inc.	MMS
51	Industrials	Non-Leader	Volatile	Applied Industrial Technologies, Inc.	AIT
52	Industrials	Non-Leader	Volatile	Ennis, Inc.	EBF
53	Industrials	Non-Leader	Volatile	AAON, Inc.	AAON
54	Industrials	Leader	Stable	Illinois Tool Works Inc.	ITW
55	Industrials	Leader	Stable	Republic Services, Inc.	RSG
56	Industrials	Leader	Stable	Canadian Pacific Kansas City Limited	CP
57	Industrials	Leader	Volatile	The Boeing Company	BA
58	Industrials	Leader	Volatile	Southwest Airlines Co.	LUV
59	Industrials	Leader	Volatile	C.H. Robinson Worldwide, Inc.	CHRW
60	Technology	Non-Leader	Stable	Amdocs Limited	DOX
61	Technology	Non-Leader	Volatile	ADTRAN Holdings, Inc.	ADTN
62	Technology	Non-Leader	Volatile	Power Integrations, Inc.	POWI
63	Technology	Non-Leader	Volatile	Cantaloupe, Inc.	CTLP
64	Technology	Leader	Stable	Roper Technologies, Inc.	ROP
65	Technology	Leader	Stable	Teledyne Technologies Incorporated	TDY
66	Technology	Leader	Stable	Motorola Solutions, Inc.	MSI
67	Technology	Leader	Volatile	Corning Incorporated	GLW
68	Technology	Leader	Volatile	Micron Technology, Inc.	MU
69	Technology	Leader	Volatile	STMicroelectronics N.V.	STM

Source: Own elaboration.

In Table 8, all the selected companies are indicated where the selection is not implemented in favor of certain companies. The selection process was random and the selected stocks include both well-established market leaders and smaller non-leader firms. This ensures a comprehensive view of different performance and risk profiles in the stock market.

Although the initial stock selection is defined as 72 company stocks, the final dataset includes 69 stocks. The three excluded stocks belong to non-leader, stable companies and are not obtained due to data inconsistencies. While the majority of the selected stocks are from non-leader companies, stability remains relatively rare within this group. This composition allows for a more comprehensive analysis of market dynamics across different volatility and leadership profiles, providing valuable insights into the predictive performance of models across diverse stock characteristics.

2.1.7. Variable Considerations

For the analysis, a rich set of variables is used to capture the details of stock market behaviors. The volume and volatility indicators are crucial for analyzing the day-to-day changes and broader trends in stock prices. On the other hand, the time-based variables are also crucial for stock analysis. The day of the month and day of the week might directly affect stock prices, and they are also included in the analysis. These variables assist in identifying patterns like whether stocks tend to rise or are more likely to fall at certain times. This approach provides investors with useful clues about the best times to buy or sell. In addition, in this analysis, the moving-average-based variables via using Simple Moving Average (SMA) are included to highlight trends. Not only moving averages but also the rolling window methods are used for analyzing volume, price, and volatility which provides a clearer picture of market dynamics. In general, this approach is effective for understanding price fluctuations.

2.1.8. Data Split

There is a list of 69 datasets from selected companies each dataset includes determined variables.

Table 9. Data division for modeling

Dataset	Time interval	The number of Observations	Percentage (%)
Train set	January 2000 - December 2019	5031	79%
Validation set	January 2020 - December 2022	756	12%
Test set	January 2023 - April 2025	583	9%

Source: Own elaboration.

Table 9 illustrates the division of the datasets into 3 core parts. The datasets used in this study contain a total of 6,370 observations and the separate subsets are a training set, a validation set, and a test set to support the modeling process. The training set includes 5,031 observations, covering the period from January 2000 to December 2019, and consists of approximately 78.99% of the data. The validation set includes 756 observations from January 2020 to December 2022, which is representing about 11.87% of the dataset. Finally, the test set only captures 583 observations from January 2023 to April 2025, which includes roughly 9.15% of the total observations. This time-based split allows for a clear and structured evaluation of model performance, where the training data is used to build the models, the validation set is used for tuning and selection, and the test set is reserved for final performance assessment.

2.2. Modeling Approach

The analytical framework involves a robust set of models to ensure a thorough examination of the predictive capabilities through different types of stocks.

Table 10. Applied econometric and machine learning models

No	Model	Type
1	Linear Regression	Linear econometric model
2	Elastic Net Regression	Regularized linear econometric model
3	Decision Tree	Non-linear, simple machine learning model
4	Random Forest	Non-linear, advanced machine learning model
5	LightGBM	Non-linear, advanced machine learning model

Source: Own elaboration.

Table 10 presents an overview of the five machine learning models applied in this study. In this study, five machine learning models were carefully selected to balance simplicity, interpretability, and predictive power across various stock market scenarios. The two linear models such as Linear Regression and Elastic Net Regression serve as foundational baselines. Linear Regression offers straightforward implementation and interpretability, which makes it a common starting point for predictive modeling. Elastic Net Regression combines L1 and L2 regularization addresses issues of multicollinearity and performs variable selection, enhancing model robustness in the presence of correlated features.

The three non-linear models (Decision Tree, Random Forest, and LightGBM) opted for their ability to capture complex, non-linear relationships in financial data. This combination of models allows for a comprehensive evaluation of both linear and non-linear approaches in predicting stock returns, facilitating a deeper understanding of which methodologies perform best under different market conditions (Bartol et al., 2022; Zeleke et al., 2024)

2.2.1. Linear Regression

An multivariate linear regression model is used which is to identify the impact of different factors and for linear predictions.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots\dots\dots + \beta_k x_k + \varepsilon_t \qquad (3)$$

In Equation 3, y is target variable (Return of Stock), and x values are predictors which are the input variables that influence target. β's are coefficients are displaying the impact or weight of each x on y. Lastly, ε_t is error term which captures everything that affects y but is not included in the model (Montgomery, Peck, & Vining, 2021).

2.2.2. Elastic Net Regression

Ridge regression, LASSO, and Elastic Net are regularization methods normally used to address multicollinearity and overfitting problems by applying penalty terms to regression models. Hoerl and Kennard introduced the ridge regression estimator in 1970 as an alternative to the ordinary least squares (OLS) estimator in the presence of multicollinearity. In ridge regression, ridge parameter plays an important role in parameter estimation (Hoerl and Kennard, 1970). LASSO (Least Absolute Shrinkage and Selection Operator) regression is a shrinkage. This model is mostly used for variable selection, and also for regression models (Tibshirani, 1996).

Elastic Net as a machine learning model is build by the combination of 2 models such as ridge and lasso. Elastic Net combines the penalties of both Lasso and Ridge regression to assist and to improve the model accuracy and handle multicollinearity by selecting and regularizing features simultaneously.

$$\hat{\beta} \;=\; arg \min_{\beta} (||\mathrm{Y} \,-\, \mathrm{X}\beta||^{2} \;+ \lambda||\beta||_{1} \;+ \lambda||\beta||^{2}) \tag{4}$$

The elastic net model formula is provided in Equation 4. As we mentioned before, elastic net regression combines the penalties of both LASSO (L1-norm) and Ridge (L2-norm) regression, using two regularization parameters, $\lambda_1$and λ_2. This combination allows it to perform variable selection like LASSO while also handling multicollinearity and improving model stability like Ridge, making it particularly useful when there are many correlated features. $\hat{\beta}$ is showing the estimated coefficients, while Y is the vector of observed target variable values and X is the matrix of explanatory variables (predictors). The regularization parameter λ controls the strength of the penalty, and $||\mathrm{Y} - \mathrm{X}\beta||^2$ displays the sum of squared residuals (the error between predicted and actual values), with $||\beta||^2$ representing the squared L2-norm of the coefficient vector and L1-norm $||\beta||_1$, which is the sum of the squares of the coefficients (Hans, 2011).

2.2.3. Decision Tree Model

A Decision Tree is a non-parametric supervised learning algorithm normally used for classification and regression tasks. In the context of regression, the model splits the input space into distinct and non-overlapping regions using a tree-like structure, where internal nodes represent decision rules based on feature values, and leaf nodes correspond to the predicted values. The goal of the decision tree in regression is to minimize the sum of squared differences between the actual and predicted values in each partition. The formula representing the prediction of a regression decision tree can be expressed as:

$$\hat{y} = \sum_{m=1}^{M} c_m \, * I(x \, \in \, R_m) \tag{5}$$

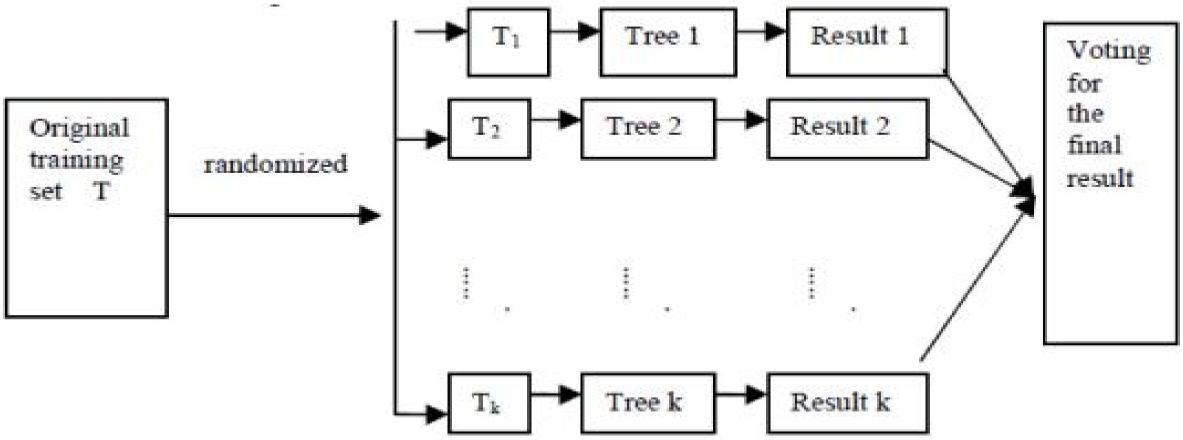
In Equation 5, M represents the number of terminal nodes (or leaves), R_m is the m-th region of the input space defined by the tree, c_m is the mean of the target variable (e.g., log return) in region R_m, and $I(x \in R_m)$ is an indicator function that equals 1 if observation $x$belongs to region R_m, and 0 otherwise. This formula shows that the tree assigns a constant value c_m to each region, effectively approximating the target function by piece-wise constants (Ying, 2015).

In the context of stock return prediction, decision trees are particularly valuable due to their ability to capture non-linear relationships and interactions among technical indicators and momentum variables without requiring complex preprocessing. They can easily handle missing data and are robust to multicollinearity, which is common in financial datasets. In our analysis, the decision tree model has not proved effective in identifying key variables influencing future log returns, particularly momentum-based features. This makes it a useful baseline model for return forecasting and a foundation for more advanced ensemble methods.

2.2.4. Random Forest Model

Random Forest (RF) is an ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. Random Forest model is mostly used to detect non-linear relationships which linear regression is not able to detect. It is a complex model which is widely used for both classification and regression tasks, and demonstrates superior performance in scenarios where high accuracy is crucial and the data may include several feature types and complexities.

Figure 2. Random forest workflow diagram



Source: Liu et al., (2012).

Figure 2 illustrates the process of the random forest algorithm. It is an ensemble method that builds multiple decision trees using bootstrap samples where random samples are drawn with replacements from the original training dataset. In this process, each tree is trained on a different bootstrap sample, and during the construction of each tree, a random subset of features is selected at every split, rather than using all available predictors. This randomization at the data and feature levels increases model diversity and reduces overfitting. For prediction, the Random Forest aggregates the outputs of all individual trees: by averaging in regression tasks and by majority voting in classification tasks, which stabilizes the variance and improves generalization compared to a single decision tree (Rigatti, 2017).

In the context of predicting stock log returns, random forest offers several advantages. It effectively captures complex non-linear interactions between features and is less sensitive to noise and overfitting compared to individual decision trees. Its robustness to outliers and high-dimensional feature spaces makes it particularly well-suited for financial data, where patterns are often subtle and noisy. In our study, the Random Forest model consistently ranked momentum variables among the most important predictors, reinforcing the relevance of trend-based features in return forecasting. The model's strong performance also demonstrates its utility as a reliable and interpretable tool in stock market prediction tasks.

2.2.5. Light Gradient Boosting (LGBM)

LightGBM (Light Gradient Boosting Machine) is a highly efficient gradient boosting framework based on decision tree algorithms, designed for speed and performance. Unlike traditional boosting methods, LightGBM grows trees using a leaf-wise (best-first) approach rather than a level-wise (breadth-first) approach, which leads to deeper trees and potentially better accuracy. The objective of LightGBM is to minimize a regularized loss function by adding weak learners in a stage-wise manner. The general form of the LightGBM prediction at iteration t is given by:

$$\hat{y}^{t} = \hat{y}^{t-1} + \eta * f_t(x) \tag{6}$$

In Equation 6, $\hat{y}^{t}$ is the updated prediction after the t-th iteration. $\hat{y}^{t-1}$ is the previous prediction, η is the learning rate, and $f_t(x)$ is the newly added decision tree model (a weak learner) trained to fit the residual errors of the previous model. This additive modeling approach incrementally improves prediction accuracy by focusing on the remaining errors at each step (Ke et al., 2017).

In the context of stock log return prediction, LightGBM is particularly powerful due to its ability to handle large-scale data and model complex, non-linear relationships with high accuracy and speed. Its built-in handling of missing values, efficient memory usage, and support for parallel learning make it ideal for financial datasets that are often large and noisy. In our results, LightGBM consistently outperformed traditional models in terms of predictive power, especially for stocks with volatile patterns. Momentum indicators again emerged as dominant features, highlighting their critical role in forecasting short-term price movements. LightGBM's balance between interpretability, speed, and performance makes it an excellent choice for return prediction in data-rich environments.

2.2.6. Cross Validation Technique

For all defined models, the cross-validation technique is utilized. Cross-validation is a fundamental technique in machine learning that helps evaluate the performance and generalizability of a model. Instead of training and testing a model on the same dataset, cross-validation involves splitting the data into several parts, or "folds." The model is trained on some folds and tested on the remaining fold, and repeating the process multiple times. This approach ensures that the model is tested on different parts of the data, and providing a more accurate estimate of its ability to perform on unseen data. Cross-validation assists detect overfitting, where a model performs well on training data but poorly on new data. A commonly used formula for calculating the average error across k folds is (Bergmeir & Benítez, 2012):

$$Average\ Error = \frac{1}{k}\sum_{i=1}^{k} Error_i \tag{7}$$

In Equation 7, k is the number of folds, and $Error_i$ is the error on the i_{th} fold. However, when dealing with time series data, standard cross-validation is not appropriate because of the time-dependent nature of the data. In time series, the order of observations matters, and using future data to predict the past would introduce unrealistic results. Therefore, a special method called "time series cross-validation" is used. In this approach, the model is trained on earlier observations and tested on later ones, respecting the timeline. Typically, we start with a small portion of the data, train the model, and then expand the training window step-by-step, forecasting the next points. This method is often called "rolling" or "expanding window" cross-validation. It is very effective because it mimics real-world forecasting situations, where only past information is available to predict the future. A simple illustration of expanding window cross-validation is:

$$Train: [1,2,3] \rightarrow Test: [4], Train: [1,2,3,4] \rightarrow Test: [5], etc. \tag{8}$$

Structure 8 displays the cross-validation structure. The training set continuously expands by including each new observation, while the model always predicts the next unseen data point. This approach ensures that future information is never used to predict past outcomes, maintaining the integrity of time series forecasting.

The initial training set is organized for the first 5 years from 2000 to 2005 (1250 training days). Then the dataset is expanded by 10 days. Repeatedly, 10 more days are added to the training set. This repetition continues until we reach the end of the training period and 5 years gives a deep understanding of historical patterns. For obtaining results of evaluation metrics, the median value is selected out of values that are collected from cross-validation. This cross-validation approach follows time order and is also beneficial for future leakage.

2.3. Evaluation Metrics

Every model in this thesis is assessed with three key measures; Root Mean Square Error (RMSE), which measures the magnitude of errors in predictions. Mean Absolute Error (MAE) is the next metric, which provides an average value of error magnitude; and lastly, R-squared (R^2) displays how well our models capture changes in stock prices. The selection of the metrics is done to provide insight into the performance of the models concerning accuracy and fit, covering different dimensions.

Table 11. Overview of applied evaluation metrics for model assessment

Evaluation Metric	Type	Describtion
R-squared	Coefficient of determination	It is an evaluation criterion for models. This metric calculates the percentage of the dependent variable's variance. It can be predicted based on the independent variables (Lewis-Beck and Skalaban, 1990).
MAE	Loss function	It is a loss function. It calculates the prediction accuracy by averaging the absolute differences between the expected and actual numbers (Robeson and Willmott, 2023).
RMSE	Loss function	It is a loss function and determines the square root of the average of the squares of the prediction errors, and it highlights greater errors and being sensitive to outliers (Aptula et al., 2005).

Source: Own elaboration.

Table 11 shows all evaluation metrics which are applied for the analysis. To understand the evaluation metrics in a detailed way, we need to dive into each metric specially. R-squared (R^2), mean absolute error (MAE), and root mean square error (RMSE) are well-known evaluation metrics used in machine learning tasks. The lower MAE and RMSE values and higher R^2 indicate better model performance (Wang et al., 2017; Chicco et al., 2021).

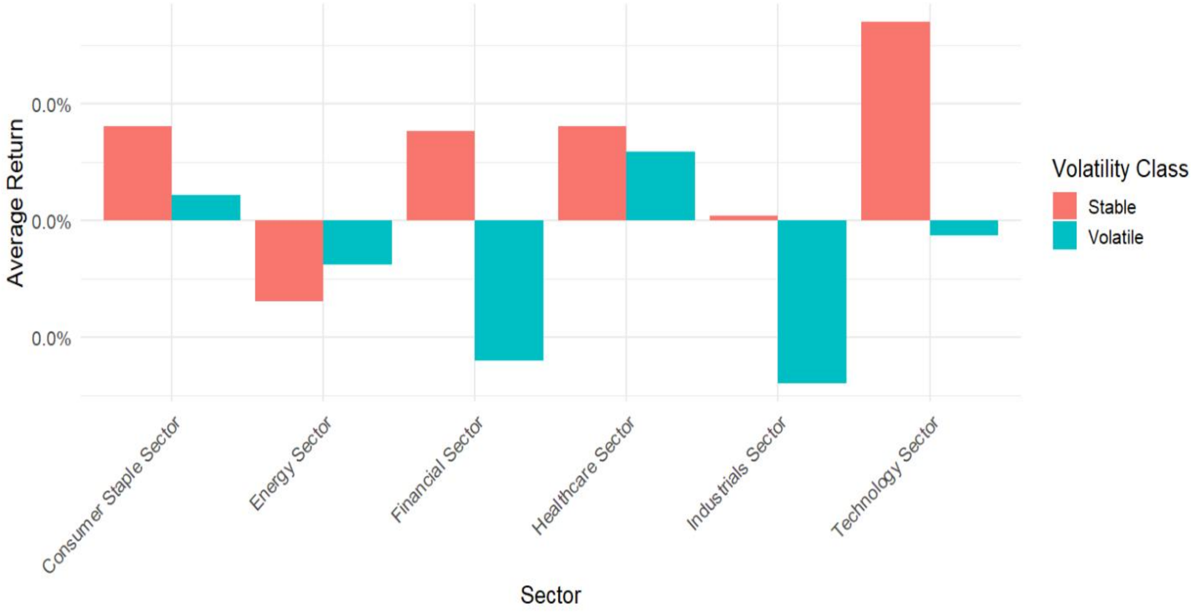
The selection of stocks, variables, and models, combined with evaluation metrics, is designed to provide a deep understanding of stock price dynamics and to validate the effectiveness of various predictive models in real-world scenarios. This detailed approach not only aims to aid in model comparison but also enriches our understanding of the financial market's complexity, making it a crucial study for investors and financial analysts alike.

2.4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is applied to reveal insightful patterns and relationships in the stock market. In this section, visualizations and summary tables are generated to provide key information in primary dimensions such as sector classification, market leadership status, and volatility levels. As a note, all EDA analysis is based on 967 stocks which is revealed after adding time constraints. These analyses help to better understand the structure of the market and guide the modeling process.

As a start of EDA, we compared the volatility levels of whole stocks and stocks that have been present in the stock market since 2000. While the whole list stocks over 3000, the stocks after the time filter become 967. The mean annualized volatility of more than 3000 stocks in the market is 0.6569 (from April 2024 to May 2025). In comparison, the companies that have existed for at least 25 years in the stock market have a lower average volatility of 0.4166 (from April 2024 to May 2025). This suggests that old companies that have been present in the stock market since 2000 are slightly less volatile than newly established companies.

Figure 3. Average return by sector and volatility class



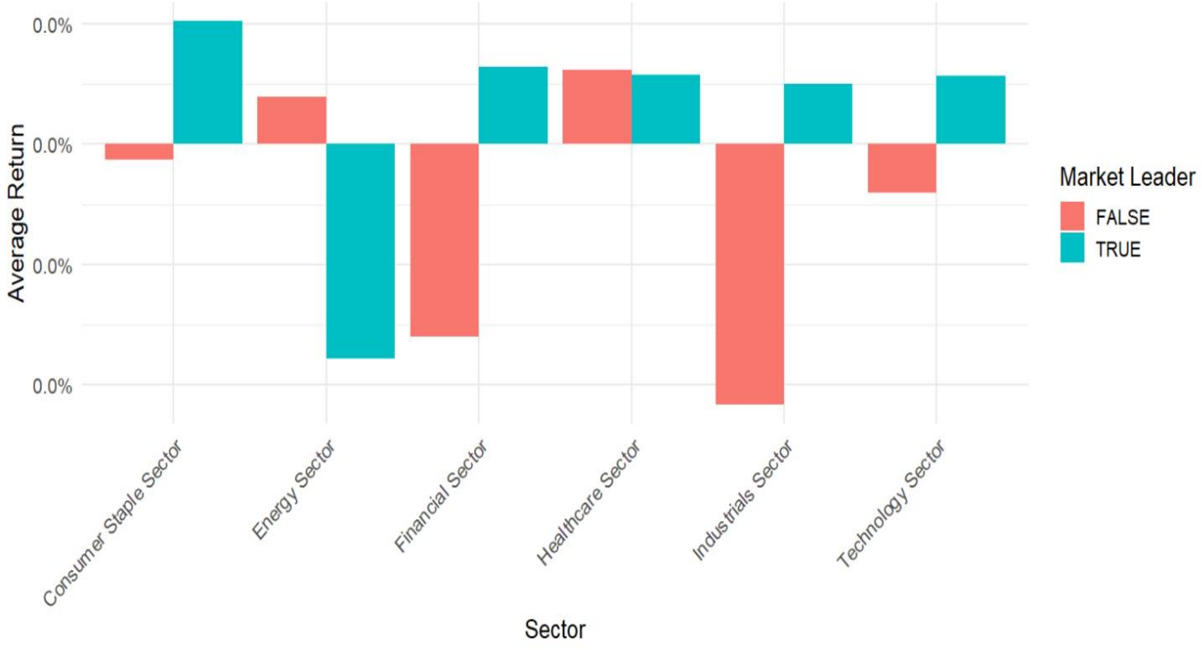
Source: Own elaboration.

Figure 3 illustrates the average daily return of stocks across six sectors, divided into stable and volatile classes. In most sectors, stable stocks (in red) demonstrate higher average returns than volatile ones (in turquoise). We observe in the Technology sector, stable stocks significantly perform better rather than volatile stocks, which shows a strong positive return compared to nearly zero for their volatile counterparts. Similarly, in the Financial and Healthcare sectors, stable stocks maintain moderate positive returns, while volatile stocks tend to perform worse or show negative averages.

On the other hand, sectors such as Energy, Industrials, and Healthcare show negative average returns for volatile stocks, reflecting the greater risk and unpredictability tied to price swings in these industries. Surprisingly, even in typically defensive sectors like Consumer Staples, stable stocks slightly outperform volatile ones, which reinforces the idea that lower volatility is often associated with more consistent and favorable returns.

Overall, this figure supports the argument that stability in stock price behavior may be linked to better average returns in many sectors which highlights the value of volatility classification in stock selection and model evaluation.

Figure 4. Average return by sector and market leadership



Source: Own elaboration.

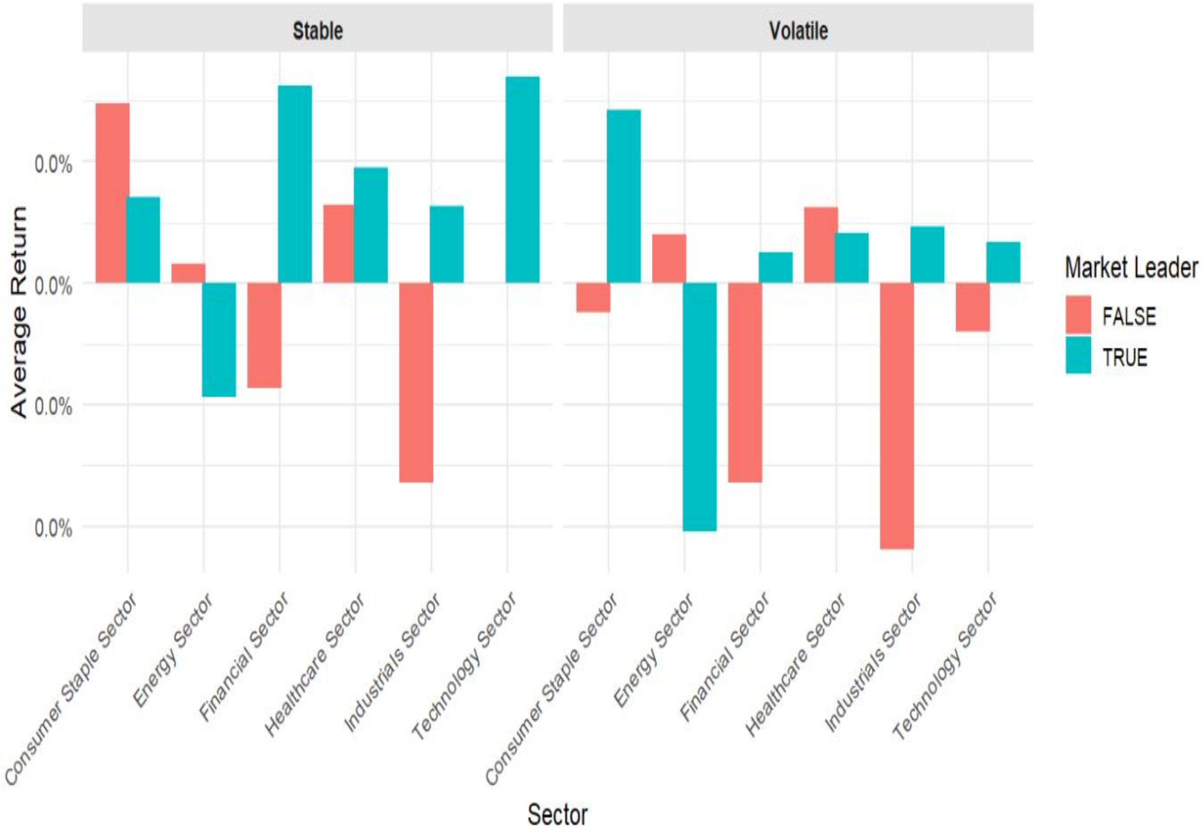
Figure 4 compares the average daily return across six sectors by distinguishing between market leaders (shown in blue) and non-leaders (shown in red). The plot reveals several insightful patterns about how market leadership status relates to return performance.

In most sectors, market leaders tend to outperform non-leaders. For example, in the Energy sector, leaders deliver significantly higher returns compared to non-leaders, who show notable negative performance. A similar pattern is observed in the Financial and Industrials sectors, where non-leaders experience negative or near-zero returns, while leaders maintain small but positive gains. These findings align with the idea that market leaders typically large, established firms which are more resilient and stable, especially in turbulent market conditions.

In contrast, the Consumer Staples and Healthcare sectors display that the difference in returns between leaders and non-leaders is minimal. It suggests that leadership status may play a less critical role in these more defensive or regulated industries. The Technology sector, known for both rapid growth and risk, shows slightly better performance for leaders, but both groups maintain positive average returns.

Overall, the figure emphasizes that being a market leader is often associated with higher and more stable returns, especially in cyclical or economically sensitive sectors.

Figure 5. Average return by sector, leadership, and volatility



Source: Own elaboration.

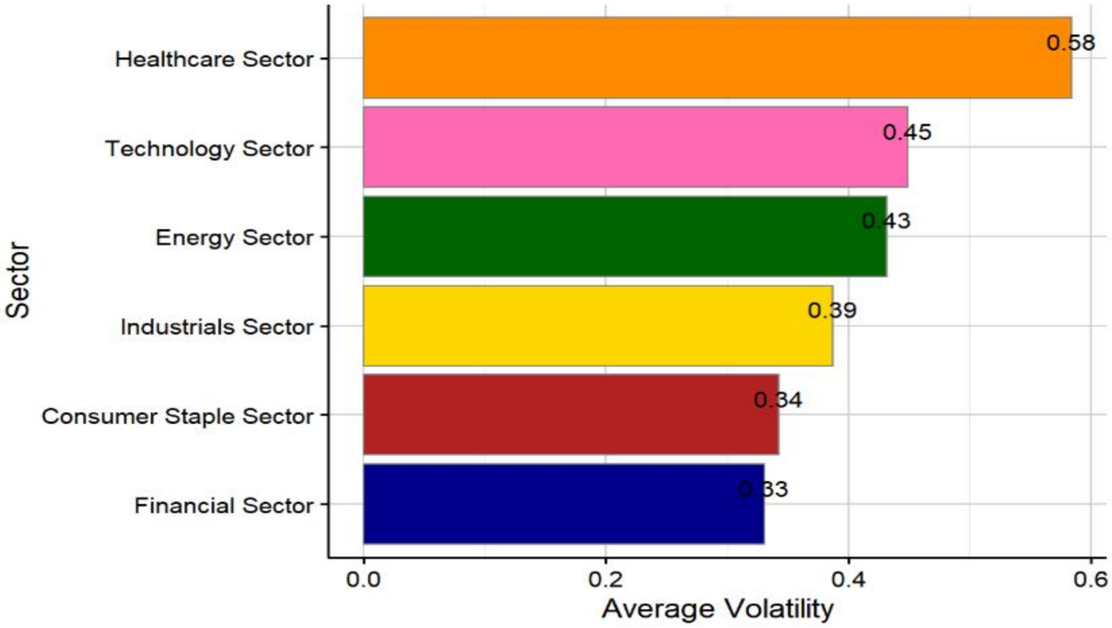
Figure 5 provides a comprehensive view of average stock returns by combining three key classification dimensions: sector, market leadership status, and volatility level. The chart is divided into two panels (Stable and Volatile) with each panel showing average returns for market leaders (blue) and non-leaders (red) across six sectors.

A clear pattern emerges: Stable market leaders tend to perform best across most sectors. For instance, in the Technology, Financial, and Energy sectors, stable leaders exhibit notably higher average returns than both their volatile counterparts and non-leaders. This reinforces the notion that price stability combined with market dominance offers a favorable return profile, especially in economically sensitive or growth-driven sectors.

In contrast, volatile non-leaders often show the weakest performance, particularly in Industrials and Financials, where they yield negative average returns. This suggests that smaller, riskier companies without a dominant market position are more vulnerable to price swings and underperformance.

There are a few exceptions. In the Consumer Staples sector, stable non-leaders slightly perform better than leaders. It is a hint for the defensive nature of the sector where even smaller players may offer consistent returns. It supports the value of a multi-dimensional stock classification approach in understanding stock behavior and building more targeted forecasting models.

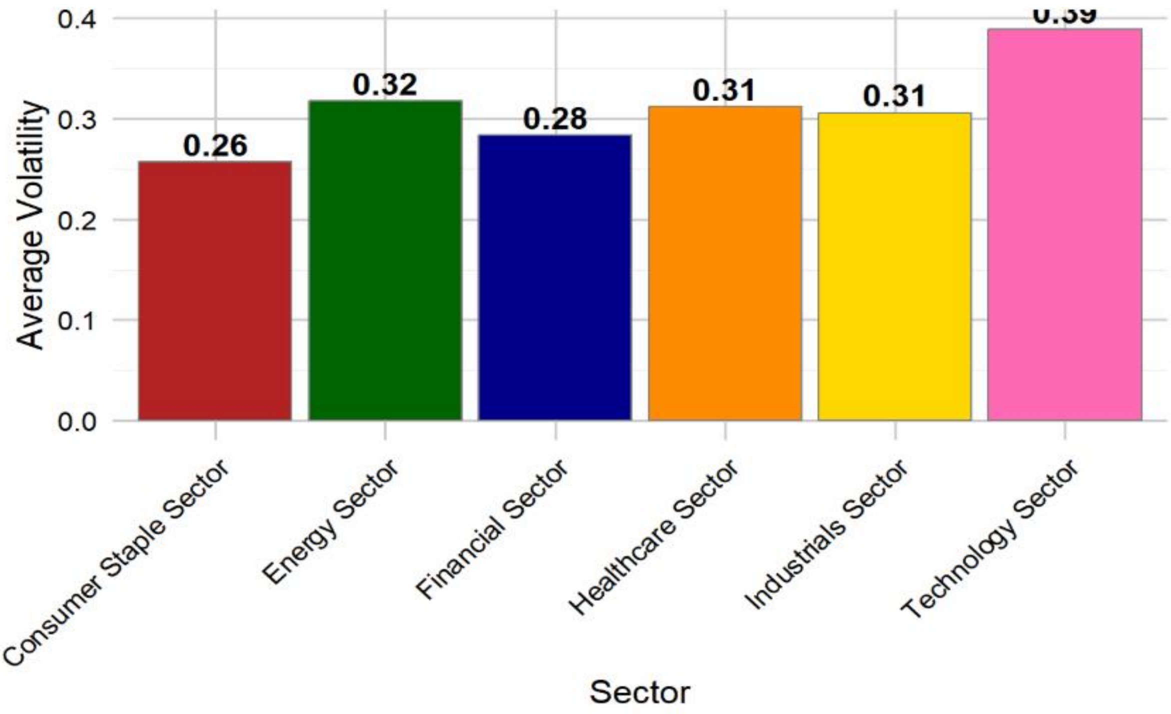
Figure 6. Average stock volatility by sector



Source: Own elaboration.

Figure 6 presents the average volatility levels across different sectors, based on a consistent 25-year historical analysis. The volatility values were calculated using the annualized standard deviation of daily returns over the past 25 years, and include 967 stocks that have been continuously listed and active in the market during this entire period. This approach ensures consistency and comparability across sectors by using a stable sample. Among the sectors, Healthcare exhibits the highest average volatility with a score of 0.58, which indicates more frequent and intense price fluctuations. In contrast, the Financial sector shows the lowest average volatility at 0.33, which reflects more stable stock price movements. These findings highlight how sector-specific characteristics such as regulatory dynamics, innovation cycles, or sensitivity to macroeconomic shifts can significantly influence the risk profile of stocks within each sector.

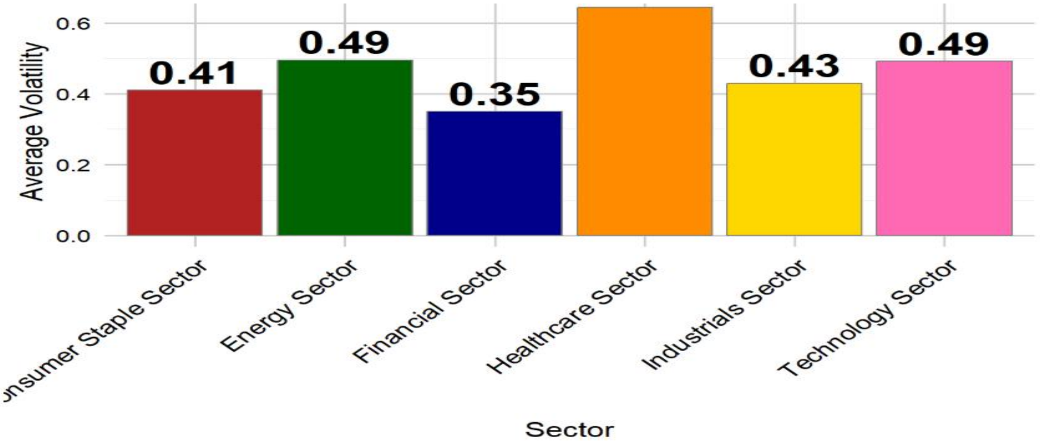
Figure 7. Average stock volatility by sector for market leaders



Source: Own elaboration.

Figure 7 focuses only on market leader companies and displays the average volatility level of each sector company. We may observe that the average volatility score is in the interval of 0.26 and 0.39 which shows leader companies are performing stable rather than non-leader companies. The technology sector leads in volatility score, followed by the energy and healthcare sectors. So these are more risky market leaders rather than other sector market leaders. In the following, the consumer staples and financial sectors show the lowest volatility for their leading firms. It shows that top-performing companies in these sectors tend to have more stable returns.

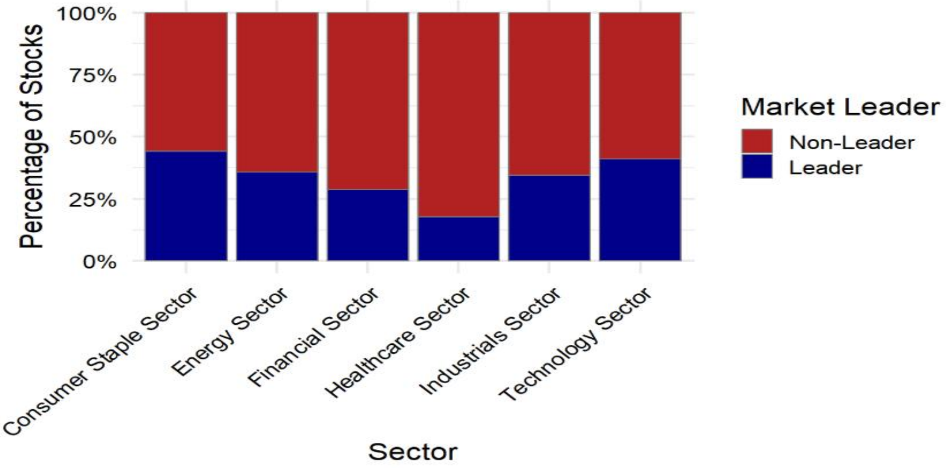
Figure 8. Average stock volatility by sector for non-market leaders



Source: Own elaboration.

In this stage, we can see the average volatility scores for companies specifically non-market leaders in Figure 8. The average volatility range for each sector's nonmarket leader companies is between 0.35 and 0.64. This approach proves that non-market leaders are more risky rather than market leader companies. The healthcare non-leader companies show the highest volatility and differentiate from other sectors. It defines that smaller firms in this sector encounter greater uncertainty. Nevertheless, the financial sector remains more stable even among non-leading firms.

Figure 9. Distribution of market leaders across sectors

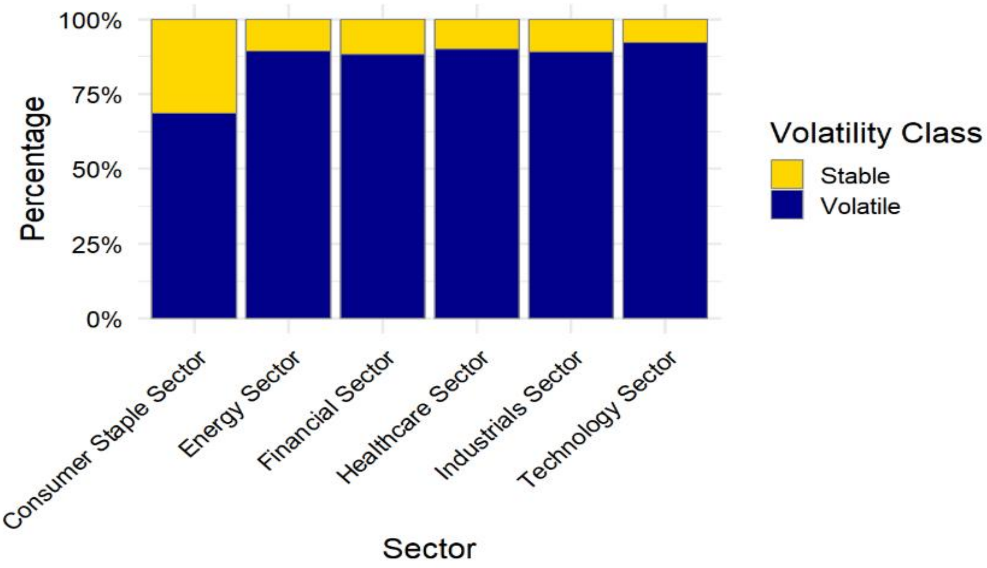


Source: Own elaboration.

Figure 9 illustrates the distribution of market leaders and non-leaders across the six analyzed sectors. While it may seem surprising that some sectors include up to 45% of stocks classified as leaders, this proportion reflects the outcome of a clear and consistent classification rule. In this study, market leaders are defined as companies with a market capitalization of at least USD 10 billion, placing them in the large-cap or mega-cap categories. This threshold is widely accepted in financial literature and aligns with typical classifications used by institutions and index providers.

Although certain sectors such as Consumer Staples or Technology show a relatively higher share of companies meeting this criterion, this is a reflection of the true market composition rather than a selection bias. These sectors include many globally dominant firms, naturally resulting in a higher proportion of large-cap stocks. Therefore, while leadership distribution varies by sector, the defined threshold remains a valid and robust approach for distinguishing leading firms from smaller or mid-sized peers in this analysis.

Figure 10. Distribution of stocks volatility across sectors



Source: Own elaboration.

Figure 10 presents the volatility distribution of stocks by sector, highlighting the proportion of volatile versus stable companies. In the figure, yellow bars represent stable stocks, while blue bars indicate volatile stocks. To classify these groups, annualized volatility was calculated based on each stock's price behavior up to April 2025, and a threshold of 25% annualized volatility was applied. Stocks with volatility below 25% were categorized as stable, while those above 25% were labeled volatile. The chart clearly shows that in most sectors, the vast majority of stocks are volatile (roughly 85% to 90%). The Consumer Staples sector stands out as an exception, with around 30% of its companies classified as stable. Overall, the visualization suggests that frequent price fluctuations are common across most sectors, while Consumer Staples, known for defensive and consistent business models, tend to exhibit greater price stability.

The analyses focused on general insights into the market till now using leadership status and volatility status. If we would like to understand the market behavior in a better way, we also need to focus on individual company trends. The investor desires to be aware of which companies dominate, which are facing less fluctuation and so on.

Table 12. Leading companies by sector in 2025 (Top 3 per Sector)

No	Sector	Company name	Symbol	Market cap
1	Consumer Staple	Walmart Inc.	WMT	$730.0 Billlion
2	Consumer Staple	Costco Wholesale Corporation	COST	$408.4 Billion
3	Consumer Staple	The Procter & Gamble Company	PG	$380.4 Billion
1	Energy	Exxon Mobil Corporation	XOM	$477.6 Billion
2	Energy	Chevron Corporation	CVX	$279.7 Billion
3	Energy	Shell plc	SHEL	$202.8 Billion
1	Financial	JPMorgan Chase & Co.	JPM	$687.6 Billion
2	Financial	Visa Inc.	V	$617.3 Billion
3	Financial	Mastercard Incorporated	MA	$472.9 Billion
1	Healthcare	Eli Lilly and Company	LLY	$707.2 Billion
2	Healthcare	UnitedHealth Group Incorporated	UNH	$490.1 Billion
3	Healthcare	Novo Nordisk A/S	NVO	$389.4 Billion
1	Industrials	General Electric Company	GE	$152.7 Billion
2	Industrials	Caterpillar Inc.	CAT	$176.1 Billion
3	Industrials	RTX Corporation	RTX	$187.4 Billion
1	Technology	Apple Inc.	AAPL	$3.698 Trillion
2	Technology	NVIDIA Corporation	NVDA	$3.495 Trillion
3	Technology	Microsoft Corporation	MSFT	$3.176 Trillion

Source: Own elaboration.

Table 12 shows the top three market leader companies based on market capitalization in each sector for 2025. This table shows which companies are leading each sector. The corporations provided are the most influential and financially robust players within their respective industries. These entities play a huge role in trends, innovation, and strategic direction through global markets.

The **Technology sector** is led by **Apple Inc. (AAPL)**, **NVIDIA Corporation (NVDA)**, and **Microsoft Corporation (MSFT)**. All other sector leaders have market capitalization in billions of USD but only Technology sector companies lead the market with trillions of United States dollars. These tech giants not only lead their sector but also represent the most valuable companies in the world. **Apple, NVIDIA,** and **Microsoft** have a market capitalization of **$3.698 trillion, $3.495 trillion,** and **$3.176 trillion**, respectively. The large gap between the technology sector and other sectors occurs because of innovation, global consumer and enterprise adoption, and expansion into AI, cloud computing, and consumer electronics.

JPMorgan Chase & Co. (JPM) company leads with a **$687.6 billion** valuation in the Financial sector. It is the largest bank in the United States according to its assets and a major global financial services provider. **Visa Inc. (V)** and **Mastercard Incorporated (MA)** are other dominating entities in the finance sector with market caps of **$617.3 billion** and **$472.9 billion** respectively. They dominate the global electronic payments industry and benefit from widespread digitization and increasing cashless transactions.

The most dominant company in the **Healthcare sector** is **Eli Lilly and Company (LLY)** with a market cap of **$707.2 billion**. This entity is popular because of its diabetes and obesity drugs, including Mounjaro and Zepbound. **UnitedHealth Group (UNH)** valued at **$490.1 billion** is another leader doing business in care and health insurance, while **Novo Nordisk (NVO)** has a market value of **$389.4 billion**, and became well-known due to its breakthroughs in GLP-1 drugs for weight loss and diabetes treatment.

In the **Energy sector**, traditional oil and gas giant organizations keep dominance. **Exxon Mobil Corporation (XOM)** is leading the Energy sector with a value of **$477.6 billion**. The following companies are **Chevron Corporation (CVX)** at **$279.7 billion**, and **Shell plc (SHEL)** at **$202.8 billion**. Their market leadership is supported by integrated operations, vast global reserves, and strategic changes toward cleaner energy investments amid energy transition efforts. **Walmart Inc. (WMT)** remains the global retail leader with a **$730 billion** market cap in the Consumer Staple sector. Walmart is growing through scale, supply chain efficiency, and digital transformation. Other dominant companies are **Costco Wholesale**

Corporation (COST) and **Procter & Gamble (PG)** with values of **$408.4 billion** and **$380.4 billion**, respectively. These companies are known for their strong consumer loyalty and global brand recognition, especially in essential goods and services.

Lastly, in the industrial sector, the 3 dominating entities are **RTX Corporation (RTX)**, **Caterpillar Inc. (CAT)**, and **General Electric Company (GE)**. They are at the top of the list due to industrial powerhouses, and their market values ranging from **$152.7** to **$187.4 billion**. These companies are strong in the market because of the critical infrastructure, aerospace, heavy machinery, and industrial technologies. These elements are essential for global economic development.

Table 13. Sector-wise lowest capitalized companies in 2025 (Bottom 3 per Sector)

No	Sector	Company name	Symbol	Market cap
1	Consumer Staple	Steakholder Foods Ltd.	STKH	$0.4 Million
2	Consumer Staple	Tantech Holdings Ltd	TANH	$1.4 Million
3	Consumer Staple	Greenlane Holdings, Inc.	GNLN	$2.1 Million
1	Energy	Antero Resources Corporation	AR	$0.6 Million
2	Energy	Trio Petroleum Corp.	TPET	$2.7 Million
3	Energy	EON Resources Inc.	EONR	$6.9 Million
1	Financial	Lion Group Holding Ltd.	LGHL	$1.5 Million
2	Financial	Sterling Bancorp, Inc. (Southfield, MI)	SBT	$248.0 Million
3	Financial	Bleichroeder Acquisition Corp. I	BACQ	$250.7 Million
1	Healthcare	Aclarion, Inc.	ACON	$0.6 Million
2	Healthcare	Revelation Biosciences, Inc.	REVB	$1.1 Million
3	Healthcare	Scorpius Holdings, Inc.	SCPX	$1.4 Million
1	Industrials	Gencor Industries, Inc.	GENC	$240.3 Million
2	Industrials	Azul S.A.	AZUL	$244.9 Million
3	Industrials	flyExclusive, Inc.	FLYX	$247.7 Million
1	Technology	Signing Day Sports, Inc.	SGN	$1.4 Million
2	Technology	WM Technology, Inc.	MAPS	$241.3 Million
3	Technology	Silvaco Group, Inc.	SVCO	$244.8 Million

Source: Own elaboration.

Table 13 shows the three lowest market capitalization companies in each sector in 2025, based on their market capitalization. This offers insights into the smallest publicly traded players in each industry. These companies often operate in niche markets, early growth stages, or under financial distress, and their relatively small market caps suggest limited investor influence and trading volume. Although we look at the companies that possess the worst market capitalization, we see industrial, technology, and financial sectors have bigger market cap valued companies rather than consumer staples, energy, and healthcare organizations. Overall, this table describes the diversity and illustrates how companies with vastly different scales of operation exist in the same industry. These smaller players often represent higher risk however may also offer high-reward potential if successful in innovation or market disruption.

Table 14. Stable companies by sector in 2025 (Top 3 per Sector)

No	Sector	Company name	Symbol	Volatility
1	Consumer Staple	The Coca-Cola Company	KO	0.16
2	Consumer Staple	The Procter & Gamble Company	PG	0.18
3	Consumer Staple	Post Holdings, Inc.	POST	0.18
1	Energy	Enbridge Inc.	ENB	0.17
2	Energy	Enterprise Products Partners L.P.	EPD	0.17
3	Energy	MPLX LP	MPLX	0.19
1	Financial	Mountain Lake Acquisition Corp.	MLAC	0.01
2	Financial	Bleichroeder Acquisition Corp. I	BACQ	0.02
3	Financial	Launch One Acquisition Corp.	LPAA	0.02
1	Healthcare	Amedisys, Inc.	AMED	0.14
2	Healthcare	Takeda Pharmaceutical Company Limited	TAK	0.18
3	Healthcare	Innoviva, Inc.	INVA	0.18
1	Industrials	Waste Connections, Inc.	WCN	0.16
2	Industrials	Republic Services, Inc.	RSG	0.17
3	Industrials	RELX PLC	RELX	0.19
1	Technology	Juniper Networks, Inc.	JNPR	0.14
2	Technology	Automatic Data Processing, Inc.	ADP	0.18
3	Technology	Amdocs Limited	DOX	0.19

Source: Own elaboration.

Table 14 identifies the top 3 most stable companies in each sector. The companies ranked by their annualized volatility (2024 April - 2025 April). Companies with low volatility experience fewer or smaller price fluctuations, which makes them attractive to risk-averse investors, long-term holders, and those seeking dividend reliability. Financial companies act differently than other sector firms. In general, the volatility range of the top 3 most stable companies is between 14%-19% for all sectors but in the finance sector there is an ultra-low volatility of 1%. This difference is because of companies' limited trading activity or capital structure. The major players in the Consumer Staple sector, t**he Coca-Cola Company (KO)** and **Procter & Gamble (PG)** also move historically steady performance with volatility values of **0.16** and **0.18**, respectively. These firms are known for their strong global brands, recession-resistant products, and stable earnings. The most stable companies in the Energy sector are **Enbridge (ENB)** and **Enterprise Products Partners (EPD)** which have the same volatility score of **0.17**. In Healthcare, **Amedisys (AMED)** and **Takeda Pharmaceuticals (TAK)** keep stable operations and market confidence, supporting volatility scores below **0.18**. Technology entities such as **Juniper Networks (JNPR)** and **Automatic Data Processing (ADP)** have dependable cash flows which exhibit volatility as low as **0.14** and **0.18**, respectively.

Table 15. Volatile companies by sector in 2025 (Top 3 per Sector)

No	Sector	Company name	Symbol	Volatility
1	Consumer Staple	China Liberal Education Holdings Limited	CLEU	6.72
2	Consumer Staple	Oriental Rise Holdings Limited	ORIS	5.25
3	Consumer Staple	Tantech Holdings Ltd	TANH	4.09
1	Energy	PTL Limited	PTLE	2.49
2	Energy	Leishen Energy Holding Co., Ltd.	LSE	1.93
3	Energy	Trio Petroleum Corp.	TPET	1.84
1	Financial	Health In Tech, Inc.	HIT	3.46
2	Financial	Mercurity Fintech Holding Inc.	MFH	1.49
3	Financial	Sezzle Inc.	SEZL	1.29
1	Healthcare	Scorpius Holdings, Inc.	SCPX	4.29
2	Healthcare	Mainz Biomed N.V.	MYNZ	3.86
3	Healthcare	Tectonic Therapeutic, Inc.	TECX	3.69
1	Industrials	Richtech Robotics Inc.	RR	2.15
2	Industrials	ZJK Industrial Co., Ltd.	ZJK	2.09
3	Industrials	Microvast Holdings, Inc.	MVST	1.98
1	Technology	Signing Day Sports, Inc.	SGN	2.34
2	Technology	QXO, Inc.	QXO	2.09
3	Technology	Quantum Computing Inc.	QUBT	1.96

Source: Own elaboration.

Table 15 displays the three most volatile companies in each sector in April 2025 via using annualized volatility. We obtain a clear view of companies that may be considered high-risk investments and the most unpredictable stock price movements. High volatility is often associated with greater risk but also higher potential returns, attracting speculative or short-term investors. The consumer staple sector shows unusually high volatility levels, with **China Liberal Education Holdings (CLEU)** recording a striking **6.72** which is far above typical industry norms. Similarly, **Tantech Holdings (TANH)** appears again with a volatility of **4.09** displaying instability in its market performance. In the Energy sector, **PTL Limited (PTLE)** leads with a volatility of **2.49** suggesting significant fluctuations in stock prices possibly due to unstable commodity markets or operational uncertainty. In Healthcare, **Scorpius Holdings (SCPX)** shows volatility of **4.29** which is high sensitivity to market or clinical news. In addition, technology companies such as **Signing Day Sports (SGN)** and **Quantum Computing Inc. (QUBT)** exhibit volatility values above **1.9** portraying the speculative nature of emerging tech ventures.

2.5. Feature Engineering

In the feature engineering part, a meaningful set of features is prepared and added to the stock price data that could improve the performance of prediction models. The core goal is to generate variables that capture the most important patterns, trends, and behaviors in the market. This is a crucial step because machine learning models learn from the information we give them. Hence, it is essential to create high-quality, and informative features for better predictions.

The target variable in this study is **Log_Return** which is provided in Table 9. In this paper, the objective is to predict whether the stock price will go up or down on this day and by how much. This measure is widely used in financial studies because it stabilizes the variance and allows easier mathematical handling of returns over time. The Log_Return is calculated as the natural logarithm of the ratio of today’s closing price to the previous day’s closing price (Martin & Wagner, 2019):

$$\mathrm{Log\_Return_t} = \ln\left(\frac{\mathrm{P_t}}{\mathrm{P_{t-1}}}\right) \tag{9}$$

To support this prediction, new variables are created and grouped into several categories:

2.5.1. Returns

Log_Return reflects the daily return of a stock based on its closing prices and it helps to detect recent stock movement patterns. In this analysis, 5, 10, and 21 working days are used to represent weekly, bi-weekly, and monthly trading trends, respectively. This is based on typical working days in financial markets.

2.5.2. Moving Averages

The moving average is a fundamental tool in technical analysis, used to smooth out short-term price fluctuations and reveal the underlying trend direction. It helps reduce the effect of random volatility and provides clearer signals about market behavior (Kirkpatrick II & Dahlquist, 2010). In this study, a 14-day simple moving average (SMA_10) is used to track short-term trends in stock prices. The 14-day window is a widely accepted standard in financial analysis, as it effectively balances reactiveness to recent price movements with the ability to filter out market noise, making it suitable for short-term market forecasting (Achelis, 2001). Additionally, the Close_to_SMA variable is created to measure how far the current closing price deviates from its 14-day average. This measure is often used to identify overbought or oversold conditions, helping to detect price reversals or entry/exit opportunities (Nau, 2014). The use of both SMA_10 and its relative measure enhances the model's sensitivity to price momentum while maintaining interpretability.

2.5.3. High-Low Features

When we analyze financial time series data with statistical models, it is a key assumption that the parameters of the model are constant over time. Nevertheless, the economic environment often changes, and it can not be reliable to assume that a model's parameters are constant. One of the popular techniques to assess the constancy of a model's parameters is to compute parameter estimates over a rolling window. Rolling window is needed to be a fixed size throughout the sample (Zivot & Wang, 2003). In this analysis, Rolling_High_Max_10 and Rolling_Low_Min_10 variables are created to track the highest and lowest prices in a recent window. Besides, Price_Range shows the daily range between high and low prices, and Close_to_Open compares the closing and opening prices. These features provide insights into daily volatility and price behavior.

2.5.4. Volume-based Features

It is obvious fact that trading volume is a strong signal of investor interest and market activity (Kadapakkam, Kumar, & Riddick, 1998). Volume_Change variables are created to measure the day-to-day change in volume, and also Rolling_Volume_Mean_5 and Rolling_Volume_Mean_10 variables to observe short-term volume trends. Another volume based created variable is Volume_to_AvgRatio which compares the current volume to the average. This factor helps to detect unusual activity.

2.5.5. Date Features

Time-based variables often reveal hidden patterns (Demirer & Karan, 2002). It is proved that the date-based features have a potential impact on the stock market and it is necessary to include them in the model for better performance (Zhang, Lai, & Lin, 2017). DayOfWeek and Month are added to observe if certain days or months have consistent behaviors. IsMonthEnd is created as a binary variable that shows the end of the month. This factor can affect prices due to fund rebalancing or reporting cycles (Demirer & Karan, 2002).

2.5.6. Return Averages

Rolling averages of past returns over 7, 14, and 30 days are calculated and added to the analysis to assist models learn from recent performance trends. This trend is whether upward or downward and in general, we use the impact of trend in forecasting.

2.5.7. Volatility Measures

Subsequently, Rolling_Return_Std_5, _10, and _21 features are calculated to measure return volatility over different time windows. In this concept, higher volatility might signal risk or uncertainty, and it is important when trying to forecast returns.

Table 16. Variable definitions and descriptions

Feature Group	Variables Created	Role
Returns	Log_Return	Response Variable
Moving Averages	SMA_14	Explanatory Variable
	Close_to_SMA	Explanatory Variable
High-Low Features	Rolling_High_Max_10	Explanatory Variable
	Rolling_Low_Min_10	Explanatory Variable
	Price_Range	Explanatory Variable
	Close_to_Open	Explanatory Variable
Volume	Volume_Change	Explanatory Variable
	Rolling_Volume_Mean_5	Explanatory Variable
	Rolling_Volume_Mean_10	Explanatory Variable
	Volume_to_AvgRatio	Explanatory Variable
Date Features	DayOfWeek	Explanatory Variable (Categorical)
	Month	Explanatory Variable (Categorical)
	IsMonthEnd	Explanatory Variable (Categorical)
Return Averages	Rolling_Return_Mean_5	Explanatory Variable
	Rolling_Return_Mean_10	Explanatory Variable
	Rolling_Return_Mean_21	Explanatory Variable
Volatility	Rolling_Return_Std_5	Explanatory Variable
	Rolling_Return_Std_10	Explanatory Variable
	Rolling_Return_Std_21	Explanatory Variable

Source: Own elaboration.

The diverse set of features is displayed in Table 16 where all variables aimed to capture both short-term and medium-term behaviors in stock movements. These created variables include price momentum, trend strength, investor behavior, and market dynamics. All of these factors might influence the return. This structured feature engineering process is for building more reliable and interpretable prediction models.

2.5.8. Handling with Multicollinearity problem

Multicollinearity is one of the problems in predictive modeling, especially with linear regression. This problem occurs when predictor variables are highly correlated with each other. This can distort the interpretation of model coefficients and reduce the model's reliability. Previous literature found that there are four measurements of multicollinearity. The first detector of multicollinearity is a pairwise correlation using a correlation matrix. A bivariate correlation of 0.8 or 0.9 is often determined as a threshold to display a high correlation between two regressors (Mason & Perreault, 1991). However, the correlations do not necessarily mean multicollinearity which is the issue in this method. The most widely used indicator of multicollinearity is the Variation Inflation Factor (VIF) or Tolerance (TOL) (Neter et al., 1996).

The VIF is the reciprocal of TOL. There is no formal value of VIF to define the presence of multicollinearity, however, a value of 10 and above often indicates multicollinearity (Weisberg, 2005). Thus, the Variance Inflation Factor (VIF) is applied to all 69 stock datasets to address this problem. A VIF score greater than 10 is considered problematic because of multicollinearity (Chan et al., 2022). Common variables with consistently high VIF scores in the majority of the 69 datasets included: Close, Open, High, Low, Adjusted Close, Simple Moving Average of 10 days, price_range, Rolling_High_Max_10, and Rolling_Low_Min_10, close to simple moving average, rolling window using mean for volume for 5 and 10 days, the standard deviation of rolling return for 10 and 21 days. For instance, in the INGR company dataset, Open, Close, and Price_Range had infinite VIFs, while in the TR company dataset, Open_TR had a VIF above 5,000.

All such problematic variables with VIF scores greater than 10 are excluded from the linear models to ensure reliable linear regression modeling. It is verified that after modification there is no multicollinearity problem anymore for linear models. Elastic net combines L1 and L2 regularization and naturally handles the multicollinearity issue. Although the elastic net model, does not require variable removal in most cases, the extremely high multicollinearity still can impact the model and may cause model results biased (Altelbany, 2021). Similarly, tree-based models (like Random Forests and Gradient Boosting) and neural networks are generally not affected by multicollinearity. These models are not affected because they do not rely on coefficient estimation in the same way as linear models. Based on the structure of models, in linear models such as linear regression and elastic net regression the problematic variables are excluded and used 13 features. Nevertheless all 26 features except symbol and date (which are non-sense variables) are maintained for non-linear models like decision tree, random forest and light gradient boosting to preserve potential nonlinear relationships and interactions. This process provides a fair and technically sound comparison across models (Chan et al., 2022). The generated diverse set of features is displayed in Table 16 where all variables aimed to capture both short-term and medium-term behaviors in stock movements. These created variables include price momentum, trend strength, investor behavior, and market dynamics. All of these factors might influence the return. This structured feature engineering process is for building more reliable and interpretable prediction models.

2.5.9. Feature Importance

Permutation Importance is a powerful and model-agnostic method that is used to evaluate how much each feature contributes to a machine learning model's predictive power. The primary idea is intuitive where we randomly shuffle the values in the dataset for each feature (so the relationship between the feature and the target is broken). After that mixture, we measure how much the model's performance drops. If performance significantly declines, it means the feature is important; if the performance stays the same, so the feature likely had little influence. This method helps us to understand which inputs truly influence predictions, offering deeper insights than just looking at model accuracy (Altmann et al., 2010).

Instead of model-specific importance metrics (e.g., Gini importance in decision trees), permutation importance is evaluating the impact of features by directly observing the effect of "disrupting" them. This approach makes the feature importance check especially valuable such as understanding which financial indicators are most predictive for different stock types. It is also helpful in avoiding biased assumptions because the importance is judged by the feature's actual contribution to unseen (test) data, rather than how it affected the model during training.

We repeat this permutation multiple times (to reduce randomness) and average the results for stability. In our case, the 10 repeats, and a random state of 42 for reproducibility are defined which means 10 times it is repeated to avoid randomness.

Momentum is analyzing past behavior and deciding the future trend. In this case, it refers to the tendency of stocks that have performed well in the recent past to continue performing well shortly, and vice versa for poorly performing stocks. In financial markets, this influence is often determined using technical indicators such as past returns over short or medium-term horizons (Chan, Jegadeesh, & Lakonishok, 1996). Momentum is a key factor in predicting log returns of stocks because it represents market participants' behavioral biases, such as herding and delayed overreaction to information. In our analysis, momentum-related variables emerged as some of the most significant predictors of future log returns across the majority of stocks. This highlights their strong explanatory power and practical relevance in forecasting models, supporting the well-documented momentum effect in financial literature (Steiner, 2009).

2.6. Best Model Selection Technique

Majority voting is a simple but powerful ensemble decision-making method where the final choice is determined by the preference of the majority among multiple options. This approach helps increase reliability by combining the outcomes of several models instead of relying on a single one in predictive modeling. Another advantage of majority voting refers to reducing the risk of overfitting or poor performance because of an individual model's limitations. Sometimes, one model can perform better in a specific stock, but it does not mean that the defined model is always better in this group of stocks. However, using the majority voting technique and analyzing numerous stocks instead of one stock, provides reliable decision-making regarding the selection of the best model for different groups of stocks. Majority voting is especially beneficial in comparative studies as this dissertation, where it ensures a fair and interpretable selection of the most suitable model for each class by aggregating multiple perspectives (Kotu & Deshpande, 2014; Jha & Kaur, 2022).

In this study, majority voting is used not to combine model outputs, but rather as a systematic decision rule to identify the preferred modeling approach (linear or non-linear) within each of the 24 defined stock sections. For each section, three representative stocks are analyzed, and the performance of both linear and non-linear models is evaluated individually. If two or more out of the three stocks in a section show better predictive results with linear models, then the section is categorized as best suited for a linear approach. In contrast, if non-linear models outperform in at least two cases, then a non-linear modeling strategy is selected as more appropriate for that section. Although, the best models selected in this mechanism, 2 sections have only 1 and 2 stocks which made data 69 instead of 72 (as it is planned). In those sections, we consider how other stocks in the same sector, leadership and volatility acted. This voting mechanism enables a balanced and evidence-driven model selection process across different combinations of sector, volatility, and leadership classification.

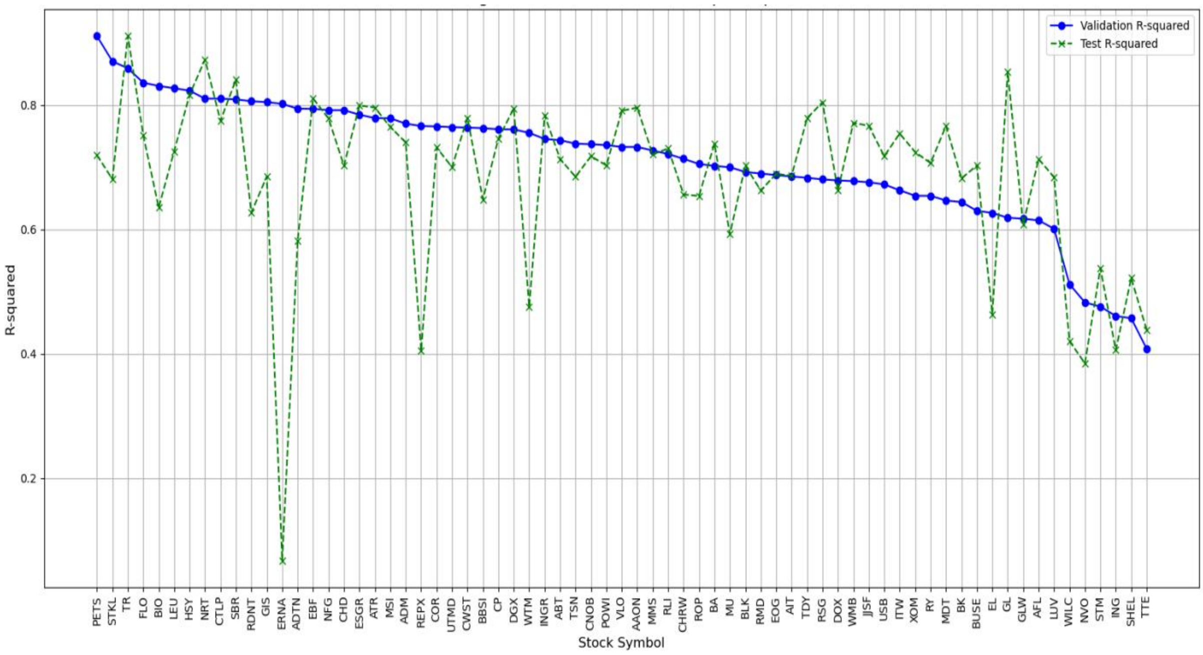
3. EMPIRICAL RESULTS

As emphasized widely in the methodology section, the data is selected based on three specific criteria, collected into 24 unique sections and also divided into three parts: training, validation, and testing. Several variables are created to enhance the predictive accuracy of the models. Standardization is applied for independent variables to adjust the scale of the variable. Five machine learning models are applied and evaluated using three distinct metrics such as R-squared, MAE, and RMSE. Currently, in this section, we present the results of the predictive modeling. Each model's results are observed individually, and followed by a comparative analysis throughout all models. The best models are identified per sector, market leadership status, and volatility status. Finally, the overall best models are defined for each of the 24 sections.

A total of 69 stocks were selected and evaluated using 5 different predictive models, resulting in 345 model–stock combinations. This comprehensive approach allows for a detailed investigation of stock market behavior across various modeling techniques and stock characteristics. The line plots are applied to show differences between train and test sets. These plots allow us to understand the performance of each model on all stocks more easily.

3.1. Model Results

Figure 11. Linear regression model results for all analyzed stocks on the validation and test samples

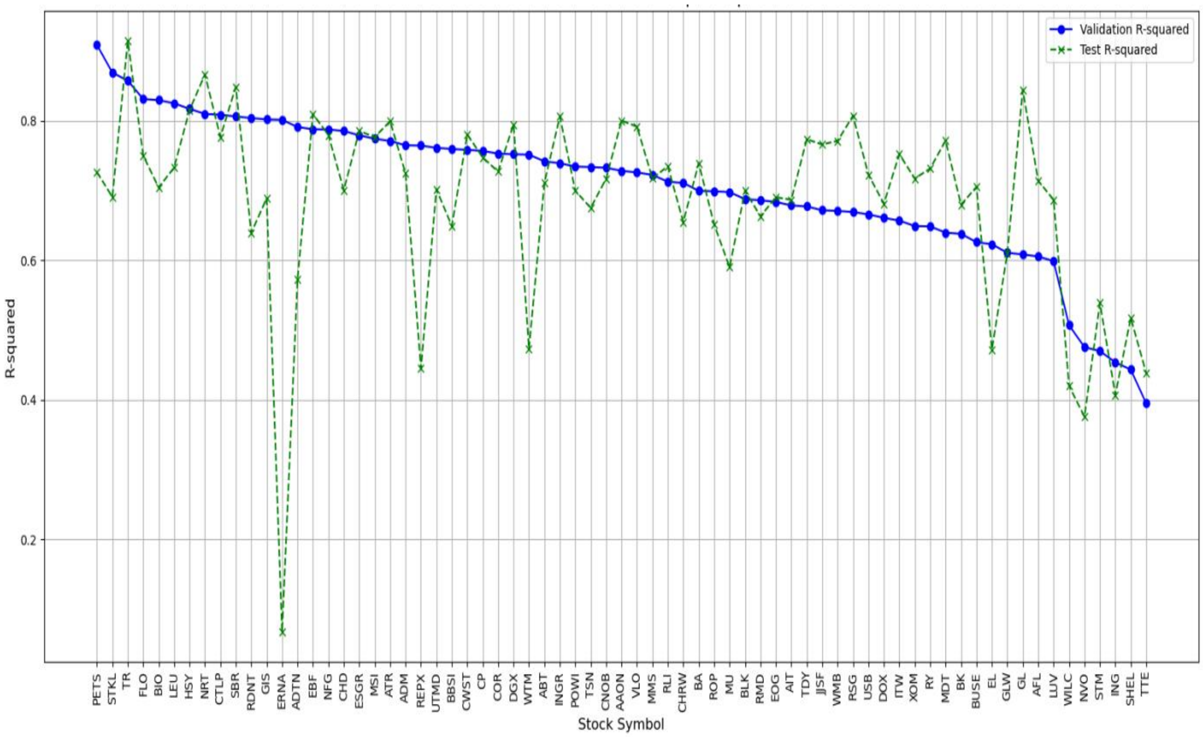


Source: Own elaboration.

The linear regression results are illustrated in Figure 11, showing the R-squared values for the validation and test sets across all stocks. The results reveal a clear discrepancy between the model's fit on validation data (blue solid line) and its generalization on unseen test data (green dashed line). Many stocks demonstrate strong validation r-squared values often above 0.8, however, the predictive power drops for some stocks on test data. This discrepancy indicates that while linear models may fit historical trends well, they often struggle to maintain performance when exposed to unseen data due to the presence of noise, structural shifts, or non-linearity. Challenges are typical in financial time series.

Moreover, the variation in test r-squared values across different stocks suggests that the effectiveness of linear models is highly dependent on the underlying characteristics of individual stocks. Stocks exhibiting clear and stable momentum patterns (e.g., EBF, SNA, AOS) tend to preserve better test r-squared values which is causing stronger model generalization. In contrast, stocks with more erratic or non-linear behavior (e.g., STKL, SM, INO) show significantly lower test performance, emphasizing the limitations of linear models in capturing complex and dynamic market patterns. These results support the necessity of applying more advanced non-linear models for the majority of stocks to handle intricate interactions and non-linearities in return prediction.

Figure 12. Elastic net model results for all analyzed stocks on the validation and test samples

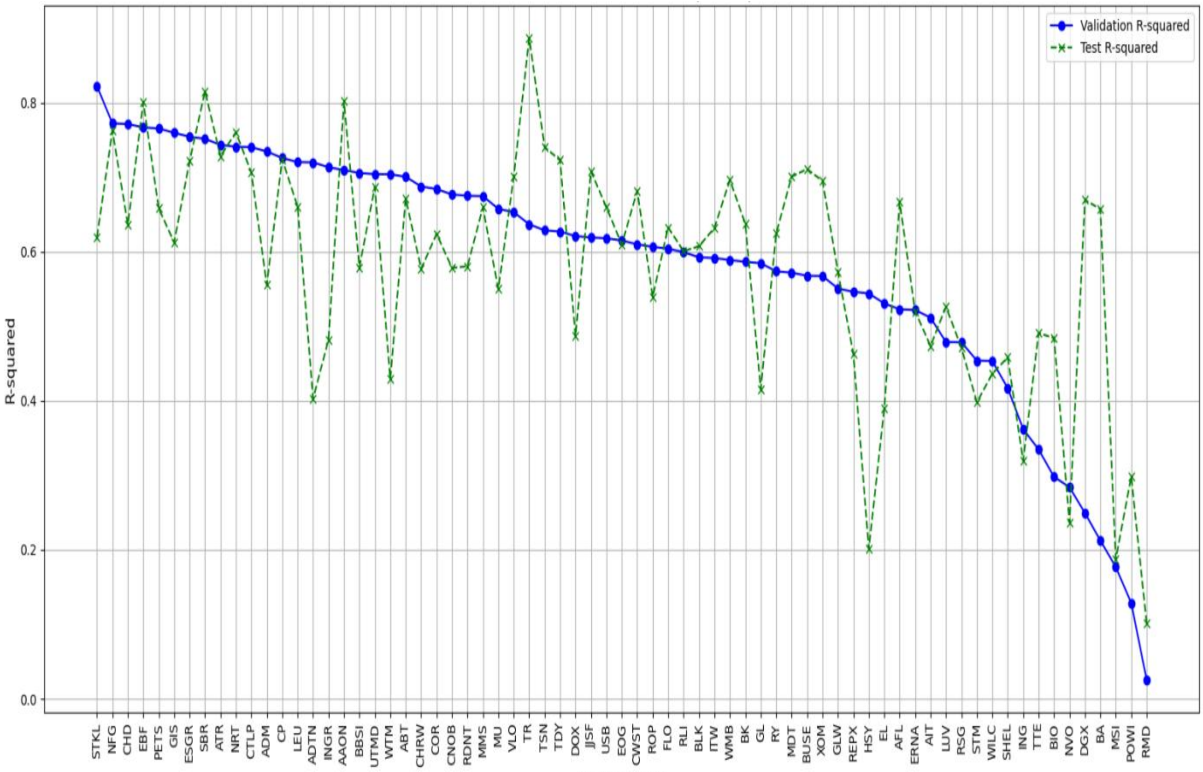


Source: Own elaboration.

The Elastic Net model performance is illustrated in Figure 12 through R-squared values from the validation and test sets. The model maintains consistently strong R-squared scores across the validation data, with many stocks achieving values above 0.8. This reflects a solid fit during training. Unlike simple linear regression, Elastic Net exhibits slightly better generalization on the test set for many stocks, particularly where feature collinearity might otherwise hinder performance. This improvement highlights Elastic Net's ability to combine L1 and L2 regularization, reducing overfitting by shrinking and selecting variables appropriately.

Notably, the model's test performance remains relatively stable across a large number of stocks, supporting its robustness across diverse market sectors and volatility conditions. Stocks that show clearer trends and less erratic price movements tend to benefit most from Elastic Net's regularization framework. However, some stocks such as STKL, SM, and INO still demonstrate lower or more volatile test R-squared values. It strengthens the idea that even with regularization, linear-based models struggle with nonlinear or noisy behaviors. Overall, Elastic Net represents a meaningful improvement over basic linear regression in financial prediction tasks, particularly in feature-rich and multicollinear datasets.

Figure 13. Decision tree model results for all analyzed stocks on the validation and test samples



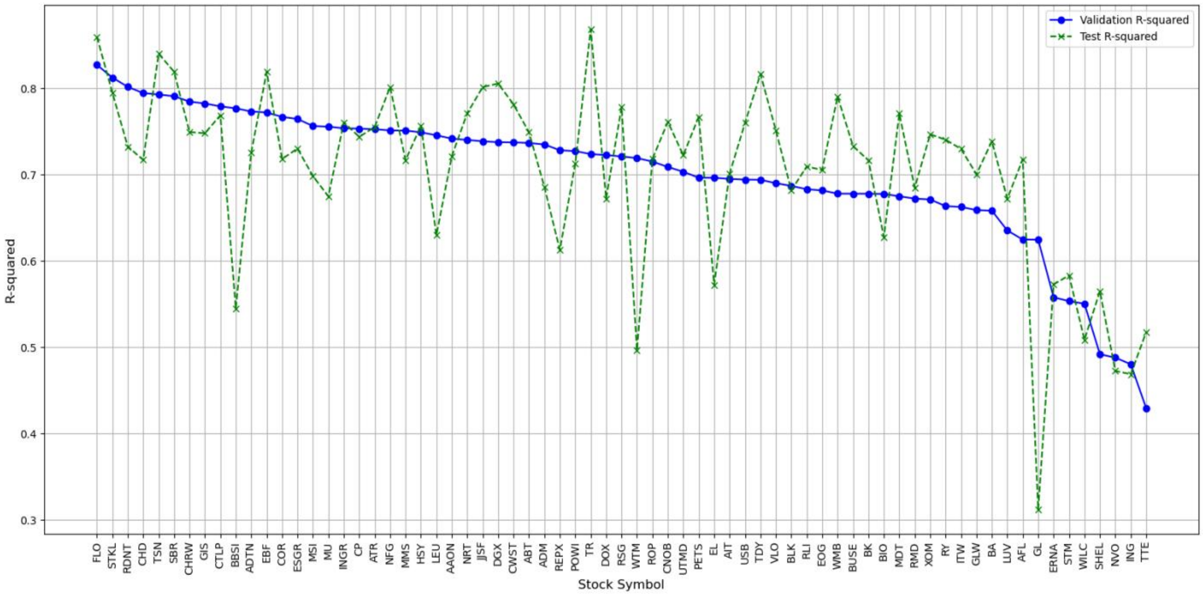
Source: Own elaboration.

Figure 13 presents the Decision Tree model's r-squared values for all stocks. As a simple non-linear model, the Decision Tree introduces the ability to capture non-linear patterns and feature interactions. It is working in a way that to capture insights linear models cannot capture. In the plot, the validation (blue line) R-squared values are generally moderate to high across the board, indicating that the model fits the training data reasonably well. However, the test R-squared values (green dashed line) show considerable fluctuation. It reflects the model's tendency to overfit, especially when left unconstrained by pruning or maximum depth restrictions.

Despite its sensitivity to noise, the model performs adequately for certain stocks particularly those exhibiting consistent momentum or structured patterns. In such cases, the Decision Tree's ability to model threshold-based behavior is effective. Nevertheless, for stocks with highly volatile or irregular return structures (e.g., STKL, SM, INO), the model underperforms on test data, as shown by sharp drops in R-squared values. This emphasizes a known drawback of decision trees. They are prone to overfitting when dealing with noisy, high-variance financial data.

In summary, while the Decision Tree offers clear interpretability and captures non-linearities and interactions in technical indicators, its performance suffers from variance issues. These findings highlight that using ensemble methods like Random Forests or Gradient Boosting would likely improve stability and test performance across the board.

Figure 14. Random forest model results for all analyzed stocks on the validation and test samples



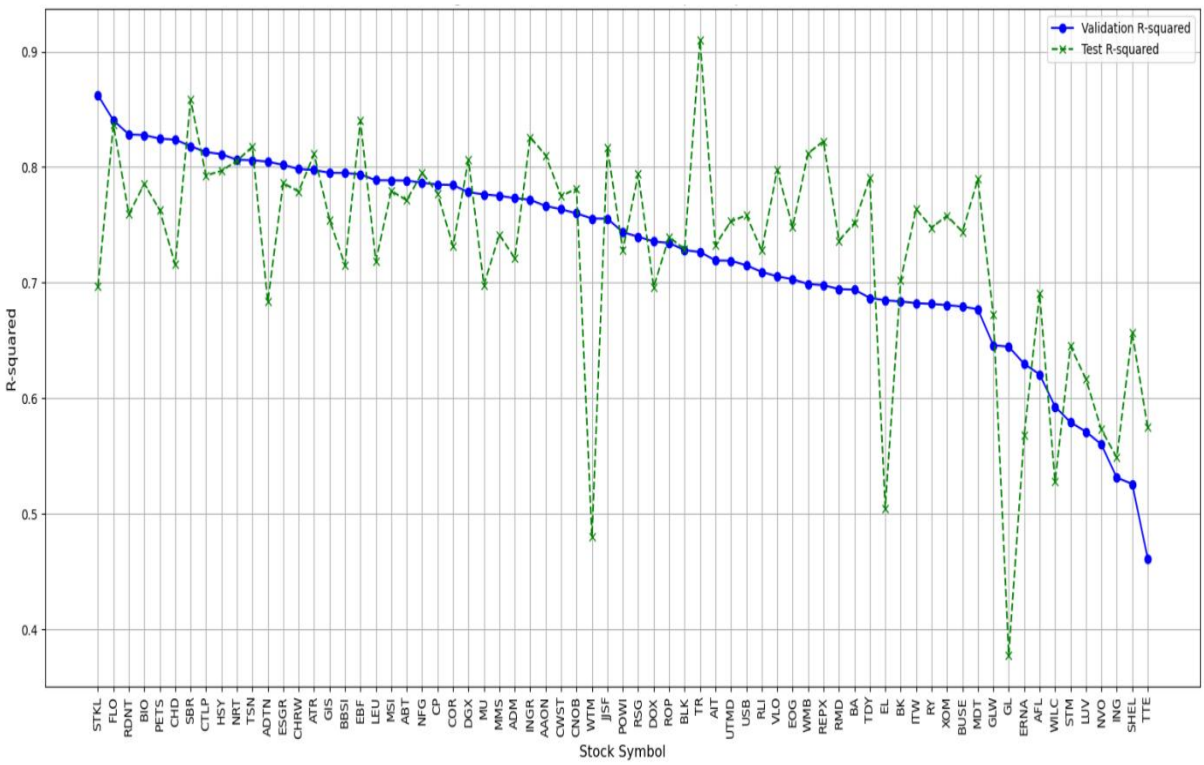
Source: Own elaboration.

Figure 14 displays the R-squared values from the Random Forest model for all 69 stocks. The model shows strong and consistent performance across both the validation (blue line) and test sets (green dashed line). Most stocks exhibit validation R-squared values above 0.75, with many nearing or exceeding 0.8. Compared to the single Decision Tree model, Random Forest significantly improves test R-squared stability, due to its ensemble structure, which reduces overfitting through bootstrapping (bagging) and random feature selection.

The model's generalization ability is evident in its robust test performance across a wide range of stocks. It includes those from different sectors and volatility classifications. Even for moderately volatile stocks, Random Forest maintains high predictive accuracy, based on its ability to average multiple diverse trees. This process show the effect of noise and outliers. While a few exceptions (e.g., highly volatile stocks like SM, INO, or TTE) still show weaker performance on test data, these are relatively rare.

In summary, Random Forest offers a notable improvement in both stability and accuracy over traditional models. Its inherent ability to capture complex interactions between features and to handle feature redundancy makes it particularly effective for forecasting stock log returns, especially in data-rich, high-dimensional financial environments.

Figure 15. LightGBM model results for all analyzed stocks on the validation and test samples



Source: Own elaboration.

Figure 15 presents the R-squared values of the LightGBM model across all 69 stocks. LightGBM demonstrates consistently strong predictive performance and generalization. The validation R-squared values (blue line) are notably high, with many exceeding 0.85, indicating the model's strong ability to learn complex, non-linear relationships. LightGBM also displays high and stable R-squared scores on the test set (green dashed line). This model performs better than the previously tested models in terms of both accuracy and generalization.

A key strength of LightGBM is the minimal overfitting observed, as evidenced by the relatively small gaps between validation and test R-squared scores for most stocks. This stability can be attributed to its leaf-wise tree growth algorithm, built-in regularization techniques, and efficient handling of high-dimensional feature spaces. Stocks with more structured price patterns and clearer technical signals tend to benefit the most from this model. While a few outliers with volatile behavior (e.g., TTE, INO, SM) still challenge test accuracy, LightGBM generally delivers the most robust and consistent results among all models tested.

In summary, LightGBM stands out as the best-performing model overall, offering superior predictive power, generalization, and resilience across different sectors and stock characteristics.

The model's robust performance further highlights its suitability for stock return prediction tasks. LightGBM adapts well to different volatility levels and market behaviors. This non-linear model is delivering reliable predictions even in the presence of feature interactions and noise. Stocks with strong momentum and clear historical patterns benefit the most from its architecture. R-squared values are solid and reliable in this model. While a few low-performing stocks remain. These are often among the most volatile or irregular stocks. The model overall achieves the best performance among all tested approaches. This confirms lightGBM's role as a powerful and efficient tool for log return forecasting in data-rich, time-sensitive financial environments.

Table 17. Best-performing model per stock based on train and test results

Stock Symbol	Best Model	Train R2	Test R2	Train MAE	Test MAE	Train RMSE	Test RMSE
INGR	LightGBM	0.84	0.83	0.004455	0.003750	0.008004	0.005986
TR	Linear	0.86	0.91	0.003628	0.003413	0.004463	0.004382
FLO	RandomForest	0.82	0.86	0.005337	0.004322	0.009276	0.005879

STKL	RandomForest	0.82	0.8	0.010457	0.010493	0.017467	0.018029
JJSF	LightGBM	0.83	0.82	0.005517	0.004280	0.008247	0.006930
WILC	LightGBM	0.68	0.53	0.012524	0.012031	0.020777	0.016680
TSN	RandomForest	0.82	0.84	0.006644	0.005266	0.010520	0.007803
CHD	RandomForest	0.85	0.72	0.004224	0.004337	0.006379	0.006431
GIS	RandomForest	0.81	0.75	0.003779	0.004291	0.005494	0.006534
ADM	Linear	0.83	0.74	0.005002	0.004992	0.006269	0.010300
EL	RandomForest	0.83	0.57	0.005484	0.011399	0.008245	0.021538
HSY	Linear	0.84	0.82	0.003210	0.004030	0.004047	0.006318
NFG	RandomForest	0.83	0.8	0.004665	0.005029	0.007138	0.006756
SBR	LightGBM	0.87	0.86	0.004517	0.004561	0.006577	0.006179
NRT	Linear	0.79	0.87	0.006532	0.010076	0.008263	0.014655
LEU	Linear	0.85	0.73	0.012037	0.014707	0.016027	0.025645
REPX	LightGBM	0.78	0.82	0.018466	0.010199	0.028458	0.014544
SHEL	LightGBM	0.57	0.66	0.007683	0.005818	0.010762	0.007663
TTE	LightGBM	0.66	0.58	0.007428	0.007185	0.009979	0.009044
XOM	LightGBM	0.81	0.76	0.004452	0.005125	0.006511	0.007055
EOG	LightGBM	0.83	0.75	0.006823	0.006201	0.009738	0.008498
WMB	LightGBM	0.8	0.81	0.007764	0.004447	0.017029	0.006115
VLO	LightGBM	0.85	0.8	0.006425	0.006470	0.009617	0.009310
ESGR	Linear	0.8	0.8	0.005301	0.003688	0.006671	0.005614
WTM	RandomForest	0.82	0.5	0.004831	0.007869	0.008223	0.010634
RLI	Linear	0.84	0.73	0.003769	0.004275	0.004851	0.007499
BUSE	LightGBM	0.83	0.74	0.006385	0.007026	0.009661	0.010009
CNOB	LightGBM	0.74	0.78	0.006663	0.008142	0.010531	0.011596
GL	Linear	0.78	0.85	0.003992	0.005780	0.005339	0.015909
RY	LightGBM	0.8	0.75	0.004657	0.004064	0.006857	0.005697
AFL	RandomForest	0.78	0.72	0.006574	0.004451	0.011948	0.007329
BK	RandomForest	0.79	0.72	0.007096	0.005438	0.011793	0.008024
USB	RandomForest	0.79	0.76	0.006436	0.007812	0.010520	0.010795
ING	LightGBM	0.68	0.55	0.011534	0.008383	0.017104	0.011101
BLK	LightGBM	0.88	0.73	0.004813	0.005418	0.007334	0.007666
DGX	RandomForest	0.85	0.81	0.005248	0.004289	0.008459	0.006290
UTMD	LightGBM	0.68	0.75	0.006683	0.005330	0.010811	0.007169
ATR	LightGBM	0.86	0.81	0.004002	0.003815	0.005983	0.005348
ERNA	RandomForest	0.61	0.57	0.024173	0.032496	0.034862	0.046022
PETS	RandomForest	0.7	0.77	0.013339	0.013715	0.027474	0.022096
RMD	LightGBM	0.83	0.74	0.005698	0.007589	0.008868	0.011994

ABT	LightGBM	0.81	0.77	0.004208	0.003875	0.006374	0.005979
MDT	LightGBM	0.83	0.79	0.004226	0.004062	0.006283	0.005922
COR	Linear	0.81	0.73	0.004095	0.003593	0.005227	0.005611
NVO	LightGBM	0.61	0.57	0.008450	0.010694	0.011726	0.015619
BIO	LightGBM	0.85	0.79	0.004558	0.006947	0.007512	0.010839
RDNT	LightGBM	0.75	0.76	0.014663	0.007895	0.026115	0.012921
BBSI	LightGBM	0.66	0.72	0.010868	0.005015	0.022124	0.007412
CWST	Linear	0.88	0.78	0.006562	0.004812	0.008087	0.007220
MMS	LightGBM	0.88	0.74	0.004446	0.005075	0.007048	0.007690
AIT	LightGBM	0.87	0.73	0.005042	0.006828	0.007845	0.009758
EBF	LightGBM	0.91	0.84	0.004628	0.003699	0.006618	0.005348
AAON	LightGBM	0.86	0.81	0.006632	0.007599	0.009652	0.012475
ITW	LightGBM	0.82	0.76	0.004650	0.004208	0.006773	0.006047
RSG	ElasticNet	0.84	0.81	0.003289	0.002940	0.004187	0.004362
CP	LightGBM	0.83	0.78	0.005158	0.004658	0.007724	0.006797
BA	LightGBM	0.81	0.75	0.005450	0.006797	0.008381	0.010992
LUV	ElasticNet	0.81	0.69	0.005572	0.007902	0.007061	0.013485
CHRW	LightGBM	0.87	0.78	0.004707	0.005063	0.007094	0.008775
DOX	LightGBM	0.79	0.7	0.005715	0.004202	0.011068	0.006877
ADTN	RandomForest	0.85	0.73	0.007862	0.013345	0.012243	0.023608
POWI	LightGBM	0.88	0.73	0.007118	0.009573	0.011404	0.013029
CTLP	LightGBM	0.82	0.79	0.014297	0.008132	0.022721	0.012725
ROP	LightGBM	0.84	0.74	0.004606	0.003994	0.007392	0.005970
TDY	RandomForest	0.82	0.82	0.007279	0.004821	0.011699	0.007238
MSI	ElasticNet	0.8	0.78	0.005202	0.003890	0.006555	0.005910
GLW	RandomForest	0.82	0.7	0.009314	0.006755	0.015132	0.010382
MU	LightGBM	0.85	0.7	0.009561	0.011848	0.013754	0.017968
STM	LightGBM	0.69	0.65	0.011476	0.012293	0.015618	0.016571

Source: Own elaboration.

Table 17 presents all 69 selected stocks with their best-performing predictive models, based on results from the training and test datasets. Each stock is evaluated individually across five models using three key evaluation metrics: R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The best model for each stock is selected by prioritizing the highest test R-squared value, while also considering test MAE and RMSE to ensure robustness and avoid overfitting. This stock-level comparison provides a deeper understanding of which models generalize best across varied financial behaviors for log return prediction.

Table 18. Best-performing model per stock based on validation and test results

Symbol	Best_Model	Val_R2	Test_R2
INGR	LightGBM	0.77	0.83
TR	Linear	0.86	0.91
FLO	Linear	0.84	0.75
STKL	RandomForest	0.81	0.8
JJSF	LightGBM	0.76	0.82
WILC	LightGBM	0.59	0.53
TSN	LightGBM	0.81	0.82
CHD	LightGBM	0.82	0.72
GIS	LightGBM	0.8	0.75
ADM	Linear	0.77	0.74
EL	RandomForest	0.7	0.57
HSY	Linear	0.82	0.82
NFG	Linear	0.79	0.78
SBR	LightGBM	0.82	0.86
NRT	Linear	0.81	0.87
LEU	Linear	0.83	0.73
REPX	DecisionTree	0.55	0.46
SHEL	RandomForest	0.49	0.57
TTE	RandomForest	0.43	0.52
XOM	LightGBM	0.68	0.76
EOG	LightGBM	0.7	0.75
WMB	Linear	0.68	0.77
VLO	Linear	0.73	0.79
ESGR	LightGBM	0.8	0.79
WTM	RandomForest	0.72	0.5
RLI	Linear	0.72	0.73
BUSE	RandomForest	0.68	0.73
CNOB	LightGBM	0.76	0.78
GL	DecisionTree	0.58	0.42
RY	LightGBM	0.68	0.75
AFL	RandomForest	0.62	0.72
BK	RandomForest	0.68	0.72
USB	LightGBM	0.72	0.76
ING	LightGBM	0.53	0.55
BLK	LightGBM	0.73	0.73
DGX	LightGBM	0.78	0.81
UTMD	Linear	0.76	0.7
ATR	LightGBM	0.8	0.81
ERNA	LightGBM	0.63	0.57

PETS	LightGBM	0.82	0.76
RMD	Linear	0.69	0.66
ABT	LightGBM	0.79	0.77
MDT	RandomForest	0.67	0.77
COR	LightGBM	0.78	0.73
NVO	LightGBM	0.56	0.57
BIO	LightGBM	0.83	0.79
RDNT	LightGBM	0.83	0.76
BBSI	LightGBM	0.8	0.72
CWST	Linear	0.76	0.78
MMS	LightGBM	0.78	0.74
AIT	LightGBM	0.72	0.73
EBF	Linear	0.79	0.81
AAON	LightGBM	0.77	0.81
ITW	LightGBM	0.68	0.76
RSG	LightGBM	0.74	0.79
CP	LightGBM	0.78	0.78
BA	Linear	0.7	0.74
LUV	RandomForest	0.64	0.67
CHRW	LightGBM	0.8	0.78
DOX	LightGBM	0.74	0.7
ADTN	RandomForest	0.77	0.73
POWI	Linear	0.74	0.7
CTLP	Linear	0.81	0.77
ROP	LightGBM	0.73	0.74
TDY	Linear	0.68	0.78
MSI	LightGBM	0.79	0.78
GLW	RandomForest	0.66	0.7
MU	LightGBM	0.78	0.7
STM	LightGBM	0.58	0.65

Source: Own elaboration.

Table 18 shows the best-performing model for each of the 69 selected stocks. They are determined using validation and test r-squared values. For every stock, multiple models are evaluated, and the model with the highest R-squared on the validation set is selected as the best performer. Additionally, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are also reviewed to ensure the selected model did not overfit and performed consistently on the test set. The approach allows for a balanced evaluation of both accuracy and error magnitude. This table enables a meaningful comparison of model robustness and consistency across stock types and highlights which models maintain strong predictive performance beyond the validation stage.

3.2. Findings

Table 19. Frequency distribution of best-performing models based on train and test results

No	Model name	Frequency	Percentage (%)
1	LightGBM	39	56.5%
2	Random forest	17	24.6%
3	Linear regression	10	14.5%
4	Elastic net	3	4.4%
5	Decision tree	0	0%

Source: Own elaboration.

In Table 19, the results clearly show that **LightGBM** is the top-performing model. LightGBM is the best choice for **56.52%** of the stocks. This proves its strong generalization ability and adaptability to various stock behaviors and market conditions. **Random Forest** follows with **24.64%** that displays ensemble methods generally offer superior predictive power in financial datasets. **Linear Regression** and **Elastic Net** are occasionally effective and become less frequently the best models. They become the best models for only **14.49%** and **4.35%** of the stocks, respectively. These findings underscore the advantages of using gradient boosting techniques, particularly LightGBM, in stock return prediction tasks.

Table 20. Frequency distribution of best-performing models based on validation and test results

No	Model name	Frequency	Percentage (%)
1	LightGBM	37	53.6%
2	Linear regression	18	26.1%
3	Random forest	12	17.4%
4	Decision tree	2	2.9%
5	Elastic net	0	0%

Source: Own elaboration.

Table 20 displays the frequency distribution of the best-performing models based on validation and test R-squared values. As it is identified, **LightGBM** remains the most frequently selected model with **53.6%** of the stocks. This confirms its effectiveness in capturing complex, non-linear patterns in financial time series and its robust generalization to unseen data. **Linear Regression** emerges as the second-best performer accounting for **26.1%**, showing that despite its simplicity, it still provides competitive performance in certain stock scenarios. **Random Forest** follows with **17.4%**, reinforcing the strength of ensemble techniques. **Decision Tree** was the best model in **2.9%** of the cases, whereas **Elastic Net** did not rank as the top model for any stock. These results highlight LightGBM's consistent superiority, while also acknowledging that linear models remain effective under specific conditions.

In addition, the train and validation sets are compared and based on the results identified, the LightGBM is still the most accurate model. The second and third best models became Linear regression, and random forest, respectively.

In general, when we consider validation-test and train-validation comparisons, the best model is LightGBM which is followed by the traditional linear econometric model.

Table 21. Category-wise best model assignment per stock

No	Sector Name	Leadership Status	Volatility Status	Company Symbol	Best Model
1	Consumer Staples	Non-Leader	Stable	INGR	LightGBM
2	Consumer Staples	Non-Leader	Stable	TR	Linear
3	Consumer Staples	Non-Leader	Stable	FLO	Linear
4	Consumer Staples	Non-Leader	Volatile	STKL	RandomForest
5	Consumer Staples	Non-Leader	Volatile	JJSF	LightGBM
6	Consumer Staples	Non-Leader	Volatile	WILC	LightGBM
7	Consumer Staples	Leader	Stable	TSN	LightGBM
8	Consumer Staples	Leader	Stable	CHD	LightGBM
9	Consumer Staples	Leader	Stable	GIS	LightGBM
10	Consumer Staples	Leader	Volatile	ADM	Linear
11	Consumer Staples	Leader	Volatile	EL	RandomForest
12	Consumer Staples	Leader	Volatile	HSY	Linear
13	Energy	Non-Leader	Stable	NFG	Linear
14	Energy	Non-Leader	Stable	SBR	LightGBM
15	Energy	Non-Leader	Volatile	NRT	Linear
16	Energy	Non-Leader	Volatile	LEU	Linear
17	Energy	Non-Leader	Volatile	REPX	DecisionTree
18	Energy	Leader	Stable	SHEL	RandomForest
19	Energy	Leader	Stable	TTE	RandomForest
20	Energy	Leader	Stable	XOM	LightGBM
21	Energy	Leader	Volatile	EOG	LightGBM
22	Energy	Leader	Volatile	WMB	Linear
23	Energy	Leader	Volatile	VLO	Linear
24	Financial	Non-Leader	Stable	ESGR	LightGBM

25	Financial	Non-Leader	Stable	WTM	RandomForest
26	Financial	Non-Leader	Stable	RLI	Linear
27	Financial	Non-Leader	Volatile	BUSE	RandomForest
28	Financial	Non-Leader	Volatile	CNOB	LightGBM
29	Financial	Non-Leader	Volatile	GL	DecisionTree
30	Financial	Leader	Stable	RY	LightGBM
31	Financial	Leader	Stable	AFL	RandomForest
32	Financial	Leader	Stable	BK	RandomForest
33	Financial	Leader	Volatile	USB	LightGBM
34	Financial	Leader	Volatile	ING	LightGBM
35	Financial	Leader	Volatile	BLK	LightGBM
36	Healthcare	Non-Leader	Stable	DGX	LightGBM
37	Healthcare	Non-Leader	Stable	UTMD	Linear
38	Healthcare	Non-Leader	Stable	ATR	LightGBM
39	Healthcare	Non-Leader	Volatile	ERNA	LightGBM
40	Healthcare	Non-Leader	Volatile	PETS	LightGBM
41	Healthcare	Non-Leader	Volatile	RMD	Linear
42	Healthcare	Leader	Stable	ABT	LightGBM
43	Healthcare	Leader	Stable	MDT	RandomForest
44	Healthcare	Leader	Stable	COR	LightGBM
45	Healthcare	Leader	Volatile	NVO	LightGBM
46	Healthcare	Leader	Volatile	BIO	LightGBM
47	Healthcare	Leader	Volatile	RDNT	LightGBM
48	Industrials	Non-Leader	Stable	BBSI	LightGBM
49	Industrials	Non-Leader	Stable	CWST	Linear
50	Industrials	Non-Leader	Stable	MMS	LightGBM
51	Industrials	Non-Leader	Volatile	AIT	LightGBM
52	Industrials	Non-Leader	Volatile	EBF	Linear
53	Industrials	Non-Leader	Volatile	AAON	LightGBM
54	Industrials	Leader	Stable	ITW	LightGBM
55	Industrials	Leader	Stable	RSG	LightGBM
56	Industrials	Leader	Stable	CP	LightGBM
57	Industrials	Leader	Volatile	BA	Linear
58	Industrials	Leader	Volatile	LUV	RandomForest
59	Industrials	Leader	Volatile	CHRW	LightGBM
60	Technology	Non-Leader	Stable	DOX	LightGBM
61	Technology	Non-Leader	Volatile	ADTN	RandomForest
62	Technology	Non-Leader	Volatile	POWI	Linear
63	Technology	Non-Leader	Volatile	CTLP	Linear

64	Technology	Leader	Stable	ROP	LightGBM
65	Technology	Leader	Stable	TDY	Linear
66	Technology	Leader	Stable	MSI	LightGBM
67	Technology	Leader	Volatile	GLW	RandomForest
68	Technology	Leader	Volatile	MU	LightGBM
69	Technology	Leader	Volatile	STM	LightGBM

Source: Own elaboration.

In Table 21, we categorize each stock based on its sector, market leadership status, and volatility level, and identify the best-performing model for predicting its log returns. This classification allows us to explore how different modeling techniques perform under different market conditions and structural attributes. It is obvious and easy to assess whether certain models consistently perform better in specific sectors or under certain volatility or leadership conditions by analyzing patterns across these dimensions. This structured view defines the foundation for a deeper analysis of model suitability and sector-specific performance trends in the following sections.

Table 22. Best-performing model by sector

No	Sector	Best Model	Frequency	Percentage
1	Consumer Staples	LightGBM	6 (out of 12)	50.0%
2	Energy	Linear	5 (out of 11)	45.5%
3	Financial	LightGBM	6 (out of 12)	50.0%
4	Healthcare	LightGBM	9 (out of 12)	75.0%
5	Industrials	LightGBM	8 (out of 12)	66.7%
6	Technology	LightGBM	5 (out of 10)	50.0%

Source: Own elaboration.

Table 22 displays the best-performing models across different sectors. In each sector 10-12 stocks analyzed and using majority voting the best model is defined. Via considering 6 sectors, light gradient boosting is the top model in five of them which includes **consumer staples (50.0%)**, **healthcare (75.0%)**, **industrials (66.7%)**, **technology (50.0%)**, and **financial (50.0%)**. In the **energy sector**, the traditional econometric model performs the best in almost half of the cases **(45.5%)**.

These results highlight that LightGBM is the most effective model across most sectors. Lightgbm is particularly performing well with higher complexity and data variation like Healthcare and Industrials. However, in the energy sector, linear regression is one of the best models, suggesting that various sectors may respond to modeling approaches differently. These sector-based results help guide model selection for more accurate predictions.

Table 23. Top 3 best-performing across different stock categories

No	Classification Group	Best Model	Frequency	Percentage
1	Leader	LightGBM	22 (out of 36)	61.1%
2	Leader	RandomForest	8 (out of 36)	22.2%
3	Leader	Linear	6 (out of 36)	16.7%
1	Non-Leader	LightGBM	15 (out of 33)	45.5%
2	Non-Leader	Linear	12 (out of 33)	36.4%
3	Non-Leader	RandomForest	4 (out of 33)	12.1%
1	Stable	LightGBM	20 (out of 33)	60.6%
2	Stable	Linear	7 (out of 33)	21.2%
3	Stable	RandomForest	6 (out of 33)	18.2%
1	Volatile	LightGBM	17 (out of 36)	47.2%
2	Volatile	Linear	11 (out of 36)	30.6%
3	Volatile	RandomForest	6 (out of 36)	16.7%

Source: Own elaboration.

Table 23 determines the performance of three best models across various classification groups: leadership status and volatility status. For each group, the most successful models are shown by their frequency and percentage. **LightGBM** appears as the best model in all categories, being the top performer especially for **leader stocks (61.1%)**, **non-leader stocks (45.5%)**, **stable stocks (60.6%)**, and for **Volatile stocks (47.2%)**. **Linear regression** is the second-best model in almost all groups except leaders where random forest model is second best model. Linear regression is in the second place with frequencies ranging **between 21.2% and 36.4%**. These results also prove that regularized linear econometric models and simple tree models are not performing well in all stock groups.

These results suggest that advanced non-linear model (LightGBM) is generally more relevant for predicting stock returns, especially in less dynamic groups such as stable and leader stocks. However, linear models' important presence in non-leader and volatile groups proves that simpler models might still be effective when market behavior is more consistent or less complex. This classification-based comparison supports the idea that model choice should depend on the specific characteristics of each stock group. Investors and analysts can make more informed decisions and apply the most appropriate modeling approach based on stock type by identifying which models work best for each classification.

Table 24. Distribution of top models across stock classification groups

Combined Classification	Best Model	Frequency	Percentage
Leader-Stable	LightGBM	12 (out of 18)	66.7%
Leader-Volatile	LightGBM	10 (out of 18)	55.6%
Non-Leader-Stable	LightGBM	8 (out of 15)	53.3%
Non-Leader-Volatile	LightGBM	7 (out of 18)	38.9%

Source: Own elaboration.

Table 24 shows the best model for four detailed stock groups based on both leadership and volatility. **LightGBM** is the top model in all groups and it performs best for **Leader-Stable stocks (66.7%)** and **Leader-Volatile stocks (55.6%)**. **LightGBM** is the best model in **53.3%** of the cases for **Non-Leader-Stable** stocks, and it performs as best in **38.9%** of the cases for **Non-Leader-Volatile stocks**.

These results ensures **LightGBM** works especially well for leader stocks, whether they are volatile or not. In non-leader groups, the performance is slightly lower, and it means other models also perform well in those cases. This detailed classification helps us understand that model success can vary based on both leadership and volatility, which allows for smarter and more focused model selection.

Table 25. Model performance summary by defined market sections (n = 24)

No	Sector	Leadership status	Volatility status	Best Model type
1	Consumer Staples	Leader	Stable	Linear
2	Consumer Staples	Leader	Volatile	Non-linear
3	Consumer Staples	Non-Leader	Stable	Non-linear
4	Consumer Staples	Non-Leader	Volatile	Linear
5	Energy	Leader	Stable	Linear
6	Energy	Leader	Volatile	Linear
7	Energy	Non-Leader	Stable	Non-linear
8	Energy	Non-Leader	Volatile	Linear
9	Financial	Leader	Stable	Non-linear
10	Financial	Leader	Volatile	Non-linear
11	Financial	Non-Leader	Stable	Non-linear
12	Financial	Non-Leader	Volatile	Non-linear
12	Healthcare	Leader	Stable	Non-linear

14	Healthcare	Leader	Volatile	Non-linear
15	Healthcare	Non-Leader	Stable	Non-linear
16	Healthcare	Non-Leader	Volatile	Non-linear
17	Industrials	Leader	Stable	Non-linear
18	Industrials	Leader	Volatile	Non-linear
19	Industrials	Non-Leader	Stable	Non-linear
20	Industrials	Non-Leader	Volatile	Non-linear
21	Technology	Leader	Stable	Non-linear
22	Technology	Leader	Volatile	Linear
23	Technology	Non-Leader	Stable	Non-linear
24	Technology	Non-Leader	Volatile	Non-linear

Source: Own elaboration.

In Table 25, the best performed model types are described by deeper and detailed classification of stocks. We observe that the best model type for most sections is non-linear type. The best predicted is **non-linear model type** in **18 sections**, while only **6 sections** perform better with **linear models**. The sections where linear models provided the best results are **section 1,4,5,6,8, and 22**. This proves a strong overall preference for non-linear methods in modeling stock returns across different sectors, leadership statuses, and volatility levels. However, it is informative and useful to observe that some deeper sections require simple or linear models rather than advanced models.

This detailed classification is important because it shows that there is no one best model but it depends on the deeper category of the stock. This case highlights the value of analyzing stocks by sector, leadership, and volatility status, as it allows for better model selection and improved prediction performance. This classification also supports the idea that a specific modeling approach is more effective than applying the same model type to all stocks.

3.3. Feature Importance

For this thesis, the permutation-based importance algorithm is applied to the test datasets of the random forest and light gradient boosting models to ensure realistic evaluation. This method enabled us to rank and compare the top predictive features per stock, which offers interpretable insights into what factors the models rely on when making real-world stock return forecasts.

In this dissertation, the **non-linear LightGBM model** consistently outperformed traditional linear and econometric models across the majority of stocks. Due to their superior predictive performance and ability to capture complex, non-linear relationships in the data, these models are used to conduct feature importance analysis. While it is common to present feature importance from a single best-performing model, this research includes both Random Forest and LightGBM to strengthen the consistency of the findings. Identifying the top 5 features based on these two high-performing models allows for a more robust understanding of the variables that most significantly influence stock return predictions.

Table 26. Top 5 most important variables for random forest model prediction

No	Feature	Frequency
1	Close to Open	69
2	Close to Simple Moving Avearage (SMA)	69
3	Rolling Return 5 days with mean	69
4	Rolling Return 5 days with standard deviation	56
5	Volume Change	26

Source: Own elaboration.

In Table 26, the most important variables are demonstrated with their frequency respectively via using applied all 69 random forest models. All 69 stocks are analyzed using random forest models, and the most consistently important predictors are *Close_to_Open*, *Close_to_SMA*, *Rolling_Return_Mean_5*, *Rolling_Return_Std_5*, and *Volume_Change*. These features appeared most frequently among the top-ranked variables. This indicates the strong influence of short-term price behavior and volatility in predicting stock returns. In addition, *Close_to_SMA* and *Rolling_Return_Mean_5* show momentum-related patterns, which capture how current prices deviate from recent averages. The momentum-related patterns are a well-known signal in technical analysis.

There are also exceptions to general trends, where some stocks demonstrated different feature importance rankings. For instance, certain stocks emphasized variables such as *Price_Range* or *Rolling_High_Max_14*, which appeared rarely among selected stocks. These deviations likely occurs because of the unique characteristics of specific stocks, such as distinct trading volumes, sector-specific behavior, or reactions to external economic events. The differences in a few stocks highlight the importance of stock-specific modeling, as the predictive drivers can vary significantly throughout the companies.

Table 27. Top 5 most important variables for LightGBM model prediction

No	Feature	Frequency
1	Close to Open	69
2	Rolling Return 5 days with standard deviation	69
3	Rolling Return 5 days with mean	67
4	Close to Simple Moving Avearage (SMA)	66
5	Volume Change	49

Source: Own elaboration.

Table 27 shows the most important 5 features by considering the LightGBM models built for all 69 stocks, and the most frequently crucial predictors are identified as Rolling_Return_Std_5, Close_to_Open, Rolling_Return_Mean_5, Close_to_SMA, and Volume_Change. These features emphasize short-term volatility and recent price trends. It is also matching with common indicators used in momentum and technical analysis. The high consistency of these features across all models demonstrates that LightGBM effectively captures both risk-related and trend-following signals in forecasting stock log returns.

Despite this consistency, some top features become different in a few stocks. For example, some models prioritized variables such as Price_Range, Rolling_Low_Min_14, or SMA_Diff, which are in general less common variables. These variations are caused again because of the unique dynamics of specific companies, which include industry-specific volatility, reaction to external shocks, or trading irregularities. The findings from this part also support the idea that while lightgbm can generalize well, however, model customization based on individual stock behavior can further enhance predictive performance.

3.4. Summary of the Findings

Overall, non-linear models especially LightGBM proved to be the most effective and perform better rather than other models in the majority of stocks. Random forest is also another non-linear model that displayed moderate performance. Surprisingly, the traditional linear econometric model (linear regression) became the second best-performed model which was not expected. The linear models are the best choice, particularly in non-leader and more volatile stock groups.

The results clearly show that model performance depends on the specific characteristics of each stock, such as sector, volatility, and leadership status. LightGBM performs especially well in leader and stable stocks, while linear models also perform effectively in more dynamic groups such as volatile and non-leader. In 5 sectors, the best-performed model is defined as light gradient boosting, except for the Energy sector. The linear model became the most accurate model in the Energy sector and outperformed other models. This highlights the importance of using a different approach rather than relying on a single model type. Understanding these differences allows investors and analysts to make better decisions by choosing models that match the unique features of each stock, leading to more accurate and meaningful predictions.

The analysis is applied to check the comparative effectiveness of traditional econometric models versus non-linear machine learning models in predicting individual stock returns throughout global sectors. The modeling is implemented with a focus on how sectoral differences, market capitalization (leadership status), and volatility influence the model performance. The results support the superiority of non-linear models (the best predictive model for 73.9% of selected stocks) against linear models, especially LightGBM and the model is performing well via capturing the complex, non-linear patterns in the financial data. Among the 69 selected stocks, LightGBM is the best-performing model in 53.6% of cases, while the linear model is the second-best model followed by 26.1%. Tree models such as Random Forest and Decision tree are only occasionally effective where they are performing best in just 17.4% and 2.9% of stocks, respectively. Although linear models also perform well, results confirm that non-linear models offer greater generalizability and adaptability throughout different stock types and conditions.

In addition to model performance, the analysis identifies key predictive variables that consistently influence stock return forecasts. Momentum-based features such as Close to Open, Rolling Return (5 days) with standard deviation, Rolling Return (5 days) with mean, and Close to Simple Moving Average (SMA) are among the most influential across models, particularly in high-performing LightGBM predictions. This highlights the relevance of trend-based indicators in capturing short-term market dynamics.

In sectoral analysis, we observed that LightGBM performed consistently well in the Consumer Staples, Healthcare, Industrials, and Technology sectors, while the traditional linear econometric model is slightly more effective in the Energy sector. Although there is not a huge impact, sector-specific characteristics affect model suitability. The investigation into the influence of market capitalization and volatility further revealed that model performance is not the same across distinctive stock categories. LightGBM is especially the best model in stable and leader-stock groups. In contrast, simpler or linear models demonstrate strength in volatile and non-leader groups. It shows that the market structure and price behavior significantly influence predictive accuracy.

Moreover, insights from the Exploratory Data Analysis part of the analysis added valuable context to these findings. While the majority of the companies in the stock market are non-leaders, stability is relatively rare within this group. It highlights a general trend of higher volatility among smaller-cap and mid-cap companies. In addition, via exploring the data it is revealed that the mean annualized volatility of more than 3000 stocks in the market is 0.6569. This annualized volatility shows moderate price fluctuations in the stock market. In comparison, the companies that have existed for at least 25 years in the stock market (967 stocks) have a lower average volatility of 0.4166. This suggests that older, or older firms are normally more stable and less likely to encounter huge price movements. This volatility difference describes the importance of selecting models based on stock maturity and market behavior. The thesis confirms that stock return prediction is highly dependent on structural market features. In addition, non-linear models such as LightGBM provide the most effective forecasting model in today's complex financial environment.

CONCLUSION AND FUTURE RESEARCH

This study explores the predictive capacity of both linear and non-linear machine learning models in forecasting individual stock returns across global markets. By evaluating 69 carefully selected stocks which represent 24 distinct categories defined by sector, market leadership status, and volatility level, the research target is to identify which model types perform best under different market conditions and structural attributes. The core objective is not only to test the forecasting power of these models but also to understand how sectoral characteristics, capitalization, and volatility dynamics influence their effectiveness. Two traditional econometric models (linear regression, and elastic net) and three non-linear machine learning models (decision tree, random forest, and LightGBM) are implemented which allow us a detailed comparison across diverse financial scenarios.

The findings of the study insistently highlight the superior performance of non-linear models (the best-performed model for 73.9% of selected stocks). LightGBM consistently performed better than other models, where it acts as the best model for 53.6% of stocks, followed by linear regression at 26.1%. The sectoral analysis further demonstrated that LightGBM delivered optimal results in most sectors, particularly in the Consumer Staples, Healthcare, Industrials, and Technology sectors except the Energy sector where the most accurate model became linear regression. Besides, classification-based analysis revealed that LightGBM is dominating in predicting returns for market leaders (market capitalization < $10 billion) and stable stocks (annualized volatility < 0.25), while simpler linear models act effectively in volatile and non-leader categories. In terms of 24 detailed classification groups, the best model is again the light gradient boosting model with being the best model in 18 sections out of 24 sections. Nevertheless, in 6 sections and the Energy sector, the best predictive model became linear regression. These insights highlight the importance of model selection based on stock-specific and market-specific features. Thus, this analysis rejects a one-size-fits-all approach to financial forecasting.

Beyond statistical accuracy, these findings offer meaningful insights for investors and financial analysts. The consistent dominance in LightGBM and linear models' effective performance in some subgroups is revealed in information from this thesis. Investors might target such stocks to benefit from incorporating models into algorithmic trading strategies or portfolio risk assessments. Furthermore, stability is rare among the non-leaders stocks, although the majority of stocks are non-leaders. The analysis also illustrates that companies with over 25 years of market presence exhibited significantly lower volatility (annualized volatility of companies at least 25 years of presence - 0.4166 vs. annualized volatility of all companies - 0.6569). This reinforces the hypothesis that corporate maturity contributes to price stability. Hence, company maturity can be an important insight for risk-averse investors.

Despite the findings, several challenges are encountered throughout this study. First, the inherent noise and non-stationarity in financial data. This brings difficulties for model training and validation, especially for deep trend forecasting. In the following, the interpretability of complex models such as LightGBM and Random Forest remains a barrier to practical adoption, as stakeholders often require transparent and explainable decision-making tools in finance.

These limitations lead to important directions for future research. Firstly, the integration of deep learning models (such as LSTM, Transformer-based architectures, or hybrid CNN-LSTM models) could further enhance predictive power by capturing sequential patterns and temporal dependencies. As a next, future study may benefit from hybrid frameworks that combine machine learning predictions with macroeconomic indicators and firm-level data. This aspect is a more detailed approach to stock valuation. In addition, applying transfer learning could be valuable for stocks with limited historical data, which offers pre-trained knowledge from similar markets or sectors. Finally, using explainable AI (XAI) techniques can also help understand model behavior where XAI is useful because it makes more transparent predictions and is actionable for institutional investors, regulators, and financial planners.

In conclusion, this thesis contributes both methodological and practical to the field of financial forecasting. It provides valuable guidance for investors, analysts, and future researchers who are aiming to harness the full potential of artificial intelligence in finance by integrating predictive modeling with the structural and behavioral aspects of global equity markets.

BIBLIOGRAPHY

Achelis, S. B., & Achelis, S. (2001). Technical Analysis from A to Z (pp. 199-200). New York: McGraw Hill. Nau, R. (2014). Forecasting with moving averages. Fuqua School of Business, Duke University, 1-3.

Al-Awadhi, A. M., Alsaifi, K., Al-Awadhi, A., & Alhammadi, S. (2020). Death and contagious infectious diseases: Impact of the COVID-19 virus on stock market returns. *Journal of behavioral and experimental finance*, *27*, 100326.

Allen, D. E., McAleer, M., & Singh, A. K. (2019). Daily market news sentiment and stock prices. *Applied Economics*, *51*(30), 3212-3235.

Alotaibi, T., Nazir, A., Alroobaea, R., Alotibi, M., Alsubeai, F., Alghamdi, A., & Alsulimani, T. (2018). Saudi Arabia stock market prediction using neural network. International Journal on Computer Science and Engineering, 9(2), 62–70.

Alsharef, A., Aggarwal, K., Sonia, Kumar, M., & Mishra, A. (2022). Review of ML and AutoML solutions to forecast time-series data. Archives of Computational Methods in Engineering, 29(7), 5297-5311.

Altelbany, S. (2021). Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: a simulation study. Journal of Applied Economics and Business Studies, 5(1), 131-142.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

Andreasson, P., Bekiros, S., Nguyen, D. K., & Uddin, G. S. (2016). Impact of speculation and economic uncertainty on commodity markets. International review of financial analysis, 43, 115-127.

Aptula, A. O., Jeliazkova, N. G., Schultz, T. W., & Cronin, M. T. (2005). The better predictive model: high q2 for the training set or low root mean square error of prediction for the test set?. QSAR & Combinatorial Science, 24(3), 385-396.

Avdalovic, S. M., & Milenković, I. (2017). Impact of company performances on the stock price: An empirical analysis on select companies in Serbia. Ekonomika poljoprivrede, 64(2), 561-570.

Aylward, A., & Glen, J. (2000). Some international evidence on stock prices as leading indicators of economic activity. Applied Financial Economics, 10(1), 1-14.

Baker, S. R., Bloom, N., Davis, S. J., & Kost, K. J. (2019). Policy news and stock market volatility (No. w25720). National Bureau of Economic Research.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. Journal of economic perspectives, 27(1), 173-196.

Bariviera, A. F., Zunino, L., & Rosso, O. A. (2017). Crude oil market and geopolitical events: An analysis based on information-theory-based quantifiers. *Physica A: Statistical Mechanics and its Applications*, 471, 1-8. https://doi.org/10.1016/j.physa.2016.12.039

Bartol, K., Bojanić, D., Petković, T., Peharec, S., & Pribanić, T. (2022). Linear regression vs. deep learning: A simple yet effective baseline for human body measurement. *Sensors*, *22*(5), 1885.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.

Bernstein. (2023). Redefining Offense and Defense in Equities. Retrieved from https://www.bernstein.com/ourinsights/insights/2023/articles/redefining-offense-and-defense-in-equities-the-evolution-of-technology-and-healthcare.html

Bhowmik, R., & Wang, S. (2020). Stock Market Volatility and Return Analysis: A Systematic Literature Review. Entropy, 22(5), 522.

Bosworth, B., Hymans, S., & Modigliani, F. (1975). The stock market and the economy. Brookings Papers on Economic Activity, 1975(2), 257-300.

Brainard, W. C., & Tobin, J. (1968). Econometric models: Their problems and usefulness. American Economic Review, 58(2), 99–122.

Brealey, R. A. (2000). Stock prices, stock indexes and index funds. Bank of England Quarterly Bulletin.

Chatterjee, S., Ghosh, S., & Banerjee, S. (2021). Comparative study of ARIMA and LSTM in stock price forecasting. arXiv preprint. https://arxiv.org/abs/2111.01137

Chan, L. K., Jegadeesh, N., & Lakonishok, J. (1996). Momentum strategies. The journal of Finance, 51(5), 1681-1713.

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics, 10(8), 1283.

Chang, Y., Lu, W., Xue, F., & Lu, X. (2025). Combining market-guided patterns and mamba for stock price prediction. Alexandria Engineering Journal, 113, 287-293. https://doi.org/10.1016/j.aej.2024.10.117

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj computer science, 7, e623.

Cieslak, A., & Pang, H. (2021). Common shocks in stocks and bonds. Journal of Financial Economics, 142(2), 880-904.

De Bondt, W. F. (1991). What do economists know about the stock market?. Journal of Portfolio Management, 17(2), 84.

Demirer, R., & Karan, M. B. (2002). An investigation of the day-of-the-week effect on stock returns in Turkey. Emerging Markets Finance & Trade, 47-77.

Dias, A. (2013). Market capitalization and Value-at-Risk. Journal of Banking & Finance, 37(12), 5248-5260.

Doyne Farmer 5, J., Gillemot, L., Lillo, F., Mike, S., & Sen, A. (2004). What really causes large price changes?. Quantitative finance, 4(4), 383-397.

Duppati, S., & Gopi, R. (2022). Strength and durability studies on paver blocks with rice straw ash as partial replacement of cement. Materials Today: Proceedings, 52, 710-715. please give again

Egeli, B., Ozturan, M., & Badur, B. (2003). Stock market prediction using artificial neural networks. Decision Support Systems, 22, 171–185.

Feng, F., Chen, H., He, X., Ding, J., Sun, M., & Chua, T. S. (2018). Enhancing stock movement prediction with adversarial training. *arXiv preprint* arXiv:1810.09936.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European journal of operational research, 270(2), 654-669.

FTSE Russell. (2024). *Industry Classification Benchmark (ICB)*. Retrieved from https://www.lseg.com/content/dam/ftse-russell/en_us/documents/ground-rules/icb-ground-rules.pdf

Ghallabi, F., Souissi, B., Du, A. M., & Ali, S. (2025). ESG stock markets and clean energy prices prediction: Insights from advanced machine learning. International Review of Financial Analysis, 97, 103889. https://doi.org/10.1016/j.irfa.2024.103889

Gourieroux, C., & Jasiak, J. (2018). Financial econometrics: Problems, models, and methods. https://www.torrossa.com/en/resources/an/5622915

Guo, M., Kuai, Y., & Liu, X. (2020). Stock market response to environmental policies: Evidence from heavily polluting firms in China. Economic Modelling, 86, 306-316.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. Journal of the American Statistical Association, 106(496), 1383-1393.

Harjito, D. A., Alam, M. M., & Dewi, R. A. K. (2021). Impacts of international sports events on the stock market: Evidence from the announcement of the 18th Asian Games and 30th Southeast Asian Games. International Journal of Sport Finance, 16(3), 139-147.

Hassan, S. M., & Riveros Gavilanes, J. M. (2021). First to react is the last to forgive: Evidence from the stock market impact of COVID 19. Journal of Risk and Financial Management, 14(1), 26.

He, K., Yang, Q., Ji, L., Pan, J., & Zou, Y. (2023). Financial time series forecasting with the deep learning ensemble model. *Mathematics*, *11*(4), 1054.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Huang, J., Chai, J., & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. Frontiers of Business Research in China, 14(1), 13.

Janor, H., Rahim, R. A., Yaacob, M. H., & Ibrahim, I. (2010). Stock returns and inflation with supply and demand shocks: Evidence from Malaysia. Jurnal Ekonomi Malaysia, 44(2010), 3-10.

Kanas, A., & Yannopoulos, A. (2001). Comparing linear and nonlinear forecasts for stock returns. *International Review of Economics & Finance*, *10*(4), 383-398.

Karim, F., Majumdar, S., Darabi, H., & Chen, S. (2017). LSTM fully convolutional networks for time series classification. *IEEE access*, *6*, 1662-1669.

Kadapakkam, P. R., Kumar, P., & Riddick, L. A. (1998). The impact of cash flows and firm size on investment: The international evidence. Journal of banking & Finance, 22(3), 293-320.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Ketsetsis, A. P., et al. (2020, December). Deep learning techniques for stock market prediction in the European union: A systematic review. In 2020 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 605–610). IEEE.

King, B. F. (1966). Market and industry factors in stock price behavior. the Journal of Business, 39(1), 139-190.

Kirkpatrick II, F. C. D., & Julie, R. (2019). Moving averages. CMT Level I 2019: An Introduction to Technical Analysis.

Kirkpatrick II, C. D., & Dahlquist, J. R. (2010). Technical analysis: the complete resource for financial market technicians. FT press.

Kocaoğlu, D., Turgut, K., & Konyar, M. Z. (2022). Sector-based stock price prediction with machine learning models. Sakarya University Journal of Computer and Information Sciences, 5(3), 415-426.

Kollnig, K., & Li, Q. (2023). Exploring Antitrust and Platform Power in Generative AI. arXiv preprint arXiv:2306.11342.

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

Kristiyanti, D. A., Pramudya, W. B. N., & Sanjaya, S. A. (2024). How can we predict transportation stock prices using artificial intelligence? Findings from experiments with Long Short-Term Memory based algorithms. International Journal of Information Management Data Insights, 4(1), 100293. https://doi.org/10.1016/j.jjimei.2024.100293

Kusumahadi, T. A., & Permana, F. C. (2021). Impact of COVID-19 on global stock market volatility. Journal of Economic Integration, 36(1), 20-45.

Lara-Benítez, P., Carranza-García, M., & Riquelme, J. C. (2021). An experimental review on deep learning architectures for time series forecasting. *International journal of neural systems*, *31*(03), 2130001.

Lee, J., Kim, R., Koh, Y., & Kang, J. (2019). Global stock market prediction based on stock chart images using deep Q-network. IEEE Access, 7, 167260–167277.

Leigh, W., Hightower, R., & Modani, N. (2005). Forecasting the New York stock exchange composite index with past price and interest rate on condition of volume spike. Expert Systems with Applications, 28(1), 1–8.

Lewis-Beck, M. S., & Skalaban, A. (1990). The R-squared: Some straight talk. Political Analysis, 2, 153-171.

Li, J.-L., & Shi, W.-K. (2025). Hybrid preprocessing for neural network-based stock price prediction. Heliyon, 10(1), e040819. https://doi.org/10.1016/j.heliyon.2024.e40819

Li, P., Wei, Y., & Yin, L. (2025). Research on Stock Price Prediction Method Based on the GAN-LSTM-Attention Model. CMC-Computers, Materials & Continua, 82(1), 610-630. https://doi.org/10.32604/cmc.2024.056651

Liu, H., Manzoor, A., Wang, C., Zhang, L., & Manzoor, Z. (2020). The COVID-19 outbreak and affected countries stock markets response. International journal of environmental research and public health, 17(8), 2800.

Liu, R., & Gupta, R. (2022). Investors' uncertainty and forecasting stock market volatility. *Journal of Behavioral Finance*, *23*(3), 327-337.

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management, Forthcoming*.

Long, J., Chen, Z., He, W., Wu, T., & Ren, J. (2020). An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market. Applied Soft Computing, 91, 106205.

Lorie, J. H., & Hamilton, M. T. (1985). The stock market. RD Irwin.

Markowitz, H. (1952). Portfolio selection. The Journal of Finance.

Martin, I. W., & Wagner, C. (2019). What is the Expected Return on a Stock?. The Journal of Finance, 74(4), 1887-1929.

Mason, C. H., & Perreault Jr, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. Journal of marketing research, 28(3), 268-280.

McKillop, D., French, D., Quinn, B., Sobiech, A. L., & Wilson, J. O. (2020). Cooperative financial institutions: A review of the literature. International Review of Financial Analysis, 71, 101520.

Mehtab, S., & Sen, J. (2019). A robust predictive model for stock price prediction using deep learning and natural language processing. arXiv preprint arXiv:1912.07700.

Mehtab, S., & Sen, J. (2020). A time series analysis-based stock price prediction using machine learning and deep learning models. International Journal of Business Forecasting and Marketing Intelligence, 6(4), 272-335.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

MSCI & S&P Dow Jones Indices. (2023). *Global Industry Classification Standard (GICS)*. Retrieved from https://www.msci.com/our-solutions/indexes/gics

Nau, R. (2014). Forecasting with moving averages. Fuqua School of Business, Duke University, 1-3.

Nayak, A., Pai, M. M., & Pai, R. M. (2016). Prediction models for Indian stock market. Procedia Computer Science, 89, 441–449.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models.

Niederhoffer, V. (1971). The analysis of world events and stock prices. The Journal of Business, 44(2), 193-219.

Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, 26(4), 164–174. https://doi.org/10.1002/isaf.1459

Onyuma, S. O. (2009). Day-of-the-week and month-of-the-year effect on the Kenyan stock market returns. Eastern Africa Social Science Research Review, 25(2), 53-74.

Palomino, F., Paolillo, S., Perez-Orive, A., & Sanz-Maldonado, G. (2019). The information in interest coverage ratios of the US nonfinancial corporate sector.

Phuoc, T., Anh, P. T. K., Tam, P. H., & Nguyen, C. V. (2024). Applying machine learning algorithms to predict the stock price trend in the stock market–The case of Vietnam. Humanities and Social Sciences Communications, 11(1), 1–18.

Plerou, V., Gopikrishnan, P., Amaral, L. A. N., Meyer, M., & Stanley, H. E. (1999). Scaling of the distribution of price fluctuations of individual companies. *Physical review e*, *60*(6), 6519.

Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. PloS one, 18(2), e0279774.

Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H. C. (2021). Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions. Electronics, 10(21), 2717.

Schmid, T., & Krennmair, P. (2022). Flexible domain prediction using mixed effects random forests.

Shanaev, S., & Ghimire, B. (2019). Is all politics local? Regional political risk in Russia and the panel of stock returns. Journal of Behavioral and Experimental Finance, 21, 70-82.

Sheth, D., & Shah, M. (2023). Predicting stock market using machine learning: Best and accurate way to know future stock prices. International Journal of System Assurance Engineering and Management, 14(1), 1–18.

Shiller, R. J. (1990). Market volatility and investor behavior. The American Economic Review, 80(2), 58-62.

Smith, D. A., & Ocampo, S. (2025). The evolution of US retail concentration. American Economic Journal: Macroeconomics, 17(1), 71-101.

Song, W., Liang, J. Z., Cao, X. L., & Park, S. C. (2014). An effective query recommendation approach using semantic strategies for intelligent information retrieval. *Expert Systems with Applications*, 41(2), 366–372. https://doi.org/10.1016/j.eswa.2013.07.007

Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., & Bhat, S. K. (2023). Forecasting stock market prices using machine learning and deep learning models: A systematic review. International Journal of Financial Studies, 11(3), 94.

Steiner, M. (2009). Predicting premiums for the market, size, value, and momentum factors. Financial Markets and Portfolio Management, 23, 137-155.

Stockanalysis. (2025). Stock analysis and market capitalization data. Retrieved January 26, 2025, from https://www.stockanalysis.com

Suarez, R. (2016). Large-cap versus small-cap, a downside risk comparison.

Tabash, M. I., Chalissery, N., Nishad, T. M., & Al-Absy, M. S. M. (2024). Market shocks and stock volatility: Evidence from emerging and developed markets. International Journal of Financial Studies, 12(1), 2.

Teixeira, D. M., & Barbosa, R. S. (2024). Stock Price Prediction in the Financial Market Using Machine Learning Models. Computation, 13(1), 3.

Teweles, R. J., & Bradley, E. S. (1998). The stock market. John Wiley & Sons.

Thakur, R. S., & Kaur, R. (2023). A comparative study of deep learning models for Indian stock prediction. Trends in Machine Learning Research, 3(4), 424–435.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

Toomet, O., & Henningsen, A. (2008). Sample selection models in R: Package sampleSelection. Journal of statistical software, 27, 1-23.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2020). Using deep learning for price prediction by exploiting stationary limit order book features. Applied Soft Computing, 93, 106401.

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. Procedia Computer Science, 167, 599–606.

Wang, X., Jiang, B., & Liu, J. S. (2017). Generalized R-squared for detecting dependence. Biometrika, 104(1), 129-139.

Wang, Y. H., Yang, F. J., & Chen, L. J. (2013). An investor's perspective on infectious diseases and their influence on market behavior. Journal of Business Economics and Management, 14(sup1), S112-S127.

Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons.

Wielechowski, M., & Czech, K. (2022). Companies' Stock Market Performance in the Time of COVID-19: Alternative Energy vs. Main Stock Market Sectors. Energies, 15(1), 106. https://doi.org/10.3390/en15010106

Yahoo Finance. (2025). Historical stock data including daily trading volumes and adjusted closing prices. Retrieved January 26, 2025, from https://finance.yahoo.com/

Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

Yu, P., & Yan, X. (2020). Stock price prediction based on deep neural networks. Neural Computing and Applications, 32(6), 1609–1628.

Weerawarana, R., Zhu, Y., & He, Y. (2019). Learned Sectors: A fundamentals-driven sector reclassification project. arXiv preprint arXiv:1906.03935.

Zeleke, A. J., Palumbo, P., Tubertini, P., Miglio, R., & Chiari, L. (2024). Comparison of nine machine learning regression models in predicting hospital length of stay for patients admitted to a general medicine department. *Informatics in Medicine Unlocked*, *47*, 101499.

Zhang, D., Hu, M., & Ji, Q. (2020). Financial markets under the global pandemic of COVID-19. Finance research letters, 36, 101528.

Zhang, J., Lai, Y., & Lin, J. (2017). The day-of-the-week effects of stock markets in different countries. Finance Research Letters, 20, 47-62.

Zivot, E., Wang, J., Zivot, E., & Wang, J. (2003). Rolling analysis of time series. Modeling financial time series with S-Plus®, 299-346.

APPENDIX

The codes are provided for reproducibility and improvement of the analysis. In this master's dissertation data selection, data preparation, Exploratory Data Analysis (EDA), and data import parts are captured in R programming while feature engineering, modeling, and evaluation stages have been applied by Python programming.

Initial source of data selection - https://stockanalysis.com/stocks/

R code for data selection and EDA - https://rpubs.com/Turgud/Master_Tesis_Data_Selection

R code for data import - https://rpubs.com/Turgud/MADataImort

Python code for feature engineering and modeling - https://github.com/TurgudValiyev/Master-Thesis/blob/main/MA%20thesis%20Stock%20Prediction.ipynb

List of tables

List of pictures