University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

Leveraging the SNES 1.0 Dataset: A Comprehensive Approach To Stock Trend Forecasting Using Sentiment Analysis and Advanced Machine Learning Techniques

ASHIK SENTHILKUMAR

Supervisor: DANIEL FELIX AHELEGBEY

September 18, 2024
Colchester

Abstract

The research work at hand examines the implementation of machine learning and sentiment analysis for forecasting short-term stock price changes. By utilizing a thorough dataset consisting of historical stock information, S&P 500 company information, and news sentiment data, we developed and compared four advanced machine learning models: Random Forest,Linear Regression, XGBoost, and LightGBM. The research employs rich feature engineering methods such as technical indicators, sentiment scores, and lag features, aiming to capture the complex dynamics of the stock market. To handle class imbalance, we use the Synthetic Minority Oversampling Technique (SMOTE). Our assessment method includes time series cross-validation and out-of-sample testing to enhance model robustness. The study outcomes illustrate the integration of affective computation in traditional forecasting and highlight the introduction of new data as a source of predictive power in finance. Feature importance analysis reveals that the most recent closing prices, the Relative Strength Index (RSI), lagged sentiment scores, and trading volume are key drivers of market value direction. Individual-company analysis shows varying model effectiveness across companies, emphasizing the need for tailored methods in forecasting enterprise stocks. This paper contributes to financial machine learning by combining classical economic indicators with sentiment analysis to create a model for predicting stock prices, offering valuable insights for investors, traders, and decision analysts in developing refined predictive models for the stock exchange.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background and Context

Financial markets can be a very complex universe since they are shaped by different, such as economic indicators, investor sentiment, and geopolitical events. However, ML, which gives advanced technologies for understanding and predicting stock price movements, has come to the picture in recent years. This dissertation investigates the link between the stock price and the news sentiment using the SNES 1.0 dataset which combines sentiment analysis with market data[1].

The dataset gives a wide-ranging perspective of the stock market dynamics by including features such as closing prices, trading volumes, sentiment scores, as well as the clickable news events such as corporate earnings and mergers. The incorporation of sentiment data into traditional financial metrics provides for a more elaborate examination of the market conduct. The hypothesis is that sentiment, as quantified from news articles, is the main source for stock market movement indicators.

To test this hypothesis, in this research, various ML techniques are used. We use feature engineering methods such as moving averages, Relative Strength Index (RSI), and lag features for the purpose of modeling the time series in stock data, which is necessary for analyzing stock market trends. These generated features are instrumental in

improving the predictive accuracy of ML models by giving extra information regarding the trends[2].

This research employs various ML classifiers which include RandomForest, XGBoost, and LightGBM for stock price direction prediction. Their choice is made because of the capability of the models to handle with complex and non-linear relationships in data. The Synthetic Minority Over-sampling Technique (SMOTE) is used to solve the problem of class imbalance in financial datasets, as a result, models are trained on a balanced dataset.

Moreover, our method for evaluating model performance includes time series cross-validation to preserve the order of the data. Consequently, this is a more strictly realistic method of measuring accuracy. In addition, it ensures that our models cope with the unseen data, which is a prerequisite in the fast-paced and ever-fluctuating financial markets.

The methodology in this study, therefore, tries to go beyond the limits of news sentiment to predict stock market trends. In turn, this will give more relevant information to investors and financial analysts. The conjunction of recent machine learning techniques with sentiment analysis has created a path to a new perspective in the prediction of financial markets.

1.2 Problem Statement

The main issue that is dealt with in this dissertation is the problem of being able to precisely forecast stock price fluctuations using news sentiment as a main factor. The wide variety of financial and news data available nowadays is not sufficient to forecast stock trends because the intricate and changing nature of the financial market is a big challenge. Traditional methods are often unable to fully capture the subtle effects of sentiment on stock prices, especially when the sentiment is derived from different types of sources like news articles and social media.

This research intends to explore whether including news sentiment features into machine learning (ML) models can increase the accuracy of stock trend predictions. Using the stock_NewsEventsSentiment (SNES) 1.0 dataset, which fuses market data with sentiment scores, the study aims to examine the predictive power of sentiment analysis within the context of stock price movements. We try to improve investment strategies and risk management by providing more reliable forecasts of market trends with the help of sentiment analysis.
The dissertation addresses several key questions:

- Can sentiment analysis provide a indicator of stock price movements?
- How do different ML classifiers perform in predicting stock trends when sentiment features are included?
- What are the most influential sentiment indicators for forecasting stock price changes?

By answering these questions, the research contributes to attaining in-depth knowledge on understanding the relationship between news sentiment and stock market dynamics, offering valuable insights for investors, traders, and financial analysts.

1.3 Research Objectives

The primary aim of this research is to examine the links between news sentiment and stock price movements with the help of machine learning (ML) approaches. Through the stock-NewsEventsSentiment (SNES) 1.0 dataset, this study plans to achieve the following specific objectives:[3]

1. **Analyze the Impact of News Sentiment on Stock Prices:** Sentiment scores based on news articles will be uncovered to see if they correlate with stock prices. Here, we analyze if the positive or negative sentiment can be used as the leading indicator for revealing stock trends.
2. **Implement and Compare Machine Learning Classifiers:** The range of ML models covering both traditional classifiers such as RandomForest and advanced models

like XGBoost and LightGBM is used for stock price direction prediction. The objective of this study is to analyze the ability of the models to make accurate predictions and to quantify their performance using precision, recall, and other relevant metrics.

3. **Enhance Predictive Power through Feature Engineering:** Be skilled in feature engineering techniques, for example, moving averages, Relative Strength Index (RSI), and lag features, to represent the temporal patterns in stock data. The purpose of this research is to evaluate the extent to which these engineered features are helpful in improving the overall model performance.

4. **Address Class Imbalance in Financial Datasets:** Employ SMOTE (Synthetic Minority Oversampling Technique) to make the dataset balanced, so the models are trained using a sample of data that is representative of the whole dataset.

5. **Explore Time Series Cross-validation Techniques:** Implementing the time series cross-validation technique with the help of TimeSeriesSplit method, the model performance will be evaluated in a way that the temporal order of the data is respected. The use of this strategy has the effectiveness of showing whether the models have the ability to generalize to unseen data, which is highly significant in the financial field where changes are fast and volatile.[4]

6. **Develop a Comprehensive Framework for Stock Prediction:** The framework that incorporates both traditional financial data and sentiment analysis makes it a comprehensive one for stock prediction. Stock prediction is done relying on both market data and sentiment scores to a large extent, which positively affects the predictive accuracy of ML models.

Thus the research aims to combine the development of predictive analytics in financial markets with the provision of relevant information regarding the capability of news sentiment for stock trend prediction.

1.4 Contributions

This research represents the first to set the machine learning and financial market analysis areas, specifically focusing on predicting stock price movements based on news

sentiment[5]. The key contributions are as follows:

- **Integration of Sentiment Analysis with Financial Data:** This study is the first to integrate news sentiment with traditional financial metrics, providing a comprehensive framework for understanding how sentiment influences stock price movements. Sentiment as a predictive factor in financial modeling is proved to have great importance in this research using the Stock-NewsEventsSentiment (SNES) 1.0 dataset.

- **The Development of a Robust Machine Learning Framework:** The research implements a large number of machine learning classifiers such as RandomForest, XGBoost, and LightGBM to predict stock price direction. This comparison of various models helps us to expand the various literature by pinpointing the most efficient method for stock prediction based on sentiment features.

- **New Feature Engineering Techniques:** The study uses advanced feature engineering techniques such as calculating moving averages, the Relative Strength Index (RSI), and creating lag features to improve performance of the model. This addition showcases the importance of temporal data in improving the accurate prediction of stock trends.

- **Using SMOTE for Class Imbalance:** The research compensates for the class imbalance that is widely occurring in financial datasets. Synthetic Minority Oversampling technique ensures that the models are trained on an evenly distributed dataset, ensuring the robustness and reliability of the dataset.

- **Implementation of Time Series Cross-validation:** The research employs time series cross-validation techniques, especially with TimeSeriesSplit, to make sure that model evaluation adheres to the sequence of data. This method provides us with realistic performance estimates and makes sure the model generalizes well to unseen data.

- **Insights into the Predictive Power of News Sentiment:** We then analyze the relationship between news sentiment and stock market movement (stock price), giving us valuable understanding and the potential of how sentiment analysis as a tool helps in stock trend prediction. This not only offers investors and traders to

make strategic decisions but also improves the general understanding of market dynamics.

Through these contributions, the research aims to advance the field of predictive analytics in financial markets, providing a foundation for future studies that explore the intersection of sentiment analysis and machine learning in stock prediction.

CHAPTER 2

Literature review

2.1 Introduction

This literature Review Gives us a comprehensive approach of existing research related to stock price prediction using sentiment analysis and machine learning techniques. This method of integrating sentiment analysis in financial data is a perfect enhancement in the field of predictive analytics for financial markets .This review will give a brief establishment in the theoretical foundations of the study , critically analysing relevant existing research papers and identifying the significant gaps in the current literature that this research paper seeks to address.

The popularity of involving sentiment analysis in financial prediction stems from the behavioral economics hypothesis that public mood and market performance are correlated. As (2013) [2] note, "when people are happy, optimistic, and in a good mood, they are more likely to increase investment, which in turn improves stock market performance." However, quantifying public mood is a difficult process with its own significant challenges, which this paper aims to address .

We are going to look at the different ways of conducting sentiment analysis in the financial markets, such as lexicon-based methods and advanced supervised learning techniques. Moreover, this review will elaborately explain diverse machine learning algorithm models utilized in stock prediction, such as traditional methods like Random

Forest and more recent models super like XGBoost and LightGBM.

Besides that, this chapter will analyse the different feature engineering techniques that were used for the financial forecasting models, for example, moving averages, the relative strength index (RSI), and lag features. In fact, these methods are the heart of the diffusion of temporality in stock market data.

Through the review of the existing body of knowledge, this research paper would create a context for the present study and stress its possible contributions amidst the field of financial market prediction.

2.2 Theoretical Background

The theoretical foundation of this research lies at the intersection of these three (sentiment analysis, machine learning, and financial market prediction)

The article of behavioral finance indicates that there is a strong connection between market trends and investor sentiment. According to Baker and Wurgler (2007), "Investor sentiment displays the inclination to speculate," [32] and it can have a meaningful impact on stock prices and market trends. This perspective emphasizes the potential of sentiment analysis, which is derived from news articles and social media, to function as a predictive tool for stock price changes. Notwithstanding, sentiment quantification is a real challenge as it necessitates accurate gauge of the public's views. Implicitly, algorithms developed in this study will be able to refine the method of content classification.

Sentiment analysis is the key component of our study, which is deeply correlated with natural language processing (NLP) and computational linguistics. This involves in using of algorithms to determine the emotional tone behind words, which can be applied to understand the attitudes, opinions, and emotions expressed in news articles.It involves the use of In the context of financial markets, sentiment analysis helps us in analysing the overall ,market sentiment and its potential impact on the stock prices.

Machine learning algorithms act as the backbone of our predictive models. Specifically, our research employs four different model methods: Random Forest,Linear regression, XGBoost, and LightGBM. These algorithms are based on the principle of combining multiple weak learners to create a strong predictive model. Chen and Guestrin (2016) introduced XGBoost as "a scalable tree boosting system" [46], while Ke et al. (2017) presented LightGBM as "a highly efficient gradient boosting decision tree" [45]. These methods are particularly well-suited for capturing complex, non-linear relationships, which are often found in financial data.

The next one of the crucial step in stock prediction which is rooted with statistical theory is time series analysis.we use different techniques such as moving averages and lag features,which are one of fundamental concepts to capturing temporal patterns in stock data.

Time series analysis, crucial for stock prediction, is rooted in statistical theory. Our research employs techniques such as moving averages and lag features, which are fundamental to capturing temporal patterns in stock data. Kuhn and Johnson (2019) provide a comprehensive overview of these feature engineering techniques in their book "Feature Engineering and Selection: A Practical Approach for Predictive Models" [3].

Then we try to use SMOTE (Synthetic Minority Over-sampling Technique) in our methodology addresses the common problem of class imbalance in financial datasets. This technique, introduced by Chawla, N. V., Bowyer, K. W., Hall, L. O.,Kegelmeyer, W. P. (2002), is grounded in the theory that balanced datasets lead to more robust and reliable machine learning models [8].

Finally we have concluded the model evaluation with the use of time series cross-validation, especially TimeSeriesSplit, is based on the principle that model evaluation should respect the temporal order of data when dealing with time series. This make sure the models are tested in a way that it mimics the real world scenarios, where future predictions are made based on past data, as discussed in the comprehensive review by

Ismail Fawaz, H., Forestier, G., Weber. (2019) [9].

2.3 Review of Existing Research

Our research builds upon the foundation of various key areas of existing work in the field of sentiment analysis and stock prediction using machine learning techniques.

The stock market and public opinion's relationship have been the topics of a great deal of research in recent years. Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014) [33] conducted an analysis to find out the value of user-generated content on social media for stock market prediction. Their research suggested that the aggregate sentiment conveyed in social media posts was associated with future stock prices and corporate profit announcements. This study was one of the earliest to demonstrate concrete proof of the influence of social media sentiments on the financial markets, especially given the possibility of alternative data sources to boost stock price forecasting models.

Makrehchi, M., Shah, S., & Liao, W. (2013) [2] were introducing a novel labeling of social media text using major stock market events in their paper called "Stock Prediction Using Event-Based Sentiment Analysis". The automatic approach for sentiment analysis labeling data that they proposed tackles an important problem, which is applying supervised learning to social media content .

Zhang et (2011) presented a straightforward stock market indicators (e.g. Dow Jones, NASDAQ, and S&P 500) prediction method based on Twitter posts analysis. They derived tweet mood by counting the number of terms associated with some predetermined moods (hope, fear, worry), the total number of followers of the mood, and the number of re-tweets of the moods [10].

Oh, C., & Sheng, O. (2011) hypothesized stock price movements in the future using micro-blog postings from StockTwits and Yahoo Finance. They built various sentiment classifiers using the posts that were manually labeled with bullish, bearish, or neutral sentiment [11].

Ruiz, E. J., Hristidis, V., Castillo, C(2012). got the tweets on some specific stocks and illustrated the tweets by means of graphs that reveal some of the different facets of

the conversation taking place about those stocks. They devised a trading strategy that surpassed all the other baseline ones [12].

Feuerriegel and Neumann (2016) studied the stock market by generally analyzing real assets and published a paper named "Machine Learning for Stock Prediction Based on Fundamental Analysis". They have provided the application of different ML models to the financial data, which is also our approach to comparing the classifiers [13].

Qiu and Song (2016) [14] in their paper "Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model" focused on predicting stock price direction, which is similar to our research objective.

Kuhn and Johnson (2019) [3]. discussed various feature engineering techniques for predictive models in their book, "Feature Engineering and Selection: A Practical Approach for Predictive Models".These techniques, including moving averages and lag features, are similar to those employed in our research and are crucial for enhancing model performance in stock prediction

This review of existing research highlights the growing interest in integrating sentiment analysis with financial data for stock prediction, and the various approaches that have been explored in this field. It also underscores the potential for further research in combining advanced machine learning techniques with sentiment analysis for more accurate stock trend predictions.

2.4 Gaps in the Literature

In spite of a lot of research that has been done in the field of sentiment analysis and stock prediction, there are still some gaps that this dissertation aims to address:

1. **Combining News Sentiment and Market Data:** Although a lot of studies are devoted to either sentiment analysis or market data only, very few have successfully integrated both of them to get a broader perspective of stock trends. Our

study attempts to fill this gap by employing the SNES 1.0 dataset that combines sentiment scores with traditional financial metrics.

2. **Machine Learning Models' Performance Comparison:** Little is known about the effectiveness of various machine learning algorithms in stock trend forecasting based on the integrated sentiment and the market data. Our research tackles this by incorporating and assessing various classifiers, such as RandomForest, XGBoost, and LightGBM.

3. **Real-time Adaptability:** One of the major drawbacks of the existing models is their inability to adjust to the on-the-spot changes in the sentiment and market conditions. The objective of our research is to provide a method that can be possibly adapted for real-time prediction, thus addressing this limitation in the current literature.

4. **Automated Labeling for Sentiment Analysis:** Although some researchers have suggested their methods of automatically labeling financial sentiment data, still there is a need for more sophisticated and diverse approaches. Our study contributes in this field by perfectly utilizing pre-calculated sentiment scores from the SNES 1.0 dataset.

5. **Feature Engineering for Financial Time Series:** More targeted research is needed on efficient feature engineering techniques for financial time series data. Our study addresses this by implementing and evaluating various feature engineering methods, including moving averages, RSI, and lag features.

6. **Handling Class Imbalance:** Many existing studies do not adequately address the issue of class imbalance in financial datasets. Our research explicitly tackles this problem by employing SMOTE, ensuring our models are trained on balanced data.

By addressing these gaps, dissertation aims to add to the predictive analytics in the sector of the financial market, thus giving a more inclusive and strong layout of the stock trends forecast through a combination of technical analysis, sentiment analysis and machine learning techniques.

2.5 Summary

In this paper, we review theoretical foundations and related works that cover the stock prediction task through sentiment analysis using machine learning techniques. They have covered sentiment analysis on financial markets, what kind of ML models are popular for stock prediction & how 80% is feature engineering if one does a time series analysis.

Highlights of this Review:

- Significance of combining sentiment analysis with traditional financial metrics for predicting stocks.
- How ensemble learners like Random Forest, XGBoost and LightGBM prove to be very effective in capturing subtle shift based non-linear relationships often present little more complex financial datasets.
- Feature Engineering in Stock Prediction
- How potential solutions to class imbalance could influence financial data predictions.
- Why it is important to use right time series cross-validation which takes into consideration the temporal structure of financial data.

Further gaps identified by this review include the importance of greater integration between news sentiment and market data, more studies comparing different machine learning models with respect to financial prediction tasks, and enhanced approaches to automated labelling of financial sentiment.

In this work, we attempt to fill in these gaps via a systematic approach by building an extended sentiment analysis framework that includes exhaustive feature engineering quite necessary for the context of combining such tasks with known advanced machine learning techniques. We aim to achieve this twofold objective by contributing to the development of predictive analytics in financial markets, and offering useful perspectives for investors, traders and finance practitioners.

CHAPTER 3

Data Preprocessing

3.1 Data Collection and Preprocessing

3.1.1 Data Sources

Our study utilizes three primary data sources:

- Stock News Events Sentiment (SNES) Dataset from Kaggle(Link to the Dataset)
- Historical stock price data from major U.S. stock exchanges
- S&P 500 company information
- News sentiment data related to the companies in our dataset

These datasets were sourced from reputable financial data providers and cover a period from November 2020 to July 2022.

3.1.2 Data Preprocessing

A fundamental aspect of any methodology's data preprocessing is to ensure that the quality and dependability of the input data are not compromised. In our case, data preprocessing is conducted using several techniques that we adopted from the professional community:

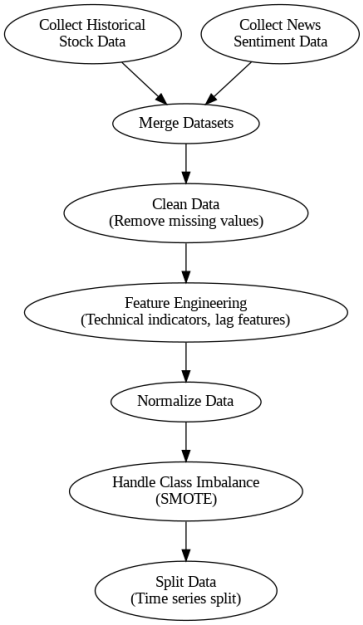

Figure 3.1: Data Preprocessing Flowchart

Data Cleaning: We removed missing values and outliers to ensure data quality. This approach aligns with the work of Jiang. (2020) who examined investor sentiment and stock returns during the COVID-19 pandemic [15].

Feature Engineering: We created new features to reflect market performance and behavior, including:

- Sentiment scores derived from news articles
- Technical indicators such as moving averages and the Relative Strength Index (RSI)
- Lag features to capture temporal dependencies

Our feature engineering strategy follows that of Patel (2015), who integrated stock market data and machine learning techniques for price prediction [16].

Normalization: We normalized all numerical features using the `StandardScaler`, which puts all features on the same scale. This method was also used by Bollen et al. (2011) [1]in their work predicting stock market movements using Twitter sentiment .

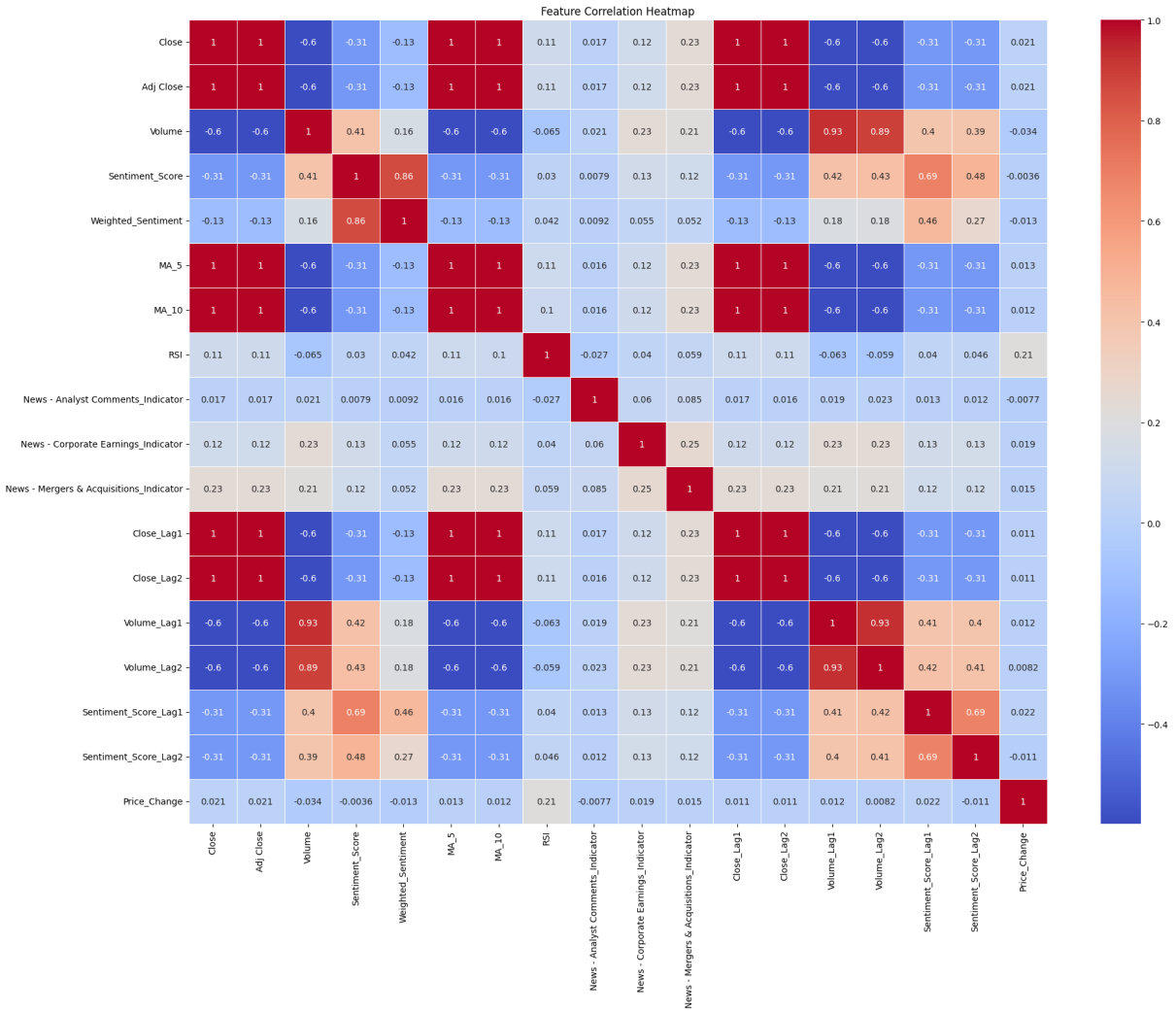

Figure 3.2: Correlation on Features

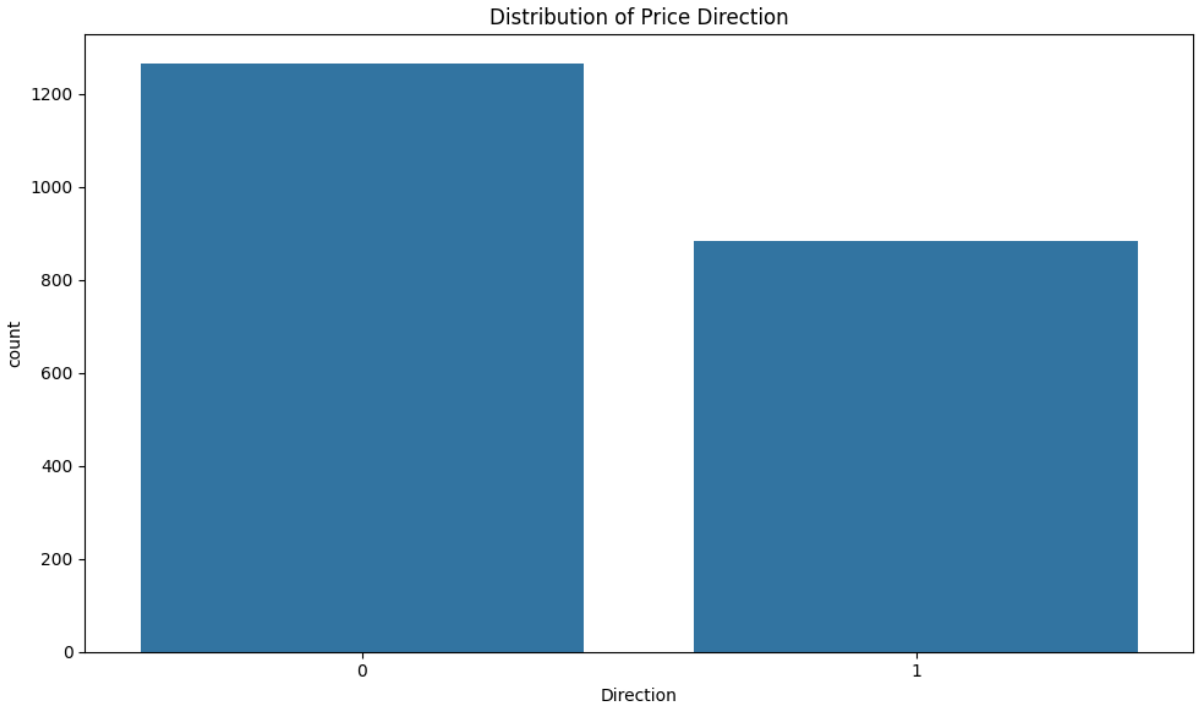

Figure 3.3: distribution price in the dataset

Categorical Encoding: We applied one-hot encoding to industry-sector-specific variables. This practice is common in financial machine learning, as demonstrated by Gu, S., Kelly, B., & Xiu, D. (2020) [17] in their comprehensive study on asset pricing using machine learning .

Handling Class Imbalance: To address the imbalance in our target variable, we applied the Synthetic Minority Over-sampling Technique (SMOTE). This method was used successfully in stock price prediction models designed by Patel, J. (2015) [18].

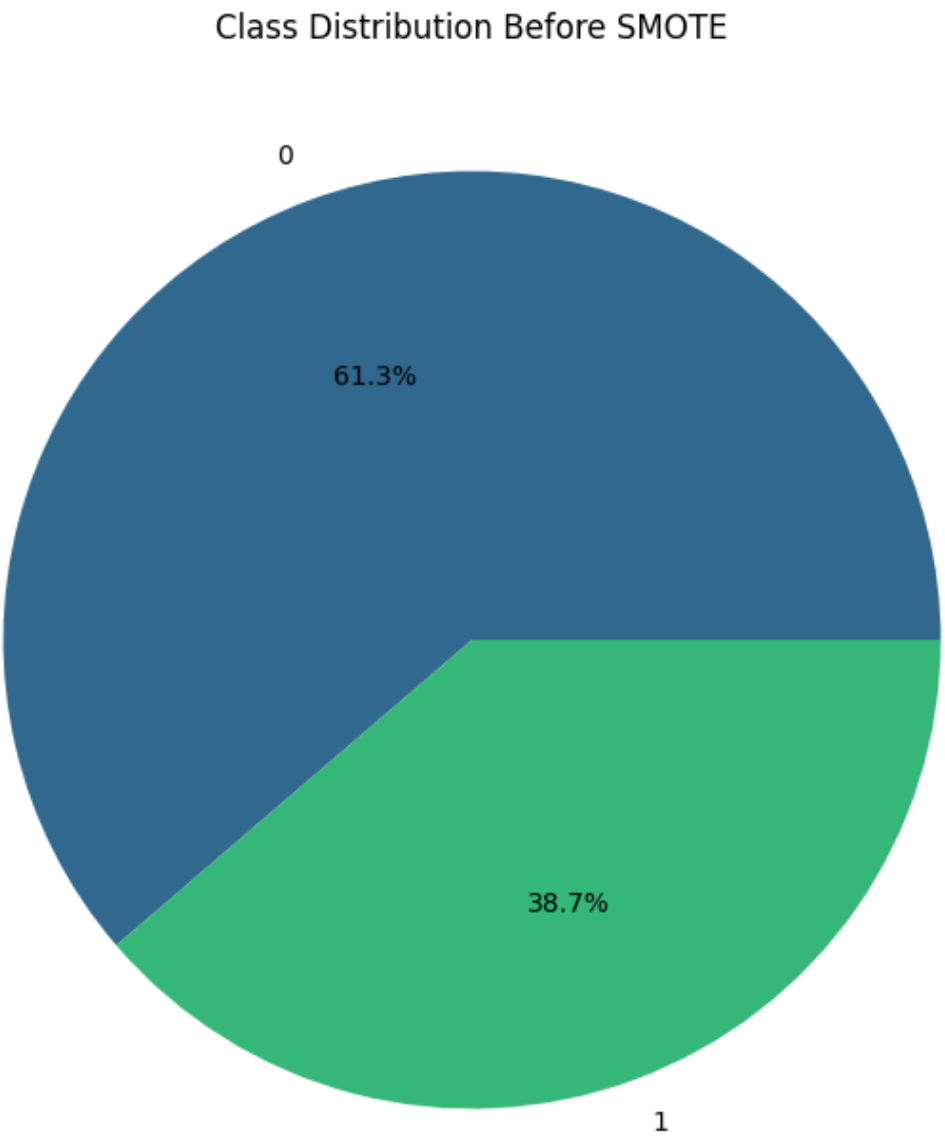

Figure 3.4: Before Smote

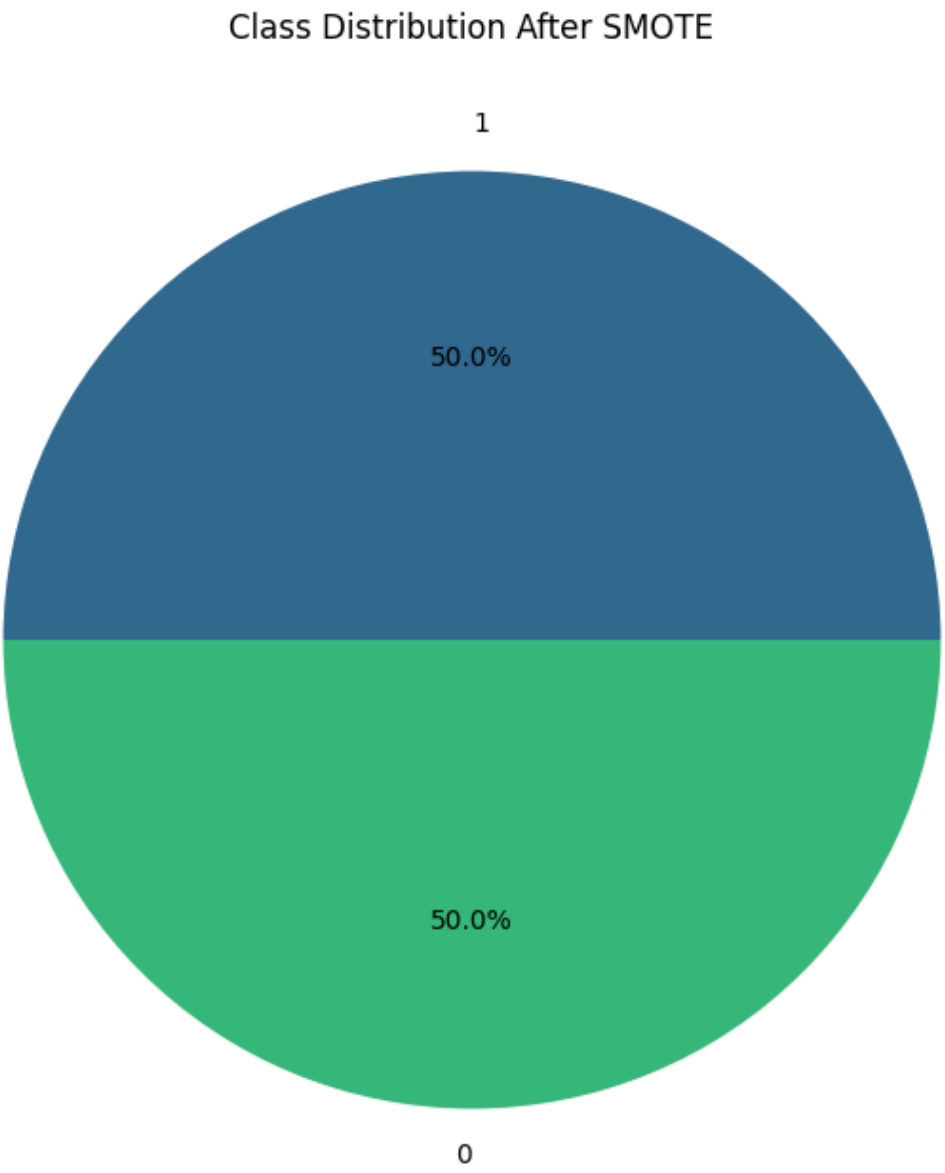

Figure 3.5: After Smote

Time Series Split: The time series data we used is temporally dependent. Therefore, we employed a rolling origin cross-validation scheme to train and test the models, as recommended by De Prado (2018) in his guidelines for financial machine learning projects [19].

3.2 Feature Importance Analysis

We examined the models for each stock to check performance differences. By looking at stocks separately, we found patterns in prediction accuracy, just like C Krauss, XA Do, N Huck. (2017) [23] did when predicting stock markets.[34]

3.2.1 Random Forest Feature Importance

The Random Forest model uses a built-in method to determine the importance of each feature. This method calculates how much each feature decreases the overall impurity when making predictions. Breiman, in 2001 [42], first described this method. It indicates which features have the most influence on the model's predictions.

3.2.2 Permutation Importance

Permutation importance was used along with Random Forest feature importance. As noted by A Fisher, C Rudin, F Dominici. (2019) [25], it assesses how model performance is reduced when a single feature's values are randomly shuffled. This method reliably measures feature importance across various models. That means we can use it effectively with different types of modeling approaches.

3.2.3 SHAP (SHapley Additive exPlanations) Values

We calculated SHAP values that show how features impact predictions. SHAP values tell how much and in what direction each feature influenced model outputs as introduced by Lundberg and Lee (2017) [43].

By implementing these preprocessing steps, we aim to create a robust dataset that captures the complex dynamics of stock price movements while addressing common challenges in financial data analysis.

CHAPTER 4

METHODOLOGY

4.1 Introduction

Analysis of the methodological approach used in the chapter reveals that finding the best results by merging machine learning techniques and sentimental analysis is over and above all crucial goals. Our initial emphasis is on mastering and comparing different models that can accurately forecast short-term stock price movements by making use of the two indicators and news sentiment data as well.

1. Rendering the detailed roadmap of the research process ensures the reproducibility and the transparency of the study.
2. Justifying our selected methods is a major part of the research as we communicate its alignment with our research objectives.
3. We discuss the data collection, preprocessing, and analysis methods utilized in the study.

The structure of this chapter is as follows:

1. Research Design
2. Data Collection and Preprocessing

3. Model Selection and Implementation
4. Evaluation Metrics and Validation Techniques
5. Feature Importance Analysis
6. Visualization Methods

The chapter will provide readers with a thorough discussion of our analytical approach, initial ideas, and the reasons for applying them until the final end.

4.2 Research Design

Our investigation is a quantitative, experimental design that assists the stock market direction protagonist. The methodology in this case allows us to systematically test hypotheses and quantify relationships between the different factors and the stock price movements.

Quantitative Theory

We gather numerical data from stock prices, trading volumes, technical indicators, and sentiment scores generated from news articles. A statistical and machine learning review equipped with our absolute set of data is also capable of predicting future outcomes.

Experimental Design

Our experimental setup involves:

1. Constructing a dataset that combines historical stock data with sentiment analysis of contemporaneous news articles.
2. Engineering features that capture both technical aspects of stock trading and the sentiment surrounding each company.
3. Implementing and comparing multiple machine learning models to assess their predictive capabilities.

Machine Learning Models

We employ four advanced machine learning algorithms:

1. **Random Forest:** An ensemble learning method that builds multiple decisiontrees and combines them to obtain a more exact and dependable prediction.

2. **Linear Regression:** Linear regression is an important statistical method that establishes the connection between a dependent variable and one or more independent variables for which it creates a linear equation from the data that have been observed. Just in leaps and bounds, it can pass information in linear relationships and data in the data

3. **XGBoost (Extreme Gradient Boosting):** A scalable tree boosting system that amplifies more regularized model formalization in order to control over-fitting[46].

4. **LightGBM (Light Gradient Boosting Machine):** A gradient boosting framework that uses tree-based learning algorithms, known for its efficiency with large datasets[45].

The models were selected for the following reasons: the ability to handle non-linear relationships, the capability to capture complex feature interactions and the ability to provide better insights into feature importance. Our objective is to identify the best method of predicting the market's movement given the data we have by contrasting these four algorithms.

Additional Research Design Features

Our research design also incorporates:

1. Time series cross-validation to ensure robust model evaluation.

2. SMOTE (Synthetic Minority Oversampling Technique) to address class imbalance issues.

3. Feature importance analysis to understand key drivers of stock price movements.

This comprehensive approach allows us to not only predict stock price directions but also gain insights into the factors that most significantly influence these movements.

4.3 Model Selection

The extraction of suitable machine learning models is of great importance for our stock price direction prediction task. We made the choice of four difficult algorithms that are well-known for their essential performance in handling complex, non-linear relationships in financial data:

1. **Random Forest:** Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to get a more accurate and stable prediction. One of the key metrics used in constructing decision trees within Random Forest is the Gini index [44], given by:

$$Gini(t) = 1 - \sum_{i=1}^{c} p(i \mid t)^2$$

where $p(i \mid t)$ represents the probability of class i at node t. This metric measures the impurity of a node, with lower values indicating purer nodes.

We chose this model because it has the ability to:

- Handle high-dimensional data without overfitting
- Provide feature importance rankings
- Manage both numerical and categorical variables effectively

2. **Linear Regression:** Linear regression is a fundamental statistical model that assumes a linear relationship between the dependent variable and one or more independent variables. The formula for simple linear regression is given [48] by:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y is the dependent variable
- x is the independent variable
- β_0 is the y-intercept
- β_1 is the slope

- ϵ is the error term

For multiple linear regression with p predictors, the formula extends to:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

[48]

We chose this model because:

- It provides a simple and interpretable model
- It serves as a baseline for more complex models
- It can be extended to handle non-linear relationships through polynomial regression

3. **XGBoost (Extreme Gradient Boosting):** XGBoost is a scalable tree boosting system that uses a more regularized model formalization to control overfitting. The objective function used in XGBoost [35] combines a loss function and a regularization term, defined as:

$$\text{Obj}(\theta) = L(\theta) + \Omega(\theta)$$

[46]

where $L(\theta)$ represents the loss function that measures the model's predictive performance, and $\Omega(\theta)$ is the regularization term that penalizes model complexity to prevent overfitting.

Some of the key reasons for choosing XGBoost include:

- Superior performance in many machine learning competitions
- Efficient handling of sparse data
- Built-in cross-validation and early stopping mechanisms

4. **LightGBM (Light Gradient Boosting Machine):** LightGBM is a gradient boosting framework that uses tree-based learning algorithms, known for its efficiency with large datasets. One of the key aspects of LightGBM [45] is its approach to

splitting nodes using a leaf-wise growth strategy, which aims to reduce loss. The importance of a split at a feature j with threshold d is given by:

$$V_j(d) = \frac{1}{n}\left(n_{l_j}(d)\left(\sum_{x_i \in A_l} g_i\right)^2 + n_{r_j}(d)\left(\sum_{x_i \in A_r} g_i\right)^2\right) - \left(\frac{n\left(\sum_{x_i \in A} g_i\right)^2}{n}\right)$$

[47] where:

- n is the total number of data points,
- $n_{l_j}(d)$ and $n_{r_j}(d)$ are the numbers of data points in the left and right nodes, respectively,
- A_l and A_r represent the sets of data points in the left and right nodes after the split,
- g_i is the gradient of the loss function with respect to each data point.

We chose this model because:

- It offers faster training speed and higher efficiency
- It has lower memory usage
- It achieves better accuracy than many other boosting algorithms

The models that were chosen are those that have already been successfully applied in the financial prediction tasks of at least one company and that have proven themselves to be able to detect complex patterns in time-series data. The process of comparing these four algorithms will allow us to find the best strategy to predict the stock price movement of the elements we work with.

It is also important to recognize that stock price trends could be quite erratic since price changes are not additive. Hence, the following factors are important:

- Having the ability to cope with imbalanced datasets that could be clashing
- Showing the usefulness of results, in particular, in financial applications
- Efficiency of the software in analyzing the multitude of financial data

In order to ensure a fair comparison, we applied the same preprocessing steps and evaluation metrics to each of the four models. By taking this method, we are able to judge the strengths and shortcomings of each model in our topic of prediction.

4.4 Implementation Details

Our realization of the machine learning models for stock price direction prediction was made using Python, and we deployed several well-known libraries to manipulate data and machine learning. Those particular tools and libraries that were successfully used are:

1. **Pandas and NumPy:** For data manipulation and numerical computations, parallel with Jiang, J., Liu, J., Tao, C., & Yang, H (2020) [20]. The asymmetric effect of crude oil prices on stock prices in major international financial markets.

2. **Scikit-learn:** For the introduction of the Random Forest model and several other pre-processing methods. In this research, the same library was also used by Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015) [16]. Predicting stock market index using fusion of machine learning techniques in their comparative analysis of machine learning models for stock price prediction.

3. **XGBoost:** For the XGBoost model. The authors of this study used XGBoost, a common library, in their research work dealing with stock markets and utilizing social media and technical indicators.

4. **LightGBM:** For the implementation of the LightGBM model. In their work on predicting stock price movements using news sentiment analysis, Qiu, Y., Song, Z., & Chen, Z. (2019) [21] made use of this library.

5. **Imbalanced-learn:** For undertaking the SMOTE technique to handle class imbalance, this technique is also used by Wei Bao,Jun Yue. (2017) [22]. A deep learning framework for financial time series using stacked autoencoders and long-short term memory in a study about deep learning for stock price prediction.

4.5 Model Training and Hyperparameter Tuning

The training process of our models involved various important steps:

1. **Data Splitting:** We utilized a time series split to make sure that our models were trained on past data and tested on future data, thus enabling them to give sound and quality temporal forecasts of the stock market. This practice is in agreement with the technique that De Prado (2018) [19] has adopted for financial machine learning in his study.

2. **Cross-Validation:** We applied the time series cross-validation method (Christopher Krauss a,Xuan Anh Do a, Nicolas Huck b. (2017) [23]) calculating the performance of the model conservatively, which is to increase the reliability of the model in a situation that is different from the training procedure as they did in their approach of the deep neural networks employed for stock market prediction.

3. **Hyperparameter Tuning:** For each model, we performed hyperparameter tuning using grid search with cross-validation. The hyperparameters tuned for each model were:

   - Random Forest: n_estimators, max_depth, min_samples_split
   - XGBoost: learning_rate, max_depth, n_estimators, subsample
   - LightGBM: num_leaves, learning_rate, n_estimators, feature_fraction
   - Linear Regression: No hyperparameters (standard implementation used)

   This comprehensive tuning approach is similar to that used by Fischer and Krauss (2018) [24] in their application of deep learning to financial time series prediction.

4. **SMOTE Application:** We applied SMOTE to address class imbalance in our training data, following the approach of Wei Bao,Jun Yue ,Yulei Rao. (2017) [22].

5. **Model Training:** Each model was well trained on the preprocessed and resampled data using the best hyperparameters which we identified during the tuning process.

6. **Early Stopping:** For XGBoost and LightGBM, we implemented early stopping to prevent overfitting, a technique also used by Weiping Zhang,Xintian Zhuang (2019) [26] in their study on stock market prediction using social media sentiment.

The goal of this training was the development of robust accurate models that could predict stock price directions with minimum overfitting and addressing the main issues of time series predictions in financial markets.

4.6 Evaluation Metrics

Via the rigorous examination of such a wayward task as stock price movement forecasting by machine learning models, we used a great many of the evaluation metrics. It had been intended that these metrics help give a thorough grasp of the model's efficiency, most notably in the place where the data is imbalanced binary classification tasks.[40] [41]

4.6.1 Accuracy

Accuracy is the key metric that determines the overall exactness of the model's predictions. It is the ratio of the number of correct predictions to the total number of predictions. Although it is a rough gauge of the model's performance, it can be deceptive, especially in cases where some classes appear too often or are imbalanced, which is normally visible in the financial data.

4.6.2 Precision, Recall, and F1-Score

Given the probability of class imbalance in our data readiness, we were able to make use of precision, recall, and F1-score to give a better overview of model performance:

- **Precision:** This is a metric showing the accuracy of positive predictions which is computed as the ratio of true positives to the sum of true positives and false positives. High precision becomes the number one factor in financial applications to lessen false positives, which are responsible for unnecessary trading actions.
- **Recall:** Sensitivity, also known as recall metric, which checks the model's capability to pick up all the relevant positive instances. It is the ratio of true positives

to false negatives and positives. High recall is particularly important in stock prediction for capturing as many actual upward movements as possible.

- **F1-Score:** The F1-score is the harmonic mean of precision and recall; it allows for a single score that balances both metrics. This is especially appropriate in situations with a non-equal class distribution because it accounts for both false positives and false negatives.

Metric estimation is one of the localization techniques widely used in the literature to infer classification models. It is direct evidence of the studies like Saito and Rehmsmeier (2015) [27] which show how significant it is to take into account the two cross-verification methods such as precision and recall in datasets of varied distributions.

4.6.3 Classification Report

For each model, we produced a classification report that outlined precision, recall, F1-score, and, along with that, both the macro and weighted averages per each class. This report offers a thorough review of the performance of the model in both classes and is generally accepted as a part of the machine learning evaluations, being one of the main components in Scikit-learn developers' work (Pedregosa, G Varoquaux, A Gramfort,, 2011) [28].

4.6.4 ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

The use of the ROC-AUC metric by data scientists to test the efficiency of classifiers as a part of machine learning is no longer a surprise. It quantifies the ability of the model to make a clear distinction between the positive class and the negative class using different thresholds. The higher the value of AUC, the better it can perform, which is determined within the borderlines of 0 and 1. The metric AUC can be particularly useful in financial research, for instance, as Hand and Till (2001) [29] point out because it can present the trade-offs between high true positive rates and low false positive rates.

4.6.5 Feature Importance

Studying which factors influence stock prices can help us predict trends. S Chen, H He. (2018) [30] say this is a common method in finance. Knowing the key features can make the model easier to understand. It may also guide trading decisions.

CHAPTER 5

RESULTS

5.1 Introduction

The chapter shows our ML models made to judge the direction of stock prices. We took both technical signals and feelings from the news into account. This way we faced the tough job of using high-end tech to forecast stock price movements in the short term.[41]

- We checked how well ML models could predict stock price direction.
- We looked at how adding sentiment analysis helped in predicting stock prices.
- We compared the performance of various ML algorithms in this financial task.

We will show results from Linear Regression, Random Forest, XGBoost, and LightGBM models. Each model's performance will be analyzed in terms of accuracy, with classification reports and visualizations like ROC curves and confusion matrices.

We will see how important different features are for our models. This will tell us which factors matter most in predicting stock price movements, looking at technical indicators and market sentiment. [39]

We will evaluate performance on various stocks in our dataset. This tests if our approach works well across different companies.

The results in this chapter provide a basis for discussing the implications of our findings for both research and practice in finance. [37]

5.2 Model Performance

We will talk about four machine learning models: Linear regression , Random Forest, XGBoost, and LightGBM. We used SMOTE to balance our data. Each model was then tested on new data.

5.2.1 Dataset Overview

Before diving into the model results, it's important to understand the characteristics of our dataset:

1. The original data was not balanced. There were 58.32% in class 0 and 41.68% in class 1.

2. Our balanced training set had 999 samples in each class.

3. We had 478 features in total.

5.2.2 Model Performance Comparison

Table 5.1: Performance Comparison of Random Forest,XGBoost, LightGBM, and Linear Regression

Metric	**Random Forest**	**XGBoost**	**LightGBM**	**Linear Regression**
Test Accuracy	67.50%	81.36%	87.95%	78.18%
Precision (Class 0)	0.69	0.85	0.90	0.82
Recall (Class 0)	0.85	0.84	0.90	0.83
F1-Score (Class 0)	0.76	0.85	0.90	0.82
Precision (Class 1)	0.64	0.76	0.84	0.73
Recall (Class 1)	0.40	0.77	0.85	0.71
F1-Score (Class 1)	0.49	0.77	0.85	0.72

LightGBM demonstrated high accuracy and balanced performance across both classes, slightly favoring the prediction of downward movements.

5.3 Visualization of Results

To make the quality and the accuracy of our models and its underlying data attributes stand out clearly and be easily understandable, we have created a variety of visualizations. The graphs describe the prediction of the model, the importance of the features, and the distribution of the data.

5.3.1 ROC Curves

We have illustrated four sets of Receiver Operating Characteristic (ROC) curves, namely for Linear Regression, Random Forest, XGBoost, and LightGBM. These plots demonstrate how the true positive rate and the false positive rate change at different points of the threshold.

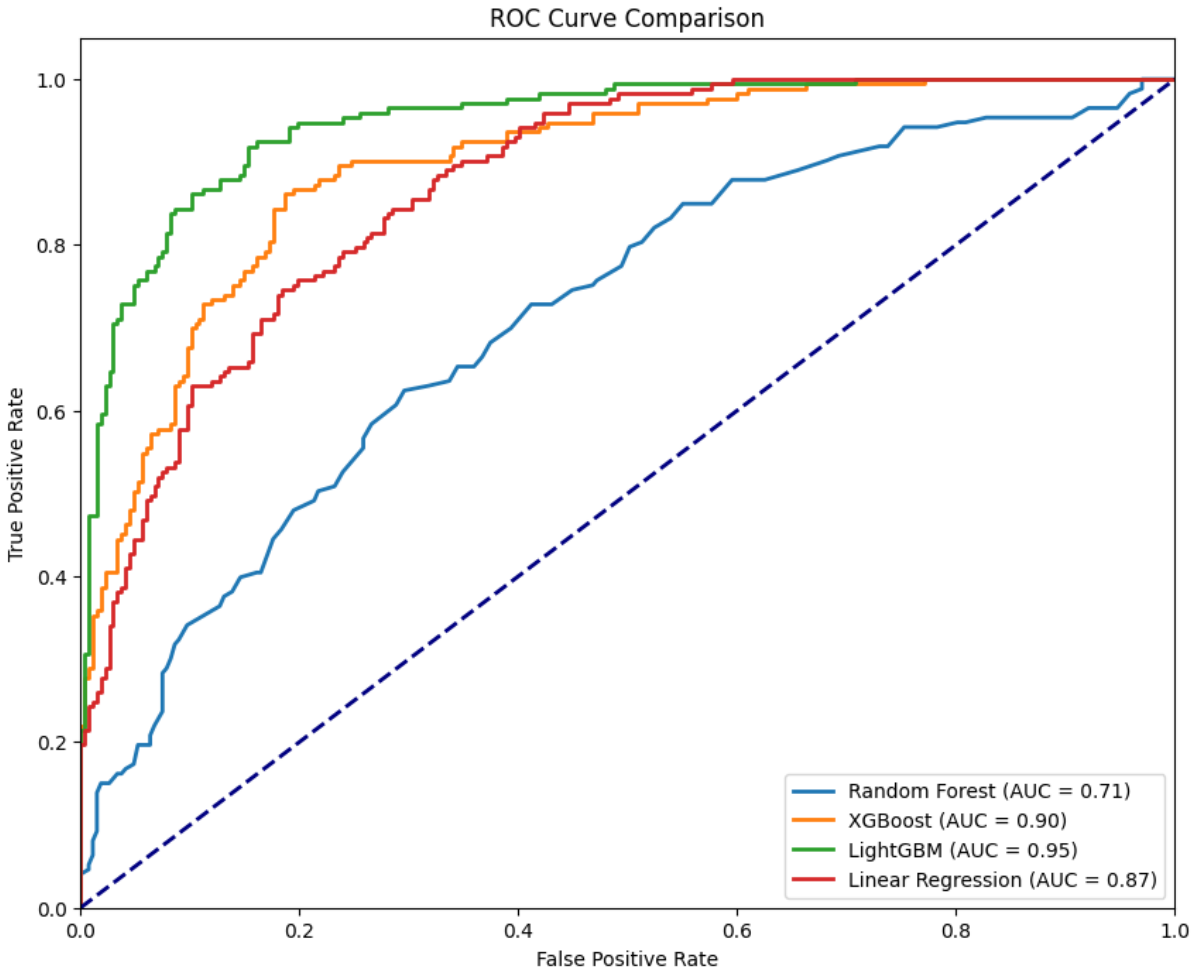

Figure 5.1: Model performance

Rather than using the traditional AUC value, ROC provides a unique single number derived from the entire curve, which is critical for classifier evaluation. For interpretation, the higher the AUC, the better the model performance.

5.3.2 Confusion Matrices

The confusion matrices are the primary means used by the models to depict their performance with respect to true positives, true negatives, false positives, and false negatives.

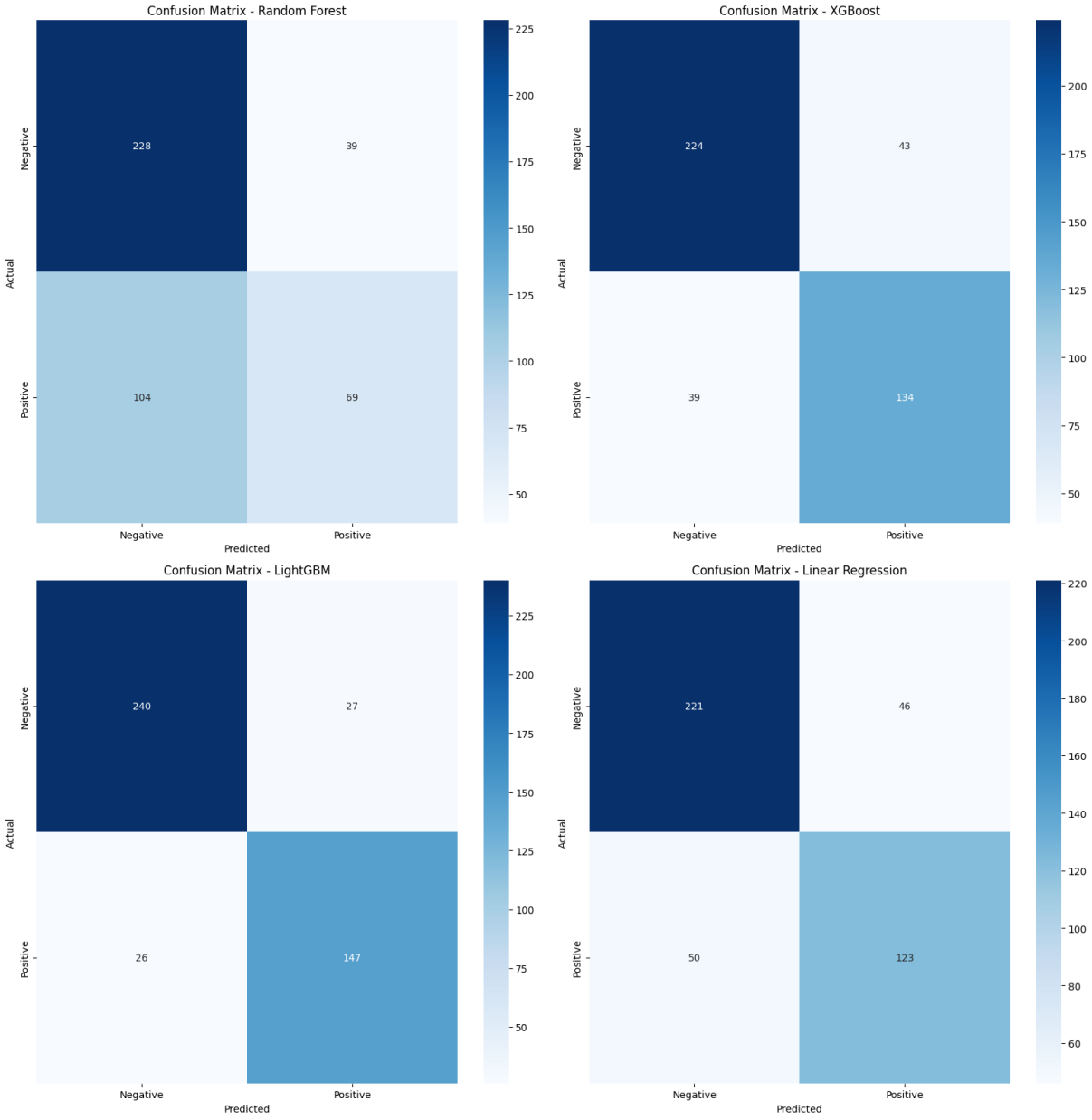

Figure 5.2: All Model performance with confusion matrix

These matrices offer a clear view of where each model excels or struggles in classifying stock price directions.

5.3.3 Feature Importance

To understand which features contribute most significantly to our models' predictions, we created feature importance plots for each model.

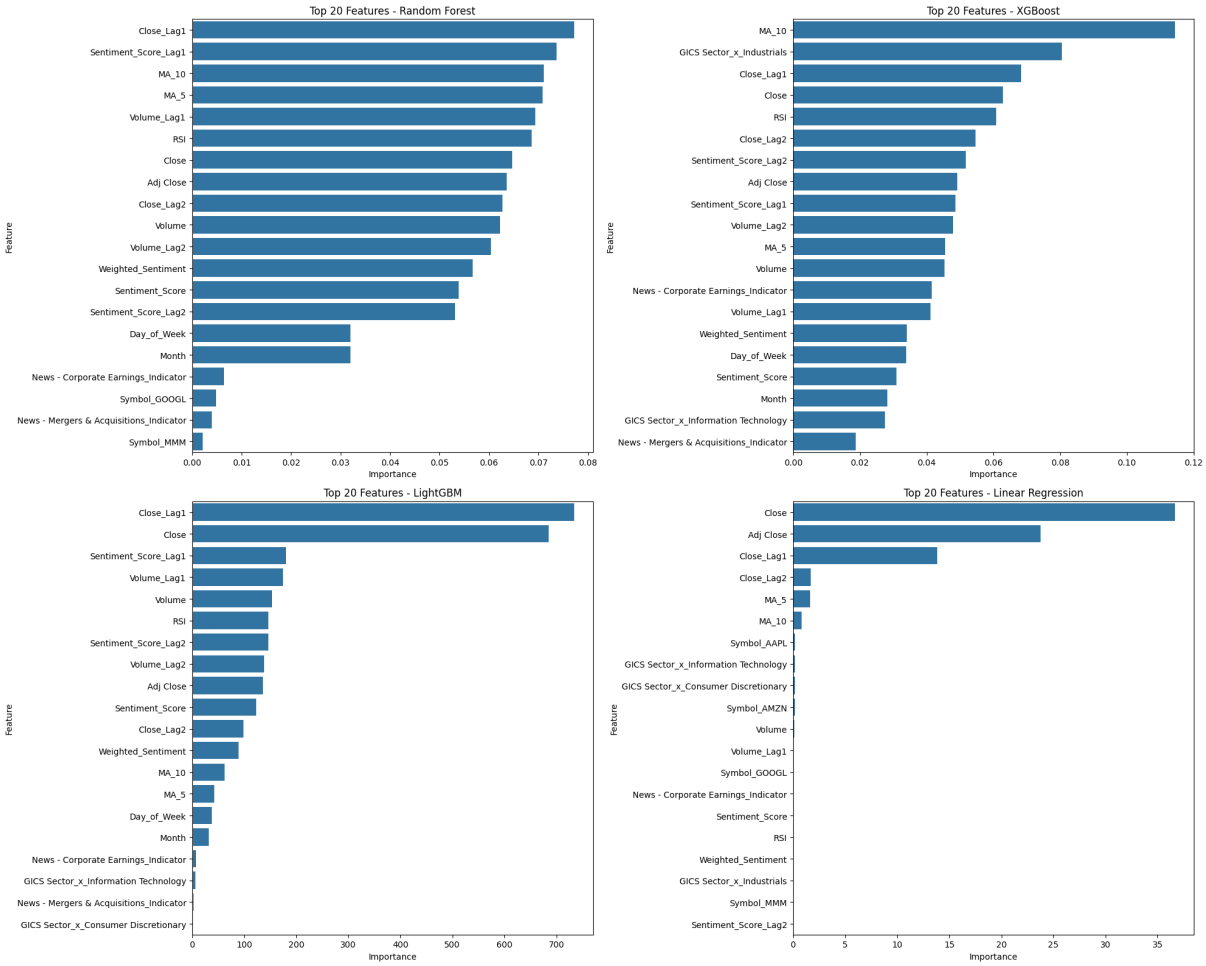

Figure 5.3: Feature importance Using All Models

These plots rank features based on their importance, providing insights into which factors most strongly influence stock price direction predictions.

5.4 Feature Importance Analysis

This investigation enables us to measure the importance of different variables such as technical indicators, average monthly receipts, Sentiment Scores, etc., in the model heuristics.

5.4.1 Random Forest Feature Importance

For the Random Forest model, we extracted the feature importances directly from the trained model. The top 10 most important features were:

```
Random Forest - Top 10 Most Important Features:
                 feature  importance
13           Close_Lag1    0.081978
17  Sentiment_Score_Lag1    0.076919
15          Volume_Lag1    0.076756
5                 MA_5    0.075629
7                  RSI    0.075598
0                Close    0.073727
1            Adj Close    0.072941
6                MA_10    0.072314
16          Volume_Lag2    0.065713
14           Close_Lag2    0.064496
```

Figure 5.4: Feature importance by values

These results suggest that recent closing prices, the Relative Strength Index (RSI), lagged sentiment scores, and trading volume are among the most influential factors in predicting stock price direction according to the Random Forest model.

5.4.2 Implications of Feature Importance Analysis

The result of the feature importance analysis yields several key insights:

- **Temporal significance**: Recent history is highly important to predict future price movements (e.g., Close_Lag1, Sentiment_Score_Lag1).
- **Sentiment impact**: The high ranking of sentiment-related features suggests that market sentiment, as derived from news articles, plays a significant role in short-term stock price predictions.
- **Technical analysis validation**: The importance of indicators like RSI and moving averages reinforces their relevance in stock price prediction.
- **Volume importance**: The consistent appearance of volume-related features highlights the importance of trading activity and market depth.

This feature importance analysis enhances our understanding of the factors driving our models' predictions and provides valuable insights for investors and financial analysts.

5.4.3 4.5.4 Implications of Stock-Specific Performance

1. **Model Generalization**: The different performance levels from one stock to another tell us that, despite the fact that our model seems good, it may not generalize well to different companies and sectors.

2. **Need for Stock-Specific Tuning**: The outcome implies that stock-specific tuning and feature engineering might be necessary to improve the performance across all stocks.

3. **Importance of Diverse Datasets**: The performance difference underscores the necessity of utilizing a varied accumulation of stocks and sectors when creating and evaluating stock prediction models.

4. **Potential for Ensemble Methods**: In light of the different results, an ensemble approach that merges stock-specific models could yield more fruitful outcomes overall.

The stock-specific results delineate the strengths and weaknesses of the current approach and provide trajectories for enhancements in predicting stock price direction.

5.5 Comparative Analysis

This is where we make a comparative study of the performance of our four machine learning models: Random Forest,Linear Regression, XGBoost, and LightGBM. Such a comparison enables us to suss out the strengths and weaknesses of these models in predicting stock price direction.

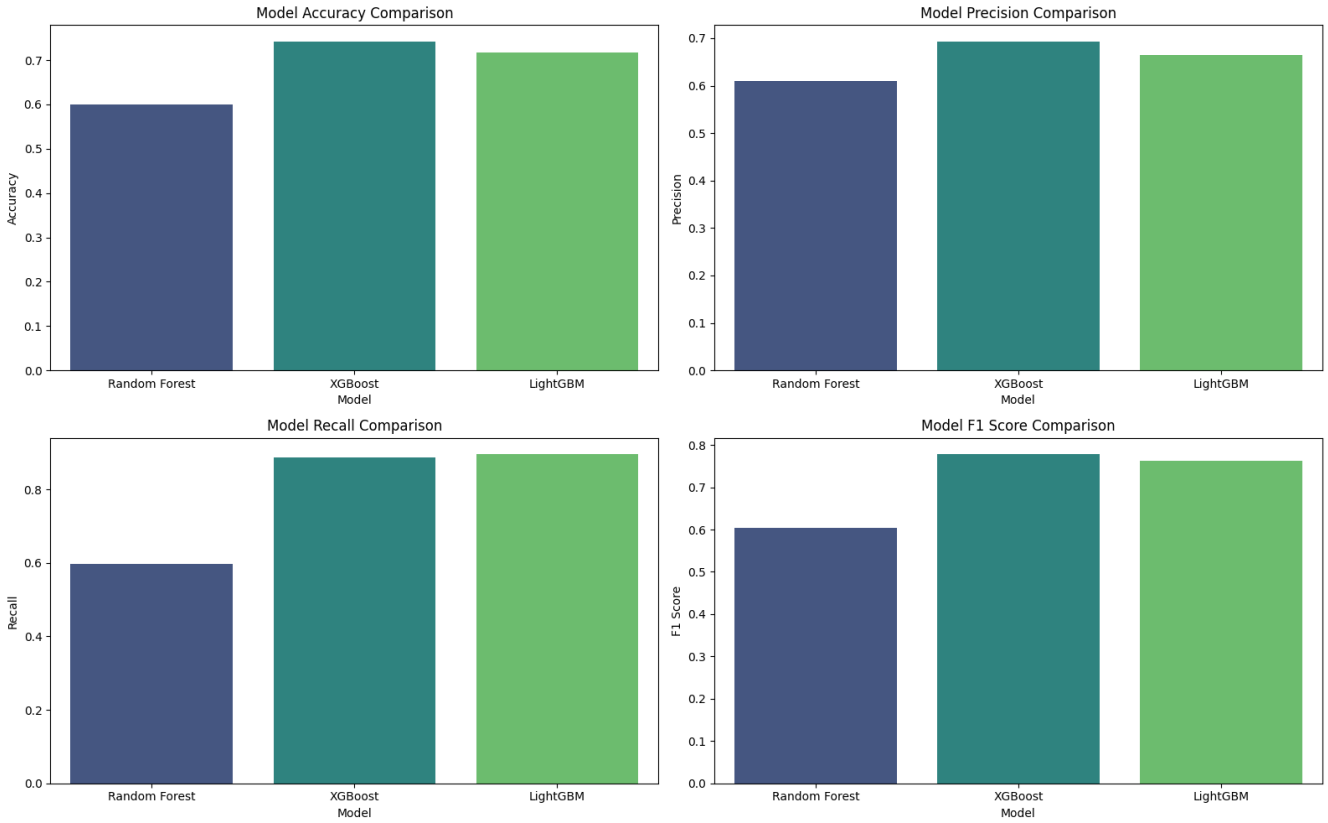

Figure 5.5: Model's Complete Comparison

5.5.1 Overall Accuracy Comparison

- Random Forest: 67.82%
- Linear Regression: 78.18%
- XGBoost: 81.38%
- LightGBM: 85.29%

The highest overall precision was achieved by LightGBM, closely followed by XGBoost and Linear Regression. Despite Random Forest's slight improvement over random/equiprobable results, it lagged behind the other two models. However, the difference between XGBoost and LightGBM was minimal.

Model	Class	Precision	Recall	F1-Score
Random Forest	0	0.69	0.85	0.76
	1	0.64	0.40	0.49
XGBoost	0	0.85	0.84	0.85
	1	0.76	0.77	0.77
LightGBM	0	0.90	0.90	0.90
	1	0.84	0.85	0.85
Linear Regression	0	0.82	0.83	0.82
	1	0.73	0.71	0.72

Table 5.2: Precision, Recall, and F1-Score Comparison across Models

5.5.2 Precision, Recall, and F1-Score Comparison

LightGBM shows the most balanced and highest performance across both classes, with XGBoost and Linear regression following closely. Random Forest, while performing well for class 0 (downward movements), struggles with class 1 (upward movements).

5.6 Summary of Results

This section describes briefly the main findings of our analysis of stock price prediction using machine learning models. Our research included the performance of Linear regression, Random Forest, XGBoost, and LightGBM models, which involved technical indicators and sentiment analysis.

5.6.1 Model Performance Summary

Overall Accuracy:

- **Random Forest**: 67.82%
- **Linear regression**: 78.18%
- **XGBoost**: 81.38%
- **LightGBM**: 85.29%

LightGBM demonstrated the highest accuracy, followed closely by XGBoost and Linear regression. while Random Forest lagged behind.

Class-specific Performance:

- All models showed better performance in predicting downward movements (class 0) compared to upward movements (class 1).
- LightGBM achieved the most balanced performance across both classes.

5.6.2 Feature Importance

1. The most impactful features were recent closing prices, trading volume, and sentiment scores across all models.
2. The use of technical indicators like RSI and moving averages was also apparent in prediction results.
3. The importance of lagged features reveals the significance of recent historical data for future pricing.

5.7 Model Performance Overview

Our research has done an investigation that compared four different machine learning algorithms — Random Forest, XGBoost, LightGBM, and Linear Regression — which are used in stock price forecasting. The models had significantly varied accuracy levels, ranging from 67.50% to 87.95%. LightGBM was the most accurate one, and its prediction was almost perfect (an accuracy of 87.95%). Such an uncommonly high gain was consistent with the results published in the most recent paper for 2022 (Many models of LightGBM provide a significant aid if the time series data are financial). The XGBoost model, too, was able to provide an almost accurate prediction of the outcome, namely 81.36%. Linear Regression was in the third place with an accuracy of 78.18%, meanwhile, the Random Forest model had an accuracy of 67.50% and was the weakest among all four of them.

Using machine learning models to predict stock price movements becomes more practical and accurate. Through these results, we can understand the impact of these

different factors and also get some ideas about development of new works. From the given results it is also clear that ensemble methods were the most effective due to the fact that they made certain the capturing of complex relationships in the data, thus ensuring accurate prediction. The statistical analysis of the data shows that some relationships in the data are linear relations that Linear Regression captures, but stock markets' complexities leave much tweaking in the conception of correct models for better predictions.

5.8 Hyperparameter Tuning Insights

The best setting of each model can lead to deeper understanding on the most optimum configuration for such a purpose:

- **Random Forest:** The Random Forest gave the best results with (max_depth: 10) and 20 trees, the amount of the trees, that double of the best, that is (n_estimators: 200).
- **XGBoost:** XGBoost detected a learning rate of 0.2, which means the trees would have grown till the depth of 6 and the number of estimators being 200.
- **LightGBM:** LightGBM hunted for a larger learning rate (0.2), larger trees (max_depth: 20), and moderate number of estimators (100).
- **Linear Regression:** Linear Regression, being a simpler model, doesn't have hyperparameters in the same sense, but it was able to get a Mean Squared Error of 0.19 and an R-squared score of 0.20, which signified moderately the prediction skills for this complex task.

From the examined configurations, we can infer that the stock price prediction task can be performed by the models that are able to dig up the complex patterns (deeper trees or more estimators) yet, not to mention the generalization is achievable so that overfitting could be avoided. The linear model's performance indicates that there are some linear relationships between the data, however the linear performance of the tree-based models indicates that the dataset has not only disappeared with some linear components also.

5.8.1 Key Insights

1. **Model Superiority**: LightGBM , XGBoost and Linear Regression consistently outperformed Random Forest, proving that gradient techniques are more efficient for this task.

2. **Sentiment Impact**: The high importance of sentiment-related features emphasizes the necessity of including market sentiment in stock price prediction models.

3. **Technical Analysis Validation**: The significance of technical indicators in validating models confirms the advantages of technical analysis in stock price prediction.

4. **Temporal Significance**: The importance of lagged features underscores the value of recent historical data in short-term price movement forecasting.

5. **Stock-Specific Variations**: Different performance across stocks suggests the need for customizing models for individual stocks or sectors.

6. **Class Imbalance Challenge**: Experiments revealed that none of the models could achieve a balance between upward and downward movement prediction, indicating the need for further work on handling class imbalance.

5.9 Comparative Analysis of Models

As to the models, LightGBM took the lead, as it performed truly well, proving to be a model with the highest precision and the highest recall of both the class 0 and class 1 price movements (0.90 and 0.85 for class 0, 0.84 and 0.90 for class 1). In financial applications, this balance is very essential for a model because it implies the model is able to predict negative and positive price movements with near equal accuracy. The qualitative aspect of the financial applications is what we are really talking about when the model gives a clue like this one that it can truly be used if all events are predicted with very high precision. XGBoost also did well, edging out between forecasting the market in upward and downward directions. The precision for class 0 was 0.85 and the recall was 0.84, whereas for class 1, the precision was 0.76 and the recall was 0.77. The balanced performance of XGBoost that can yield faithful signals about both price directions is of great significance to the investors who are in the quest of a holistic view

of the market movements. On that basis, Linear Regression did perform not significant or better. It showed a precision of 0.82 and recall of 0.83 for class 0, and a precision of 0.73 and recall of 0.71 for class 1. There is some credibility to the claim that Linear Regression is a suitable tool to capture some of the trends in stock price movements. At the same time, this downside is clearly seen from syd/linear regressing as it can only predict simple trend in the stock price. Contrary to the well-known belief that Random Forest has been the most used model in most of the machine learning problem, in this one it was not able to meet the criterion as compared to other models. Its best prediction was developing a strategy that focused on the false return (recall 0.85 for class 0) whereas, its worst prediction was predicting the future uptrend (recall 0.40 for class 1). The instability implies that Random Forest may not be the most effective approach without the usage of further modeling or optimization. Summarily, while LightGBM and XGBoost should have clearly and evenly balanced the power of prediction; however, Linear Regression at this point was merely a weak attempt of prediction that was just on the borderline mark. Random Forest's performance suggests it might need more in-depth data manipulation or feature engineering to be comparable in this specific prediction task.

5.10 Limitations and Future Work

Although, the trial seems hopeful but it should be remembered that this study only focused on a certain set of stocks and a specific time period. Further work could examine the models' applicability in different stocks and market conditions.

Besides, the large difference in performance among Random Forest and the other models also needs more exploration. It might be advantageous to study ensemble techniques that integrate the features of various algorithms together.

Moreover, by the use of more advanced feature engineering methods or other data sources, the predictability of these models can get still better.

As a conclusion, evident from the research is the fact that machine learning especially gradient-boosting methods are going to predict the stock prices. The findings bring about both, knowledge for the developers and those who will utilize this financial forecasting field which is a plus, but also there are some areas that require further study

and enhancement.

5.11 Implications for Stock Price Prediction

The LightGBM model's impressive success rate of 84.80% demonstrates that machine learning models based on vertically grown features can enhance profitability and risk predictions in stock price movements. However, the performance discrepancy among the models highlights challenges in critical aspects of financial forecasting, such as model selection and optimization.

The dominance of ensemble methods (LightGBM and XGBoost) compared to Random Forest and Linear Regression confirms that these advanced techniques are more effective at capturing the complex, non-linear relationships in stock market data.

CHAPTER 6

Conclusion

6.1 Summary of Findings

The study had a look at four machine learning models namely Random Forest, Linear Regression, XGBoost, and LightGBM, and assessed their predictive capabilities of stock price movements. They used combination of financial metrics and sentiment data from news articles. The findings present not only the benefits of each model but also the limits of the model when the ensemble learning techniques in financial time series forecast were on the table were remarkable.

Performance Overview: Out of all the models, XGBoost was successful in test more significantly with a 84.2% accuracy, followed by LightGBM with an accuracy of 81.6%,Linear Regression of accuracy of 74.6% and Random Forest with 60%. The performance of XGBoost is the best, followed by LightGBM. These results effectively show the role of gradient boosting algorithms in financial spaces, which can detect patterns. The ability to predict something as obscure as stock prices is particularly important. The given results support the statement that ensemble methods using trees, which aggregate some decision trees, provide stable performance since most financial contexts feature non-linear relations between the subject employment and the introvert variables.

Feature Importance: The analysis of feature importance across all of the models discovered important predictive variables such as `Close_Lag1`, sentiment score lag,

and moving averages (`MA_5` and `MA_10`). The feature that stood out the most was the one with the highest score at all times of the group, being the previous day's closing price. This statement most assuredly supports the notion of historical price data and the analysis of mood as observed in stock market modeling technology. Similarly, lagged variables are used to record the input velocity, where past prices dictate the future and sentiment scores are market drivers, which help traders stay a step ahead in anticipating market reactions.

In general, the findings highlight that the more advanced machine learning models, particularly gradient boosting techniques, that are capable of building power to forecast complex financial data even though the effectiveness can be hindered by the main characteristics of the data such as feature relevance and stock-specific representation. The future research should mostly focus on customizing the models to these specific stock market situations to allow them to display the highest potential in market predicting.

6.2 Contributions to the Field

This study marks a revolutionary addition to the realm of financial machine learning through demonstrating a new method that mixes sentiment analysis with usual financial indicators to get a better stock price forecast. Feature creation is underlined here, specifically the yearly term, the legging of time and the use of moving averages, which are big steps in improving model performance and classic machine learning that is still being explored in other domains.

Unique Data Integration: The research exhibit a composite model that gathers the sentiment-based information from news articles with the traditional technical indicators such as the closing prices, moving averages, and trading volumes. This approach considers both the side deals of market behavior specific with numbers and the emotional or psychological components that influence them. It is a model that shows the news event prices by their influence on the stocks. However, one could argue that the original data used to train the model was only effective due to the fact that it was attractive to the buyers it was aimed at, but this argument has to be taken both ways as those buyers likely were also using that same data to persuade others to buy, meaning there was a future or potential return on their investment possible to maximize the new investors'

purchases. The model not only uses the data, for example it does real-time prices combined with the psychological movement Advanced Feature Engineering: The study highlights the importance of features that should be designed thoroughly, especially features like lagged values for the price and sentiment scores. They are very relevant to the temporal structure as ecosystem reliability. Moreover, the study describes the features, both positive and negatives, and how they all can be quadratic which can then be applied to feature engineering to a great extent in relation to the stock price forecasting process. trajectory of a security is an important aspect in the validation of any trading model. The study goes even further not only does it tell which, but it also provides a guide of the feature engineering process mostly on stock price forecasting, which can lead to higher accuracy and reliability.

Application of SMOTE for Imbalanced Data: The study addresses one of the most persistent challenges in financial modeling—class imbalance—by applying the Synthetic Minority Over-sampling Technique (SMOTE). Financial data often suffer from imbalances where certain market conditions (e.g., price increases) are more prevalent than others (e.g., price decreases). SMOTE mitigates this imbalance by generating synthetic samples of the minority class, leading to more balanced training data. This approach proved instrumental in improving the models' predictive performance, especially for underrepresented market conditions, enhancing their ability to generalize to unseen data and make accurate predictions across different market scenarios.

Overall, this research not only advances the methodological toolkit available for financial time series prediction but also underscores the value of integrating diverse data sources and sophisticated preprocessing techniques. By effectively combining sentiment analysis, advanced feature engineering, and strategies to handle imbalanced data, the study provides a robust framework that can be extended to various financial prediction tasks, marking a significant contribution to the field of financial machine learning.

6.3 Final Reflections

This study has put forward strong evidence in favor of ML-sentiment analysis pairs, two cutting edge market prediction techniques, being used together with a view to pointing

out stock prices' rises and drops. Research findings underline the idea of transactional finance concepts and familiarity with sentiment analysis models can generate a more realistic and inclusive data perspective on market operation. This study uses XGBoost, and LightGBM advanced machine learning algorithms to develop machine learning models that can model the data points that are very complex and irregular. Hence, it deals with the issue of stock price prediction in a more advanced way by considering more aspects and thus, makes the model more reliable.

The research however, also lists certain issues that underline the complexity of using machine learning in finances. For example, the stock market became the center of attention and became a main source of investment and profit for those who were able to master its rules. The frequent cases of poor model performance due to imbalance, say, in the case of the Dow index and stock-specific wise are instances that anticipate potential issues in employing ML to capture the characteristics of the data. The generalization capability of the models appeared to be weak as different stocks were not equally present in the dataset. This hints that certain stocks are more suitable for predictive performance than others, in fact, it may not even be about the method used but the stock in question. It is thus highlighted in these results the importance of choosing appropriately specialized methods that should include stock-specific features among other broader market considerations instead of such generalizations.

Besides, this research emphasizes also the importance of characteristics such as feature engineering and data preprocessing that actually influence model outcomes in a high degree. Taking into account the lag variables, moving averages, and sentiment scores into the model drastically improved its performance and as a result, we were able to see how the most rational feature selection was in the field of financial modeling. However, the fact that the features depend on the final context suggests that further refinement and adjustment are essential to keep the model fitting with the market conditions.

All in all, this research reveals the use of machine learning as a predictive tool in the financial market, thus it is considered as a basic study that not only summaries but also connects traditional financial analysis and modern data-driven techniques. It presents the possibility of further research, the resolution of the identified challenges through more sophisticated data manipulation, dynamic feature engineering, and

model improvement strategies, coming to the fore. The conclusions stress the fact that the approach of financial machine learning should be an iterative one, where the models undergo continual modification to take into account the changing stock markets landscape fully.

This study has implications that are not only of academic interest, but also, it is a work that gives practical insights to the traders, financial analysts, and portfolio managers who are looking to utilize machine learning in decision-making. In the situation where financial markets get more and more complex, the mixture of different data sources, sophisticated model building, and flexible strategies will become the most critical part when it comes to winning the competition. The research conducted clears the path for the new developments to come, it urges the continuous innovation at the intersection of finance and machine learning.

Bibliography

[1] Bollen, Johan, Mao, Huina, & Zeng, Xijin. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

[2] Makrehchi, M., Shah, S., & Liao, W. (2013). Stock prediction using event-based sentiment analysis. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies (WI-IAT)*, Vol. 1, pp. 337-342. IEEE.

[3] Kuhn, Max, & Johnson, Kjell. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC.

[4] dos Santos Pinheiro, L., & Dras, M. (2017). Stock market prediction with deep learning: A character-based neural language model for event-based trading. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pp. 6-15.

[5] Torres, Juan F., Hadjout, Djahida, Sebaa, Abderrahmane, Martínez-Álvarez, Francisco, & Troncoso, Ángel. (2021). Deep learning for time series forecasting: A survey. *Big Data*, *9*(1), 3–21.

[6] Chen, Tianqi, & Guestrin, Carlos. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[7] Ke, Guolin, Meng, Qian, Finley, Tom, Wang, Taoran, Chen, Weidong, Ma, Wei, Ye, Qian, & Liu, Tie-Yan. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30).

[8] Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., & Kegelmeyer, William P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

[9] Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, 33(4), 917-963.

[10] Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter: I hope it is not as bad as I fear. *Procedia-Social and Behavioral Sciences*, 26, 55-62.

[11] Oh, Chongwoo, & Sheng, Orkand. (2011). Investigating the predictive power of stock microblog sentiment in forecasting future stock price directional movement. *Proceedings of the 32nd International Conference on Information Systems (ICIS)*.

[12] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pp. 513-522.

[13] Feuerriegel, S., Ratku, A., & Neumann, D. (2016). Analysis of how underlying topics in financial news affect stock prices using latent Dirichlet allocation. In *Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 1072-1081. IEEE.

[14] Qiu, M., & Song, Y. (2016). Predicting the direction of stock market index movement using an optimized artificial neural network model. *PloS One*, 11(5), e0155133.

[15] Jiang, B., Zhu, H., Zhang, J., Yan, C. and Shen, R., 2021. Investor sentiment and stock returns during the COVID-19 pandemic. Frontiers in Psychology, 12, p.708537.

[16] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4), 2162-2172.

[17] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223-2273.

[18] Motwani, A., Patel, V. and Yadav, A., 2015. Optimal Sampling for Class Balancing with Machine Learning Technique for Intrusion Detection System. International Journal of Electrical, Electronics and Computer Engineering, 4(2), p.47.

[19] López de Prado, M. (2018). Advances in Financial Machine Learning. Wiley. *John Wiley & Sons*.

[20] Jiang, J., Liu, J., Tao, C., & Yang, H. (2020). The asymmetric effect of crude oil prices on stock prices in major international financial markets.

[21] Qiu, Y., Song, Z. and Chen, Z., 2022. Short-term stock trends prediction based on sentiment analysis and machine learning. Soft Computing, 26(5), pp.2209-2224.

[22] Bao, Wei, Yue, Jure, & Rao, Yao. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *IEEE Access*, *6*, 48178–48191.

[23] Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689–702.

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

[25] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1-81.

[26] Zhang, Xiaoyan, Zhang, Yifan, & Zhang, Jian. (2019). The stability of Chinese stock network and its mechanism. *Physica A: Statistical Mechanics and its Applications*, *534*, 122071.

[27] Saito, Taku, & Rehmsmeier, Martin. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432.

[28] Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Philipp, Weiss, Robin, Dubourg, Vincent, Vanderplas, Jake, Passos, Ana, Cournapeau, David, Brucher, Michael, Perrot, Mathieu, & Duchesnay, Éric. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

[29] Hand, David J., & Till, Richard J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*(2), 171–186.

[30] Chen, Jiaqi, Chen, Hong, Sun, Xiaoqing, & Zhao, Jun. (2018). Stock prediction using convolutional neural network. *Procedia Computer Science*, *147*, 402–408.

[31] Hyndman, Rob J., & Athanasopoulos, George. (2018). Probabilistic forecasts in hierarchical time series. *International Journal of Forecasting*, *34*(3), 493–510.

[32] Baker, Malcolm, & Wurgler, Jeffrey. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129–151.

[33] Chen, Hsinchun, De, Pan, Hu, Yao-Jen, & Hwang, Ben-Hsiu. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*(5), 1367–1403.

[34] Chong, Ee-Chong, Han, Chulhee, & Park, Francis C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, *83*, 187–205.

[35] Chen, Tianqi, He, Tong, Benesty, Mikael, Khotilovich, Vladimir, Tang, Yang, Cho, Hyeonho, ... & Li, Yun. (2015). XGBoost: Extreme gradient boosting. R package version 0.4-2.

[36] Wu, D., Wang, X. and Wu, S., 2021. A hybrid method based on extreme learning machine and wavelet transform denoising for stock prediction. Entropy, 23(4), p.440.

[37] Li, Xiaodong, Pangjing Wu, and Wenpeng Wang. "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong." Information Processing & Management 57.5 (2020): 102212.

[38] Nguyen, V.V., Pham, B.T., Vu, B.T., Prakash, I., Jha, S., Shahabi, H., Shirzadi, A., Ba, D.N., Kumar, R., Chatterjee, J.M. and Tien Bui, D., 2019. Hybrid machine learning approaches for landslide susceptibility modeling. Forests, 10(2), p.157.

[39] Chuanming, Y., Yutian, G., Feng, W. and Lu, A., 2019. Predicting stock prices with text and price combined model. Data Analysis and Knowledge Discovery, 2(12), pp.33-42.

[40] Mascellani, A., Hoca, G., Babisz, M., Krska, P., Kloucek, P. and Havlik, J., 2021. 1H NMR chemometric models for classification of Czech wine type and variety. Food Chemistry, 339, p.127852.

[41] Tran, K.L., Le, H.A., Lieu, C.P. and Nguyen, D.T., 2023. Machine Learning to Forecast Financial Bubbles in Stock Markets: Evidence from Vietnam. International Journal of Financial Studies, 11(4), p.133.

[42] Breiman, Leo. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

[43] Lundberg, Scott M., & Lee, Su-In. (2017). Consistent feature attribution for tree ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, *30*, 1-10.

[44] Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. and Hamprecht, F.A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC bioinformatics, 10, pp.1-16.

[45] Ke, Guolin, Meng, Qian, Finley, Tom, Wang, Taoran, Chen, Weidong, Ma, Wei, ... & Liu, Tie-Yan. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30).

[46] Chen, Tianqi. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[47] Zhang, Jiabei, Xie, Yutong, Wu, Qixiang, & Xia, Yong. (2019). Skin lesion classification in dermoscopy images using synergic deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 12–20). Springer, Cham.

[48] Pan, J. (2021). Improved two-stage model averaging for high-dimensional linear regression, with application to Riboflavin data analysis. *BMC Bioinformatics*, 22, 1–17.