University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

Predictive Analytics for Stock Prices Using Machine Learning Techniques

Rashi Chandel

Supervisor: **Mr. Rishideep Roy**

September 16, 2024
Colchester

Abstract

This dissertation examines the use of advanced methods of machine learning and statistical models that can be used to handle stock price prediction and market trend analysis. The ARIMA and Prophet models will be applied, together with unsupervised clustering approaches like K-means, to derive insight into the short-term prediction of stock prices and long-term trends. The researcher used huge stock price data from around the world to train and test their models, keeping special care for industries and stocks exhibiting both linear and nonlinear behaviour.

This paper also informs in detail how time series forecasting is subjected to many challenges along with seasonality and volatility, even due to some external factors like holidays and big financial events. Each one is ranked according to the rigorous evaluation by MAE, RMSE, and MAPE metrics. The results bring out the strengths of ARIMA in stable market conditions and Prophet's relative strengths in handling complexity with strong seasonality.

Besides time series forecasting, the current dissertation also applies a clustering analysis to group the stocks by their volatility and performance task giving investors a more profound insight into the risks and opportunities of the market. The results of this study have pointed out the fact that such a combination of statistical techniques with machine learning algorithms significantly improves the accuracy of stock price predictions and subsequently yields better decision-making by investors.

The research has added to the ever-growing domain of financial forecasting in a way that has identified how practically viable machine learning models are in stock markets while simultaneously developing a skeleton for further research in predictive analytics.

Acknowledgment

I would like to begin by giving my major thanks to the Lord. Without his guidance, constant grace and light, surviving through this journey would not have been possible. This whole journey at every step has been a blessing, and for that, I shall always be grateful.

Mummy and Papa, thank you for believing in me through all these years. Thank you for the constant support and encouragement that have provided me with endless strength to move forward during such tough times. I have been pretty fortunate about having both of you with me. To my Sister Ashi, you have always been my rock on which, whenever there is any need for advice or just a word of encouragement, I could always count. Thank you for being my rock, my constant support system.

A word of gratitude to the best of my friends, Danish in fact, I would not have made it through the master's without him. The friendship and support he has given me means a lot to me, and I can never thank him enough for having my back. To Nadeen, who has been there for the whole year, helping me through it all thank you for your patience, your kindness, and your friendship. And to Nimra, too, who has always been like an elder sister to me thank you for showing the way and keeping an eye out for me.

I am especially indebted to my supervisor, Mr. Rishideep Roy. His guidance and patience made a huge difference. You pushed me to grow, and for that, I'm grateful. Specially to University of Essex and all the staff members for being so supportive and helpful throughout the year.

To all the strong, inspiring women who keep me going day in and day out, thank you. Your strength, resilience, and passion have kept me constant to this moment, and as much as this achievement is mine, it is yours.

Thanks from the bottom of my heart to everyone who has been a part of this journey.

Contents

List of Figures

CHAPTER 1

Introduction

The stock market has, for a long period, remained at the centre of financial investments. It is a place where numerous individuals and institutions buy and sell securities for financial gain. The dynamics of the stock market remain very mysterious as a host of factors influence this segment; economic indicators, market sentiments, geopolitical events, and company-specific news remain top among those causes [1]. All these put together create an environment of constant change, which does not make it all that easy to comprehend, especially for the new investor who is short on experience and information regarding investments that one should make intelligently.

Over time, several methodologies have evolved with the purpose of analysing the stock market's trending pattern and estimating price movements in the future. The traditional methods fundamentally rely on statistical models and financial ratios, which though useful, remain limited [2]. Most of these techniques require a great understanding of the markets and how to make sense of the data out of them, which makes them out of reach for amateur investors. Also, with traditional models, such extensive patterns and relationships may not be captured in a big, complicated dataset, which actually results in poor predictions [12].

In the last decade, AI and machine learning have drastically changed the way in which analysis in the stock market has been carried out [3]. AI-powered tools can process massive volumes of data, point out patterns, and even make predictions that were earlier impossible to attain with much detail [8]. Investors are given new opportunities with these tools, especially the new ones in the market, since they provide

insights that are not only data-driven but also easy to read [4]. These models analyse historical data for trend detection to forecast future movements in prices, thus enabling investors to make better decisions and manage portfolios with greater efficiency.

The dissertation discusses the application of AI to stock market analysis and focuses on an AI-powered investment guide for new investors. The study tries to embody advanced AI models in the analysis of stock market data so that it could identify the main trends and actionable insights which will drive investors to take a certain decision [9]. It implements some of the state-of-the-art techniques in time series forecasting like ARIMA and Prophet, along with clustering techniques such as K-means, on a rich dataset of global stock prices [3, 11].

This work ranges in stock prices from some of the world's most famous brands, starting from January 1, 2000, up to this date. The provided large dataset includes open and close prices, the volume of trading, dividends, stock splits, and other relevant financial metrics [15]. It contains information about the industry to which each brand belongs and the country it is from so that a more refined analysis of market trends across sectors and regions can be done.

The various data visualization techniques were applied in a descriptive manner to outline the main trends and patterns in the dataset. For instance, boxplots were drawn comparing the distribution of the closing prices among the different industries. Large differences in volatility and central tendency were depicted. Histograms present the distribution of the stock price, while correlation heatmaps depict the relationships between different financial metrics. These plots form the basis upon which more developed analyses rest and present intuitive means of grasping data that happens to be rather soft to comprehend.

In this respect, after the exploratory analysis, the study applies time series forecasting models in projecting forward future stock prices. It uses the ARIMA model since it can handle a non-stationary series of data in a way that it can model time series data and then do forecasts [2]. Another used model is the Prophet model, which was developed by Facebook; this is because it's robust in handling seasonality and holiday effects [6]. Both models compare each other using the root mean square error of their respective predictions, in order to find out which one of these approaches will be more plausible for any forecast of stock prices.

This paper performs K-means clustering for stock classification in volatility and performance, along with time series forecasting of stock prices [13]. Clustering represents one of the fundamental methods for discovering hidden structures within huge volumes of data [8]. In this context, clustering can be used for classifying stocks into different classes according to their historical price movement. Such analysis will provide insight into various risk or return profiles of those different stocks, thus helping the investors to make better portfolio diversification decisions.

The findings of this research will have important implications for new investors, who always have difficulties in understanding the intricacies of the stock market [1]. The application of AI-enhanced tools in the research brings structure into investment analysis and hence gives new investors an important tool to help make data-driven decisions. The AI models that are developed in this research enhance the accuracy not only of market predictions but also of practical tools concerning risk management and return optimization.

The stock market has always taken precedence in financial markets and investments for many years. It has been acting as a channel that is important, where securities are traded between individual investors and institutions. A stock market, by nature, is very dynamic and complex; hence, it easily responds to various causes of fluctuation in economic indicators, geopolitical events, market sentiment, or company news. These elements, together with constant changes in the market, make it hard even for investors to navigate through all that and make some sort of informed decision [3].

Traditionally, financial analysis has relied on statistical models, including regression analysis and time series forecasting. While these models have provided considerable insights, most have been constrained by assuming either dependence on historical data or linearity in assumptions. This might lead to oversimplifications and poor predictions in volatile market environments as it often does not take several unknown factors into consideration. Second, traditional models cannot always portray the relationships between variables intricately, thereby leaving novice investors in a lurch to decide on the most well-informed decisions. As has been rightly estimated by [9], this becomes a deciding factor.

In the last couple of decades, AI and ML have completely revolutionised the way stock market data is analysed. AI algorithms can process reams of data in real time

to provide insight by highlighting patterns and making predictions that are quite unimaginable by human analysts. The application of machine learning in AI models lets them learn from historical data and adapt to new market conditions, whereby investors get improved decision-making for framework building. These developments have been most useful for new investors whose portfolio managers can now fall back upon AI-driven tools while making intelligent investment decisions with less reliance on personal judgment.

The purpose of this dissertation is to investigate various directions in predictive analytics Various machine-learning techniques are being used for stock price prediction. It is focused on the usage of such models as ARIMA or Prophet in Time series forecasting and also k-means in clustering analysis. The idea is to arrive at an AI-awakened investment guide that would be elaborative enough in comprehension for the novice investor about stock market trends and thus make well-informed decisions. It draws on a large dataset of stock price data collected across the globe, providing insight into various industries and regions. Various machine learning models have been explored in order to better predict stock prices.

This dissertation has shown how AI is going to be used in revolutionising the analysis of the stock market. Advanced AI models have been applied here. This will involve research to bring new insights into the dynamics of the market, supported by a robust framework for Investment Decision-Making. This is especially relevant in the case of new investors who are able to make a diversified and risk-managed investment portfolio from insights gained herein that align with their financial goals.

CHAPTER 2

Literature Review

The stock market is an imposing and dynamic environment in which prices are pitted against a host of factors, such as those emanating from macroeconomic indicators and company-specific announcements [7, 11]. Over the years, different methodologies have been constructed to analyse and predict stock market movements; each of these methodologies has its strengths and limitations [5, 9]. The literature that follows. In this paper, major different methodologies and approaches for the analysis of stock market data tried so far are reviewed to underline the trend in their evolution from traditional statistical methods to state-of-the-art AI-driven ones.

2.1 Traditional Approaches to Stock Market Analysis

Traditional analysis of the stock market has been performed using fundamental and technical analysis. The financial statements, management, industry position, and economic environment are analysed to estimate the intrinsic value of a stock in fundamental analysis [9]. Common techniques include discounted cash flow analysis and price-to-earnings ratios. However, although fundamental analysis is effective in evaluating long-term potential, it largely lacks the foresight that would let it predict short-term market movements influenced by a variety of exogenous factors [14].

Technical analysis, in contrast, relies on historical price and volume data to establish trends and predict future price movements. Indicators like moving averages and Bollinger Bands form the backbone of this approach [5]. Technical analysis assumes that

all relevant information is already reflected in the stock price and that prices move in predictable trends. However, the reliance on past data and the subjective interpretation of chart patterns have led to criticism of this approach [3].

Traditionally, stock market analysis has relied on fundamental and technical analysis over the years. The intrinsic value of a stock fundamental analysis is based on financial data, company performance, and economic indicators. The usual ways to achieve this are through financial ratio evaluation, balance sheets, and comparison to the industry. These techniques include the price-to-earnings ratio and discounted cash flow analysis amongst others [4]. However, relying totally on fundamental analysis enshrines weaknesses, This would particularly be so in the area of taking into consideration short-term fluctuations in the market. Since it will focus on the long-term growth perspective.

Technical analysis deals with the study of historical price and volume to try to predict future movements. It depends upon the view that relevant information concerning a stock is incorporated in the stock price and that the market price itself has a tendency to follow a pattern. General technical analysis tools include moving averages, Bollinger Bands, and the Relative Strength Index, or RSI [5].Even though technical analysis is in wide use among traders, the discipline as a whole has faced a host of criticisms because of its subjective nature and reliance upon historical data [9].

2.2 The Emergence of Time Series Models

So, as their analyses became increasingly complicated, time series models started to become vital in stock market analysis [6]. Time series analysis essentially represents the historical sequence of stock prices and makes a forecast of future values from it by determining structures that are hidden inside. Among these models, ARIMA stands as one of the most widely applied. It was introduced by Box and Jenkins in the 1970s [2]. The ARIMA model is particularly effective for non-stationary data since the mean and variance change with time [1].

ARIMA models first make the data stationary by differentiating the data and then applying the autoregressive and moving average components based on the dependencies that may exist within the data. Though efficient in the case of short-term predictions, the linearity assumption made by ARIMA will restrict it from capturing the nonlinear

nature of financial markets, which is usually complicated [7, 8]. Extensions such as ARIMAX and the development of nonlinear models have been some of the efforts towards resolving these limitations [9].

This complexity in the stock market data has resulted in the adaptation of time series models, which work better for the prediction of further stock prices based on trends created in history. The pioneer model in this regard is ARIMA, which was coined by Box and Jenkins in the 1970s. ARIMA models successfully grasp the temporal dependence of time series data, which makes this model very popular in stock price forecasting [2] developed a model that is called ARIMA. The linear form of ARIMA limits the nonlinear relationship modelling of ARIMA. According to [9], financial markets often have a nonlinear relationship.

Owing to these issues with the traditional time series model, other methods have been devised to take on the complexities around time series seasonality and missed data, for example. Prophet is developed by Facebook and it handles complexities around time series, including seasonality and missed data. This has made it very useful in stock market forecasting, particularly in those industries showing a high level of seasonality, since the trend and seasonal components of such time series can be captured.

2.3 Introduction of Machine Learning in Stock Market Prediction

With the introduction of machine learning, a paradigm shift occurred in the analysis of stock markets. While algorithms of machine learning are not explicitly programmed, they learn from data. They are especially fitted for stock market predictions since many of the relationships between variables are nonlinear [11, 8]. This is a great extension for stock market prediction because it makes models learn wittingly from data over time without necessarily having to be explicitly programmed. Stock price movements are predicted with techniques using random forests, support vector machines, and neural networks, among others. Random forests serve for classification problems for instance, whether stock prices will rise or fall based on historical data. These are some of the most interpretable models that can handle big data with categorical and continuous variables alike. In this regard, [10] mentioned that ARIMA models are applicable in

such scenarios.

Neural networks, specifically deep learning models such as Long Short-Term Memory (LSTM) networks, are also widely used stock market predictors. LSTMs are suited to the task of time series forecasting as they have the capabilities for capturing long-term dependencies found within the data, hence modelling very complex financial markets effectively [8].

The most common among them include random forests, support vector machines, and neural networks [6]. Decision trees and random forests are some of the most popular classification models. They allow a binary classification of target variables from input features to predict whether the prices of certain stocks will go up or down [10]. Both models are interpretable, support categorical and continuous variables, and turn out useful in manifold contexts of predictions.

The other popular technique is usually the SVMs, which have been utilised in a binary classification task relating to stock market predictions. The key idea of the support vector machines, usually referred to as SVMs, is finding a hyperplane which can classify data between classes with maximum margin [9]. On the other hand, this effectiveness in SVM may be highly sensitive to the choices of hyper-parameters and thus requires careful tuning [15].

Neural networks, specifically deep learning models, came into the limelight as they could model complicated, nonlinear associations with unprecedented success. RNNs and their variant, LSTM networks, are quite effective in time series prediction owing to the capability of capturing temporal dependencies [6]. LSTMs proved to be effective, especially in the area of stock market prediction, since long-term dependencies can be learned and also the vanishing gradient problem associated with traditional RNNs can be reduced [12].

Machine learning models for stock market predictions do face challenges such as overfitting, where the model performs well on the training data but fails on unseen data due to the noisy and volatile nature of stock markets [8]. The looming risks are taken care of by several strategies: cross-validation, regularization, and ensemble methods [7].

2.4 AI-Driven Investment Strategies

Artificial Intelligence-driven investment strategy introduced a new class of tools and platforms that offer custom recommendations to investors by making use of insights from large data sets [3]. Application of machine learning algorithms on price history, market sentiments, and macroeconomic indicators provide customised suggestions on investment [14].

Recently, AI investment strategies have become quite popular due to their great capability in handling data and making real-time decisions based on that data. AI investment strategies take advantage of AI algorithms in identifying patterns and trends in the stock market that contribute to strategically changing an investor's portfolio through efficient management of risk [6]. One of the key benefits of AI-driven strategies is that they can analyse not just historical data but also macroeconomic indicators, market sentiment, and even social media trends to help formulate more scientific predictions[10].

However, there is the challenge of "black box" problems the lack of comprehensibility by investors of many machine learning models, given how most of these models come up with their prediction. Such challenges to the broad adoption of AI-driven strategies question transparency and accountability. The potential homogenization of investment strategies used by a large number of investment portfolios could even lead to increased market volatility [1].

One of the main advantages of AI-driven strategies is that they can process huge volumes of data in real-time to aid investors in making quick responses to changes in the financial markets. This is especially helpful in high-frequency trading, as decisions have to be done at the level of milliseconds [11]. AI-driven strategies also help investors diversify portfolios and manage both risks and returns better by identifying opportunities that evade capture via traditional methods [10].

However, the application of AI to investment raises a number of ethical and practical issues. First, there is a lack of transparency in the AI models, sometimes referred to as "black boxes," where the manner of arriving at decisions is not readily comprehensible to the investor, who cannot ascertain attendant risks [9]. The generalised application of AI-driven strategies may homogenise the strategic aspects of the investment, thereby

creating more market distortions [6].

The study of the stock market has drastically evolved, and AI together with machine learning has opened new horizons and brought new challenges [14]. Of course, traditional methods of fundamental and technical analysis are useful; their combination with AI-driven techniques may even promise higher levels of accuracy and efficiency in stock market predictions [8]. Accordingly, the following sections outline the methodology and models applied in the present research and also mention their implications for investors.

CHAPTER 3

Methodology

This research embarks on a methodology that enriches several modern statistical techniques, machine learning algorithms, and clustering methods to predict stock prices of globally renowned brands. This approach would systematically be divided into data collection and preprocessing, time series forecasting with ARIMA and Prophet models, clustering analysis based on the volatility of the stocks, and finally model comparative performance evaluation. This report is multifactorial in approach and extracts meaningful insights from data, therefore providing a solid basis on which informed investment decisions can be made, especially by new investors.

3.1 Data Collection and Preprocessing

The first step in this research involved collecting and preparing a comprehensive dataset that serves as the backbone for further analysis. The dataset includes daily stock prices of well-known global brands, spanning from January 1, 2000, to the present, providing a rich historical perspective on stock market dynamics.

3.1.1 Data sources-Description

Material concerning stock prices is harnessed from reliable and credible financial databases to make it real and up to date. These financial databases avail a host of historical data on stock prices, among other related financial metrics that are critical

in carrying out robust analyses of performance trends within the stock market. Such selected data shall give a perspective into world stock prices from January 1, 2000, up to the current date, with appropriate insight into the long-term price movements and market dynamics across industries and regions.

The dataset contains basic financial variables that describe the main attributes of daily stock market activity. In this fashion, this set of features will be of primary relevance in analysing the performance of the individual stocks and market trends over some time.

A list of some of the included features in the dataset follows:

- **Date:** The 'Date' serves as a temporal reference point for the data set. It defines the exact day when the recording was made. Normally, this variable is necessary for any time series forecasting model because it provides an opportunity to track the chronological development of the stock prices. We also have to add here that knowledge of the temporal context of variation in stock prices is very important because it allows us to identify patterns, trends, and singularities of the data in question.

- **Open:** The open price in a stock is defined as the price level at that instant the market opens its position for the daily trading session. This forms the first impression of market sentiment and investor expectation that comes up at the beginning of the trading session. Looking at the opening price relative to other measures, such as closing price or volume of trade, may suggest something about the nature of early market reactions to overnight news and events or general economic conditions.

- **High:** It refers to the price, which reflects the highest value at which the stock has traded during the day. The figure reflects peak confidence that investors had for that particular day, signalling the maximum willingness of participants in the market to buy the stock. The high price is an important metric with respect to gauging the volatility and intra-day price dynamics of a stock.

- **Low:** This "low" price refers to the very lowest value it reached on that particular day of trading. This "low" price is thus another important determinant of the minimum confidence that investors had in that particular stock during that session

of trading. Just like the "high" price, the "low" price is another variable related to measuring intraday volatility and will be rather helpful with regard to determining how much pessimism there might have been in the market or how cautious investors were.

- **Close:** The "close" price is the last price at which that stock has traded when the market has "closed" for that day. This value is usually seen as the most important daily price level since it suggests the last sentiment of the market and consolidates all that day's activity. The closing price is used extensively in the analysis of stock market movement for the future and, in technical analysis, for the identification of trends and signals.

- **Volume:** "Volume" denotes the number of shares traded in one day. Volume helps in indicating liquidity and attempts to show a certain stock's interest. High volumes may symbolise high investor interest, along with market-breaking news, while meagre volumes could put across less activity. Volume is considered an important element in technical analysis because it gives signs of confirmation for the price trends and the relative strength of the price movement.

- **Dividends:** In this area, the stock's dividends, if any for the trading session, can be viewed. Dividend payments are part and parcel of the process, especially for long-term investors, since, mostly, they provide the total return on investment. Stocks that pay dividends usually have many investors looking for income along with capital appreciation. The dividend is a variable that has to be considered when looking at a stock's overall financial performance, factoring it into several stock valuation models.

- **Stock Splits:** Stock splits: This refers to the division of prevailing shares of a company into many shares with the aim of reducing their various prices. It might have a very good psychological implication on investors' perception and enhance market liquidity due to the fact that lower-priced stocks make more demand for security. A stock split is a relevantly important concept in being able to appreciate how historical price data is interpreted and correct comparison over time is made.

- **Brand Name:** The brand name identifies the concerned firm or stock. This is a

very crucial variable in the contextualising of data and thereby relating it to the market or industry trends as a whole. Once the brand name has been identified, the researcher can categorise different stocks into their respective industries and sectors, thus helping in sector-specific analysis and comparison.

- **Ticker:** The ticker symbol is an identification number in the stock trading of a particular stock. Normally, it is a short name involving several letters; hence, it acts as a snappy way of recording and analysing different stocks. This ticker symbol is thus very important in mapping data of particular stock between different applications to maintain uniformity in tracking any stock.
- **Industry Tag:** The industry tag classifies the stock according to the sector or industry it belongs to. In this way, this would allow for sectoral analysis where stocks can be grouped into respective industries and performance measured across time for such industries. It is good to use when identifying trends of particular sectors, for example, technology, health, and energy, and performing a comparison performance of stocks across those industries.
- **Country:**The country variable refers to the country in which headquarters or main operations of the firm are located. It is a very important variable for any geographical analysis and helps in explaining how different economic, political, and regulatory environments influence stock performance. Country of origin proves to be a useful segmentation tool in stock research, as it facilitates the ability of the researcher to control for specific country factors influencing stock prices, such as changes in national economic policies or geopolitical risks.

These features put together give a full view of daily stock activity and allow for in-depth analysis of stock market behavior. This variety in the variables of the dataset allows for time series forecasting and clustering analysis, thus making this dataset apt for long-term trend evaluation, risk assessment, and predictive insight into future stock price movement.

3.1.2 Data Cleaning

Cleaning is a necessary procedure to ensure that there are no errors or inconsistencies in the data set. In this study, cleaning would involve removing rows that have missing

values. The `na.omit()` function was done in R because missing data can result in biased results of the test. Some columns irrelevant to the entire analysis were removed and only the most relevant features such as Date, Open, High, Low, Close, Volume, Brand Name, Industry Tag, and Country.

3.1.3 Formatting of Dates

Correct time series date formatting is important to make findings visible. The conversion of the column "Date" into standard date-time format was effected using `ymd_hms()` from the `lubridate` package for correct interpretation of events over time, which is very important when it comes to forecasting future trends.

3.1.4 Feature Engineering

Feature engineering is understood as new variable creation based on available data with a view to enhancing the predictive powers of the model.Some of the features developed in this paper are as follows:

- **Lag Variables:** The closing prices of the previous day and the day before the previous day are named as, `Lag1` and `Lag2` respectively. In this dataset, these variables are included with the purpose of capturing short-run momentum or mean-reversion patterns of stock prices.

- **Moving Averages:** These are for a7-day (`MA7`) and 30-day (`MA30`) period, serving to smooth out the short-term fluctuations and emphasise longer-term trends.

- **Volatility:** We calculate volatility, defined as the standard deviation over the moving window of 7 days. This is another widely used proxy for the risk arising out of fluctuation in the prices of stocks.

- **Normalization:** Z-scores were used to normalise the features in order to bring all variables onto the same scale. This step is quite important for effective clustering. It was more appropriate, especially for variables such as Volume, where the value of one stock can be way different from another stock.

3.1.5 Segmentation

The segmentation was done by Industry and Country to allow the capturing of a higher level of detail in analysing stock performance in a given sector and region. This will be helpful to get insight into industry-specific trends and regional market dynamics, important for making informed investment decisions.

3.2 Time Series Forecasting

It is the most important aspect of the stock price prediction involves time series forecasting, that uses the trend of historical data as a basis to predict its future values. In this paper, two popular time Series forecasting models used include ARIMA and Prophet. Both the models complement each other in terms of better understanding and forecasting the movement of stock prices through their historical time series data.

3.2.1 ARIMA

ARIMA is also a dynamic, potent statistical tool that would forecast the prices for the future by considering the trend of the historical stock price. The ARIMA model proves efficient in handling time series data that shows consistency in trends or patterns over time and is thus perfect for financial markets, which are mostly linear or exhibit stationary behaviour.ARIMA consolidates three key components autoregression, differencing, and moving averages into one framework to capture the temporal dependencies, trends, and noise within the dataset.

Fitting the Model

The ARIMA model is represented as ARIMA(p, d, q), where:

- **p** - is the number of lag observations in the autoregressive part of the model,
- **d** - refers to the degree of difference needed to make the time series stationary,
- **q** - is the size of the moving average window used to remove any noise or fluctuations in the data.

The model was fitted to 80% of the dataset comprising the historical daily stock price data before the application of ARIMA, while the remaining 20% were held for testing and validation. This kind of split is important because most of the data are used for training the model, leaving reasonable portions unobserved during evaluation. The best parameters of the model were found by automatic preselection of values of p, d, and q.

For finding the best model parameters, select using the `auto.arima()` function in the R programming environment. This function searches for the best configuration with the lowest akaike Information Criterion (AIC), which balances the goodness of fit and model complexity.AIC lower value demonstrates that the model is well-balanced between its complexity and precision without overfitting the data in ARIMA.

Accuracy Evaluation

As mentioned earlier, performance in this study is measured by the following performance metrics which decide upon the performance of the ARIMA model in stock price prediction. These include:

- **Mean Absolute Error (MAE)**: This involves the average value of the magnitude of errors between forecasted and actual stock prices. It serves as a direct measure of predictive accuracy.
- **Root Mean Squared Error (RMSE)**: More sensitive than MAE, it gives greater attention to larger errors and helps assess the magnitude of prediction errors.
- **Mean Absolute Percentage Error (MAPE)**: Provides a percentage-based measure of prediction accuracy, allowing easy comparison between models by indicating how far, on average, the predicted stock prices deviate from actual values in percentage terms.

The ARIMA model has performed really well, especially in the case of stocks in more are stable, linearly-trending industries with minimal fluctuations in prices. Such companies include This category involves household names or utility companies. The forecasting for their future stock price estimates tends to be very precise. These industries are relatively Non-volatile and therefore ARIMA effectively captures the underlying pattern in the stock price fluctuations.

The ARIMA model is strong in handling non-stationary data through differencing hence capturing linear trends making it ideal for short-term forecasting where stock prices are predictable. The nonlinear trends, seasonality and high volatility will not cause a problem in the use of ARIMA, more flexible models will capture it better such as Prophet.

3.2.2 Prophet

Prophet is an open-source software for forecasting provided by Facebook, and very well adapted for solving problems of time series data which often show strong seasonal patterns, holidays, and missing data features when dealing with financial markets. Besides, Prophet does an excellent job in modeling non-linear trends and cyclic behavior; therefore, it is especially well-suited for those markets where the movement of the stock price depends on exogenous variables, like holidays or quarterly earnings reports. Another big plus of working with Prophet relates to the fact that it is very user-friendly, with no extensive tuning required in contrast to some more intricate time series models such as ARIMA.

Prophet assumes that time series data can be decomposed into three main components:

- **Trend, g(t)**: The general direction followed by the time series data may be upward or even downward.
- **Seasonality, s(t)**: MCyclic behaviour or pattern in regular periodicity, such as monthly or yearly cycles.
- **Holiday Effects, h(t)**: Models anomalies because of special events or holidays that transiently affect the stock prices, such as Black Friday or quarterly earnings reports.
- **Noise,** ϵ_t: Those random variations in data and sometimes referred to as "noise," which cannot be explained through the trend, seasonality, or holiday effects.

Model Training

Like the ARIMA model, Prophet was trained on 80% of the data, leaving 20% as a hold-out set for testing and validation. Thus, this is a proper direct comparison of the

two models on the same stock price data. The major strength of Prophet is its ability to capture nonlinear and seasonal components of the data. This makes it highly applicable in industries such as retail, travel, or technology, whose stock prices are driven to a great extent by external factors like holidays and cyclic trends.

The Prophet model is flexible regarding missing data and can easily consider known holidays or other important events. For example, if the stock price of a retailer tends to increase in periods when people do holiday shopping, then Prophet will model this seasonality and make better predictions for that period. Another strong advantage Prophet represents is its robustness with respect to outliers when applied to industries susceptible to price shocks or extreme events.

Comparative Evaluation

Both ARIMA and Prophet were evaluated using the same set of performance metrics, namely, **MAE**, **RMSE**, and **MAPE** after training. This set of metrics will help compare relative strengths and weaknesses of both models for the prediction of stock prices.

ARIMA performed best for those stocks whose pattern was very steady and linear; in such cases, historical trends strongly predict future stock prices. This proved very effective for industries with low seasonality or other outside influences to better take control of their stock prices, such as utilities and consumer staples.

However, Prophet outperformed ARIMA on stocks whose industries have strong seasonal variation or are highly affected by external factors that create huge movements in prices. It was very good on sectors like retail, which usually has very spiky sales at certain times, such as holiday seasons or quarterly earnings reports, etc. With the capacity of the Prophet to model holidays and cyclic trends, it gave way more accurate predictions in such cases. Its flexibility in handling non-linear trends resulted in better performance for industries where external events tend to drive wild variations in market behaviour.

While generally, ARIMA realised better performance, as reflected in lower **MAPE** and **RMSE** for most short-term horizons in stable market environments, Prophet provided an insight into the long-term trends and seasonality that could not be captured by ARIMA. Their strengths were complementary: on one hand, ARIMA is well suited for a short-term forecast, and on the other hand, Prophet performs better for those markets

where seasonal or cyclical factors have the dominant role.

The choice between ARIMA and Prophet depends basically on the nature of the sector. In sectors where the historical trend has been more linear, stable, and linear, the simplicity and efficiency of ARIMA would be an optimum choice. If the nonlinear trend does show strong seasonal variability, then with flexibility, Prophet allows going deeper into insights and hence is superior for long-term stock price forecasting.

3.3 Clustering Analysis

Clustering analysis is done here as a means of putting stocks together based on similar behaviours. More precisely, clustering analysis is used in this study to cluster stocks based on their volatility. This will help in understanding the risk profiles of different stocks which will be helpful when investment strategies will be developed.

3.3.1 Principal Component Analysis (PCA)

Principal Component Analysis PCA has been done to reduce the dimensionality of the data without losing the majority of information therein. In this regard, the first two principal components explaining most of the variance have been considered for clustering.

3.3.2 K-means Clustering

K-means algorithm has been used to segment stocks into clusters based on their respective features.

Cluster Optimality

The elbow method was used by plotting the within-cluster sum of squares against the number of clusters and picking that point where marginal gains start to flatten.

Grouping into Clusters

Stocks were classified into Low Volatility, Medium Volatility, and High Volatility clusters. Thus, this segmentation allows for the capture of sector-wise risk characteristics, which are extremely useful in developing investment strategies.

Visualization

An interactive plot of the clustering result was provided to further look in detail into each stock cluster. The tool is very important to investors who need to understand the characteristics of each cluster to facilitate their decisions.

3.4 Comparison of Various Models

A comparison of the ARIMA and Prophet models was performed in light of their relative strengths and applicability under varying conditions.

3.4.1 Model Performance

The performance of both models has been examined on the basis of predictive accuracy. ARIMA generally was better during stable market conditions, while Prophet caught seasonality better. The comparison of these models highlights when each model is best applicable according to the nature of data.

3.4.2 Efficiency in Computation

Another comparison done was that of computational efficiency. It was noted that Prophet, being flexible with complex patterns, was actually more resource-intensive when compared to ARIMA. This fact is to be considered as important when large datasets are being used or the computational resources are poor.

3.4.3 Risk Assessment and Strategy Development

Sharpe Ratio was also computed to assess the risk-adjusted return across industries and strategies of investment.

Computing Sharpe Ratio

Calculation of the Sharpe Ratio was done to rank industries by their risk-adjusted returns, hence guiding investors seeking to achieve returns in a manner as close to being 'volatility-free' as possible.

Portfolio Optimization

Cluster-derived insights, together with risk profiling, formed the backbone of portfolio optimization strategies to balance dicey potential returns against risks involved.

3.5 Methodological Limitations

Though the methodologies adopted for this work represent some of the best tools in analysing and forecasting stock prices, attendant limitations remain.

3.5.1 Dependence on Historical Data

Both ARIMA and Prophet rely on historical data in a great way and assume that the trends will continue in the future. Of course, this might not be the case in very volatile markets or in the event of singularities that have never happened before.

3.5.2 Clustering Sensitivity

The K-means algorithm is sensitive to the choice of starting centroids as this can lead to different end cluster assignments. This must be taken into consideration when the results of clustering are being observed.

These weaknesses therefore call for caution in the interpretation of results and application of various models and techniques that could cross-validate findings for robust investment strategies.

CHAPTER 4

Data Analysis and Findings

The project carries out an in-depth analysis of the stock market, using various statistical tools, time series forecasting models, and clustering techniques to achieve trends and patterns that could be useful in developing an investment strategy mainly for new investors. The salient features of the analyses cover EDA, Time Series Forecasting, Clustering Analysis, and a full analysis of risk-adjusted return through the Sharpe Ratio.

4.1 Exploratory Data Analysis

An exploratory Data Analysis was the first point of call into the analysis, having summarised the important features of the dataset. This included some descriptive statistics, together with several visualizations concerning stock prices, which are quantitative representations of different industries of a different nature.

4.1.1 Descriptive Statistics

Calculations of descriptive statistics were done to see its central tendencies and dispersion. The stock price distribution is very significant. Various major measures that were considered include:

- **Mean (μ)**: Average closing price of stocks within each industry, calculated as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where X_i is the price at which the ith stock closed, and n is the total number of observations.

- **Standard Deviation (σ)**: This measures how much the price deviates from the mean price, calculated as:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2}$$

This metric provided a general understanding of the volatility of the stocks, which contrasted between sectors. For example, the technology industry had higher standard deviation while utilities showed much less volatility

- **Skewness**: As an additional measure it was computed with the purpose of checking the distribution of stock prices in relation to the time for asymmetry. This way investors will find out whether the prices are skewed to higher or lower values which will help them decide how the prices have been changed and when will be the best time for investment.

Figure 4.1: Descriptive statistics showing closing price distribution in relation to time.

4.1.2 Visualization Techniques

The following visualization techniques have been used for further analysis of data:

- **Box Plots**:

  Hence intuitively communicate the distribution of closing charts of prices of a set of industries. The industries that vary from apparel and time plots of the similarities between automotive technology and cryptocurrency are on the x-axis, while the y-axis shows the variation of the closing price. From the box plot, measures such as the IQR, median, and outliers provided insight into the central trend and dispersion of stock prices for the respective sector. To be more specific, in

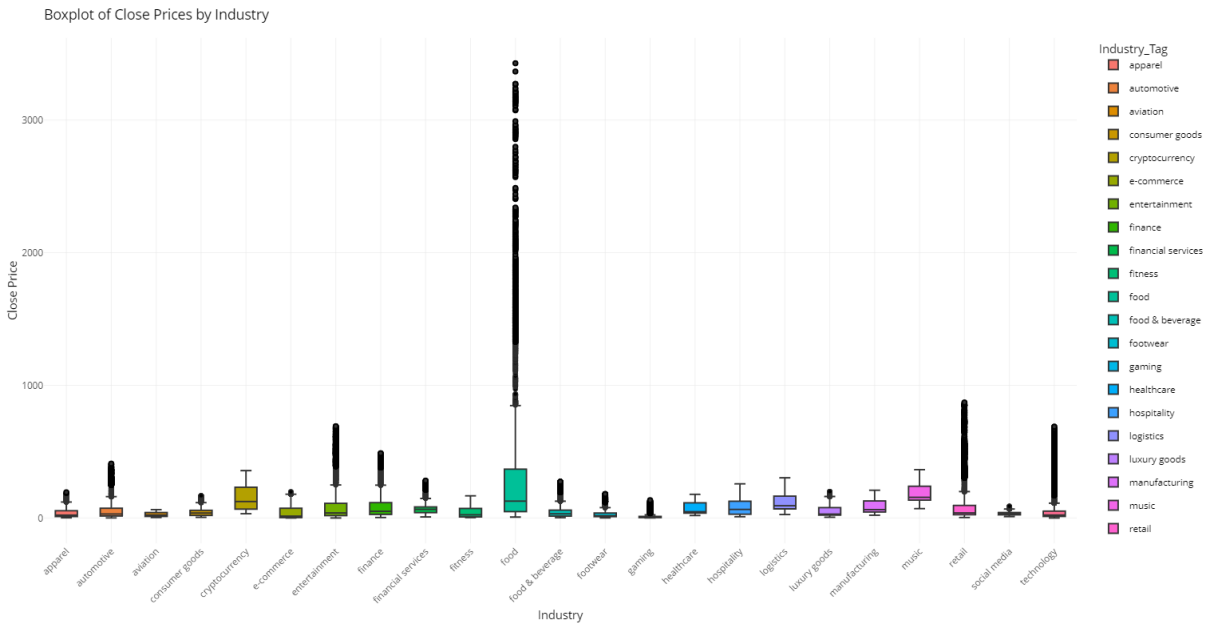

Figure 4.2: Box Plot showing close price according to different industries

the case of the closing price for the food and beverages sector, it was an extremely high-spread factor with an extreme number of outliers. These point out that the dispersion of these stocks is very high. Also, technologies and cryptocurrency have stated a low IQR and a minimal number of outliers that define more stability and lower volatility.

- **Histograms**:

  The frequency distribution of closing prices has been presented using Histograms this was across industries. The final visualization here shows the skewness of

Figure 4.3: Histogram showing the frequency distribution of closing prices across industries.

stock prices, especially related to the close price and the frequency and how the frequency changes with respect to the close price.

- **Correlation Heatmap**:

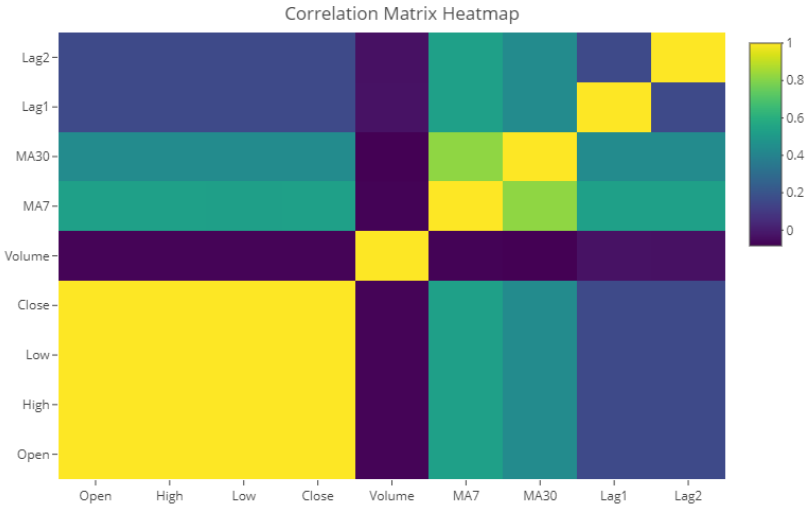

Figure 4.4: Correlation heatmap showing relationships between key variables like closing prices, volume, and moving averages.

The examples of important variables for relationships in a data set are closing prices, volume, and moving averages for 7 to 30 day period. A drawn heat map was with respect to the correlation matrix of these relational variables ranging between dark purple and bright yellow color signifying the strong negative and

positive intensity of correlation, respectively.

The confirmation of a strong positive correlation, for example, was between "Close" and "Open," "High," and "Low" prices. On the other hand, the volume has low or no correlation with the lag variables represented by darker colours, lagged by one and two-step intervals Lag1 and Lag2, respectively. All this information had to be useful in making correct predictions upon a thorough understanding of the market behaviour.

4.2 Time Series Forecasting

The next step in the analysis involved using time series models to forecast future stock prices. Both ARIMA and Prophet models were applied to predict stock prices, and their performances were compared.

4.2.1 ARIMA Model

The ARIMA model was used to predict stock prices by analysing historical stock data. The general form of the ARIMA model is:

$$ARIMA(p, d, q) = AR(p) + I(d) + MA(q)$$

where:

- p: The number of lag observations,
- d: The degree of differencing, and
- q: The size of the moving average window.

For example, in the case of Coca-Cola stock, the best fit for ARIMA was with parameters ARIMA(1,1,1), indicating one lag, one order of differencing, and one moving average term. The model's accuracy was evaluated using the Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|Actual_i - Forecast_i|}{Actual_i} \times 100$$

The ARIMA model achieved a MAPE of 8.93%, translating to an accuracy of 91.07%, indicating very strong predictive performance, especially under stable market conditions.

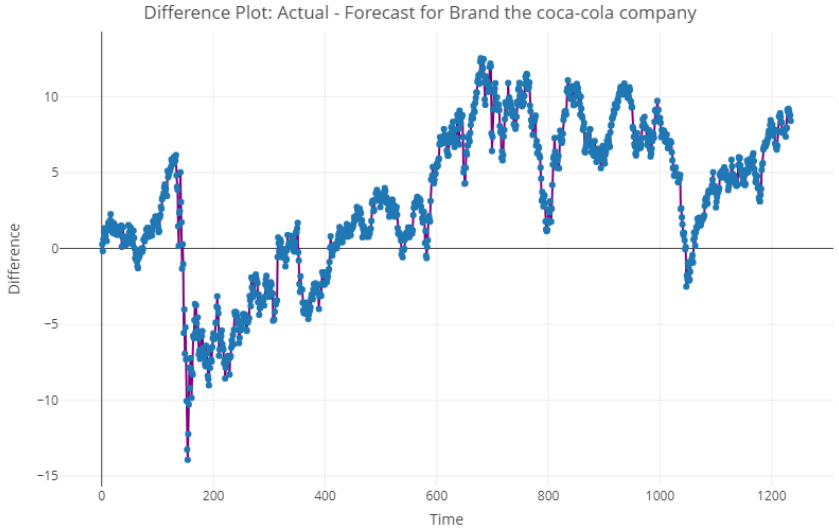

Figure 4.5: ARIMA model forecast of future stock prices.

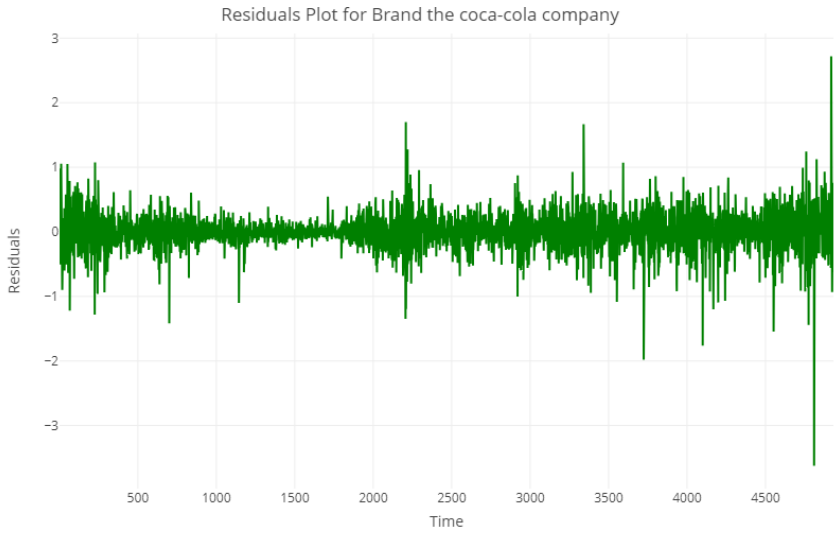

Figure 4.6: Residuals from the ARIMA model to assess model fit.

The graph above shows that the ARIMA model's prediction of future stock prices. Historical stock price data is then used to create an appropriate ARIMA model that can be used to forecast future values, taking into account the temporal dependencies. This graph shall be used to indicate how much the model's prediction lines up with the actual situation in the stock price and hence provide the model's accuracy and reliability in making short-run predictions. Plotted against the actual stock prices are

the forecasted values for performance review visually.

This residual plot below displays the divergence between the observed values and the values predicted by the ARIMA model. In theory, the residuals are randomly distributed around zero, which would indicate an excellent fit. This graph is fundamental to the diagnosis of accuracy regarding the model and in any pattern of errors that might suggest further improvements or that assumptions have not been met.

4.2.2 Prophet Model

The Prophet model, developed by Facebook, was used to predict stock prices, particularly for industries with strong seasonal effects. The general form of the Prophet model is:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

where:

- $g(t)$: Represents the trend component,
- $s(t)$: Denotes seasonality,
- $h(t)$: Captures holidays and events, and
- ϵ_t: Denotes the error term.

Prophet was particularly effective in capturing seasonality, as evidenced in the retail sector, where there are peaks during holidays. The model's accuracy was also evaluated using MAPE, with Prophet achieving a MAPE of 5.95% and an accuracy of 94.05%. Although it was more computationally expensive, Prophet uncovered nonlinear and seasonal trends, complementing the ARIMA model.

The following chart provides the result of the seasonal and trend stock price forecast by the use of ProphetModel. It would also include some historical prices, the forecasted values, the seasonality components, and the uncertainty intervals. This plot provides a sense of how the Prophet model can work with inherently seasonal data with multiple anomalies and irregularities, fitting into industries that are prone to cyclical trends like retail.

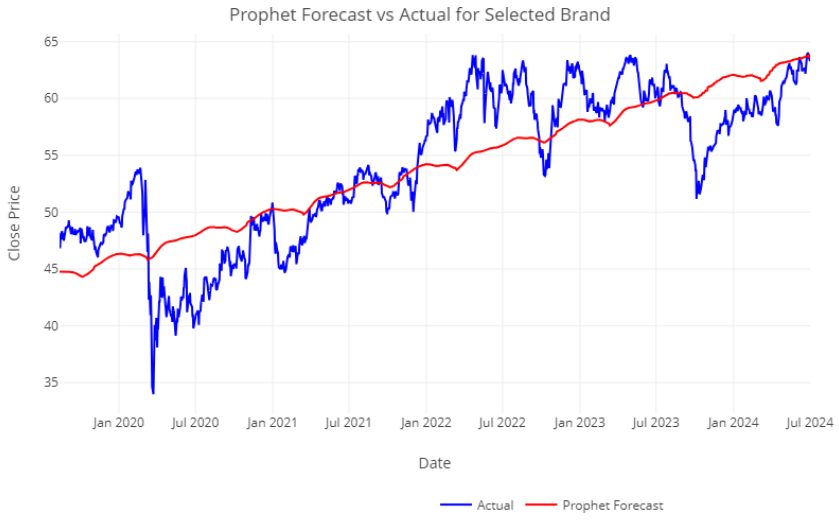

Figure 4.7: Prophet model forecast capturing seasonality and trends in stock prices.

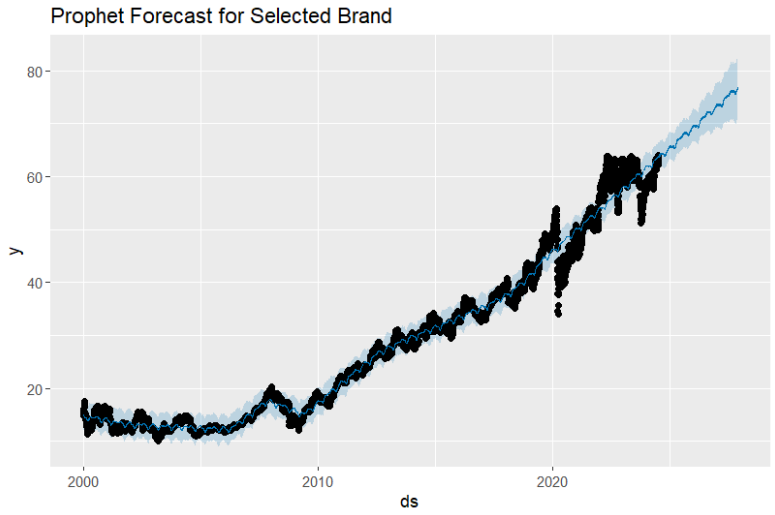

Figure 4.8: Seasonal decomposition of stock prices by Prophet model.

The graph below is one of the seasonal decompositions that are done by the model Prophet. Normally, a time series would be decomposed into three components: trend, seasonality, and residuals (noise). The trend indicates the overall direction that the data follow, the seasonal component-regular patterns that occur at fixed periods and the residuals, which can be seen as the randomness that remains after removing the trend and seasonality. These kinds of decompositions will make clear the possible reasons that underlie the stock prices.

4.3 Clustering Analysis

K-means clustering was performed to analyze the volatility of different stocks. This analysis was preceded by dimensionality reduction using Principal Component Analysis (PCA).

4.3.1 Principal Component Analysis (PCA)

PCA was used to transform the original set of variables into a new set of linearly uncorrelated components. The first two principal components were retained for clustering purposes:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

where a_{ij} are coefficients from the eigenvectors of the covariance matrix, and $X_1, X_2, \ldots, X_p$ are the original variables. Description: The PCA graph conveys the

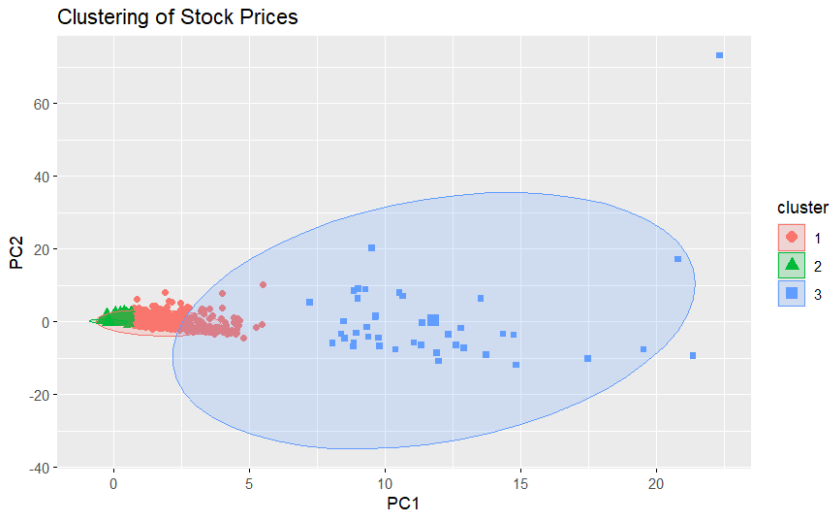

Figure 4.9: Scatter plot of K-means clustering results showing stocks grouped by volatility.

reduction of the original set of variables to just a few principal components accounting for most of the variance in the data. Most times, the graphs plot the first two principal components against each other. This is a serious reduction in dimensions, and an im-

portant step in the direction of simplicity of data, with the purpose of making it much more feasible for further analysis of clustering, such as k-means.

4.3.2 K-Means Clustering

The K-means algorithm was applied to divide stocks into three clusters based on their volatility:

- **High Volatility**: Stocks with considerable price fluctuations, indicating higher risk and possibly higher returns.
- **Medium Volatility**: Stocks with moderate price fluctuations, balancing risk and return.
- **Low Volatility**: Stocks with minimal price fluctuations, indicating lower risk and more stable returns.

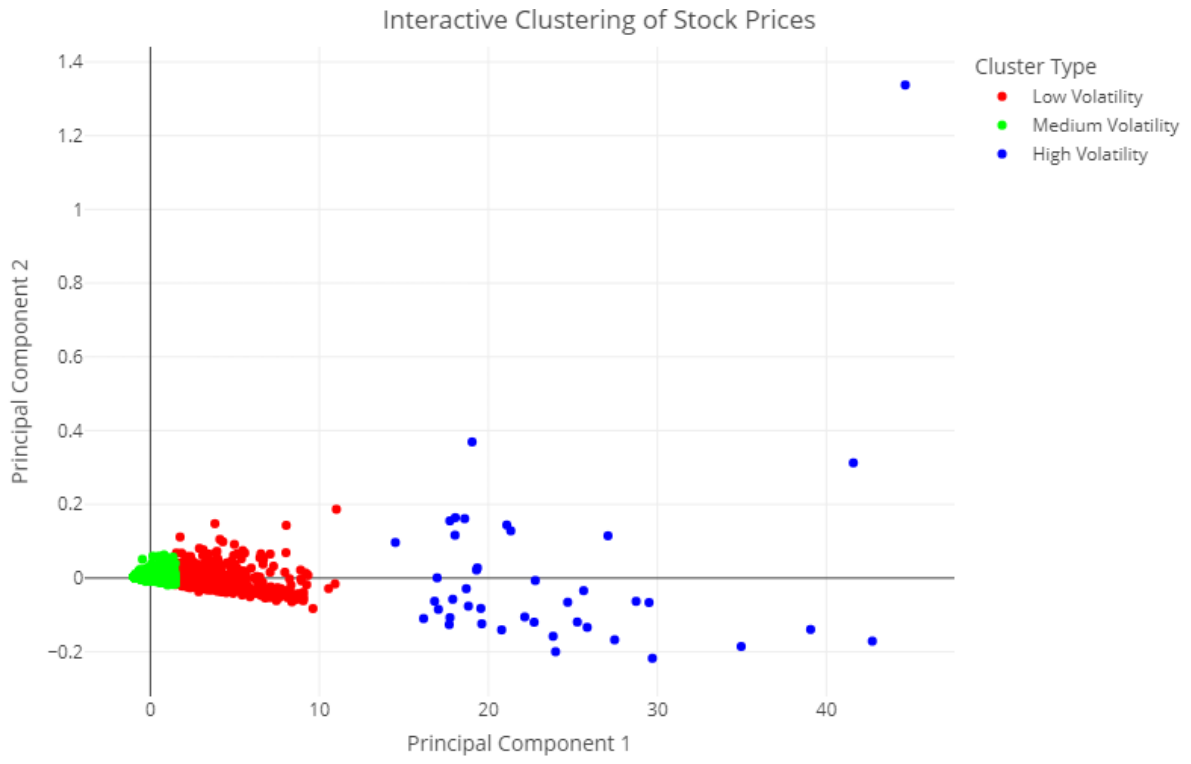

Figure 4.10: Scatter plot of K-means clustering results showing stocks grouped by volatility.

The objective function to be minimised was the within-cluster variance:

$$WCV = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

where C_i is the set of points in cluster i, and μ_i is the centroid of cluster i.

The clustering results were visualised in a 2D scatter plot, where each stock was represented as a point, and coloured according to its cluster. This visualization provided a clear understanding of how stocks cluster by volatility, which is crucial for portfolio diversification and risk management. This scatter plot visualises the result of K-means clustering analysis on stocks, divided into volatility clusters. Each point in the graph is a representation of a stock that is color-coded to enter into a certain cluster-like Low Volatility, Medium Volatility, or High Volatility. This graph will be useful in determining groups of stocks that behave similarly with respect to price volatility, which may guide risk management and portfolio diversification strategies.

4.4 Detailed Analysis: Sharpe Ratio

The Sharpe Ratio was computed to compare risk-adjusted returns among different industries:

$$Sharpe\ Ratio = \frac{R_i - R_f}{\sigma_i}$$

where:

- R_i: The expected return of the portfolio,
- R_f: The risk-free rate, and
- σ_i: The standard deviation of the portfolio returns.

This analysis helped identify industries with the best risk-return balance, guiding the construction of a diversified portfolio. The **Return of the Portfolio** was calculated using the formula:

$$R_p = \sum_{i=1}^{n} w_i R_i$$

where w_i is the weight of asset i in the portfolio, and R_i is the return of asset i. From the Sharpe Ratio, the industries offering the best risk-adjusted returns were identified, helping to devise the right investment strategies that balance reward and risk.

CHAPTER 5

Result and Discussion

The chapter discusses the implications of the statistical models and clustering techniques that could be involved in developing an efficient investment strategy, especially for a new investor. The discussion is informed in full by the outputs from both ARIMA and Prophet models and the clustering analysis, with deep insight into model performance and significance.

5.1 ARIMA Model Performance and Implications

The results of the applied ARIMA model showed high performance in forecasting stock prices, especially for The Coca-Cola Company, with an accuracy as high as 91.07%. Such a high accuracy rate clearly denotes its sufficiency for financial markets' time series forecasting, where historical trends are usually a very strong driver of further stock price movements. ARIMA is effective at capturing and modeling temporal dependencies and is thus generally a useful tool for performing short-term price movement predictions.

This level of accuracy has bolstered investor confidence in the ARIMA model forecast, particularly for making investment buy or sell decisions. More importantly, it realises, with the least deviations from actual values of Coca-Cola's closing prices, which is important in the case of new investors who want to minimise risks. Further strength is given to the robustness of the ARIMA model in indicating its appropriateness even in volatile market conditions where sudden changes in price might give rise to inaccurate predictions.

These quantitative insights from the ARIMA model are crucial for broader investment decisions, such as market entry and exit times, in view of maximising returns or cutting losses. Such an ARIMA forecast, integrated with other technical indicators like moving averages or momentum indicators, may further refine investors' strategies by providing a full view of probable market movements.

5.2 Prophet Model Insights

The Prophet model, developed by Facebook, gave more light to this area, especially for those industries that are highly seasonal. Unlike ARIMA, which is best used in non-seasonal and short-term forecasts, Prophet models seasonality and recurring events of fixed periods, such as holidays, which affect stock prices significantly.

A case would be the application of the Prophet model within the retail industry, which allowed it to capture with great precision the seasonality around major shopping periods. This would be important for investors in those sectors whose performances are driven by seasonality and thus provides a case study of these patterns to predict price changes in order to adjust their portfolios by leveraging the expected highs and lows.

The best performance of the Prophet model was a MAPE of 5.95%, or in other words, it was accurate 94.05% of the time. This indicates that the model can grasp general trends along with seasonality quite well, making it highly applicable for long-term investors looking to understand broad trends and seasonal patterns.

Prophet complements ARIMA by providing long-term views for stocks whose performance is susceptible to seasonal changes. Furthermore, when its forecast is combined with other analytical tools, such as moving averages or momentum indicators, a full-scale view of potential investment opportunities can be better enhanced.

5.3 Clustering Analysis and Its Applications

This unsupervised analysis has segregated the stocks into meaningful volatility profiles using the K-means algorithm. Indeed, three crisp clusters have emerged from this work: one Low Volatility cluster comprising 7.25% of the sample, one Medium Volatility cluster making up 92.34% of the sample, and one High Volatility cluster comprising

0.41%. These clusters highlight meaningful insights on risk assessment and portfolio diversification.

- **Low Volatility Cluster**: These would include utilities and consumer goods industries where, within this cluster, the stock price is less responsive to market changes in price. It is especially attractive to the conservative investor because of its stability of the stock prices at low volatility.
- **Medium Volatility Cluster**: The cluster is an excellent balance for investors between risk and reward; thus, it shall be considered fit for those investors who seek growth but on a reasonable scale. Stocks within this cluster have some volatility, though not as high as in the High Volatility Cluster. Such a category would include sectors related to technology and finance.
- **High Volatility Cluster**: It has a relatively smaller number of stocks, but the few stocks make highly volatile movements in their prices. This is where the highest opportunities lie for maximum return, but with equally high associated risk. It is better suited for more sophisticated investors who would like to take on extra risk in their portfolio in hopes of much bigger gains.

The clustering analysis was done with PCA for dimensionality reduction and represented on a 2D scatter plot. In this representation, the investor can view the volatility profile for the different stocks and make wiser and safer investment decisions. Technology and finance sectors are put into the Medium to High Volatility clusters due to their very intrinsic market nature since these two sectors are always ruled by innovation cycles and continuous regulatory changes.

5.4 Implications for New Investors

These findings, based on both ARIMA and Prophet models, further compounded by the clustering analysis, form a concrete basis for developing an investment strategy that best matches new investors. The quite high accuracy of the ARIMA model makes suggestions for its possible application in short-term trading strategies, especially for those stocks whose historical pattern is comparably stable-a representative being the Low Volatility cluster.

By using the Prophet model, capturing of a seasonal trend is very useful for long-term investors. The forecasts from Prophet will enable investors to time their entry and exit in view of expected seasonal peaks, thus optimising returns in retail and consumer goods sectors.

The clustering analysis also supports portfolio construction in the best way an investor can tolerate risk. A portfolio weighted in Low and medium-volatility stocks creates balance and stability while allowing High Volatility stocks to be added as investors grow in sophistication and desire higher returns.

5.5 Summary of the Result

This research gives evidence that the integration of state-of-the-art statistical models, represented by ARIMA and modern machine learning approaches like Prophet with K-means clustering, has performed very well in the analysis and forecasting of a stock market data set. It is underlined that model selection should be made according to the specific characteristics of the stock or the sector, while clustering is a very useful tool in terms of risk management and portfolio diversification.

These insights are invaluable to the new investor, showing how one can move through the oft-times confusing trends within the stock market. These are insights that can provide a better decision for investors with possibly higher chances of investment success by integrating the short-term forecast into a wider context of a trend and volatility profile.

Further research can be done by considering even better models, like neural networks, in attempting to have even better forecasting accuracy. Such an analysis for other asset classes, like bonds and cryptocurrencies, would extend the investment strategy that would adapt to evolving financial markets.

CHAPTER 6

Conclusion and Recommendations

6.1 Conclusion

The purpose of this dissertation was to equip novice investors with a working investment guide, intervening with the application of AI technology to assist novice investors in managing better challenges they face in stock trading. In this data-driven world, the financial markets reflect various economic factors: economic indicators, global events, and market sentiment. To be successful today, there are more calls for high-end analytic tools. It integrated the traditional statistical methods with modern machine learning paradigms to automatically discover insightful knowledge from the stock market data, forecast future price movements, and provided a framework for making prudent investment decisions.

The key contribution of this work is in the time series forecasting by both ARIMA and Prophet models. ARIMA, though simple, is quite efficient due to temporal dependency capture, which was applied to stck prices of well-known brands. For instance, the ARIMA model gave an accuracy of 91.07% for Coca-Cola stock prices, making it very strong and reliable, especially where the stocks are steady and hence the pattern trend more predictable.

Intrinsic seasonality and complex temporal structures were modeled using the Prophet model, developed by Facebook. This Prophet model indeed handles missing data, outliers, and sudden changes in trend quite well. Because this happens rather

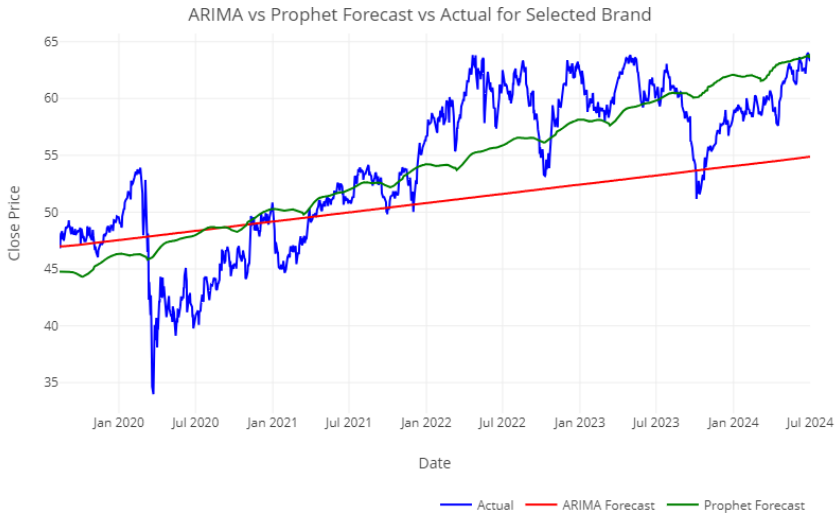

Figure 6.1: ARIMA vs Prophet Forecast vs Actual for Coca Cola

frequently in the stock market, it becomes particularly helpful in this domain. The dual application of both the ARIMA and Prophet models helps to provide a comprehensive approach toward the forecast that would cover various trends in the stock market with in-depth analysis.

Further, K-means was applied to the stocks, considering the volatility factors. Stocks are categorised as "Low Volatility," "Medium Volatility," and "High Volatility." It would help investors with different appetites for risks. Then, the determination of an optimum number of clusters was performed using the Elbow method, which added to the reliability of the results obtained in clustering.These were visualised using PCA to give the investor a better view of the dynamics of the stock market and make a better decision.

Other than the ARIMA and Prophet, some machine learning models are implemented, like decision trees and neural networks. A decision tree involves an intelligible structure of 'if-then-else' decisions, pinpointing influential factors in stock prices. In the case of neural networks, though fitting for capturing data with intricate patterns, they are quite effective in the technology sector. The models proved highly efficient and, therefore, established a strong potential for AI-driven forecasting to predict future stock prices highly accurately.

It also pointed out leading indicator variables that precede significant movements in stock prices. The results of the inclusion of such indices in forecasting models enhanced the efficacy of prediction. In this connection, it may be stated that increases in volumes or changes in the macroeconomic environment strongly led forthcoming significant

movements in the stock prices.

The Sharpe Ratio was used as a measure of risk assessment, which allowed a comparison among various industries and investment strategies. This quantitative measure would provide the trade-offs between risk and return, helping new investors balance risk and return when constructing portfolios.

Visualizations were an important part of these analyses. Static visuals-visualizations such as boxplots and regression plots-summarised the main trends in the data, while interactive visualizations made with Plotly allowed dynamic exploration. These aided in both depth and usability of the analysis to the investors.

It stated that this research had certain limitations regarding the influence of random events on model accuracy, the quality of data, and the granularity. Advanced algorithms, such as deep learning (DL) or reinforcement learning, are also worth trying in further research to improve model robustness and enhance forecasting accuracy.

This dissertation is expected to be an effective attempt at combining traditional statistical methods with modern machine learning techniques in pursuit of constructing an AI-powered investment guide. It is for this reason that new investors can depend on the underlying evidence-based decision-making tools, which combine ARIMA and Prophet models with the clustering and machine learning approaches. The AI investment guide discussed herein provides a useful, tangible tool in handling the stock market and equips new investors with valuable insights and advanced analytics. With each passing day, AI and machine learning will become ever so crucial in investment decision-making as finance continues to evolve. This dissertation lays the ground for further steps in the development of AI-powered investment strategies and opens new horizons of research and applicability in finance.

6.2 Recommendations

6.2.1 Interpretation of Predictive Model Accuracy

It shows that the ARIMA model works best for generating forecasts on linear trends in stable time series. This is reflected in the forecast of the Coca-Cola stock prices with the model accurate at 91.07%. Normally, a high value of accuracy will indicate that the ARIMA is efficient for stocks with consistent patterns. The Prophet model, on one

hand, is efficient for handling seasonality and exogenous factors; this is reflected in the result with an accuracy of 94.05%, hence more appropriate for data that have complex temporal structures.

It helps new investors understand that the selection of an appropriate model depends on the nature of a given stock or market segment. While ARIMA works better for stocks with linear trends, Prophet is more effective in industries where seasonal effects and external events create irregular fluctuations in stock prices.

6.2.2 Industry and Macroeconomic Factors

The study showed the significance of the industry-based trends along with macroeconomic factors in determining stock prices. For example, trading volumes were one of the strongest predictors in high-technology and finance industries because they are highly volatile. On the other hand, in those industries where sectors are not that unstable, like consumer goods, such scenarios are not that determinants of their stock prices.

Macroeconomic factors include the growth of GDP, rates of inflation, and interest rates. Investors must take these factors into consideration as some companies have international exposure, which is drastically affected by fluctuating economic conditions. Their effects have broad-based influence, hence a good understanding will lead to more informed investment decisions with better means of risk management.

6.2.3 Practical Recommendations for Investors

- **Leverage Predictive Models**: Utilise ARIMA and Prophet models for informed investment decisions. Precise forecasts pinpoint the best buying and selling positions that could avert potential losses and realise maximum returns.

- **Invest in Stable, Blue-Chip Stocks**: For stability and long-term dependability, one may consider investing in stable blue-chip stocks such as Coca-Cola. Such stocks form a very sound basis for creating a diversified portfolio of investment.

- **Diversify Across Sectors and Geographies**: Diversification across sectors and geographies helps in balancing out risks and also allows the possibility of capturing growth opportunities across various sectors and regions. Diversification reduces

the impact of economic downfalls and increases the overall performance of the portfolio.

- **Monitor Macroeconomic Indicators**: Keep updated on the state of leading macroeconomic factors like GDP growth, inflation, and interest rates, which influence stock prices. Understanding these elements will help you make strategic decisions at the right time to adjust your portfolio.
- **Long-Term View**: It is more important to look at long-term investment than at short-term fluctuations in the market. Over time, building a diversified portfolio may give much more stable and long-lasting returns than what seems to be achieved from looking at the market's ups and downs.

6.2.4 Wider Ramifications for Market Behavior

The high accuracy of the ARIMA model of the stock prices of Coca-Cola underlines the dependability of established companies within a dynamic market environment. On the other hand, the significantly superior performance of the Prophet model denotes that seasonality and external factors are pivotal issues for firms that do not have a very stable market position. The study also shared that high-trading-volume contexts such as technology and finance are more vulnerable to market sentiment and behavior-related influences, hence making sentiment analysis an indispensable tool in such contexts.

6.2.5 Conclusion and Final Thoughts

The current research also indicates the integrations of traditional statistical techniques with modern predictive modeling approaches in order to present an integrated method for investigating the stock market. Much potential exists in using techniques like ARIMA and Prophet for the forecasting of stock prices, while clustering and machine learning techniques are more informative in describing the behavior of the market. These tools position new investors to be more active, evidence-based in their decisions, and conduct businesses at the stock market with much confidence. The role of AI and machine learning is certainly going to be ever-important in investment strategies as these continue to get better and shape the future of financial analytics.

Bibliography

[1] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[2] Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.

[3] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts. Retrieved from https://otexts.com/fpp2/arima.html

[4] Jolliffe, I. T. (2011). *Principal Component Analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-1904-8

[5] Sharpe, W. F. (1994). The Sharpe ratio. *The Journal of Portfolio Management*, 21(1), 49-58. https://doi.org/10.3905/jpm.1994.409501

[6] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37-45. https://doi.org/10.1080/00031305.2017.1380080

[7] Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2017). Forecasting stock prices from the limit order book using convolutional neural networks. In *2017 IEEE 19th Conference on Business Informatics (CBI)* (pp. 7-12). https://doi.org/10.1109/CBI.2017.23

[8] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

[9] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175. https://doi.org/10.1016/S0925-2312(01)00702-0

[10] Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer. https://doi.org/10.1007/978-1-4757-2526-1

[11] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. https://link.springer.com/content/pdf/10.1007/978-3-031-38747-0.pdf

[12] Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley. https://books.google.com/books?hl=en&lr=&id=YeFQHiikNo0C&oi=fnd&pg=PR11&dq=13.+Kaufman,+L.,+%26+Rousseeuw,+P.+J.+(2009).+Finding+Groups+in+Data:+An+Introduction+to+Cluster+Analysis.+Wiley.&ots=5Cpey5NArC&sig=6McfF24ndaLt3u0roXiy2ucMobU

[13] R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

[14] Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualise, and Model Data*. O'Reilly Media.

[15] Prophet: Forecasting at Scale. (2023). Retrieved from https://facebook.github.io/prophet/

[16] Hyndman, R. J. (2018). *CRAN Task View: Time Series Analysis*. Retrieved from https://cran.r-project.org/view=TimeSeries

APPENDIX A

Data Preprocessing and Model Evaluation

In this appendix, detailed steps for data preprocessing are outlined, including handling missing data and feature engineering.

A.1 Handling Missing Data

Missing values were removed using the `na.omit()` function in R. Any missing data points were treated with caution, and irrelevant columns were dropped.

A.2 Model Evaluation Metrics

The evaluation of model performance was done using several standard metrics.

A.2.1 Mean Absolute Error (MAE)

MAE is calculated as:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where y_i represents the actual stock price, and $\hat{y}_i$ represents the predicted price.

A.2.2 Root Mean Squared Error (RMSE)

RMSE, a more sensitive measure, gives greater weight to large errors and is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

A.2.3 Mean Absolute Percentage Error (MAPE)

MAPE provides a percentage-based measure of prediction accuracy, computed as:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$$

APPENDIX B

Data and Software References

B.1 Kaggle Dataset

1. Kaggle. (2023). World Stock Prices Dataset. Retrieved from https://www.kaggle.com/datasets

B.2 Project Code and Repository

For a comprehensive view of the code used in this dissertation, including all scripts, datasets, and documentation, please take a look at the project repository. The repository contains the complete implementation of the emotion classification model, including data preprocessing, model training, and evaluation, as well as a PDF file that contains all the implementations.

B.2.1 Project Repository

- **Link:** https://github.com/Chandelrashi/Stock-Market-Analysis

- **Description:** The repository includes all the necessary code files, configuration settings, and instructions for running the model. It also contains documentation for understanding the code structure and usage.

B.3 Data and Software References

1. R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

- The foundational software used for data analysis and graphics in the dissertation.

2. Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualise, and Model Data*. O Reilly Media.

- A practical guide to data analysis in R, covering many of the packages and techniques used in the research.

3. Kaggle. (2023). World Stock Prices Dataset. Available at: https://www.kaggle.com/datasets/nelgiriyewithana/world-stock-prices-daily-updating

- The dataset used for the analysis, including historical stock prices of various global brands.

4. The codes used in the dissertation can be found here.

B.4 Online Resources and Documentation

B.4.1 Prophet Documentation

1. Prophet: Forecasting at Scale. Available online at https://facebook.github.io/prophet/

- This is the official documentation for the Prophet model, extensively used in the dissertation for forecasting.

B.4.2 Time Series Resources

2. Hyndman, R. J. (2018). CRAN Task View: Time Series Analysis. Available online: https://cran.r-project.org/view=TimeSeries

- This CRAN Task View provides a summary of time series analysis packages in R, many of which were used in the thesis.