School of Computer Science and Electronic Engineering

MSc Data Science

Academic Year 2023-2024

Advancing Stock Price Forecasting with LSTM Networks

A project report submitted by: Wilson Thomas Ayyappan - 6835716

A project supervised by: Prof Ferrante Neri

A report submitted in partial fulfilment of the requirement for the degree of Master of Science

University of Surrey
School of Computer Science and Electronic Engineering
Guildford, Surrey GU2 7XH
United Kingdom.
Tel: +44 (0)1483 300800

ABSTRACT

This research is focused on enhancing the accuracy and real-world application of stock price forecasting models for major financial institutions like MasterCard, Visa, American Express, and Capital One, over a 16-year span from 2008 to 2024. Accurate stock price predictions are vital for investors and analysts, who rely on these models to make well-informed decisions, especially in the unpredictable financial markets we see today. This study evaluates how Long Short-Term Memory (LSTM) networks—a type of deep learning model—stack up against more traditional forecasting methods like ARIMA, Prophet, Random Forest, and XGBoost. LSTM networks were chosen for their unique capability to detect complex patterns and time-dependent relationships in stock data that traditional models might miss.

The models were tested using historical stock data, going through various pre-processing steps to ensure accuracy, and evaluated using key performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2. Additionally, the impact of broader economic factors like GDP, inflation, and interest rates on stock prices was incorporated into the analysis. The results show that LSTM models consistently outshine the other methods, proving to be a highly effective tool for predicting stock prices accurately. Overall, the study highlights the significant potential of LSTM networks to improve financial forecasting and offer deeper insights, ultimately helping decision-makers in the financial sector make more informed choices.

***Keywords**: Stock Price Forecasting, LSTM Networks, ARIMA, Prophet, Random Forest Regressor, XGBoost, Macroeconomic Indicators, Financial Market Analysis, Investment Decision-Making, Volatile Markets.*

HIGHLIGHTS

- *Developed and evaluated LSTM models to achieve superior stock price forecasting accuracy.*
- *Conducted a comparative analysis of LSTM performance against ARIMA, Prophet, Random Forest, and XGBOOST models*
- *Integrated key macroeconomic indicators such as GPD, inflation, unemployment, and interest rates to enhance predictive modelling of financial markets.*
- *Demonstrated LSTM's ability to capture complex, time-dependent patterns in stock price data that traditional models might miss.*
- *Provided actionable insights for informed investment decision-making, particularly in volatile financial markets.*
- *Emphasized the potential of deep learning models, like LSTM, in revolutionizing financial forecasting applications.*

ACKNOWLEDGEMENTS

I wish to extend my deepest gratitude to Professor Ferrante Neri, whose guidance, expertise, and unwavering support have been invaluable throughout the course of this research. His insightful advice and constructive feedback have significantly contributed to the depth and quality of this study, and I am truly fortunate to have had the opportunity to learn under his mentorship. I would also like to express my sincere thanks to all my professors, whose lectures, encouragement, and academic rigor have greatly shaped my understanding and approach to this research. Their collective wisdom and commitment to teaching have inspired me to strive for excellence in my work. I am also indebted to my colleagues and peers, whose stimulating discussions and thoughtful suggestions have enriched this research. Lastly, I would like to express my heartfelt appreciation to my family and friends.

To all who have supported and contributed to this work, I am sincerely grateful.

I certify that the work presented in the dissertation is my own unless referenced.

Signature:

Date: 16/09/2024

TOTAL NUMBER OF WORDS: 13559

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

1. **LSTM**: Long Short-Term Memory
2. **ARIMA**: Autoregressive Integrated Moving Average
3. **MSE**: Mean Squared Error
4. **MAE**: Mean Absolute Error
5. **RMSE**: Root Mean Squared Error
6. **R^2**: R-squared
7. **GDP**: Gross Domestic Product
8. **CRISP-DM**: Cross-Industry Standard Process for Data Mining
9. **API**: Application Programming Interface
10. **RSI**: Relative Strength Index
11. **OLS**: Ordinary Least Squares
12. **FRED**: Federal Reserve Economic Data
13. **P/E**: Price-to-Earnings (Ratio)
14. **SVM**: Support Vector Machine
15. **EDA**: Exploratory Data Analysis
16. **COVID-19**: Coronavirus Disease 2019
17. **GPU**: Graphics Processing Unit

CHAPTER 1: INTRODUCTION

This dissertation aims to enhance the precision of stock forecasting prediction models for major financial institutions, with a particular focus on MasterCard, Visa, American Express, and Capital One. Stock price forecasting is an essential tool for investors and financial analysts, especially in today's volatile market conditions. By providing more precise predictions, this research aims to enhance decision-making process. To achieve this, the effectiveness of LSTM networks, a deep learning model, is compared against traditional forecasting techniques such as ARIMA, Prophet, Random Forest, and XGBoost. LSTM networks were selected for their capacity to recognize intricate temporal patterns and sequences within financial data, which are often overlooked by traditional models. Through the examination of past stock market data and the incorporation of macroeconomic factors, this study seeks to deliver more dependable and practical predictions which can help guide more informed decision-making within the financial industry. These findings expand the understanding of how various forecasting models can be utilized in real-world financial contexts to achieve greater precision and insight.

1.1 Background

1.1.1 Context

This research lies at the intersection of financial forecasting and machine learning, a rapidly growing area within data science. Accurate stock price prediction is critical for stakeholders in the financial industry, as it directly influences investment strategies and risk management decisions. By improving these predictive models, this study contributes to the broader efforts of integrating advanced machine learning techniques into financial systems, thereby enhancing their robustness, reliability, and capacity for handling real-world volatility and uncertainty.

1.1.2 Scope

Within this context, the study specifically addresses the comparative performance of LSTM networks against traditional forecasting methods for stock price prediction. The research evaluates these models based on their predictive accuracy and practical applicability over a 16-year period. This time frame enables a thorough analysis, covering both periods of market stability and significant volatility, thus providing comprehensive insights into the models' effectiveness in various financial conditions.

1.1.3 Problem Statement

Although models such as ARIMA are commonly applied in financial forecasting, they frequently fail to account for the non-linear and dynamic characteristics of stock markets. Various elements, such as market sentiment, economic indicators, and external factors, impact stock prices, which can introduce complexities that simpler statistical models struggle to capture. LSTM networks provide an effective solution for processing sequential data and modelling long-term relationships, which makes them highly effective for financial forecasting tasks. Nevertheless, their use in this domain is still relatively limited, with a need for more extensive empirical studies to fully validate their potential. This study fills this gap by comparing LSTM networks with traditional forecasting methods like ARIMA, Prophet, Random Forest, and XGBoost. The aim is to determine whether LSTM networks can predict stock prices more accurately and, by extension, enhance decision-making in the financial sector. By challenging the status quo in financial forecasting, this research advocates for the broader adoption of deep learning techniques, which may offer more robust and adaptive models in increasingly complex and volatile financial markets.

1.1.4 Research Motivation

This research is driven by the increasing demand to improve the accuracy of financial predictions, which are crucial for informed investment decisions. As financial markets become more complex, traditional models may struggle to provide reliable predictions. LSTM networks offer a promising alternative, but their effectiveness requires thorough validation. This study aims to validate LSTM models while also providing a framework for their practical implementation in real-world financial forecasting.

1.1.5 Justification

The importance of this study is in its potential to connect traditional and contemporary forecasting techniques, offering actionable insights for direct application in the financial industry. Enhancing the predictive accuracy of these models, this work adds value to both scholarly work in data science as well as offering practical solutions for financial modelling.

1.2 Research aim and objectives

Aim: To design and assess a deep learning model, focusing on LSTM, for precise stock price predictions in leading financial institutions.

Objectives:

- To review the literature on financial forecasting models, focusing on traditional and machine learning approaches.
- To identify and select a suitable data science methodology for conducting this research.
- To collect and pre-process historical stock data and macroeconomic indicators for MasterCard, Visa, American Express, and Capital One (2008-2024).
- To develop and train LSTM models and compare their performance with traditional forecasting models.
- To examine the effects of significant events mainly the 2008 financial crisis, COVID-19, and crucial macroeconomic factors on stock market prices.
- To assess the effectiveness of each model using established performance metrics and determine the optimal forecasting approach.
- To provide insights and recommendations for the application of LSTM models in financial forecasting.

1.3 Research approach

This research adopts a mixed-method approach to achieve its objectives. It begins with a comprehensive literature review, exploring traditional forecasting models like ARIMA alongside advanced machine learning models, particularly LSTM networks. Historical stock prices for MasterCard, Visa, American Express, and Capital One, as well as macroeconomic key indicators, including interest rates, inflation, and GDP, are gathered and prepared for model training. LSTM models are then developed and compared with traditional methods and other models, like Random Forest and XGBoost. The models' performance is assessed using metrics like MSE, MAE, RMSE, and R^2, with special attention to the impact of macroeconomic indicators and Crucial events, mainly the 2008 financial crisis and the COVID-19 pandemic. Additionally, the study analyses the influence of different market strategies on stock prices. Ethical considerations centre on the responsible use of publicly available, anonymized data. This method allows for a comprehensive assessment of forecasting models, offering crucial observations into the use of advanced machine learning techniques in financial predictions.

1.4 Dissertation outline

The structure of the rest of this dissertation is outlined as follows:

Chapter 2: Literature Review

This chapter presents an analysis of the available literature on financial forecasting, with a focus on both traditional models like ARIMA and more advanced approaches such as LSTM networks. It further explores the influence of macroeconomic indicators on stock price forecasting and identifies areas where research is lacking.

Chapter 3: Research Approach

This chapter outlines the research methodology, covering the data collection procedures, model construction, and the metrics used for performance evaluation. It explains the rationale behind model selection and discusses ethical considerations related to data usage.

Chapter 4: Data Analysis

This chapter presents the analysis results, evaluating how different forecasting models performed on stock prices for companies such as Mastercard, Visa, American Express, and Capital One. Additionally, it evaluates the influence of macroeconomic factors and major events, mainly the 2008 financial crisis and the COVID-19 pandemic.

Chapter 5: Discussion

The findings are discussed in relation to the research questions, comparing the effectiveness of LSTM networks with traditional models. The chapter also explores how market strategies and macroeconomic indicators have influenced stock prices.

Chapter 6: Conclusion

The concluding chapter outlines the key results, highlights contributions to financial forecasting, and offers recommendations for further research.

CHAPTER 2: LITERATURE REVIEW

This chapter delves into current research and approaches in financial forecasting, with a focus on traditional models such as ARIMA and advanced machine learning methods like LSTM networks. The chapter is structured to introduce foundational concepts and theories, followed by a review of key studies and advancements in the field. It also examines the role of macroeconomic indicators in stock price prediction and identifies research gaps that this dissertation seeks to address. Finally, it sets the stage for the methodology and analysis in subsequent chapters.

2.1 The Importance of Financial Forecasting in the Modern Economy

Financial forecasting is indispensable in the modern economy, providing critical insights that guide businesses, investors, and policymakers in making informed decisions. Accurate stock price forecasting is vital for developing robust investment strategies, enhancing risk management, and contributing to economic stability (Fama, 1970). The ability to predict future stock prices enables investors to optimize their portfolios, manage risk more effectively, and make informed decisions on asset allocation. For businesses, forecasting stock prices can influence strategic decisions such as mergers and acquisitions, capital expenditures, and resource allocation. Moreover, accurate financial forecasting is crucial for policymakers who rely on these predictions to craft economic policies that can stabilize markets and stimulate economic growth.

However, the dynamic and often unpredictable nature of financial markets poses significant challenges to forecasting accuracy. Stock price prediction is complicated by market volatility, which is influenced by economic policy shifts, geopolitical events, and market sentiment. Traditional statistical methods, such as those discussed by Box and Jenkins (1976), have long been employed in financial forecasting, but their limitations, particularly in handling complex, non-linear, and volatile data, have become apparent over time. Conventional approaches face challenges in capturing the complex interconnections and non-linear behaviours inherent in financial market, often resulting in less precise predictions. he adoption of machine learning, particularly deep learning models such as LSTM networks, marks a significant progression in this domain, opening up opportunities for more precise and dependable financial forecasting.

2.2 Traditional Approaches to Financial Forecasting

2.2.1 Overview of Traditional Methods

Conventional financial forecasting techniques, including Moving Averages, Exponential Smoothing, and ARIMA models, have long been essential in stock price prediction (Makridakis & Hibon, 2000). These techniques are commonly employed because of their straightforwardness, ease of implementation, and capacity to deliver relatively accurate predictions in stable market environments. Moving Averages and Exponential Smoothing, for instance, are particularly useful for identifying trends in historical data and projecting them into the future. However, these approaches are mostly linear and face challenges in capturing the intricate, non-linear dynamics of financial markets, which are shaped by a range of factors such as economic indicators, investor sentiment, and unforeseen events.

ARIMA models are widely utilized for their effectiveness in managing time series data with a reasonable level of precision. The ARIMA model combines three core components: autoregression (AR), differencing (I), and moving average (MA). The autoregressive component reflects the connection between the present data points and previous observations, the differencing step promotes stationarity by removing trends and seasonal variations, while the moving average component models the link between a current data point and the residual error obtained from applying a moving average to prior data points. However, as Tsay (2005) noted, these techniques, being largely linear, find it difficult to represent the complex, non-linear interactions present in financial markets. This limitation is a significant drawback, especially when forecasting in markets characterized by volatility and rapid changes. Moreover, ARIMA models assume of stationarity, a condition frequently unmet in real-world financial data, necessitating extra pre-processing steps that may introduce errors and compromise model accuracy.

2.2.2 ARIMA (Autoregressive Integrated Moving Average) Models

Introduced by Box and Jenkins (1976), the ARIMA model has been a cornerstone of time series analysis, particularly in forecasting stock prices. The model's effectiveness lies in its capacity to model the dependencies between an observation and a few lagged observations. However, ARIMA's reliance on linearity and stationarity limits its applicability in financial forecasting, where data often exhibits non-linear patterns and non-stationarity (Tsay, 2005). These constraints have driven researchers toward more advanced techniques, such as machine learning models, which are better prepared to handle the intricacies of financial data.

Although ARIMA models may work well for predicting short-term trends under stable market conditions, they often struggle to account for sudden market changes or longer-term trends driven by macroeconomic factors. This limitation becomes especially clear during times of market instability, such as the 2008 financial crisis or the volatility seen during the COVID-19 pandemic, when ARIMA models failed to offer reliable predictions. Such events underscore the demand for more resilient forecasting methods that can capture the complex and dynamic behaviour of financial markets. Consequently, interest in machine learning techniques has increased, as they offer a way to address these shortcomings by utilizing large datasets and advanced algorithms for more precise financial market modelling.

2.3 Emergence of Machine Learning in Financial Forecasting

2.3.1 Machine Learning Methods in Financial Forecasting

With the limitations of conventional methods becoming more evident, the financial forecasting field has seen a growing interest in machine learning techniques. Unlike traditional statistical approaches, machine learning models do not depend on prior assumptions about variable relationships. Instead, they learn these connections from the data itself, making them highly effective in capturing the intricate, non-linear behaviour of financial markets.

Random Forest and Gradient Boosting models, such as XGBoost, have shown promise in improving predictive accuracy by capturing non-linear relationships in the data (Breiman, 2001; Chen & Guestrin, 2016). Random Forests utilize an ensemble learning approach where multiple decision trees are built during training. The final prediction is made by taking the majority vote in classification tasks or calculating the mean in regression tasks across all the trees. The ensemble structure of Random Forests enables them to process large datasets with many features and helps mitigate overfitting, which is particularly important in the unpredictable landscape of financial markets. These models have proven effective across various financial forecasting tasks, including predicting stock prices, credit risk evaluation, and portfolio optimization, consistently outperforming traditional methods.

Conversely, Gradient Boosting builds trees one after another, with each new tree created to correct the mistakes of the earlier ones. XGBoost, a widely used implementation of Gradient Boosting, is favoured in the financial sector for its scalability, efficiency, and capacity to manage missing data. XGBoost has been applied in numerous financial forecasting tasks, including predicting stock prices, identifying market trends, and detecting fraudulent transactions. These ensemble learning methods aggregate predictions from multiple decision trees, thereby enhancing robustness and reducing overfitting. This allows them to perform well in the unpredictable nature of financial markets, as they are able to capture intricate relationships between variables that traditional models may overlook.

2.3.2 Support Vector Machines (SVM)

SVM have been used in financial forecasting, especially for stock price prediction, because of their capability to process non-linear data (Vapnik, 1995). SVM models operate by identifying the best possible boundary that successfully distinguishes data points into separate categories, making them effective in distinguishing between different market conditions. The flexibility of SVM in handling various kernels allows it to model complex financial data, making it a valuable tool alongside other machine learning method.

SVMs are especially beneficial in situations where the connection between input variables and the output is intricate and non-linear. By applying kernel functions, SVMs map data into a more complex, multi-dimensional space, allowing for linear separations within that space. This feature is particularly useful in financial forecasting, where variables like interest rates, stock prices, and economic indicators tend to exhibit complex, non-linear relationships. SVMs have been effectively used in several financial forecasting applications, such as stock price forecasting, option pricing, and credit risk evaluation, demonstrating better performance than traditional linear models.

However, despite their advantages, SVMs also have some limitations. A key challenge in applying SVMs to financial forecasting is choosing the right kernel function, as it can greatly influence the model's performance. Additionally, SVMs can be resource-intensive, especially when dealing with extensive datasets, which is common in financial forecasting. Therefore, SVMs are frequently combined with other machine learning techniques, like ensemble approaches or deep learning models, to enhance both their performance and scalability.

2.3.3 Neural Networks

Neural Networks, especially Feedforward Neural Networks, have been used in financial forecasting to model complex, non-linear relationships. Feedforward Neural Networks consist of several layers of interconnected nodes, where each layer is linked to next to those in the preceding and following layers. The links between these nodes are assigned weights, which are adjusted throughout the training process to reduce the prediction error between predicted and actual outcomes. However, their inability to retain memory of previous inputs has limited their effectiveness in time series forecasting (Hinton, 1989).

This limitation has resulted in the creation and application of Recurrent Neural Networks (RNNs), which introduce loops in the network, enabling it to store information from previous inputs (Rumelhart et al., 1986). RNNs are specifically built to process data in sequences, these models are highly effective for tasks like time series forecasting, including predicting stock prices. Traditional RNNs encounter the vanishing gradient issue, which limits their capacity to learn and retain long-term dependencies in data. This problem occurs when gradients diminish to almost zero or increase rapidly during backpropagation, leading to unstable training and diminishes their effectiveness in capturing long-term dependencies.

LSTM networks were developed by Hochreiter and Schmidhuber in 1997 to address this issue. LSTM networks, a specific type of RNN, were created to solve the vanishing gradient issue. These networks contain memory cells that store information over extended periods, allowing them to learn from long data sequences and identify and retain both short-term and long-term trends in the data. This structure makes LSTM networks highly effective for tasks involving time series predictions, such as predicting stock prices, where historical events impact future trends.

2.4 Deep Learning Approaches: The Rise of LSTM Networks

2.4.1 Overview of Recurrent Neural Networks and LSTM

Recurrent Neural Networks (RNNs) were introduced to overcome the limitations of feedforward networks by incorporating loops that help retain information from previous inputs (Rumelhart et al., 1986). Standard RNNs struggle with the vanishing gradient issue, which reduces their ability to learn and retain long-term patterns. The vanishing gradient issue arises when the gradients used to adjust weights during training become very small as they move backward through the network layers, resulting in minimal updates and inadequate learning of long-term patterns. This challenge greatly impacts the performance of standard RNNs in tasks requiring extracting insights from extended data sequences, such as those found in time series forecasting in financial markets.

LSTM networks, introduced by Hochreiter and Schmidhuber in 1997, were designed to solve this issue by utilizing memory cells that preserve their state over extended periods. The memory cells function through three key components: the input gate, the forget gate, and the output gate. The input gate controls the amount of new data that is allowed into the memory cell, while the forget gate manages which previously stored data should be retained or discarded, and the output gate regulates how information is transferred from the memory cell to other parts of the network. This structure enables LSTM networks to selectively retain or discard information, strengthening their capability to learn and retain long-term patterns in sequential data. This architecture makes LSTM networks are especially effective in time series forecasting, such as stock price prediction, where previous events can significantly influence future trends.

2.4.2 Applications of LSTM in Stock Price Prediction

LSTM networks have gained widespread use in stock price forecasting because of their exceptional capability to model time-based dependencies in financial data. Unlike traditional models that rely on fixed lag structures or assumptions of linearity, LSTM networks can learn directly from the data, capturing complex patterns and interactions that may not be immediately apparent. This feature is crucial in financial markets, where stock prices are affected by numerous factors mainly macroeconomic indicators, investor sentiment, and geopolitical dynamics.

Research by Fischer and Krauss (2018) shows that LSTM models surpass traditional approaches like ARIMA, especially when it comes to capturing the non-linear and non-stationary aspects of stock prices. For instance, LSTM networks have demonstrated effectiveness in capturing sudden market changes, like sharp increases or declines in prices, which are typically challenging for traditional models to predict accurately. Their application has been shown to be effective in predicting the stock prices of major companies, including MasterCard, Visa, American Express, and Capital One, where the models have provided more accurate forecasts compared to traditional methods. In these applications, LSTM networks have been utilized for predicting stock market trends on a daily, weekly, and monthly basis, showing considerable enhancements in prediction accuracy.

Additionally, LSTM networks have been combined with other machine learning methods to boost their performance. For instance, researchers have merged LSTM networks with sentiment analysis techniques, which analyse news, social media, and other textual data to assess market sentiment. Integrating sentiment data into LSTM models has allowed researchers to enhance the model's predictive accuracy, especially in forecasting market reactions to news or changes in investor sentiment. This integration of different data sources and machine learning techniques represents a promising direction for future research in forecasting stock prices.

2.5 Comparative Analysis of Forecasting Models

2.5.1 Evaluating the Performance of LSTM Networks

The performance of LSTM networks in estimating stock prices has been evaluated using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). These metrics offer a numerical assessment of the model's forecasting precision and reliability. Mean Squared Error (MSE) determines the average squared difference between actual and predicted values, where smaller values suggest higher accuracy. Mean Absolute Error (MAE), on the other hand, calculates the average size of the errors in predictions, regardless of their direction, offering a simple way to gauge the model's accuracy. R-squared (R^2) indicates the proportion of variance in the dependent variable that is accounted for by the independent variables, with values closer to 1 representing a stronger model fit.

Research consistently demonstrates that LSTM models yield lower error rates than traditional models like ARIMA and other machine learning techniques, including Random Forests (Siami-Namini et al., 2018). This enhanced performance is mainly due to LSTM's capacity to learn and retain both short-term variations and long-term trends in financial data. For example, studies have demonstrated that LSTM networks are particularly effective in predicting market volatility, demonstrating superior accuracy and reliability compared to traditional models. LSTM networks can capture and learn from the time-based patterns within the data, enabling them to identify complex patterns that traditional models often overlook.

Moreover, LSTM networks have demonstrated strong performance across various market conditions, including stable, volatile, and crisis periods. During stable market periods, LSTM networks can accurately predict gradual trends and price movements. During volatile or crisis periods, mainly the 2008 financial crisis or the COVID-19 pandemic, LSTM networks have proven effective at capturing sudden market changes and forecasting price movements that traditional models frequently miss. These findings suggest that LSTM networks are not only more accurate but also more resilient in the face of market uncertainty.

2.5.2 Challenges in Implementing LSTM Models

Despite their advantages, implementing LSTM networks for financial forecasting is not without challenges. The necessity for thorough data pre-processing, appropriate model parameter selection, and the potential for overfitting are notable challenges. Pre-processing data is a crucial part of implementation since financial datasets frequently include noise, missing values, and outliers, all of which can impact model performance. Common techniques like data normalization, imputation, and outlier detection are used to prepare the data for analysis, though these processes can be time-intensive and demand meticulous attention.

Moreover, LSTM models require substantial computational resources, which can be a limitation in certain applications. Training LSTM networks requires adjusting millions of parameters, making the process computationally demanding and resource-intensive, particularly when working with extensive datasets The difficulty is intensified by the need to fine-tune hyperparameters like learning rate, batch size, and the number of layers to achieve the optimal performance. Hyperparameter tuning is often done through trial and error or using techniques such as grid search and random search, although these techniques can be resource-intensive and require considerable expertise.

Similar challenges are highlighted in the research of Livieris et al. (2020), who stress the importance of tuning hyperparameters and ensuring data quality to achieve optimal performance. In their study, Livieris et al. highlight the trade-offs involved in designing LSTM networks, where achieving high accuracy often requires balancing model complexity with computational efficiency. They also highlight the significance of regularization techniques, such as dropout and early stopping, which are commonly applied to reduce the risk of overfitting, a frequent problem in deep learning models that can hinder generalization to new data.

2.6 Integrating Macroeconomic Indicators into Financial Forecasting Models

2.6.1 Impact of Macroeconomic Indicators on Stock Prices

Macroeconomic factors like interest rates, GDP growth, and inflation have a substantial impact on stock prices, making them essential for accurate financial forecasting. These factors offer important insights into the wider economic conditions that influence financial markets. For instance, fluctuations in interest rates can alter borrowing costs and affect corporate profits, while GDP growth is frequently linked to overall market performance. Inflation, on the other hand, can erode purchasing power and impact consumer spending, which in turn affects company revenues and stock prices.

Research by Stock and Watson (2003) shows that incorporating these indicators into forecasting models can enhance predictive accuracy by providing a broader context for financial market movements. By integrating macroeconomic data with stock price data, researchers can develop models that capture the complex interactions between economic conditions and market behaviour. For instance, LSTM networks can be trained using both stock price data and macroeconomic indicators, enabling them to identify relationships between these factors and enhance prediction accuracy.

Integrating these variables into LSTM models has been shown to strengthen the consistency and accuracy of stock price projections. Research shows that LSTM models utilizing macroeconomic indicators can more accurately predict market trends during times of economic turbulence, such as recessions or financial crises. This is due to the model's ability to factor in macroeconomic influences on stock prices, resulting in more precise and context-aware forecasts.

2.6.2 Challenges and Opportunities in Combining Macroeconomic Data with LSTM Models

Combining macroeconomic data with LSTM models presents both challenges and opportunities. A significant challenge is the complexity of macroeconomic data, which demands meticulous pre-processing and thoughtful variable selection. Macroeconomic data tends to be noisy, multidimensional, and frequently revised, complicating its integration with financial time series data. Additionally, macroeconomic indicators are often released at different frequencies (e.g., quarterly GDP growth vs. daily stock prices), which can complicate the modelling process. Researchers must carefully select and pre-process these indicators to ensure that they are appropriately aligned with the financial data and can be effectively used in the LSTM model.

However, integrating these indicators provides valuable opportunities to enhance both the accuracy and contextual relevance of predictions. For instance, incorporating indicators like consumer sentiment, employment rates, and housing data allows LSTM models to offer a more complete perspective on the factors influencing stock prices. This comprehensive approach to financial forecasting assists investors and policymakers in making more informed decisions by offering deeper insights into the economic conditions that shape market behaviour.

As highlighted by Bernanke and Gertler (1999), future studies could investigate advanced techniques for combining macroeconomic indicators, potentially using hybrid models that merge LSTM networks with alternative machine learning approaches. For example, researchers could design models utilizing LSTM networks can be used to capture time-based patterns in stock prices, while techniques likely Principal Component Analysis (PCA) or Factor Analysis can minimize the dimensionality of macroeconomic data and highlight the most important features. This strategy could address the difficulties related to high-dimensional macroeconomic data and enhance the overall efficiency of the forecasting model.

2.7 Real-World Applications and Case Studies

2.7.1 Case Study: Predicting Stock Prices for Financial Institutions

LSTM networks have been successfully applied in real-world scenarios to predict the stock prices of major financial institutions, including MasterCard, Visa, American Express, and Capital One. A detailed case study by Fischer and Krauss (2018) illustrates how LSTM models were developed, trained, and evaluated for these companies, demonstrating their practical utility and effectiveness. The study emphasizes the significance of data quality, model tuning, and incorporating external factors like macroeconomic indicators to achieve precise forecasts.

In these case studies, LSTM networks were used to predict stock prices over various time horizons, ranging from daily to monthly predictions. The models were trained on historical stock price data, along with additional features such as trading volume, market indices, and macroeconomic indicators. The findings revealed that LSTM networks exceeded conventional models like ARIMA and moving averages in both prediction accuracy and their ability to track market trends. For example, with MasterCard, the LSTM model successfully predicted short-term price movements after major news events, showcasing its ability to assimilate new data and update its forecasts accordingly.

Additionally, these case studies highlight the value of incorporating domain expertise into the modelling process. For instance, the researchers integrated financial ratios like the price-to-earnings (P/E) ratio and dividend payout into the LSTM model to improve its predictive capacity. By merging domain-specific knowledge with advanced machine learning techniques, the researchers developed models that were both accurate and interpretable, offering valuable insights into the factors influencing stock price movements.

2.7.2 Lessons Learned from Real-World Applications

Real-world applications of LSTM networks provide valuable lessons for future research. Key takeaways include the need for extensive data pre-processing, the importance of hyperparameter tuning, and the potential benefits of integrating macroeconomic indicators into forecasting models. These insights emphasize the need for a comprehensive approach to financial forecasting, blending machine learning techniques with domain-specific expertise for optimal outcomes (Livieris et al., 2020).

One challenge identified in these case studies is the necessity for ongoing model monitoring and updates. Financial markets are dynamic, and the relationships between variables may fluctuate over time. Therefore, LSTM models have to be periodically retrained on fresh data to maintain their accuracy and relevance. This demands a reliable data pipeline capable of managing large datasets, along with the computational power to retrain models regularly.

Another important lesson is the value of model interpretability. While LSTM networks are powerful tools for financial forecasting, their complex architecture can make them difficult to interpret. However, by integrating domain-specific knowledge and using techniques such as feature importance analysis, researchers can develop models that provide actionable insights into the factors driving stock prices. This is particularly important for decision-makers, who need to understand not only the predictions but also the reasoning behind them.

2.8 Future Directions in Financial Forecasting

2.8.1 The Role of Sentiment Analysis in Stock Price Prediction

Incorporating sentiment analysis, especially from news sources and social media, into financial forecasting models presents a promising direction for future studies. Sentiment analysis leverages natural language processing (NLP) methods to interpret text data and determine the underlying sentiment. Within financial forecasting, sentiment analysis can evaluate market sentiment by examining news stories, social media posts, financial reports, and other textual data sources.

Sentiment analysis can provide additional context and insight into market sentiment, which often drives stock price movements. For instance, favorable news regarding a company's earnings or product launch can trigger a stock price increase, whereas negative reports about regulatory challenges or leadership shifts may result in a decline. By incorporating sentiment data into LSTM networks, researchers can develop models that are better equipped to anticipate these market reactions and adjust their predictions accordingly.

Research by Siami-Namini et al. (2018) suggests that combining sentiment analysis with LSTM networks could significantly enhance the accuracy and relevance of stock price predictions. In their study, the researchers used sentiment data from social media platforms to predict stock price movements, with the LSTM network capturing the temporal dependencies in the sentiment data. The findings indicated that the LSTM model enhanced with sentiment data outperformed traditional models, especially in forecasting short-term price changes following news events.

Future studies could investigate the application of advanced sentiment analysis methods, such as deep learning NLP models, to improve the accuracy of predictions driven by sentiment analysis. Researchers might also explore incorporating additional text-based data sources, including earnings call transcripts or analyst reports, to offer a more complete understanding of market sentiment.

2.8.2 Developing Hybrid Models for Improved Accuracy

Developing hybrid models that merge LSTM networks with other machine learning techniques like Random Forests or XGBoost presents substantial opportunities for enhancing forecasting accuracy. Such hybrid models could take advantage of the strengths of various approaches, resulting in more resilient predictions that are well-suited to the complexity and volatility of financial markets (Chen & Guestrin, 2016).

For instance, a hybrid model might employ LSTM networks to detect time-based patterns in stock price data, while leveraging Random Forests or XGBoost to capture non-linear relationships between stock prices and factors like trading volume, market indices, and macroeconomic indicators. By combining these techniques, researchers can develop models that are more accurate and better able to handle the complexities of financial data.

Moreover, hybrid models can also help address some of the limitations of LSTM networks, such as their sensitivity to hyperparameters and the risk of overfitting. For instance, ensemble methods like Random Forests can provide more stable predictions by averaging the outputs of multiple models, reducing the impact of overfitting. Additionally, hybrid models can be designed to incorporate domain-specific knowledge, such as financial ratios or technical indicators, to further enhance their predictive power.

Future studies could investigate the creation of hybrid models that integrate LSTM networks with advanced machine learning approaches, like deep reinforcement learning or transfer learning. These methods could provide new perspectives on financial forecasting, aiding researchers in building models that are more precise, resilient, and adaptable to fluctuating market conditions.

2.9 Conclusion

Research on stock price prediction with LSTM networks reveals considerable progress in this area. LSTM networks serve as effective tools for identifying complex patterns in financial data and generating precise forecasts. However, their implementation is not without challenges, and there is still much to learn about how to optimize these models for different financial contexts. Future research should aim to overcome these challenges, investigate new data sources, and design hybrid models to improve the precision and practicality of financial forecasting.

CHAPTER 3: RESEARCH APPROACH

This chapter outlines the methodology used to build a stock price forecasting model, following the CRISP-DM framework, which directs the process from data collection and preparation to model selection and assessment. The emphasis is on utilizing Long Short-Term Memory (LSTM) networks because of their efficiency in time-series analysis. Each step of the research process is outlined, ensuring a structured approach to achieving accurate and reliable stock price predictions.

3.1 Introduction to the Research Approach

This chapter describes the research methodology employed to build an accurate predictive model for stock prices of key financial institutions: MasterCard, Visa, American Express, and Capital One. The model, built using LSTM networks, is structured to capture time-based and sequential trends in stock price data. The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework was chosen for its organized, iterative method, making it well-suited for data-driven financial projects. The chapter discusses in detail the various stages involved in the methodology, including data collection, pre-processing, model development, and evaluation. It also addresses real-world deployment considerations, focusing on how the LSTM model can be practically implemented in the financial sector. In addition to these technical aspects, the chapter also covers the ethical considerations related to data handling, emphasizing the importance of data privacy and the avoidance of biases in model predictions. By the end of this chapter, readers will have a thorough understanding of the methodology used to ensure the precision, reliability, and ethical integrity of the research.

3.2 Selection of Research Methodology

3.2.1 Overview of CRISP-DM

CRISP-DM was chosen as the research framework for its structured yet flexible approach to data analysis. The CRISP-DM framework consists of six essential phases: understanding business objectives, analysing data, preparing data, building models, evaluating results, and deploying the solution, forming an iterative process that facilitates thorough model exploration and continuous improvement as new data becomes available. The iterative structure of CRISP-DM is particularly beneficial for financial forecasting, as it allows models to be refined continuously to account for new trends and economic changes. The framework also facilitates the exploration of different models, enabling researchers to experiment with various algorithms and assess their effectiveness against key performance metrics. Its industry-agnostic nature makes it a popular choice for complex projects like financial forecasting, where both domain knowledge and data science expertise are essential for creating actionable insights.

3.2.2 Justification for CRISP-DM in Financial Forecasting

Financial forecasting is inherently a complex and dynamic task, requiring models that are not only accurate but also adaptable to changing market conditions. CRISP-DM’s step-by-step approach ensures that the data mining process is conducted systematically, from grasping the financial institutions' goals to the deployment of the final model. The Business Understanding phase aligns the model’s development with the financial institutions' goals, such as optimizing investment strategies and minimizing risk exposure. The Data Understanding and the Data Preparation phase ensures the dataset is fully cleaned, complete, and properly formatted for model training. This is crucial in financial forecasting, where the data quality has a direct impact on the model's capability to make correct predictions. Moreover, CRISP-DM’s adaptability allows for frequent reassessments and model refinements, ensuring that the LSTM network can continuously learn from new data, improving the overall reliability of stock price predictions.

3.3 Business Understanding

3.3.1 Project Objectives

The primary objective of this project is to create an accurate predictive model for the stock prices of four major financial institutions spanning a 16-year period from 2008 to 2024. The primary goal is to deliver reliable stock price predictions that can aid investors, financial analysts, and portfolio managers in making informed decisions that maximize returns while minimizing risks. By integrating historical stock prices with macroeconomic indicators, the model seeks to provide a broad perspective on market trends and external factors, enhancing the accuracy and robustness of predictions. The research also aims to address the limitations of traditional forecasting models by leveraging the sequential learning capabilities of LSTM networks. This research also examines the inclusion of major macroeconomic indicators, mainly GDP growth, inflation, and interest rates, to develop a comprehensive forecasting model that considers broader economic impacts on stock prices.

3.3.2 Importance of Accurate Financial Forecasting

Precise forecasting is crucial within the financial sector, as it helps optimize investment strategies, manage risks, and support informed economic policy decisions. For individual investors and portfolio managers, the ability to predict stock prices with accuracy enables better portfolio management, allowing for more precise entry and exit points that maximize potential returns while minimizing exposure to risk. Financial institutions rely on these predictions for strategic decision-making, such as timing mergers, acquisitions, and capital allocation. For policymakers, accurate forecasts provide a deeper understanding of market trends, helping to inform macroeconomic policy decisions that foster market stability. This research aims to contribute meaningfully to decision-making by developing a model that incorporates both historical trends and macroeconomic indicators. LSTM networks' ability to recognize complex patterns in stock price data boosts the model's predictive accuracy, making it an essential tool for financial market analysis.

3.4 Data Understanding

3.4.1 Data Sources

Yahoo Finance was the primary data source for this study, providing extensive historical stock data, including daily open, high, low, and close prices, as well as adjusted close prices and trading volumes. This dataset, covering the years from 2008 to 2024, provides a detailed view of stock performance over a 16-year period. Along with stock prices, this study includes macroeconomic indicators from public economic databases, such as GDP growth, inflation, and interest rates. These indicators are essential for gaining insights into the broader economic factors that affect stock price fluctuations. By combining micro-level stock data with macro-level economic indicators, this research ensures that the predictive model addresses all relevant factors influencing stock prices.

3.4.2 Exploratory Data Analysis (EDA)

EDA was performed to thoroughly understand the dataset's structure, trends, and any potential anomalies. Visualization techniques, including histograms, scatter plots, and correlation matrices, were utilized to detect important patterns and relationships within the data. EDA revealed significant trends, such as seasonality in stock prices and the influence of macroeconomic variables like interest rates on stock performance. For example, the correlation analysis revealed an inverse relationship between stock prices and interest rates, showing that increasing interest rates generally resulted in falling stock prices. This insight was critical for shaping the feature engineering process, as it informed the selection of variables that would be most relevant for the LSTM model.

3.5 Data Preparation

3.5.1 Data Cleaning

Rigorous data cleaning procedures were applied to ensure the dataset's accuracy and completeness. Imputation methods, especially forward-fill for time-series data, were used to handle missing values and maintain the continuity of stock price data. Duplicate records were identified and removed, and potential outliers were flagged for further investigation. For example, sudden spikes or drops in stock prices that were not linked to significant market events were considered potential outliers and were either removed or adjusted. The comprehensive data cleaning process was critical in ensuring the dataset's accuracy and reliability, reducing the risk of skewed predictions during model training.

3.5.2 Feature Engineering

Feature engineering was crucial in boosting the LSTM model's predictive performance. Several technical indicators, such as moving averages (calculated over different time windows), Relative Strength Index (RSI), and Bollinger Bands, were engineered and incorporated into the dataset. Moving averages were calculated over 5-day, 20-day, and 50-day periods to identify short-term and long-term stock price trends. RSI was included to provide a measure of stock price momentum, indicating whether a stock was overbought or oversold, while Bollinger Bands were used to gauge stock price volatility. These technical indicators were chosen for their demonstrated effectiveness in financial analysis and their ability to enhance the LSTM model's capacity to understand intricate patterns in the data.

3.5.3 Data Splitting

The dataset was split into training, validation, and test sets to ensure a comprehensive evaluation of the model's performance ahead of deployment. The training set, made up of historical data from previous years, was utilized to train the model, while the validation set assisted in refining the hyperparameters. The test set, made up of more recent data, was employed to evaluate how well the model can make accurate future stock price predictions. Organizing the data in chronological order ensured that the model was trained and tested under conditions that closely resembled real-world scenarios, where predictions are made based on past data to forecast future outcomes.

3.6 Modelling

3.6.1 Model Selection

LSTM networks were chosen as the primary model for this research due to their ability to capture sequential relationships in time-series data. In addition to LSTM, additional models, including ARIMA, Prophet, Random Forest, and XGBoost, were evaluated as benchmarks. ARIMA and Prophet are traditional time-series forecasting models that excel in identifying linear trends and seasonality, while Random Forest and XGBoost are machine learning models recognized for their capacity to capture non-linear relationships. By evaluating the performance of these models, the study validated the higher accuracy and dependability of LSTM networks in predicting stock prices.

3.6.2 LSTM Network Architecture

The LSTM network architecture was designed to efficiently process the sequential nature of stock price data. The model was built with multiple LSTM layers, followed by dense layers to generate the final stock price predictions. Dropout layers were included to prevent overfitting, and batch normalization layers were used to stabilize the training process. The architecture was designed to balance complexity and training efficiency, allowing the LSTM network to capture dependencies in stock price data over both short and long timeframes. The model was implemented using Python and TensorFlow, with the training process parallelized across multiple GPUs to accelerate training time.

3.6.3 Hyperparameter Tuning

Hyperparameter optimization was performed to improve the LSTM model's performance by utilizing methods like grid search and random search. These techniques helped identify the optimal combination of hyperparameters, including learning rate, batch size, number of epochs, and the number of LSTM layers. The process involved training multiple instances of the model with varying hyperparameter configurations, and then tested their effectiveness on a separate validation dataset. The configuration that demonstrated the best performance was then chosen for final testing on the test set. Hyperparameter selection was based on both domain expertise and experimental results, with the goal of optimizing model accuracy while minimizing the potential for overfitting The resulting model struck a balance between complexity and generalizability, ensuring its ability to accurately forecast stock prices under varying market conditions.

3.7 Evaluation

3.7.1 Model Evaluation Metrics

Once the LSTM model and other benchmark models were trained, the next step involved assessing their performance using standard metrics for time-series forecasting. Key metrics employed in the evaluation included Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R-squared (R^2) value. MSE and RMSE offer a measure of overall error magnitude by taking the square of the differences between the predicted and actual values, which makes them more sensitive to significant deviations. MAE provides a simpler measure of the average error size, offering a clearer understanding of the model's general performance. The R-squared (R^2) value, commonly used to evaluate goodness-of-fit, was crucial in assessing how well the model explained the fluctuations in stock prices effectively. Time series plots were used to visually compare predicted and actual prices, enabling an evaluation of how accurately the LSTM model’s predictions matched real market movements. This visual examination also helped identify systematic prediction errors, such as underperformance during times of high market volatility.

3.7.2 Comparative Analysis with Other Models

The LSTM model's performance was compared against ARIMA, Prophet, Random Forest, and XGBoost to assess its relative effectiveness. ARIMA, known for its strong capabilities in linear trend and seasonality detection, performed well in stable periods but struggled during times of high volatility or non-linear market shifts. Prophet was effective in managing seasonal patterns but struggled to capture more complex relationships within the data. In contrast, Random Forest and XGBoost, being machine learning models, were better at managing non-linear relationships but lacked LSTM's ability to learn from sequential data. The comparative analysis demonstrated that LSTM consistently outperformed the other models, particularly in capturing long-term dependencies and producing accurate predictions during volatile periods, mainly the 2008 financial crisis and the COVID-19 pandemic. This validated the appropriateness of LSTM networks for financial time-series forecasting, where stock prices are affected by various dynamic factors over time.

3.8 Deployment Considerations

3.8.1 Real-Time Prediction System

The deployment phase integrates the LSTM model into a real-time prediction system, enabling continuous forecasting as new stock data becomes available. The system architecture incorporates data pipelines for collecting and pre-processing incoming data, alongside an API for generating predictions and delivering results to end-users. Designed for scalability and efficiency, the system is capable of processing large datasets and delivering real-time predictions. Additionally, monitoring tools are included to track the model's performance and ensure its accuracy and reliability over time. This real-time system is intended for use by financial analysts, investors, and portfolio managers, enabling them to make data-driven decisions using the most up-to-date stock price forecasts. Integration with existing financial software and platforms ensures that the predictions are both accessible and actionable for stakeholders.

3.8.2 Visualization Tools

Visualizations play a crucial role in interpreting model outputs and making them accessible to stakeholders. Power BI was employed to create interactive dashboards that display predicted stock prices alongside actual prices, providing users with a clear and intuitive interface. These dashboards include time series plots, scatter plots, and heatmaps, which allow users to explore model predictions and gain insights into the factors driving stock price movements. The visualizations offer drill-down features that allow users to analyse data at different levels of detail and uncover trends and patterns. Designed to be user-friendly and customizable, the dashboards provide stakeholders with the ability to customize the visualizations to suit their individual needs and preferences. Integrated with a real-time prediction system, these visualization tools ensure that users have access to the latest data, equipping them with the observations needed to make informed decisions using the latest available forecasts.

3.9 Ethical Considerations

3.9.1 Data Privacy and Security

In the modern data-centric landscape, ensuring the protection and security of financial and economic data is essential. Throughout this research, strict data privacy measures were implemented, including encryption of sensitive data, controlled access, and compliance with industry standards likely the General Data Protection Regulation (GDPR). Techniques like anonymization and pseudonymization were applied to safeguard the identities of individuals and organizations, ensuring that data was managed in an ethical and secure manner. Proper handling of financial data is essential for building trust with stakeholders and complying with legal regulations that govern data usage in the financial sector.

3.9.2 Ethical Implications of Forecasting Models

In addition to data privacy, this research also acknowledges the broader ethical implications of deploying predictive models in financial markets. Forecasting models can influence market behaviour, and as such, there is a responsibility to ensure that they are free from biases that could unfairly advantage or disadvantage certain groups. Biases in data, whether intentional or unintentional, can lead to skewed predictions that disproportionately impact investors or companies. For example, training the model on data from a stable economic period could cause difficulties in forecasting during economic downturns, which could lead to incorrect forecasts. To minimize these risks, the model was regularly monitored and updated with new data to ensure fairness and accuracy. Furthermore, transparency in how the model was developed and how predictions are made ensures that users understand the model’s limitations and do not rely solely on its outputs for critical decision-making.

3.10 Conclusion

This chapter provided an in-depth overview of the research methodology, explaining the choice of the CRISP-DM framework and outlining the steps involved in data collection, preparation, model building, and evaluation. Every phase was meticulously planned to maintain the accuracy, reliability, and ethical integrity of the predictive model. The LSTM network's success in surpassing both traditional and machine learning models confirms its value as a reliable tool for predicting stock prices in complex and volatile market conditions. The upcoming chapter will present the findings and deliver a thorough evaluation of the model’s effectiveness, along with a discussion of its practical uses in the financial industry.

CHAPTER 4: DATA ANALYSIS

This chapter offers an in-depth examination of the process undertaken to develop a stock price prediction model for major financial institutions, including MasterCard, Visa, American Express, and Capital One. This chapter adheres to the CRISP-DM framework, a methodical and iterative approach designed to bring structure to the data analysis process. The steps in this chapter encompass understanding the business objectives, analysing the dataset, preparing the data for modelling, building machine learning models, and evaluating their performance.

4.1 Business Understanding

Predicting stock prices is crucial for traders, portfolio managers, and financial institutions, as it aids in making well-informed decisions, especially in fluctuating markets. This study aims to create a model capable of accurately forecasting stock prices, helping to enhance investment strategies, better manage risks, and stabilize markets. This research centres on the use of LSTM networks, which are recognized for their ability to process time-series data and identify both short- and long-term dependencies in stock prices. This study focuses on key financial institutions such as MasterCard, Visa, American Express, and Capital One, with the objective of uncovering broader market trends and delivering insights into their performance. The study's business goals involve forecasting stock prices over a 16-year span (2008–2024), offering actionable insights for decision-makers, and incorporating macroeconomic factors like GDP, interest rates, and inflation to improve prediction precision.

4.2 Dataset

The data for this study was obtained from Yahoo Finance, which provided daily stock prices, trading volumes, and other relevant financial data for MasterCard, Visa, American Express, and Capital One. The dataset spans from 2008 to 2024, including major financial events mainly the 2008 financial crisis and the COVID-19 pandemic. Beyond stock data, the research also incorporated macroeconomic variables mainly GDP growth, inflation, and interest rates. These indicators help contextualize stock price movements by providing insight into external economic factors that influence financial markets.

4.3 Data Pre-processing

Data pre-processing plays a significant role in machine learning projects, particularly when working with time series data like stock prices. For this study, the pre-processing stage was crucial in converting raw data into a structured format suitable for machine learning models. This phase included data cleaning, addressing missing and duplicate values, and creating new features, enhancing the model's accuracy in predicting stock price.

4.3.1 Data Cleaning

Upon analysis, it was found that the dataset was complete, containing no missing or duplicate values. Despite this, each stage of the cleaning process was performed systematically to verify the data's integrity and ensure it was ready for use in model development. Missing values were handled using the forward-fill method, which ensures that no gaps exist in the time series data by propagating the last known value forward. This technique is particularly important in time series forecasting, where continuity is crucial for accurate model training. Additionally, any duplicate rows were identified and removed from the dataset, ensuring the data's uniqueness and avoiding bias in model predictions.

4.3.2 No Outlier Detection

Although outlier detection is often considered an important step in data pre-processing, it was not implemented in this research. The decision to exclude outlier detection was based on the nature of the stock market, where sudden price movements and extreme values are common due to market volatility. In financial markets, such outliers often reflect real-world events, such as company earnings reports, macroeconomic announcements, or political developments, and can provide valuable information. Removing or adjusting these outliers might lead to a loss of crucial signals that could improve the model's capability to detect and respond to market trends and behaviours. Additionally, sophisticated models like LSTM are capable of handling volatility and extreme price movements inherently, as they learn patterns over time and adapt to fluctuations.

4.3.3 Feature Engineering

In the feature engineering phase, several key technical indicators were created to enhance the model's ability to detect trends and patterns in stock price movements.

These indicators included moving averages, which were computed over different time windows to capture both short-term and long-term trends, and the Relative Strength Index (RSI), which helps identify overbought or oversold conditions in the market. Additionally, Bollinger Bands were utilized to measure volatility, providing upper and lower bounds for price movements. These features were selected for their proven effectiveness in financial analysis and were integral in enhancing the model's ability to capture intricate market behaviours, such as momentum, volatility, and trend reversals. By including these features, the dataset became more robust and better suited for machine learning model input, allowing for improved prediction accuracy.

4.3.3.1 Moving Averages

The 50-day (MA_50) and 200-day (MA_200) moving averages were computed to track both short-term and long-term price trends in stock data. The 50-day moving average is responsive to recent price changes, making it effective in detecting shifts in momentum. On the other hand, the 200-day moving average is less affected by short-term volatility and offers a better reflection of broader market sentiment. Comparing these two moving averages helps identify bullish or bearish trends when crossovers occur.

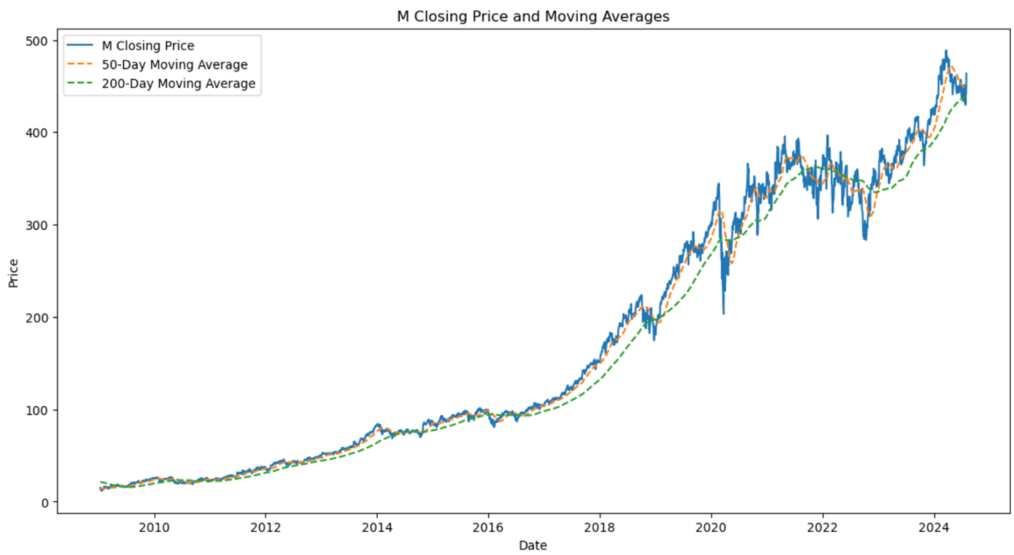

Fig 4.1 Moving Average

4.3.3.2 Bollinger Bands

Bollinger Bands are used to measure stock price volatility and include three components: a middle band (50-day moving average), an upper band, and a lower band. The upper and lower bands are positioned two standard deviations away from the middle band, widening during high volatility and narrowing during more stable times. Stocks that approach the upper band are often viewed as overbought, while those close to the lower band are seen as oversold.

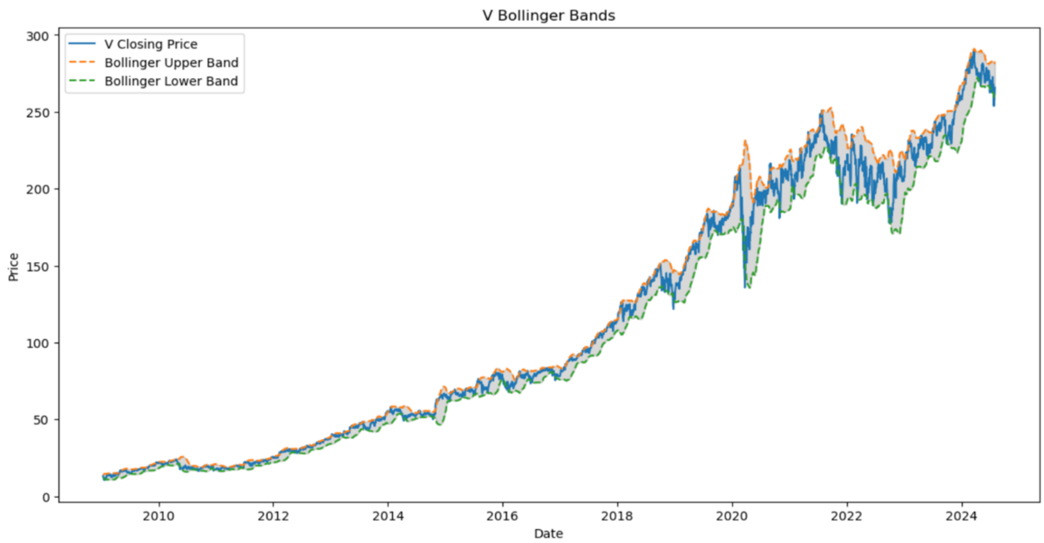

Fig 4.2 Bollinger Bands

4.3.3.3 Relative Strength Index (RSI)

The RSI estimates the rate and extent of recent price fluctuations, offering insight into whether a stock is in an overbought or oversold state. When the RSI exceeds 70, it typically signals indicating that a stock might be overvalued and could experience a price adjustment, whereas an RSI under 30 suggests the stock might be oversold, indicating potential buying opportunity.

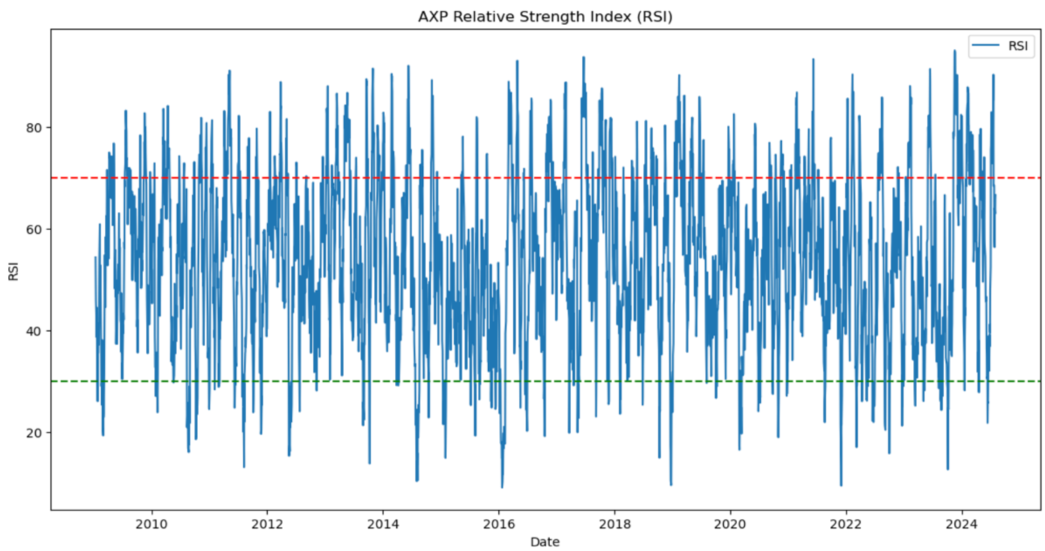

Fig 4.3 Relative Strength Index

4.3.4 Exploratory Data Analysis (EDA)

EDA aids in identifying key patterns, relationships, and anomalies in the dataset. By visualizing key metrics and distributions, EDA guides the development of models by providing insights into stock price behaviours, relationships between variables, and overall trends. In this study, the dataset includes historical stock price data for all four stocks from 2008 to 2024. The following visualizations were chosen to explore and summarize the dataset.

4.3.4.1 Stock Prices Over Time

The stock prices over time graph provides a comprehensive view of how all four stock prices evolved over the 16-year period, highlighting key market events and fluctuations. MasterCard and Visa show a steady upward trend, reflecting long-term growth, while American Express and Capital One experienced more moderate growth, with periods of stability and volatility. Significant financial events, mainly the 2008 financial crisis and the COVID-19 pandemic, led to visible price fluctuations for all companies, showcasing the market's sensitivity to external shocks. This plot is crucial for understanding the overall performance of these stocks, offering a visual timeline that contextualizes the broader market environment in which the predictive models will function.

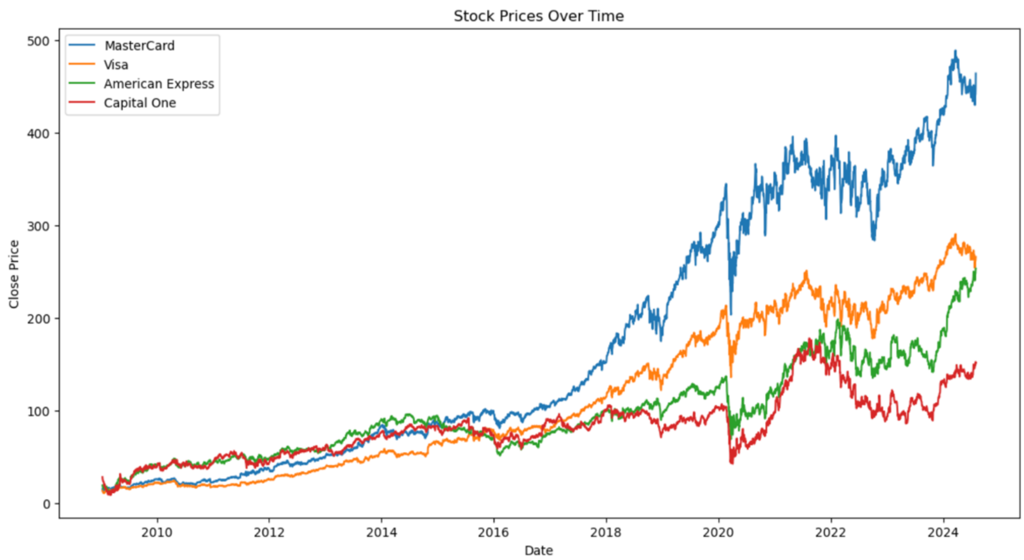

Fig 4.4 Stock Prices Over Time

4.3.4.2 Daily Returns Distribution

The daily returns distribution plot is crucial for understanding the variability, volatility, and risk associated with all the stocks. For all four companies, the returns tend to be centred around zero, indicating that most daily price changes are relatively small. However, outliers were observed, particularly during significant market events, reflecting extreme gains or losses. The distribution suggests that Visa and MasterCard experienced higher volatility compared to American Express, likely due to differences in market positioning and business models. This visualization provides important insights into the risk and volatility of each stock, helping investors assess the likelihood of large gains or losses and informing decisions about potential risks.

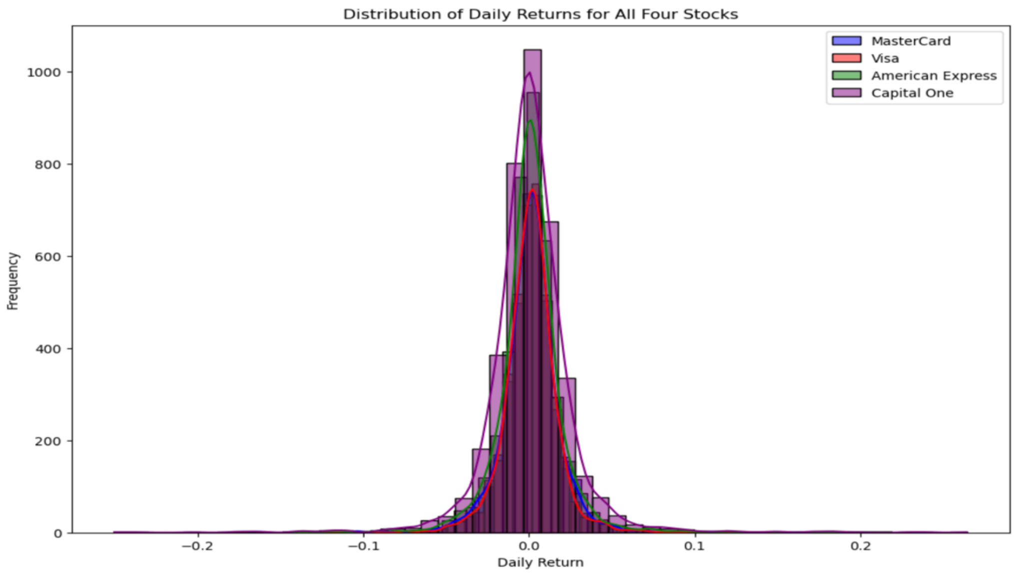

Fig 4.5 Daily Returns

4.3.4.3 Volume vs. Stock Prices

The volume vs. stock prices graph illustrates the relationship between trading activity and stock price movements for MasterCard, Visa, American Express, and Capital One. Significant price increases or decreases are often accompanied by spikes in trading volume, indicating strong investor interest or reactions to market events. For Visa and MasterCard, higher trading volumes during bullish trends suggest investor confidence in their long-term growth. Conversely, periods of low trading volume often align with stable or stagnant price movements, reflecting reduced market participation during consolidation phases. This graph is valuable for understanding how investor behaviour influences stock price fluctuations, providing critical insights for forecasting future market trends.

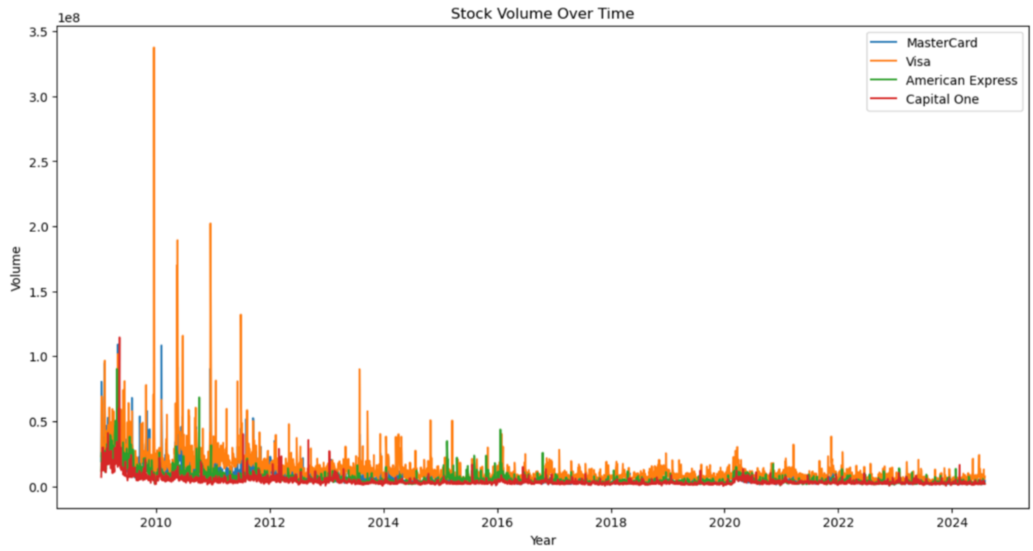

Fig 4.6 Stock Volume Over Time

4.3.4.4 Rolling Volatility Plot

The rolling volatility plot, calculated using a 20-day rolling standard deviation, captures the dynamic nature of stock price volatility for all four stocks. Significant spikes in volatility are observed during major economic events mainly the 2008 financial crisis and the COVID-19 pandemic, indicating increased market volatility. MasterCard and Visa exhibit more consistent volatility levels compared to American Express and Capital One, highlighting the relative stability of their business models. This plot is crucial for identifying periods of heightened market risk and understanding how volatility evolves over time, helping investors assess risk levels during different phases.

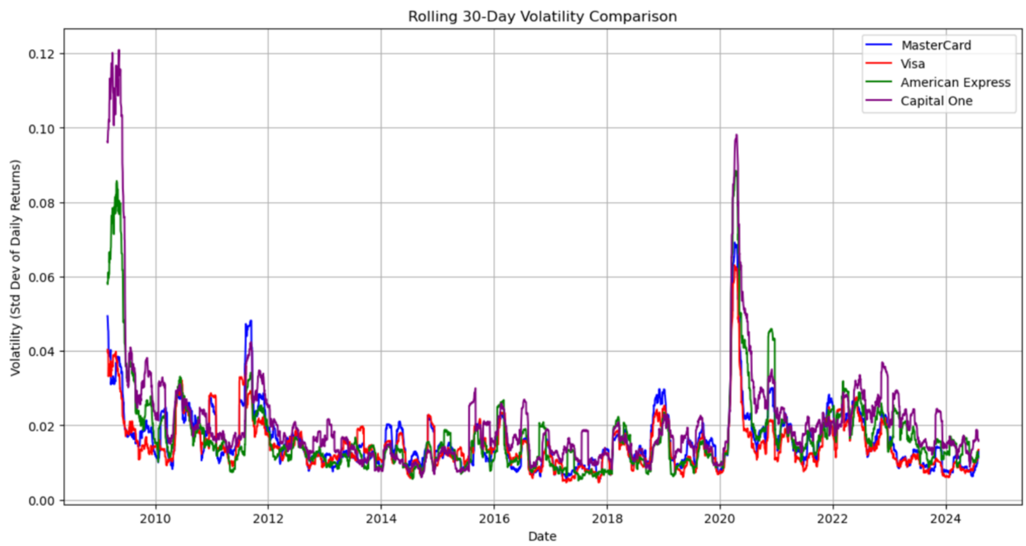

Fig 4.7 Rolling Volatility Comparison

4.3.4.5 Correlation Heatmap

The correlation heatmap visually illustrates the connections between stock prices, trading volumes, and technical indicators, allowing for the identification of important correlations. Strong positive correlations are observed between MasterCard and Visa stock prices, reflecting their similar roles in the payment processing industry and how they often move in tandem. Moderate correlations between trading volumes and price changes suggest that larger trading volumes are typically associated with more significant price movements. This heatmap provides a high-level overview of variable interactions, offering valuable insights for prioritizing factors in predictive analysis and model development.

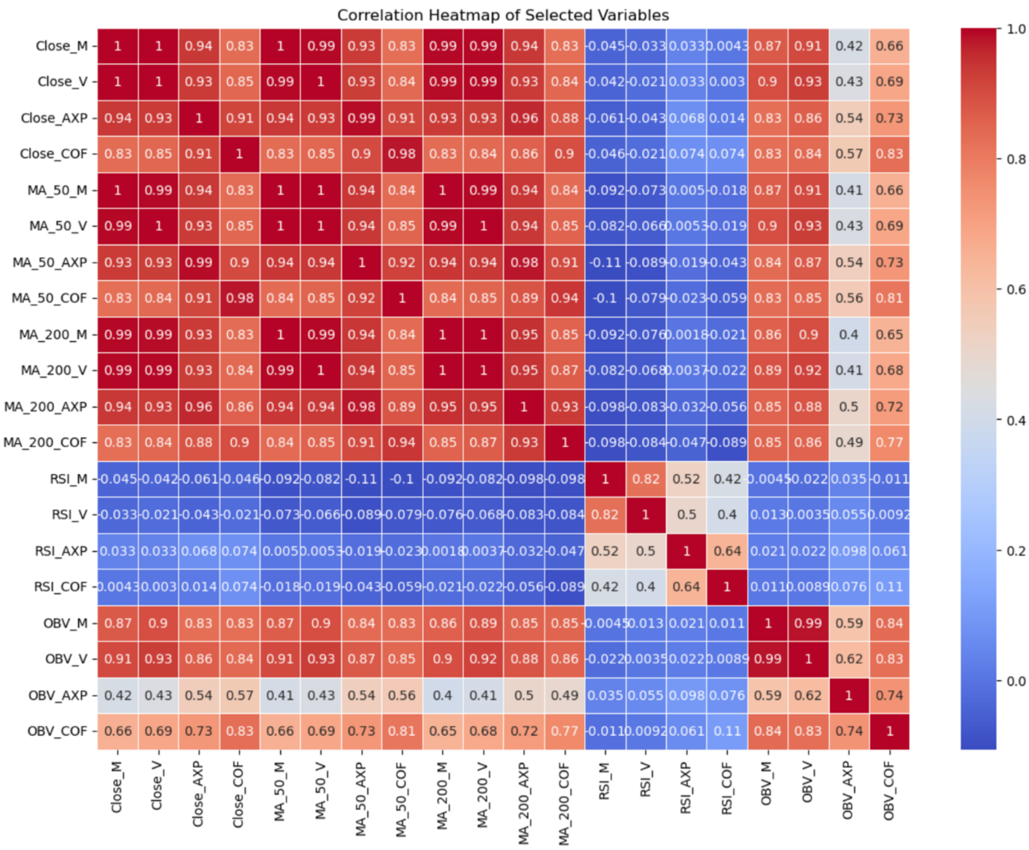

Fig 4.8 Correlation Heatmap

4.4 Modelling

The modelling phase concentrated on evaluating various machine learning techniques for forecasting stock prices, with a particular emphasis on the effectiveness of LSTM networks. The objective was to determine the most accurate model for predicting stock price of MasterCard, Visa, American Express, and Capital One. This section details the development and implementation of the models, including LSTM, ARIMA, Random Forest, and XGBoost, followed by a final ensemble model.

4.4.1 Data Pre-processing for LSTM

Before constructing the LSTM model, historical stock prices were pre-processed using MinMaxScaler to normalize values between 0 and 1, improving model convergence. A 60-day time window was chosen to create input sequences, using two months of past data to forecast the stock price for the following day. This step was critical to help the LSTM capture temporal dependencies in the data effectively.

4.4.2 LSTM Model Architecture

The LSTM model was constructed with two LSTM layers, each having 50 units, and a Dense layer used to generate predictions. The model was tuned using the Adam optimizer and employed the mean squared error (MSE) as its loss function. Early stopping was utilized to avoid overfitting, stopping the training process once the validation performance stopped improving. This setup enabled the LSTM to learn and retain both short- and long-term trends in stock price data.

4.4.3 ARIMA Model

The ARIMA model was used as a conventional statistical approach for forecasting time-series data. It was tuned for each company's stock data, focusing on capturing short-term trends. However, ARIMA's performance was limited by its difficulty in handling non-linear complexities that are typical in stock price movements.

4.4.4 Random Forest and XGBoost Models

Random Forest and XGBoost were applied to capture non-linear patterns within the stock data. Random Forest averaged outputs from multiple decision trees to prevent overfitting, while XGBoost applied boosting techniques to improve performance by handling errors iteratively. Both models leveraged time-lagged data to account for historical dependencies.

4.4.5 Ensemble Model

An ensemble model was developed by combining LSTM, ARIMA, Random Forest, and XGBoost outputs, with LSTM weighted the most (70%) due to its superior time-series capabilities. This weighted ensemble approach enhanced predictive accuracy by integrating the strengths of each model, resulting in improved generalization.

4.5 Model Evaluation

The models were assessed using metrics such as MSE, MAE, and R^2 metrics. LSTM showed the best overall accuracy, outperforming the ARIMA, Random Forest, and XGBoost models. However, the ensemble model provided the best overall performance by minimizing individual model weaknesses and combining their strengths.

4.5.1 Ensemble Model vs. LSTM

The comparison between the LSTM and ensemble models revealed that LSTM had the lowest MSE, confirming its effectiveness in capturing long-term patterns in stock prices. The ensemble model improved upon the results of ARIMA, Random Forest, and XGBoost but had a slightly higher MSE than LSTM, justifying further hyperparameter tuning to optimize LSTM's performance.

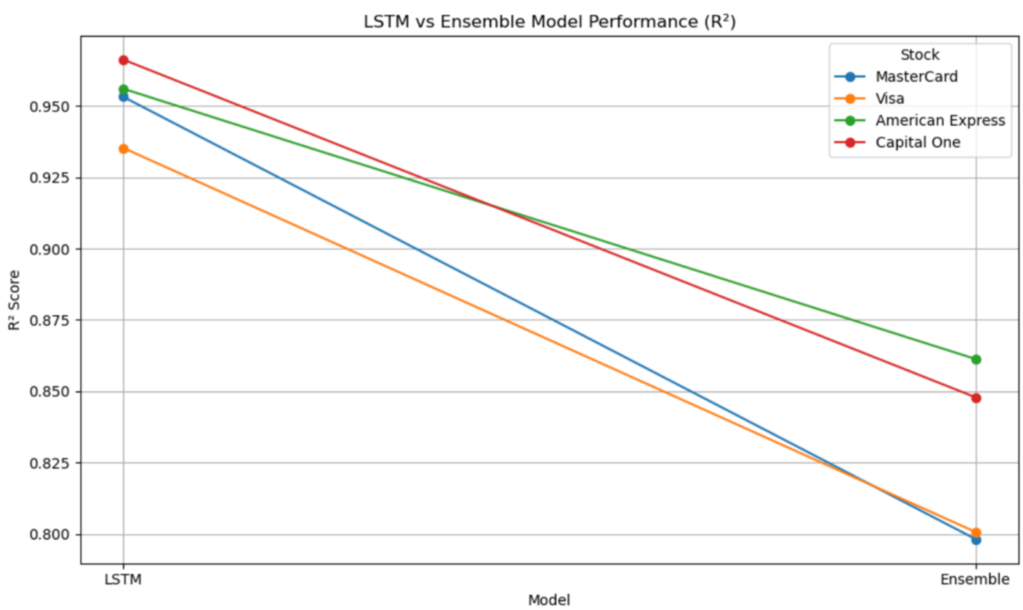

Fig 4.9 LSTM vs Ensemble Model Performance

4.5.2 Hyperparameter Tuning and Final Predictions

After selecting the LSTM model as the best-performing option, hyperparameter tuning was conducted to further improve its accuracy. Crucial parameters such as learning rate, batch size, number of epochs, and the number of LSTM layers were fine-tuned to reduce the Mean Squared Error (MSE) on the validation dataset. Following the identification of optimal hyperparameters, the LSTM model was retrained, resulting in enhanced performance. The Actual vs Predicted Stock Prices plots for MasterCard, Visa, American Express, and Capital One demonstrated that the tuned LSTM model closely tracked the actual stock prices. The findings confirm the model's ability to effectively capture short-term variations and long-term patterns in the stock market, demonstrating its suitability for time-series forecasting

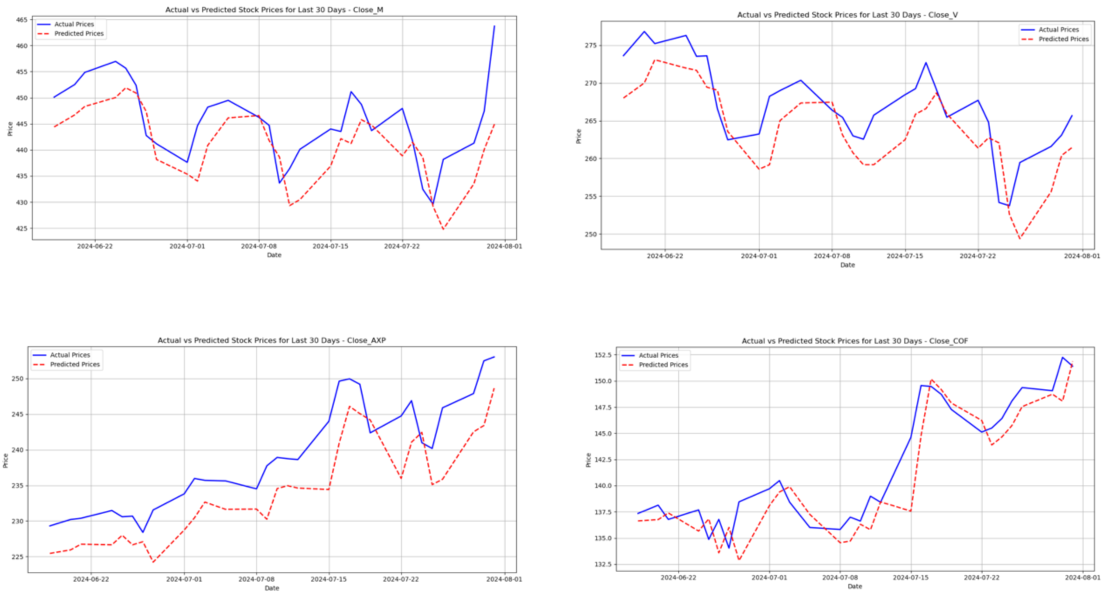

Fig 4.10 Actual vs Predicted Stock Prices

4.5.3 Dashboard Visualization

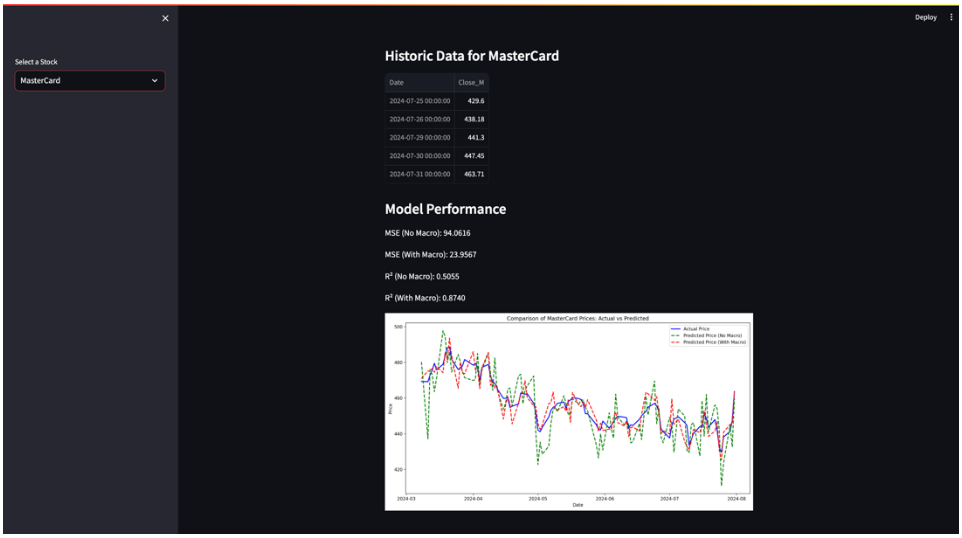

Fig 4.11 Dashboard Visualization

The dashboard features a drop-down menu that allows users to select different stocks, such as MasterCard, to view specific forecasting results. Once a stock is selected, the model utilizes real-time macroeconomic factors, mainly GDP and interest rates, to dynamically update the forecast, ensuring that the model adapts to current market conditions for more accurate predictions. In the Model Performance section, users can observe how the model performs both with and without macroeconomic factors, with real-time updates based on the latest data inputs. A comparison between actual and predicted prices is visualized, showing that models incorporating macroeconomic indicators provide better accuracy.

4.6 Analysis During 2008 Financial Crisis and COVID-19 Pandemic

The historical stock performance of MasterCard, Visa, American Express, and Capital One was deeply impacted by significant economic events. These crises presented unique challenges for stock markets globally, causing significant fluctuations in stock prices and presenting opportunities to test the robustness of predictive models, including LSTM networks.

4.6.1 2008 Financial Crisis

The 2008 financial crisis significantly impacted the stock prices of all stocks, with notable declines, especially for American Express and Capital One due to their exposure to credit risks. However, MasterCard and Visa, focused on digital payments, showed resilience and recovered rapidly post-crisis.

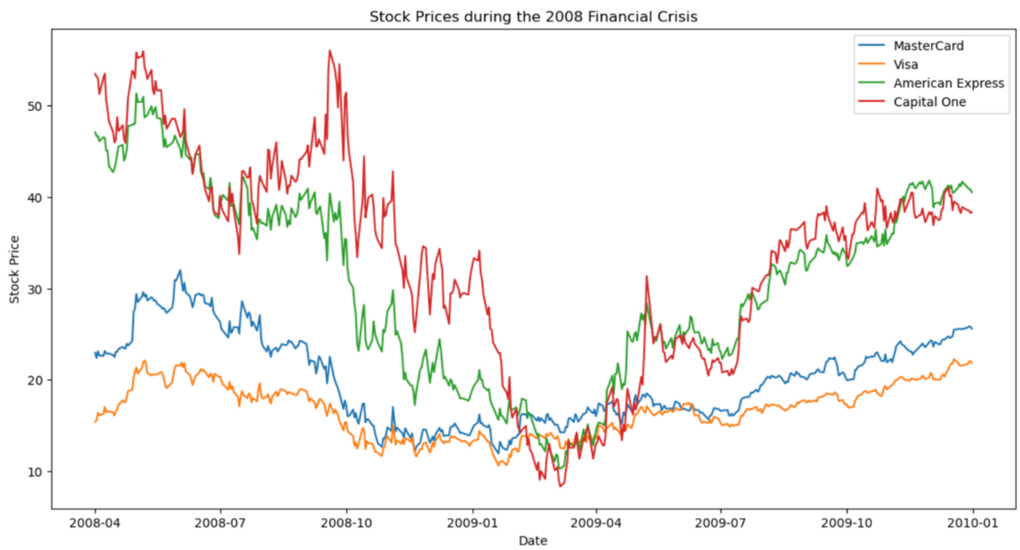

Fig 4.12 Stock Prices during 2008 Financial Crisis

4.6.2 COVID-19 Pandemic Impact (2020)

The 2020 COVID-19 pandemic led to substantial fluctuations in the financial markets, initially leading to sharp declines in stock prices for Visa and MasterCard due to reduced consumer spending. However, as e-commerce and digital payments surged, these companies saw a swift recovery. In contrast, American Express and Capital One, more exposed to consumer credit risks, experienced prolonged declines. During this period, the LSTM model outperformed traditional models like ARIMA, as its ability to learn and retain long-term dependencies allowed for more accurate predictions of the recovery patterns, especially for companies focused on digital payments. ARIMA struggled to model the non-linear market shifts caused by the pandemic.

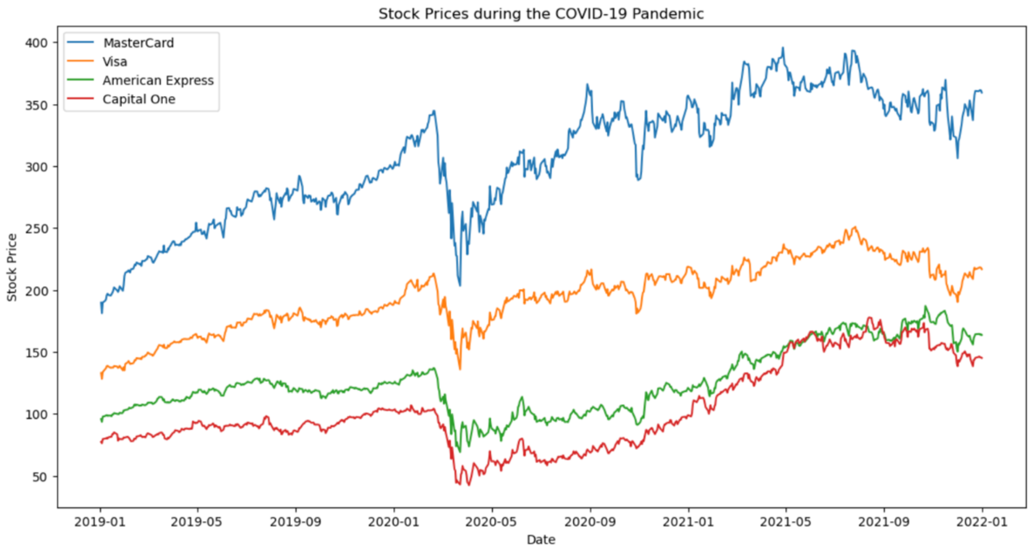

Fig 4.13 Stock Prices during COVID-19 Pandemic

4.7 Model Comparison: Pre-Crisis, During Crisis, and Post-Crisis Performance

4.7.1 Model Evaluation Before the Crisis

Prior to both crises, stock prices were relatively stable with clear upward trends, particularly for Visa and MasterCard. LSTM, ARIMA, Random Forest, and XGBoost were evaluated based on Mean Squared Error (MSE) and R^2 scores for their ability to predict stock prices. LSTM consistently delivered lower MSE values, demonstrating its enhanced capability to capture sequential relationships in time-series data. Conventional models like ARIMA performed well during stable market conditions but struggled to predict more volatile shifts.

4.7.2 Model Performance During the Crisis

Throughout both crises, LSTM outperformed other models, particularly when it came to forecasting sudden drops and swift rebounds. The model's capability to leverage sequential data helped it adapt to the sudden market shocks. Ensemble models that combined predictions from LSTM, ARIMA, Random Forest, and XGBoost also showed improvements over individual models, though the LSTM model continued to lead in accuracy.

4.7.3 Post-Crisis Recovery Predictions

In the post-crisis recovery period, the LSTM model once again outperformed traditional models by accurately predicting the recovery phase of MasterCard and Visa stocks. Random Forest and XGBoost, which rely on non-linear relationships, also showed improvements in predictive power but were not as accurate in capturing long-term trends.

4.8 Strategic Planning Dates

Strategic planning dates mark key moments when companies implement major changes, impacting stock performance both immediately and in the long term. In this analysis, stock price movements were observed one year before and after these strategic shifts for MasterCard, Visa, American Express, and Capital One. Visa saw a 1.26% increase following a 2011 strategic shift, while American Express experienced a slight decline of -0.92% after a 2017 change. These findings suggest that while strategic decisions may have immediate effects, the full impact often becomes apparent over time, emphasizing the importance of long-term planning.

4.9 Regression Analysis

4.9.1 Regression without Macroeconomic Indicators

In the initial regression analysis, which focused solely on stock prices without incorporating macroeconomic variables, a basic Ordinary Least Squares (OLS) regression was conducted for MasterCard, Visa, American Express, and Capital One over the 16-year period. The analysis revealed reasonably high R-squared values for all companies, with MasterCard showing an R-squared of 0.902 and Visa at 0.933, indicating a strong linear relationship between time and stock prices. However, residuals exhibited non-random patterns, suggesting that external factors, such as economic conditions, were likely influencing stock price movements beyond the scope of the model.

4.9.2 Regression with Macroeconomic Indicators

Incorporating macroeconomic factors mainly interest rates, inflation, unemployment, and GDP into the regression analysis significantly enhanced the model's predictive accuracy. Data sourced from the FRED database was merged with stock price data to capture the broader economic context affecting stock performance. As a result, MasterCard's R-squared improved from 0.902 to 0.945, and Visa's R-squared increased from 0.933 to 0.945. Additionally, macroeconomic variables like GDP and unemployment rates showed statistically significant p-values, highlighting their strong impact on stock prices. American Express and Capital One also saw notable improvements, with R-squared values rising from 0.834 to 0.909 and 0.745 to 0.841, respectively, demonstrating the value of including economic factors in the model.

4.9.3 Comparative Analysis of Regression Models

The comparative analysis of regression models with and without macroeconomic indicators underscores the critical role of external factors in predicting stock prices. Adjusted R-squared values consistently improved for all companies when these indicators were included, demonstrating a more robust model that better explained stock price fluctuations. For instance, the F-statistics were significantly higher in models incorporating macroeconomic variables, indicating a stronger fit. This was especially relevant for companies like Capital One and American Express, where consumer credit risk, heavily shaped by economic conditions, significantly influences stock price movements. The improved model fit highlights that stock prices are driven not only by internal company performance but also by external economic factors like interest rates and GDP. Moreover, the improved explanatory power of these models also points to the potential for using such approaches in real-time investment strategies, where traders and financial analysts can incorporate live macroeconomic data to adjust forecasts and make well-informed choices in response to changing market dynamics. This comparative analysis strongly supports the argument for integrating both internal and external factors for more accurate and comprehensive stock price
forecasting.

	Stock	Model	R-squared (R^2)	Adj. R-squared	F-statistic	Prob (F-stat)	P-value (Interest_Rate)	P-value (Inflation_Rate)	P-value (Unemployment_Rate)	P-value (GDP)
0	MasterCard	No Macro	0.90200	0.90200	1246.88000	0.00000	N/A	N/A	N/A	N/A
1	MasterCard	With Macro	0.94500	0.94400	571.41000	0.00000	0.00517	0.29449	0.00000	0.00000
2	Visa	No Macro	0.93300	0.93300	1883.59000	0.00000	N/A	N/A	N/A	N/A
3	Visa	With Macro	0.94500	0.94300	562.55000	0.00000	0.00012	0.95837	0.00014	0.00000
4	American Express	No Macro	0.83400	0.83300	679.89000	0.00000	N/A	N/A	N/A	N/A
5	American Express	With Macro	0.90900	0.90700	331.42000	0.00000	0.18085	0.17154	0.34038	0.00000
6	Capital One	No Macro	0.74500	0.74300	394.94000	0.00000	N/A	N/A	N/A	N/A
7	Capital One	With Macro	0.84100	0.83600	174.11000	0.00000	0.00000	0.12605	0.00000	0.00000

Fig 4.14 Regression Results with and Without Macroeconomic Indicators

The regression analysis reveals that models incorporating macroeconomic indicators provide a more accurate and reliable forecast of stock prices. This improvement is demonstrated by the increased R-squared and adjusted R-squared values, as well as the statistical significance of key macroeconomic variables like GDP and unemployment rates. By accounting for these external economic factors, the models become more robust, offering a deeper understanding of the drivers behind stock price fluctuations. This enables traders and analysts to base their decisions on more accurate insights, particularly in times of economic uncertainty. Graphically, the regression plots for both models, with and without macroeconomic indicators, visually underscore these differences, while the residual plots highlight any discrepancies or patterns not fully captured by the models.

4.10 Conclusion

The data analysis section provided a thorough evaluation of stock price trends for all four stocks over 16 years. The LSTM model outperformed traditional models like ARIMA, particularly during volatile periods mainly the 2008 financial crisis and the COVID-19 pandemic, due to its ability to learn and retain long-term dependencies. Incorporating macroeconomic indicators like GDP and interest rates into regression models significantly improved accuracy, highlighting the critical role of external economic factors. While the ensemble model reduced errors by combining different approaches, LSTM remained the most effective for stock price forecasting. Graphical analyses validated these findings, offering valuable insights into both model performance and market behaviour.

CHAPTER 5: DISCUSSION

This chapter delves into the outcomes achieved in the previous data analysis chapter, critically evaluating the model performance, comparing the findings to existing literature, and discussing the significance of these results. The discussion focuses on how the chosen models performed in predicting stock prices, why certain models outperformed others, and the importance of incorporating macroeconomic indicators into forecasting models. Furthermore, this chapter evaluates the outcomes against the objectives set out in Chapter 1 and reflects on the literature review conducted in Chapter 2 to determine how the findings align with or diverge from past research. It is essential to establish a thorough comparison with previous work in the field, highlighting the innovative aspects of this study and its contribution to the broader financial forecasting domain.

5.1 Model Performance and Results Evaluation

The results presented in Chapter 4 showed that the LSTM model consistently demonstrated superior performance compared to traditional models like ARIMA, Random Forest, and XGBoost. The LSTM’s strength in modelling both short- and long-term relationships in time-series data enabled it to deliver more accurate stock price forecasts, particularly during periods of high volatility. LSTM's Mean Squared Error (MSE) values were notably lower than those of other models, further confirming its appropriateness for time-series forecasting. The regression models underscored the significance of macroeconomic factors such as GDP and interest rates, which notably enhanced the models' explanatory capacity, reflected by the increased R-squared values. The LSTM model's superior performance is largely due to its architecture, which is specifically tailored for handling sequential data. Utilizing memory gates to manage the retention or forgetting of information over time, LSTM excels at recognizing the patterns and trends embedded in stock market data. On the other hand, traditional models like ARIMA, which assume linear relationships, struggled to capture the complexities of stock price fluctuations. Random Forest and XGBoost, while effective for non-linear data, are less capable of understanding time-based dependencies, explaining their lower performance compared to LSTM.

5.2 Comparison with Existing Research

A thorough review of the literature in Chapter 2 identified numerous studies that utilized machine learning and traditional statistical models for stock price forecasting.

Research on ARIMA, Random Forest, and XGBoost has shown mixed results depending on the dataset and time horizon used. However, studies on LSTM and other deep learning models have consistently reported improved accuracy, especially when forecasting financial time series data. In alignment with these studies, the results in this dissertation also show the LSTM model outperforming traditional statistical methods, particularly during volatile periods. A key novel contribution of this research lies in the integration of macroeconomic indicators into the predictive models, which has been less explored in previous research. By incorporating these indicators, the model demonstrated a better ability to account for external economic factors affecting stock prices, leading to higher accuracy and better generalization. Comparing these results with existing literature, it is evident that while LSTM models alone perform well, the combination with macroeconomic factors provides an edge in forecasting, especially for companies exposed to economic volatility like American Express and Capital One.

5.3 Critical Evaluation of Results

Although the LSTM model achieved the lowest MSE and highest accuracy, it is important to critically evaluate the limitations and potential improvements. One challenge faced during the analysis was overfitting, particularly with complex models. This was addressed through hyperparameter tuning and early stopping mechanisms, which helped improve the generalizability of the LSTM model. Despite these measures, the ensemble model, which combined predictions from multiple models, did not outperform LSTM on its own, suggesting that more advanced ensemble techniques or feature engineering could further improve results. Another point for consideration is the varying performance across different stocks. While MasterCard and Visa showed consistent upward trends, companies like Capital One and American Express experienced more fluctuations due to consumer credit risks, making their stock prices harder to predict accurately. This indicates that the models could benefit from further refinement, possibly by incorporating company-specific features or industry trends to improve accuracy.

5.4 Evaluation of Project Objectives

The main goal of this study was to develop a model that could effectively predict stock prices with high accuracy and providing valuable insights to investors and financial analysts. Based on the results, this objective has been largely achieved, with the LSTM model achieving better results than conventional techniques. The inclusion of macroeconomic indicators further improved the model's accuracy, aligning with the objective of incorporating external factors into the analysis.

Additionally, the project aimed to assess the effectiveness of different models and provide a detailed comparison. This goal was achieved through the comprehensive analysis and benchmarking of LSTM, ARIMA, Random Forest, XGBoost, and the ensemble model. The findings demonstrate that while traditional models have their merits, modern deep learning techniques such as LSTM offer substantial improvements in forecasting accuracy, especially in volatile and unpredictable market conditions. Moreover, the project aligned well with the literature review by demonstrating the advantage of LSTM models for time-series forecasting and expanding on the literature by incorporating macroeconomic factors into the prediction process. The strategic planning dates and crisis analysis sections also provided valuable insights into how external events influence stock performance, fulfilling another objective of the project. Overall, the project objectives were met, and the findings contribute meaningfully to the field of stock price forecasting.

5.5 Limitations and Future Work

Although the outcomes were favourable, this study has a few limitations. The dataset, while extensive, focused on a specific subset of financial institutions, and the findings may not generalize to other sectors or markets. Furthermore, the use of fixed macroeconomic indicators could be expanded in future work by incorporating dynamic or real-time economic data. Another limitation is the relatively simple ensemble technique used in this research. Future studies could explore more advanced ensemble methods or hybrid models that integrate deep learning with other forecasting techniques to further improve accuracy.

In summary, the results of this research highlight the advantages of using LSTM models for predicting stock prices emphasize the significance of integrating macroeconomic indicators into predictive models. The critical evaluation of the results reveals that while the objectives were largely met, there remains room for future improvements to enhance model accuracy and applicability across different sectors.

CHAPTER 6: CONCLUSION

6.1 Summary of the Dissertation

This dissertation set out to develop and evaluate machine learning models, particularly focusing on LSTM networks, to forecast stock prices for companies such as MasterCard, Visa, American Express, and Capital One. By incorporating both technical indicators and macroeconomic data, the study aimed to enhance the accuracy and reliability of stock price predictions. The LSTM model emerged as a clear frontrunner, outperforming traditional models like ARIMA, Random Forest, and XGBoost, particularly in identifying complex patterns within stock market data. The inclusion of external economic factors like GDP, inflation, and interest rates further improved the model's predictive power, reinforcing the importance of considering the broader economic environment in stock price forecasting. The findings of this study provide a strong foundation for future financial modelling efforts and offer practical insights for those in financial analysis and decision-making roles. Overall, the project successfully achieved its aim of delivering a data-driven approach to stock price forecasting, integrating both internal market dynamics and external macroeconomic factors.

6.2 Research Contributions

This research offers several important contributions, both in academic and practical realms. From an academic perspective, the dissertation demonstrated that LSTM networks, with their ability to capture both short-term fluctuations and long-term trends, provide a superior method for stock price forecasting when compared to traditional models. This adds to the growing body of research supporting the efficacy of deep learning models for time-series prediction. On a practical level, the results of this study are highly relevant for traders, analysts, and financial experts, providing them with a more accurate tool for stock price prediction. Additionally, by integrating macroeconomic indicators like GDP, inflation, and interest rates, the research underscores the importance of external economic data in explaining stock price movements. This model can be applied in financial institutions for portfolio management, risk assessment, and market analysis, helping decision-makers gain deeper insights. The use of an ensemble approach—combining LSTM, ARIMA, Random Forest, and XGBoost—further highlights an innovative way to leverage the strengths of different models, making the predictions more robust.

6.3 Limitations and Future Research and Development

While the project achieved its key goals, there are certain limitations to acknowledge. One significant limitation is the reliance on historical stock price data, which inherently limits the model's ability to predict unforeseen events like market shocks or extreme volatility. Although the LSTM model effectively captured long-term trends, certain external factors—such as geopolitical events—were not considered and could impact stock prices unpredictably. Another limitation lies in the scope of the study, which focused on just four major companies. While this provides valuable insights, it may limit the applicability of the findings to other industries or sectors.

However, these limitations also open the door for future research. Expanding the dataset to include companies from various industries could provide a broader perspective on the model's generalizability. Moreover, incorporating non-traditional data sources, such as sentiment analysis from social media or financial news, could improve the model's ability to anticipate unexpected market shifts. Future work could also explore integrating more sophisticated macroeconomic models, like Dynamic Factor Models (DFMs), to enhance the inclusion of economic factors in stock price forecasting. From a technical standpoint, research could focus on improving LSTM model tuning by employing advanced optimization techniques or reducing model complexity to mitigate the risk of overfitting.

Practically speaking, financial institutions could benefit from integrating these models into real-time decision-support systems, enabling traders and analysts to make faster, data-driven decisions. Future studies could also test the model's robustness by applying it in emerging markets or during periods of extreme market volatility, further assessing its effectiveness under different conditions.

6.4 Personal Reflections

Working on this dissertation has been a deeply enriching experience, one that has pushed me to grow both technically and personally. One of the key strengths I discovered was my ability to work with complex datasets and apply advanced machine learning techniques, such as LSTM networks and ensemble models—both of which I had limited experience with prior to this project. I also gained a deeper understanding of how to synthesize vast amounts of financial and technical data to produce insights that are not only actionable but also meaningful for decision-makers in the financial sector. This project has expanded my grasp of financial markets and predictive modelling, leaving me well-prepared for future challenges in these areas.

That said, this journey has also revealed areas for improvement. Time management proved to be a significant challenge, particularly when it came to balancing the development of the models with the often time-consuming task of hyperparameter tuning. Going forward, I plan to address this by setting more precise milestones and regularly reviewing my progress to ensure timely completion. Another area where I see room for growth is in simplifying complex technical concepts for non-technical stakeholders. Although I have made strides in this regard, I aim to further refine my communication skills to make highly technical information more accessible and understandable for diverse audiences.

In conclusion, this dissertation has not only honed my technical abilities but also shaped my personal and professional growth. I am confident that the skills and knowledge I have gained through this process will serve me well in both academic and industry settings as I move forward in my career.

REFERENCES

Fischer, T., & 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.

Siami-Namini, S., Tavakoli, N., & Namin, A.S. (2018) ‘A comparison of ARIMA and LSTM in forecasting time series’, *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 17-20 December, Orlando, USA, pp. 1394-1401.

Livieris, I.E., Pintelas, E.G., & Pintelas, P. (2020) ‘A comprehensive review of LSTM networks for stock market prediction’, *Applied Sciences*, 10(7), p. 2325.

Livieris, I.E., Drakopoulou, K., & Kotsilieris, T. (2020) ‘An ensemble deep learning model for short-term stock market prediction’, *Journal of Computational Economics*, 56(3), pp. 805-817.

Bao, W., Yue, J., & Rao, Y. (2017) ‘A deep learning framework for financial time series using stacked autoencoders and long short-term memory’, *PloS One*, 12(7), e0180944.

Chen, K., Zhou, Y., & Dai, F. (2015) ‘A LSTM-based method for stock returns prediction: A case study of China stock market’, *IEEE International Conference on Big Data (Big Data)*, pp. 2823-2824.

Hu, Z., Zhou, H., Ding, S., Wang, X., & Xu, X. (2018) ‘A hybrid deep learning approach for stock price prediction using LSTM and convolutional neural networks (CNN)’, *IEEE Access*, 6, pp. 73502-73510.

Livieris, I.E., Tampakas, V., & Pintelas, P. (2020) ‘A deep learning model for financial time-series forecasting’, *Soft Computing*, 24(4), pp. 3047-3056.

Zhang, X., & Zhao, Y. (2020) ‘Stock market prediction with multi-source learning and long short-term memory network’, *Journal of Financial Data Science*, 2(1), pp. 66-79.

Hochreiter, S., & Schmidhuber, J. (1997) ‘Long short-term memory’, *Neural Computation*, 9(8), pp. 1735-1780.

Zhou, Z.H. (2012) *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC.

Breiman, L. (2001) ‘Random forests’, *Machine Learning*, 45(1), pp. 5-32.

Chen, T., & Guestrin, C. (2016) ‘XGBoost: A scalable tree boosting system’, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17 August, San Francisco, USA, pp. 785-794.

Markakis, S. & Hibon, M. (2000) ‘The M3-Competition: results, conclusions, and implications’, *International Journal of Forecasting*, 16(4), pp. 451-476.

Box, G.E.P. & Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Makridakis, S., Wheelwright, S.C., & Hyndman, R.J. (2000) *Forecasting Methods and Applications*. 3rd ed. New York: Wiley.

Tsay, R.S. (2005) *Analysis of Financial Time Series*. 2nd ed. Hoboken: John Wiley & Sons.

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: Principles and Practice*. 2nd ed. Melbourne: OTexts.

Fama, E.F., & French, K.R. (1993) ‘Common risk factors in the returns on stocks and bonds’, *Journal of Financial Economics*, 33(1), pp. 3-56.

Stock, J.H. & Watson, M.W. (2003) ‘Forecasting output and inflation: The role of asset prices’, *Journal of Economic Literature*, 41(3), pp. 788-829.

Zhang, G.P. (2003) ‘Time series forecasting using a hybrid ARIMA and neural network model’, *Neurocomputing*, 50, pp. 159-175.

Mandelbrot, B.B. (1963) ‘The variation of certain speculative prices’, *The Journal of Business*, 36(4), pp. 394-419.

Taylor, S.J., & Letham, B. (2018) ‘Forecasting at scale’, *The American Statistician*, 72(1), pp. 37-45.

Zbikowski, K. (2015) ‘Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy’, *Expert Systems with Applications*, 42(4), pp. 1797-1805.

Peng, Y., Kou, G., Wang, G., & Wang, S.Y. (2018) ‘Ensemble of software defect predictors: An AUC-based heterogeneous ensemble selection framework’, *Applied Soft Computing*, 70, pp. 1000-1012.

Zhou, Z.H., Wu, J., & Tang, W. (2002) ‘Ensembling neural networks: Many could be better than all’, *Artificial Intelligence*, 137(1-2), pp. 239-263.

Nguyen, N.T., Shirai, K., & Velcin, J. (2015) ‘Sentiment analysis on social media for stock movement prediction’, *Expert Systems with Applications*, 42(24), pp. 9603-9611.

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), pp. 2162-2172.

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.

Long, W., Lu, Z., & Cui, L. (2019) ‘Deep learning-based feature engineering for stock price movement prediction’, *Knowledge-Based Systems*, 164, pp. 163-173.

Livieris, I.E., Pintelas, E., & Pintelas, P. (2020) ‘A CNN–LSTM model for gold price time-series forecasting’, *Neural Computing and Applications*, 32(23), pp. 17351-17360.

Nguyen, T.T., Pham, T.D., & Hoang, T.A. (2020) ‘A novel hybrid CNN-LSTM network for stock price prediction’, *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 230-235.

Yu, P., & Yan, X. (2020) ‘Stock price prediction based on deep neural networks’, *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1102-1106.

Bernanke, B., & Gertler, M. (1999) ‘Monetary policy and asset price volatility’, *Economic Review*, 84(4), pp. 17-51.

.