Comparative Analysis of ARIMA and LSTM Models for Predicting Stock Prices in the Spanish Banking Sector

By
Miloš Trifunović

A Capstone Project Report Submitted to AGE and GMU- Guglielmo Marconi University in Partial Fulfillment of the Requirements for the Degree of Master in Data Science

January, 2024

Student ID: - 22238647
Supervisor Name: - Shanthi Iyer
Date: - 24.01.2024.

uniathêna
ATHENA GLOBAL EDUCATION

Unimarconi
LA PRIMA UNIVERSITÀ DIGITALE ITALIANA

Table of Contents

ACRONYMS:

1. **ARIMA (Autoregressive Integrated Moving Average)**: A statistical analysis model used for time series forecasting. It combines autoregressive (AR) and moving average (MA) components and integrates differencing to make the time series stationary.
2. **LSTM (Long Short-Term Memory)**: A type of recurrent neural network (RNN) used in deep learning, particularly effective in predicting time series data. It is capable of learning long-term dependencies in data sequences.
3. **RSI (Relative Strength Index)**: A momentum oscillator used in technical analysis that measures the speed and change of price movements, often used to identify overbought or oversold conditions in a stock.
4. **EDA (Exploratory Data Analysis)**: An approach in data analysis that employs various techniques, including graphical representations, to maximize insight into a dataset, uncover underlying structures, extract important variables, and detect outliers and anomalies.
5. **MAPE (Mean Absolute Percentage Error)**: A measure of prediction accuracy in a forecasting model, calculated as the average of the absolute percentage errors of the predictions.
6. **MSE (Mean Squared Error)**: A risk metric corresponding to the average squared difference between the estimated values and the actual value.
7. **RMSE (Root Mean Squared Error)**: A standard way to measure the error of a model in predicting quantitative data. It is the square root of the MSE.
8. **ACF (Autocorrelation Function)**: A mathematical tool used to measure and identify the extent to which a time series is linearly related to a lagged version of itself.
9. **PACF (Partial Autocorrelation Function)**: Similar to ACF, but it shows the pure correlation of a time series with its lag, excluding the correlation contributions from intermediate lags.

LIST OF TABLES:

LIST OF FIGURES:

ABSTRACT

This research presents a comparative analysis of Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) models in predicting stock prices in the Spanish banking sector. Focused on major banks like Banco Santander, Banco Bilbao Vizcaya Argentaria, and Caixabank, the study utilizes a decade of historical stock data. It explores the efficiency and applicability of both models, integrating the Relative Strength Index (RSI) with LSTM for potentially enhanced forecast accuracy. The research aims to tailor these models to the unique dynamics of the Spanish banking sector, contributing significantly to financial econometrics and predictive analytics. Through meticulous model building, optimization, and validation, it addresses the challenges in accurately forecasting stock prices, providing insights for investors and policymakers in this economically crucial sector.

Keywords: ARIMA Model, LSTM Network, Stock Price Prediction, Spanish Banking Sector, Financial Econometrics, Time Series Analysis, Machine Learning in Finance, Predictive Analytics, Model Comparison, Relative Strength Index (RSI)

CHAPTER 1

1.0 INTRODUCTION

1.1 Introduction

Predicting stock prices is a crucial intersection of economics, mathematics, and technology. This research focuses on the Spanish banking sector, an important component of global finance, to compare two prominent predictive models: Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks. The use of these models in the Spanish banking context has not been explored yet, presenting a unique research avenue.

Accurate stock predictions are vital for investment strategies and economic stability. This study aims to refine the understanding of ARIMA and LSTM models in forecasting stock prices and contribute to the field of financial econometrics, emphasizing prediction reliability.

The core of this research is a detailed analysis of ARIMA and LSTM models in the Spanish banking environment. It introduces an innovative approach by integrating the Relative Strength Index (RSI) with the LSTM model, examining its impact on forecast accuracy. The objective is to determine the most effective model, parameters, and features for accurate stock price prediction in this sector.

While ARIMA is known for short-term forecasting, LSTM excels in capturing long-term patterns. Using historical data from major Spanish banks, this study seeks to provide new insights and fill the gap in existing literature. The research unfolds through a comprehensive literature review, a meticulous methodology, an extensive performance analysis, and concludes with a synthesis of findings and future research directions. This approach ensures a thorough examination of these models and their practical implications in the Spanish banking context.

1.2 Background

The development of sophisticated predictive models has revolutionized financial market analytics, particularly in forecasting stock prices. The Spanish banking sector, crucial to Spain's economy, faces distinct challenges and opportunities in this area, underscored by recent market fluctuations and the intricacies of financial instruments. This highlights the need for precise and reliable predictive models specifically tailored to this sector. Despite progress in financial analytics, there is a notable lack of focused comparative studies of predictive models in the context of Spanish bank stocks.

This research, titled 'Comparative Analysis of ARIMA and LSTM Models for Predicting Stock Prices in the Spanish Banking Sector,' aims to address this gap. It critically assesses and compares the accuracy of ARIMA (Autoregressive Integrated Moving Average) and LSTM (Long Short-Term Memory) models in predicting the stock prices of major Spanish banks. The objective is to identify which model, including its parameters and features, provides the most accurate predictions, offering crucial insights for investment strategy development and advancing the field of financial predictive analytics.

Furthermore, this study enhances the broader understanding of financial predictive analytics, offering practical insights for investors, financial institutions, and the evolving fintech sector. It aims to refine investment decision-making processes and encourage innovation in financial technology applications.

This research delves into ARIMA and LSTM models, each unique in feature utilization and efficiency. ARIMA, grounded in traditional statistical analysis, demonstrates robustness with limited input features, particularly historical price data (Adebiyi, Adewumi, and Ayo, 2014; Mondal, Shit, and Goswami, 2014). In contrast, LSTM, emerging from deep learning, leverages a broader range of input features to discern complex, non-linear patterns in sequential data like stock prices (Choi and Choi, 2023; Ma, 2020). This distinction in feature utilization is vital, considering the dynamic nature of the Spanish banking sector.

The study also addresses the nuances of model optimization, crucial in the unique market dynamics of the Spanish banking sector. Precise parameter selection and tuning are critical for accurate stock price predictions. The research aims to tailor model selection and optimization to the specific traits of the Spanish banking sector, thereby enhancing the accuracy and practicality of the predictive analysis.

1.3 Problem Statement

Predicting stock prices in the Spanish banking sector, a key part of Spain's economy and a mirror of global market complexities presents a significant challenge in financial analysis. Despite advancements in financial analytics, there is a noticeable gap in the literature: the absence of comparative studies focused on stock price prediction within this sector. This gap is particularly crucial given the sector's dynamic nature, which is shaped by economic policies, market sentiment, and global trends and requires more advanced forecasting methods than traditional models provide. The lack of such studies specifically tailored to the Spanish banking context hinders our understanding and impedes the development of effective predictive models. While widely used, current forecasting approaches often fail to capture modern financial markets' complex, non-linear dynamics. This issue is particularly acute in the Spanish banking industry, where unique market conditions call for more sophisticated analytical techniques.

This research aims to address this critical gap by conducting a thorough comparative analysis of two advanced models: Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks. The study focuses on assessing their effectiveness in the distinct environment of the Spanish banking sector. By bridging this gap in predictive analytics, the research seeks to provide vital insights into the applicability and efficiency of these advanced modeling techniques. The ultimate objective is to advance the field of financial econometrics, offering more robust tools for economic analysis and decision-making in a data-driven age, especially for a sector with significant national and global economic impact.

1.4 Objectives

The primary aim of this capstone project is to address the challenges outlined in the problem statement, focusing on improving stock price prediction models within the Spanish banking sector. The objectives are structured to be measurable and achievable, aligning with the project's overall scope. These specific objectives include:

1. **Comparative Analysis of Predictive Models**: Conduct an in-depth analysis of the Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) models to evaluate their effectiveness in predicting stock prices of major Spanish banks.
2. **Enhancing Forecasting Accuracy**: Investigate the integration of the Relative Strength Index (RSI) with the LSTM model, assessing its impact on forecast precision.

3. **Optimization of Model Parameters**: Analyze and fine-tune key parameters in both ARIMA and LSTM models, customizing them to fit the unique characteristics of the Spanish banking sector.
4. **Validation and Practical Applicability**: Test these models using current market data from leading Spanish banks to ensure their practicality and reliability in predicting real-world market trends and stock prices.
5. **Contributions to Financial Econometrics**: Provide valuable insights into the field of financial econometrics, focusing on the practical and theoretical applications of ARIMA and LSTM models.

By achieving these objectives, the project aims to surmount the forecasting challenges in the Spanish banking sector and enrich the field of financial predictive analytics. The expected outcomes are to offer valuable tools and insights for investors, analysts, and policymakers in their decision-making processes.

1.5 Proposed Solution

This capstone project proposes a structured methodology to improve stock price prediction in the Spanish banking sector using advanced forecasting models: Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks. The proposed solution involves these steps:

1. **Exploratory Data Analysis (EDA)**: Acquire and preprocess historical stock data from major Spanish banks, emphasizing data quality and normalization. Focus on analyzing the 'Close' price to discern trends, seasonal patterns, and volatility using various graphical methods.
2. **Model Building**:
    - Develop an ARIMA model, capitalizing on its strengths in time-series analysis with an emphasis on trend and seasonality. Optimize parameters (p, d, q) for peak performance.
    - Construct an LSTM model using the 'Close' price as the primary input, with additional exploration of the Relative Strength Index (RSI) to evaluate its influence on forecasting accuracy.

3. **Optimization**: Refine the parameters of both ARIMA and LSTM models to maximize predictive accuracy. Assess LSTM's effectiveness with both single-feature and dual-feature (including RSI) configurations.
4. **Validation and Testing**: Apply an 80-20 split for training and testing data. Evaluate model performance using metrics like Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE).
5. **Ethics, Reliability, and Validity Measures**:
   - Adhere to ethical standards, particularly in data privacy.
   - Test model reliability across various time frames and market conditions.
   - Confirm model validity through comparisons with actual market data.
6. **Final Analysis and Reporting**: Perform a concluding comparative analysis of ARIMA and LSTM models to determine their suitability for the Spanish banking sector. Prepare a detailed report on the findings and potential areas for future research.

It is essential to recognize the inherent limitations and challenges in this research. The effectiveness of the models could vary with changing market conditions, reflecting the dynamic nature of financial markets. This variability and reliance on historical data introduce a layer of complexity. Historical data, while informative, may not always predict future market trends accurately.

Data quality and availability are also pivotal; any limitations in these areas could impact the models' performance. Furthermore, the computational intensity and lower transparency of LSTM models, when compared to more straightforward models like ARIMA, are important to note. These aspects underscore the need for a balanced interpretation of the results and their real-world applicability.

By acknowledging these challenges, the research aims to provide more accurate and reliable stock price predictions, guiding stakeholders in informed decision-making. It seeks to determine the most suitable model for various market scenarios while being mindful of these limitations. This balanced approach aims to deepen understanding and pave the way for future enhancements in financial predictive analytics.

1.6 Research Questions

To steer this investigation into the effectiveness of ARIMA and LSTM models for stock price prediction in the Spanish banking sector, with a particular focus on large-cap banks, the following research questions are formulated. These questions are crucial for delving into the nuances of predictive modeling, scrutinizing the impact of feature selection, and identifying the optimal model parameters for enhanced accuracy:

1. **How do ARIMA and LSTM models perform in predicting the stock prices of large-cap Spanish banks?**
   - This question aims to assess and compare the effectiveness of ARIMA and LSTM models in the Spanish banking sector, specifically focusing on their predictive accuracy for stock prices.
2. **How does the use of closing price as the sole feature in the ARIMA model impact its predictive accuracy compared to the broader feature set in the LSTM model?**
   - This inquiry explores the effect of using just the closing price in ARIMA versus a more comprehensive range of features in LSTM. It assesses how feature selection influences the forecasting precision of these models in the context of large-cap Spanish banks.
3. **What are the optimal parameter settings for ARIMA and LSTM models to maximize their predictive accuracy for large-cap Spanish bank stocks?**
   - This question concentrates on identifying and fine-tuning key parameters in both models to enhance their accuracy in forecasting stock prices. It explores the most effective configurations for optimal forecasting precision.

These research questions are fundamental in guiding the methodological approach and analytical process of this study. They address the intricacies of predictive modeling in finance and aim to make a significant contribution to the field of financial econometrics, particularly in the realm of stock price prediction for large-cap Spanish banks.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Introduction

In the evolving field of financial econometrics, particularly in stock price prediction, there has been a significant shift towards integrating traditional statistical methods with advanced machine learning techniques. This Literature Review is dedicated to examining two central models – the Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks – specifically in their application to the dynamic Spanish banking sector, known for its economic significance and fluctuating market conditions.

The review undertakes a thorough analysis of the ARIMA and LSTM models. ARIMA has established its reputation for efficient short-term forecasting and handling non-stationary data, while LSTM is recognized for its ability to model non-linear relationships in sequential data, an essential aspect of financial time series. The review aims to evaluate the strengths, limitations, and adaptability of these models under specific market conditions, with a particular focus on large-cap Spanish banks.

This review also uniquely applies ARIMA and LSTM models to the Spanish banking sector, addressing a notable gap in existing research. The sector's economic importance and distinctive forecasting challenges underscore the urgency of this targeted study, which could establish new standards in financial econometrics.

A vital part of this analysis is investigating how technical indicators, such as the Relative Strength Index (RSI), can enhance the predictive capabilities of LSTM models. This aspect reflects the dynamic nature of financial modeling, where incorporating various data types and analytical approaches is increasingly vital.

The following sections will provide an in-depth comparative analysis of ARIMA and LSTM models. This includes evaluating their performance, understanding the impact of feature selection and model optimization, and assessing their suitability for the specific requirements of the Spanish banking sector. Moreover, the review will highlight emerging trends and future directions in financial predictive modeling. This encompasses examining hybrid models and integrating diverse data types, suggesting a potential shift in methodologies for predicting stock prices.

2.2 Historical Insights and Comparative Analysis of ARIMA and LSTM Models

The evolution of financial predictive models, particularly ARIMA (Autoregressive Integrated Moving Average) and LSTM (Long Short-Term Memory) networks, mirrors the growing complexity of financial markets. Recent studies offer diverse assessments of these models' effectiveness in stock price forecasting.

Kobiela et al. (2022) explored NASDAQ stock data, employing Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE) for evaluation. Their findings underscored ARIMA's proficiency in longer-term forecasts, especially when using historical price values as the sole feature. ARIMA outperformed LSTM in extended prediction windows, such as thirty days and three months, suggesting its suitability in scenarios with limited data features and longer forecasting horizons.

Conversely, Alzheeva and Kochkarov (2020) found a distinct advantage in using LSTM models for Russian stock analysis, where they achieved a 65% reduction in RMSE errors compared to ARIMA. This highlights LSTM's enhanced capability in managing complex, non-linear time series data. This observation aligns with the findings of Siami-Namini, Tavakoli, and Namin (2018), who noted LSTM's significantly lower error rates over ARIMA. Similarly, Ashok and Prathibhamol (2021) reported superior performance from LSTM models.

These contrasting results underscore the context-dependent effectiveness of predictive models in stock price forecasting. ARIMA shows strengths in long-term forecasts with limited data inputs, while LSTM is more adept at handling environments rich in complex, non-linear data. This variability, influenced by the nature of the financial data and the specific market context, is particularly pertinent to the dynamic Spanish banking sector.

These historical insights highlight the necessity of a tailored approach in model selection, considering dataset characteristics and the intended prediction timeframe. This strategy is especially critical in the Spanish banking sector, where market specificity and data intricacies are key factors. Considering the unique Spanish banking context, a nuanced understanding of these models' performance is essential for precise and effective forecasting.

2.3 Feature Utilization and Model Efficiency in ARIMA and LSTM

This subchapter explores how ARIMA and LSTM models differ in their approach to feature utilization and model efficiency in stock price forecasting, further distinguishing their respective applications.

The ARIMA model, rooted in traditional statistical analysis, shows robust efficiency in scenarios with limited data inputs. It typically relies on a single feature, usually historical price values, making its approach streamlined yet effective for longer-term predictions. ARIMA excels in processing time-series data, with a particular focus on the historical progression of stock prices. Kobiela et al. (2022) highlighted ARIMA's ability to leverage its strengths in time-series analysis effectively, even with restricted data features, rendering it a viable tool for certain forecasting scenarios.

On the other hand, the LSTM model, derived from deep learning, requires a more extensive range of input features for optimal training and performance (Roondiwala, Patel, and Varma, 2017). To improve predictability, incorporating technical indicators like the Relative Strength Index (RSI) has been proposed. The significance of such indicators in stock price forecasting has been emphasized by researchers such as Agrawal, Khan, and Shukla (2019) and Htun, Biehl, and Petkov (2023). Additionally, Wong et al. (2002) demonstrated the efficacy of these indicators in generating positive returns in certain markets. The ability of LSTM to concurrently analyze multiple variables provides a nuanced understanding of market dynamics, making it particularly suitable for situations where a variety of factors influence the forecasting outcome.

2.4 Model Building and Optimization

The construction and optimization of ARIMA and LSTM models are fundamental in the field of time series forecasting, such as stock price prediction. This section delves into the methodologies and optimization techniques for these models as presented in various academic studies.

2.4.1 ARIMA Model Building and Optimization

Creating an ARIMA (Autoregressive Integrated Moving Average) model is a process that starts with the careful identification of its parameters: p (autoregression), d (differencing), and q (moving average). The initial estimation of these parameters often relies on analyzing the data's time plot, as well as using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). These analytical tools are instrumental in identifying preliminary values, thereby guiding the structural formation of the model (Siami-Namini, Tavakoli, and Namin, 2018; Ariyo, Adewumi and Ayo, 2014; Ma, 2020).

In optimizing ARIMA models, the Akaike Information Criterion (AIC) plays a crucial role. It helps balance the model's complexity with its ability to fit the data effectively. A lower AIC value is indicative of a model that aptly captures data patterns without being overly complex. To achieve this balance, various combinations of (p, d, q) are tested, with the selection of the optimal model based on the minimum AIC value (Mondal, Shit, and Goswami, 2014).

An additional step in the validation of the ARIMA model includes using the Ljung-Box test. This test is crucial for checking the presence of autocorrelation in the residuals of the model. The absence of significant autocorrelation is a desirable property, as it suggests that the model has successfully captured the underlying patterns in the data without leaving any systematic structure unaccounted for.

2.4.2 LSTM Model Building and Optimization

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, require the tuning of several hyperparameters for effective model building. The initial and crucial step is determining the number of memory units or blocks, which significantly influences the model's ability to capture long-term dependencies in time series data. The number of memory units dictates the model's complexity and depth, impacting its predictive capability (Siami-Namini, Tavakoli, and Siami Namin, 2018).

Other essential hyperparameters in LSTM models include the number of layers, which adds depth to the model, and the dropout rate, used to prevent overfitting by randomly omitting a fraction of neuron activations. Furthermore, the choice of optimizer (such as ADAM) and loss function (like MSE) during training is critical. These parameters are optimized by minimizing the difference between predicted and actual values in the training set. The model's performance is evaluated using metrics like RMSE or MSE, with the ideal model configuration achieving the lowest error rates. Different combinations of hyperparameters, like the number of epochs and learning rate, are experimented with to identify the most effective model setup.

2.5 Methodological Gaps in Forecasting Spanish Banking Stocks

The application of ARIMA and LSTM models in financial predictive modeling, particularly within the Spanish banking sector, reveals distinct methodological gaps that call for a more specialized approach. These gaps broadly encompass issues of data relevance, model applicability, the absence of sector-specific analysis, and inconsistent results.

A primary limitation in current methodologies is the reliance on models that used datasets from markets like the American or Russian stock exchanges, which significantly differ from the Spanish banking context. Such models often fail to accurately represent the unique economic indicators, regulatory frameworks, and market behaviors typical of the Spanish banking sector. This discrepancy raises serious questions about the accuracy and relevance of directly applying findings from these models to the Spanish environment. Additionally, this gap highlights the need for models that more accurately reflect the specific conditions of the Spanish banking sector, ensuring the reliability and applicability of predictive models.

Another critical issue is the scarcity of studies directly comparing ARIMA and LSTM models within the Spanish banking context. This gap limits the ability to assess their relative performance and suitability accurately. Many existing studies generalize their findings, overlooking the specific market dynamics of Spanish banks, thus diminishing the applicability and relevance of their conclusions to the Spanish banking sector.

Furthermore, the literature exhibits inconsistent findings regarding the superiority of either ARIMA or LSTM models. This inconsistency creates ambiguity in their application to the Spanish banking sector and hinders the development of a clear strategy for model selection and optimization tailored to the specific needs of Spanish banking stocks.

To bridge these methodological gaps effectively, research that is intricately aligned with the unique characteristics and challenges of the Spanish banking sector is crucial. Such research should include collecting and analyzing market data specific to Spain, ensuring a comprehensive understanding of the market's nuances and complexities. It is also imperative to thoroughly examine the architecture of the LSTM model, focusing on its adaptability and optimization to accurately reflect and forecast the dynamics of the Spanish banking environment. A systematic and critical comparison of ARIMA and LSTM models' performance within this context is equally essential. Undertaking focused and contextually tailored research can significantly enhance predictive models' accuracy, efficiency, and relevance in the Spanish banking sector. This advancement will contribute to the field of financial econometrics and empower more informed decision-making and strategic planning within this crucial economic domain.

2.6 Challenges in Stock Price Forecasting

Forecasting stock prices in the Spanish banking sector entails distinct challenges due to financial markets' inherent volatility and complexity. Ariyo, Adewumi, and Ayo (2014) highlight these aspects, pointing out the inherent difficulties in achieving accurate forecasts. Although the ARIMA model is effective for short-term forecasting, it faces challenges with the non-stationarity typical in financial data. This is particularly pertinent in a market influenced by Spain's economic dynamics and regulatory changes. The process of selecting parameters (p, d, q) in ARIMA adds another layer of complexity, especially when applied to intricate market patterns.

LSTM models are acknowledged for their ability to process complex, non-linear data patterns effectively. However, a practical challenge arises in ensuring their accuracy and adaptability to the specific nuances of the Spanish banking market. This includes careful consideration in selecting and processing input features, optimizing model parameters for this particular context, and accurately interpreting model outputs. Given the dynamic and often unpredictable nature of financial markets, achieving precision in these models is a formidable task. Emphasizing accuracy and interpretability is crucial for their practical application and utility in making informed predictions about Spanish banking stocks.

Furthermore, the issue of overfitting in machine learning models, including neural networks, poses a significant barrier to their ability to generalize effectively to new data sets. As noted by Toprak, Çağil, and Kökçam (2023) and Kobiela et al. (2022), selecting and optimizing these models demand considerable time and computational resources.

Data quality and relevance are also pivotal in determining the accuracy of predictive models. Given the dynamic nature of financial markets, regular updates to these models are necessary to ensure their continued relevance and accuracy amidst market changes, as Ma (2020) discussed.
Despite these challenges, there remains a notable gap in research specifically addressing stock price prediction in Spanish markets. This lack of targeted investigation limits the current understanding and presents significant opportunities for future research to explore this area more deeply.

2.7 Advanced Methodological Framework and Techniques

This research introduces a structured and innovative methodology aimed at enhancing the accuracy of stock price predictions within the Spanish banking sector, with a focus on leading institutions such as Banco Santander, Banco Bilbao Vizcaya Argentaria, and Caixabank. Leveraging a decade of historical stock data from Yahoo Finance, each step of the methodology is designed to address the research problem effectively and meet the objectives of the project.

Proposed Methodology:

- **Data Collection**: Collect data from Yahoo Finance for three major Spanish banks
- **Exploratory Data Analysis (EDA)**: The research begins with comprehensive EDA, prioritizing data integrity and key processes such as managing missing values. Techniques including trend analysis, seasonality identification through time series decomposition, and volatility assessment using standard deviation are utilized for an in-depth understanding of the market.
- **Model Building**: The study involves developing an ARIMA model, customized to address time-series complexities, and an LSTM model. ARIMA's parameters (p, d, q) are determined using ACF and PACF analysis. In parallel, an LSTM model is developed, with the initial focus on the "Close" price, with subsequent incorporation of the Relative Strength Index (RSI) to comprehensively evaluate the technical indicator’s impact.

- **Optimization**: Both the ARIMA and LSTM models undergo extensive optimization, including adjustments in components and architecture, to better align with the dynamics of the Spanish banking market.
- **Validation and Testing**: An 80-20 data split is used for validation purposes. The study employs metrics such as Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) to rigorously test the models' predictive capabilities.
- **Ethics, Reliability, and Validity Measures**: Ethical standards, particularly concerning data privacy, are strictly adhered to. Reliability and validity are systematically evaluated across different market conditions using quantitative analysis.
- **Final Report**: The research culminates in a comprehensive comparative analysis of the ARIMA and LSTM models, assessing their effectiveness and discussing implications for future research.

Python and Jupyter Notebook are extensively utilized throughout the research for their advanced data processing capabilities, enabling sophisticated data analyses and model optimization.

2.8 Advantages of Proposed Methodology

The methodology proposed in this research offers a structured and innovative approach, bringing several key advantages to the accuracy of stock price predictions within the Spanish banking sector. These advantages collectively enhance the effectiveness and impact of the project:

- **Comprehensive Exploratory Data Analysis (EDA)**: The methodology begins with an in-depth EDA using advanced techniques to analyze market trends and volatility. This thorough initial assessment is pivotal for accurately interpreting the complex financial dynamics of the Spanish banking market.
- **Customized Model Development**: The ARIMA and LSTM models are specifically tailored to address the unique time-series characteristics of the Spanish banking sector. Incorporating the Relative Strength Index (RSI) in the LSTM model aims to augment its predictive capacity, offering a more holistic view of market trends. This customization ensures the models are finely tuned to the nuances of the Spanish banking environment.

- **Extensive Model Optimization**: The rigorous optimization process applied to both models ensures their alignment with the distinctive market conditions in Spain. This step increases the models' reliability and applicability, making them more suited to the specific forecasting needs of the sector.
- **Rigorous Testing and Validation**: The use of metrics such as MSE, MAPE, and RMSE in testing these models serves to validate their predictive accuracy, providing a robust framework for evaluating their performance.
- **Ethical Standards and Technical Robustness**: Adherence to ethical considerations, particularly regarding data privacy, lends credibility to the research. Utilizing Python and Jupyter Notebook for data processing underpins the methodology's technical robustness, supporting advanced analysis and model optimization.
- **Filling a Research Gap and Providing a Versatile Framework**: This research addresses a significant gap in the study of the Spanish banking sector and offers a methodological framework that is adaptable to other financial markets. The methodology's versatility and technical proficiency render it a valuable asset for both academic research and practical industry application, ultimately enhancing decision-making and strategic planning in the field of financial econometrics.

2.9 Recent Developments and Future Directions in Financial Predictive Modeling

Financial predictive analytics is currently undergoing an exciting transformation, marked by the integration of machine learning techniques with traditional econometric models. A notable trend in this area is the growing interest in hybrid models that merge the strengths of ARIMA and LSTM, as highlighted by researchers such as Ma (2020) and Kobiela et al. (2022). These hybrid models represent a novel approach that boosts prediction accuracy by combining statistical methods with deep learning techniques.

One significant avenue for future research is the development of ARIMA-LSTM hybrid models. These models harness ARIMA's strength in linear trend analysis and LSTM's capability to process complex, non-linear data. This synergy creates a more encompassing and flexible tool for stock price forecasting, which is particularly pertinent in the volatile environment of the Spanish banking sector.

Furthermore, recent advancements in predictive models have begun to incorporate a variety of data types, including economic indicators, market sentiment, and social media insights, as observed by Swathi, Kasiviswanath, and Rao (2022). This expansion in data utilization acknowledges the multifaceted nature of financial markets and the need for a more holistic forecasting approach. Incorporating diverse data sources significantly bolsters the accuracy and robustness of predictions in the current era of abundant data.

This trend towards integrating multiple data types is in line with broader developments in data science, where amalgamating different analytical methods and data sources often yields more comprehensive and reliable outcomes. The future trajectory of financial predictive modeling appears to be continuing in this direction, with a focus on developing hybrid models and integrating a broader spectrum of data types. Such advancements are anticipated to lead to more precise and nuanced market forecasts, further enriching the field of financial econometrics.

CHAPTER THREE

3.0 Data Collection and Modelling Analysis

3.1 Methodology Overview: ARIMA and LSTM Model Application in Spanish Banking

This chapter outlines the methodology to address the significant challenges in accurately predicting stock prices in the Spanish banking sector and the gap in comparative studies. It focuses on a comprehensive analysis of two advanced predictive models: Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks, tailored to the unique dynamics of the Spanish banking sector influenced by economic policies, market sentiment, and global trends.

3.1.1 Tailored Model Development

The methodology includes developing ARIMA and LSTM models, each fine-tuned to the Spanish banking sector's specific requirements:

- **ARIMA Model Application**: Leveraging ARIMA for its proficiency in linear trend analysis and short-term forecasting, the model is calibrated to reflect the Spanish banking market's historical price patterns.
- **LSTM Model Deployment**: The LSTM model, recognized for handling complex, non-linear dynamics, is utilized to capture and forecast the intricate nature of stock prices in the sector.

3.1.2 Comprehensive Data Collection and Analysis

The research uses a decade of historical stock price data from major Spanish banks like Banco Santander, Banco Bilbao Vizcaya Argentaria, and Caixabank, sourced from Yahoo Finance. This dataset underpins the thorough exploration of market trends and forecasting potentials for both ARIMA and LSTM models.

3.1.3 Integration with Literature Review Insights

Insights from the literature review are incorporated to ensure that the models are grounded in theoretical and empirical findings, enhancing their applicability to the Spanish banking sector.

3.1.4 Addressing Data Specificities

The methodology also involves meticulous handling of data peculiarities, including missing values and anomalies, to maintain the data's integrity and the accuracy of the models.

3.1.5 Metrics and Algorithms Implementation

Various metrics and algorithms are implemented for optimizing and validating the ARIMA and LSTM models. Performance metrics such as MSE, MAPE, and RMSE are utilized to assess predictive accuracy.

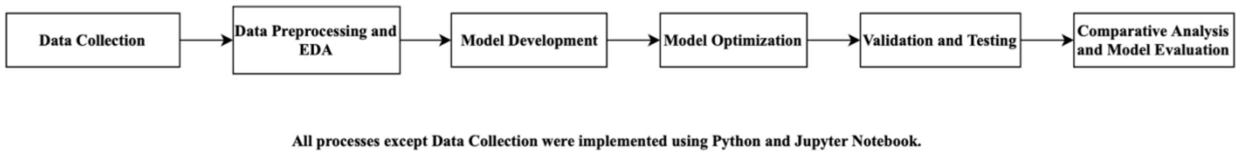

Figure 1: Research Methodology Overview

This methodology, aimed at developing and comparing ARIMA and LSTM models, addresses the research gap in predictive analytics for the Spanish banking sector. It seeks to provide insights into these models' applicability and efficiency, contributing to the field of financial econometrics and informed decision-making in this critical economic area.

3.2 Data Collection

The effectiveness of ARIMA and LSTM models in predicting stock prices in the Spanish banking sector hinges on a robust data collection process. Sourced exclusively from Yahoo Finance, this research utilizes a decade's worth of stock prices from three major Spanish banks, spanning from December 20, 2013, to December 20, 2023.

3.2.1 Data Sources and Scope

The datasets encompass 2560 observations for each of the three key Spanish banks, providing a decade-long perspective of market trends and cycles. This extensive timeframe ensures a comprehensive view for predictive modeling.

3.2.2 Data Features

Key features in the datasets include Date, Open, High, Low, Close, Adjusted Close, and Volume. Additionally, the Relative Strength Index (RSI) is calculated as an extra feature, offering further insights into market dynamics.

3.2.3 Data Collection Process

Data was efficiently downloaded from Yahoo Finance by selecting specific start and end dates. This method secured data that is accurate and relevant for the study. The data utilized in this study, being secondary public data, underwent a thorough evaluation to confirm its relevance for the research objectives. This type of data, already collected and available publicly, offers a practical and efficient means for comprehensive analysis in the context of the Spanish banking sector.

The subsequent phase involves a thorough Exploratory Data Analysis (EDA). This critical step ensures data integrity and includes managing missing values, conducting trend analysis, identifying seasonality through time series decomposition, and assessing volatility with standard deviation. The EDA provides deeper insights into market trends and anomalies, crucial for the effective application of the ARIMA and LSTM models.

3.3 Data Preprocessing and Exploratory Data Analysis (EDA)

In this research, a thorough data preprocessing and EDA are essential to ensure the accuracy and effectiveness of the ARIMA and LSTM models in forecasting stock prices in the Spanish banking sector. This subchapter elaborates on the comprehensive steps undertaken in this phase.

3.3.1 Managing Missing Values

The first step in data preprocessing involves addressing any missing values in the datasets. Missing data, if not handled correctly, can lead to biased predictions. Depending on the nature and extent of missing values, various techniques are employed, including forward filling, backward filling, using the mean of the available data, or linear interpolation methods. These techniques ensure the integrity and continuity of the time series data, which is crucial for accurate modeling.

3.3.2 Trend Analysis

For identifying trends within the stock price data, visual techniques such as line charts are used. These graphical representations provide a clear view of the stock price movements over time, helping to identify any long-term trends that might be present in the data. This visual trend analysis is an essential step in understanding the overall direction and behavior of the stock prices.

3.3.3 Seasonality Assessment

The data is also examined for seasonal patterns using methods like seasonal decomposition. This process helps in identifying and understanding any recurring patterns or cycles in the stock prices, which are common in financial time series data. Recognizing these patterns is vital for developing models that can accurately predict future stock prices, taking into account these seasonal influences.

3.3.4 Volatility Assessment

Volatility assessment is conducted by calculating statistical measures such as the standard deviation of stock prices. This assessment provides insights into the stability and risk associated with the stock prices, which is an important factor in financial modeling and risk management.

3.4 Feature Engineering

In addition to utilizing the historical "Close" price, the Relative Strength Index (RSI) is incorporated as an additional feature in the LSTM model. Calculated using its established formula, the RSI serves as a momentum indicator, instrumental in identifying market conditions that are overbought or oversold. The integration of RSI into the LSTM model is anticipated to significantly bolster the model’s capability to more accurately reflect market dynamics. This approach is supported by the findings of researchers such as Agrawal, Khan, and Shukla (2019) and Htun, Biehl, and Petkov (2023), who have highlighted the effectiveness of such indicators in predictive modeling. Furthermore, studies by Wong et al. (2002) have demonstrated the potential of these indicators to yield positive returns in certain market scenarios, underlining the practical value of incorporating RSI in forecasting models.

$$RS = \frac{Avg.Gain}{Avg.Loss}$$

$$RSI = 100 - \frac{100}{1 + RS}$$

Figure 2: RSI Formula

3.4.1 Data Scaling

Data scaling is particularly important for LSTM models, which are sensitive to the scale of input data. The Min-Max scaling method is employed to normalize the data. This scaling process ensures that all input features contribute equally to the model training, enhancing the model's ability to learn from the data effectively.

3.4.2 Data Splitting

Prior to model training, the dataset is divided into training and testing sets. An 80-20 split is used, where 80% of the data is allocated for training the model, and the remaining 20% is reserved for testing and evaluating the model's performance. This separation is crucial for assessing the model's generalization ability and its effectiveness in predicting unseen data.

This comprehensive approach to data preprocessing and EDA lays a strong foundation for the subsequent modeling and analysis, ensuring that the ARIMA and LSTM models are trained and tested on high-quality, well-prepared data.

3.5 Model Building Strategies

This subchapter outlines the structured strategies employed in building the ARIMA and LSTM models, crucial for predicting stock prices in the Spanish banking sector. The approach adheres to best practices in model construction, ensuring clarity, precision, and relevance to the specific context of this research.

3.5.1 ARIMA Model Construction

The development of the ARIMA model begins with verifying the stationarity of the time series data. Non-stationary data is addressed through differencing or appropriate transformations, ensuring the data's suitability for ARIMA modeling. Preliminary parameters for the ARIMA(p,d,q) model are estimated using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), following guidelines from Ma (2020) and Siami-Namini, Tavakoli, and Namin (2018). Furthermore, this research will employ a rolling ARIMA approach, which involves periodically re-estimating the model parameters over the dataset to capture any potential changes in the underlying data patterns. This dynamic method is particularly useful in the context of financial time series data, where market conditions can evolve rapidly.

3.5.2 LSTM Model Development

For the LSTM model, data normalization is an initial critical step, implemented using the MinMaxScaler. This normalizes the input features, making them suitable for the LSTM's processing. The architecture of the LSTM model is thoughtfully designed, initially selecting one layer with four neurons for single-feature model and forty neurons for dual-feature model (involving 'Close' price and RSI). This strategic choice, inspired by Siami-Namini, Tavakoli, and Namin (2018), aims for a simpler architecture to prevent overfitting, particularly given the limited features used.

The time steps in the LSTM model are set to one to reflect the selected lag in the series. The Adam optimizer is chosen for its effectiveness in optimization, and Mean Squared Error (MSE) is used for loss measurement. Dropout regularization, initially omitted to maintain model simplicity, can be introduced later if overfitting is observed. The learning rate and batch size are set at 0.001 and 1, respectively, providing a starting point for the training process. The Dense layer, a critical component of the model, is configured with a value of one, with all these parameters subject to adjustments during training. An initial Epoch value of 100 is chosen to balance between adequate training and computational efficiency.

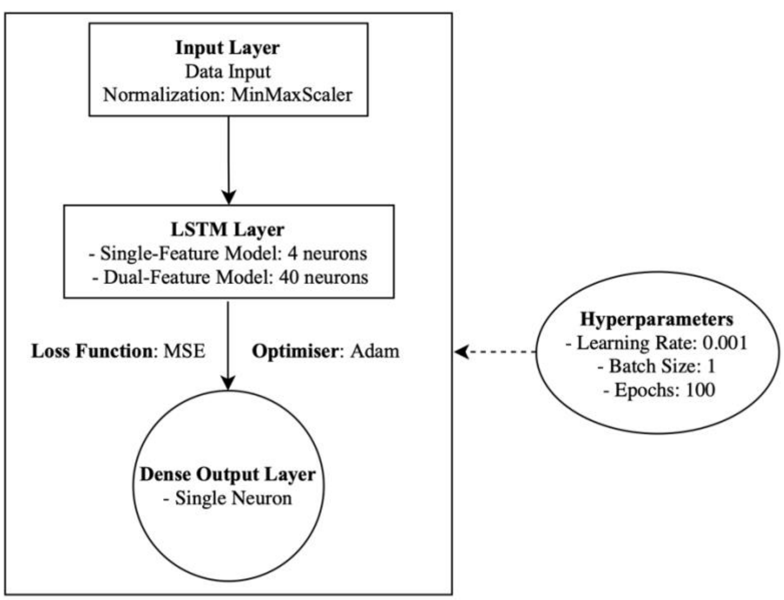

Figure 3: The initial LSTM Model Architecture

This strategic approach in model building for ARIMA and LSTM combines theoretical understanding with practical considerations, aligning with the research's aim to provide robust and accurate predictive tools for the Spanish banking sector.

3.6 Optimization Techniques for Model Accuracy

This subchapter outlines the optimization techniques employed to enhance the accuracy of ARIMA and LSTM models, specifically for forecasting stock prices in the Spanish banking sector. These techniques are key to refining the models to align with the market's unique dynamics.

3.6.1 ARIMA Model Optimization

In optimizing the ARIMA model, the primary objective is to identify the configuration that best fits the data while avoiding overfitting. This involves selecting a model with the lowest Akaike Information Criterion (AIC) value, a method recommended by Mondal, Shit, and Goswami (2014). To streamline this process, the `auto_arima` Python library is employed. This tool automates the evaluation of various parameter combinations (p, d, q) and selects the model that offers the optimal balance between accuracy and simplicity based on AIC, along with other criteria like the Bayesian Information Criterion (BIC) and the Ljung-Box test.

3.6.2 LSTM Model Optimization

For the LSTM model, hyperparameter tuning is crucial. Parameters such as the number of LSTM cells, learning rate, batch size, and epochs are strategically adjusted. Balancing learning complexity and overfitting prevention, the learning rate is fine-tuned for stable training. Batch size and epochs are optimized to balance computational efficiency and learning effectiveness. To mitigate overfitting risks, Dropout regularization is considered, which aids in model generalization by deactivating neurons randomly during training.

Implementing these optimization techniques for both ARIMA and LSTM models is done with careful consideration of their distinct characteristics and the Spanish banking market's requirements. The aim is to develop models that are theoretically robust and practically capable of making accurate real-world predictions.

3.7 Validation and Testing Methodology

This section describes the approach for validating and testing the ARIMA and LSTM models, crucial for ensuring their accuracy in predicting stock prices within the Spanish banking sector.

3.7.1 Data Split

The datasets are strategically divided using an 80-20 split, allocating 80% for model training and 20% for testing. This split ensures a robust learning phase while reserving a substantial portion for an objective performance assessment.

3.7.2 Evaluation Metrics

The effectiveness of the ARIMA and LSTM models is determined using three principal metrics. Mean Squared Error (MSE) quantifies the average of the squares of errors, pinpointing the model's accuracy where lower values indicate better performance. Mean Absolute Percentage Error (MAPE) offers an understanding of prediction accuracy in relative terms, ideal for comparing performance across datasets. Root Mean Squared Error (RMSE) provides error magnitude in the same units as the predicted values, emphasizing larger errors and enhancing interpretability.

3.8 Ethics, Reliability, and Validity Measures

In this research, strict adherence to ethical standards is maintained, particularly regarding data privacy. The use of publicly available secondary data from Yahoo Finance ensures compliance with ethical norms. The ARIMA and LSTM models' reliability is tested across different time frames and market conditions, assessing their consistency and adaptability. For validity, model predictions are compared with actual market data, confirming their accuracy and practical applicability in the Spanish banking sector.

3.9 Comparative Analysis and Model Evaluation

The comparative analysis of the ARIMA and LSTM models in this research is focused on their effectiveness in predicting stock prices, specifically the "Close" price, within the Spanish banking sector. The evaluation involves a detailed comparison using key performance metrics: Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error

(RMSE). These metrics provide a clear and quantitative measure of each model's predictive accuracy. The analysis is designed to highlight the relative strengths and limitations of each model, offering insights into their suitability for accurately forecasting stock prices. The results from this comparative evaluation will be instrumental in determining the more suitable model for practical forecasting applications in the Spanish banking sector.

3.10 Technological Tools and Software Utilization

This research leverages Python for its powerful data analysis and machine learning capabilities, with Jupyter Notebook as the interactive platform for code execution and visualization. The ARIMA model is developed using Python's statistical libraries, while the LSTM model utilizes the Keras library, a high-level neural networks API, for efficient deep learning model construction and training.

3.11 Conclusion

The comprehensive methodology and systematic approach outlined in the preceding sections have laid a solid foundation for the application and evaluation of ARIMA and LSTM models in predicting stock prices in the Spanish banking sector. This research has methodically addressed significant challenges, bridging the gap in comparative studies within this sector. Key highlights include:

- **Tailored Model Development**: Both ARIMA and LSTM models were meticulously developed and fine-tuned to align with the unique dynamics of the Spanish banking sector, leveraging their respective strengths in handling linear trends, short-term forecasting, and complex, non-linear dynamics.

- **Robust Data Collection and Analysis**: Utilizing a decade of historical stock price data from major Spanish banks, this research has provided an extensive base for exploring market trends and potential forecasting capabilities, enhancing the models' applicability and relevance.

- **Incorporation of Comprehensive EDA and Preprocessing Steps**: The thorough EDA and data preprocessing, including managing missing values, trend analysis, seasonality assessment, volatility assessment, and feature engineering, have ensured the integrity and quality of the data used in the models.

- **Strategic Model Optimization and Validation**: The optimization techniques applied to both models, including hyperparameter tuning and diagnostic checks, have been instrumental in enhancing their accuracy and reliability. The validation and testing methodologies, including the use of key performance metrics like MSE, MAPE, and RMSE, have provided a clear understanding of each model's predictive capabilities.
- **Ethical Considerations and Practical Relevance**: Adhering to ethical standards, particularly in terms of data privacy, and testing the models across various market conditions, have bolstered the research's reliability and validity. The comparative analysis has offered valuable insights into the models' effectiveness in real-world scenarios.
- **Technological Implementation**: The effective utilization of Python, Jupyter Notebook, and other relevant tools and libraries like Keras has facilitated efficient model development and evaluation, demonstrating the practical application of advanced technological tools in financial econometrics.

In summary, this research has not only contributed to the field of financial econometrics by providing a detailed comparative analysis of ARIMA and LSTM models but also aided in informed decision-making within the Spanish banking sector. The insights garnered from this study hold the potential to guide future research and practical applications in stock price prediction and other related financial areas.

CHAPTER FOUR

4.0 Data Analysis, Modeling, Validation, and Performance Evaluation

4.1 Introduction

In this chapter, an exhaustive analysis is conducted on the three datasets from prominent Spanish banking institutions - Banco Santander (SAN), Banco Bilbao Vizcaya Argentaria (BBVA), and

Caixabank (CABK). The focus encompasses a detailed comparative evaluation of ARIMA and LSTM models, the integration of the Relative Strength Index (RSI) to potentially enhance forecasting precision, and the strategic optimization of model parameters. Rigorous validation and evaluation processes are undertaken to ascertain the practical efficacy and reliability of these models in real-time market scenarios.

4.1.1 Dataset Examination

This section delves into the examination of the three datasets, which comprise 2560 observations each. The each dataset encompasses 6 features, each providing crucial insights into the stock market dynamics:

- **Date**: The date when the data was obtained
- **Open**: The opening price of the stock at the beginning of the trading day.
- **High**: The highest price at which the stock traded during the trading day.
- **Low**: The lowest price at which the stock traded throughout the day.
- **Close**: The final trading price of the stock at the end of the trading day.
- **Adjusted Close**: The closing price of the stock, adjusted for any corporate actions such as dividends or stock splits.
- **Volume**: The total number of stock shares that were traded during the day.

index	Date	Open	High	Low	Close	Adj Close	Volume
0	2013-12-20	3.750	3.759	3.700	3.755	2.472463	43110444.0
1	2013-12-23	3.756	3.789	3.725	3.789	2.494851	11507213.0
2	2013-12-24	3.740	3.769	3.735	3.745	2.465879	4981350.0
3	2013-12-27	3.730	3.795	3.710	3.774	2.484974	19022863.0
4	2013-12-30	3.783	3.799	3.750	3.782	2.490241	8422452.0

Table 1: Caixabank Dataset description

index	Date	Open	High	Low	Close	Adj Close	Volume
0	2013-12-20	8.726	8.800	8.656	8.784	5.305657	152109404.0
1	2013-12-23	8.786	8.829	8.739	8.829	5.332839	39792306.0
2	2013-12-24	8.819	8.865	8.806	8.865	5.354582	9512777.0
3	2013-12-27	8.838	8.970	8.838	8.970	5.418005	24693413.0
4	2013-12-30	8.970	9.000	8.896	8.950	5.405923	24247404.0

Table 2: Banco Bilbao Vizcaya Argentaria Dataset Description

index	Date	Open	High	Low	Close	Adj Close	Volume
0	2013-12-20	5.982521	6.006085	5.917483	6.006085	3.700885	213163188.0
1	2013-12-23	6.011741	6.041903	5.959899	6.041903	3.722956	92522668.0
2	2013-12-24	6.049444	6.089032	6.027764	6.080549	3.746769	15662554.0
3	2013-12-27	6.098458	6.137103	6.070180	6.137103	3.781618	60934937.0
4	2013-12-30	6.144644	6.171978	6.106941	6.126735	3.775230	56726461.0

Table 3: Santander Dataset Description

4.1.2 Data cleaning

This section details the process of data cleaning, which is fundamental for ensuring the accuracy of subsequent analyses. In all three datasets, a single observation with missing values was identified. To address these gaps, a linear interpolation method was applied, filling the missing values in a manner consistent with the surrounding data. This approach was chosen for its effectiveness in preserving the continuity and integrity of the time series.

4.2 Exploratory Data Analysis

This section rigorously delves into Exploratory Data Analysis (EDA), emphasizing trend analysis, seasonality and volatility assessment. These methodologies are employed to meticulously dissect underlying patterns, discern cyclical trends, and evaluate the stability within the stock data of prominent Spanish banking institutions.

4.2.1 Trend Analysis

In this subsection, Trend Analysis is conducted by plotting line charts for each of the three datasets, representing major Spanish banks.

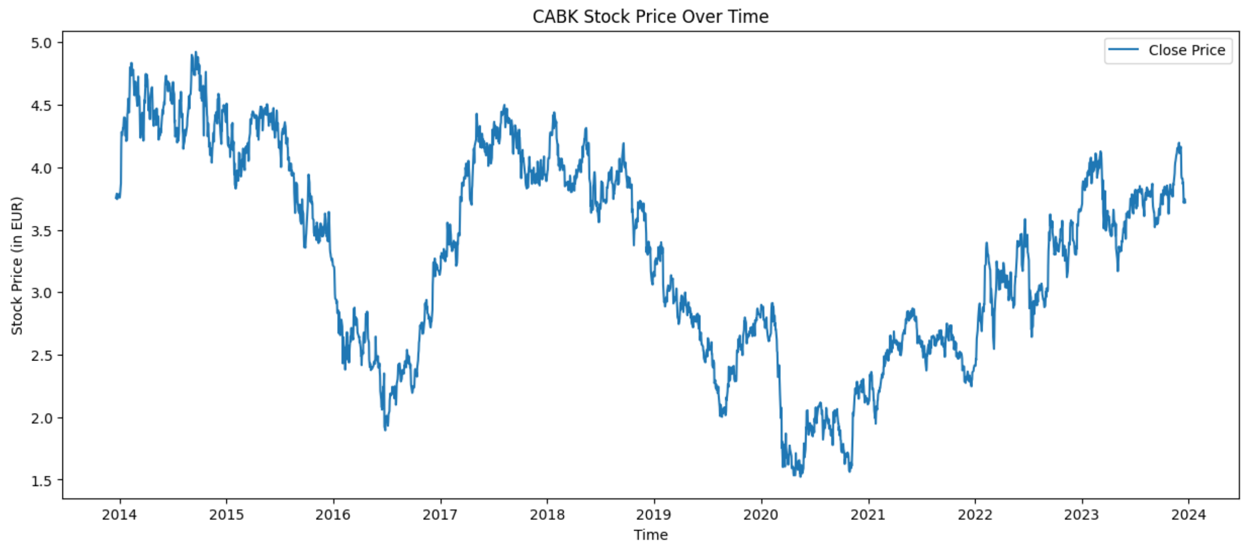

Figure 4: A decade of CABK historical price

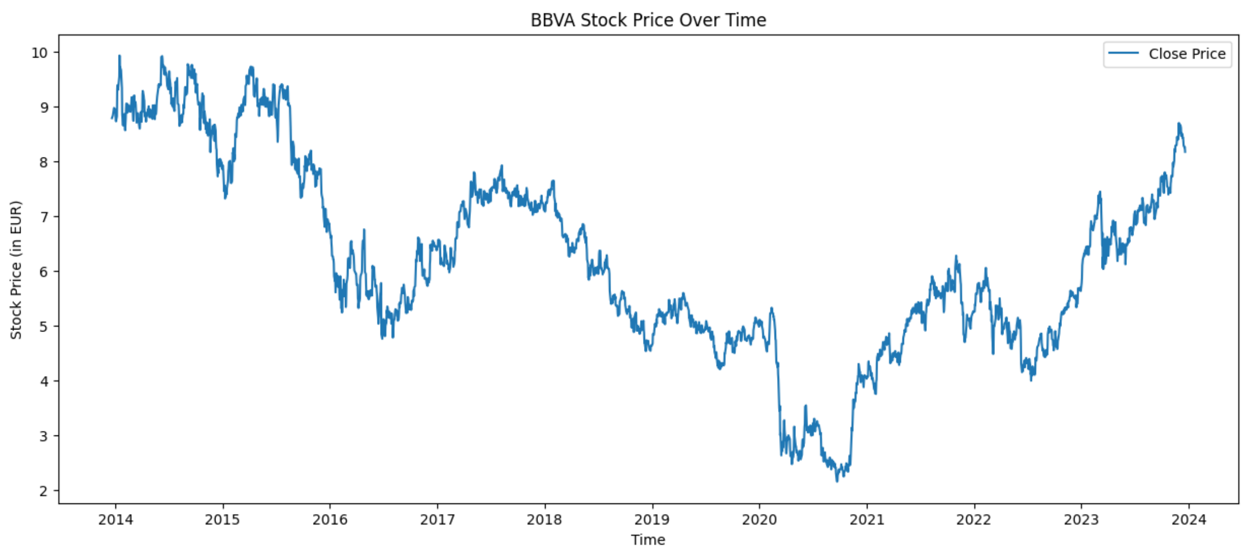

Figure 5: A decade of BBVA historical price

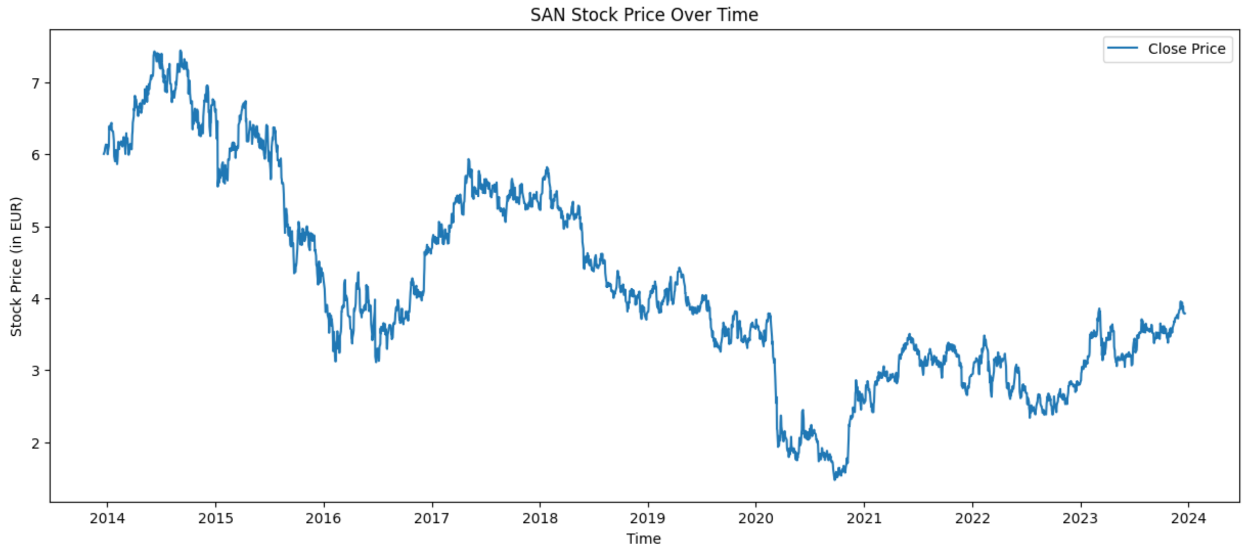

Figure 6: A decade of SAN historical price

The observed trends in the stock data are characterized by volatility, marked by significant fluctuations without a distinct long-term upward or downward trajectory. Notably, all three stocks exhibit a pronounced downturn around the same period, reaching a nadir in 2020, aligning with the global financial markets' response to the pandemic. Following 2020, there is an observable recovery in all stocks, potentially indicative of a broader market resurgence or a sector-specific rebound.

4.2.2 Seasonality Assessment

This section examines potential seasonality within the stock data of the three major Spanish banks.

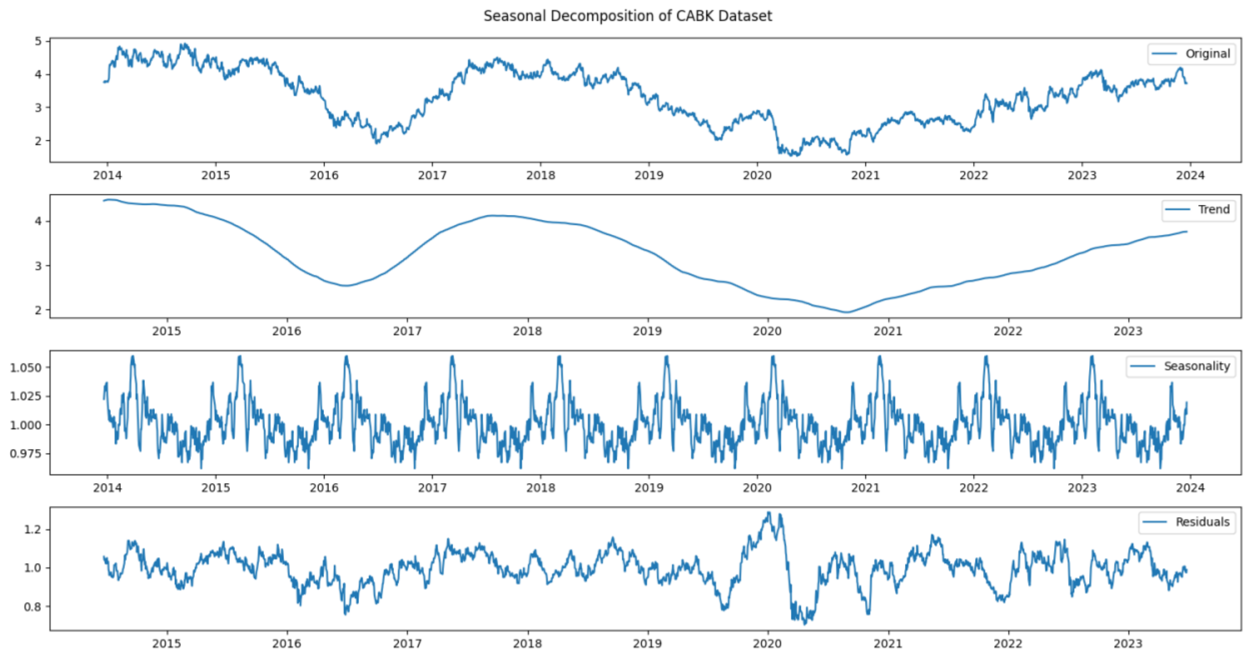

Figure 7: Seasonal decomposition of CABK Dataset

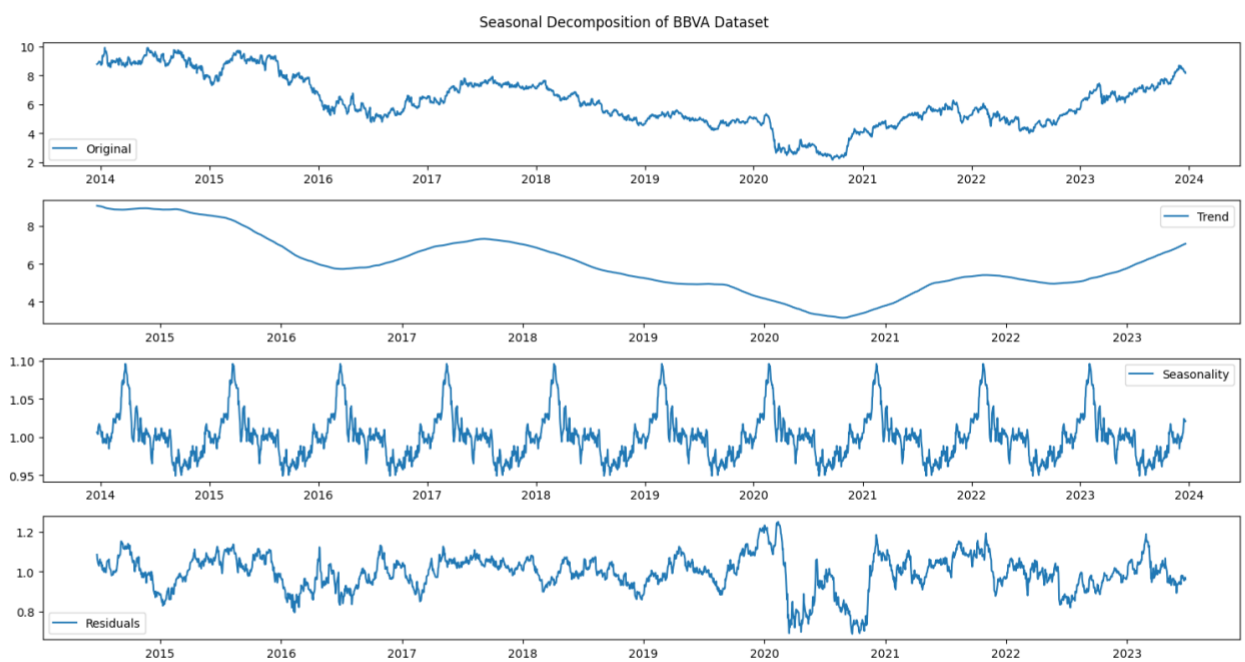

Figure 8: Seasonal decomposition of BBVA Dataset

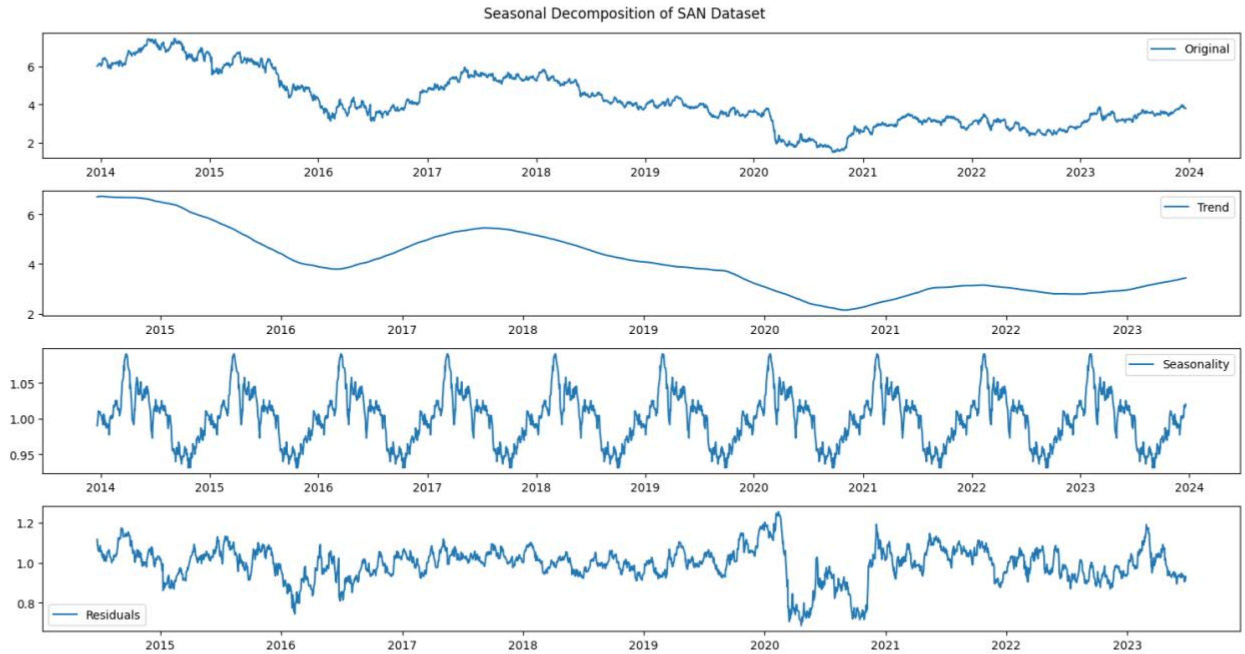

Figure 9: Seasonal decomposition of SAN Dataset

The analysis reveals fluctuations in the datasets that might suggest seasonality. Yet, upon thorough examination, the original data from the three banks do not demonstrate regular, significant oscillations that align with the strength of any extracted seasonal component. It appears that the stock prices fluctuate due to a variety of factors, and the identified seasonal patterns do not predominantly drive these price movements. This indicates that seasonality, while present, may not be the principal factor influencing the stocks' performance.

4.2.3 Volatility Assessment

In this section, the volatility of stock prices for key Spanish banks is evaluated using the standard deviation as a measure. The annualized volatility for Caixabank is calculated at 32.74%, indicating notable price fluctuations over the course of a year. Similarly, Banco Bilbao Vizcaya Argentaria (BBVA) and Banco Santander (SAN) exhibit annual volatilities of 32.2% and 32.54%, respectively. These figures, derived from standard deviation calculations, underscore a significant level of volatility in the stock prices of these major Spanish banks, reflecting the dynamic and fluctuating nature of their market performance.

4.3 Feature Engineering

This section elucidates the process of feature engineering, a crucial step in enhancing model performance by selecting, creating, and transforming relevant features. It encompasses the methodologies for feature selection and creation, data scaling to normalize feature values, and the strategic splitting of data into training and testing sets, all fundamental for refining the predictive models.

4.3.1 Feature Creation and Selection

The 'Close' price is identified as the primary feature for both models, due to its direct relevance to stock price prediction. Additionally, the Relative Strength Index (RSI) is created using its established formula, serving as an auxiliary feature to potentially enhance the LSTM model's predictive capabilities.

4.3.2 Feature Scaling

The MinMaxScaler is employed to scale both features for LSTM within a range of 0 to 1, ensuring uniformity and comparability across the different variables used in the models.

4.3.3 Data Splitting

An 80-20 split ratio is applied, allocating 80% of the data for model training and the remaining 20% for testing and evaluating the model's performance, ensuring a comprehensive assessment of its predictive capabilities.

4.4 ARIMA Model Building

ARIMA is a statistical model used in timeseries analysis and this research employed ARIMA model for stock price forecasting. The sections below will present the development of ARIMA model, its architecture, validation and performance.

4.4.1 Stationarity and Autocorrelation

This section delves into the analysis of stationarity and autocorrelation, foundational aspects of time series modeling. The Augmented Dickey-Fuller test, applied to assess stationarity, yielded p-values of 0.283414, 0.385892, and 0.385892 for the CABK, BBVA, and SAN time series, respectively. These results indicated that all three time series are not stationary, necessitating differencing for appropriate modeling.

Further, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analyses were conducted to establish initial parameters for the ARIMA models. The ACF for all datasets did not exhibit a clear, gradual decline, implying that a significant Moving Average (MA) component may not be necessary for the ARIMA model. Conversely, the PACF plots showed a sharp cut-off after the first lag, suggesting that an Autoregressive (AR) term of order 1 could be sufficient to capture the autocorrelation present in the series. Further analysis will be conducted in the next section.

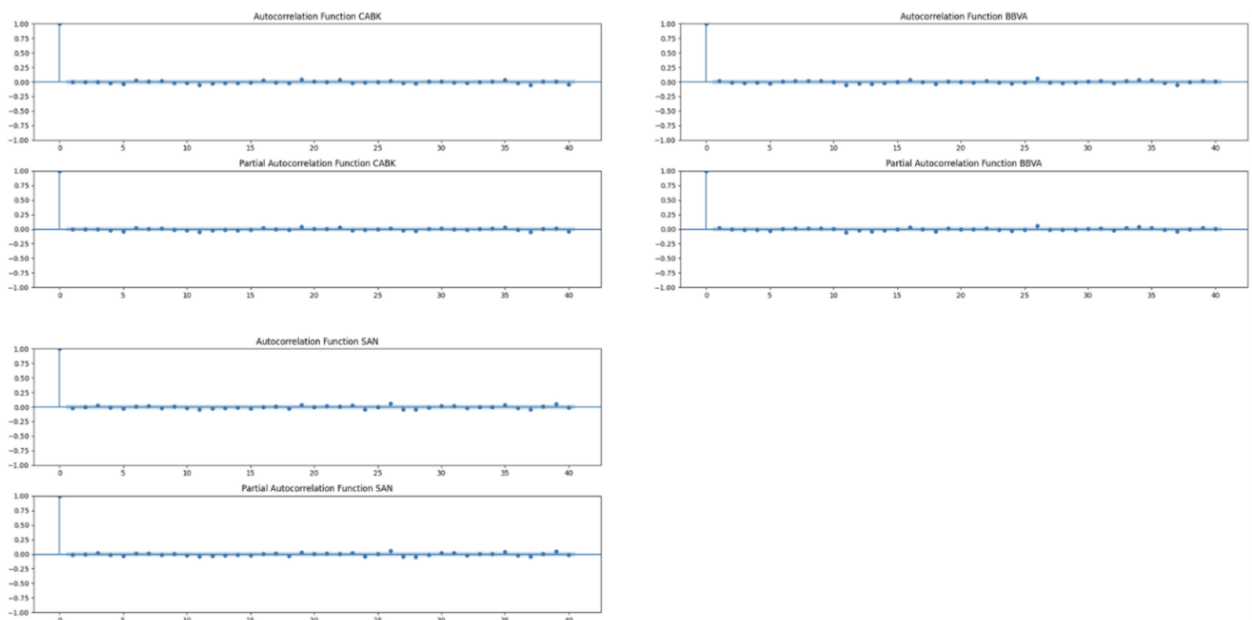

Figure 10: ACF and PACF of all three time series

4.4.2 ARIMA Model Parameters

This section outlines the determination of ARIMA model parameters. Utilizing the `auto_arima` Python library, the optimal parameters identified for each time series were ARIMA (0, 1, 0), commonly referred to as a random walk model. This model selection, predicated on achieving the lowest Akaike Information Criterion (AIC), indicates a better fit to the data with minimized complexity, thereby reducing the likelihood of overfitting.

Additionally, the `auto_arima` library incorporates the Ljung-Box test to evaluate autocorrelation in the residuals for a fixed number of lags. The p-values for the CABK, BBVA, and SAN time series were found to be 0.23, 0.72, and 0.80, respectively. These results, with p-values exceeding the 0.05 threshold, suggest an absence of significant autocorrelation in the residuals, affirming the model's adequacy in capturing the time series dynamics.

4.4.3 ARIMA Model Validation and Performance

To assess the model's efficacy, predictions were made using the test data. The performance metrics for the ARIMA model are presented in the table below:

Metric	CABK Time Series	BBVA Time Series	SAN Time Series
MSE	0.004655	0.013060	0.003876
RMSE	0.068229	0.114280	0.062265
MAPE	1.509603	1.412556	1.517837

Table 4: The performance of ARIMA models

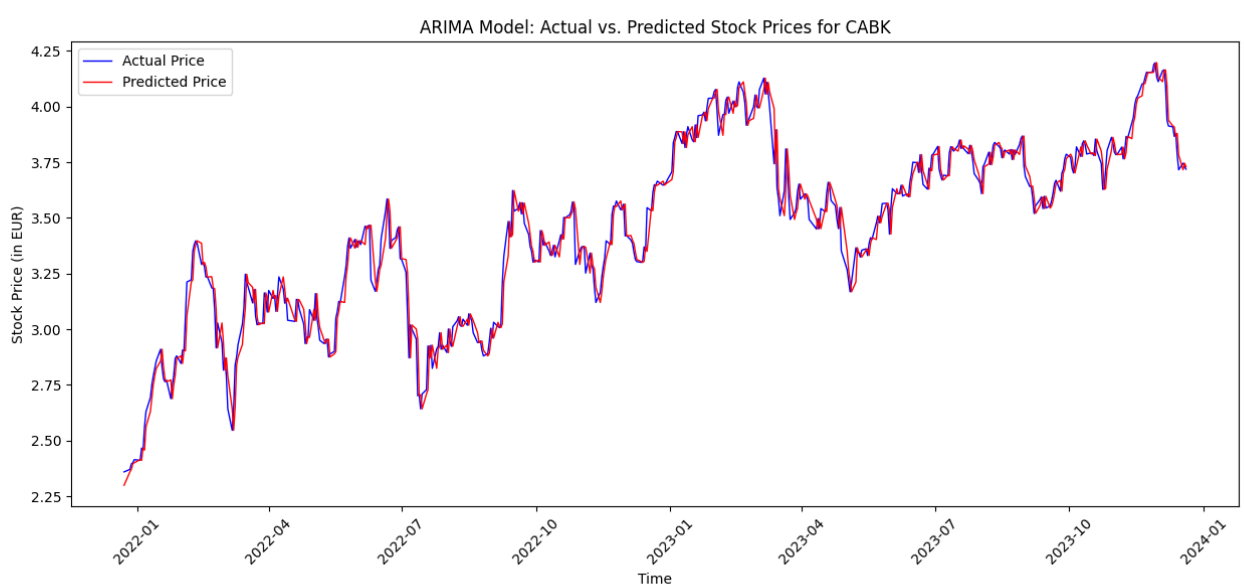

Figure 11: ARIMA Model Plot of Actual vs. Predicted Stock Price of CABK

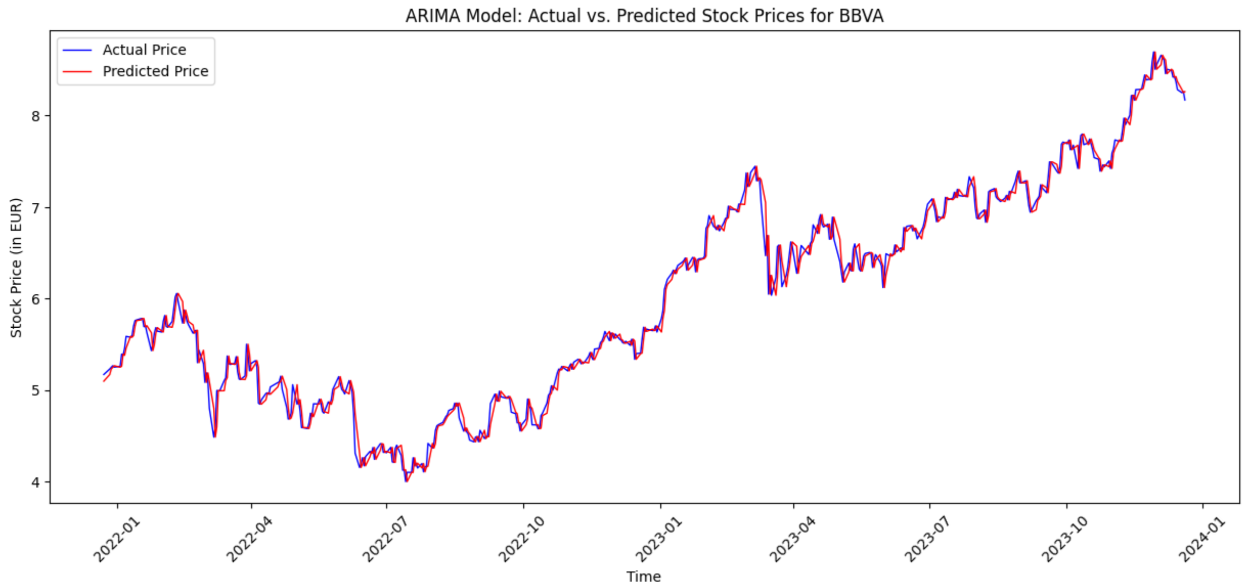

Figure 12: ARIMA Model Plot of Actual vs. Predicted Stock Price of BBVA

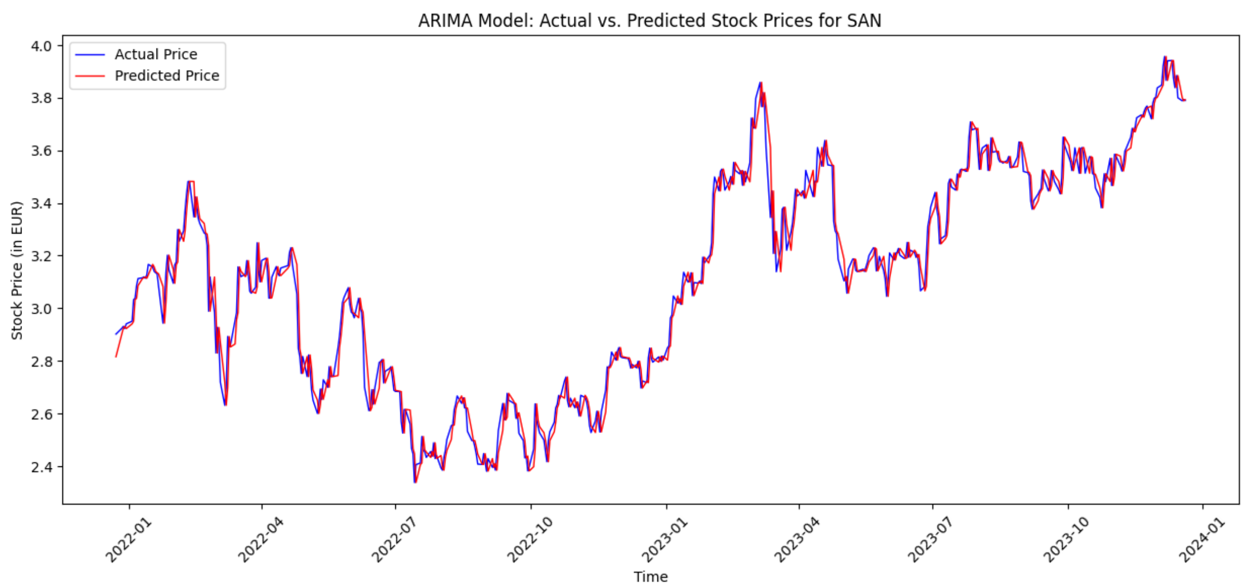

Figure 13: ARIMA Model Plot of Actual vs. Predicted Stock Price of SAN

4.5 LSTM Model Building

The sections below focus on employing the Long Short-Term Memory (LSTM) model, adept at learning long-term dependencies, for stock price forecasting. It will outline the model's development, architecture, validation, and performance evaluation in the context of financial time series analysis.

4.5.1 Initial LSTM Model Validation and Performance

In alignment with the methodology outlined earlier, the initial architecture and parameters of the LSTM model were applied. Subsequent to its application on test data, the performance of the LSTM model is systematically presented in the following table:

Metric	CABK Time Series	BBVA Time Series	SAN Time Series
MSE	0.004886	0.013353	0.004526
RMSE	0.069905	0.115556	0.067278
MAPE	1.563520	1.447847	1.667631

Table 5: The performance of LSTM models with initial architecture and parameters

4.5.2 LSTM Model Hyperparameter Tunning

Hyperparameter tuning was diligently conducted, resulting in the selection of optimal parameters for peak performance. For the CABK dataset, 4 units and 100 epochs were chosen, aligning with the initial model's configuration. For the BBVA dataset, the configuration included 14 units and 150 epochs, while for the SAN dataset, it comprised 34 units and 150 epochs.

4.5.3 LSTM Model Validation and Performance after Hyperparameter Tunning

The models were validated against the test data, with their performance outcomes detailed in the subsequent table:

Metric	CABK Time Series	BBVA Time Series	SAN Time Series
MSE	0.004879	0.013156	0.003955
RMSE	0.069850	0.114703	0.062889
MAPE	1.560905	1.419623	1.529908

Table 6: The performance of LSTM models after hyperparameter tunning

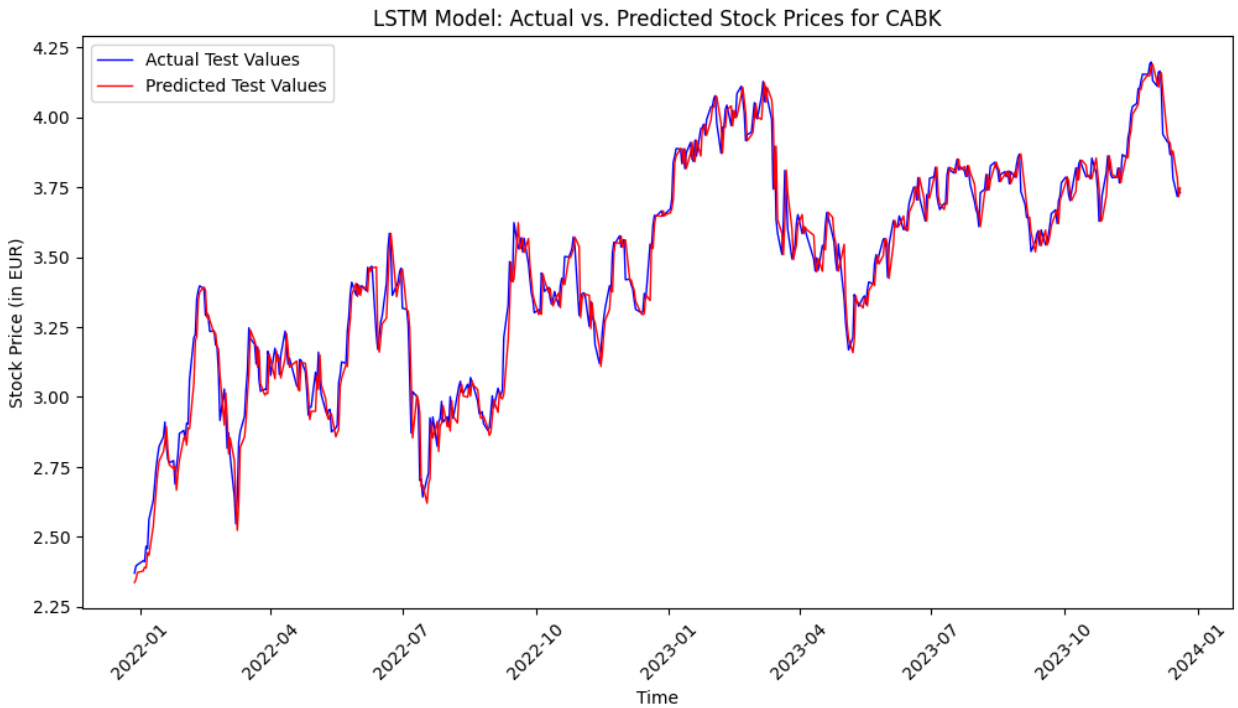

Figure 14: LSTM Model Plot of Actual vs. Predicted Stock Price of CABK

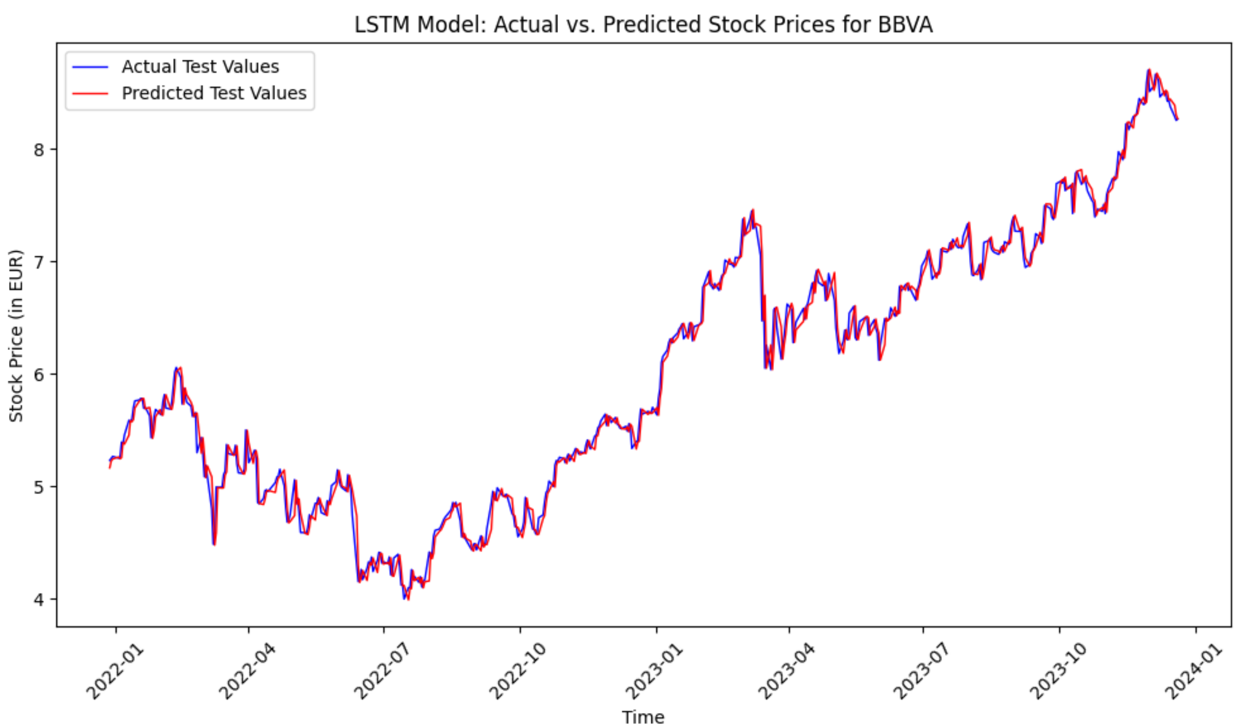

Figure 15: LSTM Model Plot of Actual vs. Predicted Stock Price of BBVA

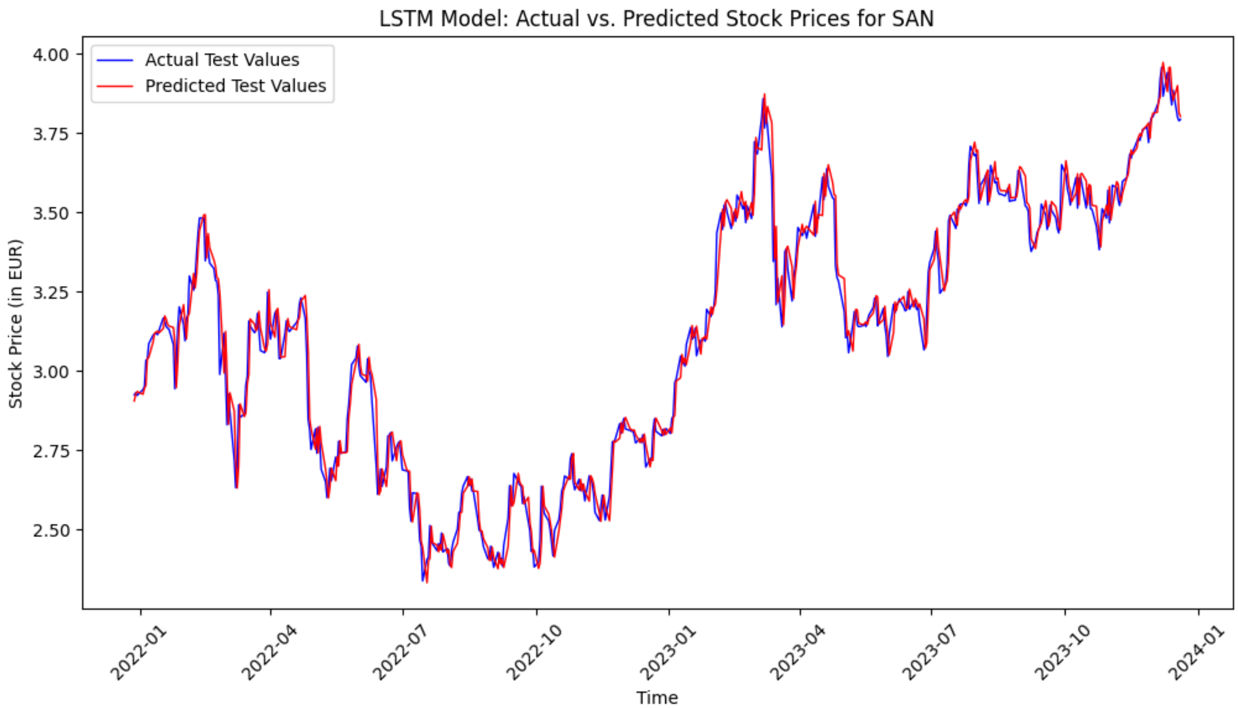

Figure 16: LSTM Model Plot of Actual vs. Predicted Stock Price of SAN

4.5.4 Two Features LSTM Model Hyperparameter Tunning

In accordance with the established methodology, the Relative Strength Index (RSI) was incorporated as a secondary feature into the LSTM model. Employing the initial architecture and parameters as a foundation, hyperparameter tuning was meticulously executed to determine the optimal model configuration. For the CABK dataset, 60 units and 150 epochs were chosen, aligning with the initial model's configuration. The configuration of the BBVA dataset included 50 units and 150 epochs, while the SAN dataset comprised 80 units and 150 epochs.

4.5.5 Two Features LSTM Model Validation and Performance

The performance metrics of the tuned models are presented in the table below:

Metric	CABK Time Series	BBVA Time Series	SAN Time Series
MSE	0.004911	0.013806	0.003970
RMSE	0.070085	0.117499	0.063015
MAPE	1.547387	1.503831	1.553404

Table 7: The performance of LSTM models with two features after hyperparameter tunning

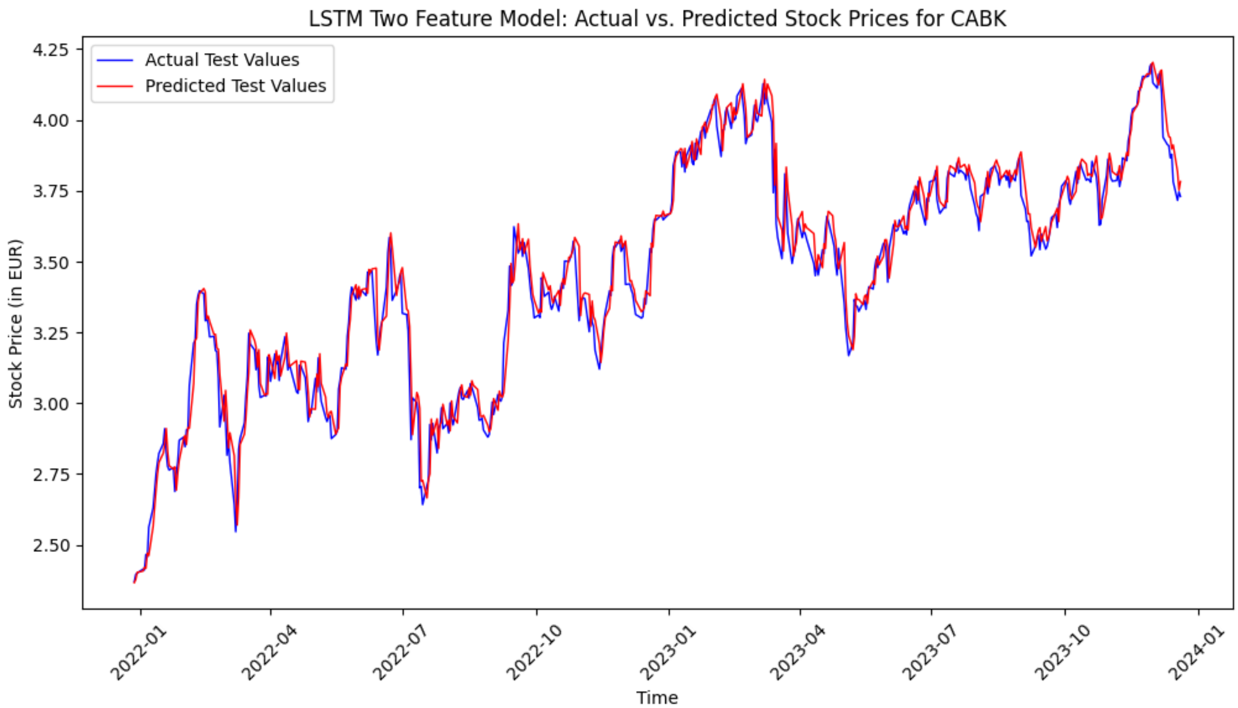

Figure 17: LSTM Two Feature Model Plot of Actual vs. Predicted Stock Price of CABK

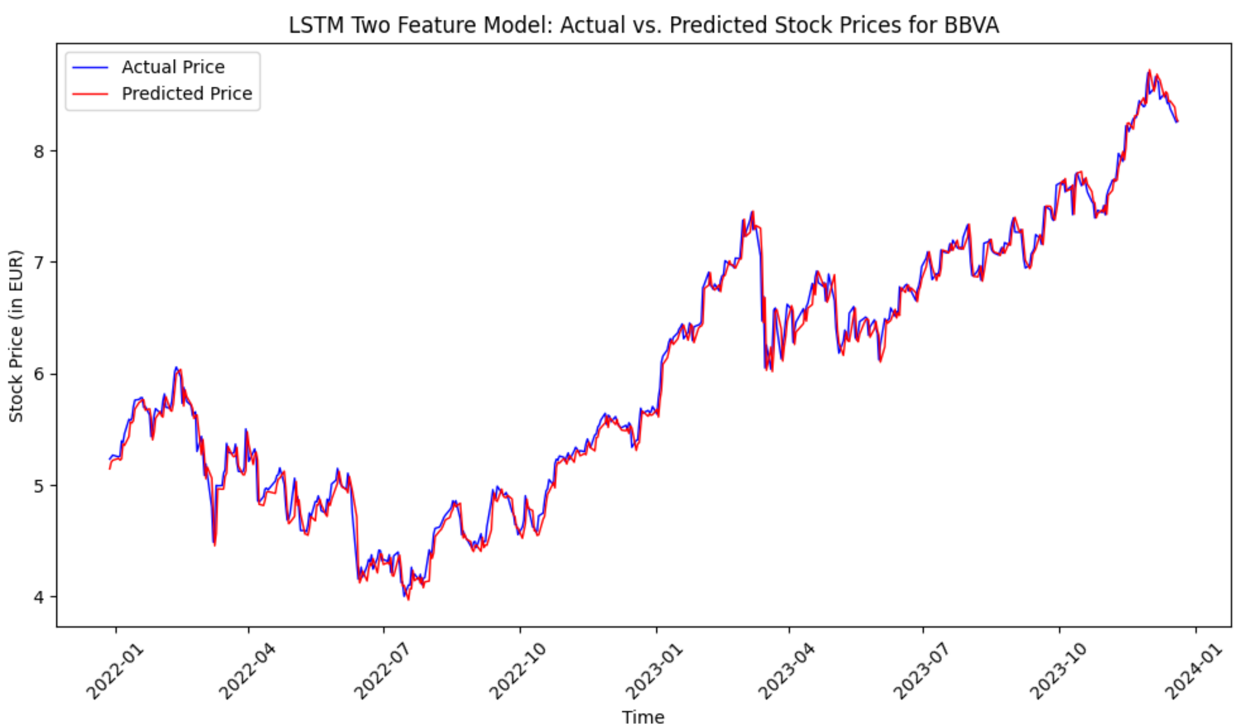

Figure 18: LSTM Two Feature Model Plot of Actual vs. Predicted Stock Price of BBVA

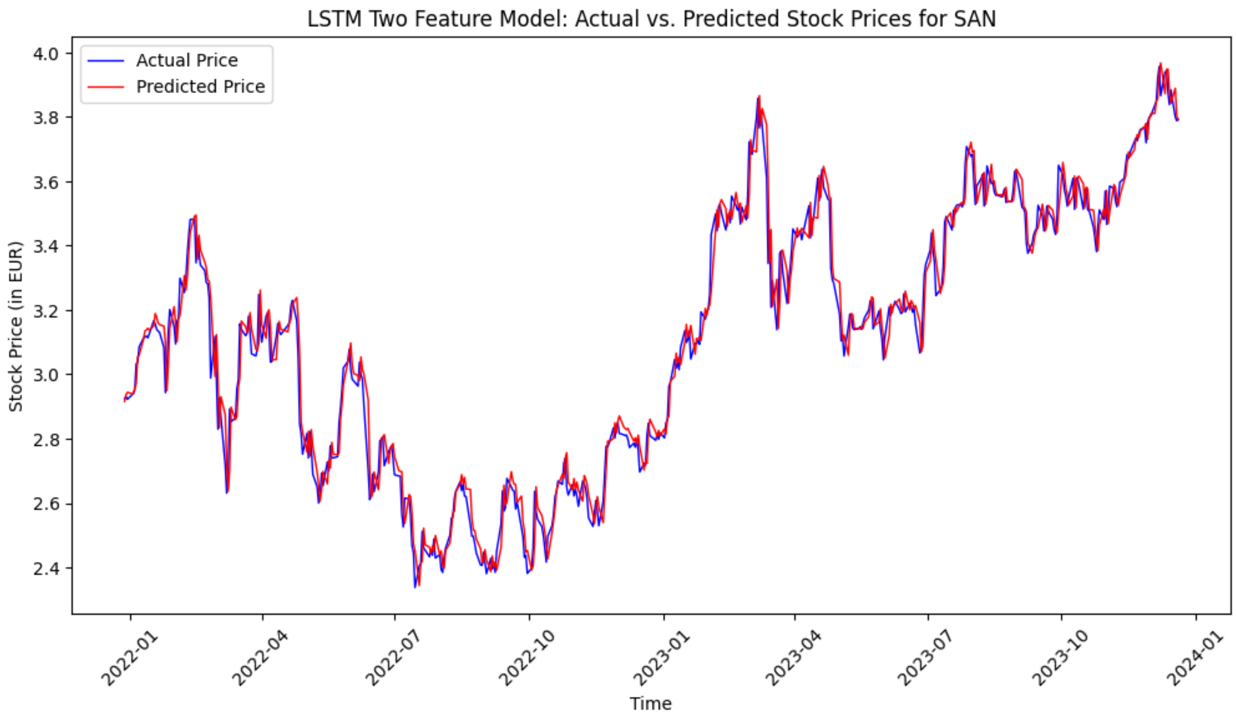

Figure 19: LSTM Two Feature Model Plot of Actual vs. Predicted Stock Price of SAN

4.6 Comparative Analysis of ARIMA and LSTM models

A comparative analysis of the models was conducted, focusing on their Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) scores. This analysis encompasses the ARIMA model and LSTM models with both one and two features. The performance metrics for each dataset are delineated in the tables below:

Model	MAPE	MSE	RMSE
ARIMA	**1.509603**	**0.004655**	**0.068229**
One Feature LSTM	1.560905	0.004879	0.069850
Two Features LSTM	1.547387	0.004911	0.070085

Table 8: The comparative analysis of CABK dataset

Model	MAPE	MSE	RMSE
ARIMA	**1.412556**	**0.013060**	**0.114280**
One Feature LSTM	1.419623	0.013156	0.114703
Two Features LSTM	1.503831	0.013806	0.117499

Table 9: The comparative analysis of BBVA dataset

Model	MAPE	MSE	RMSE
ARIMA	**1.517837**	**0.003876**	**0.062265**
One Feature LSTM	1.529908	0.003955	0.062889
Two Features LSTM	1.553404	0.003970	0.063015

Table 10: The comparative analysis of SAN dataset

In this study, the ARIMA (0, 1, 0) model, essentially a random walk representation, outperformed the LSTM models in stock price forecasting. This indicates that the ARIMA model's linear approach was more congruent with the time series data at hand, highlighting its suitability for linear analysis in this specific instance.

At its core, the ARIMA (0, 1, 0) model treats the differences between sequential observations as white noise, implying unpredictability in future values. This approach suggests that future stock prices are best forecasted as the most recent observed value plus random variations, a characteristic typical of a random walk pattern. This implies limited predictability based on historical data.

The simplicity of the ARIMA model offers substantial benefits, including easier interpretation, reduced overfitting risk, and lower computational needs. These advantages make it particularly adept for data exhibiting simple, non-predictive patterns akin to a random walk.

However, it is important to note that these findings are specific to the datasets and context of this particular research. While the ARIMA model showed superior performance in this scenario, it does not imply its universal superiority over LSTM models in all stock price forecasting contexts.

4.7 Conclusion

Chapter four encompassed a comprehensive analysis of datasets from Banco Santander, Banco Bilbao Vizcaya Argentaria, and Caixabank. Key stock market features were examined and missing values addressed through linear interpolation. Exploratory data analysis highlighted significant stock price volatility and fluctuating trends, particularly a downturn in 2020 followed by a recovery, with minimal influence from seasonal patterns. In feature engineering, the 'Close' price and RSI were emphasized, alongside data normalization and splitting strategies.

ARIMA model analysis demonstrated its superiority in forecasting compared to LSTM, with the ARIMA (0, 1, 0) model aligning well with the data's linear nature. This suggests its effectiveness for simple, non-complex patterns akin to a random walk.

The findings of this research significantly contribute to the field of financial econometrics, particularly in the analysis of Spanish banking stocks. By demonstrating the efficacy of the ARIMA model in a specific economic context, this study provides a nuanced understanding of model applicability and offers insights into the dynamic nature of financial markets. It underscores the importance of model selection based on data characteristics, enhancing predictive accuracy and informing investment strategies within the Spanish banking sector.

CHAPTER FIVE

5.0 CONCLUSION

Chapter five marks the conclusion of this research, revisiting the initial aims, problem statements, and objectives. This chapter synthesizes findings from previous analyses to provide comprehensive insights and responses to the research questions. It offers a critical evaluation, a comparison with existing literature, and recommendations for future research. Ensuring coherence, this chapter builds on earlier chapters, culminating in a conclusive summary of the study's key aspects: results interpretation, comparative analysis, future scope, and overarching conclusions.

5.1 Discussion

5.1.1 Interpretation of Results

This study's results indicate that the ARIMA (0, 1, 0) model, which operates as a random walk, was more effective in forecasting stock prices within the Spanish banking sector than LSTM models. This finding directly addresses the project's core objectives and research questions.

The primary goal of assessing the ARIMA and LSTM models' effectiveness in predicting stock prices for major Spanish banks has been achieved, with a clear preference for the ARIMA model. This outcome directly responds to the initial research question about the comparative efficacy of these models in the Spanish banking context. The findings demonstrate that simpler, linear models like ARIMA may be more suitable under certain market conditions, underscoring the criticality of model selection based on data attributes.

In terms of feature selection's impact on prediction accuracy, a key element of the second research question, the study revealed that the ARIMA model's focus on historical stock prices as the main feature led to more accurate forecasts. This underscores the importance of feature relevance, particularly in markets with complex dynamics. Additionally, adding RSI as an additional feature in the LSTM model did not improve prediction accuracy.

The third research question, pertaining to optimal model parameters, was explored through parameter tuning for both ARIMA and LSTM models. The ARIMA (0,1,0) emerged as the most effective model. For the LSTM model with one feature, the optimal range was identified as four to thirty-four units and one hundred to one hundred fifty epochs. For the two-feature LSTM model, the optimal parameter range included fifty to eighty units and one hundred fifty epochs.

5.1.2 Comparison with Existing Studies

The findings of this study stand in contrast to several existing studies where LSTM models have shown higher efficacy in stock price prediction, often in markets with complex, non-linear data patterns. The preference for ARIMA in the context of the Spanish banking sector underscores the uniqueness of this study and its contribution to the field. It suggests that the effectiveness of predictive models can significantly vary based on the specific characteristics of the market and the nature of the data.

This research contributes to the ongoing discourse in financial econometrics regarding the applicability of traditional statistical methods and advanced machine learning techniques. By demonstrating the effectiveness of a simpler model in a specific economic context, it provides a nuanced perspective on the relative strengths of ARIMA and LSTM models in financial predictive analytics.

5.2 Future Scope

5.2.1 Prospective Avenues for Continued Research

Given the dynamic nature of financial markets and the rapid evolution of predictive modeling techniques, future research can explore several avenues. One potential area is the development of hybrid ARIMA-LSTM models that could leverage the strengths of both methods—ARIMA's efficiency in handling linear trends and LSTM's proficiency in modeling non-linear relationships. Another interesting direction would be the application of these models in different sectors or geographical regions, to assess their versatility and adaptability to various market conditions.

5.2.2 Refinement and Enhancement of Methodological Approaches

The methodology of this project could be enhanced by incorporating additional data types, such as macroeconomic indicators or sentiment analysis from financial news and social media. This would potentially improve the models' predictive accuracy.

5.2.3 Broader Application Spectrum of Predictive Models

The findings of this study could be applied to other areas within the financial sector, such as risk management, portfolio optimization, and algorithmic trading. Additionally, these predictive models could be valuable for regulatory bodies for monitoring market trends and for financial analysts in making investment recommendations.

5.2.4 Broadening the Research Scope

Expanding this project to include a broader range of financial instruments, like bonds, commodities, or derivatives, could provide a more comprehensive understanding of the financial markets. A larger dataset encompassing longer time frames or additional banking institutions could also offer deeper insights.

5.2.5 Integration with Contemporary Financial Technologies

Integration of the predictive models with real-time data feeds and trading platforms could facilitate automated trading strategies. Additionally, combining these models with blockchain technology could innovate how financial data is stored and accessed, enhancing transparency and security in financial transactions.

5.2.6 Project's Potential Impact on Industry and Society

The project's contribution could significantly influence the financial industry by offering enhanced tools for stock market analysis and prediction. On a societal level, this could lead to more stable financial environments and better-informed investment decisions by individuals.

5.2.7 Recommendations for Future Development by Peers

Future researchers could focus on optimizing these models to better handle market anomalies and extreme conditions. The exploration of other technical indicators could possibly enchase model's performance. Practical applications could involve the development of intuitive interfaces that democratize access to these sophisticated financial models, making them more accessible to a diverse user base.

5.3 Conclusion

5.3.1 Revisiting Project Aims

The primary objective of this research was to conduct a comparative analysis of the Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) models in predicting stock prices within the Spanish banking sector. This study aimed to determine which model, including its parameters and features, offers the most accurate predictions for major Spanish banks, thereby contributing to the field of financial econometrics.

5.3.2 Synopsis of Key Discoveries

The key finding of this research is the notable performance of the ARIMA (0, 1, 0) model, which outperformed the LSTM models in forecasting stock prices for large-cap Spanish banks. This was particularly evident in the context of the Spanish banking sector's unique market dynamics and data characteristics. The study revealed that the simpler, linear approach of the ARIMA model was more effective in this specific context, suggesting its suitability for markets where data patterns exhibit less complexity.

5.3.3 Implications and Impact of Research Outcomes

The implications of these findings are significant for both the academic domain of financial econometrics and the practical world of financial forecasting. In academia, it challenges the prevailing notion of the supremacy of complex machine learning models like LSTM in all scenarios, highlighting the relevance of traditional models like ARIMA in certain contexts. Practically, this research provides valuable insights for financial analysts, investors, and policymakers in the Spanish banking sector, suggesting that simpler models can be equally, if not more, effective in certain market conditions. This could guide more efficient and accurate forecasting practices, aiding in investment decisions and risk management strategies.

5.3.4 Limitations of the Research

The research, while comprehensive, had its limitations. The exclusive focus on the Spanish banking sector means the findings may not directly apply to other sectors or markets with different characteristics. Additionally, the reliance on historical stock price data might not fully capture future market dynamics or unforeseen economic events, which could affect the models' predictive accuracy.

5.3.5 Final Thoughts

This project has been a significant learning journey, providing deep insights into the complexities of financial predictive modeling and the nuances of model selection based on specific market conditions. The experience has highlighted the importance of a balanced approach in choosing between traditional statistical methods and advanced machine learning techniques, underscoring the value of context in financial econometrics. The findings from this study contribute meaningfully to the ongoing discourse in the field and offer a foundation for future explorations and innovations.

REFERENCES:

- Agrawal, Manish, Khan, D.A.U. and Shukla, D.P.K. (2019) 'Stock price prediction using technical indicators: a predictive model using optimal deep learning', International Journal of Recent Technology and Engineering (IJRTE), 8(2), pp. 2297–2305. Available at: https://doi.org/10.35940/ijrteB3048.078219.
- Alzheev, A.V. and Kochkarov, R.A. (2020) 'Comparative analysis of arima and lstm predictive models: evidence from russian stocks', Finance: Theory and Practice, 24(1), pp. 14–23. Available at: https://doi.org/10.26794/2587-5671-2020-24-1-14-23.
- Ariyo, A.A., Adewumi, A.O. and Ayo, C.K. (2014) 'Stock price prediction using the arima model', in 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. Cambridge, United Kingdom: IEEE, pp. 106–112. Available at: https://doi.org/10.1109/UKSim.2014.67.
- Ashok, A. and Prathibhamol, C.P. (2021) 'Improved analysis of stock market prediction:(Arima-lstm-smp)', in 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE). NaviMumbai, India: IEEE, pp. 1–5. Available at: https://doi.org/10.1109/ICNTE51185.2021.9487745.
- Banco bilbao vizcaya argentaria, s. A.(Bbva. Mc) Stock price, news, quote & history - yahoo finance (no date). Available at: https://finance.yahoo.com/quote/BBVA.MC/ (Accessed: 24 November 2023).
- Banco santander, s. A.(San. Mc) Stock price, news, quote & history - yahoo finance (no date). Available at: https://finance.yahoo.com/quote/SAN.MC/ (Accessed: 24 November 2023).
- Caixabank, s. A.(Cabk. Mc) Stock price, news, quote & history - yahoo finance (no date). Available at: https://finance.yahoo.com/quote/CABK.MC/ (Accessed: 24 November 2023).
- Choi, J.W. and Choi, Y. (2023) 'A study of prediction of airline stock price through oil price with long short-term memory model', International Journal of Advanced Computer Science and Applications, 14(5). Available at: https://doi.org/10.14569/IJACSA.2023.0140509.

- Htun, H.H., Biehl, M. and Petkov, N. (2023) ‘Survey of feature selection and extraction techniques for stock market prediction’, Financial Innovation, 9(1), p. 26. Available at: https://doi.org/10.1186/s40854-022-00441-7.
- Kobiela, D. et al. (2022) ‘ARIMA vs LSTM on NASDAQ stock exchange data’, Procedia Computer Science, 207, pp. 3836–3845. Available at: https://doi.org/10.1016/j.procs.2022.09.445.
- Largest spanish (Ibex) banks stocks by market cap (no date) Simply Wall St. Available at: https://simplywall.st/stocks/es/banks/market-cap-large (Accessed: 24 November 2023).
- Ma, Q. (2020) ‘Comparison of arima, ann and lstm for stock price prediction’, E3S Web of Conferences. Edited by S.O. Oladokun and S. Lu, 218, p. 01026. Available at: https://doi.org/10.1051/e3sconf/202021801026.
- Mondal, P., Shit, L. and Goswami, S. (2014) ‘Study of effectiveness of time series modeling (Arima) in forecasting stock prices’, International Journal of Computer Science, Engineering and Applications, 4(2), pp. 13–29. Available at: https://doi.org/10.5121/ijcsea.2014.4202.
- Roondiwala, M., Patel, H. and Varma, S. (2015). Predicting Stock Prices Using LSTM. International Journal of Science and Research (IJSR) ISSN, [online] 6, pp.2319–7064. Available at: https://www.ijsr.net/archive/v6i4/ART20172755.pdf.
- Siami-Namini, S., Tavakoli, N. and Siami Namin, A. (2018) ‘A comparison of arima and lstm in forecasting time series’, in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, FL: IEEE, pp. 1394–1401. Available at: https://doi.org/10.1109/ICMLA.2018.00227.
- Swathi, T., Kasiviswanath, N. and Rao, A.A. (2022) ‘An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis’, Applied Intelligence, 52(12), pp. 13675–13688. Available at: https://doi.org/10.1007/s10489-022-03175-2.
- Toprak, Ş., Çağil, G. and Kökçam, A.H. (2023) ‘Stock closing price prediction with machine learning algorithms: petkm stock example in bist’, Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 11(2), pp. 958–976. Available at: https://doi.org/10.29130/dubited.1096767.

- Wong, W.-K., Manzur, M. and Chew, B.-K. (2003) 'How rewarding is technical analysis? Evidence from Singapore stock market', Applied Financial Economics, 13(7), pp. 543–551. Available at: https://doi.org/10.1080/09603100220000020906.