Enhancing Stock Price Prediction: Investigating the Impact of Lagged Observations in the ARIMA Model

Binay Oli[1]

[1] Harrisburg University

Author Note

Binay Oli, Department of Data Science, Harrisburg University of Science and Technology. The author would like to thank Professor Arnold Miles and the the members of the thesis committee for their guidance and feedback throughout the research process.

The authors made the following contributions. Binay Oli: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Binay Oli, 326 Market St. Harrisburg, PA 17101. E-mail: boli1@my.harrisburgu.edu

Abstract

Stock Price Prediction is very important for investors as it helps them make buy, hold or sell decisions. This study aims to investigate the impact of increasing the number of lagged observations in the Autoregressive Integrated Moving Average (ARIMA) model on the prediction accuracy of average trading prices for value stocks.

ARIMA (Auto Regressive Integrated Moving Average) model is a time series forecasting model which uses historical data to come up with future prediction. This model can be applied to a wide range of domains, including finance, economics, and weather forecasting. This model has 3 different components and on this study, we are investigating the impact of lagged observations.

This study looks at closing price of 10 value stocks each from 10 different sectors over the period of 10 years (2009 to 2019).

At the end of the study, it was discovered that indeed when the number of lagged observations is increased, the model performance improves. This is a significant finding as it helps investors make calculated investing decisions on an everyday basis.

Keywords: Prediction, lagged observations, ARIMA

Word count: 134

Enhancing Stock Price Prediction: Investigating the Impact of Lagged Observations in the ARIMA Model

Introduction

The stock market is very volatile and unpredictable. Over the years, many authors and scholars have spent hundreds of hours studying the stock market and have tried to come up with a prediction model that would correctly predict future prices. In this regard, authors have proposed models such as ARIMA (Auto-Regressive Integrated Moving Average) model which is based on time-series analysis. Similarly, they have proposed models based on neural networks such as LSTM (Long Short Term Memory). This paper studies both of these models and applies them to historical data of value stocks to see which of these two models can accurately predict their prices.

Literature Review

ARIMA (Auto-Regressive Integrated Moving Average) model is a time series forecasting technique. It has three components namely Auto-regression, Integrated, and Moving Average. Auto-regression refers to a model that shows a changing variable to regresses on its own lagged, prior, values. Integrated represents the differentiation of raw observations to allow for the time series to become stationary (i.e., data values are replaced by the difference between the data values and previous values). And Moving Average incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Similarly, LSTM (Long Short term Memory) model is a recurrent neural network technique that stores several short-term memory in the memory cells and combines them to create a long-term memory. The memory cells then progress through time to use the previous input to show the next output.

Yan, Weihan, and Chang (2021) talks about how deep learning is being used to analyze big data. This model training has become very vital and the study of influencing factors is essential to improving the prediction accuracy of forecasting models. It is especially important to understand the in-built systems within the models and to be able to vary them to see if the prediction accuracy of the model gets improved. In the ARIMA model, there are three parameters namely, the lag order (p), the degree of difference, and q is the order of the moving average. The major benefit of the ARIMA model is that it only requires the prior data of the time series to make a forecast which makes it very efficient for short-term forecasting. It is very important in today's world to look at digital data to see if a trend can be spotted to predict the future in every industry. The stock market relies on data very heavily, thus, it is no exception. The stock market is very volatile with prices of stocks changing every second of every day. The study done by Assous, Rousan, Najjar, and Najjar (2020) takes into account 6 key international indices and compares them to Saudi's financial market using various exponential equations separately such as linear, logarithmic, quadratic, cubic, power, and so on. They find out that variables used in the ARIMA model (lag order, moving average window, etc.) are the most useful to build the prediction model. This shows us how important the ARIMA model is in providing future predictions for the financial markets. According to Kulshreshtha and A. (2020), the ARIMA model can be a very important tool in forecasting the prices of S&P 500 stocks when combined with the Long Short Term Memory (LSTM) and Recurrent Neural Network (RNN) technique. Using LSTM and RNN techniques, a new LSTM-ARIMA model can be designed to incorporate both linear and non-linear portions of the time series. Gorecka and Szmit (1999), in their paper, propose that system performance can be improved by using a hybrid model created using ARIMA and LSTM. In this article, Tiwari and Chaturvedi (2021) talks about the stock market's complex and dynamic nature and uses the LSTM model to make forecasts and do a comparative analysis.

Machine learning algorithms are being used in every industry to solve complex problems. As per D. Xiao and Su (2022), machine learning can be used to forecast the prices of stocks and the future trends of capital market prices. In their paper “Research on Stock Price Time Series Prediction Based on Autoregressive Integrated Moving Average”, the authors talk about how ARIMA can be used to forecast the Capital markets. The ARIMA model is used not just for stock market prediction but also in many other fields such as temperature prediction, price prediction for electricity, and wind speed. This gives ARIMA model the credibility that it is indeed able to provide accurate forecasts. Jiang (2022) explain the importance of financial data analysis for investors. They also explain how the financial markets are complex non-linear dynamic systems that are affected by multiple factors at the same time. They propose a stock selection model which accounts for stock prediction and stock scoring.

Although the ARIMA model has been widely used in several different industries, there are also other time series models which compete with the ARIMA model. For example, DE and ABC models are examples of other time series models which are used to predict the stock prices. The data we see in the stock market are in sequential format. Thus, performing a time-series analysis makes a lot of sense. In this article, Kumar, Kumar, and Kumar (2021) show that by using a combination of differential evolution (DE) and artificial bee colony (ABC) algorithm, we can find an even more efficient prediction of the prices compared to the traditional ARIMA model (Kumar et al. (2021)). Similarly, Adebiyi, Adewumi, and Ayo (2014) use New York Stock Exchange (NYSE) data to discuss the forecasting performance and ability of the ARIMA model. They claim that empirically speaking, neural networks were able to produce superior results when compared to the traditional ARIMA model (Adebiyi et al. (2014)). Islam and Nguyen (2020) also compare the ARIMA and artificial neural network model for stock price prediction and shows that neural networks are superior to the ARIMA model in their prediction accuracy (Islam and Nguyen (2020)).

Many authors in the modern day refer to a hybrid model of ARIMA and LSTM.

According to Temur and Yildiz (2021), a hybrid model using ARIMA and LSTM can be used to reduce the number of errors in the final stock prices. This is very important as this model combines two different techniques into one which is more efficient than both of the two models when looked at individually. As the size of the data sets has increased tremendously, it is very important to build models which are compatible with these data sets and can handle them well.

In addition to creating a hybrid model with LSTM and ARIMA, C. Xiao, Xia, and Jiang (2020) talks about Support Vector Machine (SVM) and how it can use to propose a cumulative auto regressive moving average to make basic predictions of the stock market. It claims that SVM when added to ARIMA can make a better prediction model altogether. It is very important to understand that the prediction accuracy of these models needs to be compared very deeply to come to a conclusion about which model is better by using the same data sets. It is often seen that authors use different data sets on different prediction models which oftentimes makes it difficult to compare the models.

Yu and Yan (2020) concludes that the model created at the end of the paper with Deep Neural Networks (DNNs) has a higher prediction accuracy. Comparing this with the results from ARIMA model will help understand which model would have the better results.

Saxena and Kamnge (2020) compare the different price forecasting models in this paper such as autoregressive integrated moving average (ARIMA), single exponential smoothed model (SES), double exponential smoothed model (DES), and Damped trend linear exponential smoothed model. They go ahead and test the results of each of these models and conclude that ARIMA(1,1,2) is the best prediction model among them. In this paper, the author also talks about how they created 3 three phases to create a prediction model. Phase I was the Identification phase where they transformed the data to stabilize variance and obtain a stationary series. Similarly, they examined the data to identify the potential model that could be used. In phase two, the authors estimated parameters in the potential models and selected the best models using suitable criteria. Then, they went ahead and tested the residuals to see if they had a meaning or if they were just white noise. Finally, in phase three, they created the forecasting model. It is important to see that the authors make a lot of estimations while building the models thus the model might only be compatible with a particular type of data and might not apply to the whole stock market.

Daradkeh (2022) look at the impact of news events and sentiment trends to present a hybrid data analytical framework that integrates convolutional neural networks and bidirectional long short-term memory (CNN-BiLSTM). This is a very important topic in behavioral economics as we know that investors trade based on news and emotions which might sometimes skew the market in a completely unpredictable direction. This is a huge risk for the prediction models as they are not able to determine the behavior of investors based on historical data.

According to Temür and Yıldız (2021), a hybrid model is better at predicting stock prices than a traditional ARIMA model. They go on to propose an LSTM-ARIMA hybrid model and test it to get the results with the lowest errors and compare it to the ARIMA model. This is a novice approach as it helps understand the shortcomings of the ARIMA model. Chen et al. (2021) propose a new prediction model - CNN-BiLSTM-ECA which combines Convolutional Neural Network (CNN), Bidirectional Long Short-term Memory (BiLSTM) network, and Attention Mechanism (AM). They claim that the results were much better compared to traditional prediction models. This is an important point to understand because the authors go further in explaining that machine learning has been successfully used to simulate the non-linear relationships in financial time series. This is a major advancement that helps investors analyze non-linear data as well. They go further to elaborate that the experimental results clearly show that the LSTM model is much better at forecasting the stock prices when compared to the SVM model. Furthermore, they look at various machine learning and deep learning methods, such as support vector machine

(SVM), random forest (RF), Bayesian classifier (BC), decision tree (DT), multilayer perceptron (MLP), convolutional neural network (CNN), bi-directional long-short term memory (BiLSTM), the embedded CNN, and the embedded BiLSTM and try to find out if they can indeed be used to predict the prices of all stocks or if they can be used to predict the prices of a certain category of stocks (Zhang et al. (2021)). In the end, they conclude that this model is much better overall. In addition, to support this point, even more, Van-Dai, CHUAN-MING, and Tadesse (2020) shows that the portfolio constructed using the LSTM model outperformed the traditional regression models and this portfolio beat the S&P 500 on both active returns and Sharpe ratios. The active returns and Sharpe ratios are two of the most important factors investors look at to see if the stocks are profitable now and if they would be profitable in the future. In addition to the forecasting models stated above, LOLEA, PETRARIU, and GIURGIU (2021) compares the ARIMA model to Prophet, KNN, and Neural Networks in terms of stock market forecasting and seeks to explain which method might be the best for predicting the stock prices. Finding a model that works for all stocks is essential as it helps investors decide what stocks to invest their money in and if they can gain a sizable profit from it in the future.

From the above, it can be observed that there are many different models out there which are trying to solve the same issue - forecasting stock prices. As the prices of stocks are affected by a multitude of factors and the same factors affect the value stocks and growth stocks differently, this paper will look at the value stocks only. In this paper, we will randomly pick 10 value stocks included in the S&P 500 index and look at their prices over 10 years from 2009 to 2019. This period is specifically chosen to make sure that the outlier years could be avoided and the data would be a little more consistent. The components of the ARIMA model would be consistently altered to create the best prediction model and then this model will be compared to the LSTM model to see which among the two models is better at predicting the stock prices or if there is a need for a hybrid model.

Research Questions and Hypothesis

Research Question

The research questions that this research is looking at is “Does increasing the number of lagged observations in the ARIMA model improve the prediction accuracy of average trading prices of value stocks on a 10-year period?”. This research focuses on value stocks and aims to improve the accuracy of average trading price forecasts over a 10-year period from 2009 to 2019. By incorporating more historical data through increased lagged observations, the study seeks to determine if the ARIMA model cab provide enhanced predictions. The findings of this study will have practical implications for investors, guiding their investment decisions and potentially improving risk management practices.

Hypothesis

This study is trying to test how ARIMA model is better at predicting stock prices more accurately. Does increasing the number of lagged observations in the ARIMA model indeed increases the prediction accuracy of the average trading prices of value stocks on a 10-year period? A positive relationship between the number of lagged observations in the ARIMA model and prediction accuracy is expected. The target variable for this study is the average trading process of value stocks. The metrics used to evaluate the prediction accuracy are mean absolute error (MAE), mean squared error (MSE) and Root Mean Squared Error (RMSE). A lower value of the above metrics indicates higher prediction accuracy, as it reflects a smaller deviation between the predicted and actual values.

Methods

The stock market is very volatile and unpredictable. It is very difficult to say if the price of a stock is going to go up or down in the future. However, there is research done in this area to see if a model can be devised to predict the prices of stocks. For this study, 100 value stocks of the S&P 500 index have been looked at for a period from Jan 1, 2009 to Dec 31, 2019. This data has been fitted through ARIMA model and results have been compared with with the actual prices of stocks. Visualizations have been created to show and compare the results. Details have been outlined in the below sections.

Participants

The stock prices of 10 S&P 500 stocks from 10 different sector over 10 years from 2009 to 2019 has been used for this study. The participants for this study have been taken from https://finance.yahoo.com. The data is taken from Jan 1, 2009 till Dec 31, 2019 for each of the above stocks.

The participants have been selected based on the market capitalization of the company. This means that the companies with largest market capitalization have been looked at and studied. These companies have been dissected based on sectors.

The participants that have gone through stock splits in this period, they will be excluded from this study. Similarly, if a participant has gone through a major financial event, they will also be excluded.

Only companies which are more than 20 years old with a global presence will be looked at. These companies have been to be headquartered in the United States.

The list of stocks chosen for this research shown in Table 1 and 2.

Procedure

Data has been downloaded from https://finance.yahoo.com. CSV files have been created to store the data, which has been loaded into R Studio. The data has been tested using the ARIMA model to forecast the prices for 2020 and also tested using the LSTM model. Once the forecast is obtained from both models, several visualizations have been created and compared side by side. Actual data for these stocks from 2020 has also been downloaded and visualized. All three visualizations have been compared to identify the forecast that comes the closest to the actual data in 2020.

The project has run for 12 weeks, starting with data collection from the website and storing it in CSV format in the first week. In the second week, the closing price for each of these stocks for each trading day for the period between Jan 1, 2010, and Dec 31, 2019, has been examined. In the third week, the data has been tested using the ARIMA model, and forecasts have been visualized. In the fourth week, the data has been tested using the LSTM model, and forecasts have been visualized. Actual results for 2020 have been downloaded from https://finance.yahoo.com in the fifth week, and visualizations have been created. In the sixth week, visualizations from the ARIMA and LSTM models have been compared with the actual data visualizations to determine which model is better at predicting stock prices. In the seventh week, parameters of the ARIMA model have been altered to check if this improves the forecast's accuracy.

Preprocessing has been performed on the collected stock price data to ensure its suitability for analysis. Techniques such as interpolation or forward/backward filling have been used to handle any missing values. Only the closing stock price data has been used in the analysis. The data has been organized into a time series format, with each data point representing a specific date.

Measures

For this research project, secondary data from the stock market will be used, being obtained from https://finance.yahoo.com. The data is already existing and will not involve conducting surveys or preparing questionnaires. The news for the shortlisted companies has been examined, identifying any major changes such as mergers and acquisitions, changes in leadership, lawsuits, etc., that have occurred in the last 10 years. Companies which have gone through major changes have been excluded from the analysis.

Data Description

The list of the 100 value stocks used in the analysis are shown in Table 1 and Table 2.

The stock price data was downloaded from Yahoo Finance from Jan 1, 2009 to Dec 31, 2019. This data had several variables such as Date, Open (price at which the stock price opened on that day), High (highest price the stock was traded on that day), Low (lowest price at which the stock was traded on that day), Close (price at which the stock closed on that day) and Volume (number of shares that were traded on that day).

For this analysis, only the 'Close' variable which is the closing stock price data is used. The data was consistent and clean. Thus, no additional cleansing was required. The dataset consists of the closing prices for the selected value stocks from 2009 to 2019. For this study, we are looking at 10 value stocks each from 10 different sectors. The dataset consists of a total of 100 value stocks from 10 different sectors. In total the number of rows of the data set is 2587 which means each stock has a totally of 2587 different values for closing prices for 2587 different days. The days where the stock market is closed is completed removed from the analysis as there is no closing price during this day. During the processing step, 3 value stocks namely PSX, ABBV and DELL were dropped due to lack of data.

The data is obtained from Yahoo Finance. There are no missing values, and the data is fairly clean.

The Summary Statistics for the various sectors are shown in Table 3 to 12.

This research mainly focuses on 4 different variables namely - Stock Price, Lagged Stock Price, Lag Intervals, MAE, MSE, and RMSE. Stock price is the target variable that the model is trying to predict. This is the closing price of the stock at a given day. Lagged Stock Price is the historical stock price data that we have collected from Yahoo Finance.

Lag Intervals variable is created by shifting the original stock prices by the specified lag value from 5 to 1000 in an increment of 5 (example - 5, 10, 15, 20,. . .). MAE (Mean Absolute Error) is a variable calculated for each lag intervals and added to the corresponding row. MSE (Mean Squared Error) is a variable calculated for each lag interval and added to the corresponding row. RMSE (Root Mean Squared Error) is a variable calculated for each lag interval and added to the corresponding row

Data Visualization

The data visualization section of the report focuses on presenting a comprehensive overview of the change in closing stock prices from 2009 to 2019 for each sector. This analysis provides valuable insights into the performance and trends within the different sectors over the specified time period. To visualize the data, we have included one visualization from each sector, showcasing the closing stock price changes over the ten-year period. Each visualization provides a clear representation of the stock price fluctuations, enabling a quick comparison and identification of key trends. The visualizations are designed to be easy to interpret and highlight the relative performance of stocks within each sector. The use of color, labels, and axes ensures that the information is conveyed effectively to the readers.

By examining the visualizations, it becomes apparent how different sectors have experienced varying levels of growth, stability, or volatility over the ten-year period. For example, the value stocks in the Energy sector have remained relatively stable while the technology sector has done very well. The value stocks in the Materials and Utilities sectors are also performing well. These visualizations help in identifying any major events or market trends that might have influenced the stock prices within each sector. For instance, if there are significant spikes or drops in stock prices during certain years, it could indicate the impact of economic factors, industry-specific events, or regulatory changes on the sector. The visualizations also facilitate comparisons across sectors, allowing readers to assess the relative performance and growth rates. This comparative analysis can aid in identifying sectors that outperformed or under performed during the specified period, offering valuable insights for investment strategies or further analysis. Overall, the data visualization section serves as a powerful tool for presenting a comprehensive picture of the change in closing stock prices across sectors from 2009 to 2019. The visualizations provide a clear and concise representation of the data, enabling readers to gain insights into the performance of different sectors and make informed decisions based on the trends observed.

An example of visualization by sector is shown below for Consumer Discretionary Sector. Over the years most of the value stocks in this sector have increased in price. For example, for KSS the stock price in early 2010 was around 20 USD and in 2020 the stock price was around 100 USD. The visualizations for the 10 sections analyzed in this study are included in the Appendix A – Figures. These visualizations show how the value stocks have performed compared against other value stocks in the same sector.

Data Analysis

The major part of the research has been conducted in Python and and R using R (Version 4.1.2; R Core Team, 2021) and the R-packages *apaTables* (Version 2.0.8; Stanley, 2021), *corrplot2021* (Wei & Simko, 2021), *dplyr* (Version 1.0.7; Wickham, François, Henry, & Müller, 2021), *forcats* (Version 0.5.1; Wickham, 2021a), *ggplot2* (Version 3.4.0; Wickham, 2016), *jpeg* (Version 0.1.10; Urbanek, 2022), *kableExtra* (Version 1.3.4; Zhu, 2021), *knitr* (Version 1.39; Xie, 2015), *magick* (Version 2.7.4; Ooms, 2023), *pander* (Version 0.6.5; Daróczi & Tsegelskyi, 2022), *papaja* (Version 0.1.0.9999; Aust & Barth, 2022), *PerformanceAnalytics* (Version 2.0.4; Peterson & Carl, 2020), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *quantmod* (Version 0.4.22; Ryan & Ulrich, 2023a), *readr* (Version 2.1.1; Wickham, Hester, & Bryan, 2021), *rmarkdown* (Version 2.14; Xie, Allaire, & Grolemund, 2018; Xie, Dervieux, & Riederer, 2020), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.1.6; Müller & Wickham, 2021), *tidyr* (Version 1.1.4; Wickham, 2021b), *tidyverse*

(Version 1.3.1; Wickham et al., 2019), *tinylabels* (Version 0.2.3; Barth, 2022), *TTR* (Version 0.24.3; Ulrich, 2021), *xts* (Version 0.13.1; Ryan & Ulrich, 2023b), and *zoo* (Version 1.8.9; Zeileis & Grothendieck, 2005).

These programming languages offered robust capabilities for data analysis and processing, enabling comprehensive investigations into the dynamics of value stocks. To initiate the research, a dataset encompassing 10 value stocks from each of the 10 distinct sectors was acquired from Yahoo Finance. The dataset covered the time period from January 1, 2009, to Dec 31, 2019. A careful curation process was undertaken, where all columns except for the "Close" column, representing the closing price of the stocks on each specific day, were removed. This refined dataset was then transformed into a new dataframe comprising two essential columns: date and closing price. To gain deeper insights into the behavior of value stocks, visualizations were meticulously prepared. A dedicated visualization was generated for each sector, facilitating a comprehensive examination of stock price changes and the identification of any outliers or peculiar trends.

The research then proceeded with a sector-specific analysis, starting with an in-depth exploration of the Consumer Discretionary sector. Summary statistics were calculated for each of the 10 value stocks within this sector and stored in a dataframe. These statistics provided valuable insights into the performance and characteristics of the selected value stocks. To experiment with the impact of lagged observations on the analysis, a lag interval was defined. Subsequently, an ARIMA model was developed and tested using one value stock from the Consumer Discretionary sector, namely Ford (F). It is important to note that the selection of Ford as the exemplar stock was arbitrary, as the intention was to utilize all stocks for training, testing, and comparison purposes.

Within the ARIMA model, a lagged dataset was created iteratively based on specific values within the defined lag interval. This process involved constructing separate lagged training and testing datasets, encompassing data from 2009 to 2018 and 2019, respectively.

Using the trained ARIMA model, the stock prices for the year 2019 were forecasted and compared to the actual values. To evaluate the accuracy and performance of the model, metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were calculated for various lagged observations. These metrics were stored in a dataframe, providing comprehensive information on the model's predictive capabilities. To assess the relationship between the number of lagged observations and the resulting errors, line graphs were plotted, depicting the variations in MAE, MSE, and RMSE values. This analysis aimed to determine if the errors exhibited any noticeable patterns or tendencies as the number of lagged observations in the ARIMA model increased.

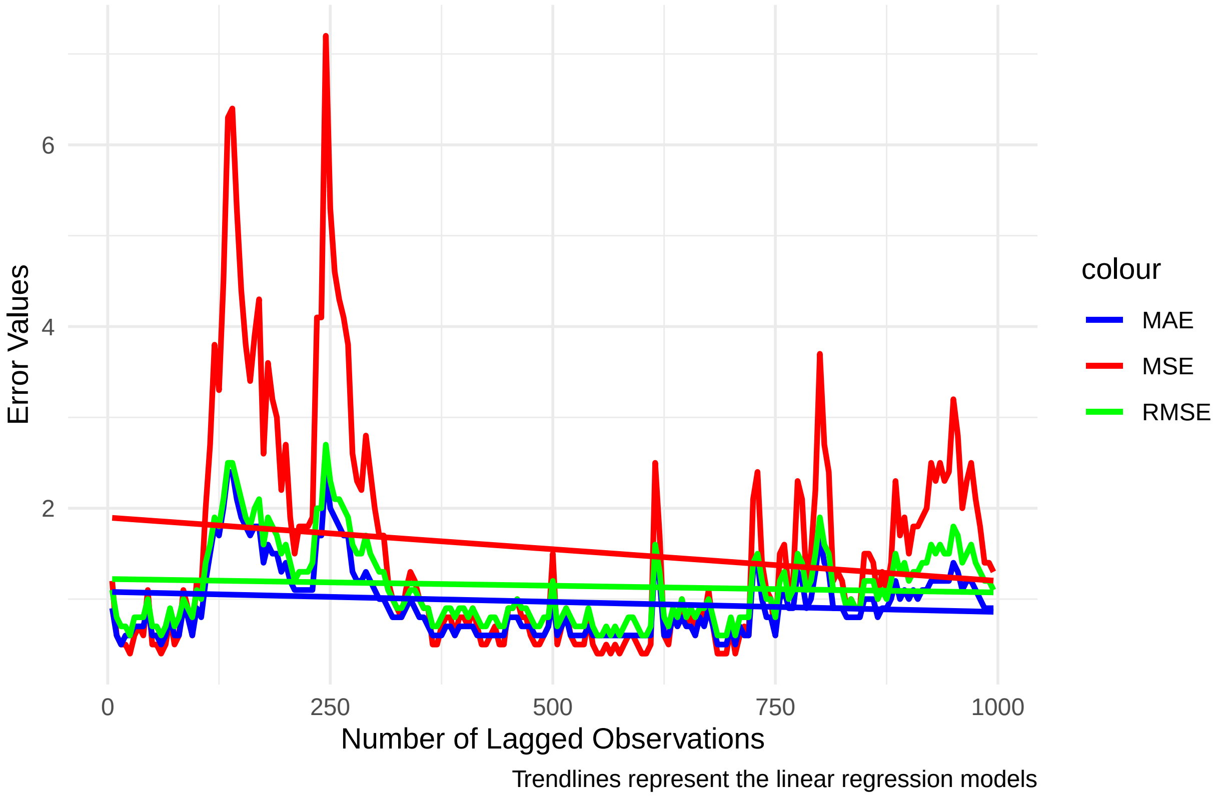
The research process described above was repeated for every sector and each value stock within the dataset, allowing for a comprehensive examination of value stock dynamics across all sectors.

In conclusion, the research undertaken in this study successfully explored the dynamics of value stocks using Python and R and provided valuable insights into the behavior of value stocks and the impact of lagged observations on the predictive performance of the ARIMA model.

Results

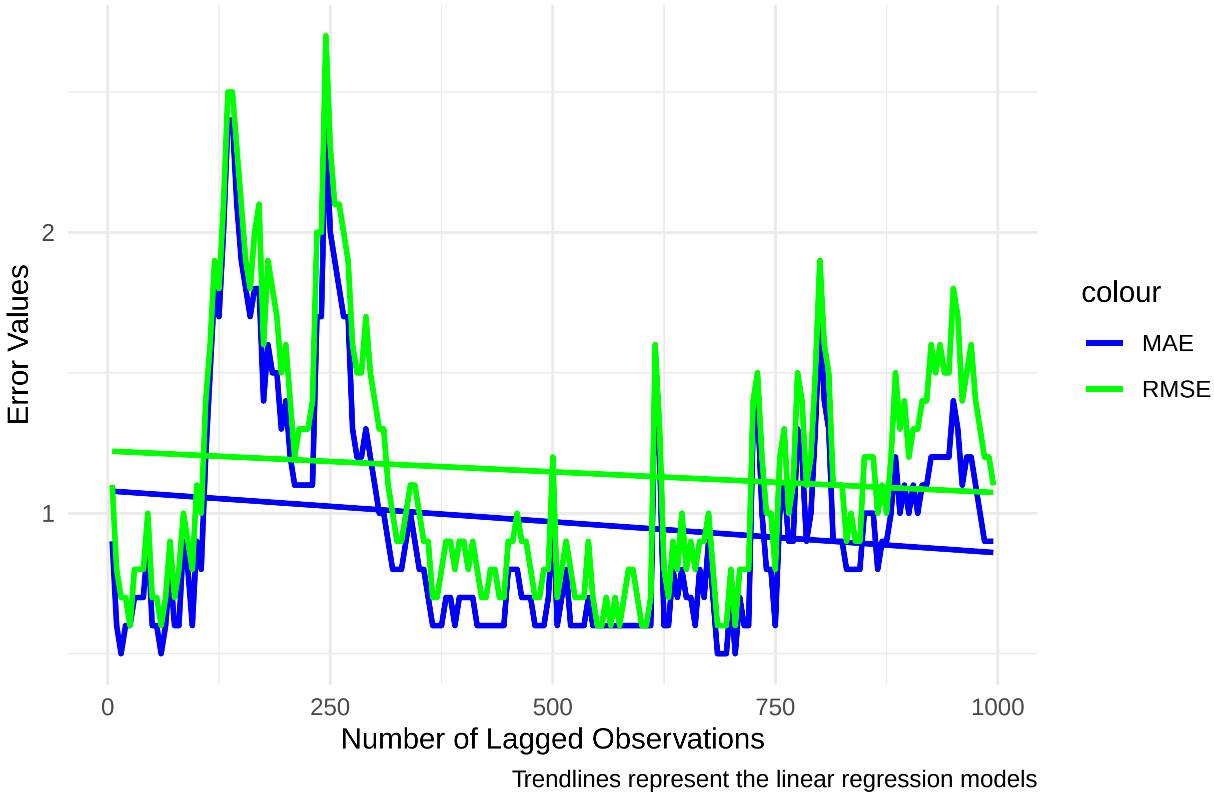
The first step of the analysis was to look at the data and preprocess it. In this step, correlation matrix was analysed to see if there was any correlation between the value stocks in the same sector. The correlation matrix showed that there is a strong correlation between the value stocks in the same sector. Because of this, it is often observed that these stocks perform in a similar way during economic cycles. The main component of this study is to compare the results of the models. This study compares the Number of lagged observations of the ARIMA model with the model's MSE, MAE and RMSE.

Figure 1: F − Forecast Results



Looking at the graph, the value of MSE is much higher compared to RMSE and MAE, thus, the metric MSE is removed and only MAE and RMSE are taken into account. This process is repeated for all 100 Value stocks. For the purpose of this study, the above 2 graphs are shown here as a sample. From the remaining value stocks, a sample of 18 Value stocks are chosen and shown in Appendix A - Forecast Results.

Figure 2: F − Forecast Results



In this study, the relationship between the number of lagged observations and the mean absolute error (MAE) and root mean squared error (RMSE) values in forecasting was examined. The datasets we created from the result of the ARIMA model has been analyzed, with each dataset corresponding to a specific stock symbol and containing information on lagged observations, MAE, and RMSE.

The results revealed a clear trend, whereby an increase in the number of lagged observations was associated with a consistent decrease in both the MAE and RMSE values. This suggests that the inclusion of more historical data points in the forecasting process leads to improved prediction accuracy.

Line graphs with trendlines were created for each dataset to visually represent this relationship. A sample is shown in Appendix A - Forecast Results. The line graphs depicted the decline in MAE and RMSE values as the number of lagged observations increased. The trendlines, which represented the linear regression models, further emphasized this downward trend. A visual examination of these charts revealed a consistent correspondence between an increase in the number of lagged observations and a decrease in both error metrics. These findings support the notion that incorporating a greater number of past observations enhances the accuracy of the forecasting model.

The results underscore the importance of including an appropriate number of lagged observations in time series forecasting. By doing so, forecast accuracy, as measured by MAE and RMSE, can be significantly improved. These findings have implications for researchers and practitioners in the field of forecasting, providing valuable insights for making informed decisions regarding the selection of lagged observations for their models.

Discussion

The main component of this project is to see if increasing the lagged observations in the ARIMA model increases the prediction accuracy of the model. The metrics used in this study to analyze this are Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean squared Error (RMSE).

Summary of Results

If the MSE value increases when the number of lagged observations is increased, this means that the accuracy of the model is reducing. If the MAE value increases when the number of lagged observations is increased, this means that the accuracy of the model is reducing. If the RMSE value increases when the number of lagged observations is increased, this means that the accuracy of the model is reducing. To sum up, if the errors are increasing when the number of lagged observations is increased, this means the accuracy of the model is deteriorating.

Looking at the charts for the value stocks for the Consumer Discretionary sector, the following observations can be made. As the number of lagged observations in the ARIMA model is increased, the value of MAE and RMSE show a downward trend. This suggests that as the number of lagged observations is increased the model performance increases, thus, increasing the efficiency of stock price prediction.

Comparison with previous Literature. To understand the significance of the findings of this study, it is important to compare it with what previous studies have discovered. Many researchers have explored how adding more past observations in time series forecasting models, like ARIMA, can affect their accuracy. While this study didn't directly compare its results to specific studies, the general consensus in the literature supports the idea that including more past observations can improve the accuracy of these models.

This study's findings are consistent with existing research, emphasizing the importance of including an adequate number of past observations in the ARIMA model. It was observed that as the number of lagged observations was increased, the MAE and RMSE values decreased. This suggests that considering a longer history of data improves the model's ability to predict stock prices more effectively.

However, it's important to note that further investigation and validation are needed. Conducting comparative analyses with relevant studies can help us gain a more comprehensive understanding of the research field and verify the findings of our study.

Limitation

The dataset used in this analysis is limited this study only looked at 100 value stocks. Although an attempt was made to cover 10 different sectors, it is still not an enough sample for the stock market. Because of this, this model cannot be generalized for all stocks in the market. Assumptions were made that the past patterns will continue in the future which is not always the case as stock market is very volatile and is impacted by a multitude of factors. The models were only evaluated based on MAE and RMSE which might not be enough to capture the full range of forecasting errors. Similarly, only 10 years of the stocks' history was taken to train and test the models. It can be argued that different time horizons produce different forecasting results.

Future Directions

This study could be an stepping stone for developing other forecasting models for predicting the stock prices for investors to help them make buy, sell or hold decisions. Future models could use 10,000 stocks from various sectors to analyze the models or use a time frame of more than 10 years. This model only looks at number of lagged observations in the ARIMA model. New Models can be created which look at the other components of the ARIMA model as well such as Moving Average, Integrated and so on. Similarly, this study can be expanded into forecasting things in other domains such as economic forecasting, climate forecasting, demand forecasting and so on.

Importance/Complications

This study explores the relationship between lagged observations and forecast accuracy, which is crucial for improving the reliability of predictions. Understanding how the inclusion of past data affects forecast performance can lead to more accurate and precise forecasting models, benefiting various fields such as finance, economics, supply chain management, and environmental forecasting. Accurate forecasting enables better decision-making, resource allocation, and planning. By examining the impact of lagged observations on forecast accuracy, this study provides insights that can inform practitioners in selecting appropriate historical data to improve the reliability of their forecasts. It has practical implications for organizations and individuals relying on forecasts to make informed decisions.

The study's findings heavily rely on the quality and availability of the dataset used.

Complications may arise if the dataset contains errors, missing values, or inconsistencies. Additionally, limited availability of comprehensive datasets or restricted access to proprietary data may limit the generalization and robustness of the study's conclusions. The study's analysis is based on the assumption that the relationship between lagged observations and forecast accuracy is stable and linear. However, this assumption may not hold true in all scenarios. Factors such as structural changes, non-linear patterns, or external shocks can complicate the relationship, leading to less straightforward interpretations of the findings.

References

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of ARIMA and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edsgao&AN=edsgcl.424531452&site=eds-live&scope=site&custid=s4786267

Assous, H. F., Rousan, N. A.-., Najjar, D. A.-., & Najjar, H. A.-. (2020). Can international market indices estimate TASI's movements? The ARIMA model. *Journal of Open Innovation: Technology, Market, and Complexity*, *6*(2). Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edsgao&AN=edsgcl.632211894&site=eds-live&scope=site&custid=s4786267

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from https://github.com/crsh/papaja

Barth, M. (2022). *tinylabels: Lightweight variable labels*. Retrieved from https://cran.r-project.org/package=tinylabels

Chen, Y., Fang, R., Liang, T., Sha, Z., Li, S., Yi, Y., … Song, H. (2021). Stock price forecast based on CNN-BiLSTM-ECA model. *Scientific Programming*, 1–20. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edb&AN=151328672&site=eds-live&scope=site&custid=s4786267

Daradkeh, M. K. (2022). A hybrid data analytics framework with sentiment convergence and multi-feature fusion for stock trend prediction. *Electronics (2079-9292)*, *11*(2), 250. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edb&AN=154817098&site=eds-live&scope=site&custid=s4786267

Daróczi, G., & Tsegelskyi, R. (2022). *Pander: An r 'pandoc' writer*. Retrieved from https://CRAN.R-project.org/package=pander

Gorecka, A., & Szmit, M. (1999). Exchange rates prediction by ARIMA and neural networks models. *International Advances in Economic Research*, *5*(4), 512. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=

edsinc&AN=edsinc.A58562927&site=eds-live&scope=site&custid=s4786267

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools.* Retrieved from https://CRAN.R-project.org/package=purrr

Islam, M. R., & Nguyen, N. (2020). Comparison of financial models for stock price prediction. *Journal of Risk & Financial Management*, *13*(8), 1–19. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edb&AN=145188361&site=eds-live&scope=site&custid=s4786267

Jiang, Y. (2022). Application and comparison of multiple machine learning models in finance. *Scientific Programming*, 1–9. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edb&AN=155931063&site=eds-live&scope=site&custid=s4786267

Kulshreshtha, S., & A., V. (2020). An ARIMA- LSTM hybrid model for stock market prediction using live data. *Journal of Engineering Science & Technology Review*, *13*(4), 117–123. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edb&AN=146329186&site=eds-live&scope=site&custid=s4786267

Kumar, R., Kumar, P., & Kumar, Y. (2021). Multi-step time series analysis and forecasting strategy using ARIMA and evolutionary algorithms. *International Journal of Information Technology: An Official Journal of Bharati Vidyapeeth's Institute of Computer Applications and Management*, 1–15. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edssjs&AN=edssjs.8A5FE3D6&site=eds-live&scope=site&custid=s4786267

LOLEA, I.-C., PETRARIU, I.-R., & GIURGIU, A. (2021). ARIMA vs. MACHINE LEARNING IN TERMS OF EQUITY MARKET FORECASTING. *Annals of the University of Oradea, Economic Science Series*, *30*(2), 299–308. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=155700210&site=eds-live&scope=site&custid=s4786267

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames.* Retrieved from https://CRAN.R-project.org/package=tibble

Ooms, J. (2023). *Magick: Advanced graphics and image-processing in r*. Retrieved from https://CRAN.R-project.org/package=magick

Peterson, B. G., & Carl, P. (2020). *PerformanceAnalytics: Econometric tools for performance and risk analysis*. Retrieved from https://CRAN.R-project.org/package=PerformanceAnalytics

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ryan, J. A., & Ulrich, J. M. (2023a). *Quantmod: Quantitative financial modelling framework*. Retrieved from https://CRAN.R-project.org/package=quantmod

Ryan, J. A., & Ulrich, J. M. (2023b). *Xts: eXtensible time series*. Retrieved from https://CRAN.R-project.org/package=xts

Saxena, K. K., & Kamnge, J. S. (2020). Comparative study of exponential smoothing models and box-jenkins ARIMA model of partitioned data of daily stock prices of the CRDB bank in tanzania. *Bulletin of Pure & Applied Sciences-Mathematics*, *39E*(1), 1–30. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edb&AN=144597434&site=eds-live&scope=site&custid=s4786267

Stanley, D. (2021). *apaTables: Create american psychological association (APA) style tables*. Retrieved from https://CRAN.R-project.org/package=apaTables

Temur, A. S., & Yildiz, S. (2021). Comparison of forecasting performance of ARIMA LSTM and HYBRID models for the sales volume budget of a manufacturing enterprise. *Istanbul University Journal of the School of Business*, *50*(1), 15. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edsinc&AN=edsinc.A671758032&site=eds-live&scope=site&custid=s4786267

Temür, A. S., & Yıldız, Ş. (2021). Comparison of forecasting performance of ARIMA LSTM and HYBRID models for the sales volume budget of a manufacturing enterprise.

Istanbul University Journal of the School of Business Administration, *50*(1), 15–46. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=151811006&site=eds-live&scope=site&custid=s4786267

Tiwari, S., & Chaturvedi, A. K. (2021). A survey on LSTM-based stock market prediction. *Ilkogretim Online*, *20*(5), 1671–1677. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=eue&AN=150195390&site=eds-live&scope=site&custid=s4786267

Ulrich, J. (2021). *TTR: Technical trading rules*. Retrieved from https://CRAN.R-project.org/package=TTR

Urbanek, S. (2022). *Jpeg: Read and write JPEG images*. Retrieved from https://CRAN.R-project.org/package=jpeg

Van-Dai, T., CHUAN-MING, L., & Tadesse, D. A. (2020). Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading. *Applied Sciences (2076-3417)*, *10*(2), 437. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edb&AN=141573652&site=eds-live&scope=site&custid=s4786267

Wei, T., & Simko, V. (2021). *R package 'corrplot': Visualization of a correlation matrix*. Retrieved from https://github.com/taiyun/corrplot

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2021b). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., …

Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Bryan, J. (2021). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Xiao, C., Xia, W., & Jiang, J. (2020). Stock price forecast based on combined model of ARI-MA-LS-SVM. *Neural Computing & Applications*, *32*(10), 5379–5388. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edb&AN=142999341&site=eds-live&scope=site&custid=s4786267

Xiao, D., & Su, J. (2022). Research on stock price time series prediction based on deep learning and autoregressive integrated moving average. *Scientific Programming*, 1–12. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=aci&AN=156056507&site=eds-live&scope=site&custid=s4786267

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://bookdown.org/yihui/rmarkdown

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://bookdown.org/yihui/rmarkdown-cookbook

Yan, X., Weihan, W., & Chang, M. (2021). Research on financial assets transaction prediction model based on LSTM neural network. *Neural Computing and Applications*, *33*(1), 257–270. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edssjs&AN=edssjs.A14D034D&site=eds-live&scope=site&custid=s4786267

Yu, P., & Yan, X. (2020). Stock price prediction based on deep neural networks. *Neural Computing and Applications*, *32*(6), 1609–1628. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edssjs&AN=edssjs.B2DE4BEA&site=eds-live&scope=site&custid=s4786267

Zeileis, A., & Grothendieck, G. (2005). Zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, *14*(6), 1–27. https://doi.org/10.18637/jss.v014.i06

Zhang, P., Yang, J. Y., Zhu, H., Hou, Y. J., Liu, Y., & Zhou, C. C. (2021). Failure in stock price prediction: A comparison between the curve-shape-feature and non-curve-shape-feature modes of existing machine learning algorithms. *International Journal of Computers, Communications & Control*, *16*(6), 1–17. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=aci&AN=154113923&site=eds-live&scope=site&custid=s4786267

Zhu, H. (2021). *kableExtra: Construct complex table with 'kable' and pipe syntax.* Retrieved from https://CRAN.R-project.org/package=kableExtra

Appendix A - Forecast Results

Figure 3 : ABBV − Forecast Results

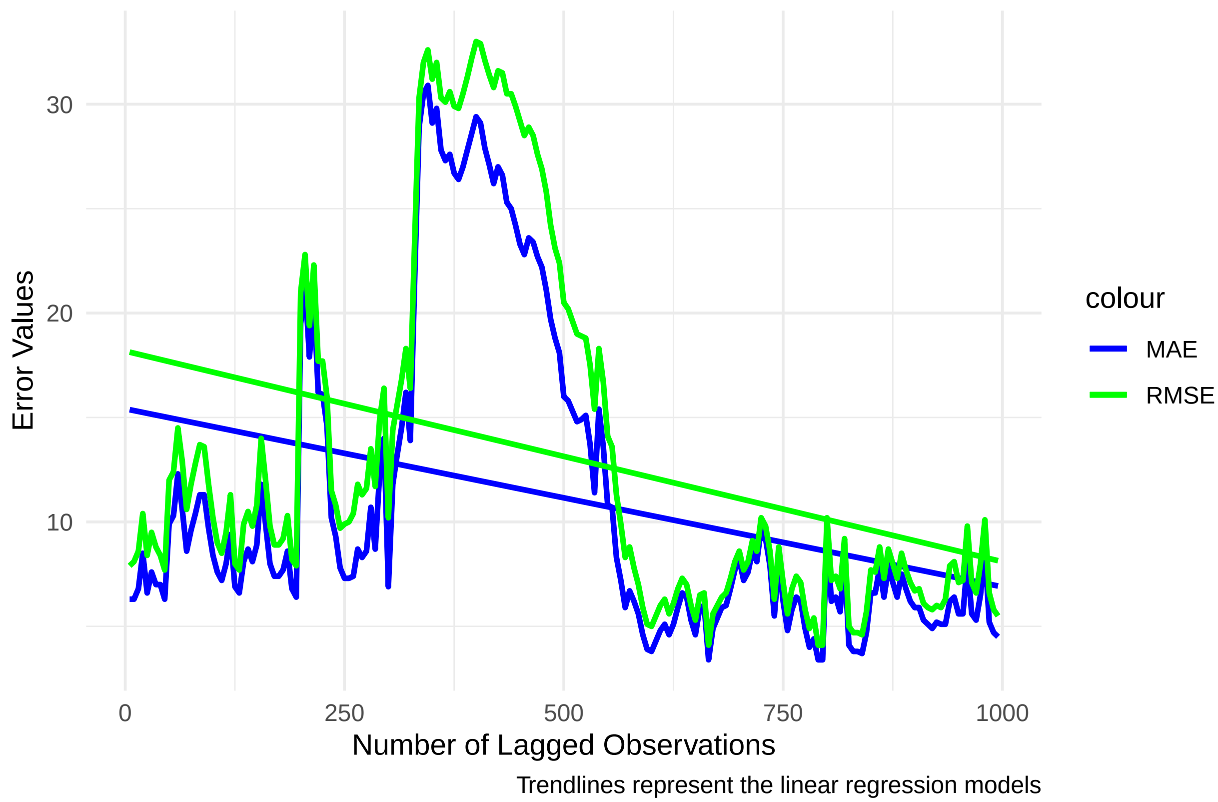


Figure 4 : D − Forecast Results

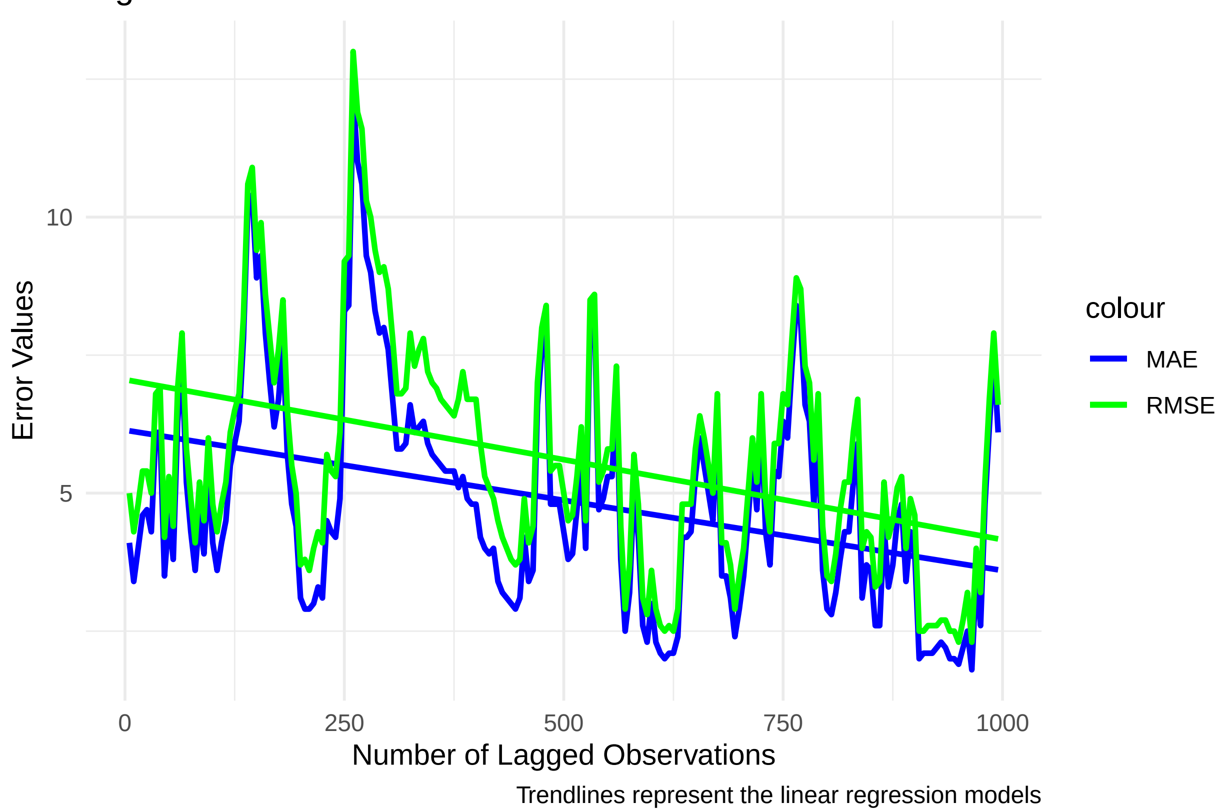


Figure 5 : TXN − Forecast Results

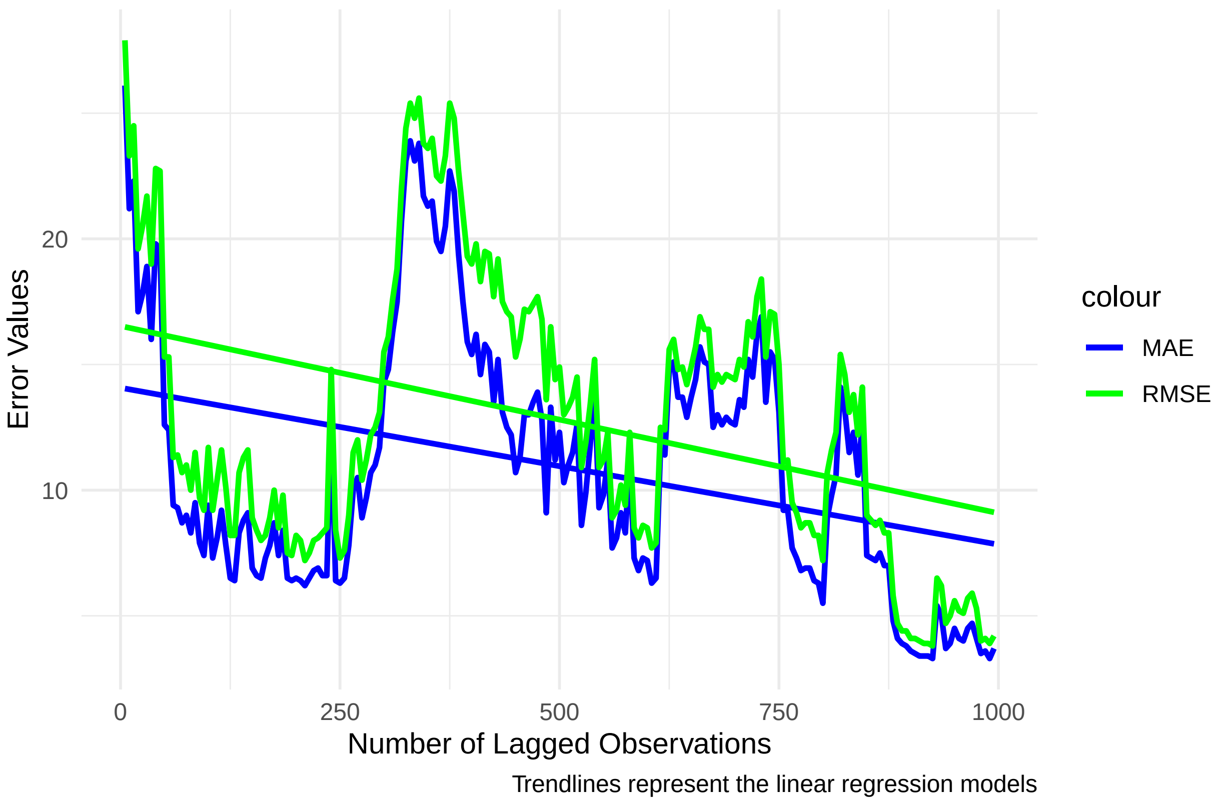
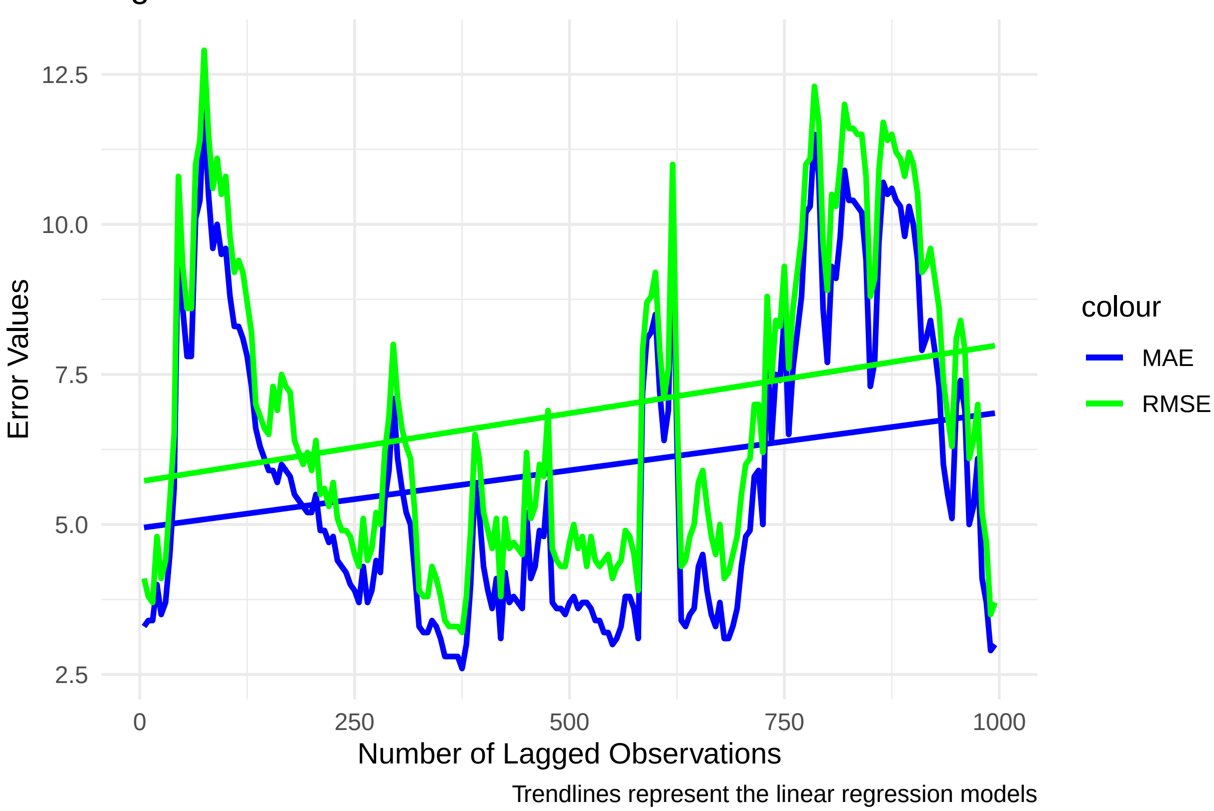


Figure 6 : K − Forecast Results



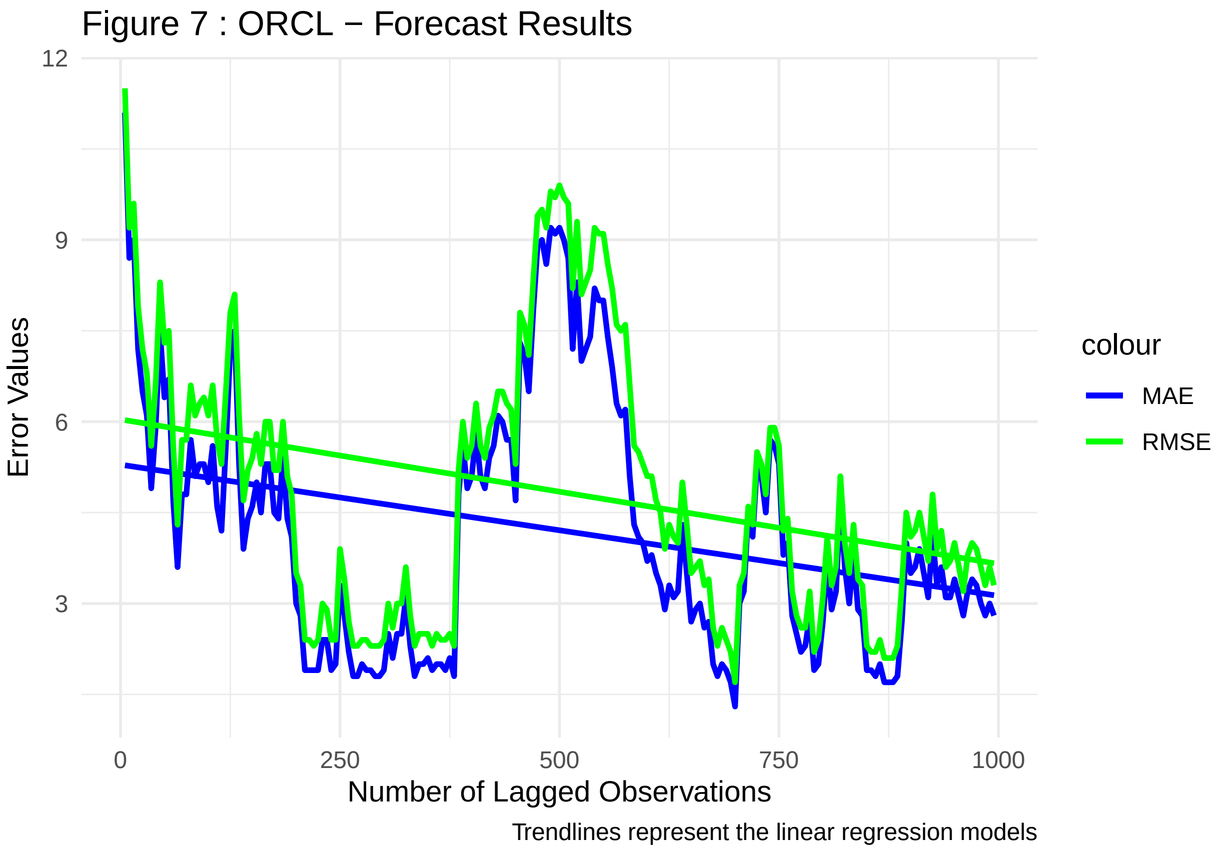


Figure 7 : ORCL − Forecast Results
Error Values
12
9
6
3
0
250
500
750
1000
Number of Lagged Observations
Trendlines represent the linear regression models
colour
MAE
RMSE

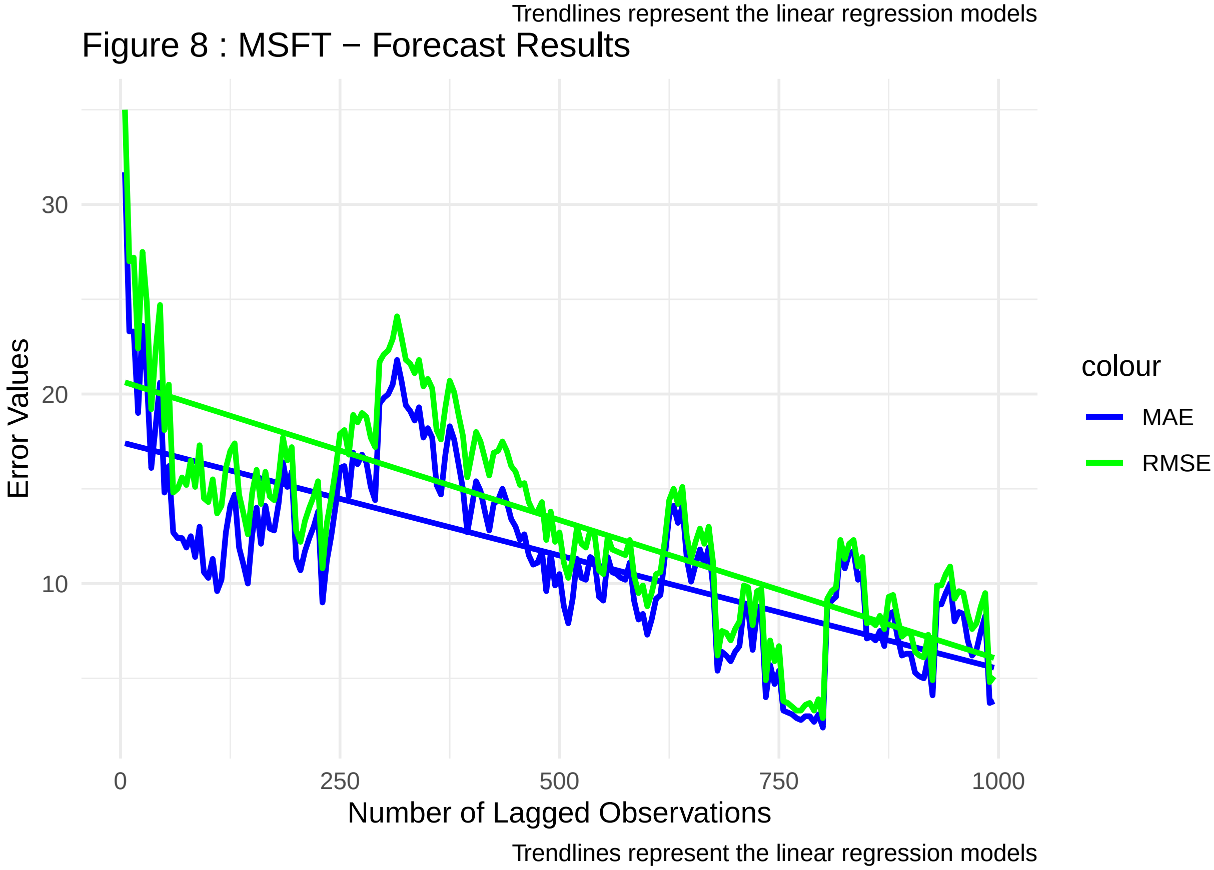
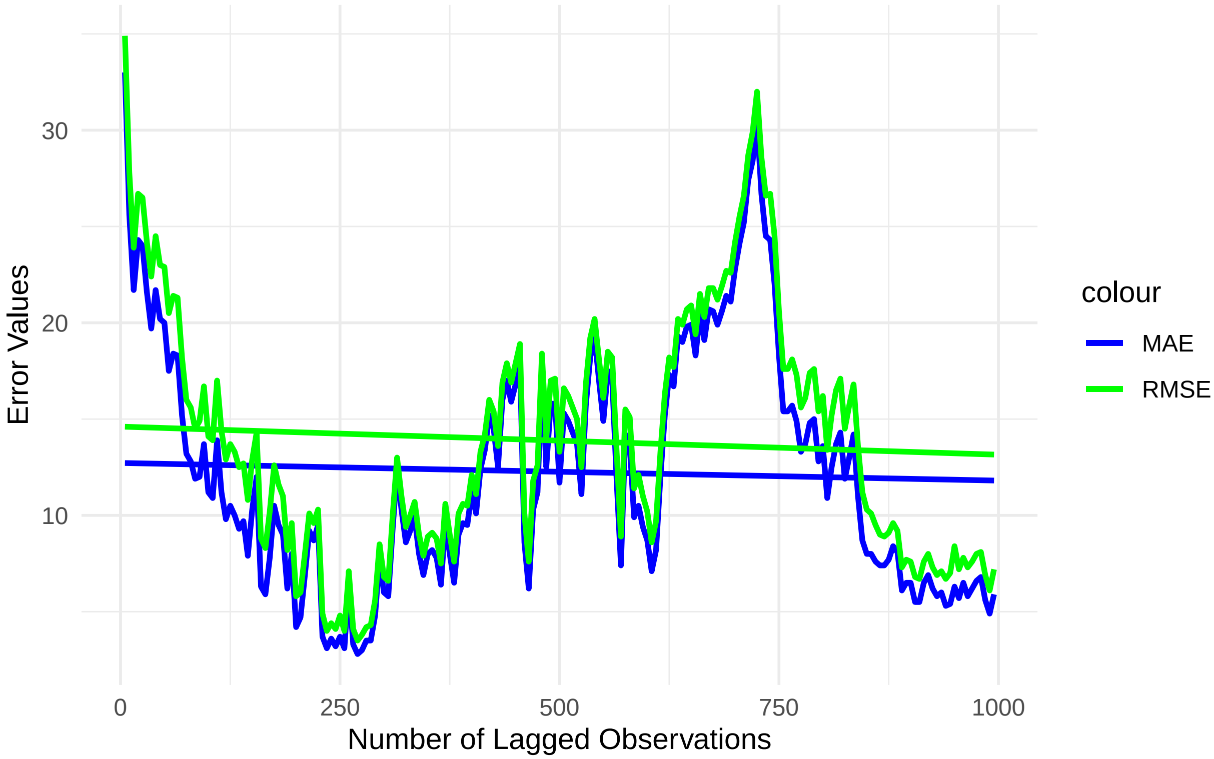


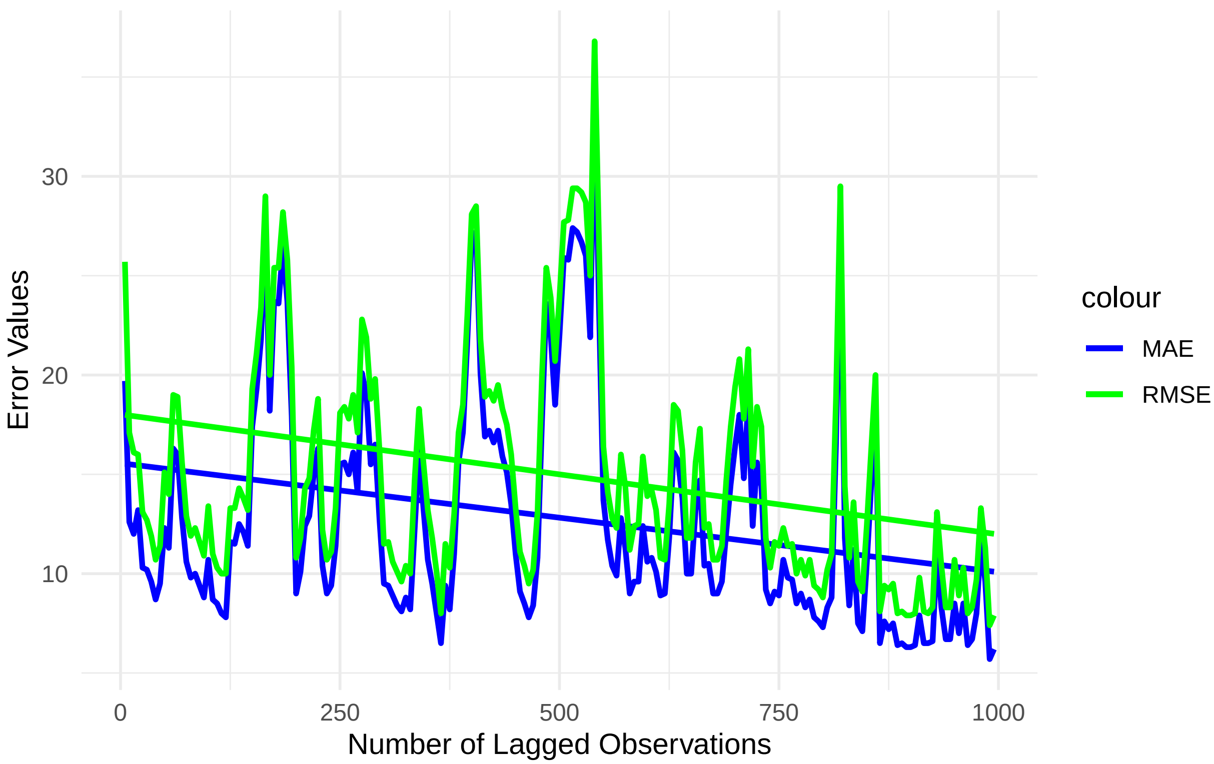
Figure 8 : MSFT − Forecast Results
Error Values
30
20
10
0
250
500
750
1000
Number of Lagged Observations
Trendlines represent the linear regression models
colour
MAE
RMSE

Figure 9 : ARE − Forecast Results



Trendlines represent the linear regression models

Figure 10 : AMGN − Forecast Results



Trendlines represent the linear regression models

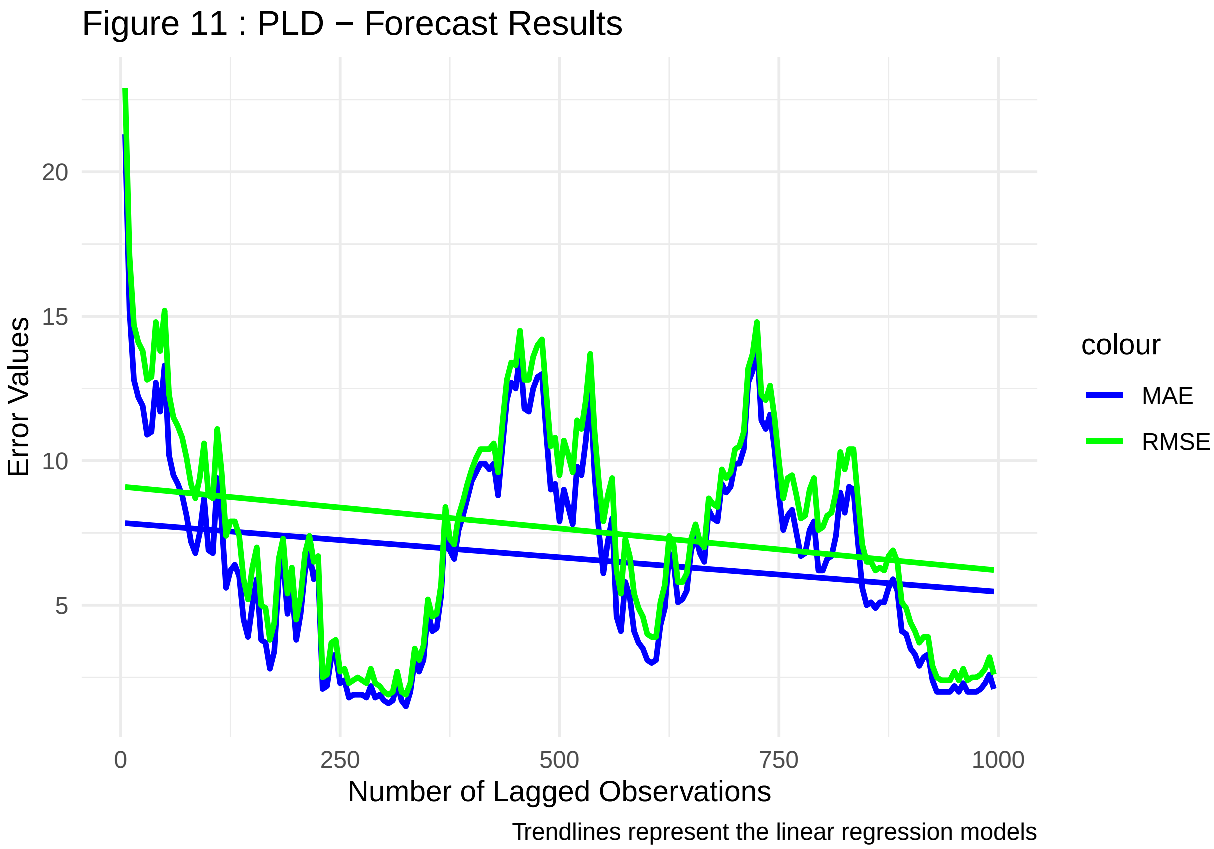


Figure 11 : PLD − Forecast Results
Error Values
20
15
10
5
0
250
500
750
1000
Number of Lagged Observations
colour
MAE
RMSE
Trendlines represent the linear regression models

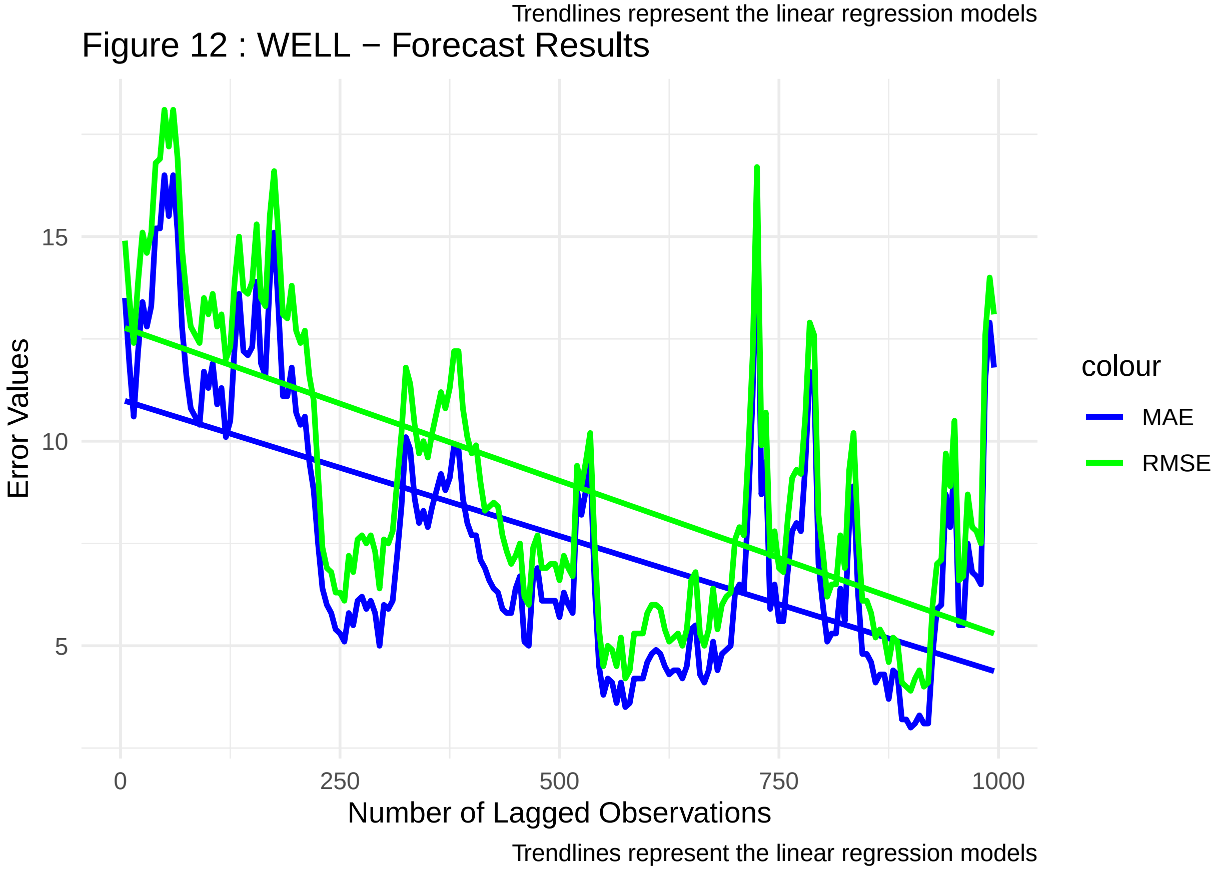
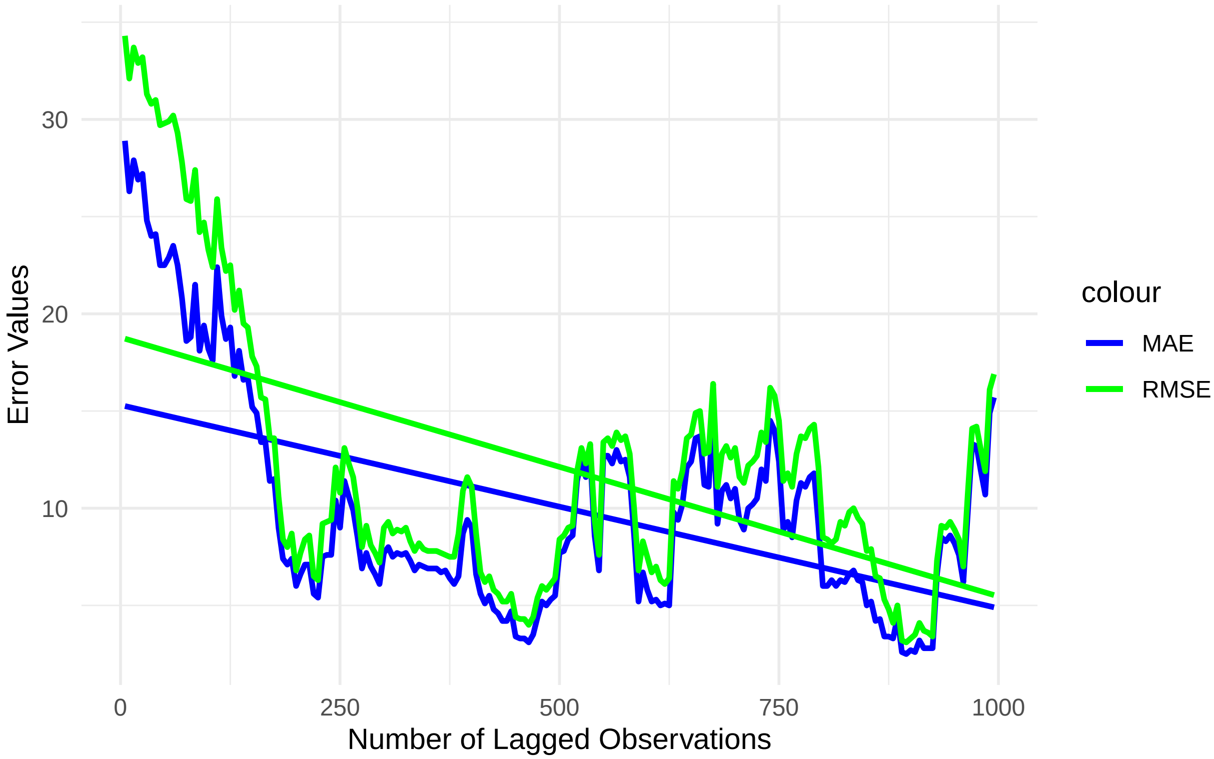


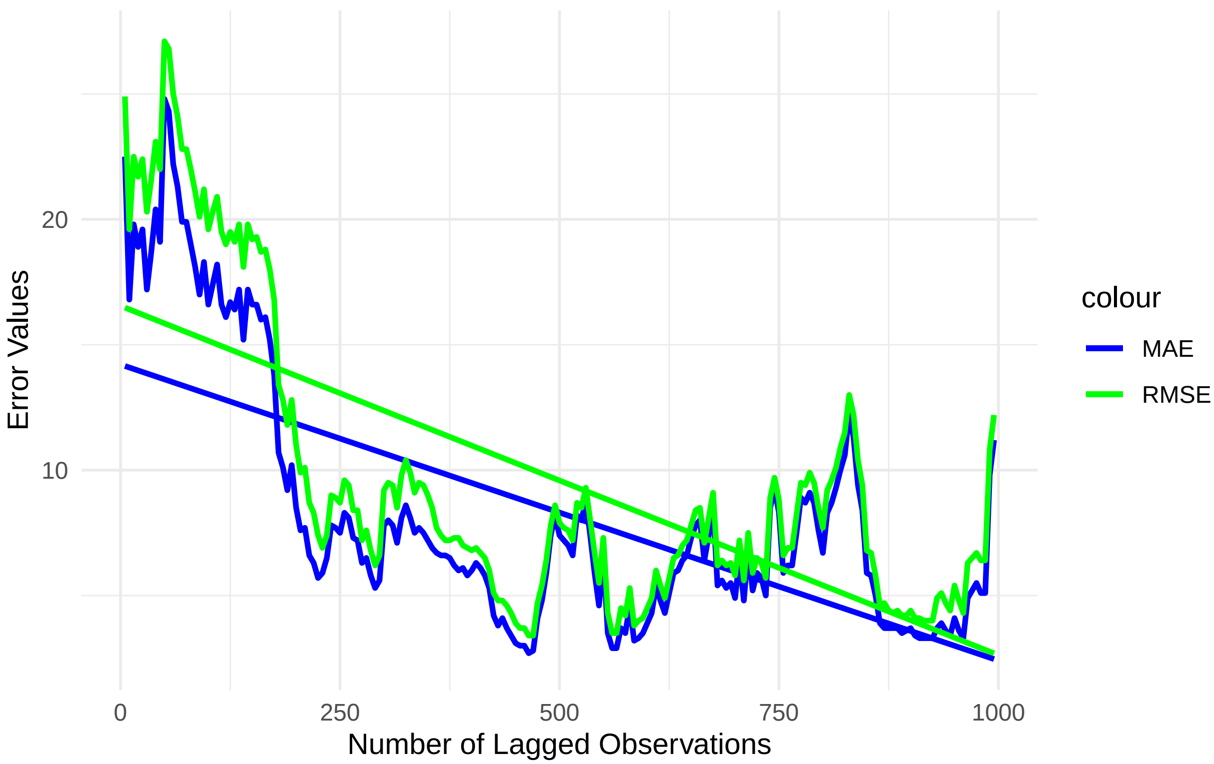
Figure 12 : WELL − Forecast Results
Error Values
15
10
5
0
250
500
750
1000
Number of Lagged Observations
colour
MAE
RMSE
Trendlines represent the linear regression models

Figure 13 : HSY − Forecast Results



Trendlines represent the linear regression models

Figure 14 : PG − Forecast Results



Trendlines represent the linear regression models

Figure 15 : KMB − Forecast Results

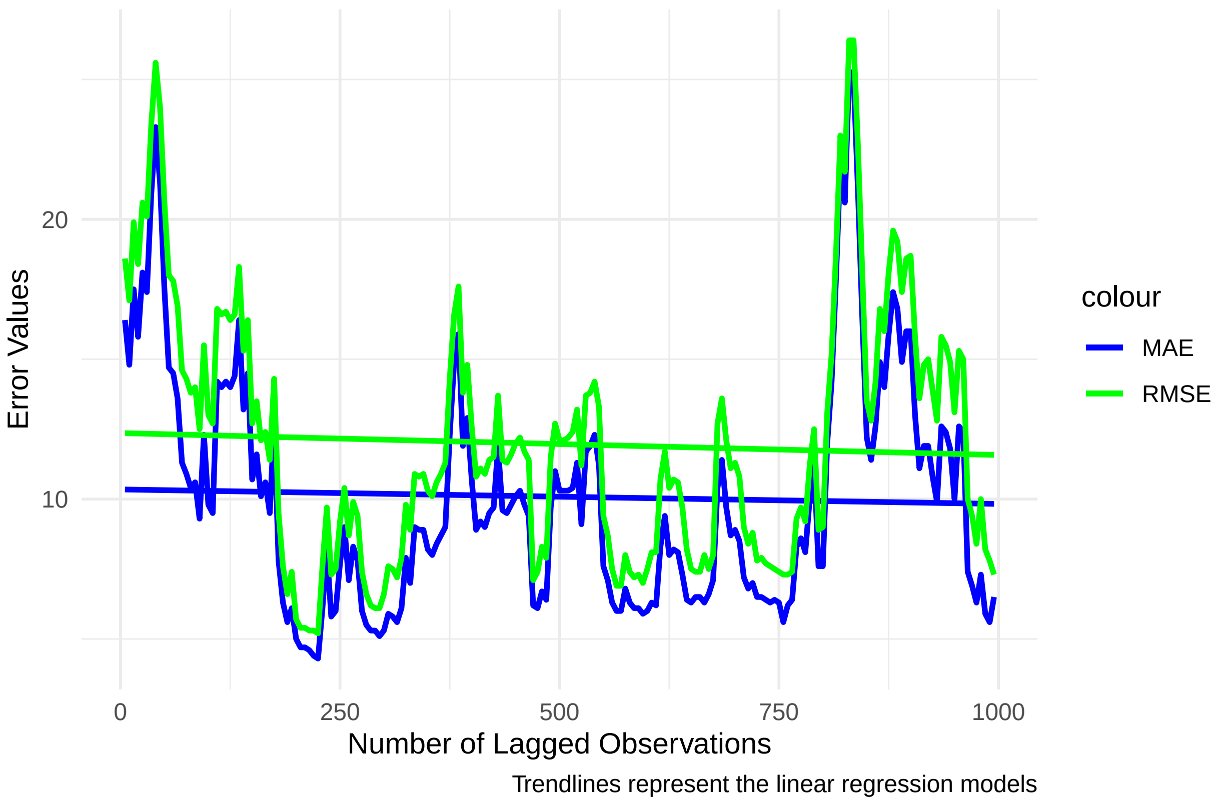


Figure 16 : KSS − Forecast Results

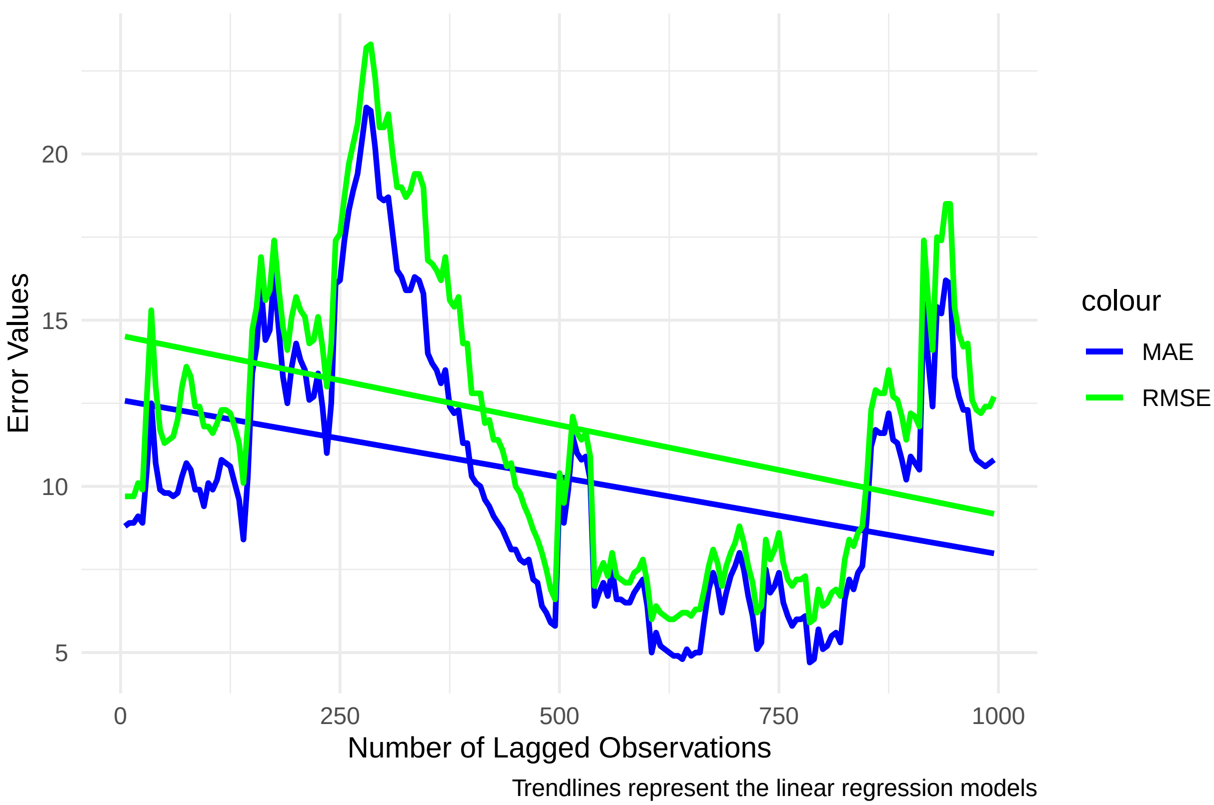
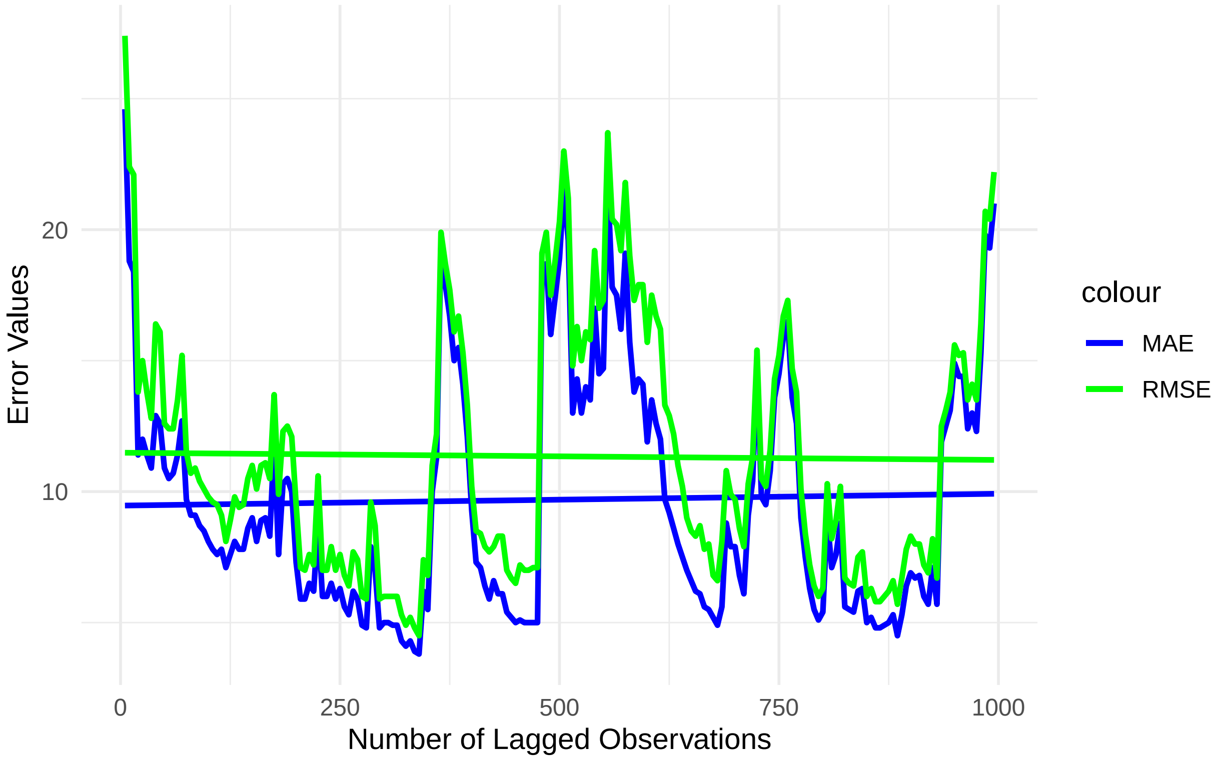
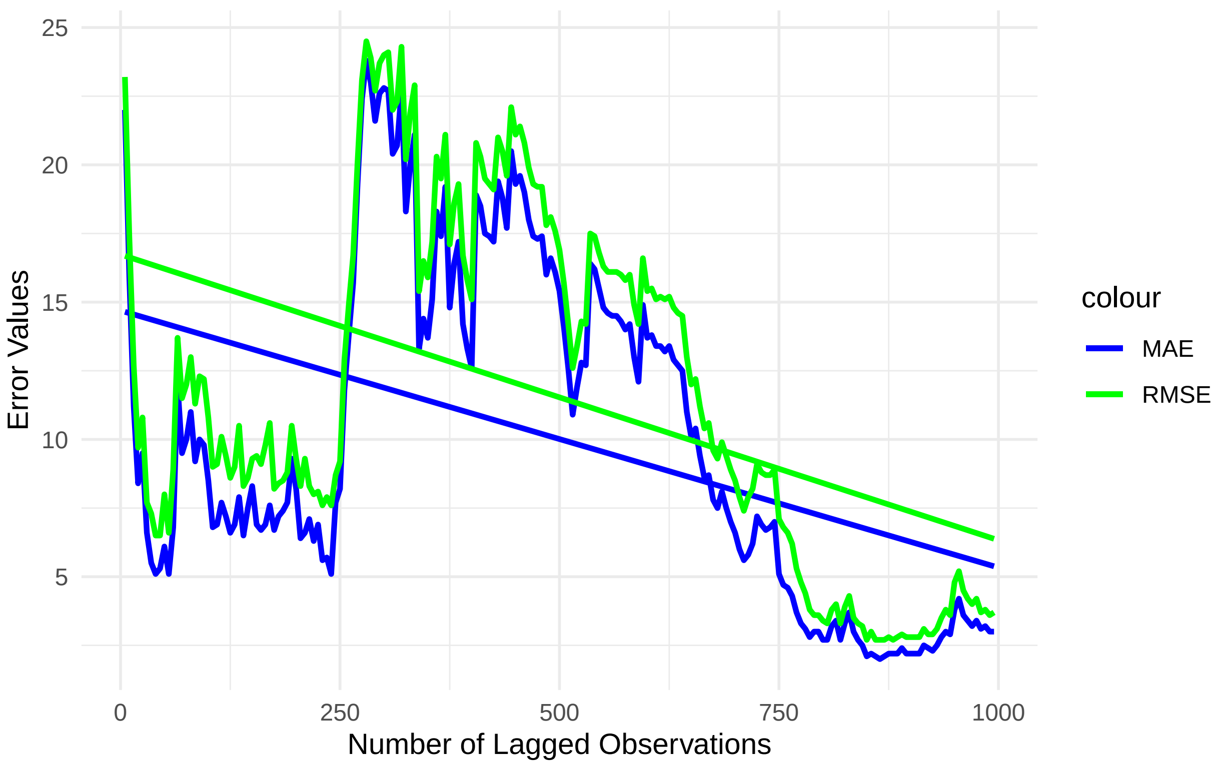


Figure 17 : HAS − Forecast Results



Trendlines represent the linear regression models

Figure 18 : BBY − Forecast Results



Trendlines represent the linear regression models

Figure 19 : CLX − Forecast Results

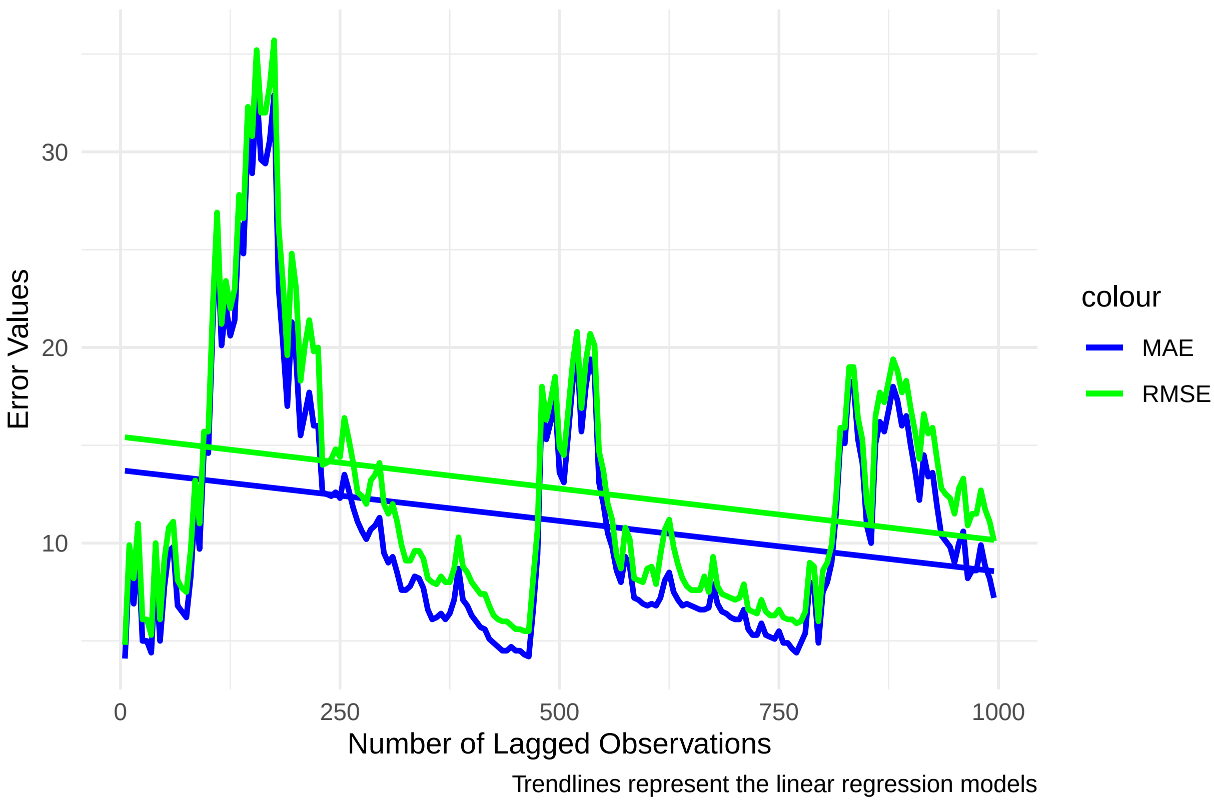


Figure 20 : O − Forecast Results

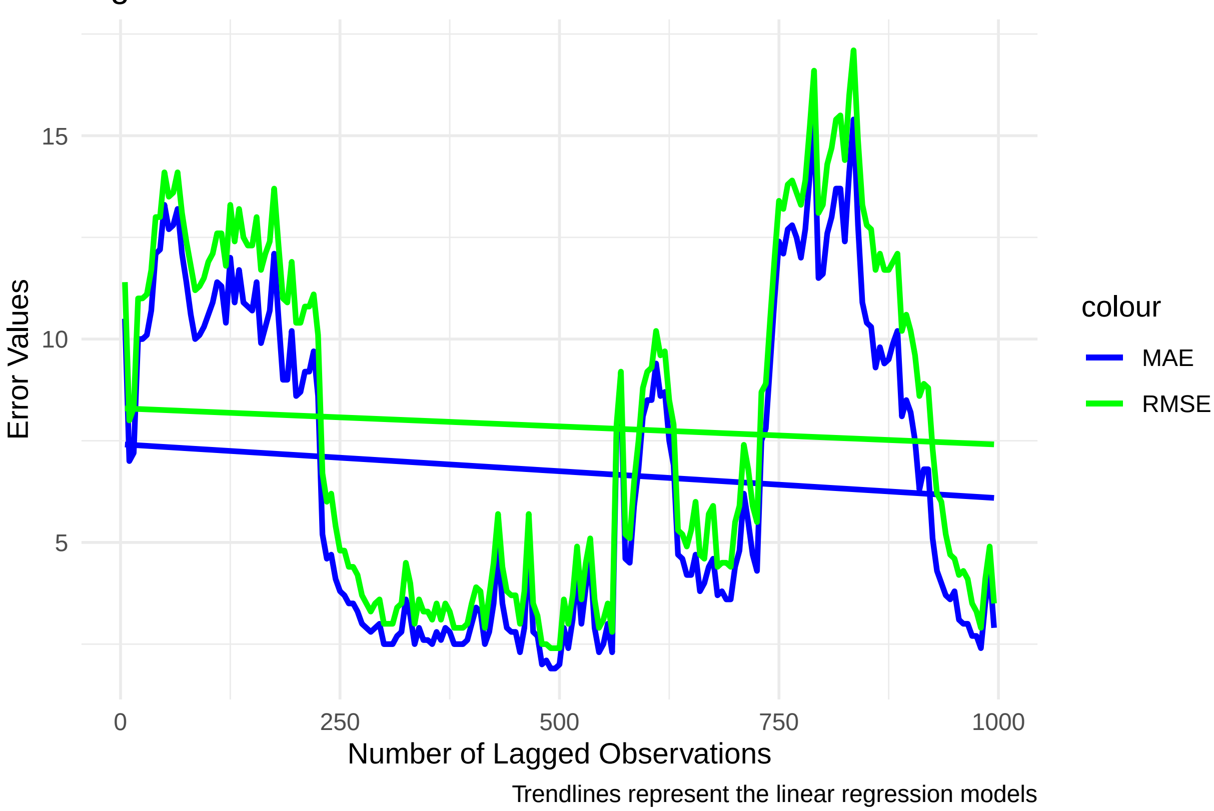


Table 1

Value Stocks Chosen from Sectors 1 to 5

Consumer.Discretionary	Consumer.Staples	Energy	Financials	Healthcare
F	GIS	XOM	BAC	BMY
BBY	CPB	CVX	C	ABBV
CCL	CLX	COP	GS	AMGN
GPS	KO	DVN	JPM	BIIB
HAS	CL	HES	MET	JNJ
KSS	HRL	MRO	MS	MRK
LYV	K	OXY	PRU	PFE
M	KMB	PSX	STT	SNY
MGM	PG	VLO	WFC	TMO
SBUX	HSY	PXD	AXP	UNH

Table 2

Value Stocks Chosen from Sectors 6 to 10

Industrials	Information.Technology	Materials	Real.Estate	Utilities
GE	CSCO	DOW	SPG	NEE
MMM	DELL	ECL	PLD	DUK
CAT	HPQ	IP	EQIX	D
DE	INTC	APD	PSA	AEP
EMR	IBM	LYB	AVB	EXC
HON	MSFT	FCX	WELL	SO
ITW	ORCL	ALB	BXP	ED
LMT	STX	WRK	VTR	SRE
UPS	TXN	PPG	O	XEL
UNP	WDC	NEM	ARE	PEG

Table 3

Summary Statistics of Consumer Discretionary sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
F	12.1	12.2	3.2	1.6	18.8
BBY	40.8	36.3	17.2	11.3	88.6
CCL	43.8	41.1	11.7	17	71.9
GPS	27.4	25.7	8.6	9.9	46.6
HAS	62.5	54	26.5	21.3	126.1
KSS	52.6	51.5	9.5	33.6	82.1
LYV	24.8	21.9	17.8	2.5	73.5
M	34.5	32.6	15.2	6.6	72.8
MGM	19.9	20.4	8.2	1.9	38
SBUX	40.3	38.7	22.2	4.1	99.1

Table 4

Summary Statistics of Consumer Staples sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
GIS	46.9	48.8	10.3	23.6	72.6
CPB	42.1	41.5	9.2	24.9	67.6
CLX	100.1	91.3	33.4	45.9	166.4
KO	39	40.7	7.7	18.9	55.8
CL	58.1	63.6	13.3	27.5	77.5
HRL	25	24.3	11.7	7.3	45.9
K	60.6	61.6	9.3	35.8	87
KMB	96.1	104.2	27.5	41.6	142.7
PG	77.6	78.8	15.7	44.2	126.1
HSY	84.3	91.9	29	30.8	161.4

Table 5

Summary Statistics of Energy sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
XOM	81.3	82	9.6	56.6	104.4
CVX	104.1	107.7	17.7	56.5	134.9
COP	54.7	55	12.4	26.8	86.8
DVN	52.7	56.8	16.9	18.6	93.1
HES	62	59.7	13	34.4	101.1
MRO	22.6	19.8	8.3	6.7	41.7
OXY	75.8	76.5	13.5	37.3	110.9
PSX	80.4	80.5	19.3	29.4	123.3
VLO	50	50.2	28.1	14.2	124.4
PXD	128	136.9	51.7	12.1	233.1

Table 6

Summary Statistics of Financials sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
BAC	17.4	15.7	7.4	3.1	35.5
C	49.7	48.6	13.7	10.2	80.1
GS	172.8	167.5	42.7	59.2	273.4
JPM	64.6	57.5	27.9	15.9	139.1
MET	40.3	42	7.9	10.8	55.7
MS	32.4	30.9	10.8	12.4	58.9
PRU	75.3	78.8	22.3	11.3	126
STT	62.1	64.3	18.9	14.9	112.7
WFC	42.3	46.2	12.1	8.1	65.9
AXP	70.7	73.1	26.2	10.3	128.6

Table 7

Summary Statistics of Healthcare sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
BMY	45.3	48.9	15.4	17.5	76.8
ABBV	67.5	63.9	17.9	33.7	123.2
AMGN	122.4	124	53.6	45.1	243.2
BIIB	215.2	245.5	110.7	42.1	476
JNJ	95.3	98.2	29.1	46.6	148.1
MRK	49.2	50.6	14.7	20	87.5
PFE	27.1	28.8	8	11.1	43.9
SNY	42.3	42.2	7	25	57.3
TMO	123.5	119.7	74.3	32.4	328
UNH	113.1	81.6	79.2	16.4	296

Table 8

Summary Statistics of Industrials sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
GE	122.5	120.9	40.2	40	197.7
MMM	136.4	140.6	50.9	41.8	258.6
CAT	94.2	90.6	29.4	22.2	170.9
DE	95.3	85.9	34.8	24.8	179.8
EMR	56.1	56.8	10.4	24.9	78.5
HON	90.1	89.4	42	22.1	182
ITW	90.2	84.1	41.1	26.2	181.3
LMT	181.5	165.2	100.9	58.2	397
UPS	90.5	97.4	21.2	38.3	134.1
UNP	87.1	84.2	42.4	16.8	181.4

Table 9

Summary Statistics of Information Technology sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
CSCO	28.2	25.1	10.4	13.6	58
DELL	22.7	23	5.2	12	35.4
HPQ	16.3	16.1	4.7	5.3	26.4
INTC	31.1	28.6	11.1	12.1	60.1
IBM	150.3	148.3	26.8	78.4	206.3
MSFT	53.2	41.5	33.7	15.1	159
ORCL	37.2	38	10	13.9	60.2
STX	35.4	36.6	16.7	3.1	68.8
TXN	56.1	46.8	31.7	13.8	131.7
WDC	57.3	50.4	24.6	12	114.3

Table 10

Summary Statistics of Materials sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
DOW	50.4	50.5	4.2	40.7	59.7
ECL	99	106.4	43.9	29.9	208.6
IP	37.6	41.1	12	3.8	61.6
APD	116.6	113.1	44.1	40.5	241.3
LYB	73.7	81.4	26.6	15	121.5
FCX	26	25.2	13.5	3.7	60.9
ALB	66.5	63	24.4	15.8	144.6
WRK	48.1	48.3	10.1	27	70.3
PPG	79.5	95.1	33.4	14.2	133.8
NEM	38.9	37.8	12.7	15.6	72.1

Table 11

Summary Statistics of Real Estate sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
SPG	143.9	154.5	44.6	24.3	227.6
PLD	43.8	40.3	17.2	9.7	92.4
EQIX	251.4	214.9	144.7	42.3	583
PSA	167.2	171.5	55.9	48	276.3
AVB	145.7	145.4	40.8	40.1	222
WELL	61.1	62.8	13	27.3	92.5
BXP	108.8	115.4	24.5	31.5	144.7
VTR	63.2	63.3	12.2	23.2	94.7
O	44.9	43.4	13.8	14.6	79.4
ARE	89.5	78.2	29	30.7	162.5

Table 12

Summary Statistics of Utilities sector's 10 stocks

Ticker	Mean	Median	Std.Dev	Min	Max
NEE	25.9	24	12.3	10.4	60.7
DUK	69.9	71.7	13.9	35.4	97.2
D	61.3	67.6	15	27.3	84.9
AEP	53.6	52.8	17.4	24.3	95.7
EXC	27.7	27.3	4.6	18.2	41.2
SO	44.2	44.6	6.9	26.8	63.9
ED	63.3	60.6	14.5	32.7	94.7
SRE	87.7	96.1	28.8	36.7	152
XEL	34.6	31.2	12	16.2	65.8
PEG	39.8	37.6	9.2	24	63.3