Forecasting Hourly Prices of Bist30 Stocks

Group 2: İrem Betül Koçak and Burak Çetiner

January 15, 2024

1 Introduction

In this project, we aim to create one or more machine learning model to forecast hourly stock values in BIST30. It is very difficult task due to the nature of the problem. The difficulty of this task caused by the complexity of the problem. In other words, there are many variables that may affect the stock prices. However, we cannot include all the variables that may have an effect on the stock prices to the model and the magnitude of their effects. Economical and political institutions such as government, central bank, employer's unions etc. can be listed as the agents whose decisions may have strong influence on the stock market. On the other hand, it is not really possible to make causal inferences about these decisions in a mathematical way. Due to the complexity of this problem, we decide to use and test different machine learning algorithms and to make use of statistical inferences. We know that "there is no correct model" for this task. Therefore, we want to design an experimental setup to compare different machine learning algorithms and their success in terms of forecasting 30 stock values with minimum WAPE (Weichted Mean Absolute Percentage Error).

In order to interpret any dataset and make forecasts about future values, we have to understand the characteristics of the data. We should examine the data, indicate any trends and seasonality if there are any. On the other hand, we need expert knowledge and causal relationships which may have strong or weak influence on the future values. Shortly, it is not enough to use only the statistical tools to interpret any dataset. For this purpose, we firstly gather information about the current situation of the Turkish stock market and try to understand the current trends. Descriptive analysis is a good starting point in this sense. However we should also be familiar with the nature of the problem and the current dynamics of the market.

We first state that we are not trying to build a generic stock market forecast algorithm but we are dealing with the Turkish stock market in the time range of December 2023-January 2024 and our aim is to forecast BIST30, in other words 30 stocks in the Turkish stock market.. After specifying the time and place, we can then discuss the current dynamics of the Turkish stock market by examining macro and micro economic indicators.

The Turkish economy is passing through a high inflation phase which has a direct effect on stock market investments. According to past data, more accurately during 2022 and 2023, there is a strong positive trend for most of the stocks in the Turkish stock market. The index of the market increased regularly in this time period until December 2023. According to some analysts, this upward direction of the market index is caused by both high inflation rate and speculations which may lead to a "stock market bubble". The trend in the time period which we make forecasts (December-January), actually justify this stock market bubble argument. Due to these dynamics of Turkish stock markets, we will argue that both complex models and long term data may fail to detect current situations in the stock market and simpler models such as linear regression may have

more accurate forecasts. We will discuss these cases and comparison of used algorithms during this project in the 4. Experiment Design and Results section.

2 Literature Review

Firstly, we review the literature for the appropriate algorithms to apply to this problem. Our findings show that some of the most commonly used algorithms to forecast stock prices as follows; ARIMA, Random Forest, Linear Regression, Ridge and Lasso Regression. We will discuss the advantages and disadvantages of each algorithm in the following section. However, we want to emphasize one thing in this part. As the reader may notice linear regression choice may seem a bit interesting and too simple to work well in such a complex problem. Despite the fact that linear regression may have less complexity, we will argue why it may be more suitable in the context of forecasting BIST30 in this time period.

2.1 ARIMA

ARIMA (Autoregressive Integrated Moving Average) is a commonly used algorithm for time series forecasting, including in stock market analysis. ARIMA can capture a wide range of standard temporal structures in data, especially it is expected to fit data with trends and seasonality [1]. Moreover, ARIMA can be adapted easily to time series data which increases flexibility due to its straightforward implementation. It is widely used to interpret time series analysis. While ARIMA is not always the best-performing model, it can provide good forecasting accuracy, especially for datasets exhibiting linear relationships or where historical patterns persist into the future. Despite all these advantages of ARIMA, it also has limitations, such as its inability to handle more complex relationships or irregular patterns in data that other models like machine learning algorithms or advanced neural networks might capture more effectively. It's often beneficial to complement ARIMA with other techniques or models to improve forecasting accuracy, especially in highly volatile or nonlinear stock market data. After stating these advantages and disadvantages of ARIMA, we try to implement ARIMA to our stock market dataset. However, we were not satisfied with the output values of ARIMA and decided not to submit these outputs as daily forecasts. As the literature suggest, it has some disadvantages which lead us not to use this algorithm in the further forecasts.

2.2 Random Forest

Random Forest is an ensemble method combining multiple decision trees used for classification and regression [2]. This algorithm tends to reduce overfitting compared to individual decision trees and can provide more robust predictions. Random Forest is effective in capturing non-linear relationships and complex patterns. As the patterns in stock markets might not follow linear trends, Random Forest can be useful to make accurate forecasts. Moreover, Random Forest can provide insights into feature importance. It is helpful to indicate the most influential variables in making predictions. This can offer valuable insights into market dynamics. This algorithm can be considered as successful to handle large datasets with a high number of features, which can be beneficial when dealing with extensive financial data. Furthermore, the algorithm is robust to outliers and noise.

However, Random Forest algorithm has some weaknesses about interpretability. In other words, these models are "black box". Thus, it is difficult to understand why the model reaches a certain decision or result. On the other hand, the tuning of the parameters is somewhat complicated. The performance of Random Forest heavily depends on the quality of the input data and the relevance

of the features. Random Forest may also struggle to capture sudden shifts or changes in market behavior that are not explicitly represented in the features.

2.3 Linear Regression

Linear regression is a commonly used method for predicting a variable by using independent variables [3]. The linear regression analysis is called univariate analysis if there is only one independent variable, whereas it is called multivariate analysis if there are more than two independent variables. Linear regression indicates a linear relationship between dependent variable and independent variables. It has some advantages such as simplicity, computational efficiency and indicating linear relationships. Linear regression is straightforward and easy to understand. It provides a simple relationship between the independent and dependent variables, making it easy to interpret. The coefficients in a linear regression model have clear interpretations. This can be useful for understanding the impact of different features on stock values.

On the other hand, there are limitations of linear regression. Linear regression assumes a linear relationship between the independent and dependent variables and if the actual relationship is nonlinear, the model may fail to create accurate forecasts of stock values. Moreover, linear regression is sensitive to outliers and noise. These models may not capture complex relationships or interactions between variables which are usually a common case in the stock markets. Another limitation about linear regression is the risk of overfitting and underfitting. Linear regression assumes that the residuals are independent. In financial time series data, this independence assumption may be violated due to autocorrelation. Furthermore, macroeconomic indicators may have an effect on stock prices as aforementioned.

Despite all of these limitations and disadvantages of the linear regression models that are stated above, we still argue that it is one of the most suitable machine learning algorithms to forecast BIST30 in the Turkish stock market during December 2023-January 2024. Although linear regression is not a strong model to capture nonlinear dynamics such as stock markets, in the given time range, the Turkish stock market has a certain negative trend for most of the stocks. However, we are also aware of there are also some outliers.

2.4 Ridge and Lasso Regression

Ridge and Lasso regression are two techniques used in linear regression to prevent overfitting and handle multicollinearity. Both methods add a regularization term to the linear regression cost function to keep the coefficients of the features within certain bounds. A regularization term proportional to the square of the coefficients is added in Ridge regression whereas in Lasso regression, a regularization term proportional to the absolute value of the coefficients is added to the cost function. The advantages of Ridge regression can be listed as follows; it helps prevent overfitting by penalizing large coefficients, it is useful data with multicollinearity, and can handle a large number of features. On the other hand, Ridge regression does not perform variable selection, in other words it includes all features in the model. Lasso regression performs feature selection by pushing some coefficients to exactly zero and it is useful when there are many irrelevant or redundant features. However, Lasso regression can be easily affected by the scale of the features. It may have problems with highly correlated features and it is not suitable with a high number of features compared to the number of observations. Both Ridge and Lasso help in preventing overfitting, which is crucial in stock market forecasting where models need to generalize well to unseen data [4]. Ridge regression, in particular, is effective in handling multicollinearity, a common issue in financial data where certain features may be highly correlated. On the other hand, The coefficients of features might not

be as easily interpretable, which could be a concern if you need to explain the factors driving the stock market predictions. Other than that, if the relationship is highly nonlinear, these methods may not perform as well.

3 Descriptive Analysis

Before we start training the model, we look at how stock prices change over time and try to capture the trend of stock prices. To do this, we will look and analyze the historical price data of these stocks. All the data, i.e. historical data for all stocks, and other data which is used in training the model is retrieved from Yahoo Finance [5].

We observed a decrease in nearly all stock prices during the first half of 2023, followed by an increase thereafter. This trend aligns with the political instability experienced in the early part of the year. After the general election in Turkey in May 2023, a sense of balance was restored. However, as shown in Figure 4, the main concern was the sharp devaluation of the Turkish Lira. Consequently, the pricing based on the Turkish Lira also increased. So, we can say that currency is an important factor in the pricing.

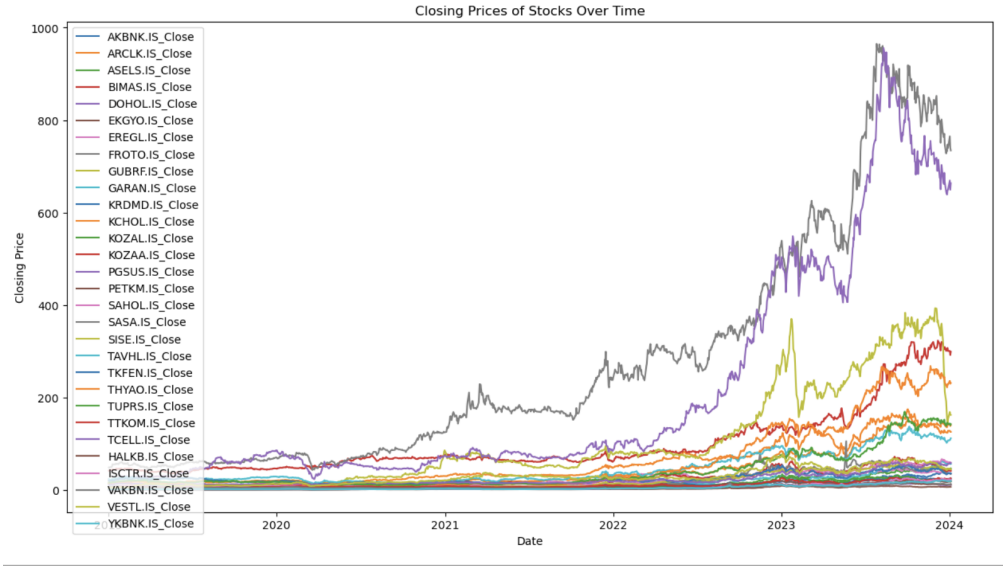

Figure 1: Closing prices of Bist30 stocks over five years.

In Figure 1, it's challenging to discern which stocks are more volatile or expensive, so we turn to the box-plot of the closing prices of these 30 stocks in Figure 2. We note significant deviations from the median in the stock prices of FROTO and PGSUS. These deviations likely causing from the companies' strategies and the overarching dynamics of global markets. For instance, FROTO experienced a dramatic increase in 2021, aligning with its investments in European markets. In contrast, PGSUS saw a sharp rise in 2022, reflecting a post-pandemic surge in demand for airline stocks, counter to its significant drop in early 2020. Other stocks, such as BIMAS, GUBRF, and THYAO, also deviate from the median in the box-plot. However, determining whether they follow an increasing or decreasing trend over the years is not possible. Therefore, in Figure 3, we present a time-series graph of prices, excluding the two outliers, FROTO and PGSUS.

Figure 3 reveals that despite some downward trends, the maximum prices for all stocks generally peak in 2023. Nevertheless, as we will explore in the forecasting section, there is also a noticeable

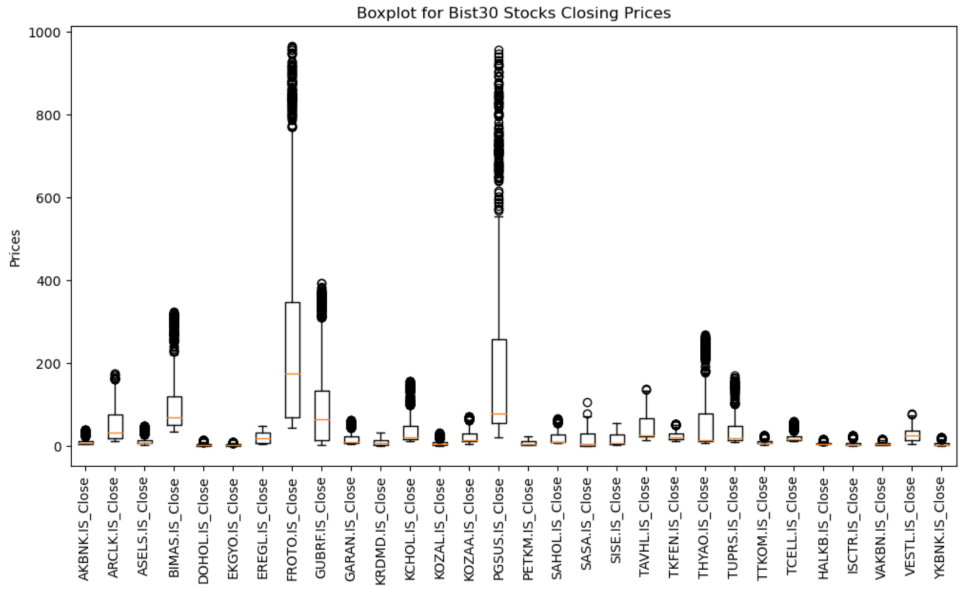

Figure 2: Boxplot of closing prices of Bist30 stocks over five years.

decreasing trend in all stock prices towards the end of 2023. This observation leads us to consider other variables which would explain this increasing trend.

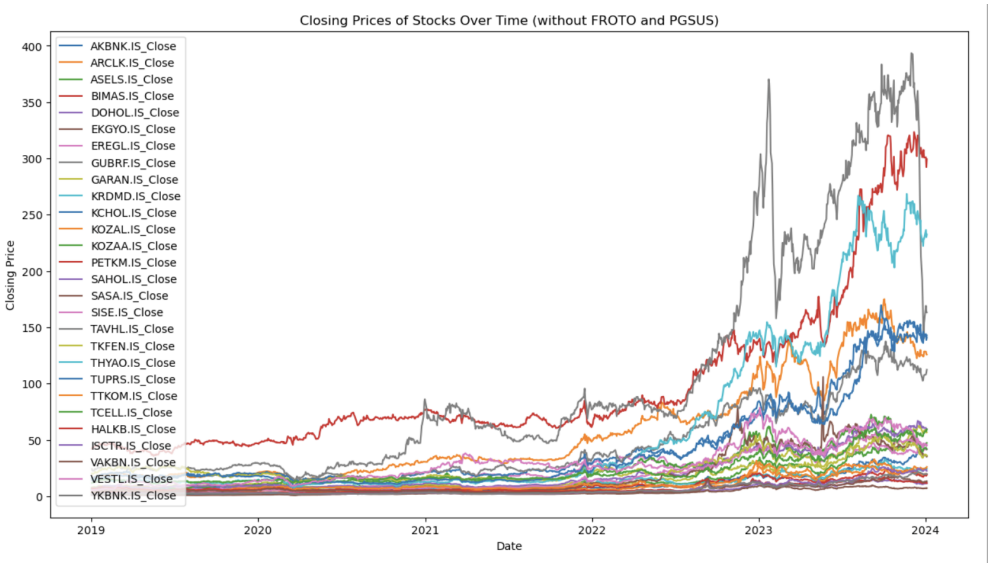

Figure 3: Closing prices of Bist30 stocks over five years (Excluding FROTO and PGSUS).

3.1 Variable Selection

Choosing the right algorithm for the model is important, but selecting sensible variables for forecasting prices is equally crucial. For training the model with different variables, we need data that is updated on a daily basis. As a result, we exclude variables like inflation rates, interest rates, and GDP of Turkey, focusing instead on variables that are updated on an hourly basis. We consider

only the closing prices of these variables for forecasting the next day's trends.

Firstly, we choose the USD-TRY (U.S. Dollar - Turkish Lira) currency exchange rate, as it fluctuates throughout the day and is a determining factor. The transactions occur in USD as they are open to global markets, but we analyze the data in terms of TRY. Therefore, when there is a peak in the USD-TRY exchange rate, we also observe a peak in the prices of all stocks.

We wished to use the BIST 30 index to forecast the prices. However, we encountered difficulties in retrieving the index data from Yahoo Finance and other data sources. Therefore, we decided to look at the Turkish Stock Price instead, and we found that these data can be useful in forecasting the prices, as presented in Figure 5.

Finally, we examined the Google Trend data of all stocks and decide to neglect it since we observed that it is not very robust data. The number of searches for indexes is not correlated with the price of that stock, as the searching behavior of people is reactive and could be influenced by steep increases or decreases in some stock prices. Therefore, this information is not useful and could deteriorate the forecasting. An example is shown in Figure 6: while the price of PGSUS has decreased, searches for this stock have increased. Notably, there is no significant increase in Google Trends for PGSUS after the year, despite a considerable rise in the stock's price. However, we observe a similar upward trend in the stock prices of both FROTO and PGSUS, accompanied by a steep increase in searches for them. Thus, it can be said that an increase in interest as reflected in Google Trends searches does not necessarily correspond to a significant increase or decrease in stock price, but is rather associated with multiple factors.

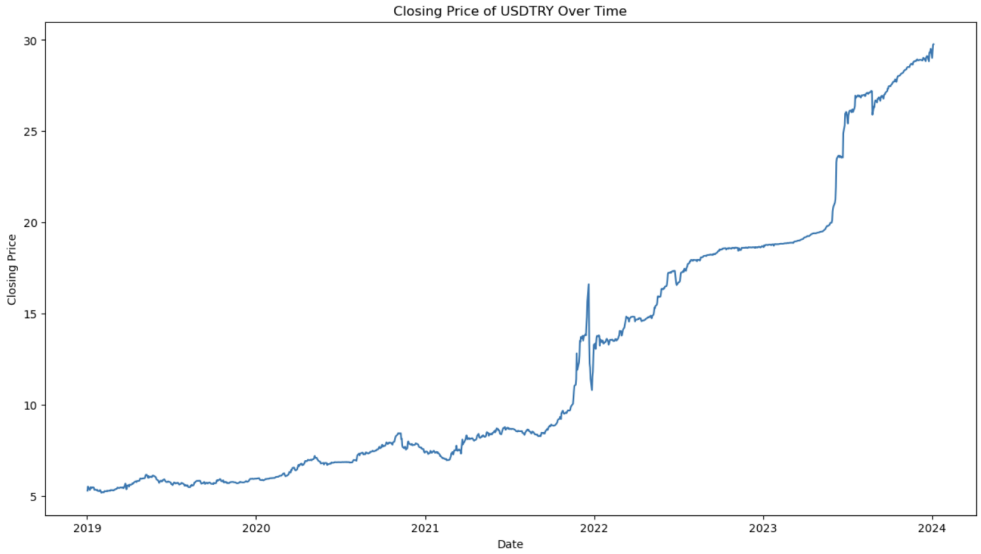

Figure 4: USD-TRY Currency change over five years.

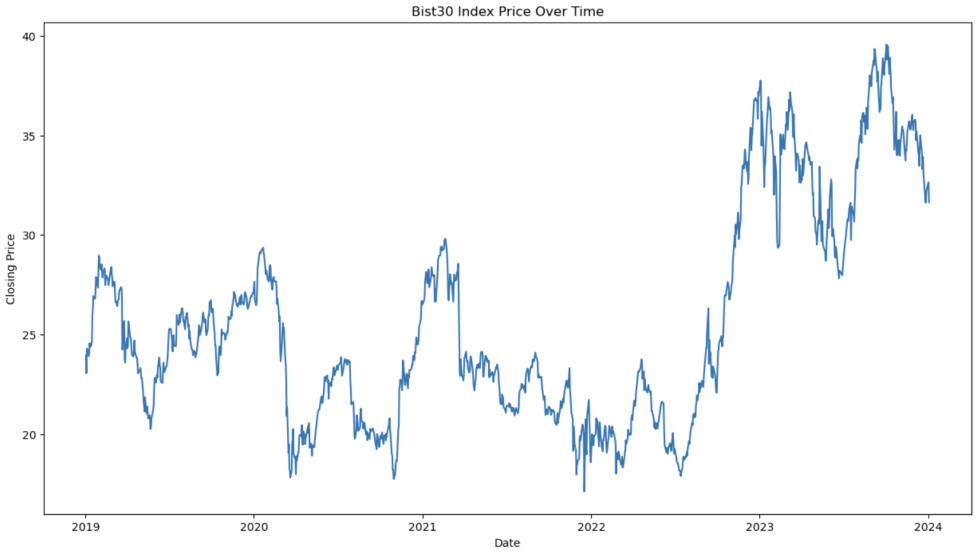

Figure 5: Turkish Stock Data Price change over five years.

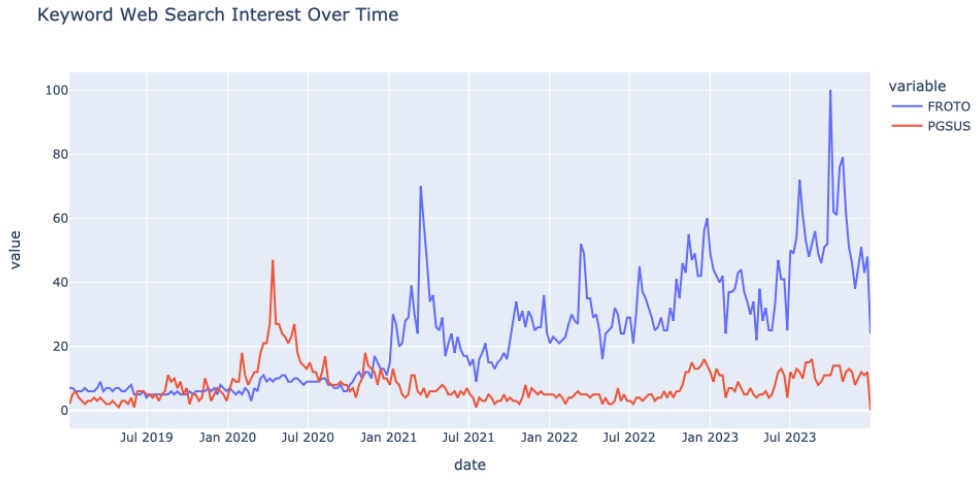

Figure 6: FROTO and PGSUS Google Trend.

4 Experimental Design and Results

We are asked to forecast for the next day for 14 days for each hours in a day, we try to variate the algorithms, timespan and variables we use to compare the WAPEs of our results. WAPE is calculated by the difference between the forecasted and actual prices of each stock in each hours.

We summarize the algorithms we use for forecasting hourly stock prices, the variables we include in the model, and the timespan of the historical data. We also add the day-by-day WAPE scores for week 1 in Table 1, for week 2 in Table 2, and for week 3 in Table 3. We have 3 N/A WAPE scores in total, which are caused by the erroneous formats we uploaded the forecasted values into the system. We will compare and discuss the different combinations of these algorithms, timespans, and variables.

When we look at the WAPE scores for each day, we see that the least successful combination in forecasting stock prices is the univariate model that only includes historical data. After the first day, we did not train the model on only one variable again. On the second day, we used linear regression including currency variable, and saw that the WAPE decreased significantly. To determine whether this was caused by the change of algorithm or the number of variables, on the third day we used the same algorithm, Random Forest, and observed that it was probably caused by the lack of training data. In the first week, we continued to use linear regression with different timespans and saw that the WAPE was always low. So the main inference from the first week is that a bivariate model is better than a univariate model. However, because the differences between the WAPEs on other days were not very high, we do not comment on the algorithm, but we note that both linear regression and random forests are effective at forecasting prices in a bivariate model. In the meantime, we also calculated prices using ARIMA, but we did not use ARIMA since it showed an increasing trend on some stocks which were nearly certain to decrease the next day.

In the second week, we also continued to apply linear and Random Forest algorithms, but this time with the variable 'Turkish Stock Price'. As the second week marked the beginning of 2024, the market produced sequential prices for the stocks, resulting in the lowest WAPE scores with both algorithms in our multivariate model.

In the last week, we decided to apply Ridge and Lasso regression to see if they would yield better results. However, we could not obtain results on two days due to incorrect formatting. On the other two days when these algorithms were tested, they did not yield good results. Therefore, on the last day, we reverted to the combination used on the second day and achieved very good results with a WAPE value of 0.01946549.

It is not feasible to compare the algorithms and variables we used over just 14 days in a highly unpredictable season of the year. The same algorithm can yield varying results, as observed on the second and last days, and on the fourth and fifth days. However, we can conclude that in a country like Turkey, which experiences trends influenced by volatile currency indexed in USD, the least complex model may not produce the most successful result, but it can yield a relatively good result. We are therefore satisfied with the results of non-univariate linear regression or random forest models.

Table 1: Week 1 Algorithms, Variables and WAPE.

Date	25.12.2023	26.12.2023	27.12.2023	28.12.2023	29.12.2023
Algorithm	Random Forest	Linear Regression	Random Forest	Linear Regression	Linear Regression
Timespan	Long Term	Long Term	Long Term	Short Term	Short Term
Variable	Historical Data	Historical Data Currency	Historical Data Currency	Historical Data Currency	Historical Data Currency
WAPE	0.05595971	0.02412649	0.02135586	0.023427597	0.01896627

Table 2: Week 2 Algorithms, Variables and WAPE.

Date	2.01.2024	3.01.2024	4.01.2024	5.01.2024
Algorithm	Linear Regression	Linear Regression	Random Forest	Linear Regression
Timespan	Short Term	Long Term	Long Term	Long Term
Variable	Historical Data Currency	Historical Data Currency TR Stock Price	Historical Data Currency TR Stock Price	Historical Data Currency TR Stock Price
WAPE	0.02064153	N/A	0.01983863	0.01897757

Table 3: Week 3 Algorithms, Variables and WAPE.

	8.01.2024	9.01.2024	10.01.2024	11.01.2024	12.01.2024
Algorithm	Lasso Regression	Ridge Regression	Ridge Regression	Lasso Regression	Linear Regression
Timespan	Long Term	Long Term	Long Term	Long Term	Long Term
Variable	Historical Data Currency TR Stock Price	Historical Data Currency TR Stock Price	Historical Data Currency TR Stock Price	Historical Data Currency TR Stock Price	Historical Data Currency
WAPE	N/A	0.02217095	N/A	0.03939937	0.01946549

References

[1] Prapanna Mondal, Labani Shit, and Saptarsi Goswami. Study of effectiveness of time series modeling (arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, 4(2):13, 2014.

[2] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[3] N Amral, CS Ozveren, and D King. Short term load forecasting using multiple linear regression. In *2007 42nd International universities power engineering conference*, pages 1192–1198. IEEE, 2007.

[4] Sanjiban Sekhar Roy, Dishant Mittal, Avik Basu, and Ajith Abraham. Stock market forecasting using lasso linear regression model. In *Afro-European Conference for Industrial Advancement: Proceedings of the First International Afro-European Conference for Industrial Advancement AECIA 2014*, pages 371–381. Springer, 2015.

[5] Financial data. Yahoo Finance. Retrieved data programmatically using Python.