BOĞAZİÇİ ÜNİVERSİTESİ

Statistical Learning For Data Mining

Author:
EGE ERDİL (Group 12)

15 January 2024

1) Introduction

As a final project of the Special Topic of Statistical Learning for Data Mining course, the hourly prediction of the stock market of BIST 30 companies has been established. It is difficult to predict stock prices because it requires examining a lot of different aspects, market patterns, and historical data. Hourly data on BIST 30, the Turkish stock market index of the Borsa Istanbul, offers a comprehensive view of price changes during the trading day. A combination of technical analysis, fundamental factors, and machine learning algorithms are used to predict the hourly stock prices of BIST 30.

To spot patterns and trends, technical analysts examine past price charts, trading volume, and a range of technical indicators. For hourly stock price prediction, analysts often use tools such as Moving Averages, Relative Strength Index (RSI), and Bollinger Bands. These indicators aid in determining the market's momentum, volatility, and probable turning points. Moreover In this project MACD (Moving Average Convergence Divergence) had been implemented. Exponential Moving Average and Simple Moving Averages have been implemented as moving averages.

The primary focus of fundamental analysis is the companies listed in the BIST 30 index and their underlying financial performances. Stock prices can be impacted by a number of factors, including economic indicators, geopolitical events, and earnings reports. To make educated forecasts, traders frequently monitor interest rates, company-specific news, and macroeconomic indicators. However, in this project implementing economic indicators and earning reports were not appropriately easy to implement. The main economic indicator used in the project was the USD/TRY rate. The currency pair that represents the exchange rate between the US dollar (USD) and the Turkish lira (TRY) on the foreign exchange market is called USD/TRY (United States Dollar to Turkish Lira). Inflation rate tried to be converted to a meaningful value and used as a parameter however throughout the trials it concluded that the effect of it was not crucial to the project.

Data preprocessing is essential before applying learning algorithms. This involves purifying and altering the data, handling values that are missing, and standardizing variables. Trading volume, pertinent indicators for technical analysis, and previous price data are necessary for hourly share prices prediction. To guarantee generalization, machine learning models are tested on untested data after being trained on historical data. Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are examples of common evaluation metrics. A good model should have low error rates and be able to accurately represent price trends. In our project the aim was minimizing the hourly stock prediction errors with the metric WAPE. The overall difference between predicted and observed values is measured by the Weighted Absolute Percentage Error, or WAPE. The computation of WAPE involves adding up all observed and predicted values, then figuring out the difference in error between the two. A lower value indicates a more accurate model.

Because financial markets are dynamic, it is inherently difficult to predict hourly stock prices. Stock prices can be impacted by outside variables such as unanticipated news, market mood,

and world economic events. Furthermore, historical performance does not guarantee future outcomes, and models must be updated continuously to reflect changing market conditions.

2) Literature Review

To understand which machine learning models can be used to successfully forecast next day's hourly stock prices, research has been made. Some of the machine learning algorithms even though the project result did not depend on them have been analyzed to understand what kind of approaches exist. (While finding results, since some of the algorithms were not taught in the course, they have not been used.) These approaches were Neural Networks (Deep Learning), ARIMA (AutoRegressive Integrated Moving Average), Long Short-Term Memory (LSTM) Networks and Prophet. While creating the model linear regression, random forest, knn algorithm, extreme gradient boosting approaches have been examined.

2.1) Course Excluded Approaches

While creating the model that fits well to the dataset 2 approaches have been analyzed to have better understanding overall in the stock price prediction. These approaches were LSTM and Prophet.

LSTM was the most commonly used approach for stock prediction. Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture created to solve the problems of gathering dependency over time in data that is serial. LSTMs are a popular option for stock price prediction because they work especially well for tasks involving time series data. However, in BIST 30, while implementing stock prices to the data the results were not differentiate from other algorithms in a positive way.

Ashish Vishwakarma and Alok Singh(2020) stated in their article that 'The method for predicting time series data using this additively based model fits non-linear trends with

Figure 1) An example of Prophet Approach

weekly, daily, and annual seasonality in addition to the effects of holidays. When there are multiple seasons of historical data and strong seasonal effects in the time series, the model performs best.' This method has been developed by Facebook Data Scientist and with the

scikit library can be used. However, it can be seen in the figure 1 in BIST 30 stock prediction this method gave relatively bad outcomes. It was not able to fit the residuals.

2.2) Course Included Approaches

Since our aim was to find the least error the number of approaches have been used. First of all linear regression has been used. By fitting a linear equation to the observed data, linear regression is a statistical technique used to model the association between a dependent variable and several independent variables. Finding the best-fitting line that minimizes the sum of the squared differences between the values that are seen and those that are predicted is the aim. A common tool for forecasting and comprehending the relationships between variables is linear regression.
Secondly, random forest methods have been used. As part of its ensemble learning process, Random Forest builds a large number of decision trees during training and produces the mean prediction (regression) of each tree separately. An ensemble of decision trees, typically trained using the "bagging" technique, makes up the "forest" it constructs.
Extreme gradient boosting methods have been analyzed. Yue Yang, YangWu, Peikun Wang,Xujiali(2021) mentioned in the article that ’As a decision tree promotion model, XGBoost combines multiple tree models into one powerful classifier.A tree is grown by continuously adding new trees, continuously dividing features, and fitting the residuals from the previous training each time a new tree is added.The final predicted value of the sample is calculated by adding the scores for each of the multiple trees that were acquired during training.’ Which have been the best performing model in general for the BIST30 stock prediction.

3) Feature Selection

Since the data given by our professor only had one parameter which was the hourly price of the stocks. Firstly, we had to decide what kind of parameters we can utilize in our models and how we can find those instances. What i have been concluded, there were parameters such as opening and closing prices of the stock for the specific hour, low and high values of the specific hour, simple moving average and exponential moving averages with different window sizes, RSI, MACD, volume of the stock in that hour, Turkish lira - Dollar exchange rate in that time period.
The US dollar's relative value in relation to the Turkish lira on the foreign exchange market is represented by the USD/TRY currency pair. Its price fluctuations and dynamic nature are influenced by a number of geopolitical and economic factors. This is why as an economic indicator I have decided to implement my predictions.
One other variable that has been considered in the algorithm is the volume of the stocks traded in that time period. When discussing stocks and financial markets, volume is the total number of shares or contracts that are traded in a specific security or market over a predetermined amount of time. It is an important metric that sheds light on the volume of trade, liquidity, and intensity of price changes.

Some technical analysis approaches have been used such as RSI, MACD, SMA and EMA. Moving Average Convergence Divergence, or MACD for short, is a well-liked momentum indicator that technical analysts use to spot probable trend reversals, validate existing trends, and produce buy or sell signals. The difference between two Exponential Moving Averages (EMAs) of the price of an asset serves as the basis for Gerald Appel's MACD. Relative Strength Index, or RSI for short, is a momentum oscillator used to gauge how quickly and how much prices move in financial markets.A statistical computation called a Simple Moving Average (SMA) is employed to examine data over a specific time frame. It is frequently used in time series analysis and is especially helpful for reducing data fluctuations so that patterns or trends can be seen. An example of a weighted moving average that gives more weight to recent data points in the computation is the exponential moving average, or EMA. The exponential moving average (EMA) gives greater weight to the most recent observations than the simple moving average (SMA), which gives equal weights to all data points within a given time period.
While trying to use Google trends he have seen some correlation and tried to implement to our code, however in finalized data we have seen that the values did not significantly improve our model or did not significantly decrease our model accuracy.

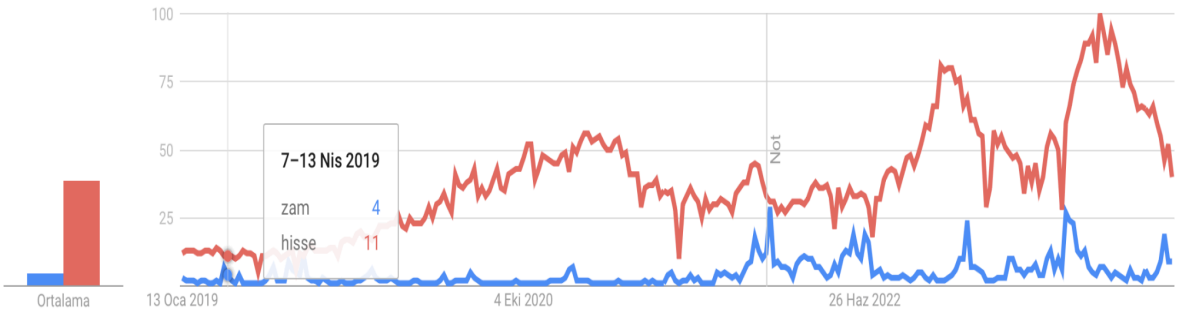

Figure 2) Google Trends example

After technical indicators it has been researched that if the seasonality have meaningful interpretation over predictions. Timestamps have been divided into years, months. days and hours to be analyzed. The figure below it can be seen that there is not a significant meaning that can expose from this approach.

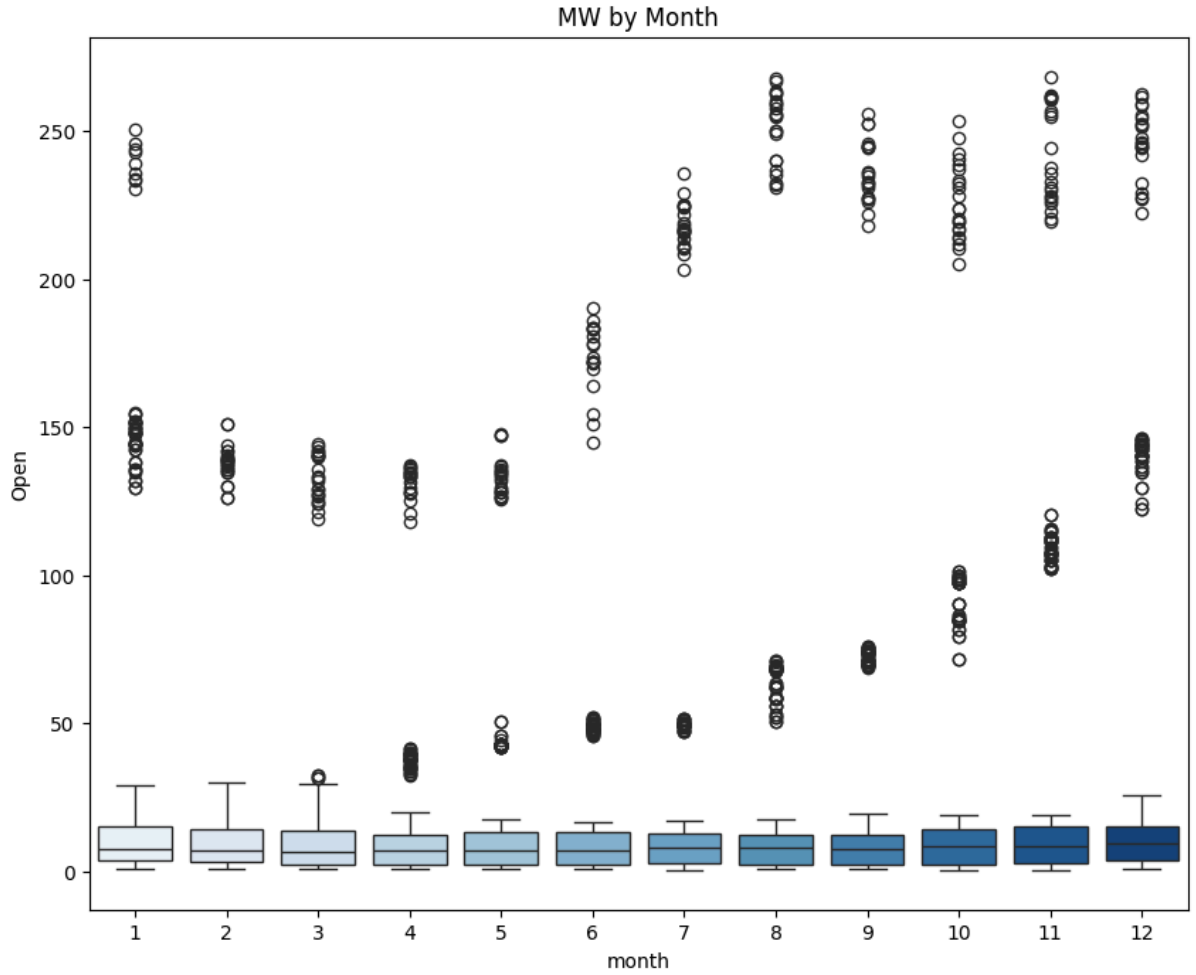

Figure 3) THYAO data historical analyze from 2015 to 2024 with divided by month

4) Descriptive Analysis

First all of predicting stock prices considering clustering entails giving careful thought to both the predictive analysis's goals and the underlying data dynamics. When there are observable trends or similarities in the past performance of particular groups, clustering stocks can be useful. Predictive models may gain from capturing common traits and patterns by identifying these clusters, improving the precision of subsequent value projections. However, clustering may introduce noise and reduce prediction accuracy if the stocks show different behaviors or react to different market drivers. Evaluating the heterogeneity between clusters and the homogeneity within them is crucial. Furthermore, it is imperative to carefully assess the predictive ability of individual stocks in comparison to grouped clusters, since certain stocks may have particular factors affecting their future values.

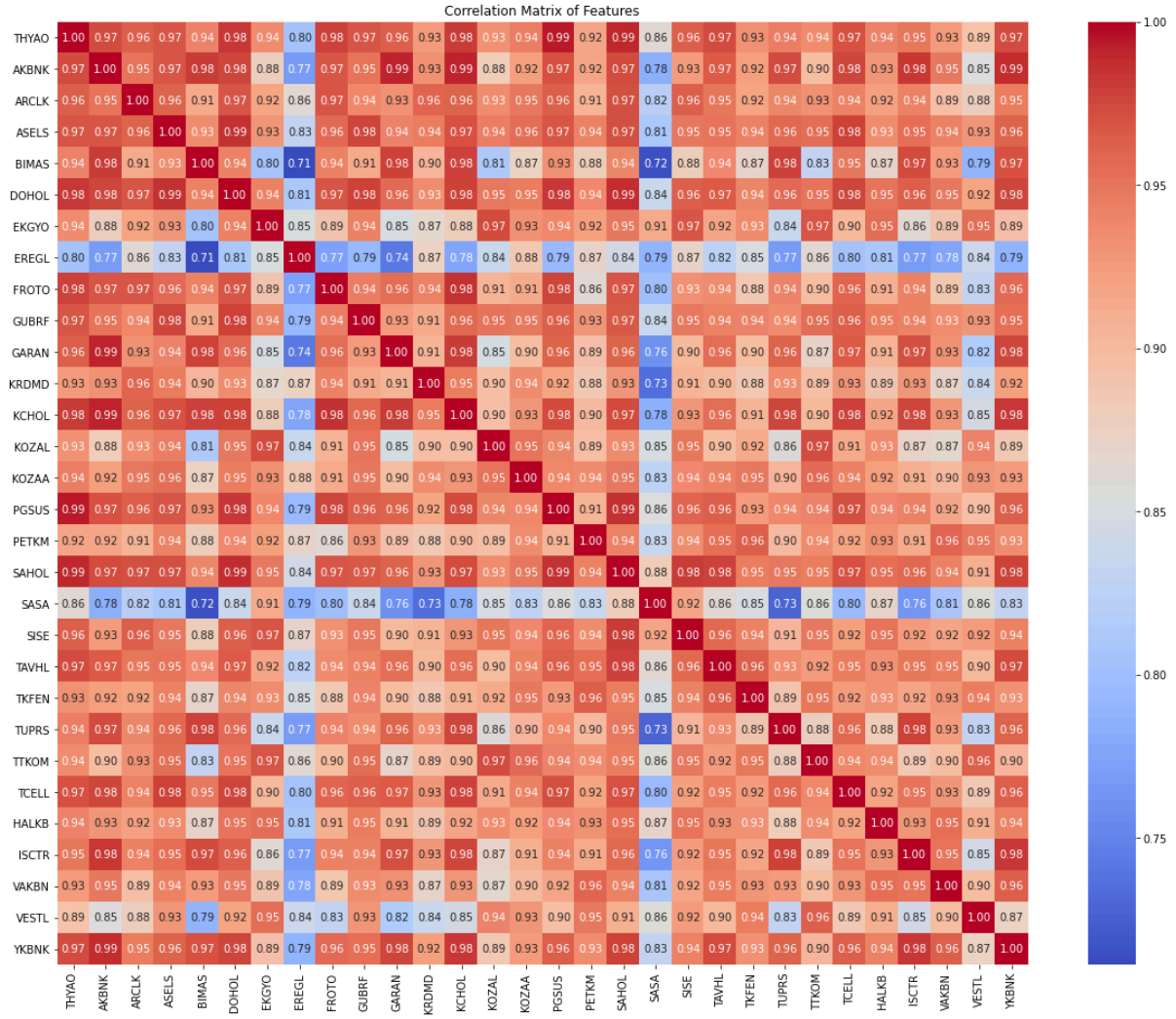

Figure 4) Correlation towards BIST 30 stocks

After analyzing and trying the clustered methods together, I have concluded that analyzing each stock by itself gives better prediction values. And analyzed every stock by its own historical data and given other parameters.

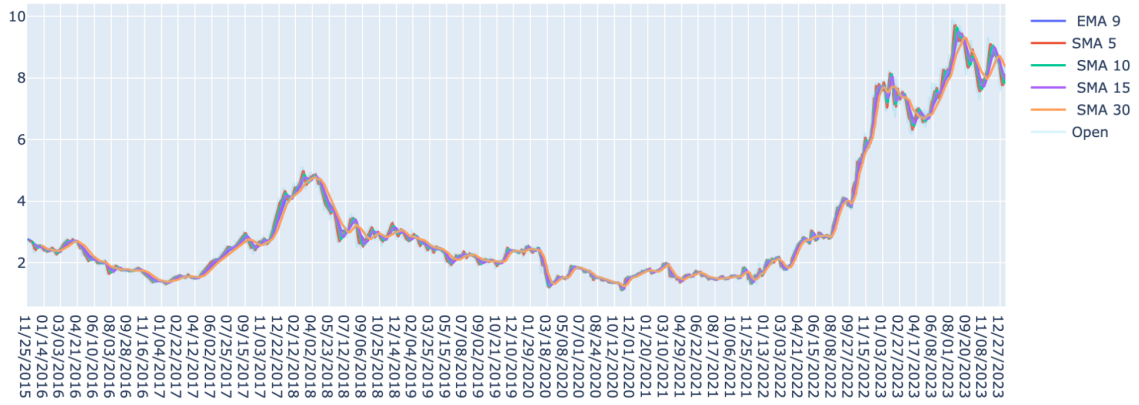

Figure 5) THYAO data historical analyze from 2015 to 2024 with technical indicators

Afterwards each dataset have been selected and analyzed by its own parameters.

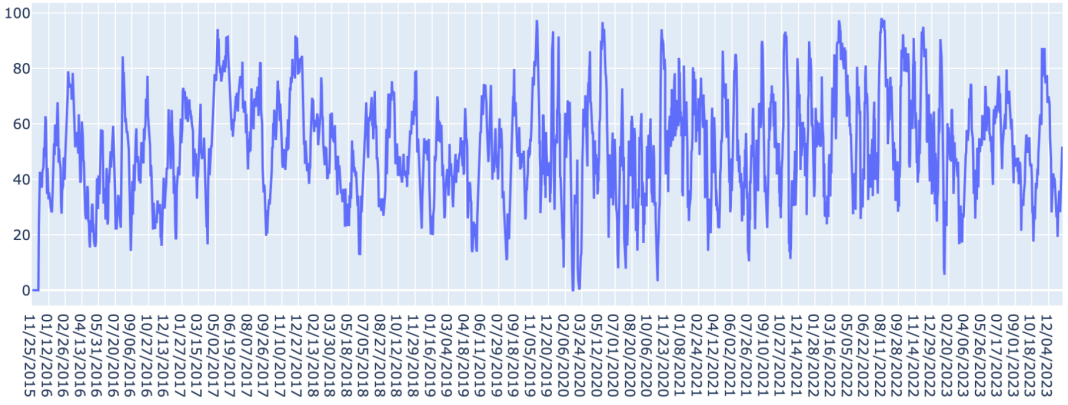

Figure 6) GUPRF data historical analyze from 2015 to 2024 RSI

5) Model Selection

XGBoost's ability to handle intricate relationships in the data and maximize predictive accuracy is what makes it superior to random forest and linear regression when it comes to predicting Turkish stock prices. The decision tree-based ensemble learning algorithm XGBoost is highly effective at identifying complex patterns and non-linear dependencies found in financial time series data. It cleverly merges the forecasts from several incompetent learners to create a strong and precise model. Through regularization techniques, XGBoost can effectively handle overfitting and tends to provide better generalization performance than

random forest. Conversely, the assumption of a linear relationship between features and the target variable in linear regression may oversimplify the intricate dynamics of stock prices. XGBoost's flexibility to adjust to the complex dynamics in movements in stock prices, utilized feature relationships, and manage abnormalities in data sets made it an effective option for economic forecasting tasks, outperforming traditional predictive models along with other ensemble methods.

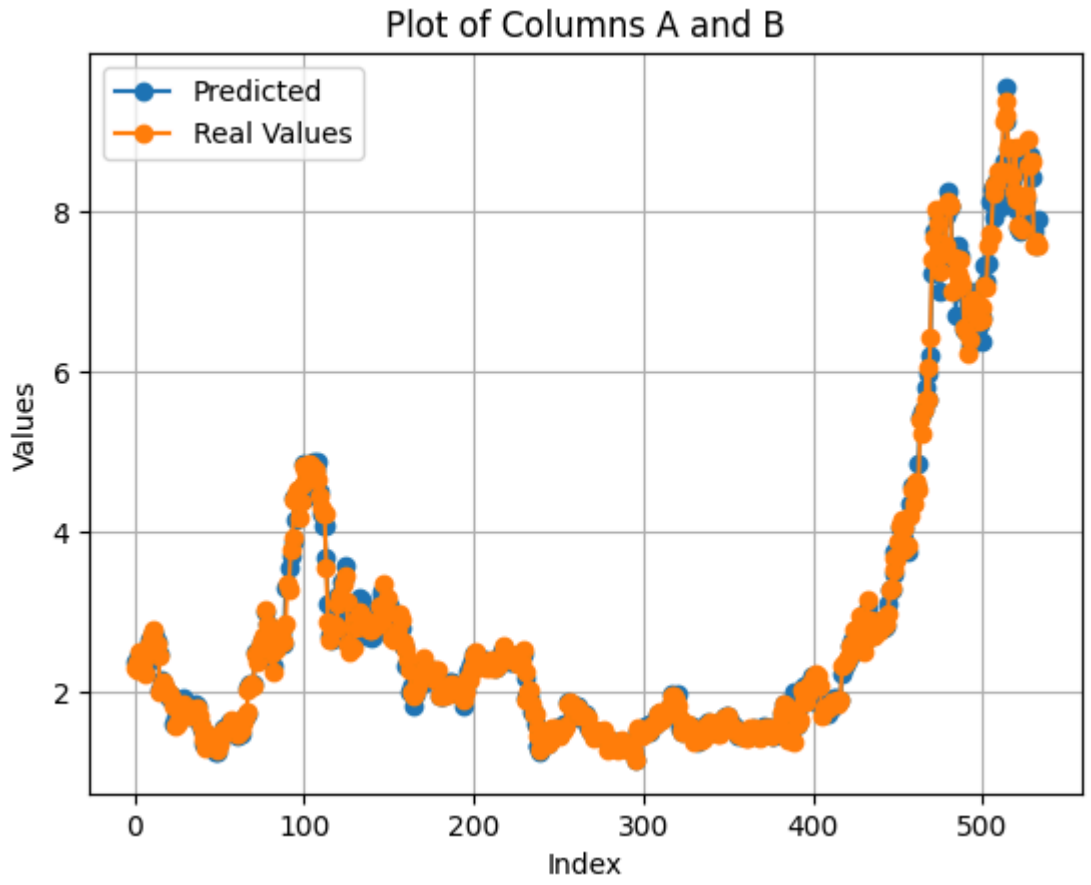

Figure 7) Example of Predicted and Real Values

It has been very challenging for me to realize that not randomly splitting our dataset to training and test sets could really affect our conclusion in a bad way. To guarantee that the model gains knowledge from a representative sample of the data, instances are chosen at random for both the training and test sets. This randomness keeps the model from being unduly impacted by particular trends or outliers and aids in capturing a wide variety of patterns found in the dataset. The training set may unintentionally be biased toward particular patterns when the split is not random, which could result in a model that performs well on the training data but poorly on fresh, untested data. As it may be seen in the data the wrong splitting leads our model to learn very different structures.

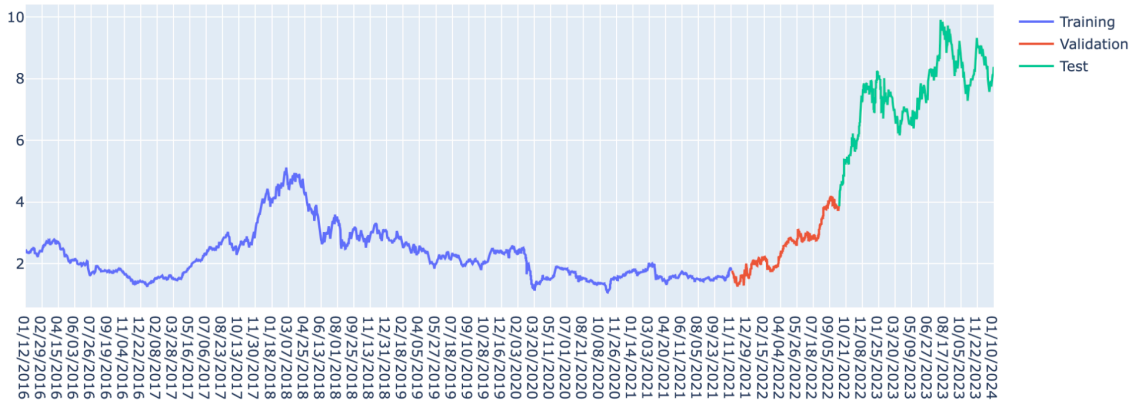

Figure 8) Mistaken Splitting of training, validation and test set

However, because of this mistake I have made, I have realized that the stock prices have been influenced by external economic reasons such as the USD TRY rate.

	Garan Error Examples
mean_squared_error	0,02648848647
wape	0,01463707807

Finally, in the project it has been selected the method WAPE to be calculated to rank our success. Mean squared error and wape is calculated lastly.

6) Results and Discussion

Leaderboard	WAPE
5,5	0,2164488402
5,9	0,02646374234
7,2	0,02935249772
8,433333333	0,02579134644
7,1	0,02072269441
6,366666667	0,02119062942
6,7	0,01843554702
5,3	0,02910553944
6,966666667	0,02270391491
6,633333333	0,02018617123

4,866666667	0,01838456874
6,451515152	0,04079868108

Last day of my submission my WAPE value was 0.01838456874. Throughout the process of submitting our predicting values I have been improved. However there are some other ways to enhance the solutions that I was not able to experiment with.

To improve the quality of the solutions neural network models can be used. However since it is out of topic I was not able to implement this approach to my solutions. Neural network models offer a powerful way to improve the quality of solutions across a range of fields. When it comes to capturing complex relationships, neural networks frequently perform better than traditional algorithms because of their ability to learn intricate patterns and representations from data.

Adding features that are based on global and economic indicators has a lot of potential to improve model robustness and accuracy. A more thorough depiction of the market dynamics can be attained by carefully adding pertinent economic indicators, such as interest rates, inflation rates, and macroeconomic trends, into the dataset. It takes a sophisticated comprehension of these indicators' influence on stock prices as well as the capacity to identify their temporal correlations in order to convert them into meaningful values.

References:

Yang, Yue & Wu, Yang & Wang, Peikun & Jiali, Xu. (2021). Stock Price Prediction Based on XGBoost and LightGBM. E3S Web of Conferences. 275. 01040. 10.1051/e3sconf/202127501040.

Ashish Vishwakarm, Alok Sing,Stock Price Prediction Using Sarima and Prophet Machine Learning Model L.B.H.S.S Trust’s Institute of Computer Application, Mumbai University, Mumbai

Chen Tianqi,Guestrin,Carlos. Tree Boosting System.2016.XGBoost:A Scalable