BOGAZICI UNIVERSITY

STATISTICAL LEARNING FOR DATA MINING

IE 582

STOCK MARKET FORECASTING CHALLENGE

Authors:
Selahattin Seha Cirit,
Ömer Yiğit Yıldırım

15 January 2024

1. Introduction

A stock market is an open marketplace where shares of companies that are publicly traded can be purchased and sold. The ownership of the corporation is represented by the stocks, commonly referred to as equities or a share. The bridge that allows the purchase and sale of shares is the stock exchange.

How and why does the stock market have an important role in the business world? Briefly, the stock market enables companies and individuals to raise capital. Individuals can profit from the market with a relatively low required capital compared to starting a business. Also, stock markets are sources of information on the state of the economy. One can have an insight into the future of an economy.

For many experienced investors and analysts, improving one's ability to forecast stock prices is crucial. Because of shifting market patterns and a lot of background noise, it can be difficult to forecast how the stock market will move. The complex process by which stock values fluctuate is influenced by several factors, including market news, quarterly earnings reports, and changes in consumer behavior. To help them make decisions, traders utilize a variety of technical indicators based on daily stock data. It's still challenging to forecast daily and weekly market fluctuations even with this data. Accurately anticipating stock fluctuations is exciting and difficult in the fast-paced world of business. (Gandhmal & Kumar, 2019)

This project aims to predict hourly stock values using data from the Borsa Istanbul Stock Exchange Market. Making 300 correct daily predictions is the aim, using 30 stocks and 10 trading hours every day. The dataset, which covered the months of January 2018 through November 2023, made sophisticated methods like feature engineering necessary. After testing several models, the Weighted Mean Absolute Percentage Error throughout the validation period supported the Light Gradient Boosting Machine Regressor's superior performance in price prediction. To start off our modeling journey we have created a heatmap to identify similarities between stocks in order to create some insight of the features in other words shares.

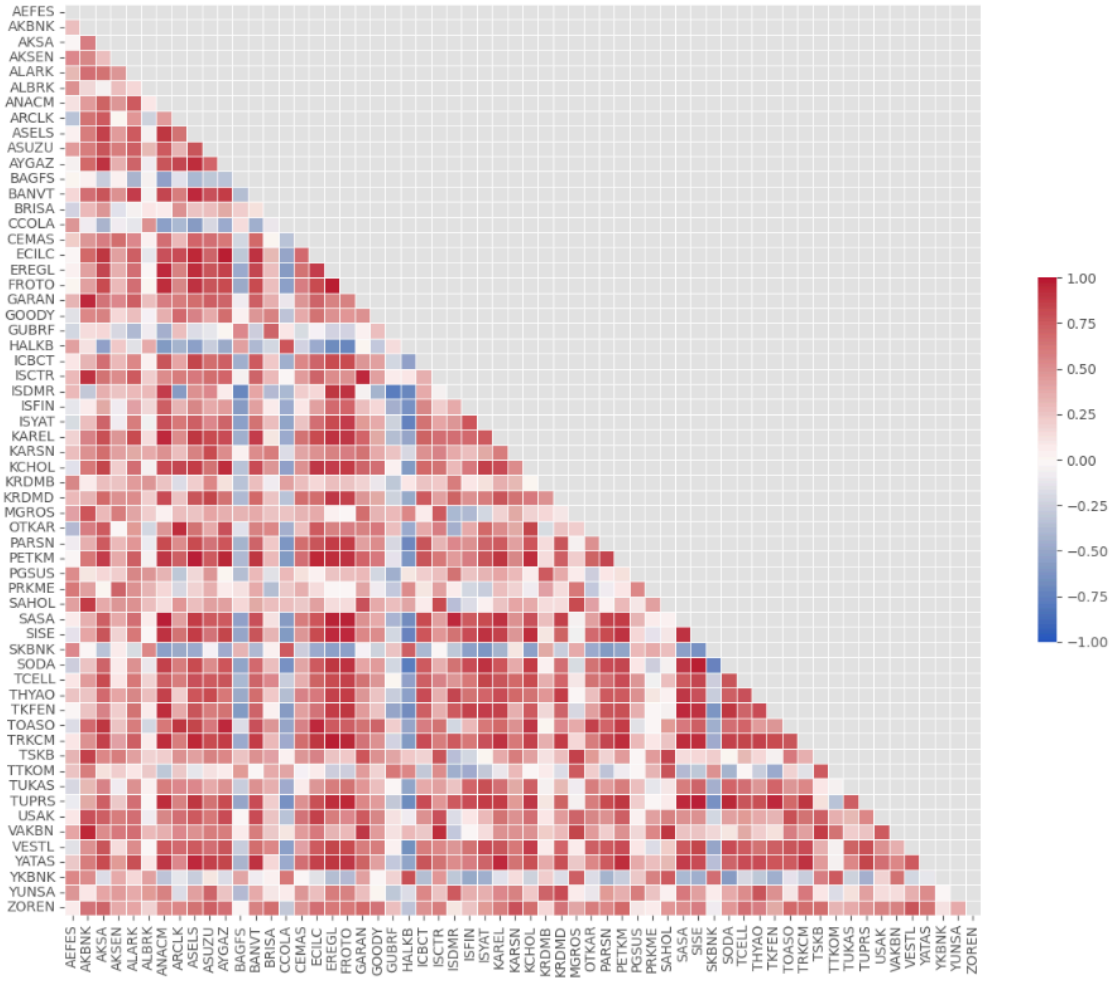

Figure. Heatmap of the Features

2. Related Literature

In order to construct a model, several machine-learning applications in the literature were analyzed. Ahuja & Kumar, have tried the three most well-known regression techniques Linear Regression, Random Forest, and Support Vector Regression for the purpose of forecasting the stock price. According to their results for predicting the stock prices of Tesla, they have determined that Linear Regression yields the best results since it minimizes MAE (Mean Absolute Error) in the closing result prediction. Based on their findings, it can be concluded that linear regression outperforms all other machine learning algorithms in terms of MAE which is 1.51. Furthermore, the random forest regressor shows a 1.94 mean absolute error following linear regression, which is higher than linear regression but lower than SVR (support vector regression) (Ahuja & Kumar, 2023).

An interesting approach to stock price prediction has been made by Christopher Bockel-Rickermann (2022). Bockell-Rickermann has used search engine query volumes in order to predict day-ahead stock returns which was actually discussed in our course. A gradient-boosted decision tree model is implemented to predict day-ahead stock returns. Three predictor sets were created: Historical Returns, Internet Search Volumes, and Generic

Information. Gradient Boosted Decision Tree models have run using the combination of these predictors and the results were evaluated by the Area under the ROC curves. (Bockel-Rickermann, 2022)

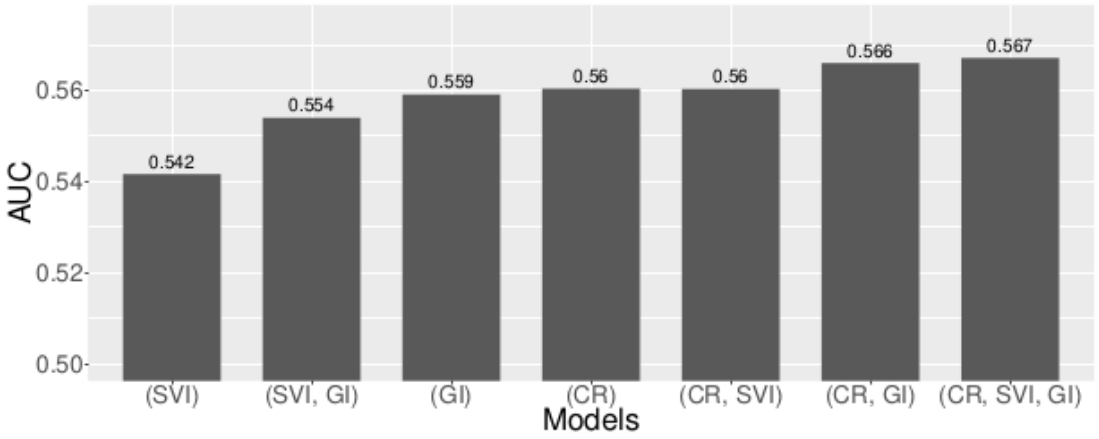

Figure. Performance Combinations of Three Predictors

Another interesting case we have analyzed is the M5 Accuracy Competition. Although the field of forecasting is different, the Machine Learning implementations of the competitors are notice worthy. The competition winner used a combination of different LightGBM models that were equally weighted. There were in total 110 models and two variations were taken into consideration for each type of model: a non-recursive forecasting approach and a recursive approach. A total of 220 models were constructed, and the average of six models, each with a unique training set and learning methodology, was used to anticipate each series. Negative log-likelihood was maximized in the optimization process. This is thought to be useful for processing data that have a probability mass of zero and a non-negative, highly right-skewed distribution (Makridakis et al., 2022). We thought this winning approach can be useful for our task of predicting stock prices.

3. Modelling Approach

In this project, financial data set are provided from Borsa Istanbul Stock Exchange Market and the main goal is to build a model that predicts hourly prices of the given stocks accurately. There are 30 stocks to be predicted and every stock market day includes 10-trading hours. Therefore, in total, 300 single prediction points have to be provided in daily manner throughout the forecasting challenge.

Initially, the data was provided for the 2018-01 & 2023-11 period. It includes 30 stocks and 10 trading hours. Due to the fact that hourly price information is not enough to explain prices in a modeling approach, feature engineering was applied in a variety of approaches. After the feature engineering stage, some of the regression learners were choiced to test their capabilities to make a final decision about the learner and their capabilities were tested through the time series metric that is called as Weighted Mean Absolute Percentage

Error on validation period. Light Gradient Boosting Machine Regressor with N target gave the best performance than other regressors.

Final performance was measured in the live competition environment through the 15 trading days that was between December 25, 2023 and January 12, 2024. Then, teams were ordered with their average daily rank.

Modeling approximation for the project will be described in detail in the following parts: Data Processing, Modelling and Evaluation.

4. Data Processing Phase

Since we will make a 10-hour price forecast for the next day with the hourly data we have, the share information for the same day can be used as a feature. For this reason, with the help of Yahoo Finance's Python library, various information of the 30 stocks we have were extracted with the help of api. With this API, daily information such as opening, high, low and closing prices, share volume and dividend distribution information for the relevant days of the relevant stock were obtained on a daily basis. Then, since financial markets and commodities are globally interrelated, different financial market indexes such as Nikkei 225, NYSE, and commodity information in US dollars such as dollar, euro currency equivalents in Turkish lira, crude oil, gold were extracted with Yahoo Finance API. Since

Financial market indexes are high-risk investment instruments, locals can also invest in companies in financial markets with the function of protecting the value of money. At the same time, companies can conduct various monetary policies with local indicators in Turkey. For this reason, monthly and annual inflation rates are obtained from the T.C. Central Bank. This variable is also used as data in the model. There is missing information about some stocks in the original data. Due to the fact that the number of missing values are not very high, they are linearly interpolated.

In financial markets, domain components provide different technical analysis. While these are sometimes technical indicators such as RSI and MACD, sometimes they are Fibonacci resistance levels and sliding averages. In floating markets, 5-day, 20-day, 60-day and 200-day averages are generally used on trading days. Throughout the studies, the storage, median, standard deviation, minimum and maximum values of stocks, commodities and financial market indices in the form of a sliding window were calculated for the days mentioned above. In this way, statistical information for the day is organized at the model stage.

In addition, the Lag variable was used for all individual feeling information in the time series loop of the time series problem of our problem. This information was added daily for a period of 1 day to 5 days. The adequacy of the time series includes basic variables such as year, month, day, day of the week, quarter information, day of the year, week of the year, and specific variables such as whether the relevant day is Monday morning or after Friday, which will generally explain the volumetric movements in financial markets, have also been added.

Financial price information is non-stationary data that contains a trend in nature. Therefore, the model trained with past price information may not produce the same sensitive

predictions today. Therefore, the hourly change will be estimated throughout the study in the light of various studies in the literature. Thus, the target was saved from the trend effect. Final price predictions will be obtained based on change rates hourly recursively or closing prices before the prediction day.

Financial price information is non-stationary data that contains a trend in nature. Therefore, the model trained with past price information may not produce the same sensitive predictions today. Therefore, the hourly change will be estimated throughout the study in the light of various studies in the literature. Thus, the target was saved from the trend effect. Final price predictions will be obtained based on change rates hourly recursively or closing prices before the prediction day.

In addition, since shares of the same market are priced with similar developments, information on all stocks will be used together in the same set. In this way, it is aimed to reduce noise learning.

Throughout the validation studies, the best result in the validation set was obtained by adding the estimated hourly changes to the closing prices. In addition, it is aimed to build models for different targets and combine the predictions taken with different loss functions to reach the final model.

5. Modeling, Hyperparameter Tuning and Evaluation

In the modeling stage, LightGBM regressor, CatBoost regressor and Random Forest regressor have been trained with the main processed dataset. LightGBM regressor and RandomForest regressor trained with multiple targets. Although it seems like global forecasting approach which is targets interact with each other in theory, it is just local model with N target (N=30) due to insufficient multi-target support of these regressors. However, so as to have a multi-target prediction property, CatBoost regressor have been trained with multi objective loss.

Because the selected learners has large number of hyperparameters, a hyperparameter tuning scheme is required to obtain better prediction performances. Random grid search method is employed for the random forest regressor with relatively narrow space and restricted iterations for the sake of simplicity. Max depth, min samples split, min samples leaf and max features are adjusted in the fine-tuning process. Optuna library is employed for LightGBM hyperparameter tuning stage. Optuna provides an elegant and automated way to search for the best hyperparameters by optimizing an objective function. It is designed to be user-friendly and highly adaptable to different machine learning frameworks. It has some key features like employing bayesian optimization, parallelization, extension ability and integrability. In order to tune the boosting model, a parameter search 3 hour is made with provided space. The parameter search space included number of estimators, max depth, bagging fraction, feature fraction, min child samples and more. During the search mean of WMAPE is returned.

The best performance is observed with the LightGBM regressor. LightGBM has outperformed CatBoost and Random Forest in terms of test performance and submission performance that is evaluated in the leaderboard system that is enabled to the teams. WMAPE is used as performance metrics and their average is taken to determine team ranks.

6. Results and Discussions

The following table shows the performance of learner models implemented. As it is mentioned above, one of the best performers happened to be the ensemble of Light GBM implementations with different targets, according to the WMAPE performance metric. It is noticed that LightGBM is capable of reaching better performance values than random forest with a tuning approach using Optuna. As a result the model marked with green is chosen as the final submission due to its high submission performance.

Learner Model Type	Additional Predictors	Target	Is Multitarget	Tuning Approach	Validation Performance	Submission Performance
Random Forest	FALSE	Hourly Percentage Change	FALSE	Random Grid Search	2.13%	None
Random Forest	OOB Stats	Hourly Percentage Change	FALSE	Random Grid Search	2.25%	None
LightGBM	FALSE	Hourly Percentage Change	FALSE	Default Parameters & Objective="Regression_l1"	1.84%	None
LightGBM	OOB Stats	Hourly Percentage Change	FALSE	Default Parameters & Objective="Regression_l1"	1.97%	None
LightGBM	FALSE	Hourly Price Difference in TRY	FALSE	Tuned Params with regression loss	1.72%	None
LightGBM	FALSE	Hourly Log Return	FALSE	Optuna 3h	1.60%	None
CatBoost	FALSE	Hourly Percentage Change	TRUE	Own Tuner with 1000 iterations	1.82%	None
Ensemble LGBM	FALSE	PCT Change, Log Return, TRY Difference	FALSE	Tuned Params with regression_l1, mape and regression loss	1.64%	1.75%

Table. Modeling Approaches Summary

7. Conclusion & Future Work

In conclusion, the challenge of forecasting hourly stock values using data from the Borsa Istanbul Stock Exchange Market was taken on by this project. Following extensive evaluation, the Light Gradient Boosting Machine Regressor was found to be the most accurate model. The project integrated insights from different strategies that have been discussed in related literature into the data processing stage. This required utilizing various technical analyses, handling financial data, and addressing missing values. The favorable results from actual trading days highlight the importance of machine learning for financial forecasting, particularly when it comes to adjusting to the particularities of stock markets. This project adds to our knowledge of stock price prediction and emphasizes how crucial it is to refine models in order to adapt to shifting market conditions.

Involvement of Search Engine Volumes as predictors and enlargement of the ensemble with equally weighted models and their with-replacement and without-replacement versions can be briefly summarized as our future research topics as they have been covered in the “Related Literature” section of this paper.

8. Code

https://github.com/ssehacirit/IE582-Semestr-Project

9. References

Ahuja, R., Kumar, Y., Goyal, S., Kaur, S., Sachdeva, R. K., & Solanki, V. (2023). Stock price prediction by applying machine learning techniques. *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*. https://doi.org/10.1109/esci56872.2023.10099614

Bockel-Rickermann, Christopher. (2022). Predicting Day-Ahead Stock Returns using Search Engine Query Volumes: An Application of Gradient Boosted Decision Trees to the S&P 100. 10.48550/arXiv.2205.15853.

Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of Stock Market Prediction Techniques. *Computer Science Review*, *34*, 100190. https://doi.org/10.1016/j.cosrev.2019.08.001

Kondiparthy, O., Nimmagadda, S., Venkata, G. K., Gera, J. R., Burugari, V. K., & Kailasam, S. (2023). A concise survey on predicting stock prices. *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*. https://doi.org/10.1109/icaccs57279.2023.10113004

Larsen, J. (2010). Predicting Stock Prices Using Technical Analysis and Machine Learning.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 Accuracy Competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013