IE582 - Fall 2023 Project

Group - 7

Dilan UĞUŞ - 2022700111

Rabia GÜL ÇELİK - 2021700180

Introduction

In the modern phase of big data, extracting meaningful insights from large datasets is increasingly crucial across various fields. Stock market prediction, in particular, is a complex task due to the volatile and unpredictable nature of financial markets. Yet, it remains a critical aspect, as accurate predictions can lead to significant economic benefits, inform investment strategies, and enhance market understanding. This project focuses on the financial sector, specifically Borsa İstanbul, utilizing advanced data analytics and learning techniques to forecast stock prices.

The task involves predicting the next day's hourly average prices for 30 companies listed on Borsa İstanbul. This involves generating ten predictions per trading day, covering the hours from 9 AM to 6 PM. To achieve this, the project employs a deep learning approach, specifically using LSTM (Long Short-Term Memory) neural networks, a type of recurrent neural network that is well-suited for time series data like stock prices.

The project's methodology encompasses data preprocessing, model training, and forecasting. The data preprocessing phase involves meticulous cleaning and preparation of the data, including removing duplicates and handling missing values. This phase is fundamental to the success of the model, as high-quality, clean data is essential for effective learning applications.

The core of the project is the modeling, which is responsible for initializing, training, and deploying the LSTM model. LSTM networks are particularly suited for this project due to their ability to remember long-term dependencies, a critical feature when dealing with complex financial time series data. The forecasting part complements this by focusing on generating accurate intraday forecasts, which are essential for short-term trading strategies and risk assessment.

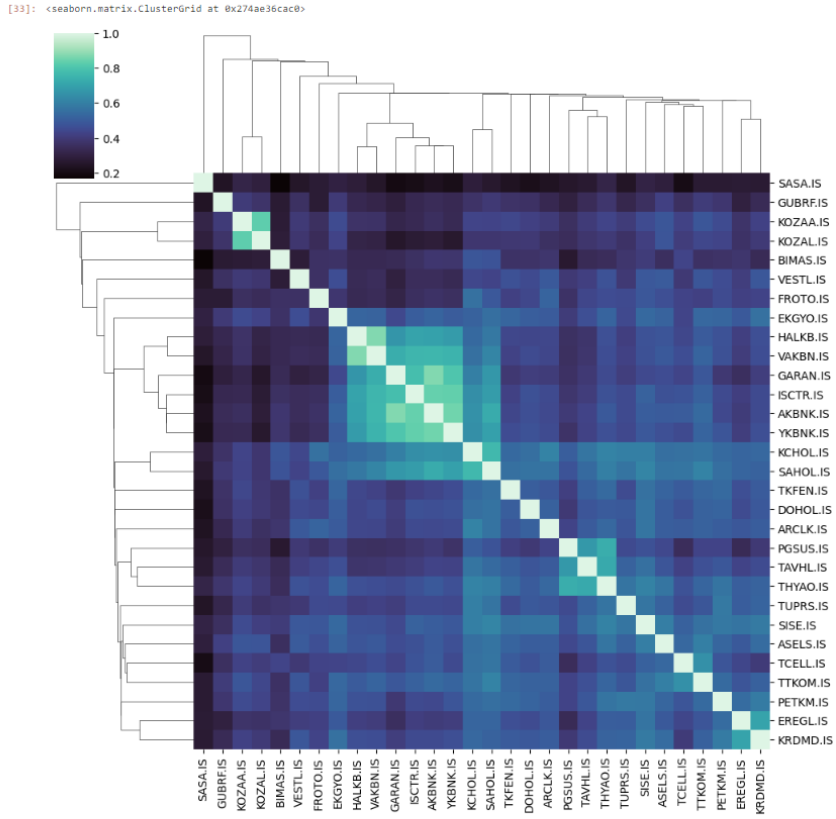
Data Analysis

The dataset for this project includes historical hourly average prices of selected stocks from Borsa İstanbul, covering approximately four years. This extensive dataset enables a comprehensive analysis and more accurate forecasting.

Each file contains time series data for stock prices in Borsa İstanbul, structured as follows:

- **Timestamp**: The date and time of the record, with hourly. The timestamp is timezone-aware (UTC+3).
- **Price**: The average price of a stock at the given timestamp.
- **Short Name**: The identifier symbol of the stocks.

The plot that has been provided in Figure 1 appears to be a cluster map of stock returns, which is a quantitative tool used to measure how stock prices are related to one another. Here’s a detailed interpretation of this figure and its potential implications:



```
[33]: <seaborn.matrix.ClusterGrid at 0x274ae36cac0>
```

Figure 1: Cluster map of Stock Prices

High Correlation Coefficients: When two different stocks have a correlation coefficient close to 1, it implies that their prices move in a similar direction during the same period. For example, AKBNK.IS and YKBNK.IS has a very high correlation of 0.852951, suggesting that they often move in the same direction, which is typical for stocks that might be in the same industry or affected by similar economic factors.

Low or Negative Correlations: A low or negative number implies that the stocks move independently or in opposite directions. For example, SASA.IS seems to have low correlations with the rest of the stocks, with a value of 0.226124 with AKBNK.IS, suggesting that it doesn't move together with AKBNK.IS.

High correlations within certain groups of stocks could indicate a sectorial relationship. For example, if AKBNK.IS and GARAN.IS both belong to the banking sector, which the high correlation of 0.872859 suggests, that it might be unwise for an investor looking for diversification to invest heavily in both, as they are likely to be affected by the same economic factors. This correlation matrix can be a valuable tool for understanding the relationships between different stocks, informing portfolio construction, risk management, and investment strategy development.

The effectiveness of the predictive models is evaluated using the Weighted Mean Absolute Percentage Error (WMAPE), a sophisticated metric that accounts for varying sizes and importance of different data points in the dataset. This metric is particularly appropriate for the financial domain, where the impact of forecasting errors can vary significantly across different stocks and market conditions.

This metric aligns well with the LSTM model's capacity to handle complex time series data, allowing for a nuanced understanding of the stock market dynamics.

In conclusion, this project represents a sophisticated and practical application of statistical learning and data mining techniques in the realm of financial analysis, with the ultimate goal of developing a forecasting product that can be applied in various business scenarios such as inventory management or service level optimization. The effective use of LSTM neural networks to analyze and predict stock market trends in Borsa İstanbul demonstrates the potential of deep learning in financial analysis. This approach is expected to assist businesses in reducing waste and better understanding the risk implications associated with market uncertainty.

Related Literature

During the literature survey, a thorough examination of current stock market prediction systems has been conducted. Forecasting stock returns has emerged as a vital area of research.

In a well-known current paper[1], the author underlines the significance of the stock market as an indicator of a country's economic health and acknowledges the inherent risk and potential for substantial returns in stock market investments. The use of LSTM, a type of artificial neural network adept in time series analysis, is justified by its ability to handle data with multiple input and output time steps effectively, making it a suitable choice for stock price prediction. The introduction of the ARO algorithm as a means to optimize LSTM's hyperparameters is a novel approach, aiming to enhance the accuracy of stock market predictions. The paper's methodology, including the use of the DJIA index stocks as a dataset and the application of evaluation criteria such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R2 score, reflects a comprehensive and rigorous approach to model comparison. The findings that the LSTM-ARO model outperforms its counterparts in prediction accuracy signal a promising direction for the application of metaheuristic algorithms in financial forecasting.

In another study[2], Shastry et al.'s study stands out for its focus on the National Stock Exchange of India Limited (NSE), offering a localized perspective on stock prediction. The LSTM model described in the paper is structured to learn context-specific temporal dependencies, a crucial feature for capturing the intricacies of stock market data. By utilizing historical data for training, the model incorporates a range of share features like opening price, day high, day low, closing price, total trade quantity, and turnover. The authors underscore the LSTM's ability to address long sequences and its superiority over other deep learning methods in handling the problem of vanishing gradients, a common challenge in recurrent neural networks.

In Sidra et al.[3], the research involves collecting and processing historical data from the NSE of India over five and a half years to develop a robust forecasting framework. The authors hypothesize that machine learning and deep learning models can learn from the features of past movement patterns of daily NIFTY 50 index values to accurately forecast future index values. This approach contrasts with the traditional methods which often struggle to accurately predict the dynamic and fast-changing patterns in stock price movements. The paper utilizes long short-term memory (LSTM) network-based models, leveraging their capability to overcome issues like vanishing and exploding gradient problems common in recurrent neural networks (RNNs). These LSTM models are designed to handle multivariate time series forecasting using a walk-forward validation method, which is both realistic and practical for real-world applications.

The paper was presented at the CAINE 2019 conference by Achyut Ghosh et al.[4], discusses the challenges and methodologies in predicting stock prices using LSTM models. The research proposes a framework employing LSTM models alongside algorithms for calculating companies' net growth. This approach allows for the analysis and prediction of future growth based on past stock prices. The methodology includes preprocessing the data from the Bombay Stock Exchange (BSE), feature selection focusing on date and closing price, training LSTM models with various layers and activation functions, and evaluating the predictions against actual prices. The findings reveal that LSTM models significantly reduce error levels across different periods and sectors, suggesting their effectiveness in long-term stock price prediction.

The paper by Ferdiansyah et al.[5], presented at the International Conference on Electrical Engineering and Computer Science (ICECOS) 2019, focuses on predicting Bitcoin prices using Long Short-Term Memory (LSTM) networks. This research acknowledges the need for automated tools capable of predicting Bitcoin's future market behavior, given its significant fluctuations and influence on the stock market. LSTM, known for its ability to maintain information over long periods, is particularly suited for time series data like Bitcoin prices. The study demonstrates the LSTM model's capability to predict future prices by training and testing the model on historical Bitcoin data from Yahoo Finance.

APPROACH

The dataset consists of hourly stock prices of 30 different stocks traded on Borsa Istanbul. The objective of this project is to create a day-ahead forecast across 30 companies for hourly average prices. LSTM model is used to make these predictions for the following reasons: We considered the sequential nature of time series data, the need to capture long-term dependencies in stock price movements, and the ability to handle non-linear patterns. The flexibility of LSTMs in learning features from historical data, mitigating the vanishing gradient problem, and their proven success in capturing temporal lags and patterns in financial time series were crucial factors and the literature review supported the suitability of LSTM for stock price prediction tasks. Overall, the LSTM model aligns well with the characteristics of the problem at hand, offering a robust solution for forecasting hourly stock prices.

Data Preprocessing: Historic prices are stored in distinct cvs files for different time intervals for four years in longitudinal format. In the first step of preprocessing, all CSV files are read and stored in a data frame. The input data frame size was 446182 by 2. Then missing or undefined values are identified and excluded from the dataset to ensure integrity and the accuracy of analyses. At the end of preprocessing, we have a data frame with 446182 rows and 2 features.

Feature Engineering: In the feature engineering stage, we introduced day sine, day cosine, year sine, and year cosine features alongside stock prices to enhance the LSTM model's understanding of temporal patterns in stock price movements. The day sine and cosine characteristics encode the cyclic nature of daily fluctuations, enabling the model to discern recurring patterns within 24 hours. Similarly, the year sine and cosine attributes furnish the model with insights into seasonal variations spanning different years, facilitating the identification of enduring trends. By integrating these temporal features, our model gains sensitivity to the time of day and year, enhancing its capacity to capture both short-term seasonality and long-term trends inherent in the stock price data. As a result of the feature engineering process, we obtained a data frame with five features, including stock prices, day sine, day cosine, year sine, and year cosine.

Modeling, Tunning & Evaluation: In the model creation and tuning phase, a unique LSTM model was generated for each of the 30 stocks, utilizing different scalers. The LSTM model consists of a sequential neural network with two LSTM layers for learning temporal dependencies, followed by global max pooling to capture key features and a dense layer for producing the final output. In recognition of the unique characteristics inherent to each stock, the necessity arose to employ distinct models and scalers tailored to individual stocks. This approach acknowledged that there was no singular global optimal model applicable for making accurate predictions across all stocks, underscoring the importance of customizing the modeling process for each stock's specific behavior and trends. Therefore, 30 different models are created, and fitted. The resulting models were then exported, each tagged with the corresponding stock name. Evaluation of the test scores for these models provided insights into their predictive performance. For the test dataset, the most recent 90 days of stock prices were isolated, leaving the rest for training purposes. The following figure shows test predictions and real stock values for SASA. As can be seen from the figure test results can capture increases and decreases of stock movements with 0.0161 mape for SASA.

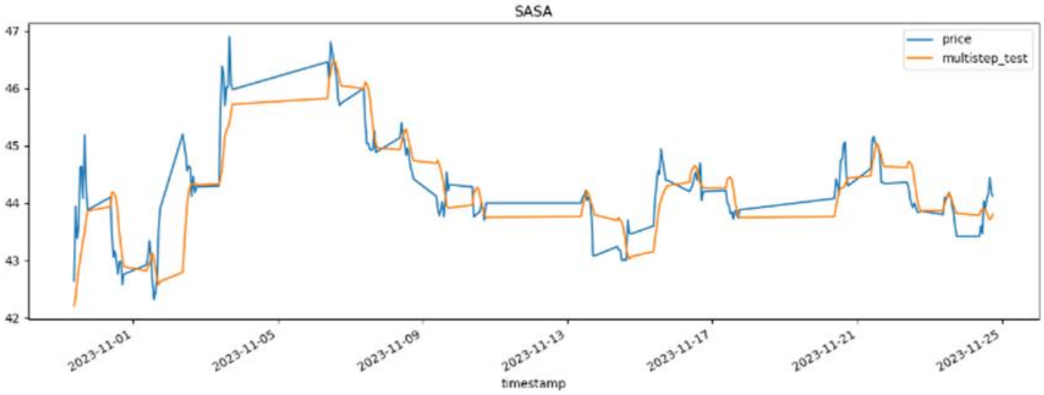


Figure 2: Stock Values for SASA

Subsequently, a grid search was conducted for each model to fine-tune the parameters of the LSTM architecture, specifically focusing on layer 1 and layer 2 and epoch number parameters. The outcome of the grid search revealed optimal parameters, which were stored for each stock in a dictionary. To finalize the models, they were reconstructed using these identified best parameters of each stock. The following two figures show MAPE numbers for each stock before and after GridSearch.

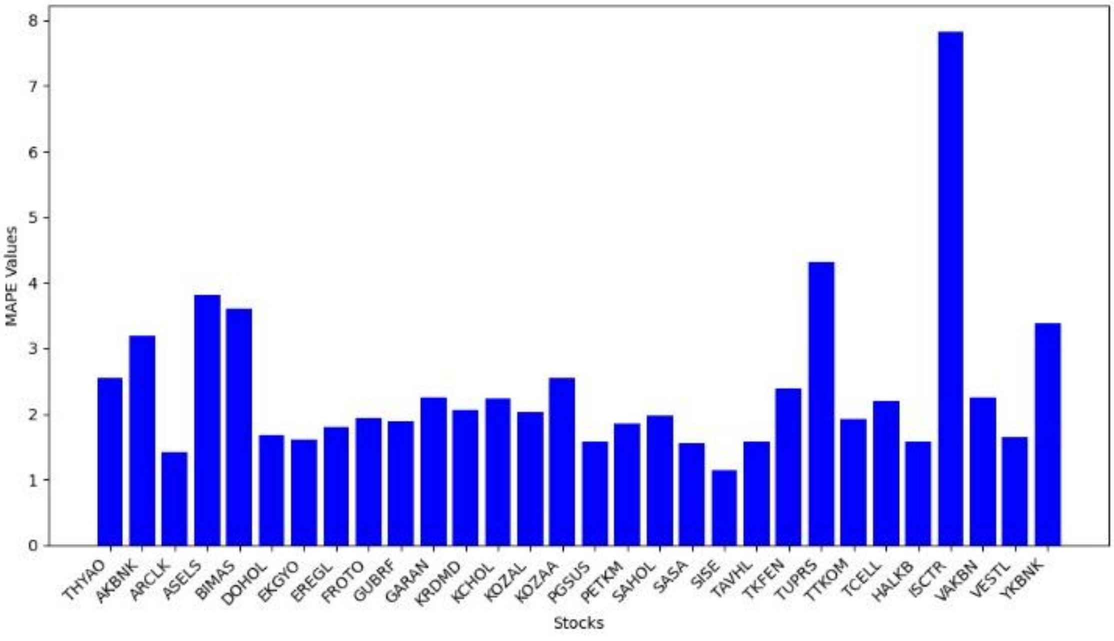


Figure 3: MAPE Results for Different Stocks Before GridSearch

After GridSearch is done, the best parameters are stored in the best parameters dictionary for each stock and used in creating final models for each stock. The highest improvement is seen on ISCTR stock, from 0.0687 mape to 0.03011 mape. The lowest mape is achieved for HALKB stock while the highest mape is obtained when predicting ASELS stock.

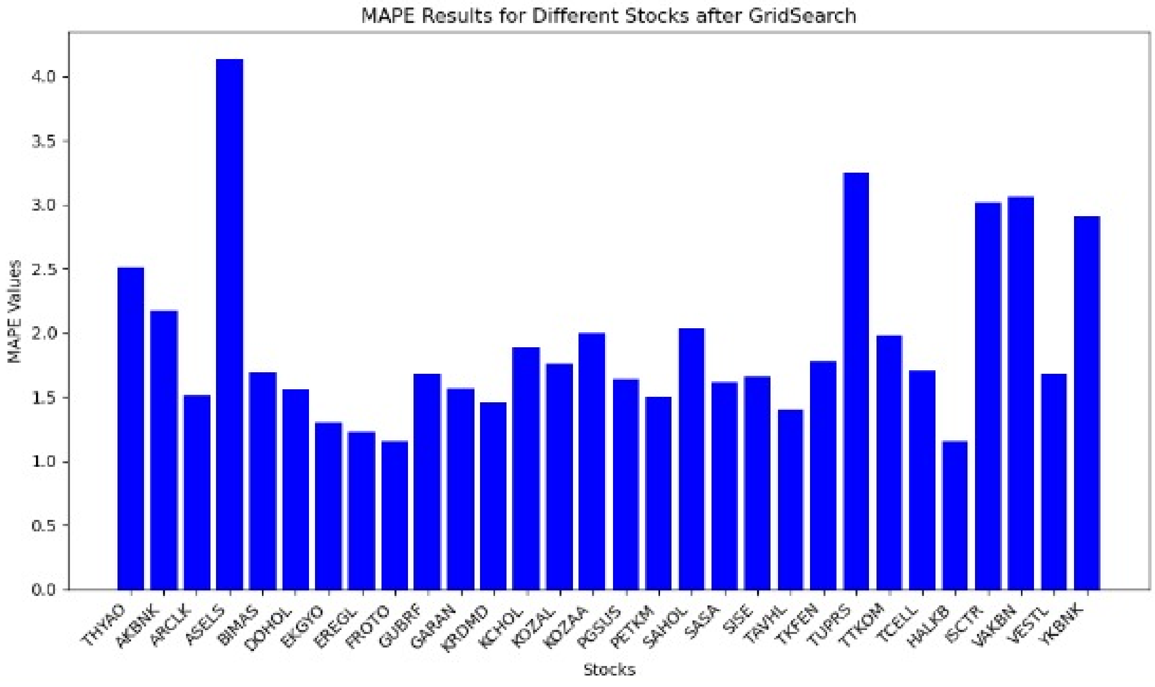
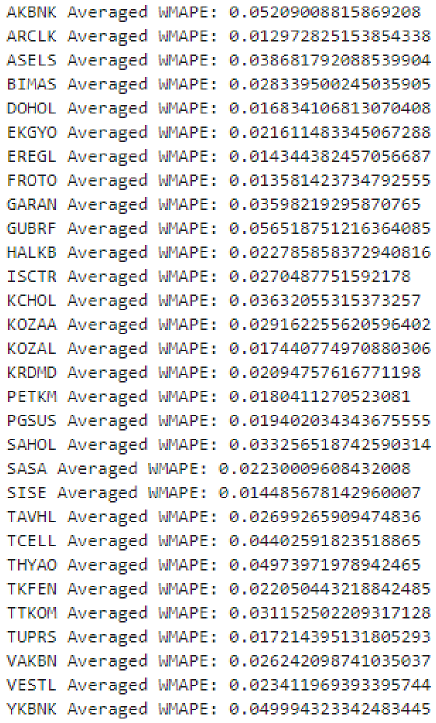


Figure 4: MAPE Results for Different Stocks After GridSearch

The average mape for 30 stocks was 0.0234 before GridSearch and it decreased to 0.0193 after parameter tuning. The models are then prepared for deployment. In the prediction phase, the relevant model and corresponding scaler are imported from the folder, allowing for accurate and tailored forecasting for each stock.

Results

In this study, the performance of a forecasting model over 14 days was evaluated by using the Weighted Mean Absolute Percentage Error (WMAPE) metric for 30 stocks. Results of each day included individual WMAPE values for stocks such as AKBNK, ARCLK, and ASELS. Distinguishing trends emerged in the day-to-day performance of individual stocks, as certain ones consistently showcased lower WMAPE values, indicating more accurate predictions. In contrast, others exhibited higher volatility in their forecasts, underscoring the variability in predictive outcomes for these specific stocks.



```
AKBNK Averaged WMAPE: 0.05209008815869208
ARCLK Averaged WMAPE: 0.012972825153854338
ASELS Averaged WMAPE: 0.038681792088539904
BIMAS Averaged WMAPE: 0.028339500245035905
DOHOL Averaged WMAPE: 0.016834106813070408
EKGYO Averaged WMAPE: 0.021611483345067288
EREGL Averaged WMAPE: 0.014344382457056687
FROTO Averaged WMAPE: 0.013581423734792555
GARAN Averaged WMAPE: 0.0359821929587O765
GUBRF Averaged WMAPE: 0.056518751216364085
HALKB Averaged WMAPE: 0.022785858372940816
ISCTR Averaged WMAPE: 0.0270487751592178
KCHOL Averaged WMAPE: 0.03632055315373257
KOZAA Averaged WMAPE: 0.029162255620596402
KOZAL Averaged WMAPE: 0.017440774970880306
KRDMD Averaged WMAPE: 0.02094757616771198
PETKM Averaged WMAPE: 0.0180411270523081
PGSUS Averaged WMAPE: 0.019402034343675555
SAHOL Averaged WMAPE: 0.033256518742590314
SASA Averaged WMAPE: 0.0223000960843200８
SISE Averaged WMAPE: 0.014485678142960007
TAVHL Averaged WMAPE: 0.02699265909474836
TCELL Averaged WMAPE: 0.04402591823518865
THYAO Averaged WMAPE: 0.0497397197894246５
TKFEN Averaged WMAPE: 0.022050443218842485
TTKOM Averaged WMAPE: 0.031152502209317128
TUPRS Averaged WMAPE: 0.017214395131805293
VAKBN Averaged WMAPE: 0.026242098741035037
VESTL Averaged WMAPE: 0.023411969393395744
YKBNK Averaged WMAPE: 0.049994323342483445
```

Figure 5: Averaged WMAPE for each Stock

The evaluation results for the Average Weighted Mean Absolute Percentage Error (WMAPE) across various stocks reveal distinctive performance metrics as shown in Figure 5. Notably, ARCLK exhibited the most accurate predictions with an impressively low Averaged WMAPE of 0.01297, indicating a high level of precision in its forecasting. The daily WMAPE value of this stock is plotted in Figure 6. Other stocks such as EREGL, FROTO, and SISE also demonstrated commendable accuracy with Averaged WMAPE values below 0.02.

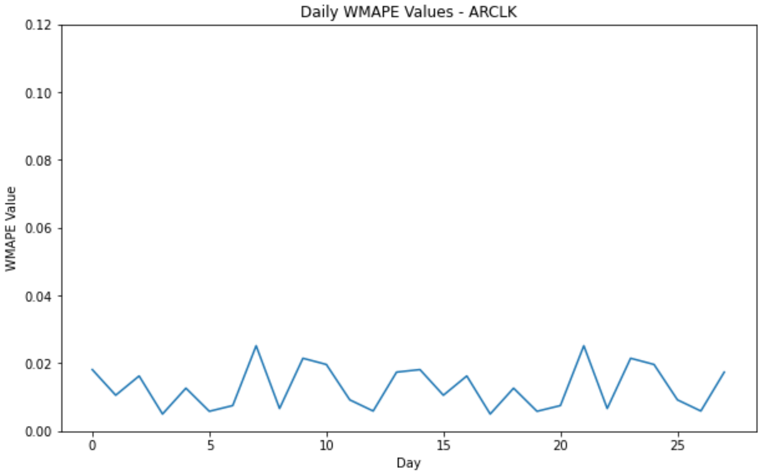


Figure 6: Daily WMAPE value of ARCLK

On the other hand, GUBRF stands out with a relatively higher Averaged WMAPE of 0.05652, suggesting increased volatility and less accurate predictions compared to the aforementioned stocks. The daily WMAPE value of this stock is plotted in Figure 7. YKBNK and THYAO also showed higher Averaged WMAPE values of 0.04999 and 0.04974, respectively, reflecting a higher degree of variability in their predictive performance.

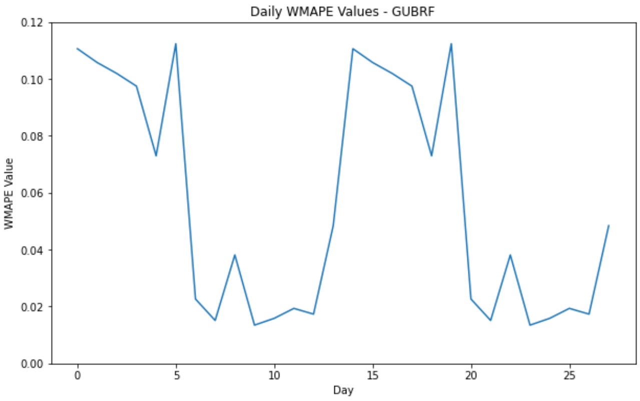
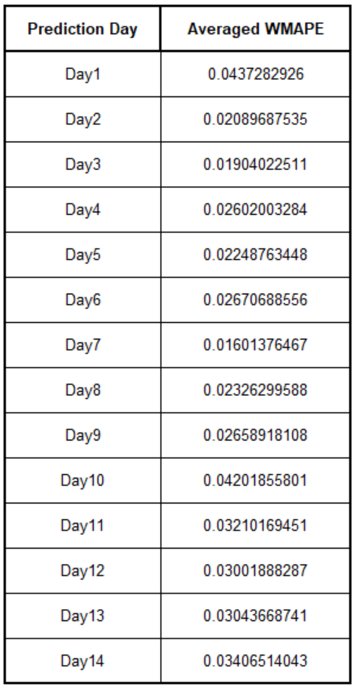


Figure 7: Daily WMAPE value of GUBRF

The average WMAPE for the entire stock was computed for each day, providing a comprehensive view of the stability and accuracy of the model. Days with lower average WMAPE values suggested better overall predictive performance for the entire stock predictions. The analysis also can make possible the identification of outliers—stocks with significantly higher WMAPE values than the daily average—conducting consideration for further investigation into the reasons behind these deviations.

The evaluation results for the Average Weighted Mean Absolute Percentage Error (WMAPE) across the prediction days exhibit varying levels of predictive accuracy as shown in Figure 8. Notably, Day 7 stands out with the lowest Averaged WMAPE of 0.01601, indicating a high degree of precision in the predictions made on that day. Day 3 and Day 14 also demonstrate commendable accuracy, with Averaged WMAPE values of 0.01904 and 0.03407, respectively.

Conversely, Day 1 and Day 10 present relatively higher Averaged WMAPE values of 0.04373 and 0.04202, respectively, suggesting a lower level of accuracy in the predictions for those days. Day 6, Day 9, and Day 4 also exhibit higher Averaged WMAPE values, indicating a comparatively increased degree of variability in their predictive performance.



Prediction Day	Averaged WMAPE
Day1	0.0437282926
Day2	0.02089687535
Day3	0.01904022511
Day4	0.02602003284
Day5	0.02248763448
Day6	0.02670688556
Day7	0.01601376467
Day8	0.02326299588
Day9	0.02658918108
Day10	0.04201855801
Day11	0.03210169451
Day12	0.03001888287
Day13	0.03043668741
Day14	0.03406514043

Figure 8: Daily Averaged WMAPE value for Whole Stocks

In conclusion, this 14-day analysis provides valuable insights into both individual stock performance and the overall effectiveness of the forecasting model. The trends observed and the presence of outliers offer ways for model refinement and optimization, enhancing its predictive capabilities for future assessments. These results provide insights into the daily fluctuations in predictive accuracy, offering valuable information for refining forecasting models and making informed decisions based on the reliability of predictions for each day.

Conclusions and Future Work

Throughout the competition period, day-ahead forecasts were generated for each stock following the methodology outlined in previous sections, resulting in an average Mean Absolute Percentage Error (MAPE) of 0.028 by the conclusion of this period. Notably, the performance varied across individual stocks, with the initial week exhibiting superior forecasting accuracy compared to the subsequent week. The observed decline in performance during the second week suggests the possibility of a pattern change that the model did not promptly detect. Specifically, the model demonstrated its most robust performance in stock ISCTR, with a MAPE of 0.027, as depicted in the above figures, while facing its most challenging prediction task in stock GUBRF, with a MAPE of 0.056. It is noteworthy that the MAPE outcomes of the developed model closely align with findings documented in the literature review, particularly referencing [1].

The model's performance could be further enhanced by incorporating additional stock market-related features that may influence stock prices. Considerations such as supply and demand dynamics, trading volumes, and broader economic indicators could provide valuable context to the model. Additionally, integrating external data sources, such as Google Trends statistics specific to each stock, could capture the impact of public interest and sentiment on stock movements. While we have successfully introduced time-dependent features like day sine, day cosine, and year sine to capture temporal patterns, exploring more sophisticated time-related features, such as intra-day trends, seasonal anomalies, or macroeconomic events, could provide a richer representation of the underlying dynamics influencing stock prices. Furthermore, experimenting with sentiment analysis on financial news articles or social media discussions related to each stock might offer insights into market sentiment, contributing to a more comprehensive predictive model.

Code

https://github.com/BU-IE-582/fall-23-ugusdilan/tree/main/PROJECT

References

[1] Gülmez, B. (2023). Stock price prediction with optimized deep LSTM network with artificial rabbit's optimization algorithm. Expert Systems with Applications, 227, 120346.
[2] Pramod, B. S., & Pm, M. S. (2020). Stock price prediction using LSTM. Test Engineering and Management, 83, 5246-5251.
[3] Mehtab, S., Sen, J., & Dutta, A. (2021). Stock price prediction using machine learning and LSTM-based deep learning models. In Machine Learning and Metaheuristics Algorithms, and Applications: Second Symposium, SoMMA 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 2 (pp. 88-106). Springer Singapore.
[4] Ghosh, A., Bose, S., Maji, G., Debnath, N., & Sen, S. (2019, September). Stock price prediction using LSTM on the Indian Share Market. In Proceedings of 32nd international conference on (Vol. 63, pp. 101-110).
[5] Ferdiansyah, F., Othman, S. H., Radzi, R. Z. R. M., Stiawan, D., Sazaki, Y., & Ependi, U. (2019, October). A lstm-method for bitcoin price prediction: A case study Yahoo finance stock market. In 2019 International Conference on Electrical Engineering and Computer Science (ICECOS) (pp. 206-210). IEEE.