IE582- Stock Prices Prediction Project

Elif Atılgan- Group 5

1. Introduction:

This project investigates the prediction of hourly average stock prices for 30 companies listed on the Borsa İstanbul Stock Exchange. Utilizing statistical learning and data mining techniques, the aim is to generate accurate forecasts for the upcoming trading day, specifically targeting the crucial hours between 9 AM and 6 PM. While traditional financial data will serve as the primary foundation for analysis, we acknowledge the potential value of incorporating external data sources such as Google Trends to enrich our prediction models. To address the time-series nature of the data, we will employ strategic feature engineering techniques, transforming the raw hourly price data into features capable of capturing both temporal patterns and pertinent market dynamics. This approach will enable the application of robust classification methods, extending their utility beyond conventional domains and opening new avenues for financial forecasting.

The primary objectives of this project are:

- To develop accurate models for predicting hourly average stock prices on the Borsa İstanbul using statistical learning and data mining techniques.
- To demonstrate the effectiveness of feature engineering in enabling the application of classification methods to time-series data.
- To contribute to the advancement of financial forecasting methodologies within the context of the Borsa İstanbul.
- To explore the potential benefits of incorporating external data into the prediction process.

2. Literature Review:

When it comes to stock prices prediction there are two types of predictors: internal and external. For extracting features using internal dynamics, the article of Nikhil Kohli (2020) is used as an inspiration. In his article, he mainly talks about two things:

a. **Technical Indicators:** Bollinger Bands, Exponential Moving Average (EMA), Average True Range (ATR), Average Directional Index (ADX), Commodity Channel Index (CCI), Rate-of-Change (ROC), Relative Strength Index (RSI), William's %R, and Stochastic %K.
b. **Lead-Lag Effects:** lagged stock prices from previous many days.

While Kohli's emphasis on established technical indicators (Bollinger Bands, EMAs, ATR, etc.) remains foundational, closer examination reveals further promising dimensions for feature extraction:

- *Leveraging Broader Market Dynamics:* Kohli (2020) emphasizes the influence of major market indexes like QQQ, S&P, and DJIA on individual stock prices. By incorporating lagged prices and technical indicators from these indexes, we can capture the broader market sentiment and its impact on specific companies. This akin to transitioning from examining a single company to analyzing the entire market landscape, gleaning valuable insights that might otherwise remain hidden.
- *Integrating Global News Sentiment:* Kohli's (2020) vision for incorporating news sources adds further depth to the predictive framework. Sentiment analysis of relevant news articles, financial reports, and social media trends can provide a real-time pulse of public opinion and market expectations. This can be likened to eavesdropping on the market's conversations, unearthing hidden anxieties and excitement that influence investor behavior.
- *Embracing Macroeconomic Factors:* External factors like interest rates, inflation, and economic indicators exert significant pressure on stock prices. Integrating these data points into our feature set fosters a holistic understanding of the broader economic context shaping market movements. This akin to stepping back and observing the forces of the global economy pulling and pushing individual stocks on their complex trajectories. In fact, another study by Buğra Bağcı (2020) revealed

that stock prices in Istanbul Stock Market are significantly affected by inflation rate, gold prices, industrial production index, money supply, exchange rate, credit volume, and internal debt stock.

3. Approach:

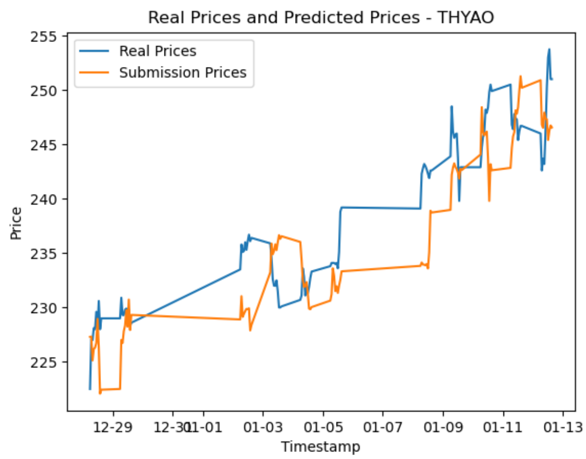
In this project, my main approach entailed a preprocessing of financial data for individual companies that are present in the data. This involved a detailed approach to extracting time-based features and addressing potential data gaps or missing values. Having stored all the data in an object called "hourly_data.”, the dataset was enriched by appending the information from an additional CSV file, "data2," representing the latest day's data that needed to be consistently integrated into the existing dataset. This integration ensured a continuous and updated representation of financial information.

After that, attention turned to the project's core, where the XGBoost machine learning method was applied iteratively. Tailored specifically for each company within the dataset, these XGBoost regression models were carefully constructed with the goal of minimizing prediction errors. During the preprocessing stage, moving averages and lag prices were created, which are subtle traits that are important indicators for identifying temporal trends in financial data. The dataset was strategically partitioned into training and test sets, with the training set serving as the crucible for training the XGBoost models. These models were then used to forecast hourly average stock values for the following day, using today's predictors as tomorrow's important features after being adjusted to the particular financial environment of each company. This comprehensive and iterative approach not only highlighted the nuanced intricacies of financial time-series forecasting but also displayed a strategic methodology for handling varied financial datasets. The project's output provided a comprehensive and detailed evaluation of each model's predictive ability for the range of organizations that were taken into consideration.

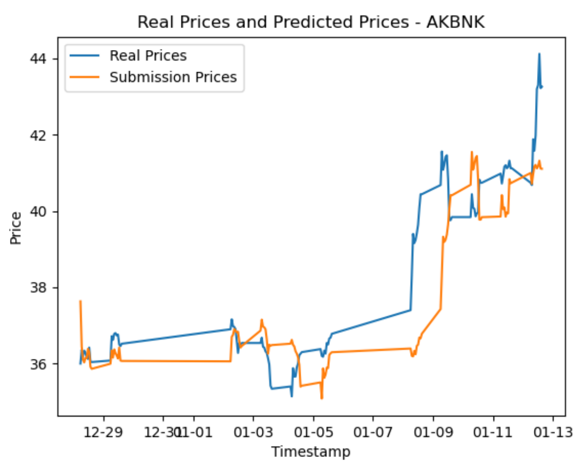
4. Results:

Average WMAPE value of my predictions is 0.01987 if we disregard the first two days due to malfunction in my computer that is irrelevant of my approach. My best WMAPE that is for 3rd of January is 0.01539 and my worst one that is for 8th of January is 0.03111. The results for each day and for each company can be summarized via the following plots:

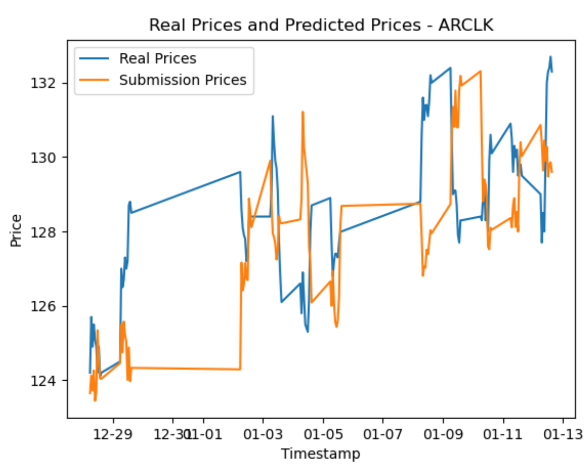
- **THYAO**



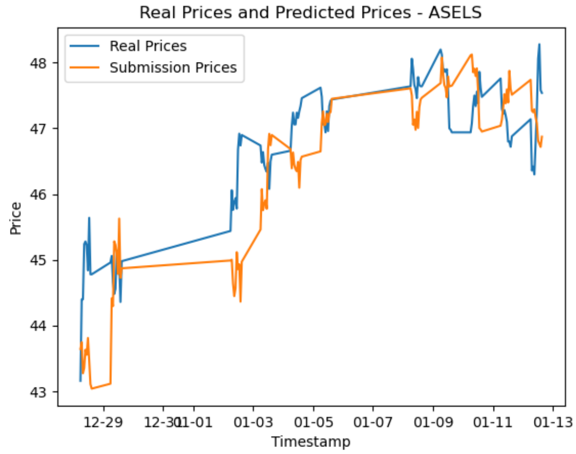
- **AKBNK**



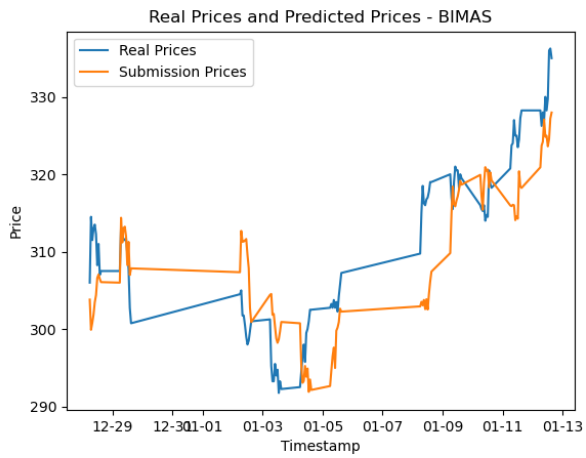
- **ARCLK**



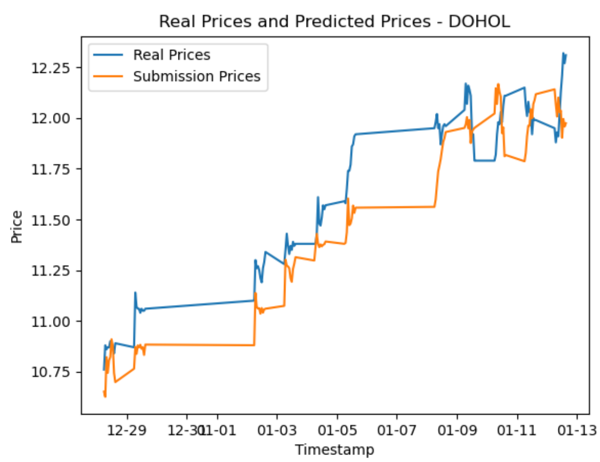
- **ASELS**



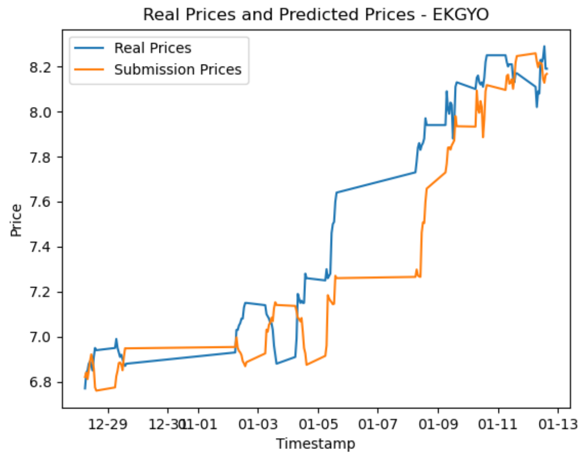
- **BIMAS**



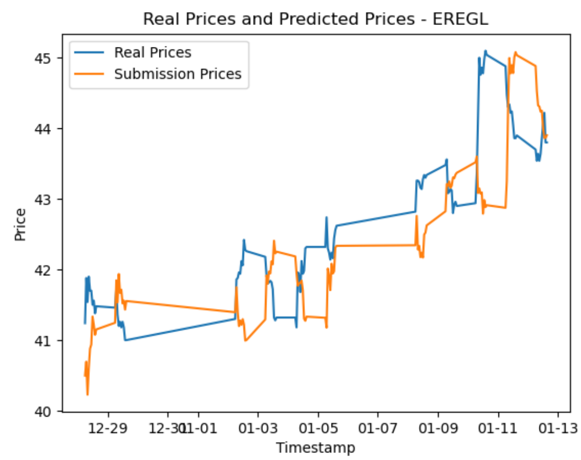
- **DOHOL**



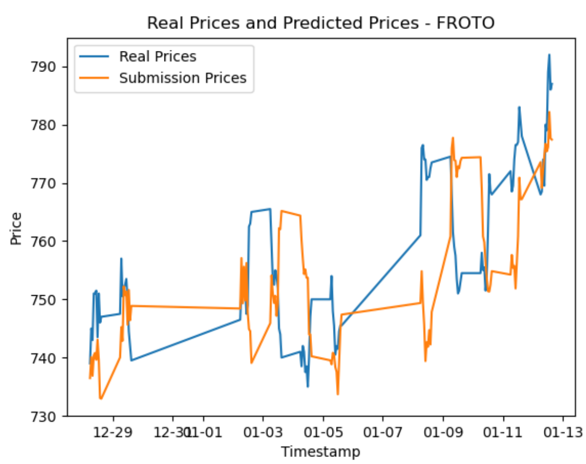
- **EKGYO**



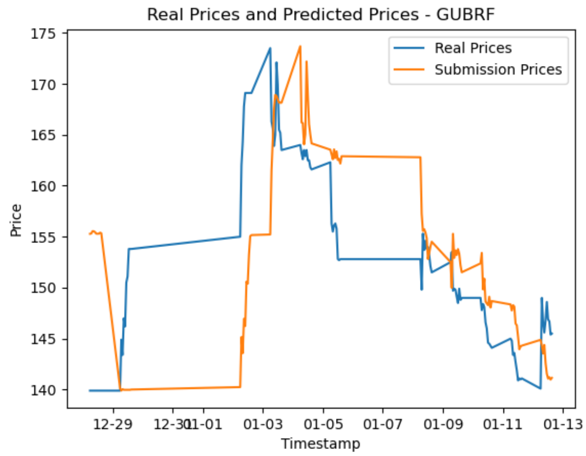
- **EREGL**



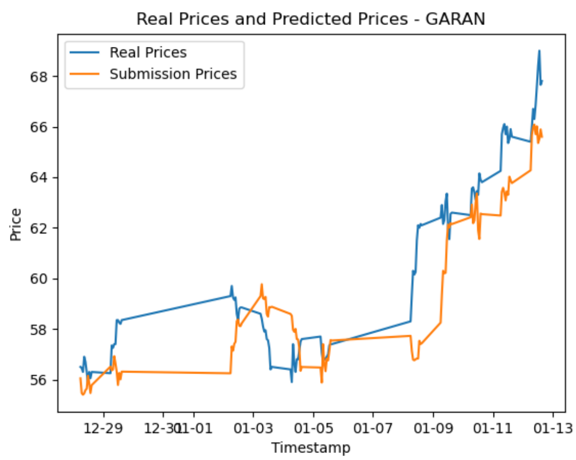
- **FROTO**



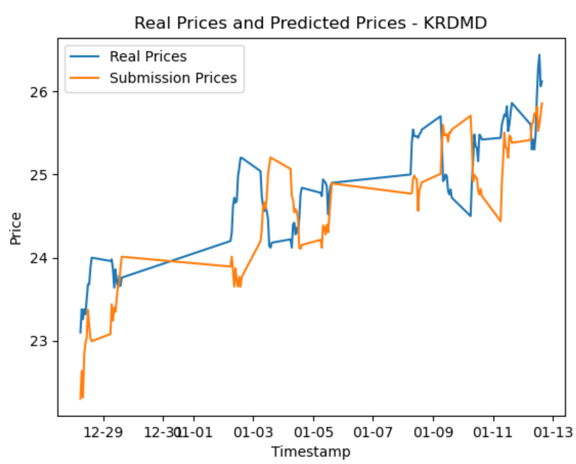
- **GUBRF**



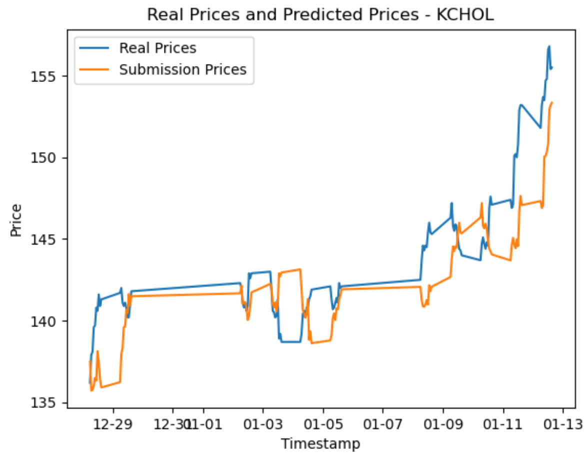
- **GARAN**



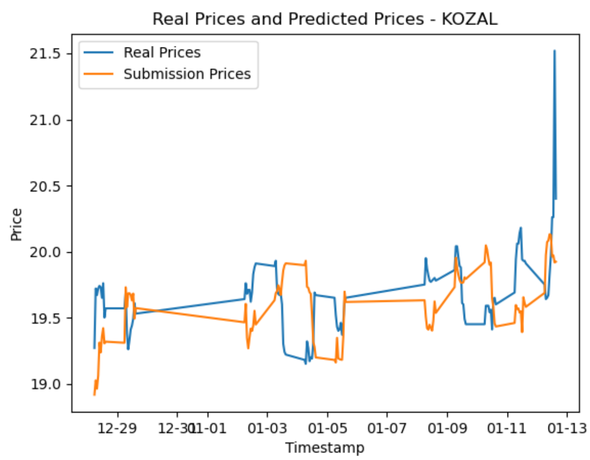
- **KRDMD**



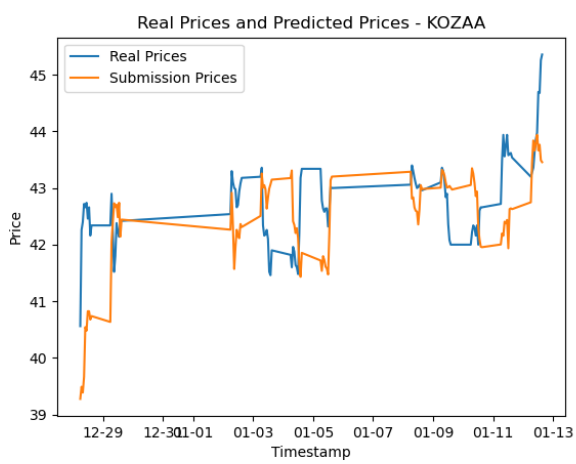
- **KCHOL**



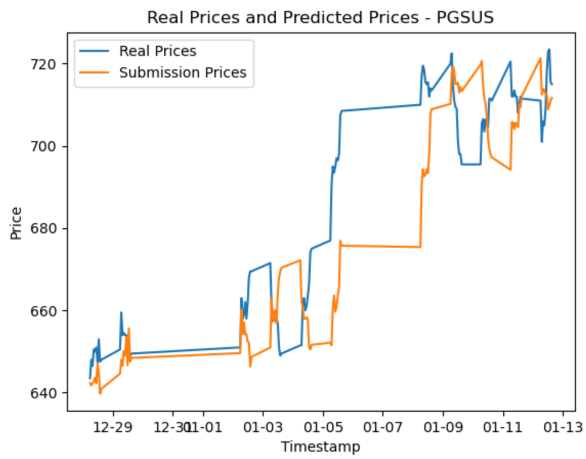
- **KOZAL**



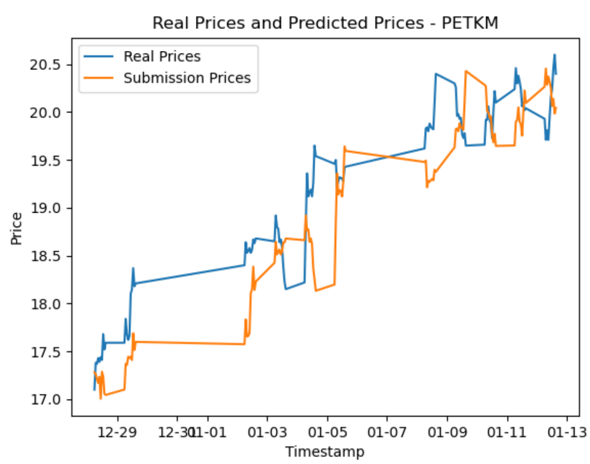
- **KOZAA**



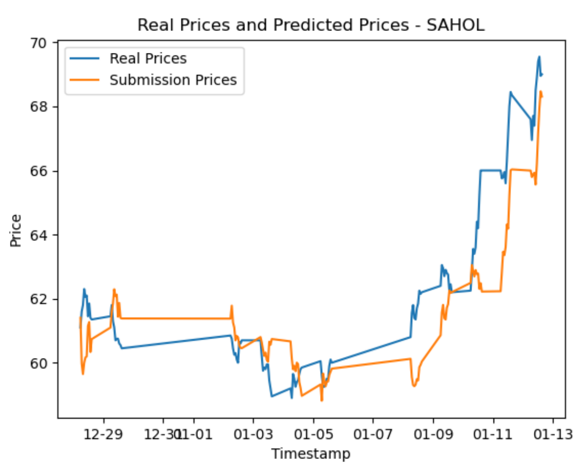
- **PGSUS**



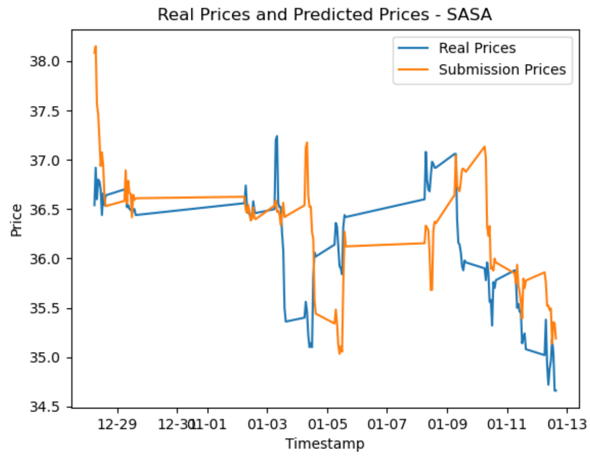
- **PETKM**



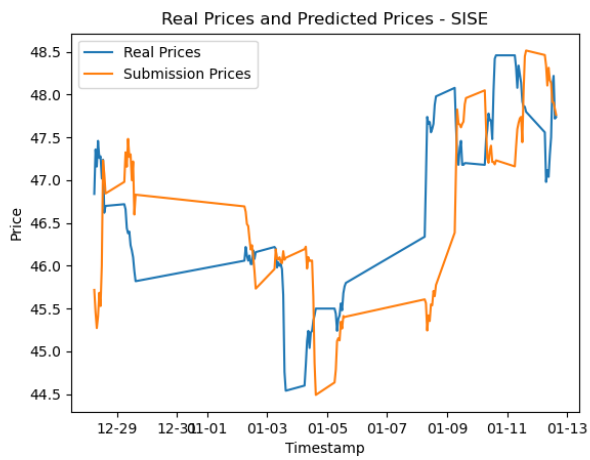
- **SAHOL**



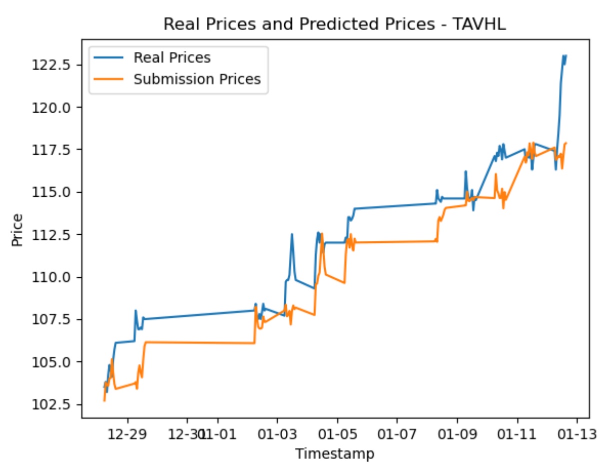
- **SASA**



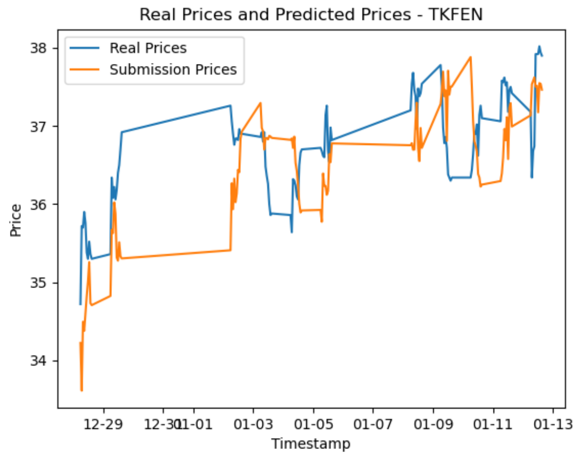
- **SISE**



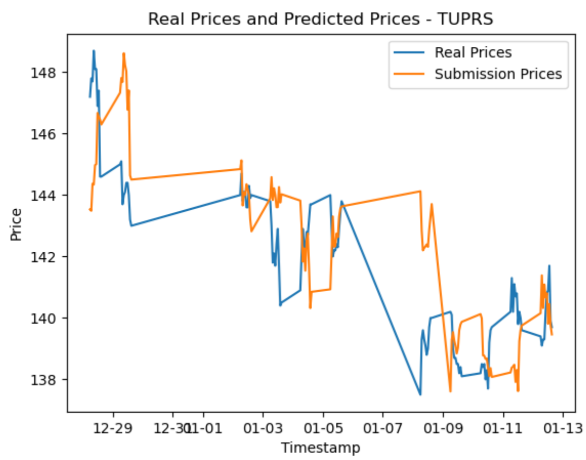
- **TAVHL**



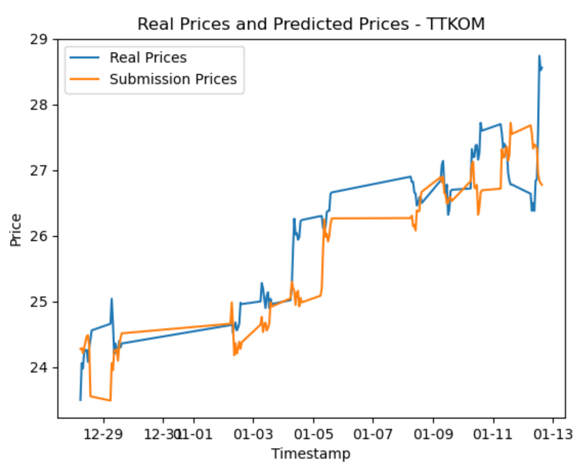
- **TKFEN**



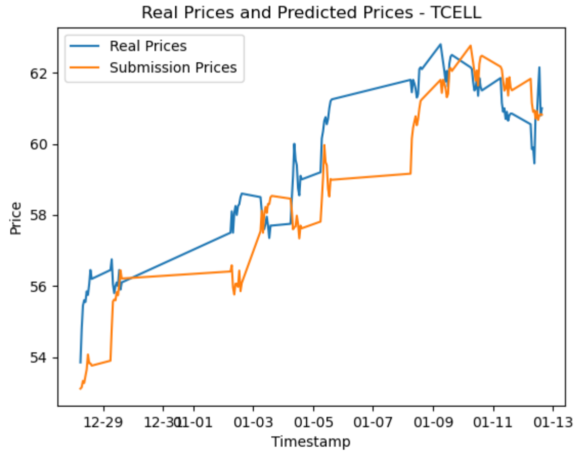
- **TUPRS**



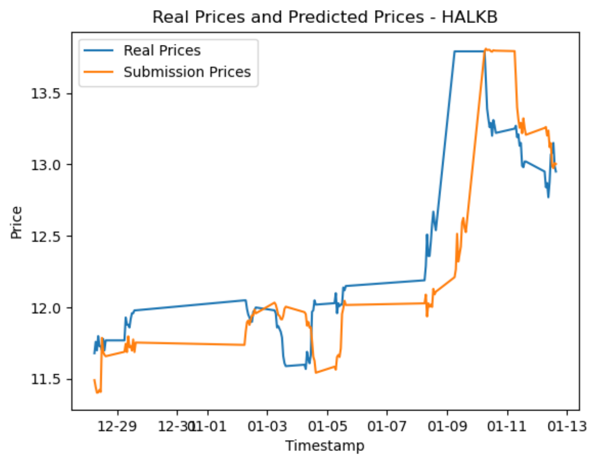
- **TTKOM**



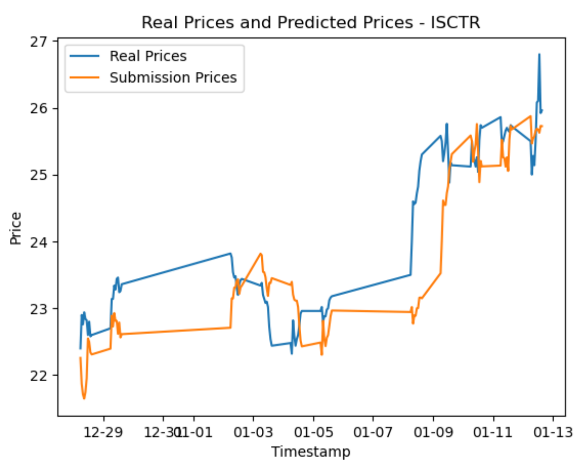
- **TCELL**



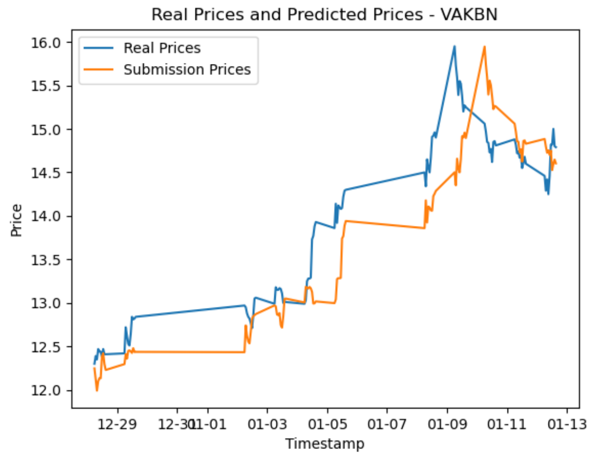
- **HALKB**



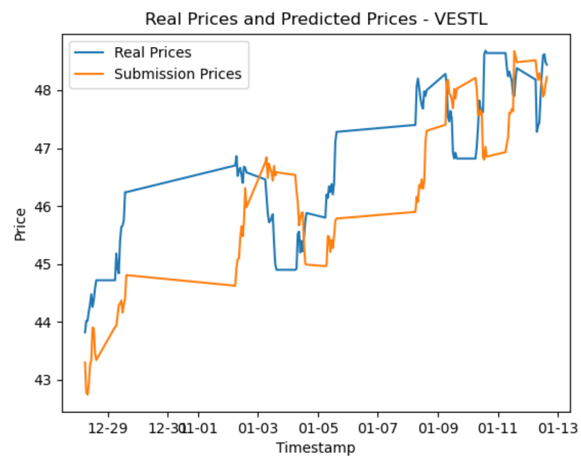
- **ISCTR**



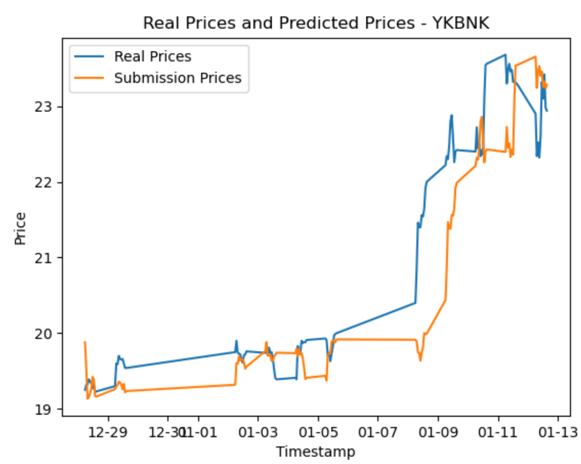
- **VAKBN**



- **VESTL**



- **YKBNK**



5. Conclusions and Future Work:

The XGBoost regression models developed in this study have demonstrated commendable accuracy in predicting hourly average stock prices based on historical data and relevant features. The implemented methodology, incorporating lagged prices and moving averages, has provided valuable insights into the temporal patterns of stock prices for various companies. However, as with any predictive model, there exist opportunities for refinement and enhancement to achieve even greater precision.

a. ***Predicting Future Predictors:***

   The reliance on predictors from the last day poses a significant limitation for forecasting future stock prices. A promising avenue for future work involves developing models to predict these predictors themselves. This could involve leveraging advanced time-series forecasting techniques or exploring machine learning algorithms specifically designed for predicting the evolving nature of financial features. By improving the accuracy of predictor forecasts, the overall predictive power of the model is expected to increase.

b. ***Incorporating External Variables:***

   The inclusion of external variables, such as Google Trends data and inflation rates, could further enhance the model's predictive capabilities. Integrating these external factors as additional predictors may provide a more comprehensive understanding of the complex dynamics influencing stock prices. Unfortunately, due to time constraints and unforeseen events, such as data loss on my computer, the incorporation of these variables was hindered. Future iterations of this research could prioritize the inclusion of external factors and explore their impact on forecasting accuracy.

In summary, even if the present model shows encouraging results, it will not be possible to adjust to the dynamic nature of financial markets without ongoing refining and expansion of the predictor set. By strengthening the model's predictive ability, the suggested improvements hope to provide a stronger foundation for upcoming stock price forecasting projects.

6. Code: https://github.com/BU-IE-582/fall-23-elifatilgan/tree/main/Project