Predicting Sectoral Growth in Turkey from Macroeconomic Indicators[1]

Team Members: Ahmet Yasin Aytar, Halil Ibrahim Ergül, Yasser Zouzou

1. Introduction

Considering the complicated outlook that Turkey's economy goes through these years, there has been a significant increase in the general public's interest in the stock market. It became tremendously prevelant for an average person to get involved in the stock market. For instance, according to observations of a newspaper (Habertürk), the number of young people who buy and sell shares in the stock market increased by 340%, exceeding 357 thousand. Therefore, informed decisions about buying, selling, or holding onto stocks is essential in bringing profits. What is more, making informed and solid predictions of stock value can also be important for businesses, as the value of their stocks can impact the company's access to capital and its policy of stocks. Overall, predicting stock value is important for both individual investors and businesses as it can impact decision-making and financial outcomes. These reasons motivated us to use of machine learning models for the purpose of this task.

The problem this project is dealing with falls mostly under the domain of Stock Market Prediction though some important divergences. For one thing, our objective is not to predict stock price per se, rather; being informed through the literature, we treat the market capitalization of companies (calculated by their stock prices) as their success/growth rate and we aim at predicting sectoral growth that different Turkish companies takes part in. Second, the task that this project focuses is not a prevalent field in the literature which generally focuses on stock prices of major indexes instead of looking at how sectors would behave differently with market values of stocks that are influenced through macroeconomic variables.

A number of research studies have looked at how various macroeconomic factors, such as gross national product (GNP), consumer price index, money supply, interest rate, and exchange rate, influence stock prices in Southeast Asian countries. According to Wongbangpo and Sharma (2002), the results of these studies show that stock prices in these countries are inclined to be negatively affected by inflation, and the relationship between interest rates and stock prices varies depending on the country. In some countries, higher interest rates are associated with lower stock prices, while in others, higher interest rates are associated with higher stock prices. Poon and Taylor (1991) examined the impact of macroeconomic variables on stock prices in the United Kingdom and they interestingly found that these variables had no effect on stock market prices.

[1] Our GitHub Repository: https://github.com/YZouzou/CS512-project

2. Problem description

The ultimate problem is that this domain mostly focuses on indicators derived from fundamental and technical analysis to assess the performance of various companies and the growth of different industries such as energy and textile (Soni, Payal & Tewari, Yogya & Krishnan, Deepa, 2022, pp. 1-4). However, the literature also suggests that the stock market is bound tightly to a country's economic growth; hereby emphasizing the importance of macroeconomic factors in understanding sectoral growth (Barth, James & Weng, Bin & Martinez, Waldyn & Tsai, Yao-Te & Li, Chen & Lu, Lin & Megahed, Fadel, 2018). In this respect, our aim is to predict the sectoral growth (which we measure by the market value of stocks for companies belonging to a certain sector) of different industries using the macroeconomic indicators in Turkey. Hence, this project differs from conventional stock prediction tasks as our research focuses on the question of to what extent macroeconomic factors might possibly help in predicting sectoral growth.

3. Method:

3.1. Data collection and preparation

The micro economic indicator datasets used in this study were manually collected mainly from the Turkish Statistical Institute (TUIK) and the Turkish Central Bank (TCMB). A summary of the datasets is provided below:

- **Total imports and exports (TUIK):** Total value of imports and exports in Turkish Lira (TL) and US Dollar (USD).
- **Export and import unit value indices (TUIK):** Computed using the base year as 2010 (2010=100).
- **Labor force statistics (TUIK):** Unemployment rates by age groups (%).
- **Consumer price index (TUIK):** Consumer price index (CPI) measures the change in the price of a market basket. CPI changes represent inflation changes. (Ref)
- **Gross domestic product (TUIK):** Gross domestic product (GDP) is defined as "the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period." (Ref)
- **Industrial Production Index (TUIK):** The industrial production index measures the changes in the output of different industries. The 2015 based (2015=100) industrial production index was used. (Ref)
- **Consumer Credits (TCMB):** Measured in thousands TRY
- **Current account deficit (TCMB):** The difference between the export and import values measured in million USD.
- **Money supply (TCMB):** "The money supply is the total amount of money—cash, coins, and balances in bank accounts—in circulation." (Ref)
- **Interest rates (TCMB)**
- **Real effective interest rate (TCMB)**: "Real effective exchange rate is the nominal effective exchange rate (a measure of the value of a currency against a weighted average of several foreign currencies) divided by a price deflator or index of costs." (Ref)

- **Retail sales (TCMB):** Retail sales reflect the spending in various retail categories. (Ref)
- **Gold and oil prices**

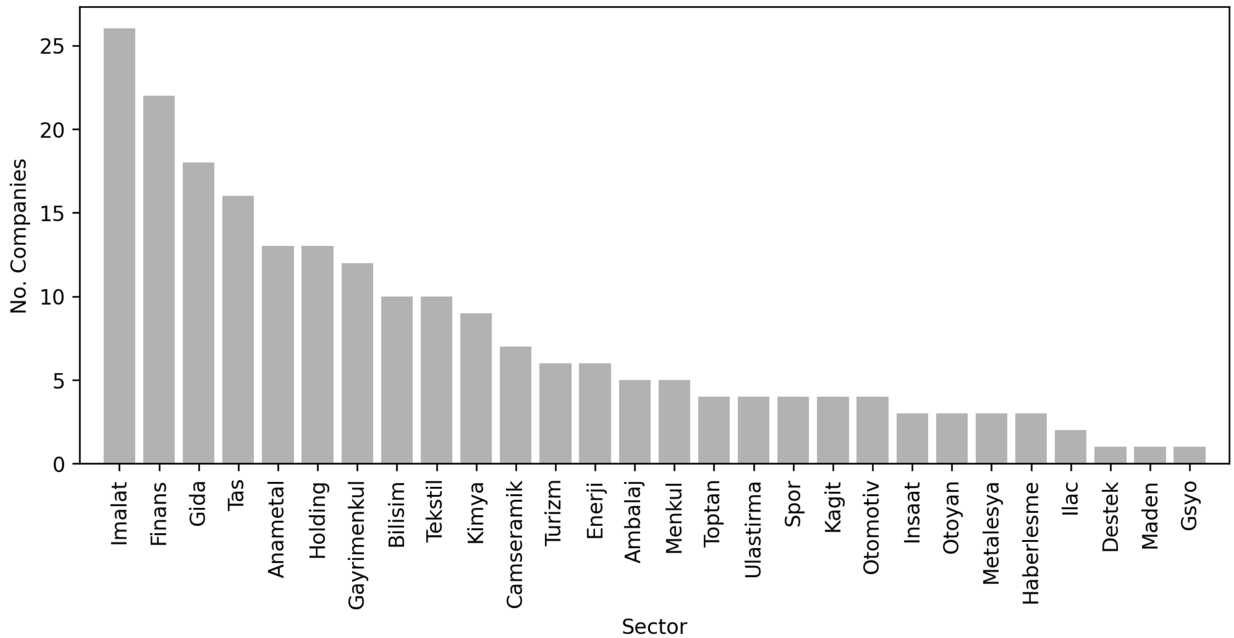

Fig. 1. Number of companies in each sector

The market capitalization value was used as a proxy for sectoral growth in this study. Market capitalization is the total value of a company's outstanding shares (Ref). The reason behind using market capitalization instead of directly using stock prices is that stocks are often subjected to splits which change the number and prices of shares, without altering the total market value. Thus, market capitalization is a more consistent value than stock prices. Market capitalization data was collected from Istanbul Borsa's official website. The dataset includes monthly market capitalization values of 215 companies from 2009 until the end of 2019. We chose to exclude the years after 2019 to avoid the drastic market changes that occurred post-2019 due to the pandemic. These companies belong to 28 different sectors as shown in Fig. 1. Time series of these companies based on their sectors are shown in Fig. 2 for the first 15 sectors by company count.

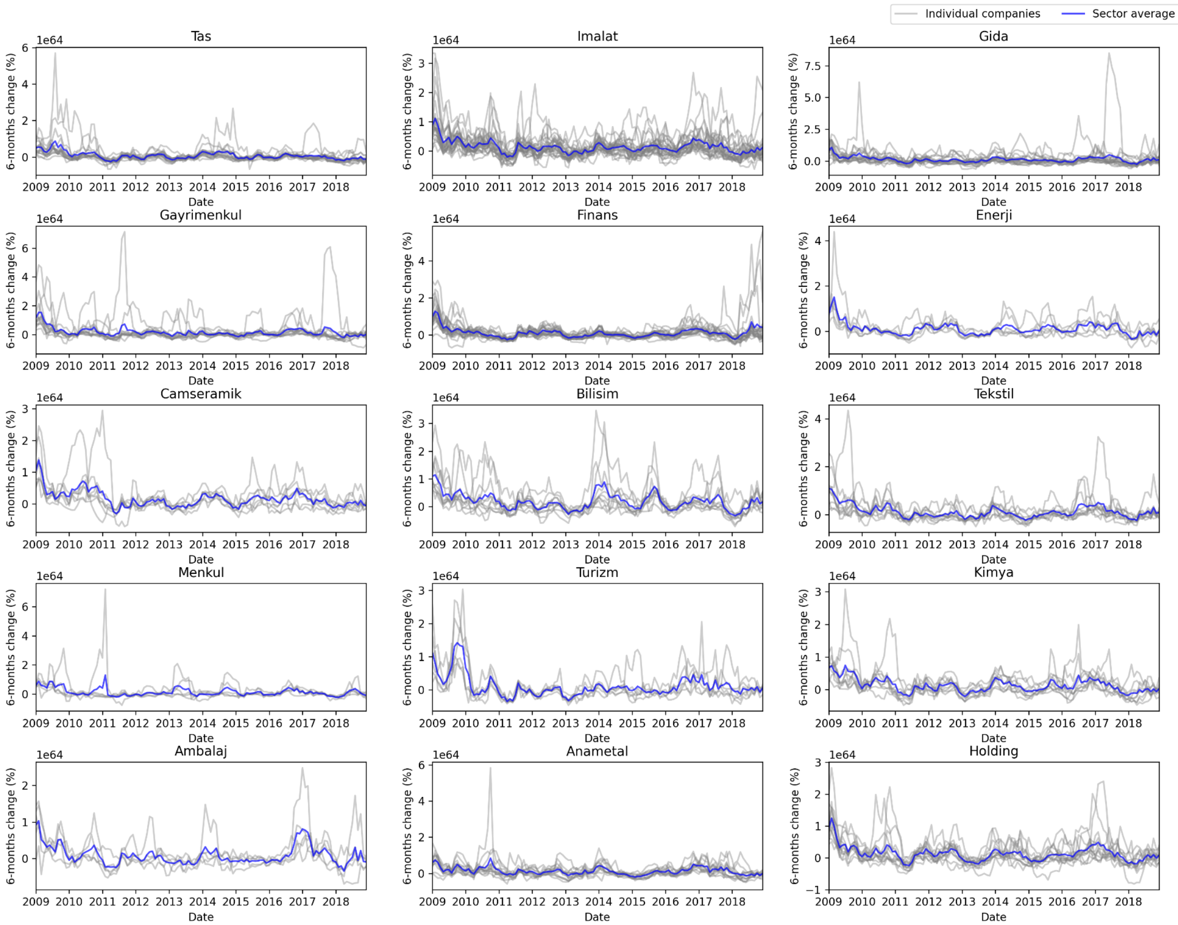

Fig. 2. Time series of companies in different sectors

The datasets collected were cleaned, brought into a unified format, and merged into one dataset for features and one dataset for the target variable.

3.2. Feature Engineering and Data Preprocessing

Some macroeconomic indicators had several versions that were highly correlated. The redundant indicators were eliminated based on advice from a financial expert. The main reduction in features occurred in the industrial production indices, which were reduced from 300+ to 7 indices.

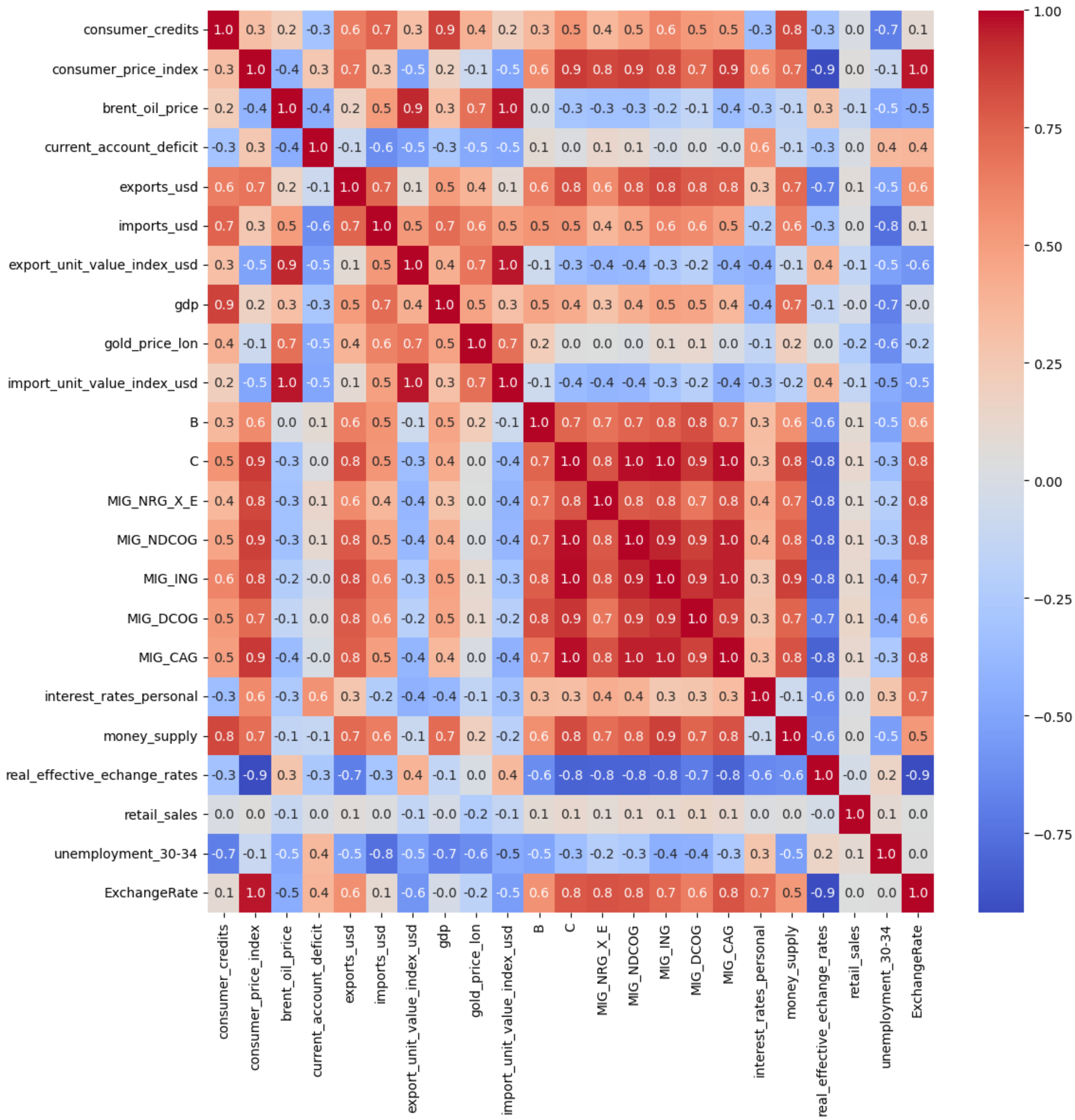

Fig. 3. Correlation heatmap

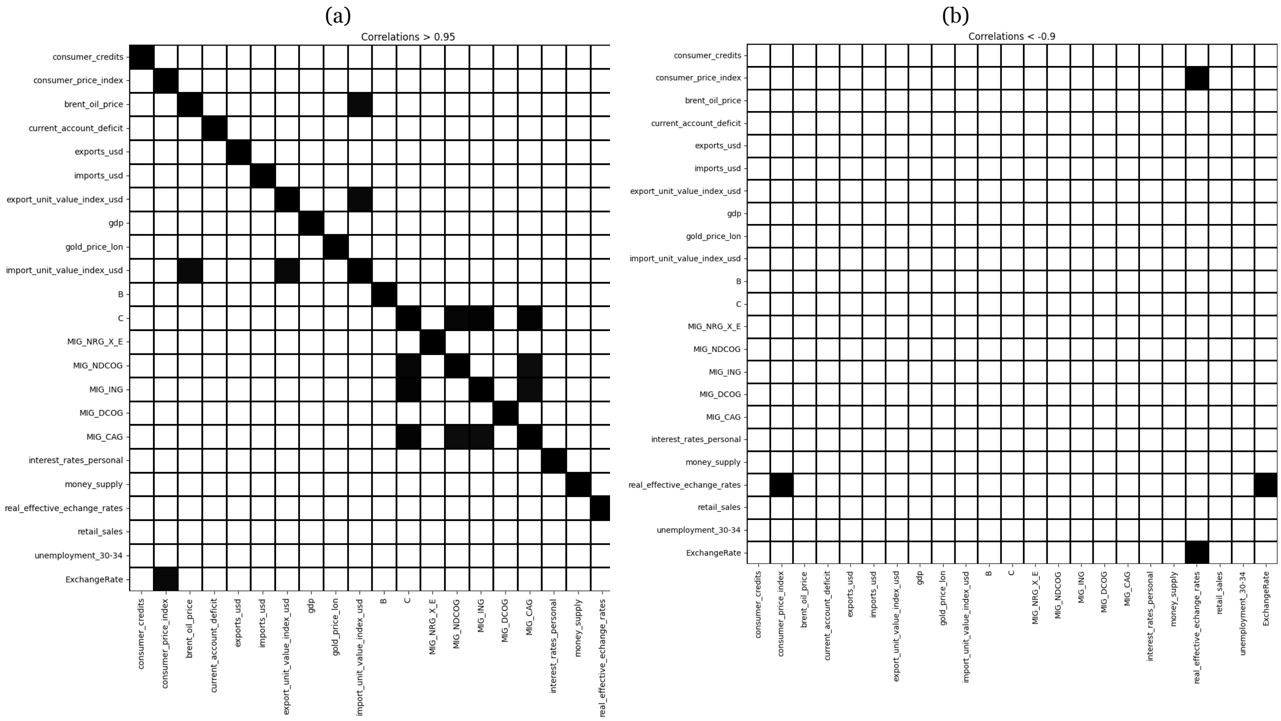

Fig. 4. Binary correlation map. (a) Positive correlations higher than 0.95. (b) Negative correlations higher than 0.9.

Subsequently, correlations between the different indicators were explored (Fig. 3). It can be observed that there are several variables that are very highly correlated. These features were labeled as “highly correlated features” and were eliminated when using models that are affected by multi-correlation, such as linear regression.

Two methods for outlier detection were tested: (1) Rolling mean plus/minus a threshold equal to the rolling standard deviation multiplied by a user-defined scalar (2) Rolling median plus/minus a threshold equal to the rolling interquartile range (IQR) multiplied by a user-defined scalar. As seen in Fig. 5, these methods fail to work with market data due to the lack of regularity. Long periods of minimal change shrink the bounds and make the method vulnerable to detect any small changes as outliers. In addition, these outliers do not represent wrong data values, they are real data that the deployed model will observe. Therefore, the decision was made to not remove any outliers to keep the model results realistic.

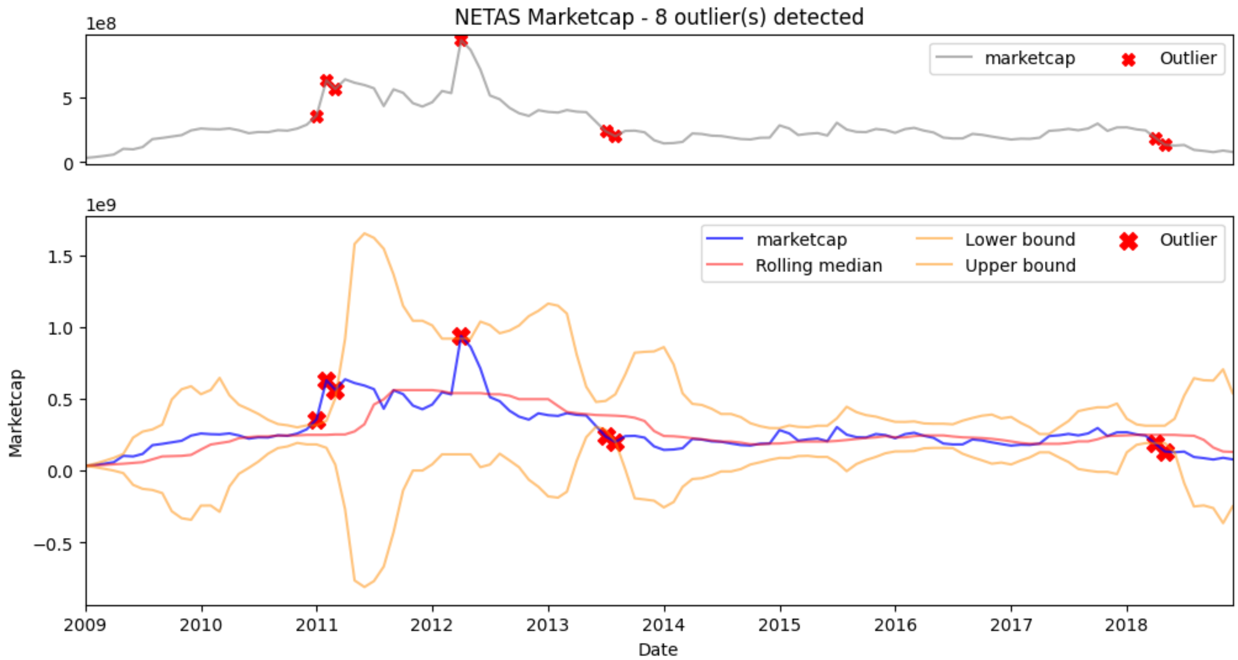

Fig. 5. Outlier detection sample for the telecommunications company NETAS

The macroeconomic indicators used in this study have very different scales. Therefore, the use of a scaling method was inevitable. It is acknowledged that the standard scaler, i.e., centering the data and dividing by its standard deviation, assumes that the data can be approximated by a normal distribution. However, this is not the case for this study as the data is time-series data, which can exhibit various distributions. Moreover, minmax scaler for our dataset has been proven to be inefficient because it tends to assigning a lot of 1 and 0 because of our feature distributions. Therefore, alternative scaling methods should be employed. After thorough research, it was determined that the power transformer with Yeo-Johnson method is a suitable choice for this study, as the only assumption it makes is that the data is continuous which is not a bad assumption for time-series data and it does not transform altogether the distribution of our features. Additionally, this method is appropriate for both positive and negative data, and it can also handle zero and missing values. Though minor differences and no difference with Random Forest, this scaling method improves MSE of some models like Gradient Boosting Regressor (for general model including all sectors and models per large three sectors MSE decreased on average by 0.18).

Additional features that incorporate the change in a certain indicator during the previous N months were added during the modeling phase to capture the trend of these indicators. Finally, the sector and company definitions were used as categorical variables to allow the model to differentiate between different companies and sectors. Without these categorical variables, the model would be predicting all company market values using the same input, which is the macroeconomic indicator data. These categorical variables were encoded either using one-hot encoding or numeric encoding depending on the model used and what suits it.

Finally, the target variable, i.e., market capitalization, was converted into a percentage change between two different months. For example, for the models predicting the market capitalization 4 months ahead, our training target would be the percentage change between four months ahead and the current time, where the current time corresponds to the date of the macroeconomic indicators used for prediction. It may be argued that this method will make the model treat a multibillion dollar company and a small company equally, rendering it unrealistic. Although it is true that the model will treat companies with different values equally in this case, the value of a company is of no interest for the investor. An investor is only interested in the changes in company stock prices as it is what determines their profit and loss.

3.3. Modeling

In order to model the data, each target value was treated as one data point. For instance, to predict the target on 07/2015, macroeconomic indicators from 03/2015 were used as predictors for the case of 4 months ahead predictions. Moreover, predictions for different time periods ahead were made. In addition, two approaches were compared, using one general model for all sectors and using individual models for each sector. The last two years of the data were left out as a test dataset.

For model evaluation, the mean absolute error (MAE) was chosen. MAE treats small and large errors equally unlike the root mean squared error (RMSE), which penalizes large errors more than small ones, which is undesirable when dealing with highly fluctuating data.

As previously stated, the objective of this study is to predict sectoral growth rates using macroeconomic indicators using real-world data. This problem inherently contains a significant amount of nonlinearities and complexities owing to the nature of financial data. Consequently, the problem cannot be approached using parametric models such as linear regression, as these models make strong assumptions, have a limited number of parameters, and are unable to represent all possible relationships in highly non-linear data. This is the rationale for adopting non-parametric models such as boosting models, tree-based models, and SVM models.

4. Results and discussion

4.1. Baseline model

Despite the notion that non-parametric models would be more suitable for the task at hand, a preliminary model was constructed utilizing linear regression as a baseline model, as it had been established as such in the progress report. Additionally, the standard scaler was employed as a baseline scaler. As anticipated, the scores obtained were not optimal. The model was trained using the initial 8 years of data, with the subsequent 2 years utilized for evaluating the model's performance. The results of the baseline model are shown in Table 1. As depicted, the mean absolute error for a 6-month lagging interval is 0.74, which is not considered optimal.

Additionally, it is acknowledged that only examining model evaluation metrics is inadequate for assessing the performance of the model. The objective of the project is to furnish investors with valuable insights and information via forecasting sectoral growth utilizing macroeconomic indicators, and it is acknowledged that correctly predicting general trends is of greater importance than precisely predicting individual points in financial predictions. Therefore, the model results should be supported with informative visualizations. To this end, visualizations of the predicted and actual values were performed for each model in order to evaluate the model's ability to capture general trends.

(a) (b)

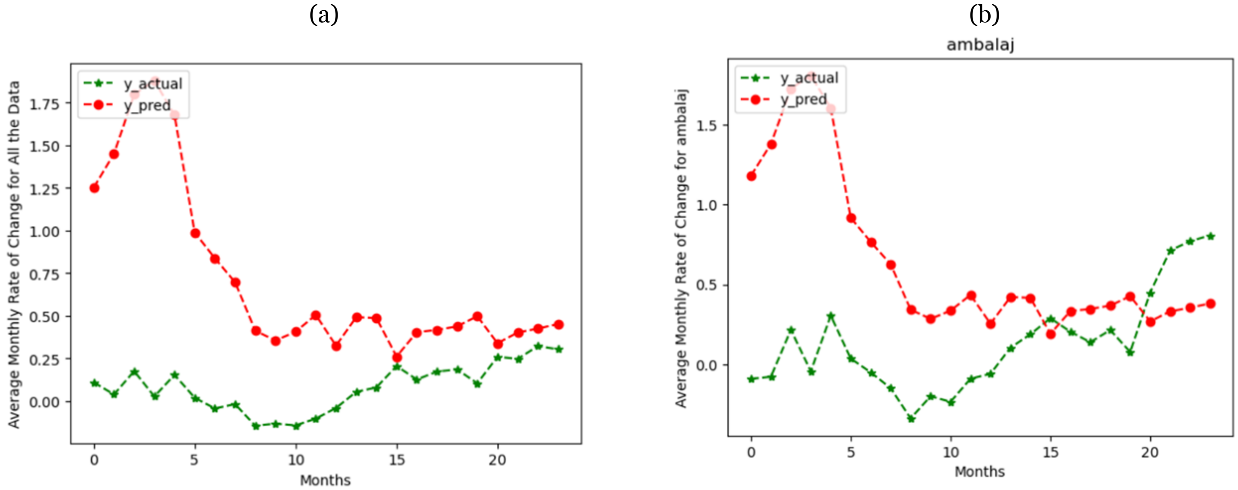

Fig. 6. Baseline model predictions (a) Average prediction over all data (b) Predictions for the “Ambalaj” sector

Fig. 6a is generated by averaging all the data for each month in the test data and comparing it to the mean of all the data for each month in the predicted y values. In this model, a lagging time of 6 months was utilized. In other words, the model tries to forecast sectoral growth rates for 6 months in the future. As evidenced in the figure, the Linear regression model is performing poorly on the test data. Furthermore, when visualizing the predicted versus actual values for different sectors, the results remain largely unchanged (Fig. 6b).

4.2. General and Sectoral Models

Our experiments showed that the performance of one general model was comparable to that of individual models for each sector (Fig. 7). This may be explained by the fact that the model is able to differentiate between different sectors using the sector categorical feature. Furthermore, the general model may be advantageous in the cases of sectors with a limited number of companies, as some of its learnings from other sectors may be generalizable to that sector.

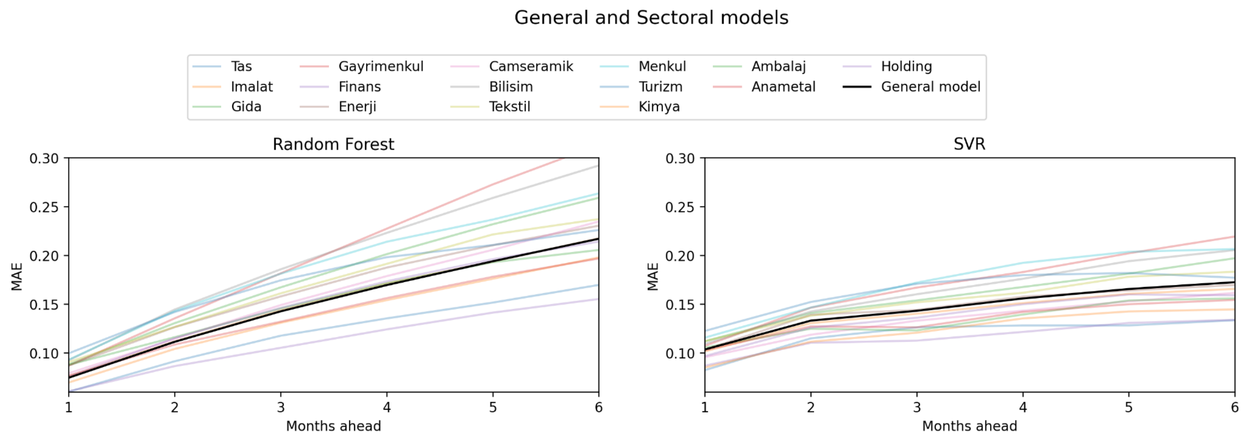

Fig. 7. 5-fold cross validation MAE average of sectoral models compared to the general model for 1 to 6 months ahead predictions using SVR and Random Forest models

4.3. Model Results

After experimenting with various models such as SVR, Random Forest, and XGBoost, it was determined that XGBoost is the optimal model for the task at hand as it effectively handles non-linear data. This outcome is not surprising as XGBoost is well-known for its capability to uncover non-linear relationships, which is essential in this case. For the sake of brevity, the details of these experiments will not be discussed in this report.

Model	No. months ahead											
	1	2	3	4	5	6	7	8	9	10	11	12
Base	0.100	0.206	0.224	0.259	0.529	0.742	0.840	0.960	0.826	0.729	0.908	1.119
XGB1	0.116	0.200	0.231	0.253	0.308	0.401	0.442	0.501	0.600	0.620	0.639	0.718
XGB2	0.103	0.147	0.186	0.235	0.294	0.388	0.354	0.367	0.452	0.447	0.485	0.595

Table 1. MAE values for different months ahead

Subsequently, a model was constructed using an XGBoost regressor and the power transformer with Yeo-Johnson method. This model is referred to as XGB 1. As observed from the results in Table 1, the MAE for a 6-month lagging interval has been improved from 0.74 to 0.40 for this model. Furthermore, a visualization of the predicted versus actual values for test data can be seen in Fig 8.

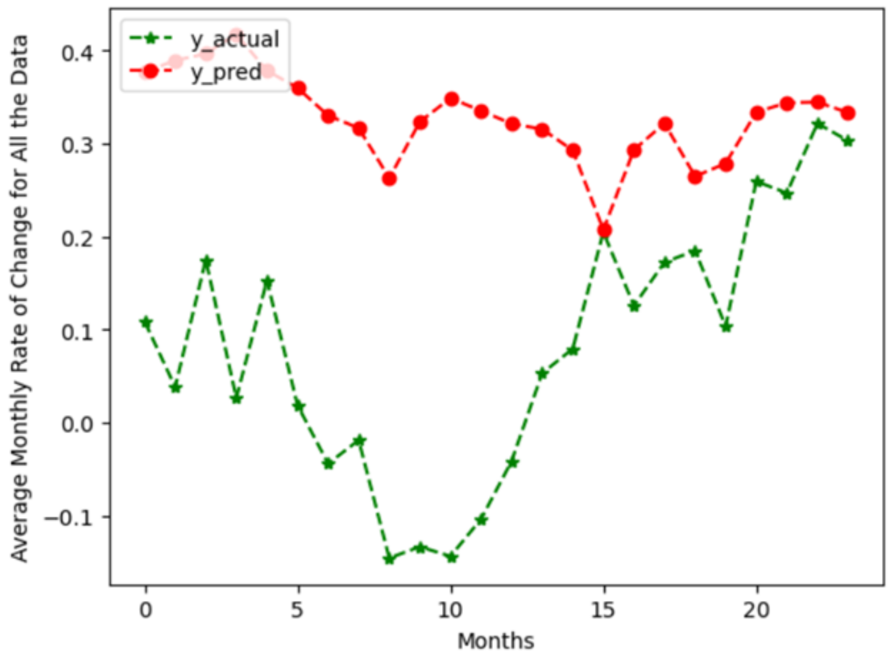

Fig. 8. Average prediction over all data using XGB 1 model

Although the model has improved in capturing general trends, it appears that there is still room for further improvement, particularly for the last 15 months of the test set. This is because the figure above is constructed by starting from the last month and progressing to the first month in the test set. Consequently, the model performed well in predicting the first 10 months of the test set, but struggled to accurately predict the remaining 14 months.

In light of this, new features were derived from the macroeconomic indicators in the feature set to enhance performance. A variety of different features were experimented with, such as the rate of change of every feature over 3, 6, 9, and 12-month periods to obtain the trends of indicators, which are crucial metrics that reflect a country's economic situation. Additionally, some features were transformed using one-hot-encoding. For example, the Brent oil price is one of the features, and it can take values between 0 and 150. Thresholds were established to generate several 0-1 features. For this indicator, 110, 80, and 60 were set as thresholds by examining the distribution of the features and determining critical values. As a result, four new features were generated, one of which is 1 if the value of the feature is between the determined thresholds. This process was repeated for some features. However, the output obtained for the model with all the derived features did not yield better results. Therefore, a general approach for feature derivation was investigated.

After conducting experiments with the data, it was determined that the model with the rate of change of 6 months for every feature without original features yielded the best result among all the models. This model is referred to as XGB 2. In terms of consistency, a 6-month lagging time for the target value was used for this model. Additionally, since all the features represent the rate of percentage change over 6 months, there is no need to scale the data. Furthermore, for this model, the test set was separated as the last 3 years of data for reasons explained below. As seen in Table 1, the performance of the model was improved. Also, a visualization of the predicted versus actual values for test data shows a significant improvement in the capability of the model to capture the trends of market capitalization (Fig. 9a).

(a) (b)

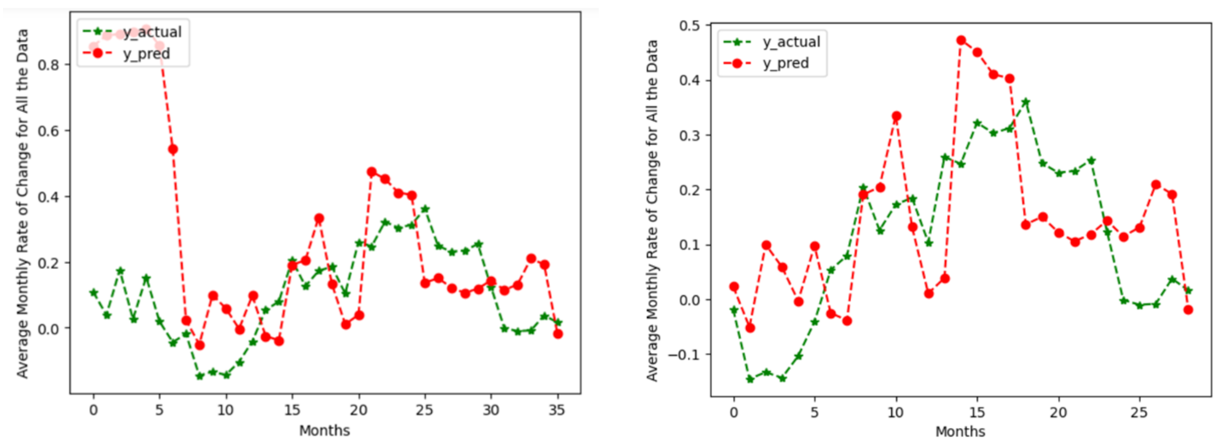

Fig. 9. Average prediction over all data using XGB 2 model (a) Before trimming the test dataset (b) After trimming the test dataset

As seen from the figure above, the model is quite effective in capturing general trends. However, there is a shift in the last 7 months of the test set. The reason for this shift is that during the last 7 months of 2018, the Turkish Lira experienced a significant depreciation against other currencies, which was not anticipated. This event had a significant impact on the growth rates of sectors in Turkey. Initially, we decided to use data from 2009 to 2019 to eliminate the disturbances caused by the 2008 financial crisis and the 2020 coronavirus pandemic, due to their unstable nature. Therefore, for the sake of simplicity, the late 2018 exchange rate crisis can be excluded from the test set, resulting in the test set ending at the sixth month of 2018. The results reported for the XGB 2 model are based on this trimmed dataset. The increase in the test dataset size for this model as mentioned earlier was to allow for this trimming without affecting the ratio of the test dataset. Fig. 9b shows a visualization of this model's predictions after trimming the last six months of 2018. Fig 10. shows the predictions of XGB 2 for the "Finans" and "Ambalaj" sectors.

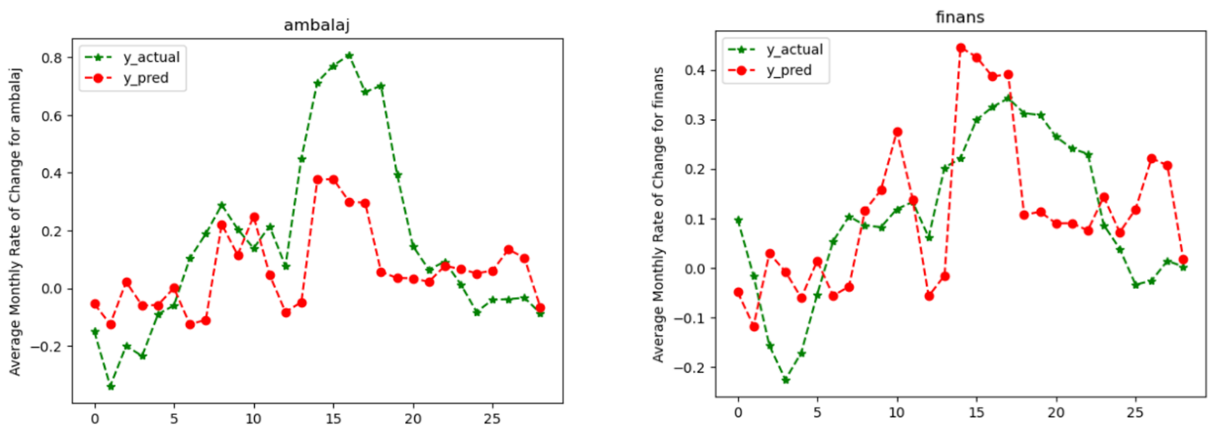

Fig. 10. Average prediction over “Ambalaj” and “Finans” sectors using XGB 2 model

5. Conclusion

In conclusion, we were able to significantly improve the baseline model's performance, reducing the mean absolute error from 0.70 to 0.28, and even higher for some sectors. Despite the challenges presented by data collection and correction, and the non-linearities and disturbances in the real-world data, we were able to produce meaningful results, particularly in terms of predicting general trends of growth rates. We discovered that XGBoost is a particularly effective model for exploring relationships among non-linear data and that the power transformer with Yeo-Johnson method performs well on this type of financial and continuous data.

In the future, we plan to further investigate different combinations of feature trends, as we recognize that every feature has an optimal lagging time for trends. By utilizing domain knowledge and technical approaches, these optimal times can be identified. After identifying these optimal lagging trend times for each feature, we intend to convert these features using one-hot encoding by determining optimal thresholds. We believe that generating these features and converting all of the features into 0-1 features by setting optimal thresholds can significantly improve the model's performance. This is because it is possible to establish very critical thresholds by optimally segmenting the feature values, yielding crucial information about certain sectoral growth rates, which would not be captured otherwise. For example, it is possible that when the rate of change for unemployment over 6 months exceeds 15%, certain sectoral growth rates are greatly affected. This work requires a thorough understanding of macro-indicators and a careful analysis of the available data, but it has the potential to significantly improve the results obtained.

Furthermore, we can also experiment with using LSTM models that can learn the optimal lagging times on their own, instead of having them defined by domain experts. Although automatically generated features using deep learning models generally outperform manually engineered features, these models lack interpretability, which makes them hard to use in a critical sector like finance. However, a study comparing both manual and automatic feature generation will still produce interesting results in this field.

Literature Research

1. Büyük, Ahmet, and Hasan Abd. “International Journal of Economic Perspectives, 2010, Volume 4, Issue 4, 601-610.”
2. Ça, Efe. 2010. “Testing Long-Run Relationship between Stock Market and Macroeconomic Variables in the Presence of Structural Breaks: The Turkish Case.” (48).
3. Erdem, Cumhur, Cem Kaan Arslan, and Meziyet Sema Erdem. 2005. “Effects of Macroeconomic Variables on Istanbul Stock Exchange Indexes.” *Applied Financial Economics* 15(14): 987–94.
4. Hasan, Nawzad et al. 2019. “Macroeconomic Determinates Of Stock Price for Industrial Companies Listed In Istanbul Stock Exchange.” 23(02).
5. Kandir, Serkan Yilmaz. 2008. “Macroeconomic Variables, Firm Characteristics and Stock Returns: Evidence from Turkey.” (16).
6. Kumbure, Mahinda Mailagaha, Christoph Lohrmann, Pasi Luukka, and Jari Porras. 2022. “Machine Learning Techniques and Data for Stock Market Forecasting: A Literature Review.” *Expert Systems with Applications* 197: 116659.
7. Soni, Payal, Yogya Tewari, and Deepa Krishnan. 2022. “Machine Learning Approaches in Stock Price Prediction: A Systematic Review.” *Journal of Physics: Conference Series* 2161(1): 012065.
8. Vijh, Mehar, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. 2020. “Stock Closing Price Prediction Using Machine Learning Techniques.” *Procedia Computer Science* 167: 599–606.
9. Weng, Bin et al. 2018. “Macroeconomic Indicators Alone Can Predict the Monthly Closing Price of Major U.S. Indices: Insights from Artificial Intelligence, Time-Series Analysis and Hybrid Models.” *Applied Soft Computing* 71: 685–97.

Appendix

Halil: Conducting extensive literature research on various methods and results, collecting the data sources into tabular format, and explanatory data analysis. Building general model and model per large three sectors using RF and Gradient Boosting Regressor with different scaling methods.

Yasser: Cleaning and merging the datasets, exploratory data analysis, and conducting general and sectoral model comparisons using SVR, RF, and Linear regression.

Ahmet: Acting as a domain expert, collecting the data sources, exploring feature importances, and building the primitive model as well as the XGBoost models. Focusing on trend analysis with feature engineering.