IE 582 - STATISTICAL LEARNING FOR DATA MINING

Project - Group 8 (Murat Uğur Akkalkan - Sergen Tuğ Toraman)

Introduction

In the ever-evolving landscape of finance, leveraging big data for predictive analytics has become a pivotal skill. This project embarks on a challenging journey into the stock prices, focusing on Borsa İstanbul, where the task is forecasting the next day's hourly average prices for 30 companies.

The result will span from 9 AM to 6 PM, required to provide 10 predictions per day for each of the 30 companies, thereby presenting a comprehensive outlook for the unfolding market trends.

To reach the forecast result, following steps are applied:
* Data Gathering: We use yahoo finance (yfinance) library on python.

* Data Preprocessing: We process stock names to get data from library and exclude null data points.

* Feature Engineering: We create meaningful features from the data for enhancing the predictive power of models.

- RSI, or Relative Strength Index, is a momentum oscillator that measures the speed and change of price movements. It is a widely used technical indicator in financial markets, particularly in stock trading. The RSI is calculated based on the average gain and average loss over a specified period, typically 14 periods. The formula for calculating RSI is as follows:

$$RSI = 100 - \frac{100}{1+\frac{Average\ Gain}{Average\ Loss}}$$

  The RSI ranges from 0 to 100 and is usually displayed as an oscillator. Traditionally, an RSI reading above 70 is considered overbought, indicating that the asset may be due for a price correction. Conversely, an RSI reading below 30 is considered oversold, suggesting that the asset may be undervalued and due for a potential upward correction.

- Moving Average (MA) is a statistical calculation used to analyze data points by creating a series of averages of different subsets of the full data set. The Moving Average is calculated by adding up a set of values and dividing the sum by the total number of values. It provides a smoothed line that represents an average value over a specified time period.

$$MA = \frac{Sum\ of\ prices\ for\ specific\ periods}{Number\ of\ periods}$$

These two indicators are the most commonly used parameters at stock market.

Also, we define percentage change from one hour to another.

To sum up, our feature space is as follows:

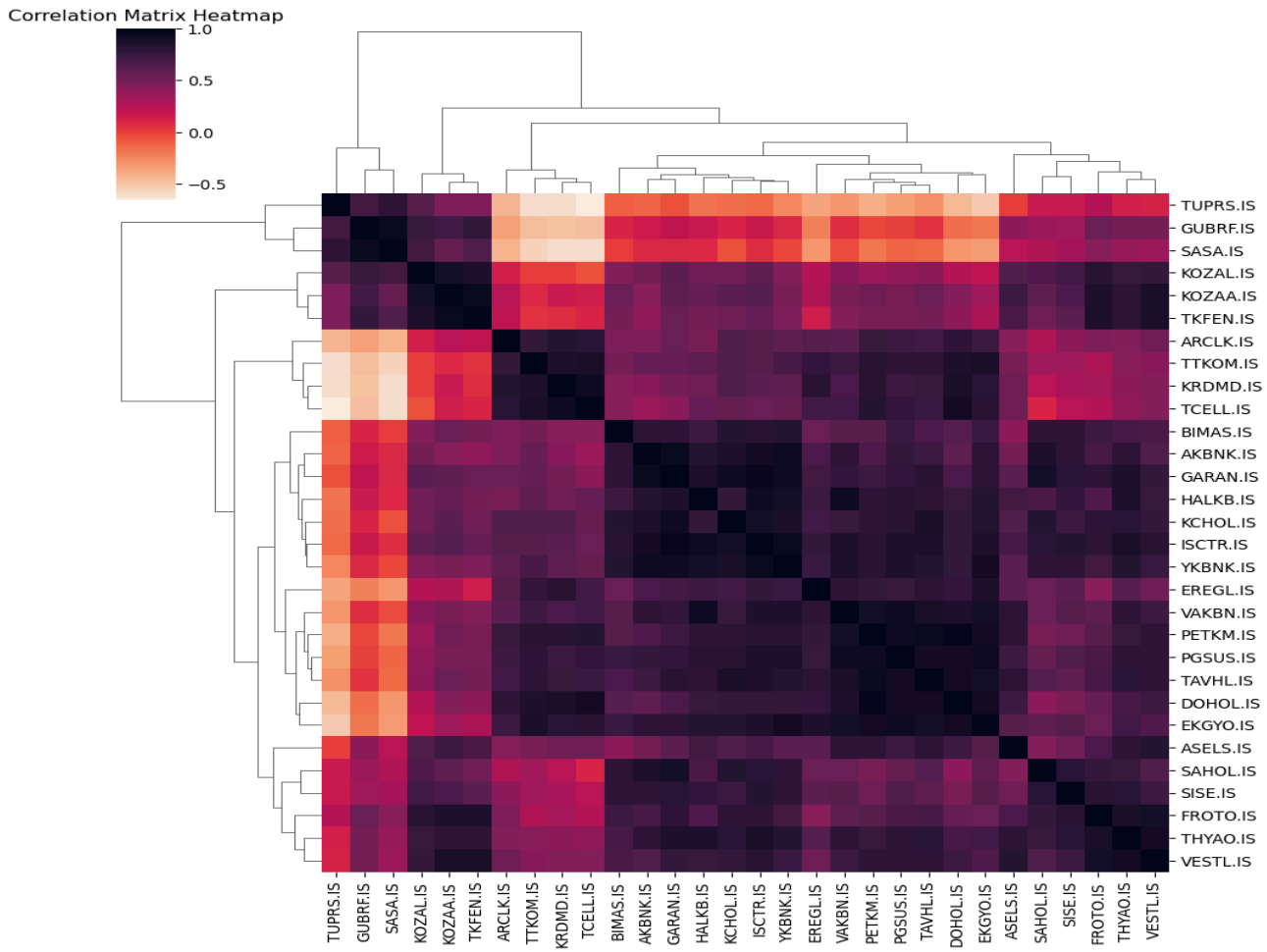
- Adj Close: Adjusted closing Price
- Volume: Volume of stock
- 10h_close_pct: last ten hours closing change percentage
- 10h_future_pct: next ten hours closing change percentage (target value)
- MA_50: Moving Average of last 50 hours
- RSI_50: RSI of last 50 hours

10h_future_pct is the target value others are the features. We planned to predict 10 hour of price changes by using the feature space.

Descriptive Analytics

In this part, some descriptive statistics are found.

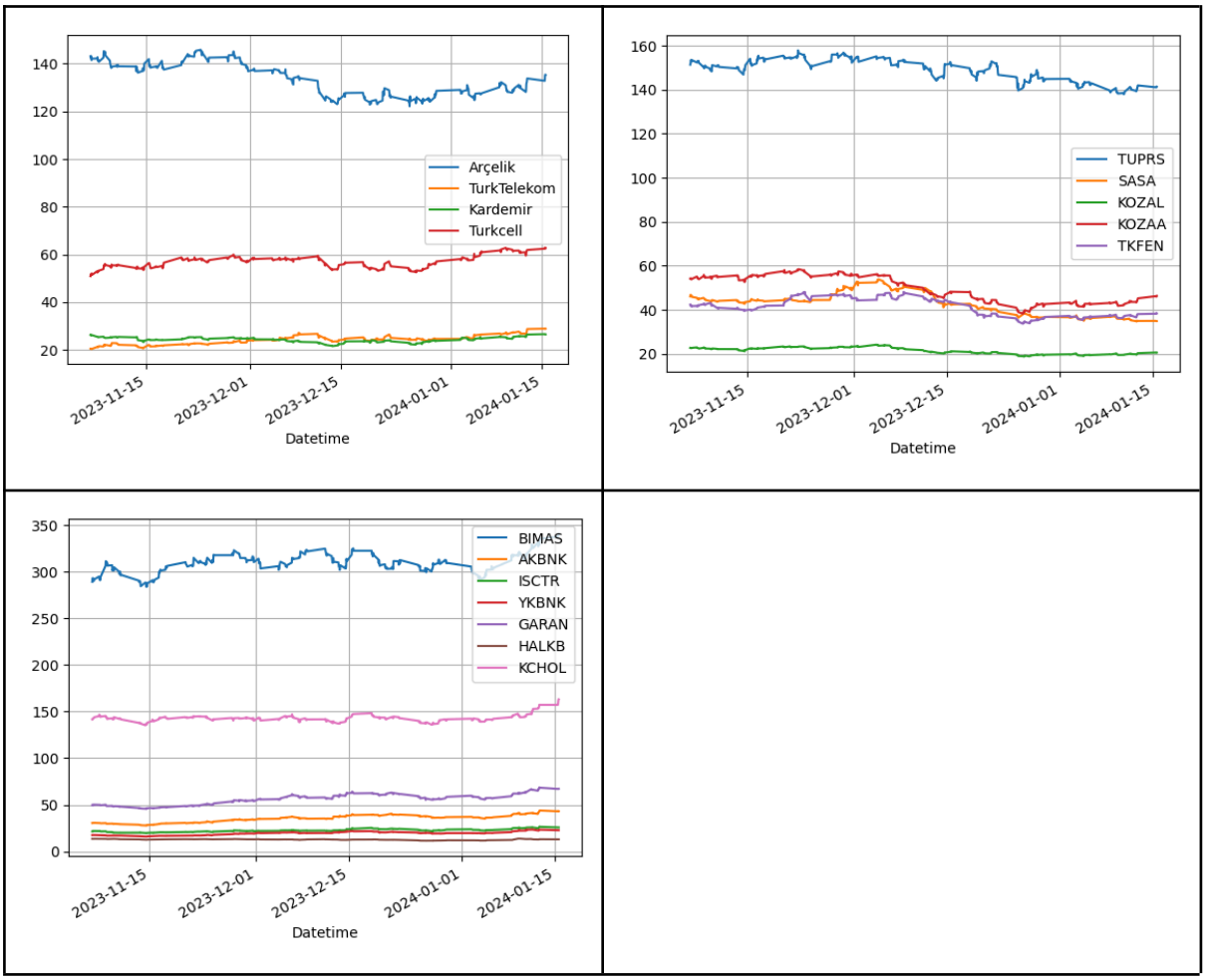
Correlation of stock prices: As seen in the following heatmap figure; some stocks trends are positively correlated.



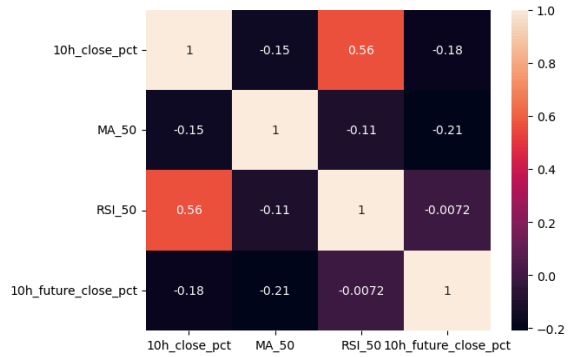
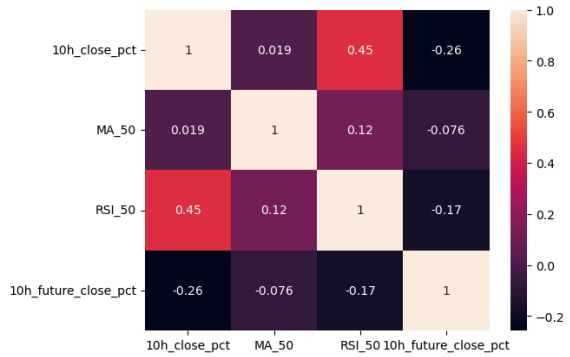
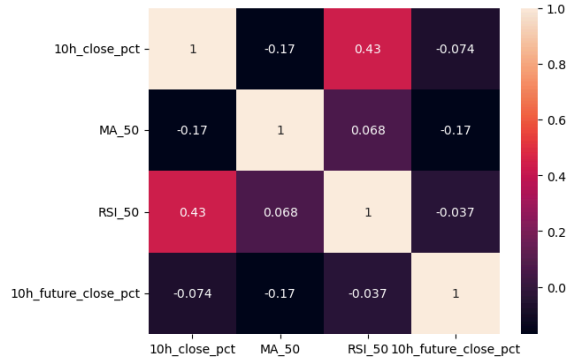
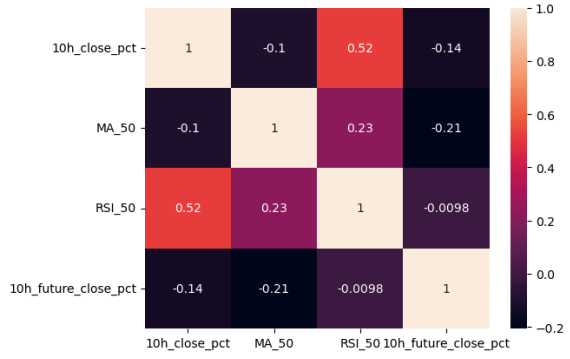
Here are the some positive correlated stocks.

- TUPRS, SASA, KOZAL, KOZALA, TEKFN
- ARCLK, TTKOM, KRDMD, TCELL
- BIMAS, AKBNK, GARAN, HALKB, KCHOL, ISCTR, YKBNK

Line-trend graph also shows how these stock are correlated can be seen at “Adjusted Prices” of them.



Also, how feature space are correlated to each other can be seen at the following graphs. Arçelik, Kardemir, Türk Telekom ve Turkcell are selected for the graphs.

ARCLK	TTKOM
	
KRDMD	TCELL
	

Approach

* Train Data Selection: Train data is selected as 20% of the all data.

* Modelling and Model Selection: We apply some algorithms to train our models:

- DecisionTreeRegressor
- GradientBoostingRegressor

We tune their parameters checking the performance metrics on separate validation data and the test data provided. MSE stands for Mean Squared Error is used as a performance metric for the model.

Also, 5-fold cross validation is applied both for test and train data.

Cross validation scores for DecisionTreeRegressor are as follows. They are very close to 0, that means model fits better.

Cross-Validation Scores on Training Set: [0.0001538 0.00019929 0.00013325 0.00015929 0.00017762]

Cross-Validation Scores on Test Set: [0.00022068 0.00019783 0.00014185 0.00013794 0.00017082]

MSE on Train Set: 0.00012739676916724527

MSE on Test Set: 0.0001103333711930507

Besides, Cross validation scores for GradientBoostingRegressor are as follows. They are very close to 0, and it is better than DecisionTreeRegressor.

Cross-Validation Scores on Training Set: [1.087e-04 1.331e-04 1.027e-04 8.7173e-05 1.167e-04]

Cross-Validation Scores on Test Set: [2.773e-04 1.100e-04 1.103e-04 4.223e-05 1.173e-04]

MSE on Train Set: 9.25e-05

MSE on Test Set: 9.081e-05

Conclusion and Future Work

In the course of this project, we delved into the challenging realm of financial forecasting for Borsa İstanbul, employing advanced statistical learning and data mining techniques. The task was to predict the next day's hourly average prices for 30 companies, spanning the critical hours from 9 AM to 6 PM.

We also use “Decision Tree” & “Gradient Boosting” and “cross validation” algorithms to predict next day stok prices.

Future work should focus on refining models, exploring additional features, and embracing emerging technologies to stay at the forefront of data-driven financial analysis. Firms financials, balance_sheet and cashflow can also be used in analysis and prediction.

Besides, portfolio analysis can be applied in order to maximize the profit. That is to say; algorithms can learn which stocks to invest and can maximize the profit.