JPX Tokyo Stock Exchange Prediction

Sriya Gorrepati
Yingying Deng

TeamID: B6

Overview

Motivation:

In financial market,for investors,how to buy stocks, when to buy and when to sell them, is very important, especially important for retail investors. because these retail investors can’t obtain historical and real-time data.

- This competition will provide financial data for the Japanese market, allowing retail investors to analyze the market to the fullest extent.
- The goal is building portfolios from the predicted stocks (around 2,000 stocks).

Dataset

- The dataset contains of multiple files such as supplemental files, train files which includes stock_prices, secondary_stock_prices,stock_list, finances, trades etc.
- We focused on core data files(train file and supplemental file). stock_prices which contain 2332531 rows and 8 columns and stock list which contain 4417 rows and 2 columns.
- The stock prices are calculated day wise.
- Stock information provided is for each company.
- Terms to know: Target- Average return, Volume- no.of shares traded.

Preprocessing

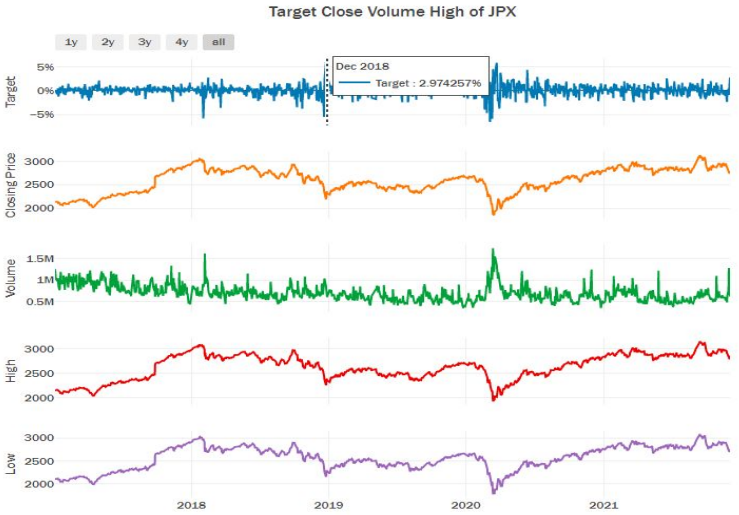
- Huge dataset.
- Divided the project into 2 different approaches by using different datasets and training and testing on multiple data.

Approach:1

- Merge stock_prices and secondary_stock_prices in train file as train data.
- Merge stock_prices and secondary_stock_prices in supplemental file as test data.
- Data normalization: use Z-scores to scale the values
- **In investing and trading, Z-scores are measures of an instrument's variability and can be used by traders to help determine volatility.**
- If some values of columns are missing and NAN , we will set 0 to them.

Exploratory Data Analysis

- The graph explains about the target, close, high, volume stocks of the dataset throughout the 4 years from 2017 Jan- Nov 2021.
- We have divided it into ranges to get a better understanding.
- When we hover over the particular graph, it gives us the range and year of that attribute.
- Can select the no.of years from the above



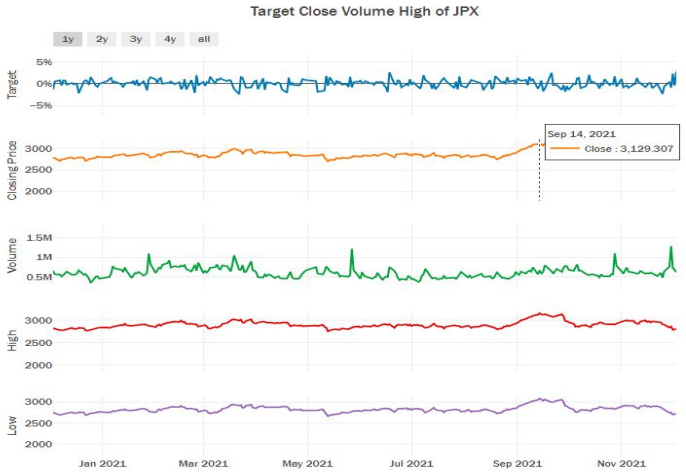
Preprocessing

Approach.2:

- Dropped the rows or columns containing the missing values.
- The target variables we considered are the ‘close price.’
- Merged two datafiles Stock_prices from train data file and Stock list into stocks.
- After dropping and merging there are, 2324923 rows and 12 columns.
- Trained and tested on the stocks dataframe.
- The train-test ratio is 70:30.
- Data normalization is done using min-max scalar.

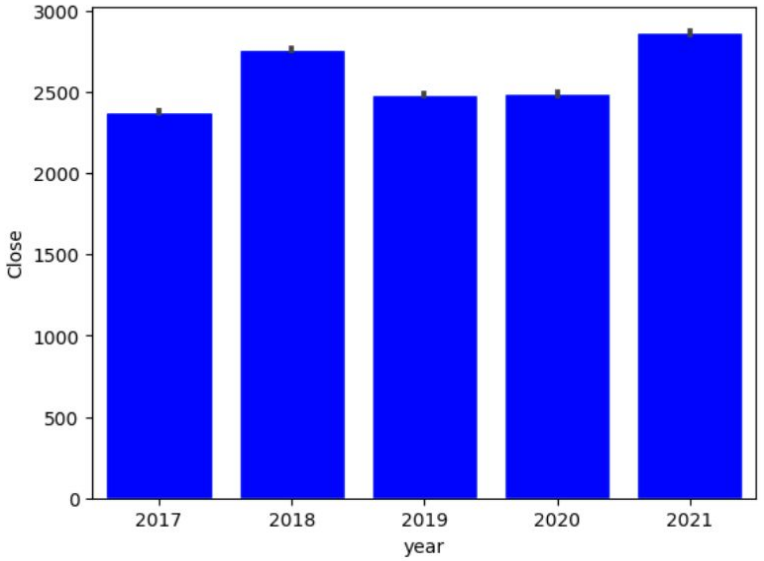
Exploratory Data Analysis

- This graph gives us the statistics of the stocks from the recent year 2021.



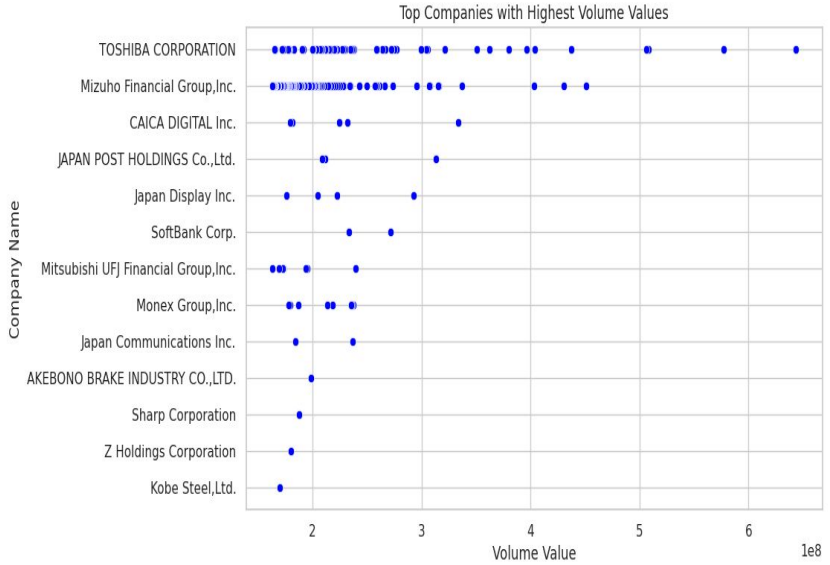
Closing Price vs Year

- Here, the average closing price from all the years is calculated.
- From the following graph we can observe that 2021 is having the highest average closing price.



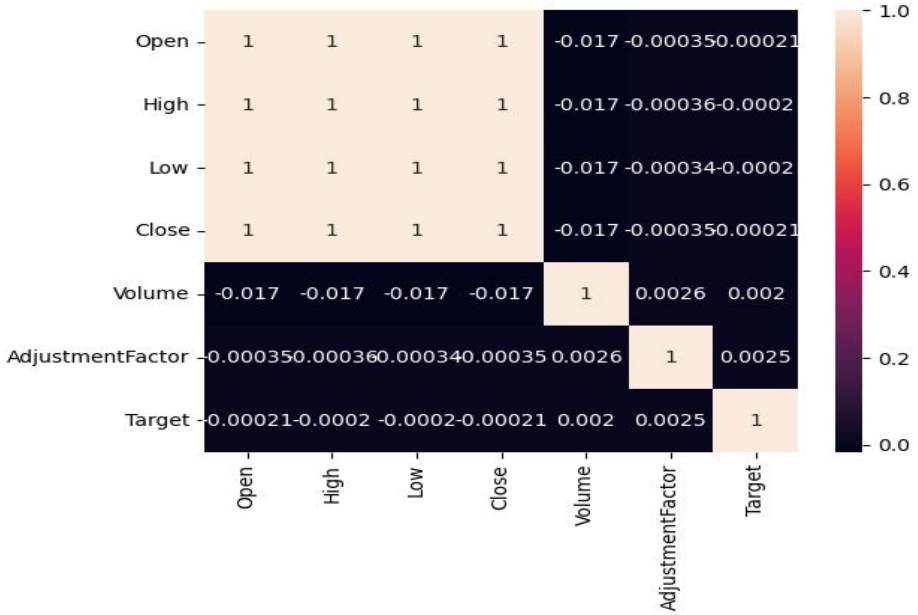
Companies with highest volume

- Average no.of shares traded by the companies.
- Took the average of 20 volume samples and calculated.
- Toshiba has the highest stock trading rate.



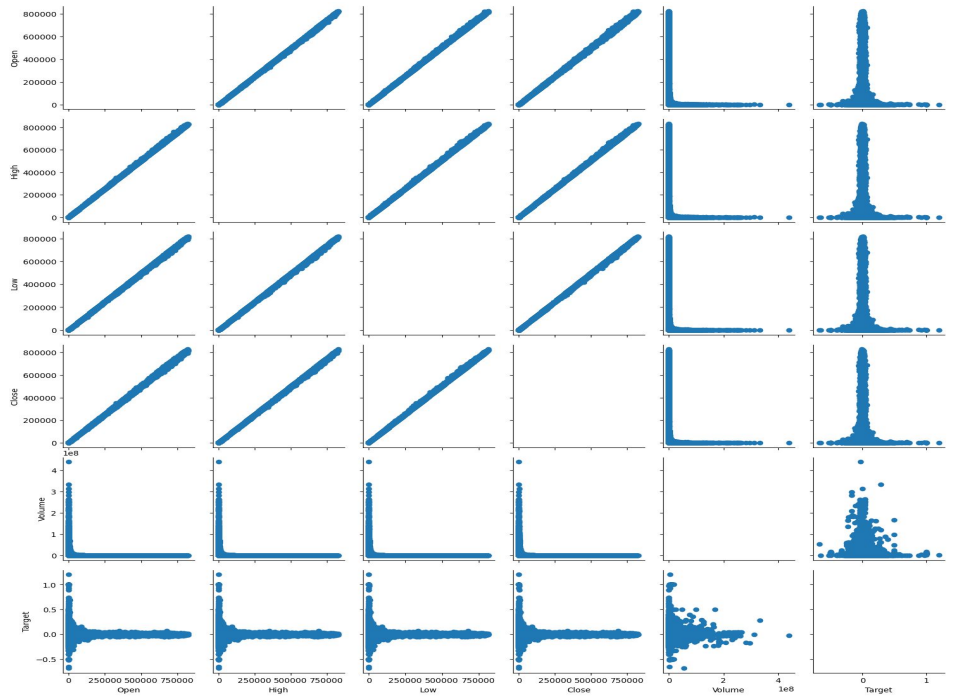
Data Correlation

- two variables analysis to pick up features as input data (headmap the correlation coefficient and use PairGrid to plot the scatter graphs)



Scatter graphs

1. Between Open,High,Low,Close prices , there is a strong correlation.
2. The trading volume decreases rapidly as the price rises. Generally, stocks with high prices have low trading volume.
3. Target presents a normal distribution.
4. Choose open,high,low, close and volume as input features, Target as output



Competition Evaluation

Sharpe Ratio of the daily spread returns

The competitors need to rank each stock active on a given day and calculate the total returns for the portfolio.

Assuming 1: the 200 highest (e.g. 0 to 199) ranked stocks as purchased and the lowest (e.g. 1999 to 1800) ranked 200 stocks as shorted.
Assuming 2: the stocks are weighted based on their ranks
Assuming 3: the stocks were purchased the next day and sold the day after that.

Target	Close_shift1	Close_shift2	rate
0.000730	2738.0	2740.0	0.000730
0.002920	2740.0	2748.0	0.002920
-0.001092	2748.0	2745.0	-0.001092
-0.005100	2745.0	2731.0	-0.005100
-0.003295	2731.0	2722.0	-0.003295

The Target and the rate calculated match. (so I will use the Target that is calculated to rank the stocks.

calculate competition scores

1. Rank the stocks (use target column to rank)

2.Weighted the top 200 stocks (S_{up})

3. Weighted the bottom 200 stocks(S_{down})

4.daily spread return (R_{day}) is the result of subtracting S_{down} from S_{up}

5.

$$Score = \frac{Average(R_{day_1-day_x})}{STD(R_{day_1-day_x})}$$

Approach.1

Models

1. Linear Regression
2. LSTM(Long Short Term Memory)
3. LightGBM(light gradient-boosting machine)

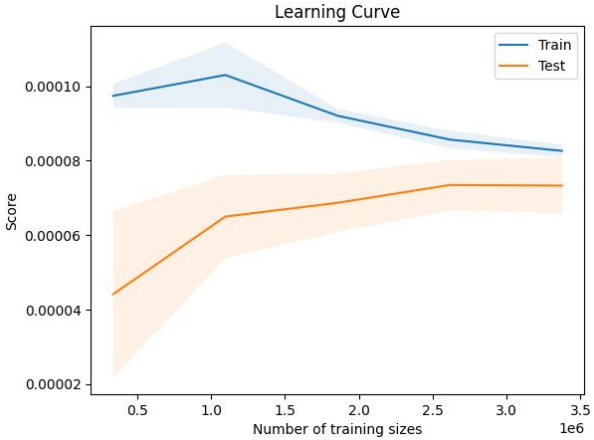
Results

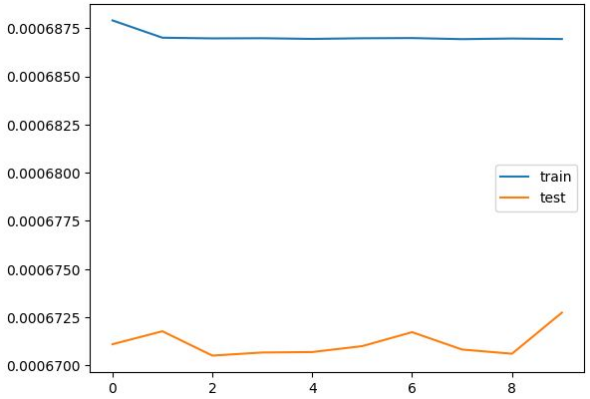
1.evaluation metric: rmse(root mean square error)

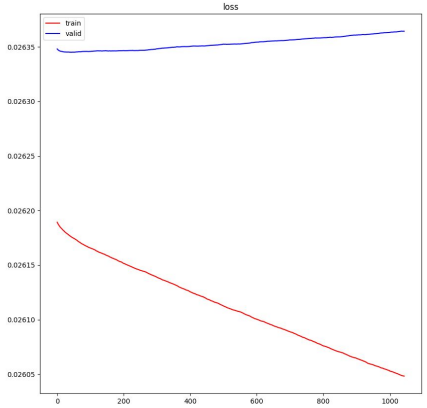
2. Competition score

Model	RMSE	Competition Score
Linear Regression	0.026096574110249224	0.08333
LSTM	0.025937180177782314	0.27103
LGBM	0.025912137193322568	0.13008

Model performance: loss function







model 1 : linear regression

model 2 : LSTM

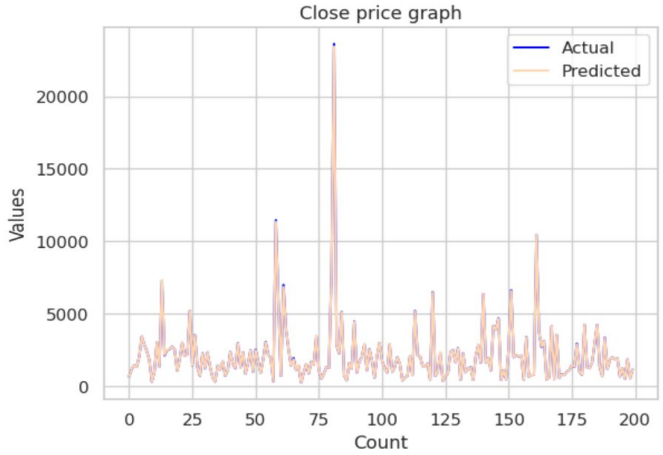
model 3: LGBM

Approach 2

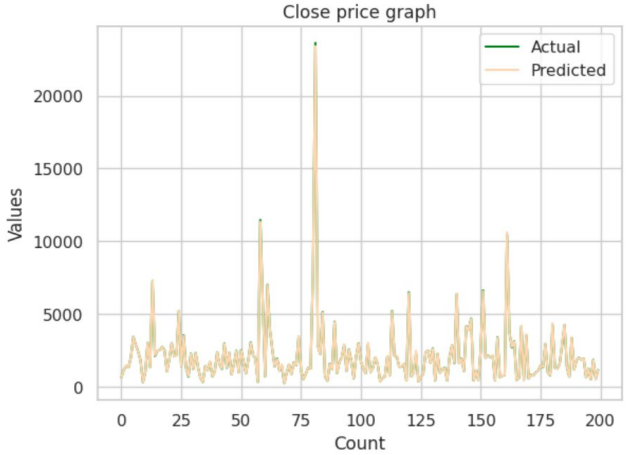
Models

- Linear Regression
- LSTM
- Lasso Regression
- Adaboost

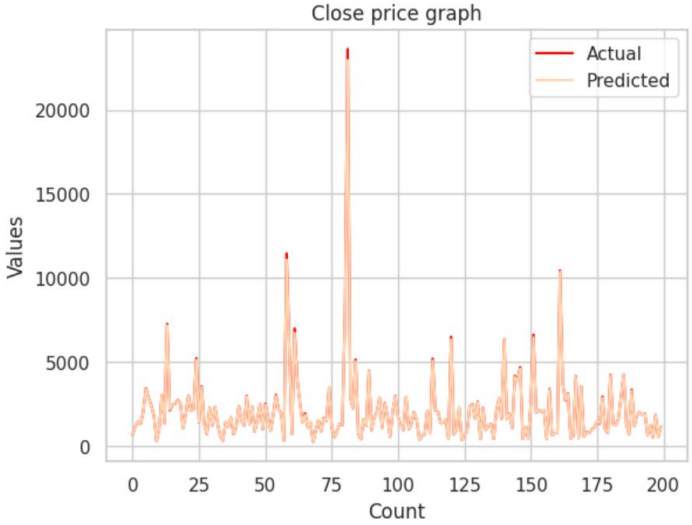
Actual Vs Predict values



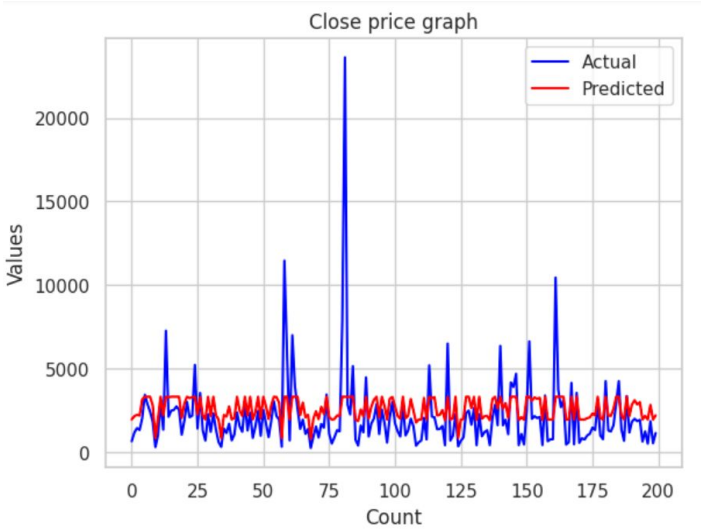
Linear Regression for 200 samples



Lasso Regression



Adaboost



LSTM

Results

Model	MSE	R-Square
Linear Regression	1239.16	0.99
Lasso Regression	2692.27	0.99
Adaboost	4784.55	0.99
LSTM	11233.34	0.13

Conclusion

- The LSTM works better for approach 1 where as in approach 2 Linear regression works better.
- The performance could change with the train-test data.
- Approach 1 gives the better accuracy scores.
- In approach 2 both the predicted and tested values align with each other.

Challenges and Future work

Challenges:

- Due to the huge dataset, preprocessing has been a challenge.
- Understanding the concept and stock market terms.
- The runtime is so long and is hard to execute quickly.

Future work:

- We want to explore the dataset more and obtain more useful information.
- Try few more regression models.
- Targeting on a single sector.

Thank you

Questions?