DS 5220: Supervised Machine Learning and Learning Theory I
Fall 2019

Project Milestone

Project Title:

Short-term stock market price prediction

Team Number: 4

TA: Ewen

Team Members:

Farhanur Rahim Ansari (001376195)

Vidhey Oza (001059237)

Problem Description

What is the machine learning problem you are trying to solve?
Predict prices of stocks listed on NYSE and NASDAQ on a short-term basis using technical indicators.

Is it a classification or regression problem?
Since we are predicting closing stock prices, which are continuous values, this is a regression problem.

Why is it important?
Predicting stock prices is of high importance to stock-brokers as well as individual traders. On top of that, when intraday prices are predicted, the traders have an almost real-time analysis of the stock based on the trades done on the day, which is a great value addition.
Technical analysis of stock prices has been done for long, and technical indicators have been used to predict the direction in which the price will go. If a machine learns how these indicators "indicate" the change in prices, the analysis that has been done manually till now can be automated.

Related Work
Technical Analysis: these are some common blogs and websites that teach the importance of technical analysis for stock trading.

- StockCharts website: https://school.stockcharts.com/doku.php?id=technical_indicators
- Investopedia: https://www.investopedia.com/

Research Papers: there are many research papers that explore stock price prediction using machine learning techniques, but there are few that work specifically on intraday price prediction.

- Patel, Jigar et al. "Predicting stock market index using fusion of machine learning techniques." Expert Syst. Appl. 42 (2015): 2162-2172.
- Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A.,Menon, V.K., Soman, K.P.: Stock price prediction usinglstm, rnn and cnn-sliding window model. 2017 Interna-tional Conference on Advances in Computing, Communi-cations and Informatics (ICACCI) pp. 1643–1647 (2017)
- Geva, Tomer and Zahavi, Jacob, "PREDICTING INTRADAY STOCK RETURNS BY INTEGRATING MARKET DATA AND FINANCIAL NEWS REPORTS" (2010). MCIS 2010 Proceedings. 39.

Dataset

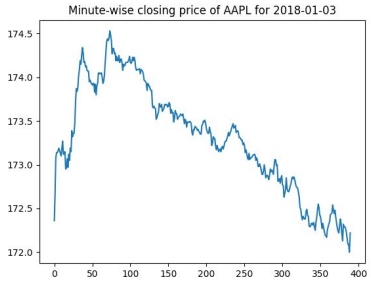
Statistics:
Number of records:
Minute-wise records from 1-1-2018 to 11-11-2019. Total 181578 records.

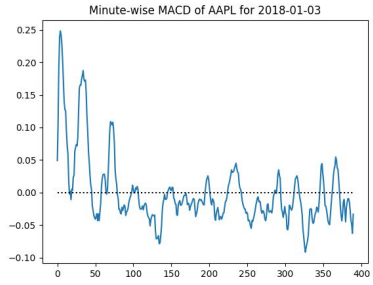
Number of features:
5 main features - open, high, low, close, volume.
22 engineered features (technical indicators) - MACD, RSI, WR, MFI, ROC, SMA, WMA, EMA, HMA, CCI, ADL, CMF, OBV, EMV, ATR, Mass Index, Bollinger L-Band, Bollinger H-Band, Ichimoku A, Ichimoku B, Aroon Index, ADX

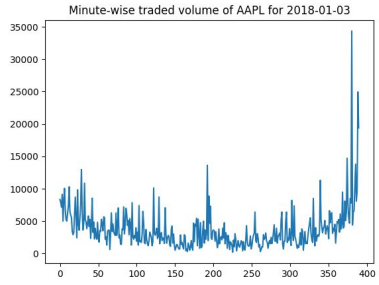
Feature Description:

- Bollinger L-Band and H-band are categorical features, while the rest are continuous features.
- Since indicators are usually formed from moving averages, the first and the last few values from the entire dataset are not computed (represented as NaN in dataset). These rows are removed before training.
- Usually data is only available for the 5 main features. Hence, we derived these features by using their defining formulae.
- Technical indicators are of many types. Some of them have fixed values which when crossed signal a major change in stock price movement (like RSI). Others are simply developed to gauge the volatility of the price or the volume of stock and hence the value exchanged during transactions (like Mass Index or Bollinger Bands).
- Some of these indicators (like MACD) are derived from differences of multiple indicators like moving averages. This is used to balance the signals generated by these indicators.

Sample graphs of features:







Approach & Methodology

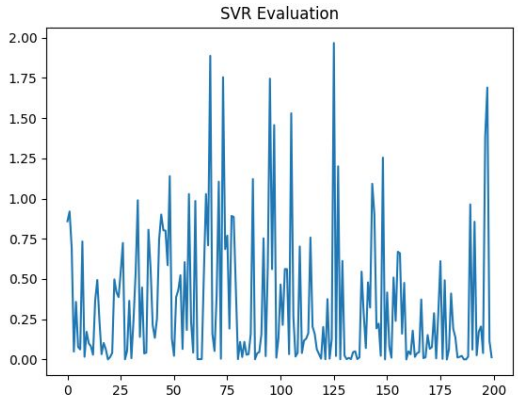
General approach:

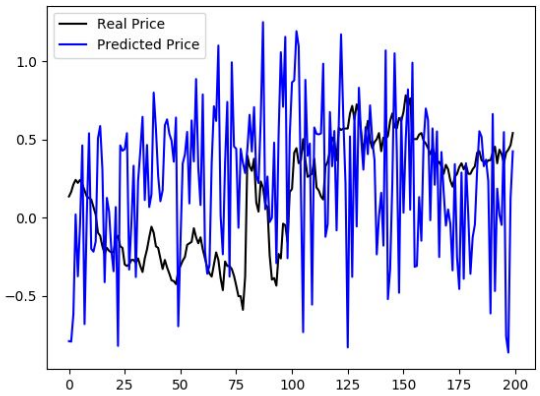
1. We extract stock price data (5 main features) from the internet. There are many free-to-use sources available including Python APIs.
2. We then performed data cleaning, since there were many unwanted features like ticker name, date, timestamp etc.
3. We then perform feature engineering and create 22 new features to have a total of 27 features in the dataset.
4. This data is then normalized using Z-score normalization.
5. The first and the last few values from the entire dataset are not computed (represented as NaN in dataset) based on how these indicators are calculated. These rows are removed before training.
6. Using this dataset, we train different models and compare accuracies based on multiple metrics.

Work done till now:

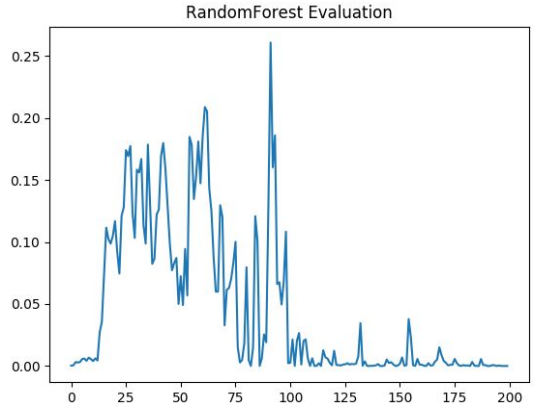
- We performed preliminary data analysis on AAPL stock price. This dataset has minute-wise data points, hence for data given for a particular minute, we predict the closing price for the next minute.
- We trained 3 different models. The metrics are given below.

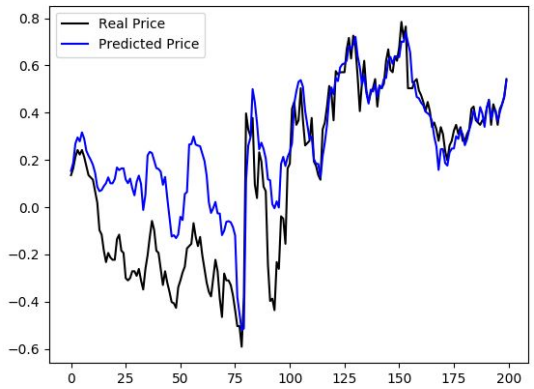
Support Vector Regressor: MSE = 0.34803



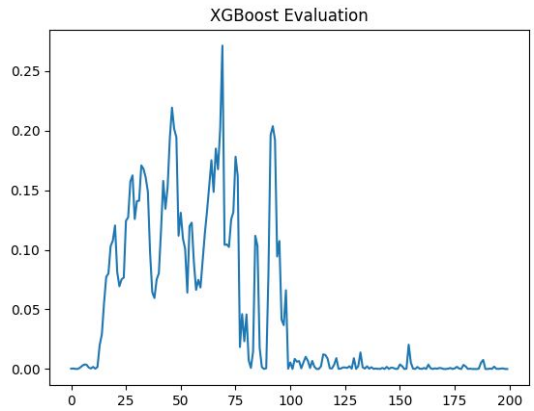


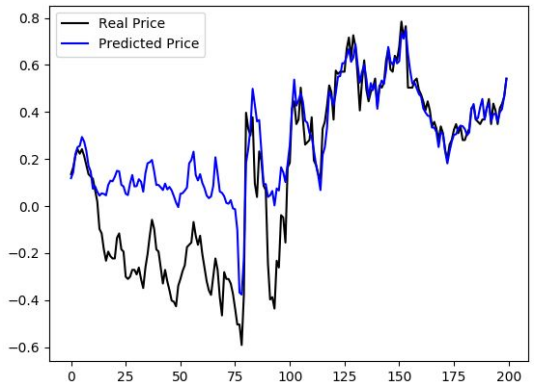
Random Forest: MSE = 0.04613





XGBoost: MSE = 0.04761





Challenges encountered:

- Since we are engineering new features, it is important to select the right indicators as features for the dataset. We looked at many research papers, did our own analysis by reading blogs and articles and selected the 22 features as described in the previous section.
- Generally closing prices don't fluctuate within a minute. So including the closing price of past minute in the training set is not a good idea, since the model might just copy the previous closing price to minimize errors. Hence, we decided not to include all of the 5 main features in trainX.
- Since indicators are based on a sliding window of data points, the number of data points is also a hyperparameter that can be tuned as per needs of the end-user as well as to increase performance. So selecting the right value is also a challenge.

Remaining Work

- We have only worked on one stock (AAPL). We are planning to perform experiments with a total of 5 stocks.
- We are currently working on minute-wise data. We also want to explore whether changing the frequency of data points affects the performance in any way.
- Other hyperparameters are also to be tuned, like the sliding window of indicators, C and gamma for SVR, etc.
- We also want to explore other models like AdaBoost, neural networks, etc.
- We have currently reported only a few of all the accuracy metrics that we want to show. Based on their values, the hyperparameters will be tuned as well.

Team Member Contribution

Vidhey: responsible for preliminary data analysis, data extraction, feature engineering
Farhan: responsible for feature engineering, model training, hyperparameter tuning