MIE1628 Big Data Final Project

Apple Stock Price Time Series Prediction

Random Forest Model

Wenjia Wu 999151643
6/30/2019

1. PROBLEM

A time series data is a data set indexed in time order and the order of data is very important. A stock price data across a period of time is a time series data.

The scope of the project is to train a model that can predict stock price of Apple in the future inside a cluster environment. For this particular project, Spark Scala is chose to be the programing language to be run on Databricks, a web-based platform for working with Spark. The prediction will be the price in one day, one week, two weeks, one month and four months.

2. DATA

The data can be accessed via Yahoo Finance, and it includes information about the daily prices of the Apple stock. The S&P 500 stock price information was introduced as well to reflect the changing in market and provide more information of the big environment to be fed into the model. A subset of data from 2008 to 2018 was taken considering three reasons. As shown in the **Figure 1** below, Apple stock price took off around 2008. There is a financial crisis in 2008 and everything was recovering. Apple release its first iPhone in Jun 29, 2007, it then play a very important part in Apple's revenue and profit, which was reflected on the price of the stock. This subset contains around 2700 entries.

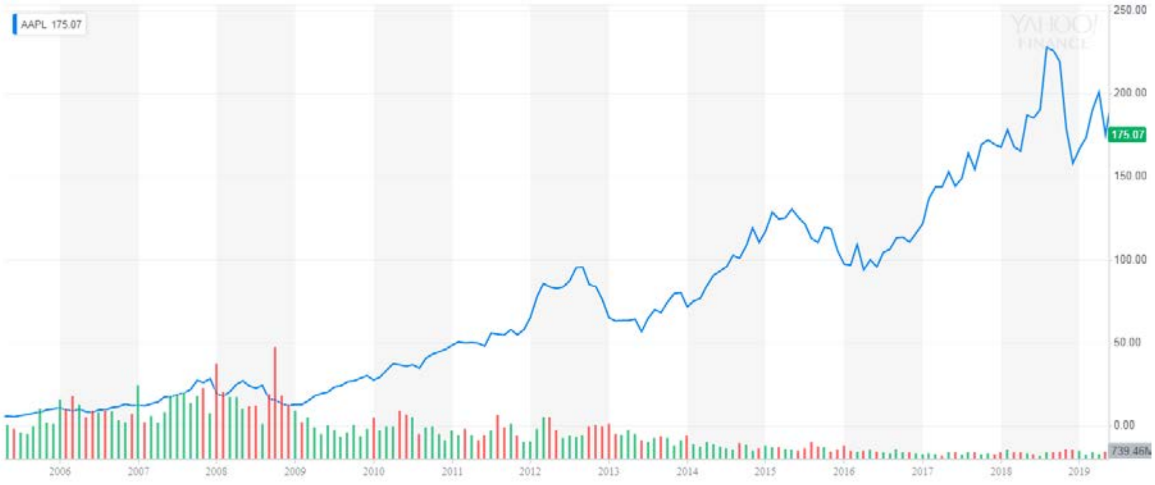


Figure 1 AAPL Stock Price From 2000 to 2019[1]

3. MODEL

The model selected for this project is Random Forest Regression. The Random Forest Regression is a parallel ensemble that leverages the parallelization of the model to enhance performance by decision tree. The model uses a number of decision trees to form its "forest" and then bootstrapped information is fed into these short trees to predict. The regression result is then calculated by averaging the results from the short decision trees.

Thanks to the bootstrapping method and parallel structure, Random Forest can prevent over fitting to certain extend and since it utilize short trees with parallel structure, it can be relatively fast in cluster environment. Hence, Random Forest was chosen for this project.

4. FEATURE ENGINEERING

Feature engineering is the keystone model builds on, as there are only 7 features in the original data set, date, open and close price for that day, highest and lowest price during the day, adjusted close price and volume, more features need to be engineered to enrich the input. Hence, after normalize the prices, the means and standard deviations of return and volume across different ranges of time, from daily result to annually result, were generated to capture the changing both in variance and across spam of time. With both vertically and horizontally information captured, the model can very well predict the variation across different time spams.

However there is a limitation to the decision tree based model, the model cannot predict a data point that is not inside the range of the train data set, as shown in **Figure 2** below. Due to the nature of stock price data, it tends to be non-stationary which means it will eventually have data points out of previous data's range. In order to address this, differences between each period were introduced. Hence, instead of predicting future prices directly, the model is tasked to predict the price differences between the future and now. The differences are most likely to be included in the training data set and thus the data set becomes stationary. As the result 88 features in total were generated.

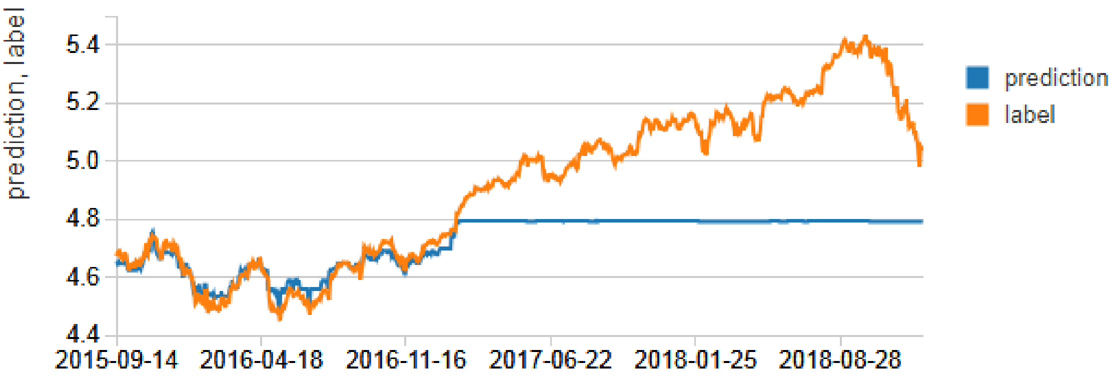


Figure 2 Data Points Outside of Train Range

5. MODEL PREPARATION

The data set now have 88 features and 5 different target variables for 5 different periods listed at the beginning of the report. The data set is split into 70% of training set and 30% of testing set. Due to the nature of the time series data, order of the data points matters, random cross validation or random split validation are not suited for validate the model. Rolling cross validation is recommended for time series data but due to the time spam of this project and no available library can be used, manual tuning is chose. Hence, to perform manual tuning the training data set is split into two data sets with 20% of data for validation. The validation set is used when tuning parameters of the model to make sure steady performance.

6. MODEL TUNING

There are two most important hyper parameters to be tuned in Random Forest, maximum depth and number of trees. The maximum depth limit how deep one single decision tree goes and the number of trees indicate how big the forest will be. RMSE is used here to measure performance of the models. RMSE measures the distance between predictions and the label, it is the most simply and fast way to tell how well one model performs and available in Spark.

One thing to be noted is that, in order to preventing over fitting, the parameter that enables same or close performance of training and validation set should be chose. As illustrated in **Figure 3** below, as the maximum depth increasing, RMSE of train set keep declining while the valid set increasing, which is the process of the model try to fit the train set as better as possible, which will result in over fitting. Hence, in this case, maximum depth of 3 should be chose.

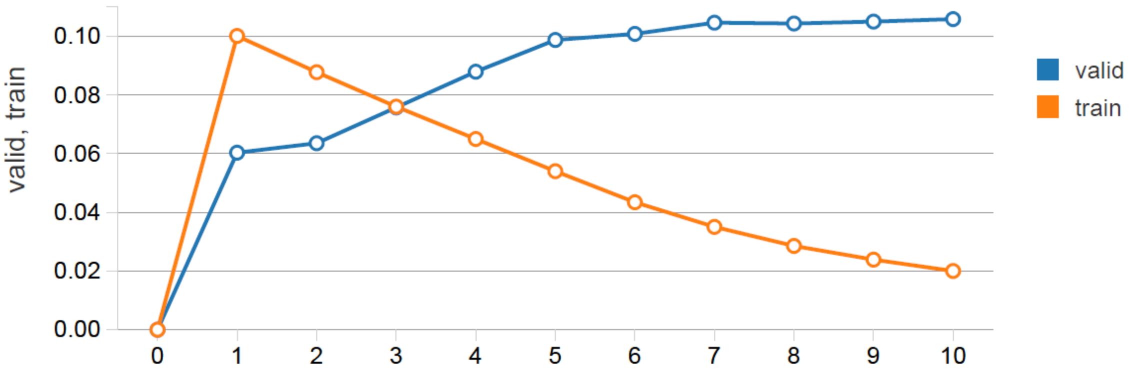


Figure 3 Maximum Depths vs. RMSE

Then each of the 5 models is tuned separately to get their sets of parameters that have best performance. Now with parameters set, the models are fed with testing data set to evaluate performance.

7. FEATURE IMPORTANCE

Chart 1: The top three features of each model

Model	1st Feature	2nd Feature	3rd Feature
1 Day	SP500 Weekly Volume Difference (13%)	Return (13%)	Biweekly Volume Standard Deviation (11%)
1 Week	Open Price (13%)	SP500 Adjusted Close (13%)	Weekly Volume Mean (13%)
2 Weeks	4 Months Return Standard Deviation (12%)	Annual Volume Standard Deviation (11%)	Last 4 Months Price Difference (11%)
1 Month	Monthly Volume Mean (17%)	Annual Volume Mean (16%)	Annual Return (11%)
4 Months	4 Months Volume Standard Deviation (18%)	SP500 Annual Volume Standard Deviation (15%)	SP500 4 Months Return Standard Deviation (12%)

The **Chart 1** above shows what is the top 3 feature importance are and how important it is, please refer to **Appendix** A to E for top 10 important features for each model. One observation can easily be made that the market clearly has a big impact on the price which is represented in the model by S&P500. It is also true in real life because market represents how well the whole economy is doing and individual company's performance can be easily affected by the economic environment it is in.

The other conclusion can be made here is that when the prediction period grows, feature importance tend to become more and more concentrated in fewer features, which can be seen clearer in the donut charts in **Appendix** A to E. The reason might be as the prediction time length grows, the less short time variation can affect the price and the bigger environment's influence it becomes.

8. MODEL EVALUATION

There are two performance measurement matrices used here, one is RMSE mentioned before to ensure models are neither over fitting nor under fitting and reflect how well a model can predict to a certain level. However, RMSE can only be compared to each other as it is scale dependent, in order to address this issue, another matrix, SMAPE is introduced to measure the error relatively to the average of prediction and label. Since it is percentage based, the matrix is scale independent, which is suited to compare different machine learning models and other time series analysis models.

8.1 RMSE Results

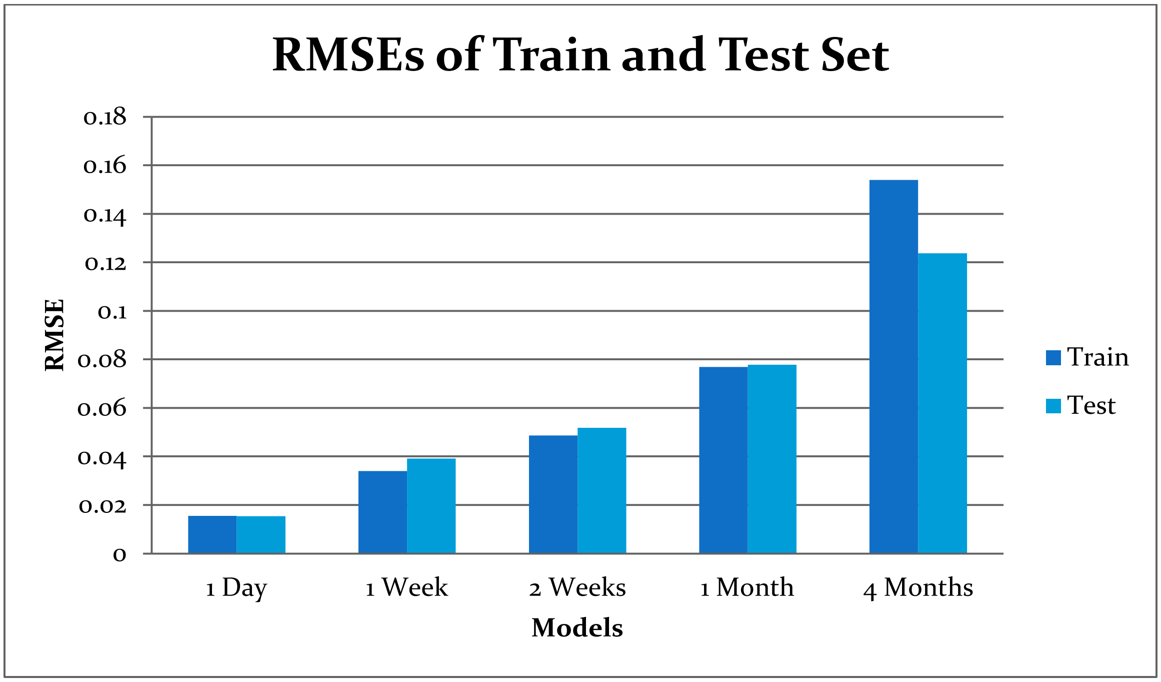


Figure 4

The RMSEs of both train and test data set in each models is shown in **Figure 4** above, as it shows, after optimized, the models are mostly neither over fitting nor under fitting. Except the model predicting prices in 4 months, it is slightly under fitting as the RMSE of test set is smaller than train set, which means the model could be tuned to be slightly more fitting to the training data set so the performance across train and test set are close to each other and reaches a more stable and better predicting ability. However, with two major hyper parameters tuned, this is the best the model can do. Considering the difficulty in predicting 4 months price in the future, this model can be accepted.

8.2 SMAPE Results

Chart 2: SMAPE Results across Different Prediction Models

Model	1 day	1 week	2 weeks	1 month	4 months
ARIMA	1.051	2.522	3.798	5.663	9.753
Holt-Winters	1.005	2.465	3.636	5.695	10.364
Linear Regression	0.454	2.033	3.445	6.733	27.796
Random Forest	1.074	2.941	3.914	6.187	9.629
Gradient boosting Tree	1.260	2.988	3.643	6.380	21.069

As shown in **Chart 2**, the SMAPE of 5 different models prepared by the team are presented. Each model has its own specialty. In regards of predicting short period of time, Linear Regression performs the best. As to predict tomorrow’s price, Linear Regression has an outstanding performance with more than 50% less error than the other models. And it performs the best in 1 week prediction as well. When prediction time spam grows, Linear Regression stops showing its edge and share the similar performance as the other models. The statistic models ARIMA and Holt-Winters perform very well in 1 month prediction. When it comes to predict price in 4 months, Random Forest and ARIMA are better choice.

9. CHALLENGES AND FUTURE IMPROVEMENTS

There are several challenges have been encountered in the process of the project. One thing which was mentioned earlier in the report is the validation of the model. As there is no rolling cross validation library available, the project could potentially benefitted from a rolling cross validation for automatic parameter grid searching.

The other challenge is to enrich the input feature. Financial market is very complicated and the stock price of an individual company is influenced by many things. Hence, more data set should be incorporated into this project like tweeter data with mentions and hash tags that is related to Apple. The other alternative could be news web scrapping to capture any news that has influence on the stock of Apple. With regards to the seasonality of time series data, the company's financial statements and product releasing events could be introduced to the model to address the seasonal changes.

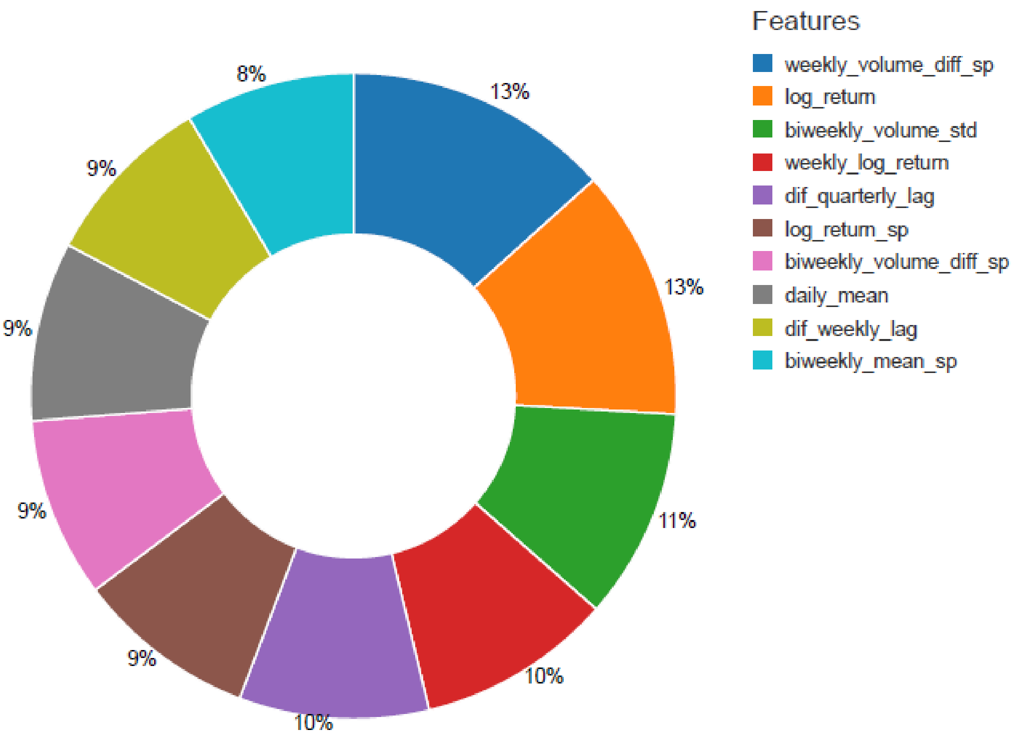
One more recommendation for the future is the neural network models and deep learning models can be explored to see if they can address the prediction better.

10. REFERENCE

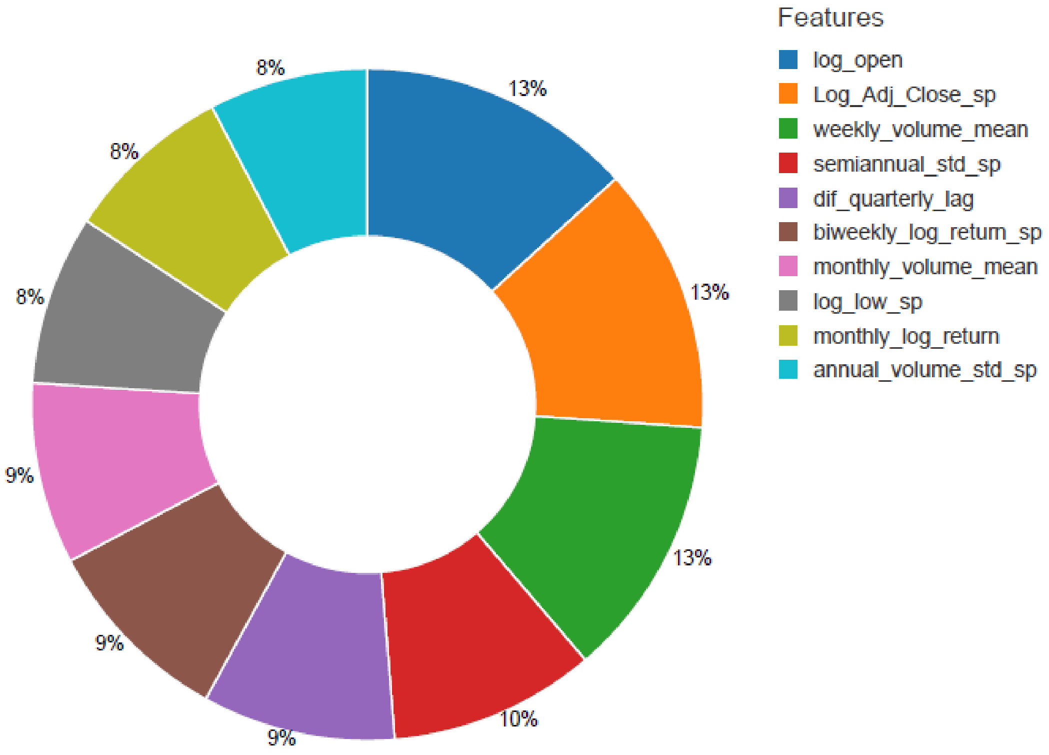
[1] Finance.yahoo.com. (2019). Yahoo is now part of Oath. [online] Available at: https://finance.yahoo.com/chart/AAPL#eyJpbnRlcnZhbCI6Im1vbnRoIiwicGVyaW9kaWNpdHkiOjEsImNhbmRsZVdpZHRoIjo3LjgxMDY1MDg4NzU3Mzk2NCwidm9sdW1lVW5kZXJsYXkiOnRydWUsImFkaiI6dHJ1ZSwiY3Jvc3NoYWlyIjp0cnVlLCJjaGFydFR5cGUiOiJsaW5lIiwiZXh0ZW5kZWQiOmZhbHNlLCJtYXJrZXRTZXNzaW9ucyI6e30sImFnZ3JlZ2F0aW9uVHlwZSI6Im9obGMiLCJjaGFydFNjYWxlIjoibGluZWFyIiwicGFuZWxzIjp7ImNoYXJ0Ijp7InBlcmNlbnQiOjEsImRpc3BsYXkiOiJBQVBMIiwiY2hhcnROYW1lIjoiY2hhcnQiLCJ0b3AiOjB9fSwibGluZVdpZHRoIjoyLCJzdHJpcGVkQmFja2dyb3VkIjp0cnVlLCJldmVudHMiOnRydWUsImNvbG9yIjoiIzAwODFmMiIsImV2ZW50TWFwIjp7ImNvcnBvcmF0ZSI6W10sInNpZ0RldiI6e319LCJyYW5nZSI6eyJwZXJpb2RpY2l0eSI6eyJpbnRlcnZhbCI6Im1vbnRoIiwicGVyaW9kIjoxfSwiZHRMZWZ0IjoiMjAwNS0wNS0wMlQwNDowMDowMC4wMDBaIiwiZHRSaWdodCI6IjIwMTktMDUtMDFUMDQ6MDA6MDAuMDAwWiIsInBhZGRpbmciOjB9LCJjdXN0b21SYW5nZSI6eyJzdGFydCI6MTExNTAwNjQwMDAwMCwiZW5kIjoxNTU2NjgzMjAwMDAwfSwic3ltYm9scyI6W3sic3ltYm9sIjoiQUFQTCIsInN5bWJvbE9iamVjdCI6eyJzeW1ib2wiOiJBQVBMIn0sInBlcmlvZGljaXR5IjoxLCJpbnRlcnZhbCI6Im1vbnRoIn1dLCJzdHVkaWVzIjp7InZvbCB1bmRyIjp7InR5cGUiOiJ2b2wgdW5kciIsImlucHV0cyI6eyJpZCI6InZvbCB1bmRyIiwiZGlzcGxheSI6InZvbCB1bmRyIn0sIm91dHB1dHMiOnsiVXAgVm9sdW1lIjoiIzAwYjA2MSIsIkRvd24gVm9sdW1lIjoiI0ZGMzMzQSJ9LCJwYW5lbCI6ImNoYXJ0IiwicGFyYW1ldGVycyI6eyJ3aWR0aEZhY3RvciI6MC40NSwiY2hhcnROYW1lIjoiY2hhcnQifX19fQ%3D%3D [Accessed 30 Jun. 2019].

11. APPENDIXES

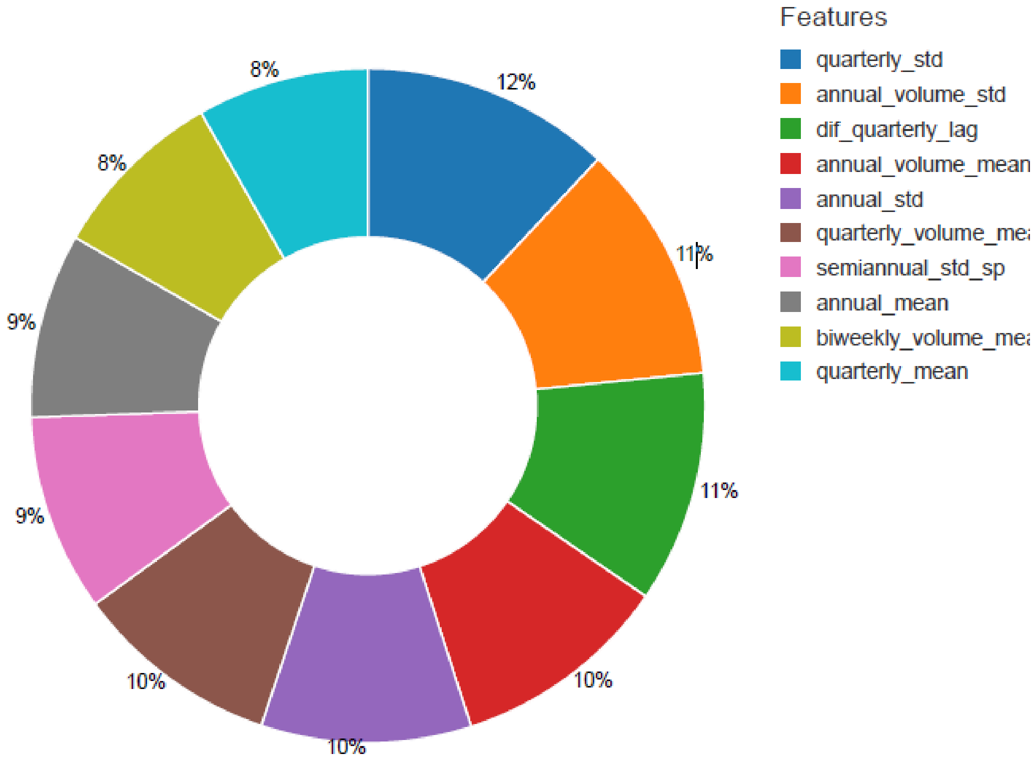
Appendix A: 1 Day Prediction Feature Importance



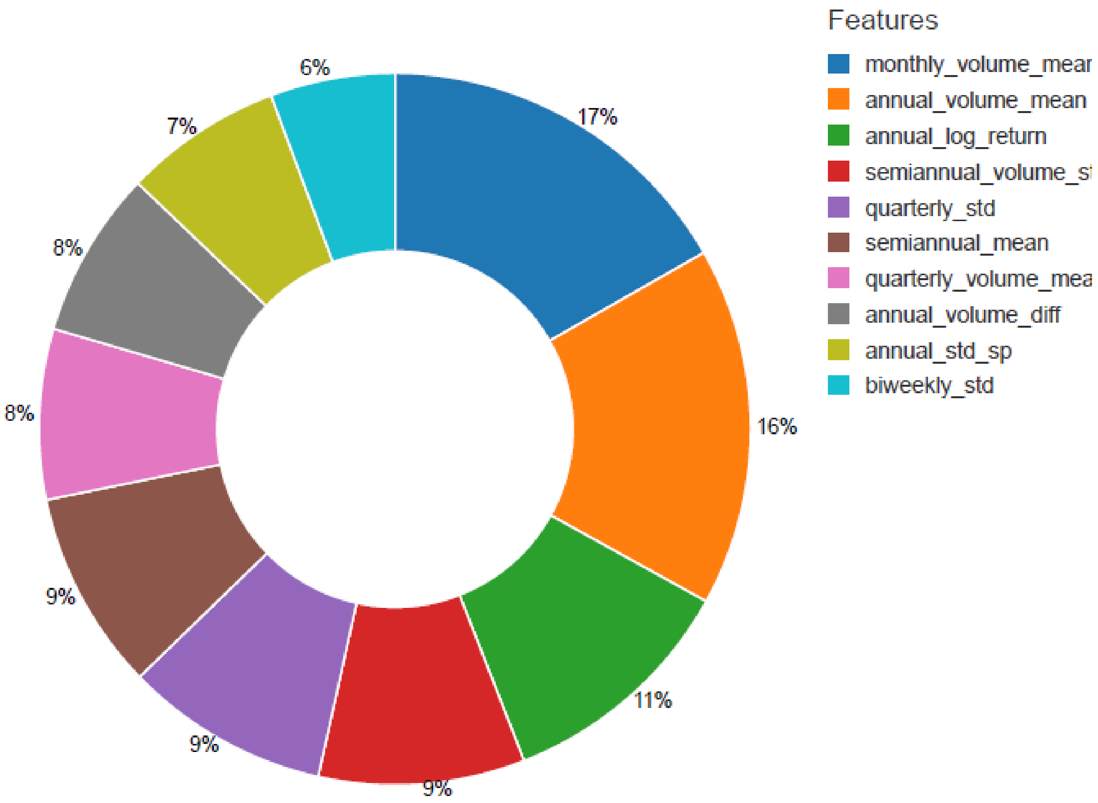
Appendix B: 1 Week Prediction Feature Importance



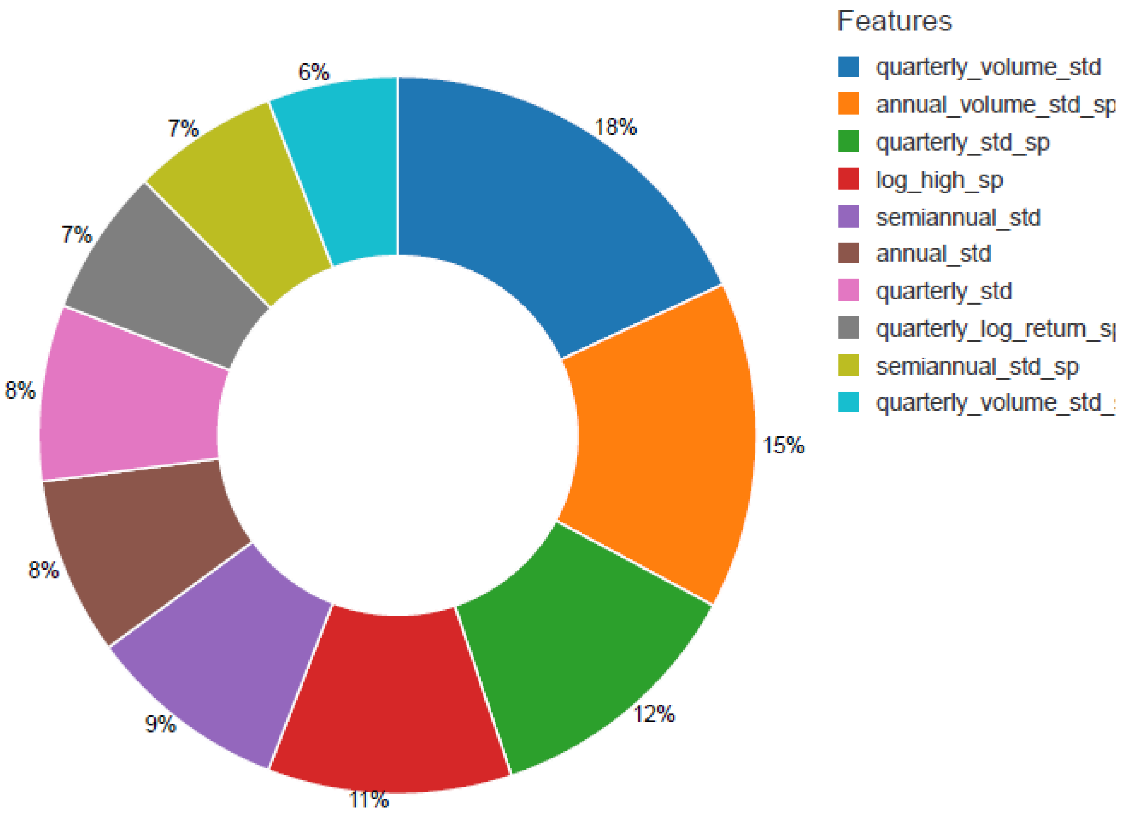
Appendix C: 2 Weeks Prediction Feature Importance



Appendix D: 1 Month Prediction Feature Importance



Appendix E: 4 Months Prediction Feature Importance



Appendix F: Code

Please refer to following pages for spark code with the following order: Random Forest, ARIMA, Holts-Winters, Linear Regression, GBT