Brief Overview: Name: Tony Jiang

I have been very interested in the concept of automated stock trading for a long time but had never really had any chance to do any project on it. After learning from the past couple lessons about different models of machine learning, I thought maybe I can apply my knowledge and practice into something I am interested in. Thus, I decided to use the Nasdaq datasets on yahoo finance to test whether we can predict statistical values of a trading date based on the statistical datas of the previous trading date. In the dataset (Nasdaq.csv), each row of the file contains several statistical datas (such as Open, High, Close, Volumn,etc) of a specific trading data. I also manually pasted the same statistical datas from the previous trading data on to the same row. The entire dataset spans for the past last 5 years, and I arbitrarily used the last 25 dates as the testing set and rest of the dataset as the training set.

Independent and Dependent Variables:

There are 6 features (independent variables) that the decision trees to take into consideration. They are all statistical values of the previous trading date: *continuous value

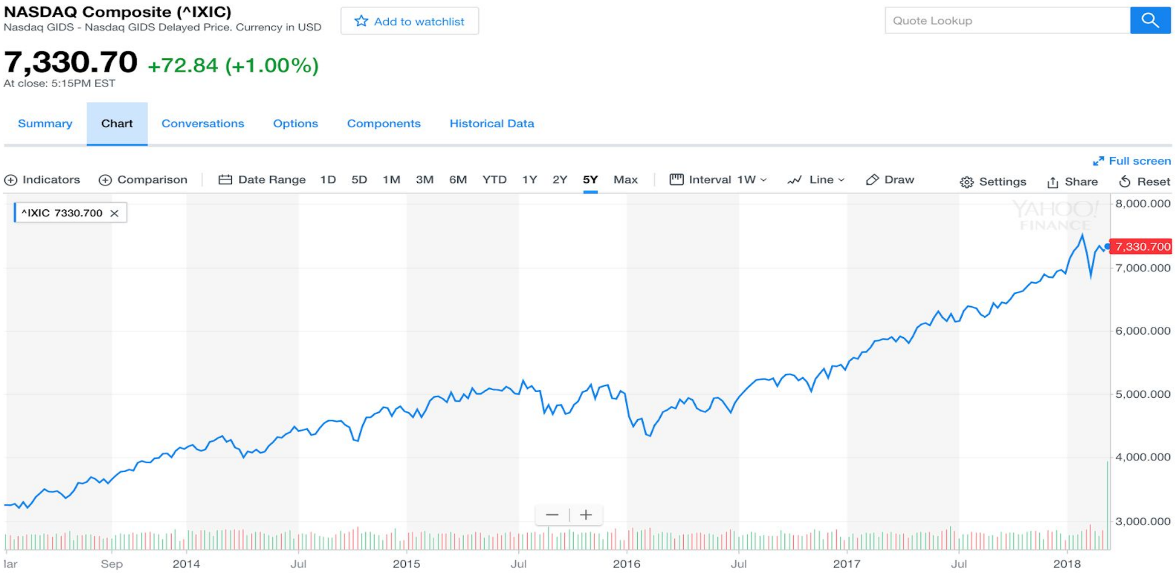
feature_names = ['Yesterday_Open', 'Yesterday_High', 'Yesterday_Low', 'Yesterday_Close', 'Yesterday_AdjClose','Yesterday_Volumne']

The dependent variable could be any statistical value of the current trading date. For my tests, I decided to use the High value of the current trading date:

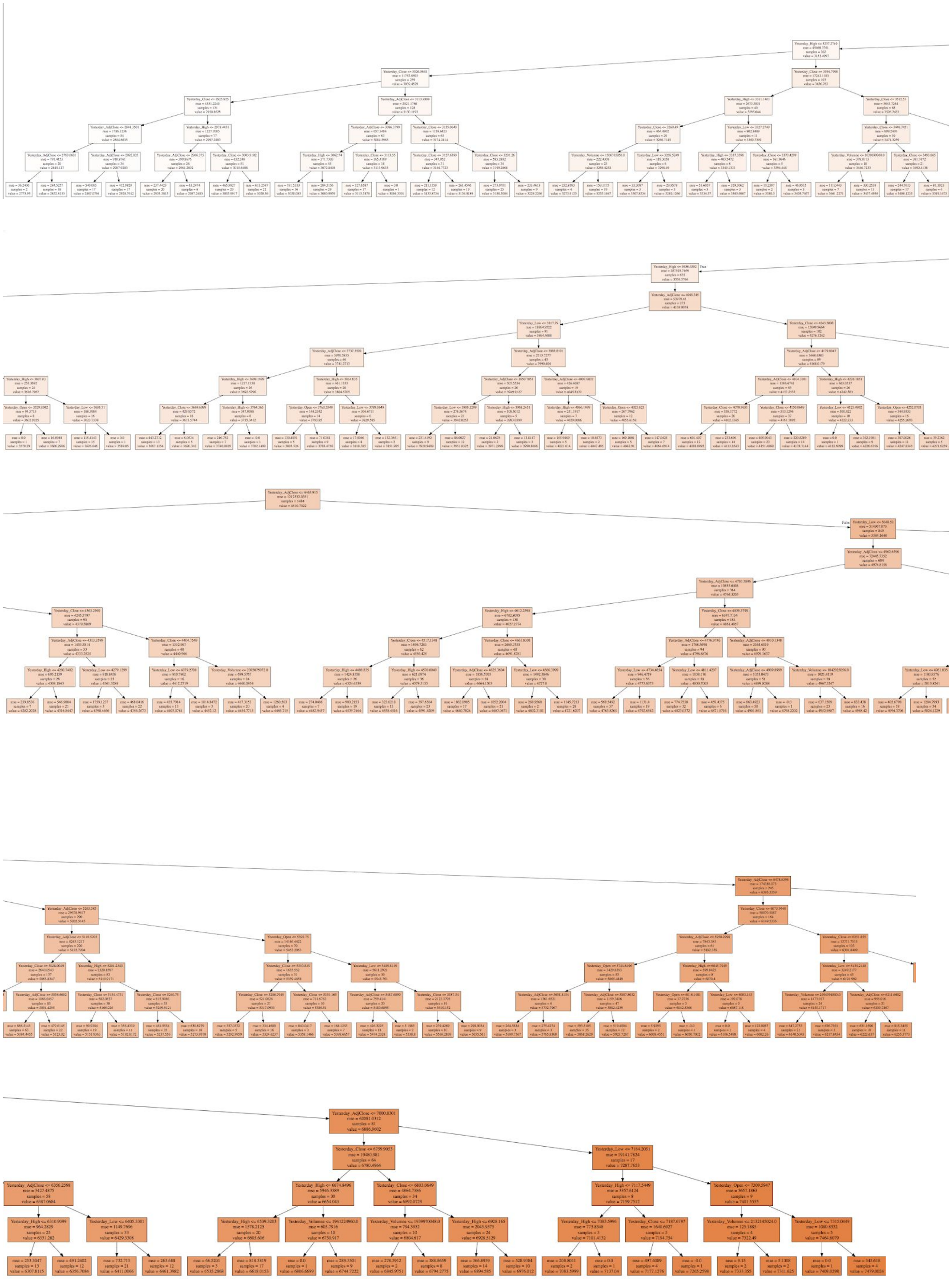
target_names = ['Today High'] *continuous value

Generic Data Graphics:

The following graph illustrates the Nasdaq closing value of all trading dates for the past 5 years



Decision Tree Image and Description:


Yesterday_AdjClose <= 7000.8301
mse = 62081.0312
samples = 81
value = 6886.9602
Yesterday_Close <= 6759.9053
mse = 19480.981
samples = 64
value = 6780.4964
Yesterday_Low <= 7184.2051
mse = 19141.7824
samples = 17
value = 7287.7653
Yesterday_AdjClose <= 6356.2598
mse = 3427.4875
samples = 58
value = 6387.0684
Yesterday_High <= 6674.8496
mse = 5946.3589
samples = 30
value = 6654.043
Yesterday_Close <= 6803.0649
mse = 4864.7386
samples = 34
value = 6892.0729
Yesterday_High <= 7117.5449
mse = 3357.6124
samples = 8
value = 7159.7512
Yesterday_Open <= 7309.5947
mse = 5657.1863
samples = 9
value = 7401.5555
Yesterday_High <= 6310.9399
mse = 964.2829
samples = 25
value = 6331.282
Yesterday_Low <= 6405.3301
mse = 1149.7696
samples = 33
value = 6429.3308
Yesterday_High <= 6539.3203
mse = 1578.2125
samples = 20
value = 6605.606
Yesterday_Volume <= 1941224960.0
mse = 605.7916
samples = 10
value = 6750.917
Yesterday_Volume <= 1939970048.0
mse = 794.3932
samples = 10
value = 6804.617
Yesterday_High <= 6928.165
mse = 2045.9575
samples = 24
value = 6928.5129
Yesterday_High <= 7083.5996
mse = 773.8348
samples = 3
value = 7101.4132
Yesterday_Close <= 7187.6797
mse = 1640.6927
samples = 5
value = 7194.754
Yesterday_Volume <= 2132145024.0
mse = 125.1885
samples = 4
value = 7322.49
Yesterday_Low <= 7315.0449
mse = 1080.8332
samples = 5
value = 7464.8079
mse = 253.3047
samples = 13
value = 6307.8115
mse = 491.2402
samples = 12
value = 6356.7084
mse = 732.715
samples = 21
value = 6411.0066
mse = 263.688
samples = 12
value = 6461.3982
mse = 66.5201
samples = 3
value = 6535.2868
mse = 818.3819
samples = 17
value = 6618.0153
mse = 0.0
samples = 1
value = 6806.6699
mse = 289.3501
samples = 9
value = 6744.7222
mse = 278.3912
samples = 2
value = 6845.9751
mse = 388.8659
samples = 8
value = 6794.2775
mse = 366.8939
samples = 14
value = 6894.585
mse = 528.9384
samples = 10
value = 6976.012
mse = 208.8011
samples = 2
value = 7083.5999
mse = 0.0
samples = 1
value = 7137.04
mse = 497.4089
samples = 4
value = 7177.1276
mse = -0.0
samples = 1
value = 7265.2598
mse = 9.15
samples = 2
value = 7333.355
mse = 5.1308
samples = 2
value = 7311.625
mse = 0.0
samples = 1
value = 7408.0298
mse = 343.618
samples = 4
value = 7479.0024

Here are the interpretation of the first 2 layers of the regression tree:
The first layer box states that :
Yesterday_AdjClose <= 4463.915; mse = 1217532.0351; samples = 1484; value = 4610.7022
This means that out of all the 1484 data values for high values, the average is 4610.7022. The tree will further break down into two boxes in the second layers. The first box represents all of the data that have Yesterday AdjClose Value less than or equal to 4463.915. The second box, similarly, represents all of the data that have High values greater than 4463.915.

On the second layer, the left box states that:
Yesterday_High <= 3636.4502 mse = 287593.7169 samples = 635 value = 3576.5766
This means that our of 635 data here, the average High values are 3576.5766 and it would further branch out base on the condition Yesterday High <= 3636.4502.

The right box on the second layer states that:
Yesterday_Low <= 5648.52mse = 514967.073samples = 849value = 5384.1648
Here, it means that out of the 849 data, the average of their high values are 5384.1648, and it would further branch out base on the condition: Yesterday Low value <= 5648.52

DT and RF feature importances:

1. The following is the Decision Tree feature importances of the 6 features:

```
[4.92583996e-04   9.09353968e-02   1.95303670e-01   8.95871836e-03
7.04285977e-01   2.36542763e-05]
```

Order: ['Yesterday_Open', 'Yesterday_High', 'Yesterday_Low', 'Yesterday_Close', 'Yesterday_AdjClose', 'Yesterday_Volumne']

2. The following is the Random Forest feature importances of the 6 features:

```
[ 3.09105237e-02,   3.02474340e-01,   5.84258076e-02,   3.13251016e-01,
2.94897595e-01,   4.07169488e-05]
```

Order: ['Yesterday_Open', 'Yesterday_High', 'Yesterday_Low', 'Yesterday_Close', 'Yesterday_AdjClose', 'Yesterday_Volumne']

Final Thoughts:

Overall, I think it’s a successful mini project. For both decision tree and randomized forest, I am able to get CV Values in the 90s with the optimal max_depth and number_estimator(Max_depth = 7 for both and n_estimator = 200). Nevertheless, as you can see from the result (printed in the program), though my predicted value is mostly in the range + - 100 from the actual value, the High values fluctuates just as much every day in the dataset. So, I can’t really use my predicted value to accurately predict whether the market would increase or decrease based from yesterday’s values. Nevertheless, It gives us a range of difference in values from one trading day to another, and validates the strength of machine learning. I am quite pleased with the overall result.