Machine Learning For Trading
Project 3
Sriram Madapusi Vasudevan
902916993

Method used for learning:

The method used reading in all the 100 csv files. Splitting them into sets of 21 and create 5 factors to predict the price on the 26th day. The final Y stored was a difference between the 21st day price and 26th day price.

The features used were:

1. Mean
2. Standard Deviation
3. Relative Strength Index
4. Rate of Change
5. Slope of the 21 set

It is to be noted is that the sets of 21 are usually overlapping. 1-21,2-22,3-23 are the sets and so on till we reach the end of the csv. This means that usually the first 21 and last 21 of any csv file will not be of much use, since the entire 21 days before set wont be available.

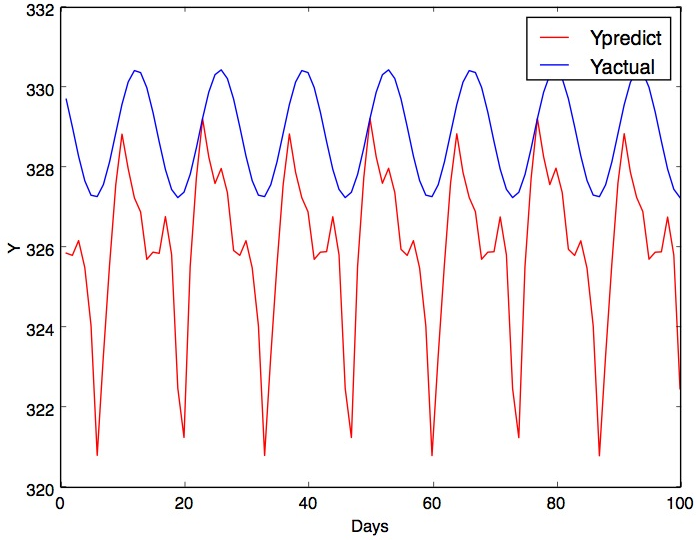
Mean :
Average of the 21 values of Y

Standard Deviation:
Standard Deviation shows how much variation or dispersion from the average exists.

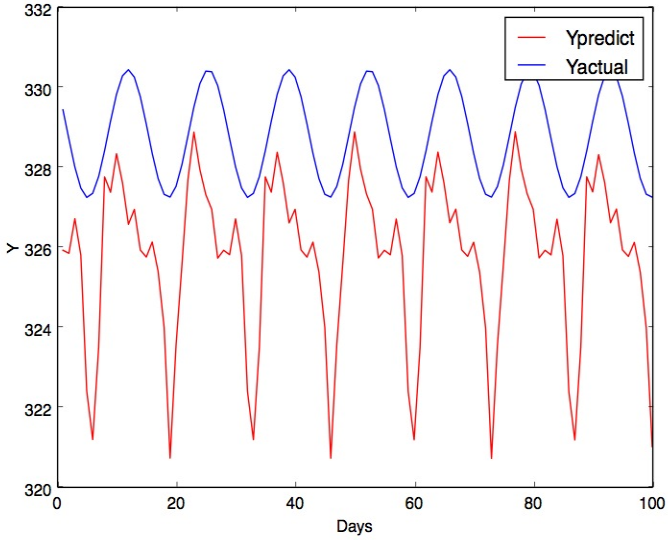
Relative Strength Index:
Relative Strength Index (RSI) is a momentum oscillator that measures the speed and change of price movements. RSI oscillates between zero and 100.

Rate of Change:
The Rate-of-Change (ROC) indicator, which is also referred to as simply Momentum, is a pure momentum oscillator that measures the percent change in price from one period to the next. The ROC calculation compares the current price with the price "n" periods ago.

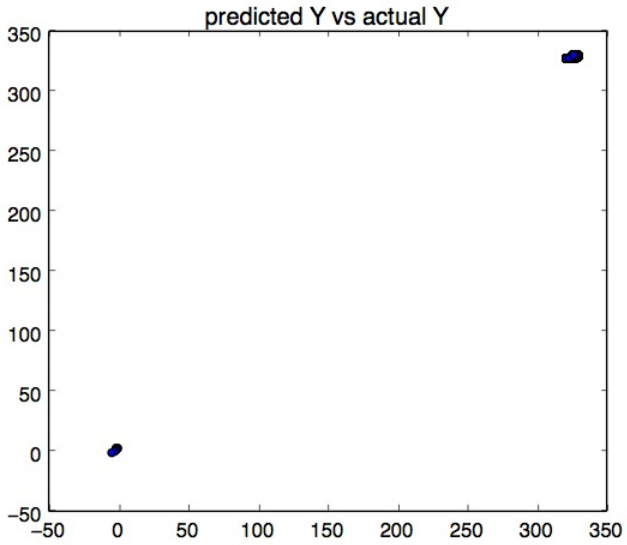
Slope:
Slope is obtained by using linear regression on 21 set of values, this is based on the assumption given the smaller dataset of 21 values. The relationship will be roughly linear. But this is an approximation at best.



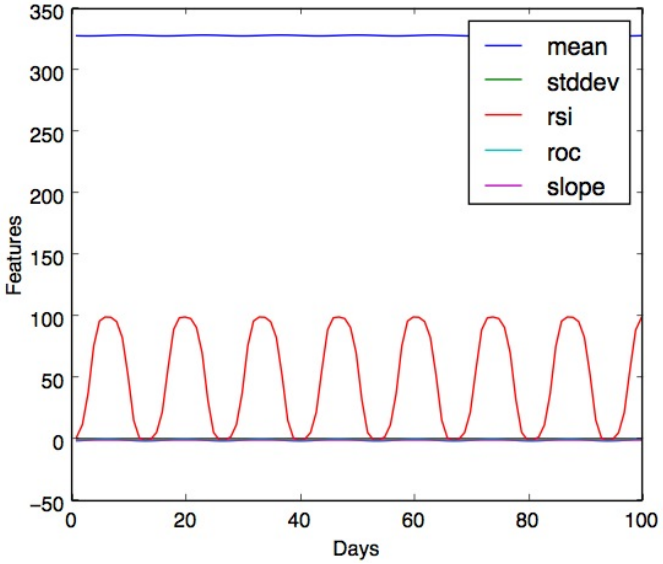
292 first 100 days



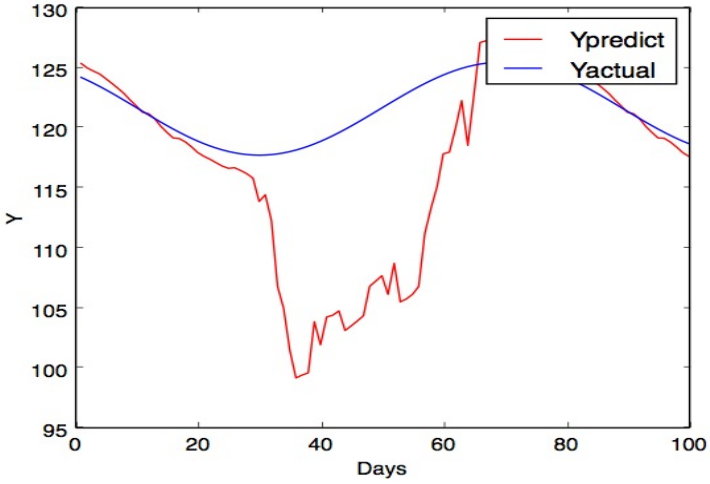
292 last 100 days



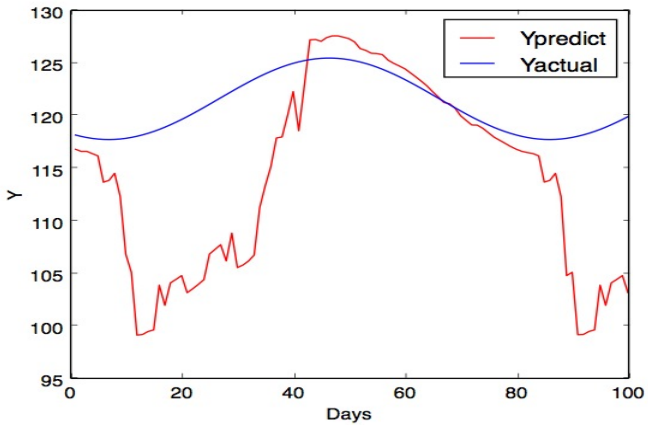
scatter plot 292



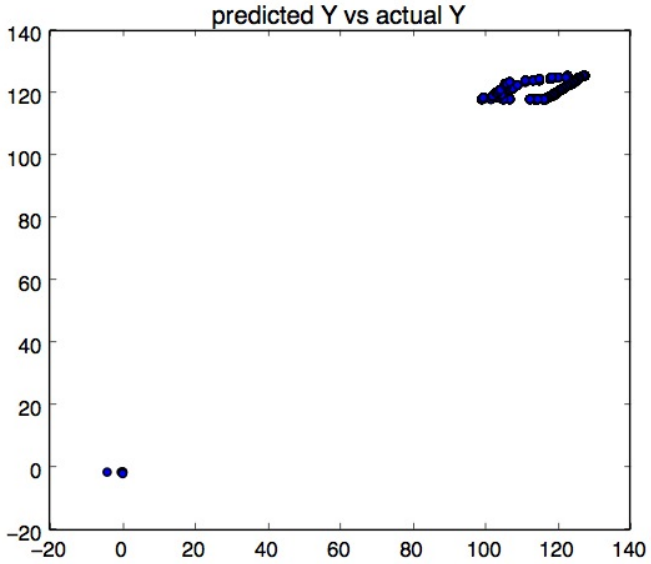
5 features plot 100 days 292



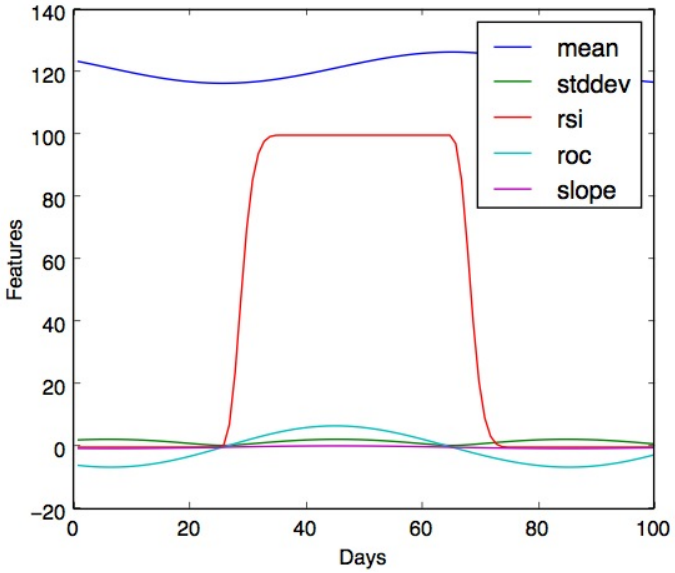
132 first 100 days



132 last 100 days



scatter plot 132



5 features plot first 100 days 132

Note:
The shown charts constitute the best runs after running for 20 times

Correlation:
0.93446 for MLT292
0.89732 for MLT132

RMS Error:
1.8782 for MLT292
5.9245 for MLT132

The Method that I used was the Random forest learner where I constructed 50 trees, but I ran into the problem where the constructed tree size was not large enough due to redundant data. So after changing code to make sure that I considered trees of sizes greater than 200,000, better results started coming in.

The other problem using Random forest learners is not having a deterministic outcome but that it will perfectly capture the trend, which can be clearly seen as the correlation is quite high, but the error encountered was high as well.

One way to decrease RMSE would have been to increase number to trees to 100 and higher. But it takes way longer to run.

Conclusion:
So, even though the above method provides good correlation, I think RMSE needs to be decreased a lot.