Final Project

Maximizing Portfolio Returns using Machine Learning

DS502
Statistical Methods for Data Science
Team 8

Authors:
Harsh Pathak
Krushika Tapedia
Lei Li
Mukund Khandelwal

Supervisor:
Dr. Fatemeh Emdad

December 5, 2017

Topic: Maximizing Portfolio Returns using Machine Learning

Abstract

This project involves researching and applying methods to forecast stock prices and build an investment portfolio whose composition can be adjusted to maximize future returns. The goal is to show that using advanced statistical methods, we can derive approximate adjusted close prices of commodities and apply results to manage portfolio built on the lines of Modern Portfolio Theory by Harry Markowitz. This has been done by applying and examining results of regression and classification models on historical data of past 3 years till November 24th, 2017. Upon examination of the results, it becomes clear that while our models worked reasonably good for Bank of America and Microsoft commodities, they were relatively less robust with ANI Pharmaceuticals commodity. Through showing that models used can forecast prices to a reasonable extent, this research allows us to get involved in an active area of research in Stock Forecasting and implement models and techniques that are more advanced in capturing price variations and momentum.

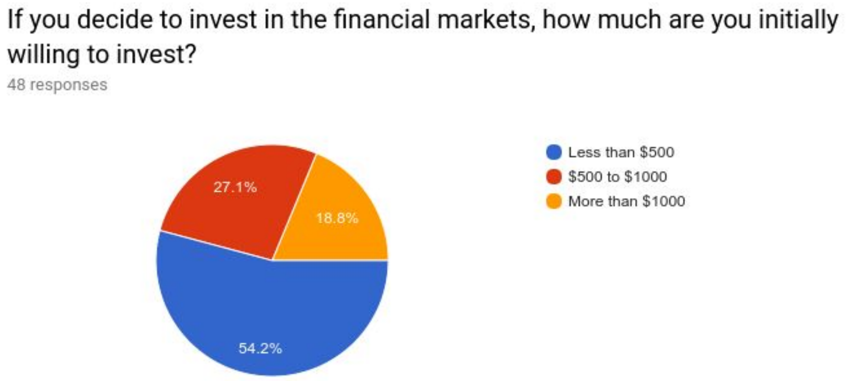
Introduction

As graduate students, investing today can lead to significant returns in future. However, the main issue is with the limitation of available capital and knowledge of markets. Our main objective is to help students choose stocks that not only fits within their budget, but also help them set forward to achieve their investment objectives. Hence, we make use of advanced machine learning techniques that help us predict stock prices and based on that, we develop a portfolio that diversifies the risks and achieves maximum return possible.

We conducted an online survey to assess what graduate students expect as an entry level investor and were able to assess their investment objectives which helped us shape our project.

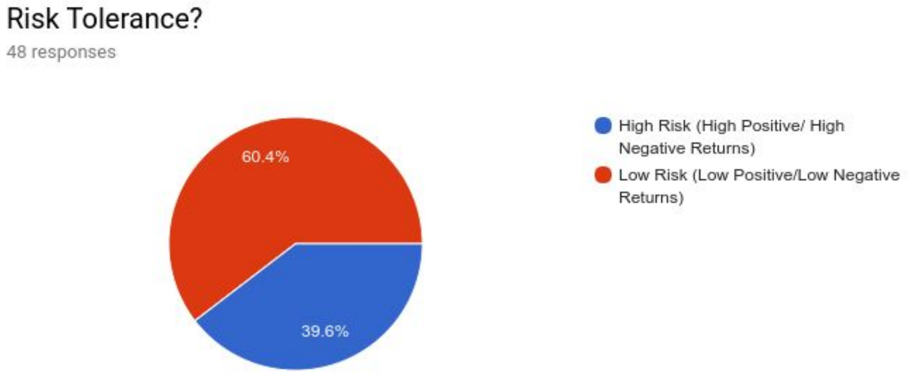
Through our online survey, we were able to gather deep insights on how graduate students weigh their investment objectives. Based on our survey,

- 53.8% of the students were willing to invest not more than $500
- 61.5% of the students wanted to invest for long term (Our definition of long term was relative to the time graduate students, on an average, spend in college)
- 80.8% of the students would rely on past performance of the market to make investments
- 59.6% of the students would prefer a low risk investment



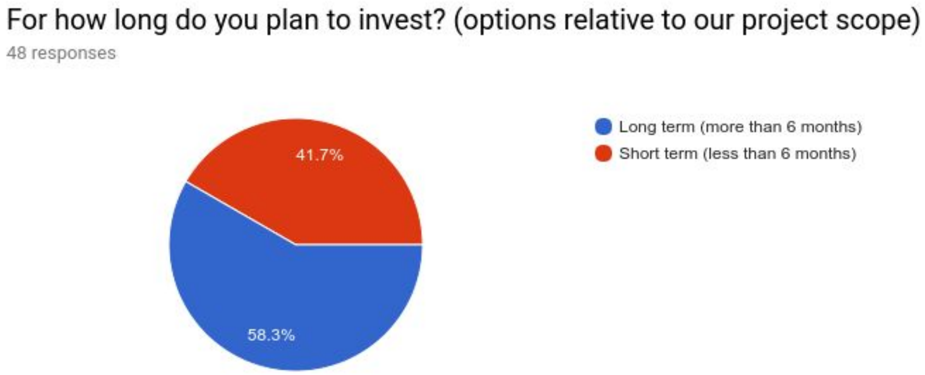
If you decide to invest in the financial markets, how much are you initially willing to invest?

48 responses



Risk Tolerance?

48 responses



For how long do you plan to invest? (options relative to our project scope)

48 responses

Our research and study on financial market through a vast array of resources available on the internet helped us choose three stocks to build our portfolio:

Bank of America, Microsoft and ANI Pharmaceuticals

One major factor that made us choose these stocks was the fact that they all belong to different industries and had correlation coefficient that was relatively small. Although one could opt for other financial instruments as well such as bonds, CDO, CMO and futures, we limited our scope of understanding to stocks as they would help us gain deeper insights into the financial sector, and later, modelling for other instruments could be incorporated based on further expansion of project scope.

On the other side, with Portfolio Management, we made sure that the scope is limited to only the main features of "what it takes to build a portfolio". Reason being that Portfolio Analysis, in itself, is an area of research with full time Master's program dedicated to teach students that skill. Our aim was to not just predict stocks, but also build a product that can be applied to real life situations. And therefore, we built a portfolio, that rebalances daily, to help investors maximise returns. We also did deploy "Time series"[6] methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

Literature Review

Many researchers have attempted to predict the market and have showcased many potential ways that one can achieve it. We kept our literature limited to recent studies though Asep Juarna in his paper describes how the trend of stock prices changing by means of prediction or extrapolating an index of stocks preceded by curve fitting computation along some periods. Curve fitting computation uses 62 days data (in accordance with value interval of 341 - 365) of IDX30 stock index published by Indonesian Stock Exchange (IDX). Curve fitting computation uses the least square method; this computation gives optimal polynomial having degree of 58 with RMSE (root mean square error) of 0.400052 or approximately 0.1133% of median of the IDX30[1]. Murtaza Roondiwala presented a recurrent neural network (RNN) and Long Short-Term Memory (LSTM) approach to predict stock market indices. LSTM's have proved their efficiency in time-series data application like prediction of next word in a text, or speech. Similar results were found in this experiment[2]. Elisabeth Woschnagg, shared the complete machine learning approach to attempt the Stock prediction problem, discussing various regression techniques and accuracy measures preferred[3].

Method

Description of the Data Set

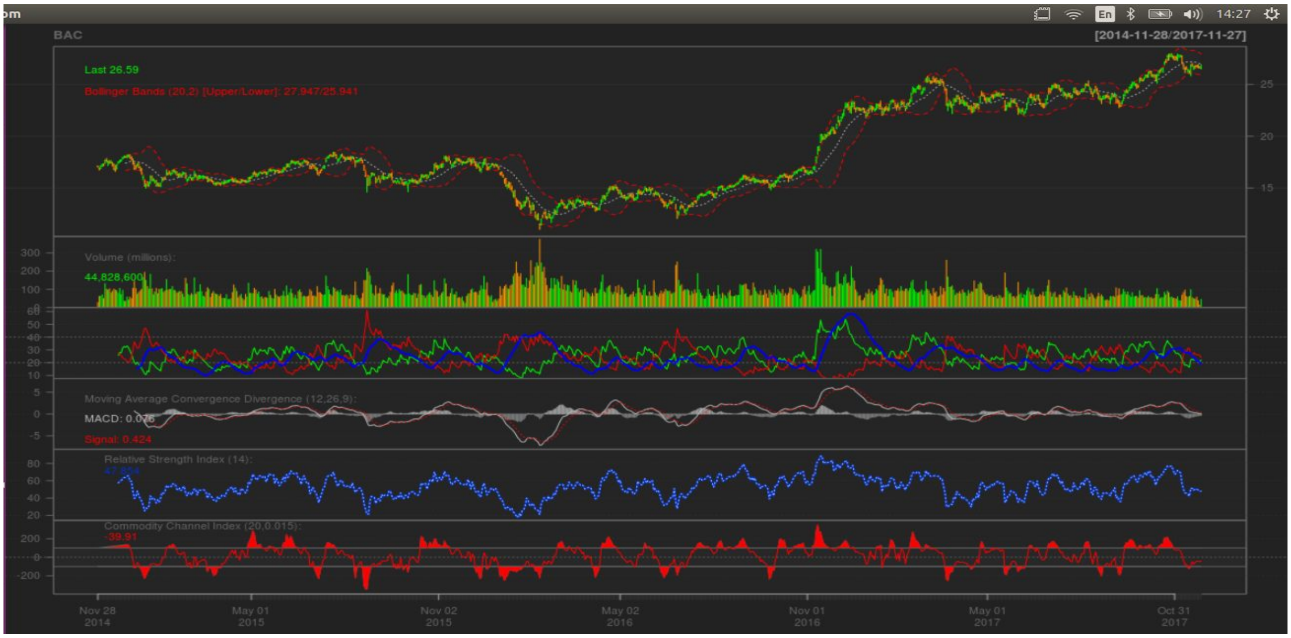
The data set is taken from Quandl, Google and Yahoo Finance website. Initially, we have decided to work on 5 stocks, belonging to different industries so as to minimize correlated risk. The time frame considered for the historical data used for forecasting will be one month. In addition to this primary data, we would perform feature engineering to add more information that can directly/indirectly affect the prices of stocks[4].

The data set consists of the following basic features:

1) **Date/Timestamp** : Date and timestamp during which the open, high and close value of a particular stock are recorded.
2) **Open** : 'Opening Price' is the price at which a security first trades upon the opening of an exchange on a given trading day. The price of the first trade for any listed stock is its daily opening price.
3) **Close** : 'Closing Price' is the price of the final trade before the close of the trading session.
4) **High** : It is the highest price of the stock during the day of trading.
5) **Adj Close:** An 'Adjusted Closing Price' is a stock's closing price on any given day of trading that has been amended to include any distributions and corporate actions that occurred at any time prior to the next day's open. The adjusted closing price is often used when examining historical returns or performing a detailed analysis on historical returns.
6) **Volume :** Volume is the number of shares or contracts traded in a security or an entire market during a given period of time. That is, when buyers and sellers agree to make a transaction at a certain price, it is considered one transaction. If only five transactions occur in a day, the volume for the day is five.

Preprocessing: Technical Analysis

TA indicators represent a statistical approach to technical analysis as opposed to a subjective approach. By looking at money flow, trends, volatility, and momentum, they provide a secondary measure to actual price movements and help traders confirm the quality of chart patterns or form their own buy or sell signals[5]. Below is the images close prices variation with other TA indicators



1. The average directional index (**ADX**) measures the strength of a prevailing trend and whether movement exists in the market. The ADX is measured on a scale of 0 to 100. A low ADX value (generally less than 20) can indicate a non-trending market with low volumes,

whereas a cross above 20 may indicate the start of a trend (either up or down). If the ADX is over 40 and begins to fall, it can indicate the slowdown of a current trend. This indicator can also be used to identify non-trending markets, or a deterioration of an ongoing trend. Although market direction is important in its calculation, the ADX is not a directional indicator.

2. The Bollinger Band (**BBANDS**) study created by John Bollinger plots upper and lower envelope bands around the price of the instrument. The width of the bands is based on the standard deviation of the closing prices from a moving average of price.
3. Developed by Donald Lambert and featured in Commodities magazine in 1980, the Commodity Channel Index (**CCI**) is a versatile **indicator** that can be used to identify a new trend or warn of extreme conditions.
4. A **simple moving average** (**SMA**) is an arithmetic moving average calculated by adding the closing price of the security for a number of time periods and then dividing this total by the number of time periods.
5. The Relative Strength Index (**RSI**), developed by J. Welles Wilder, is a momentum oscillator that measures the speed and change of price movements. The **RSI** oscillates between zero and 100. Traditionally the **RSI** is considered overbought when above 70 and oversold when below 30.

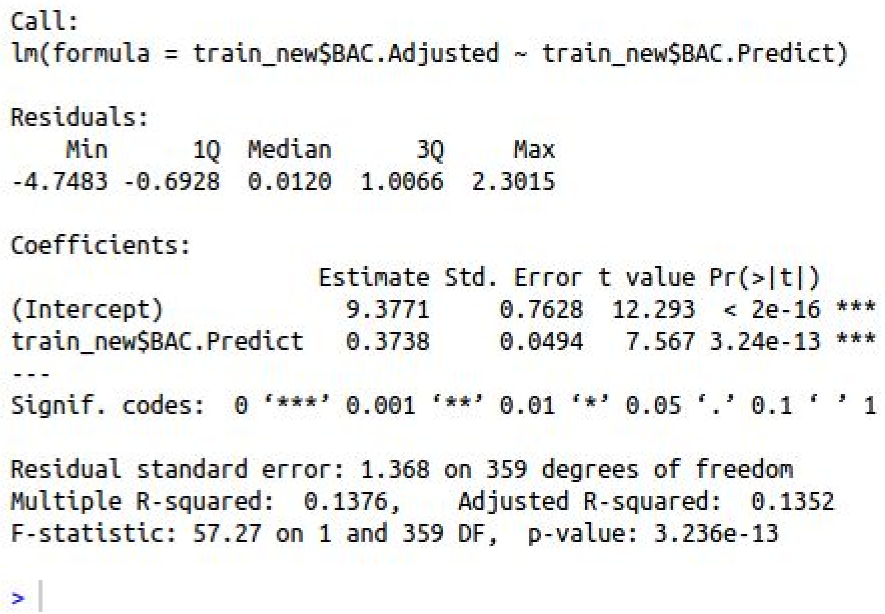
Stock Prediction Methods

Method 1 : Simple Linear Regression

Simple linear regression is a approach for predicting a quantitative response Y on the basis of a single predictor X. In this project, we use predicted stock value to predict adjusted close value.

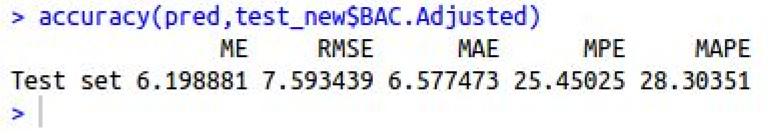
1) Bank of America

Below is the summary for Linear regression on Bank of America, which shows t-value =7.567 which is not large enough and also p-value is greater than 0.05 which makes the ANIP.Predict variable significant. The coefficient reported was 0.3786 with 0.048 Standard error. The R-squared : 13% of the variation in Y (BAC.ADJUSTED) can be described using X (BAC.PREDICT).

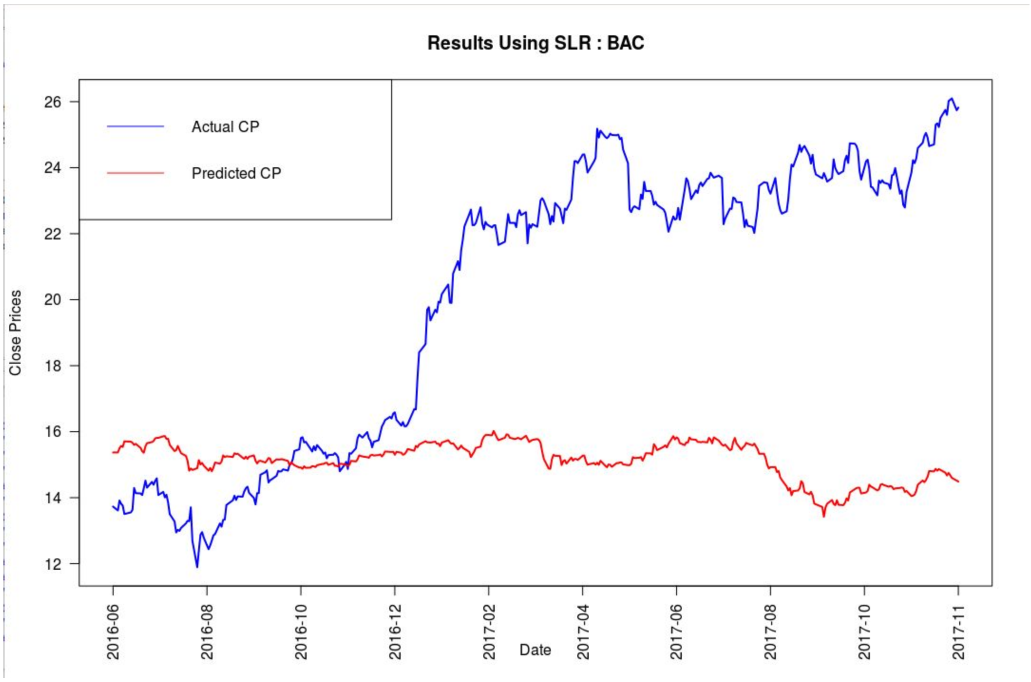


```
Call:
lm(formula = train_new$BAC.Adjusted ~ train_new$BAC.Predict)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.7483 -0.6928  0.0120  1.0066  2.3015 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)             9.3771     0.7628  12.293  < 2e-16 ***
train_new$BAC.Predict   0.3738     0.0494   7.567 3.24e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.368 on 359 degrees of freedom
Multiple R-squared:  0.1376,	Adjusted R-squared:  0.1352 
F-statistic: 57.27 on 1 and 359 DF,  p-value: 3.236e-13

> 
```

Below we are reporting the **RMSE = 7.59** and other accuracy parameters of the model.

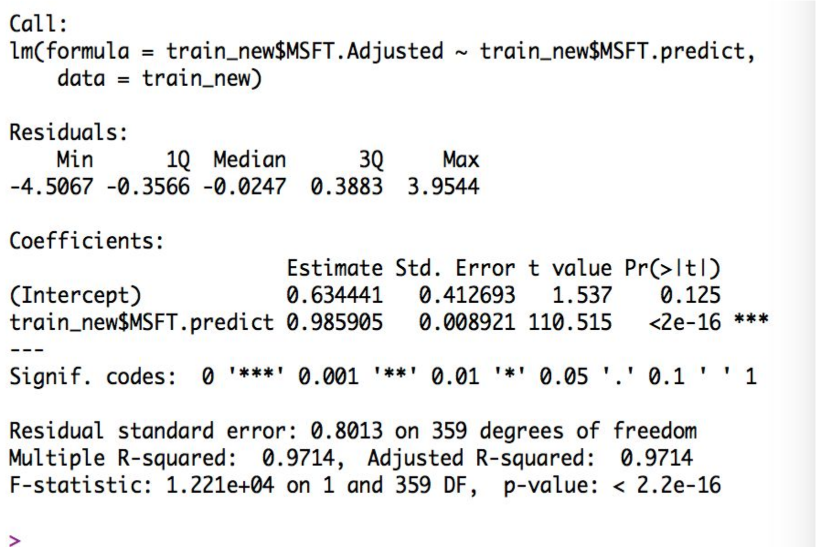


```
> accuracy(pred,test_new$BAC.Adjusted)
               ME     RMSE      MAE      MPE     MAPE
Test set 6.198881 7.593439 6.577473 25.45025 28.30351
> 
```



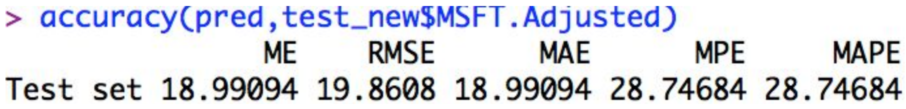
2) **Microsoft**

Below is the summary for Linear regression on Microsoft Pharmaceuticals, which shows t-value =110.5 which is not large enough and also p-value is less than 0.05 which makes the ANIP.Predict variable significant. The coefficient reported was 0.986 with 0.008 Standard error. The R-squared : 97% of the variation in Y (MSFT.ADJUSTED) can be described using X (MSFT.PREDICT).



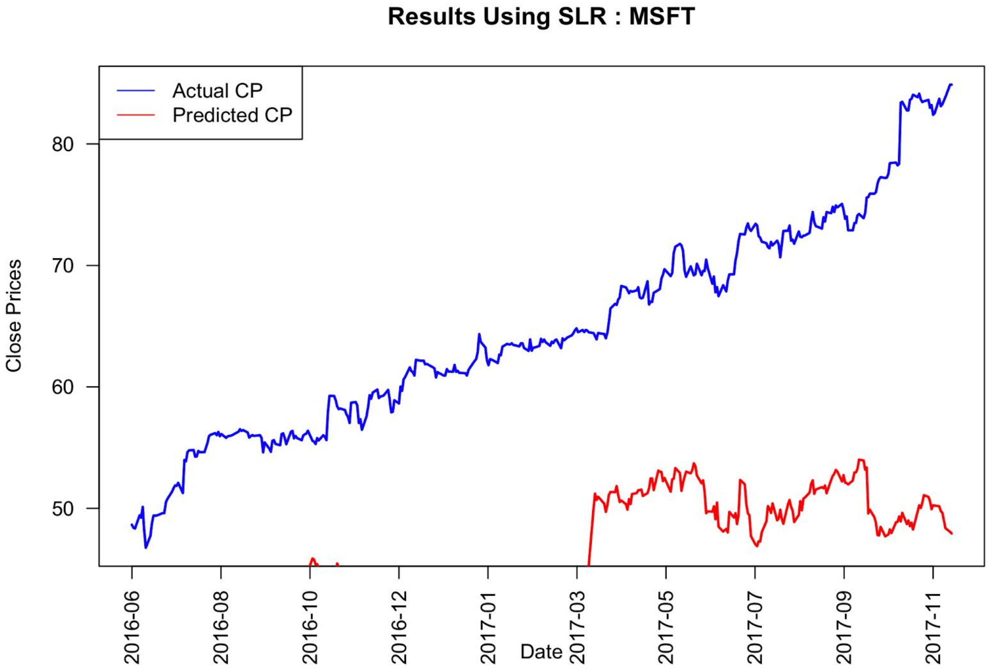
```
Call:
lm(formula = train_new$MSFT.Adjusted ~ train_new$MSFT.predict,
    data = train_new)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5067 -0.3566 -0.0247  0.3883  3.9544 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)            0.634441   0.412693   1.537    0.125    
train_new$MSFT.predict 0.985905   0.008921 110.515   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8013 on 359 degrees of freedom
Multiple R-squared:  0.9714,	Adjusted R-squared:  0.9714 
F-statistic: 1.221e+04 on 1 and 359 DF,  p-value: < 2.2e-16

>
```

Below we report the RMSE= 19.86 and other accuracy parameters.



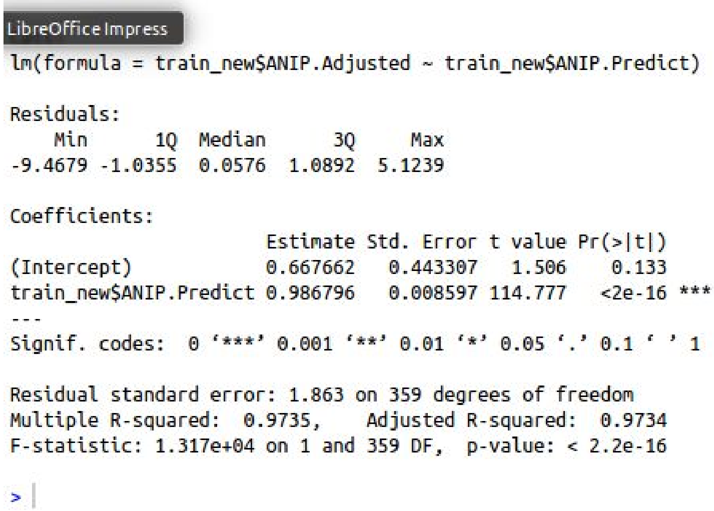
```
> accuracy(pred,test_new$MSFT.Adjusted)
               ME    RMSE      MAE      MPE     MAPE
Test set 18.99094 19.8608 18.99094 28.74684 28.74684
```

Below is the Prediction plot of the Close Price vs Time for the test data. From the plot we can see that the prediction is going off from the actual close price.



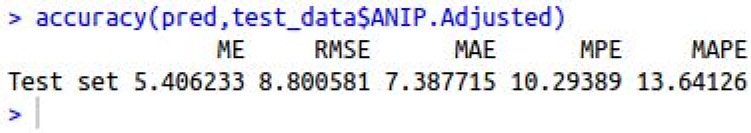
3) **ANI Pharmaceuticals**

Below is the summary for Linear regression on ANIP Pharmaceuticals, which shows t-value =114.77 which is large enough and also p-value is less than 0.05 which makes the ANIP.Predict variable significant. The coefficient reported was 0.986 with 0.008 Standard error. The R-squared : 97% of the variation in Y (ANIP.ADJUSTED) can be described using X (ANIP.PREDICT).



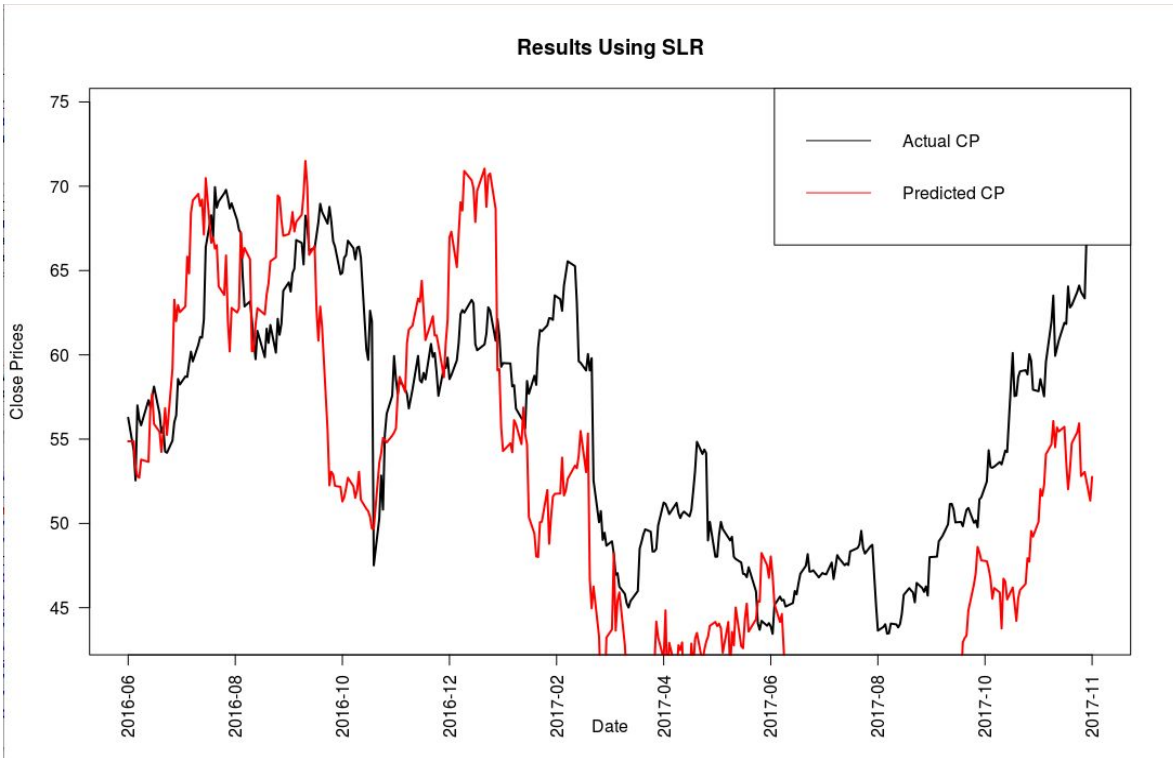
```
lm(formula = train_new$ANIP.Adjusted ~ train_new$ANIP.Predict)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.4679 -1.0355  0.0576  1.0892  5.1239 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)            0.667662   0.443307   1.506    0.133    
train_new$ANIP.Predict 0.986796   0.008597 114.777   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.863 on 359 degrees of freedom
Multiple R-squared:  0.9735,	Adjusted R-squared:  0.9734 
F-statistic: 1.317e+04 on 1 and 359 DF,  p-value: < 2.2e-16

> 
```

Below we are reporting the **RMSE = 8.80** and other accuracy parameters of the model.



```
> accuracy(pred,test_data$ANIP.Adjusted)
               ME     RMSE      MAE      MPE     MAPE
Test set 5.406233 8.800581 7.387715 10.29389 13.64126
> 
```

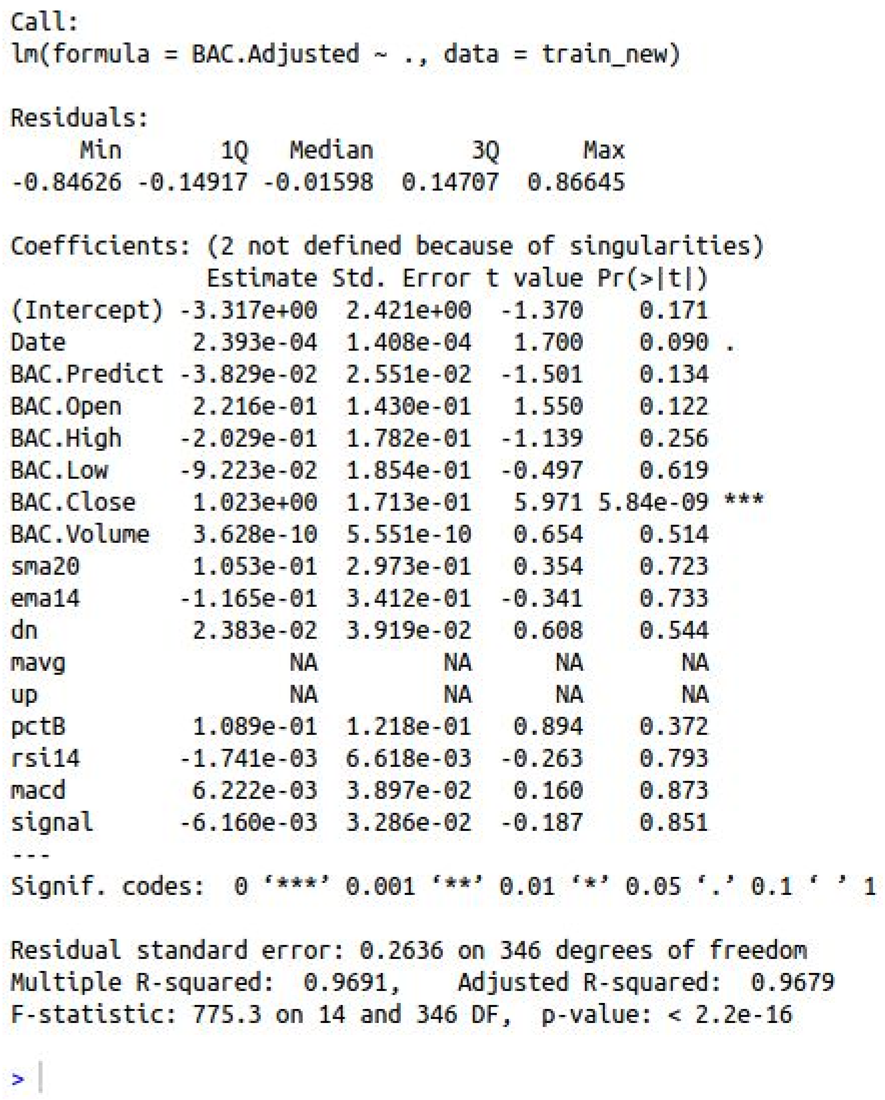
Below is the Prediction plot of the Close Price vs Time for the test data. From the plot we can see that the prediction is going off from the actual close price.



<u>Method 2: Multiple Linear Regression</u>

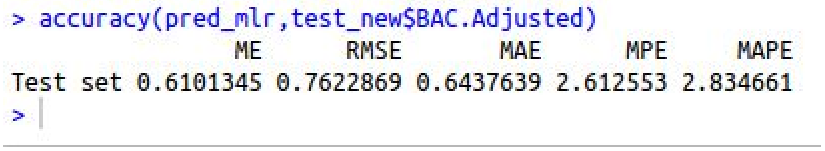
1) **Bank of America**

Below is the summary for Multiple Linear regression on Bank of America , where it hard to determine the significant contributors from the p-value. The R-squared : 96.91% of the variation in Y (BAC.ADJUSTED) can be described using X-matrix (BAC.PREDICT+TA INDICATORS).



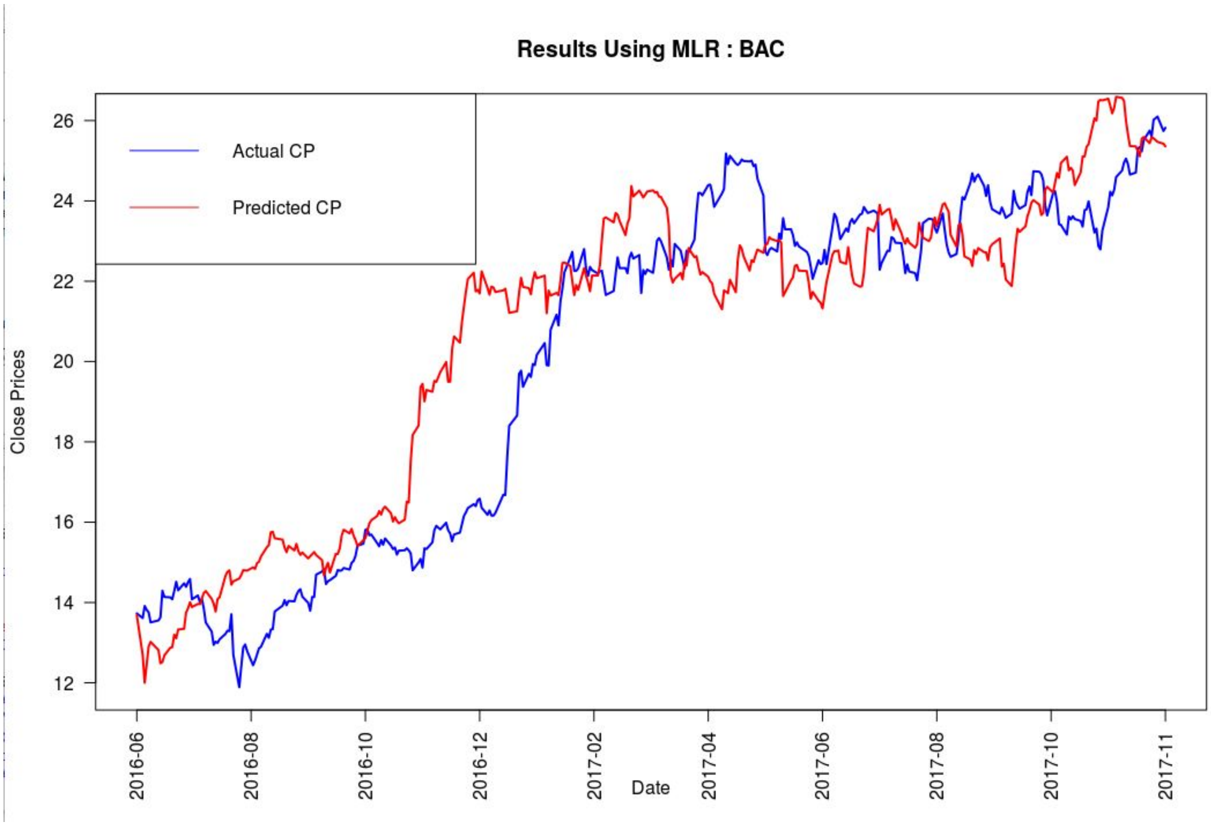
```
Call:
lm(formula = BAC.Adjusted ~ ., data = train_new)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.84626 -0.14917 -0.01598  0.14707  0.86645 

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.317e+00  2.421e+00  -1.370    0.171    
Date         2.393e-04  1.408e-04   1.700    0.090 .  
BAC.Predict -3.829e-02  2.551e-02  -1.501    0.134    
BAC.Open     2.216e-01  1.430e-01   1.550    0.122    
BAC.High    -2.029e-01  1.782e-01  -1.139    0.256    
BAC.Low     -9.223e-02  1.854e-01  -0.497    0.619    
BAC.Close    1.023e+00  1.713e-01   5.971 5.84e-09 ***
BAC.Volume   3.628e-10  5.551e-10   0.654    0.514    
sma20        1.053e-01  2.973e-01   0.354    0.723    
ema14       -1.165e-01  3.412e-01  -0.341    0.733    
dn           2.383e-02  3.919e-02   0.608    0.544    
mavg                NA         NA      NA       NA    
up                  NA         NA      NA       NA    
pctB         1.089e-01  1.218e-01   0.894    0.372    
rsi14       -1.741e-03  6.618e-03  -0.263    0.793    
macd         6.222e-03  3.897e-02   0.160    0.873    
signal      -6.160e-03  3.286e-02  -0.187    0.851    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2636 on 346 degrees of freedom
Multiple R-squared:  0.9691,	Adjusted R-squared:  0.9679 
F-statistic: 775.3 on 14 and 346 DF,  p-value: < 2.2e-16

> 
```

Below we are reporting the **RMSE = 0.762** and other accuracy parameters.



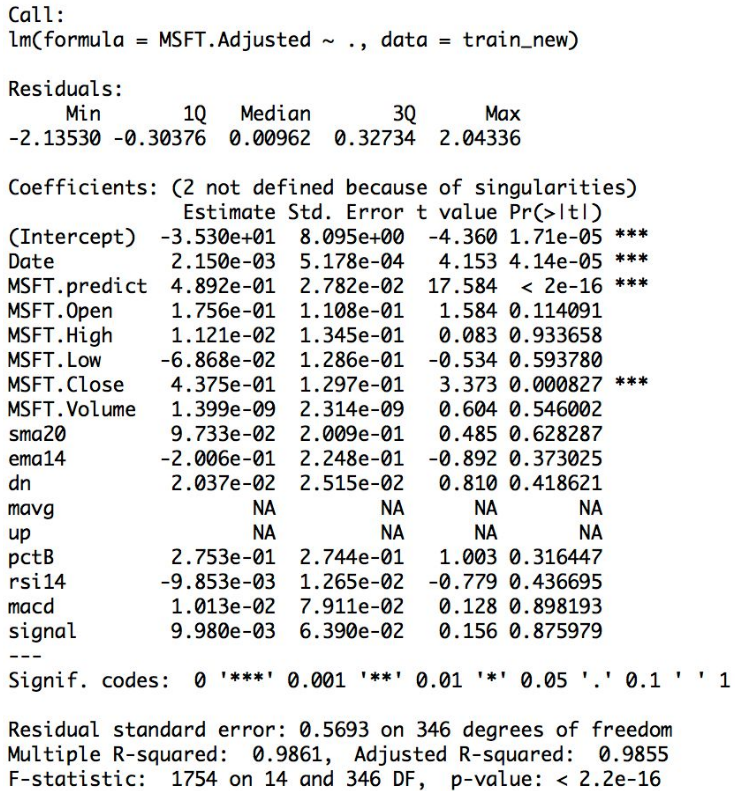
```
> accuracy(pred_mlr,test_new$BAC.Adjusted)
                ME      RMSE       MAE      MPE     MAPE
Test set 0.6101345 0.7622869 0.6437639 2.612553 2.834661
> 
```

Below is the plot showing the fit on the test data set. The Prediction plot of the Close Price vs Time. We can infer from the plot that the TA indicators we choose give a better fit to the test data in comparison to SLR.



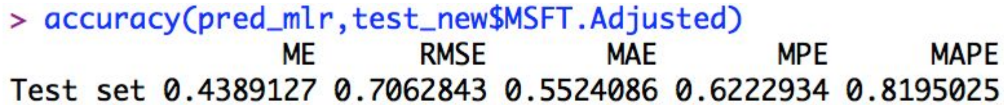
2) **Microsoft**

Below is the summary for Multiple Linear regression on Microsoft, where it hard to determine the significant contributors from the p-value. The R-squared : 98.61% of the variation in Y (MSFT.ADJUSTED) can be described using X-matrix (MSFT.PREDICT+TA INDICATORS).



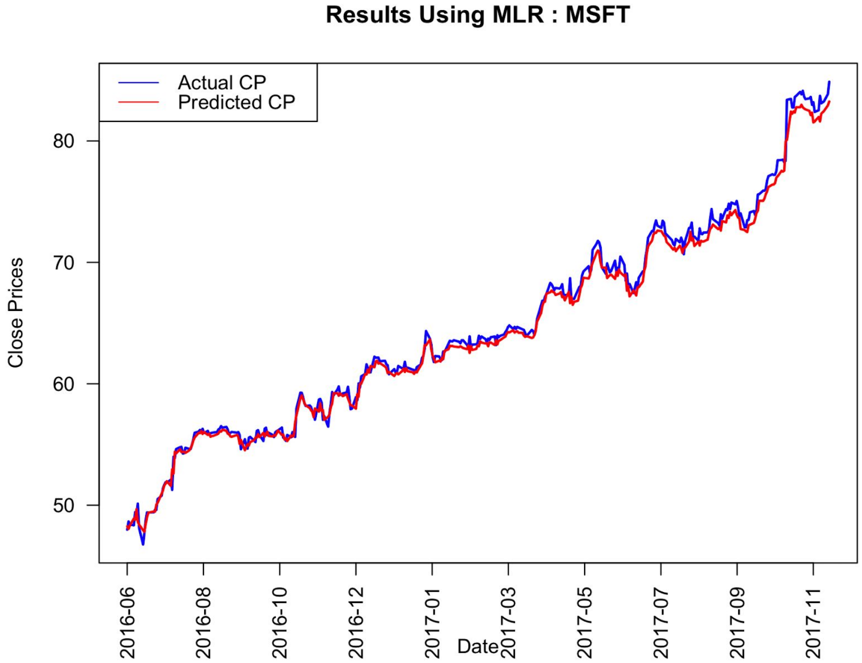
```
Call:
lm(formula = MSFT.Adjusted ~ ., data = train_new)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.13530 -0.30376  0.00962  0.32734  2.04336 

Coefficients: (2 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -3.530e+01  8.095e+00  -4.360 1.71e-05 ***
Date          2.150e-03  5.178e-04   4.153 4.14e-05 ***
MSFT.predict  4.892e-01  2.782e-02  17.584  < 2e-16 ***
MSFT.Open     1.756e-01  1.108e-01   1.584 0.114091    
MSFT.High     1.121e-02  1.345e-01   0.083 0.933658    
MSFT.Low     -6.868e-02  1.286e-01  -0.534 0.593780    
MSFT.Close    4.375e-01  1.297e-01   3.373 0.000827 ***
MSFT.Volume   1.399e-09  2.314e-09   0.604 0.546002    
sma20         9.733e-02  2.009e-01   0.485 0.628287    
ema14        -2.006e-01  2.248e-01  -0.892 0.373025    
dn            2.037e-02  2.515e-02   0.810 0.418621    
mavg                 NA         NA      NA       NA    
up                   NA         NA      NA       NA    
pctB          2.753e-01  2.744e-01   1.003 0.316447    
rsi14        -9.853e-03  1.265e-02  -0.779 0.436695    
macd          1.013e-02  7.911e-02   0.128 0.898193    
signal        9.980e-03  6.390e-02   0.156 0.875979    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5693 on 346 degrees of freedom
Multiple R-squared:  0.9861,  Adjusted R-squared:  0.9855 
F-statistic:  1754 on 14 and 346 DF,  p-value: < 2.2e-16
```

Below we report the accuracy of the model , where **RMSE = 0.70**



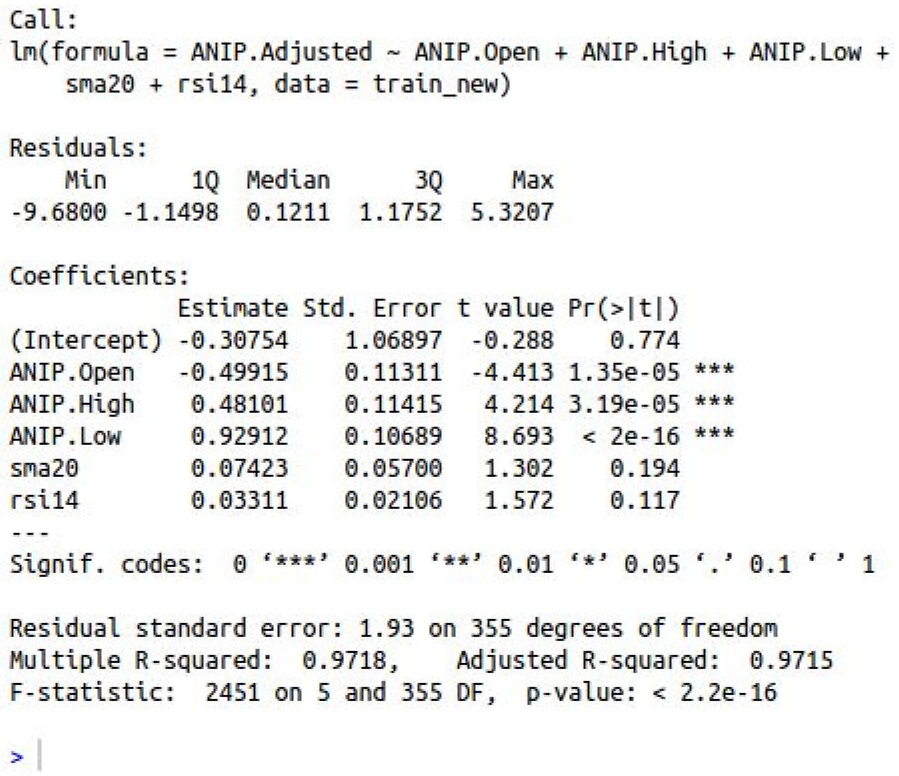
```
> accuracy(pred_mlr,test_new$MSFT.Adjusted)
                ME      RMSE       MAE       MPE      MAPE
Test set 0.4389127 0.7062843 0.5524086 0.6222934 0.8195025
```

Below is the plot showing the fit on the test data set. The Prediction plot of the Close Price vs Time. We can infer from the plot that the TA indicators we choose give a better fit to the test data in comparison to SLR.



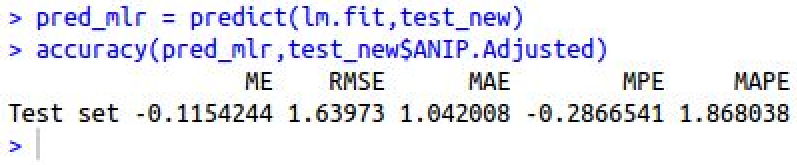
3) ANI Pharmaceuticals

Below is the summary for Multiple Linear regression on ANIP Pharmaceuticals, where ANIP.Low has p-value less than 0.05 which makes the ANIP.Low variable more significant. The R-squared : 97.18% of the variation in Y (ANIP.ADJUSTED) can be described using X-matrix (ANIP.PREDICT+TA INDICATORS).



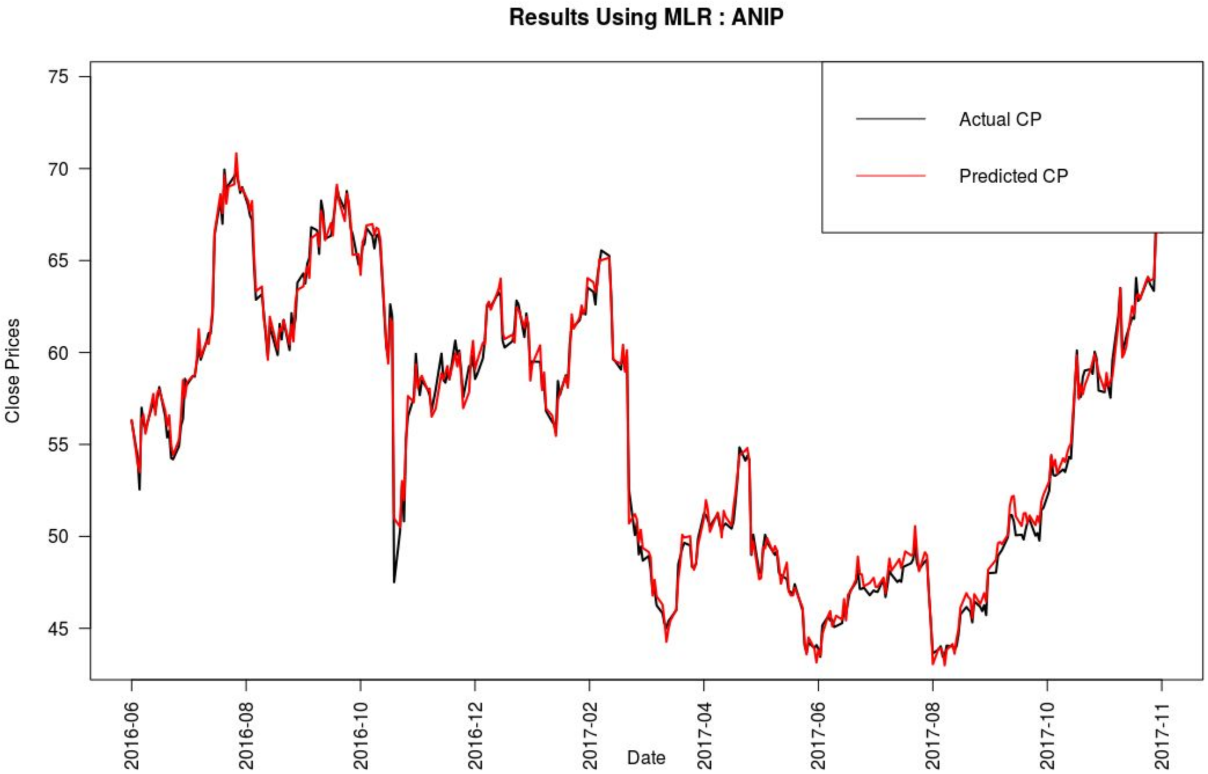
```
Call:
lm(formula = ANIP.Adjusted ~ ANIP.Open + ANIP.High + ANIP.Low + 
    sma20 + rsi14, data = train_new)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.6800 -1.1498  0.1211  1.1752  5.3207 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.30754    1.06897  -0.288    0.774    
ANIP.Open   -0.49915    0.11311  -4.413 1.35e-05 ***
ANIP.High    0.48101    0.11415   4.214 3.19e-05 ***
ANIP.Low     0.92912    0.10689   8.693  < 2e-16 ***
sma20        0.07423    0.05700   1.302    0.194    
rsi14        0.03311    0.02106   1.572    0.117    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.93 on 355 degrees of freedom
Multiple R-squared:  0.9718,	Adjusted R-squared:  0.9715 
F-statistic:  2451 on 5 and 355 DF,  p-value: < 2.2e-16

> 
```

Below we are reporting its **RMSE= 1.64** and other accuracy parameters. This shows that the accuracy of MLR is better than SLR.



```
> pred_mlr = predict(lm.fit,test_new)
> accuracy(pred_mlr,test_new$ANIP.Adjusted)
                 ME    RMSE      MAE        MPE     MAPE
Test set -0.1154244 1.63973 1.042008 -0.2866541 1.868038
> 
```

Below is the Prediction plot of the Close Price vs Time. We can infer from the plot that the TA indicators that we choose give a better fit to the test data.

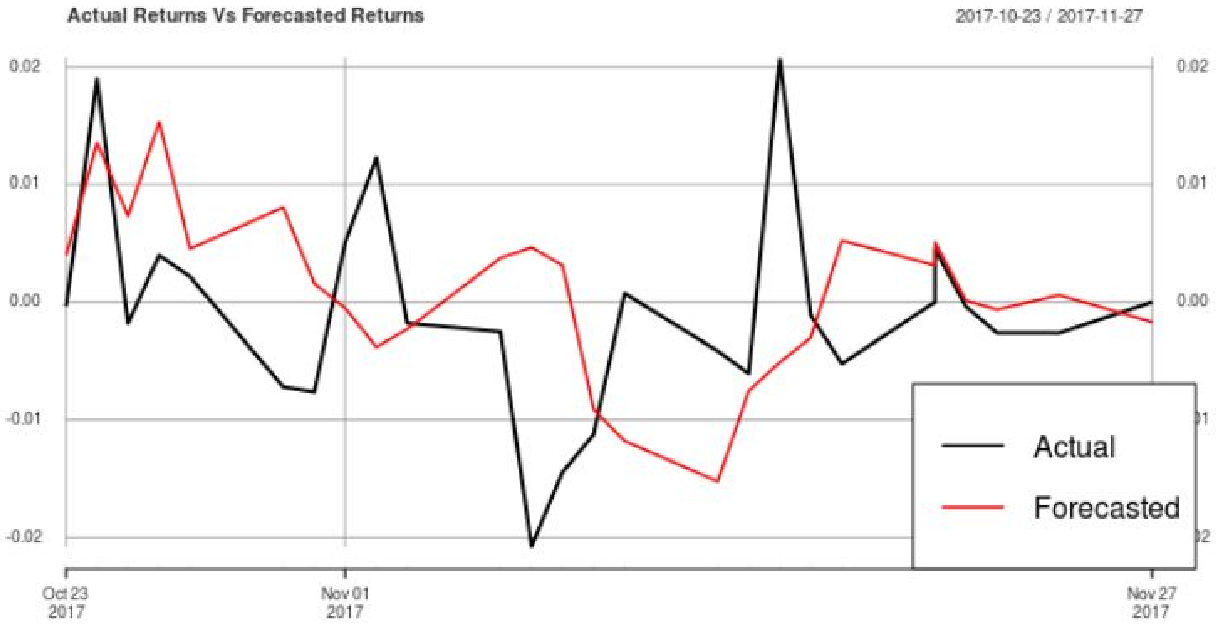


Method 3 : ARIMA

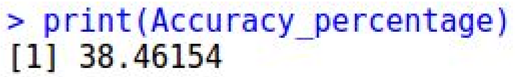
In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

1) Bank of America

   Below we are showing ARIMA results that are forecasted as the log returns.

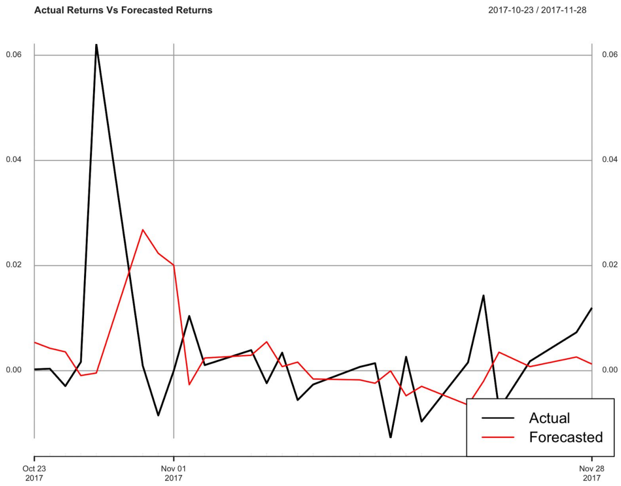


Below we are reporting the accuracy of the ARIMA.

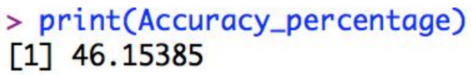


```
> print(Accuracy_percentage)
[1] 38.46154
```

2) Microsoft
   Below we are showing ARIMA results that are forecasted as the log returns.



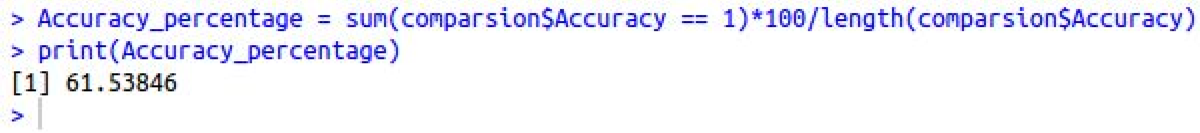
Below we are reporting the accuracy of the ARIMA as **46.15.**



```
> print(Accuracy_percentage)
[1] 46.15385
```

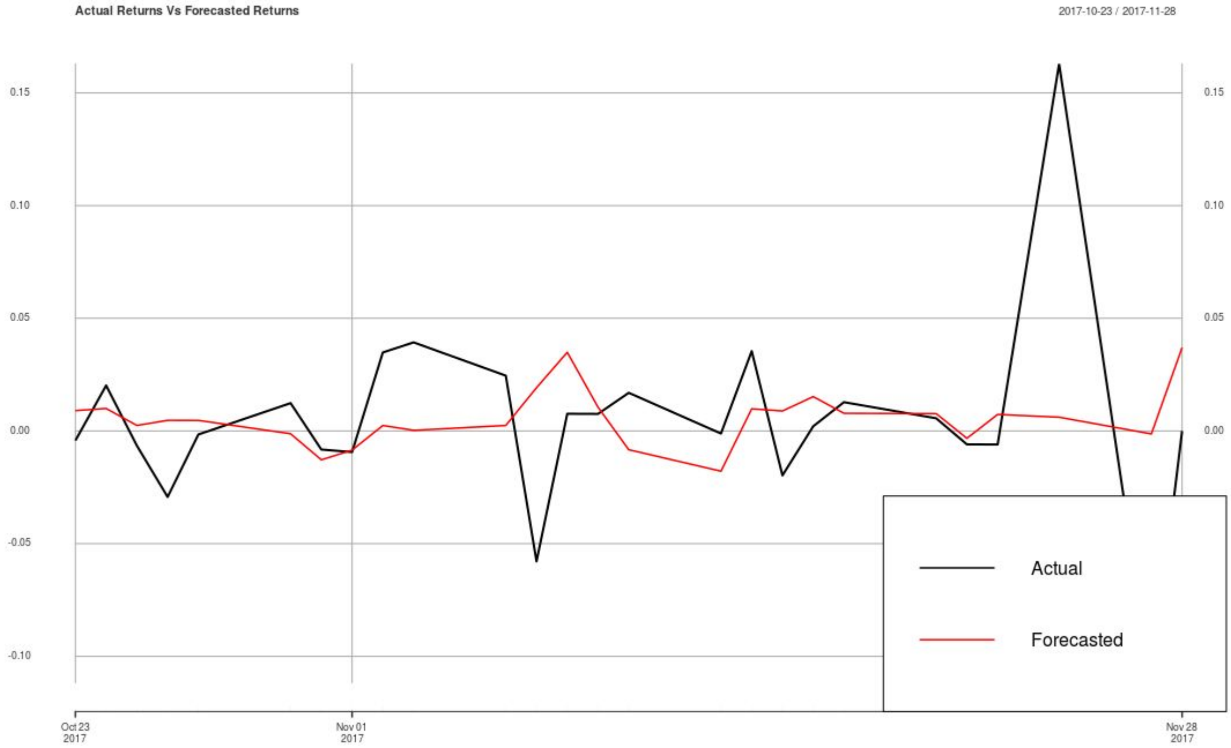
3) **ANI Pharmaceuticals**

Below is the accuracy for ARIMA.



```
> Accuracy_percentage = sum(comparsion$Accuracy == 1)*100/length(comparsion$Accuracy)
> print(Accuracy_percentage)
[1] 61.53846
> 
```

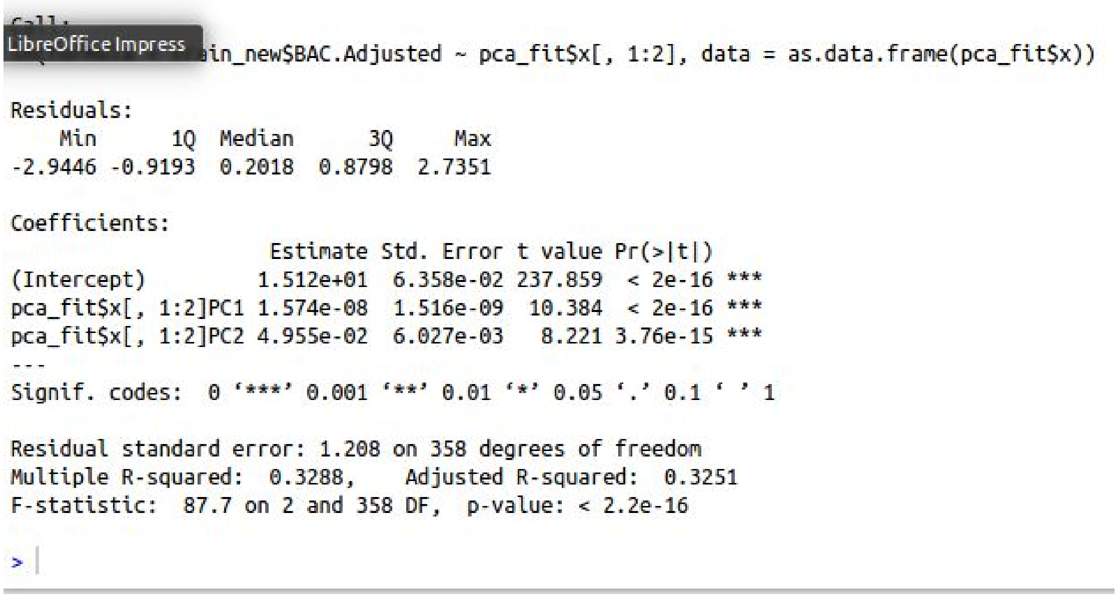
Below we are showing ARIMA results that are forecasted as the log returns.



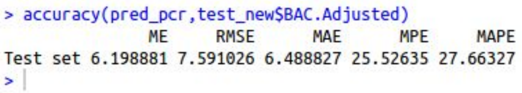
Method 4 : Dimension Reduction PCA

In below experiments we used the same data frame as in MLR and applied PCA to reduce the dimensions. We just use the first two PCA's as X-matrix and then build a regression model. Below are the results for the same.

1) Bank of America

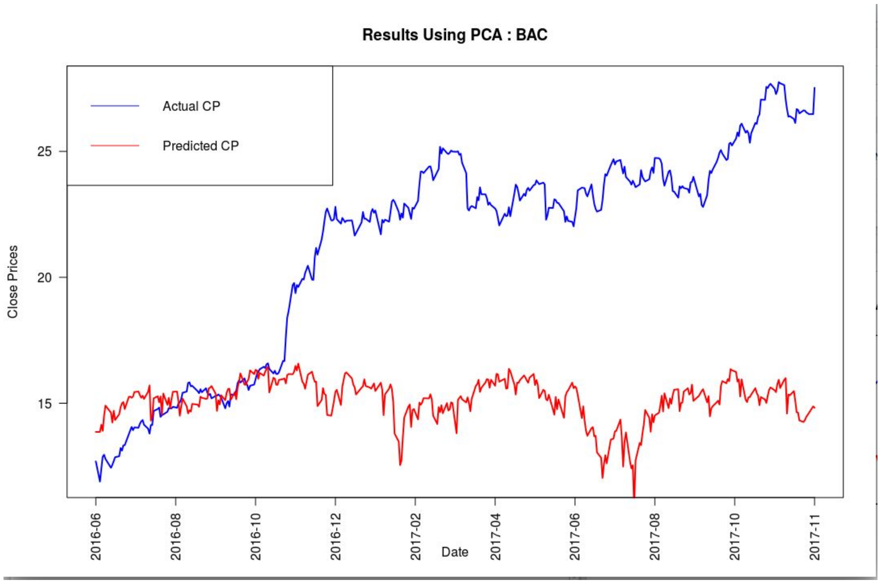


```
Call:
          ain_new$BAC.Adjusted ~ pca_fit$x[, 1:2], data = as.data.frame(pca_fit$x))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9446 -0.9193  0.2018  0.8798  2.7351 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.512e+01  6.358e-02 237.859  < 2e-16 ***
pca_fit$x[, 1:2]PC1 1.574e-08  1.516e-09  10.384  < 2e-16 ***
pca_fit$x[, 1:2]PC2 4.955e-02  6.027e-03   8.221 3.76e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.208 on 358 degrees of freedom
Multiple R-squared:  0.3288,	Adjusted R-squared:  0.3251 
F-statistic:  87.7 on 2 and 358 DF,  p-value: < 2.2e-16

>
```

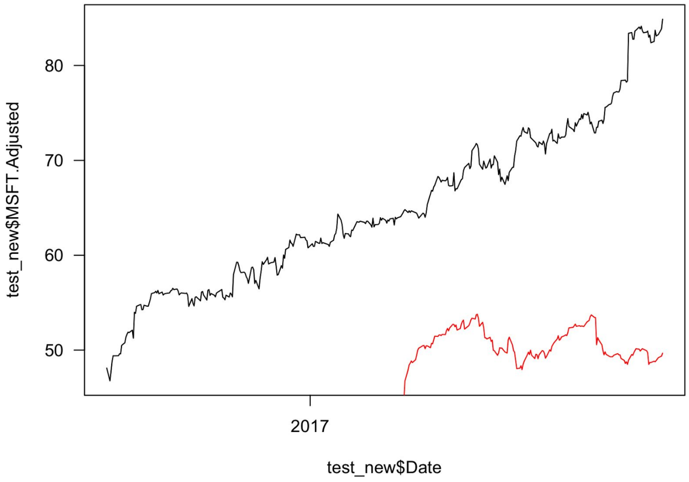


```
> accuracy(pred_pcr,test_new$BAC.Adjusted)
               ME     RMSE      MAE      MPE     MAPE
Test set 6.198881 7.591026 6.488827 25.52635 27.66327
>
```

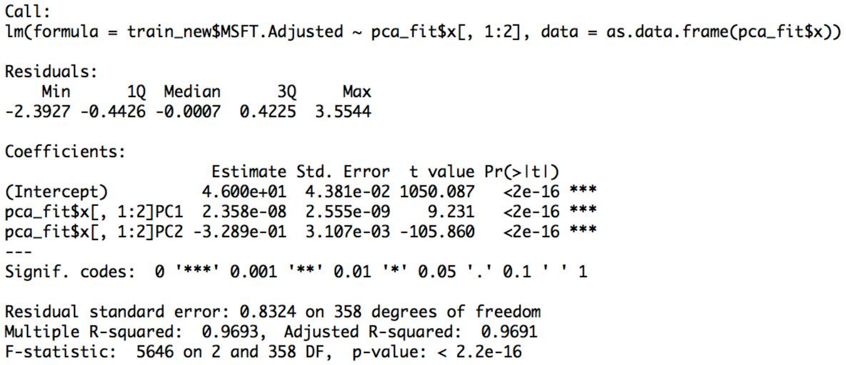
Below is the Prediction plot of the Close Price vs Time.



2) Microsoft

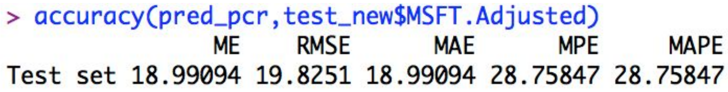


Below we can see that both the components of the PCA i.e PC1 and PC2 are significant as p-value is too low.



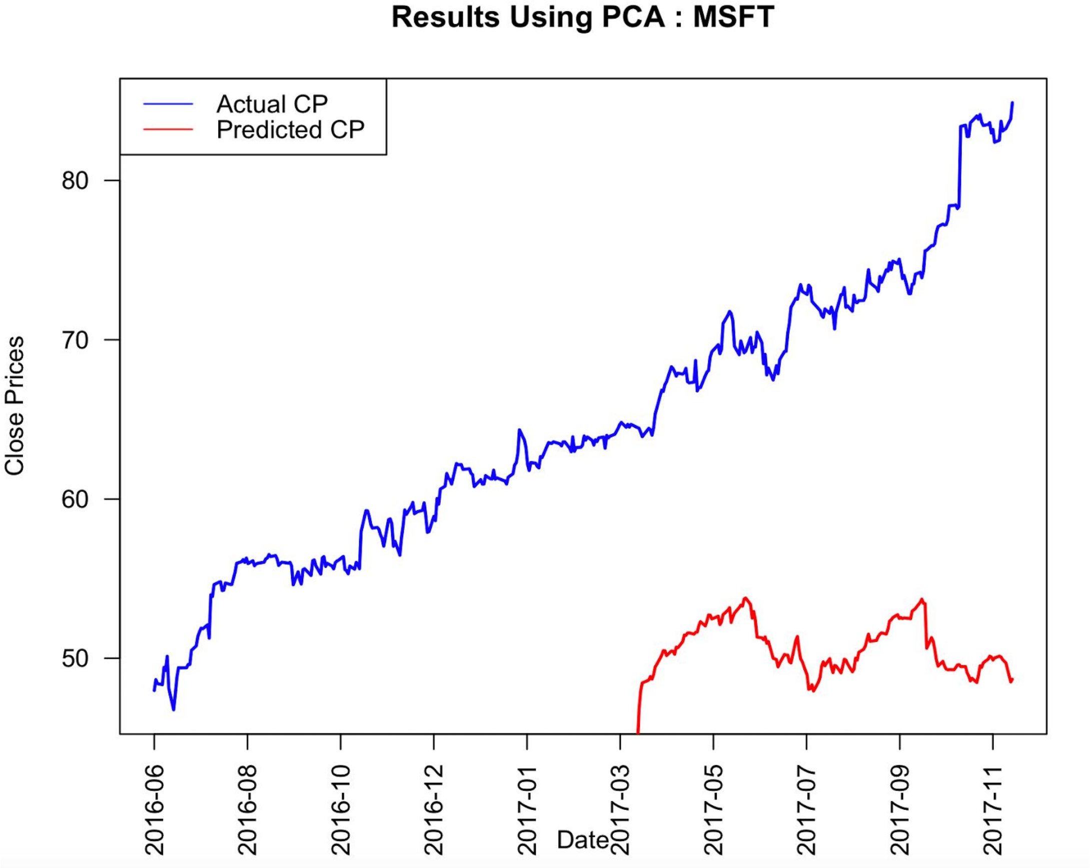
```
Call:
lm(formula = train_new$MSFT.Adjusted ~ pca_fit$x[, 1:2], data = as.data.frame(pca_fit$x))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3927 -0.4426 -0.0007  0.4225  3.5544 

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)    
(Intercept)          4.600e+01  4.381e-02 1050.087   <2e-16 ***
pca_fit$x[, 1:2]PC1  2.358e-08  2.555e-09    9.231   <2e-16 ***
pca_fit$x[, 1:2]PC2 -3.289e-01  3.107e-03 -105.860   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8324 on 358 degrees of freedom
Multiple R-squared:  0.9693,	Adjusted R-squared:  0.9691 
F-statistic:  5646 on 2 and 358 DF,  p-value: < 2.2e-16
```

Below we are reporting the RMSE = 19.82

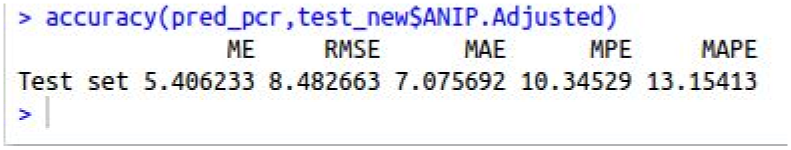


```
> accuracy(pred_pcr,test_new$MSFT.Adjusted)
               ME    RMSE      MAE      MPE     MAPE
Test set 18.99094 19.8251 18.99094 28.75847 28.75847
```

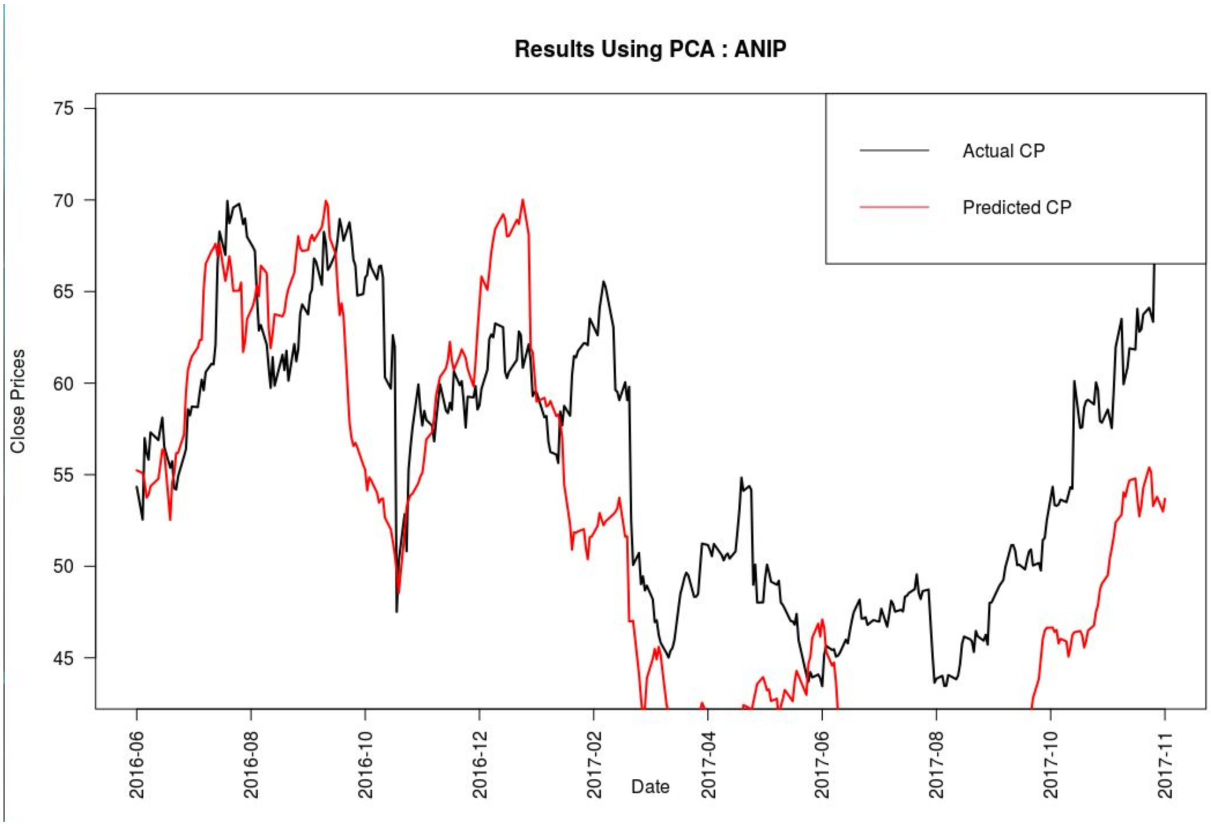
Below is the Prediction plot of the Close Price vs Time.



3) ANI Pharmaceuticals
Below we are using 2 components to fit the regression model and we can see only PC1 and PC2 is significant. And hence the accuracy is reported with **RMSE = 8.48.**



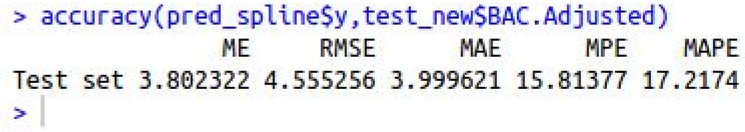
```
> accuracy(pred_pcr,test_new$ANIP.Adjusted)
                ME     RMSE      MAE      MPE     MAPE
Test set 5.406233 8.482663 7.075692 10.34529 13.15413
> 
```

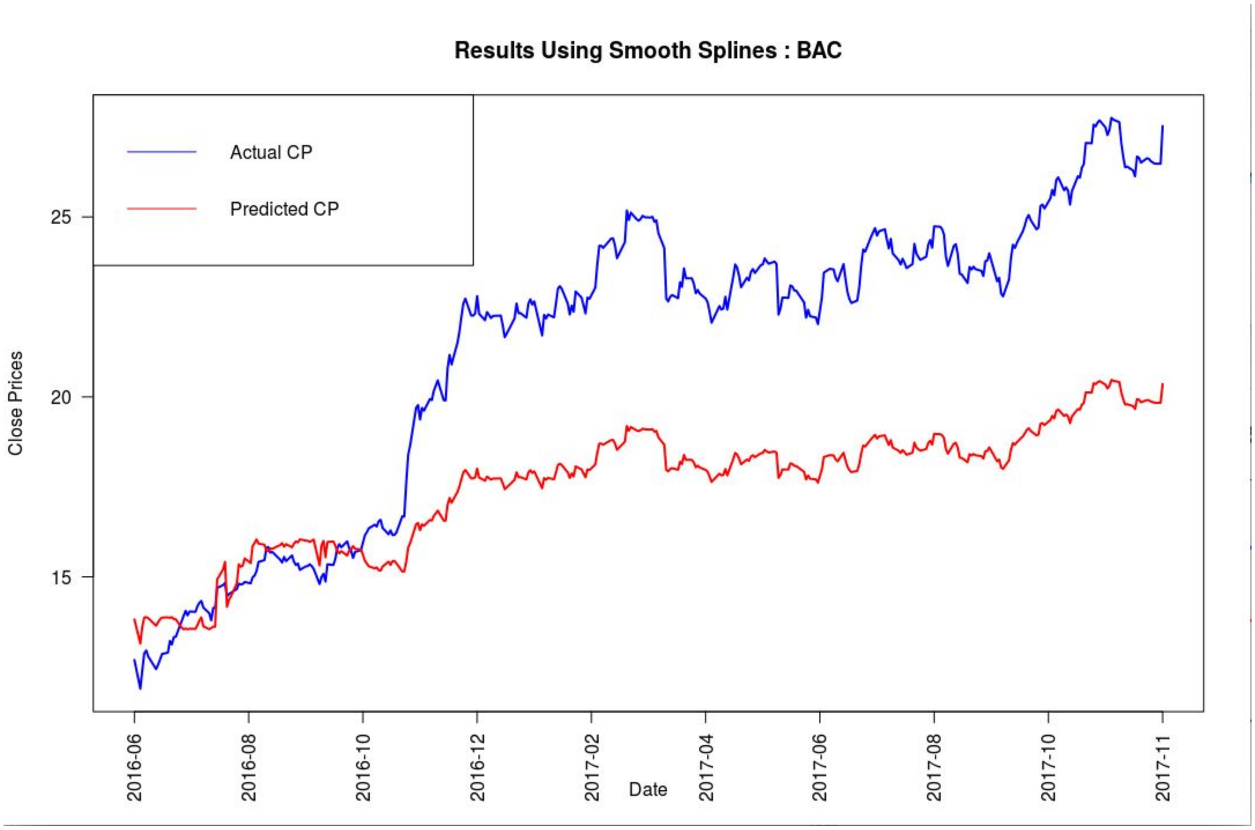


Method 5 : Splines

We applied smooth splines with cross validation equals to 10 to automatically choose the best number of knots. And then we build regression model. Below are the results of the same.

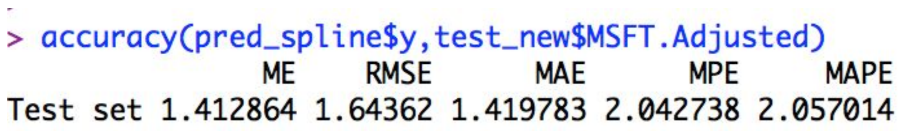
1)Bank of America


```
> accuracy(pred_spline$y,test_new$BAC.Adjusted)
               ME     RMSE      MAE      MPE    MAPE
Test set 3.802322 4.555256 3.999621 15.81377 17.2174
> 
```



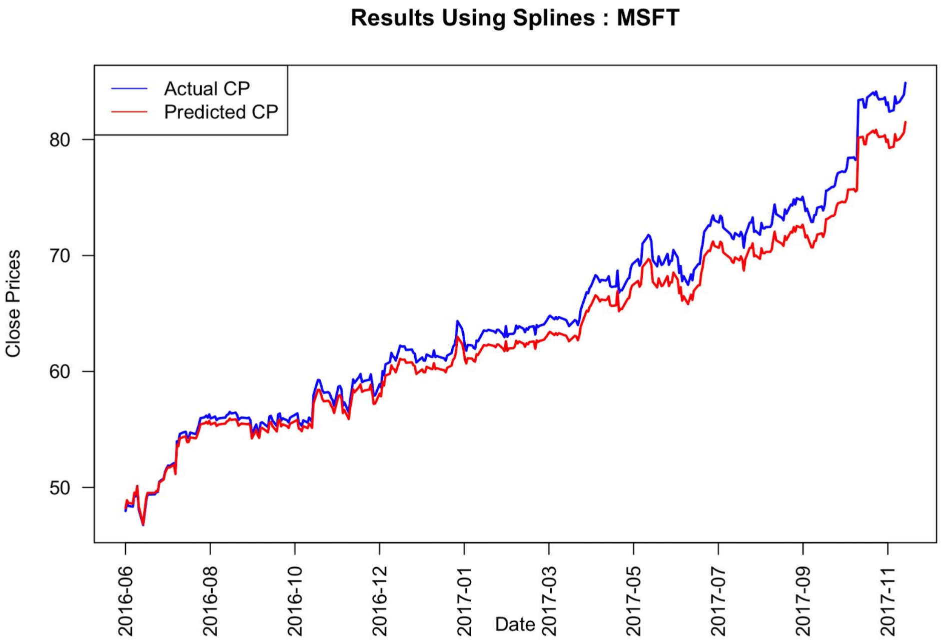
2)Microsoft

Below we are reporting **RMSE = 1.64.**

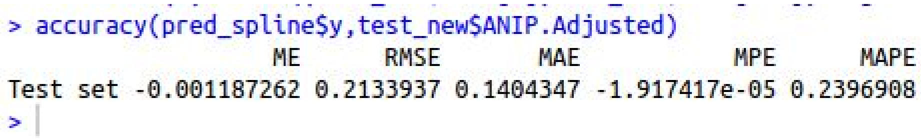


```
> accuracy(pred_spline$y,test_new$MSFT.Adjusted)
                ME    RMSE      MAE      MPE     MAPE
Test set 1.412864 1.64362 1.419783 2.042738 2.057014
```

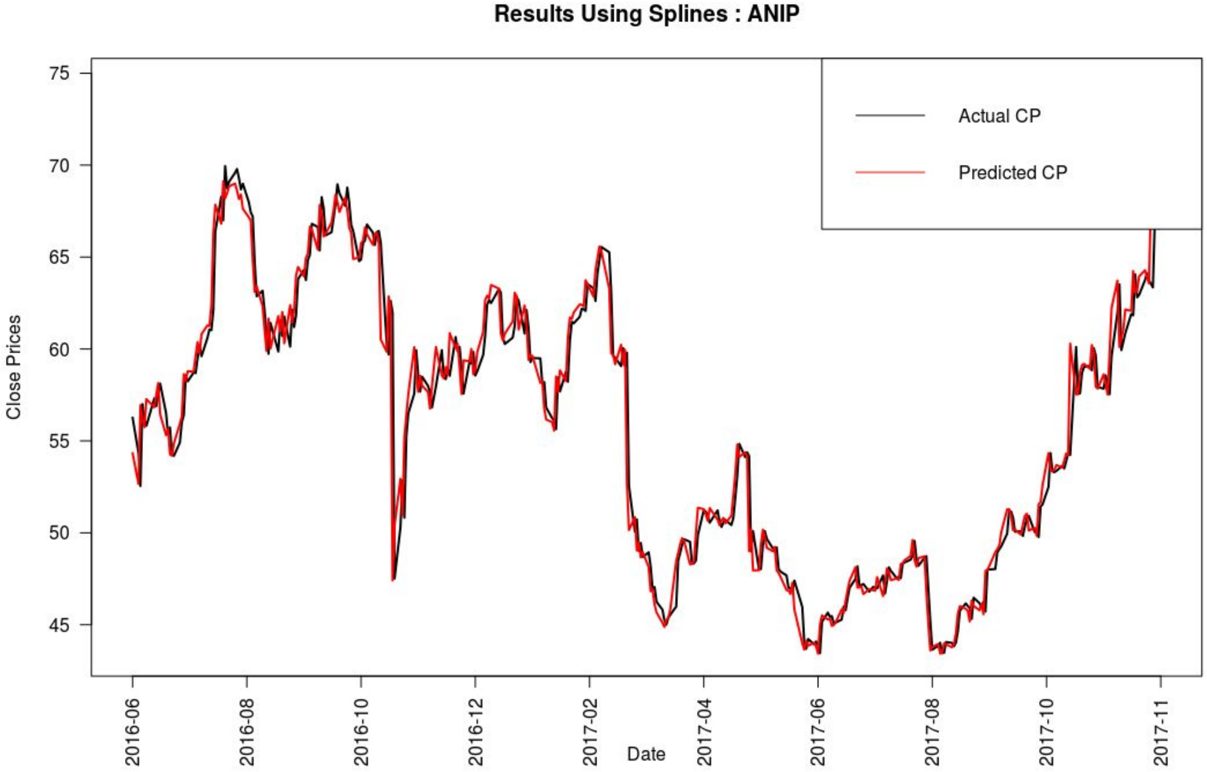
From the plot below we can define that the



3)ANI Pharmaceuticals



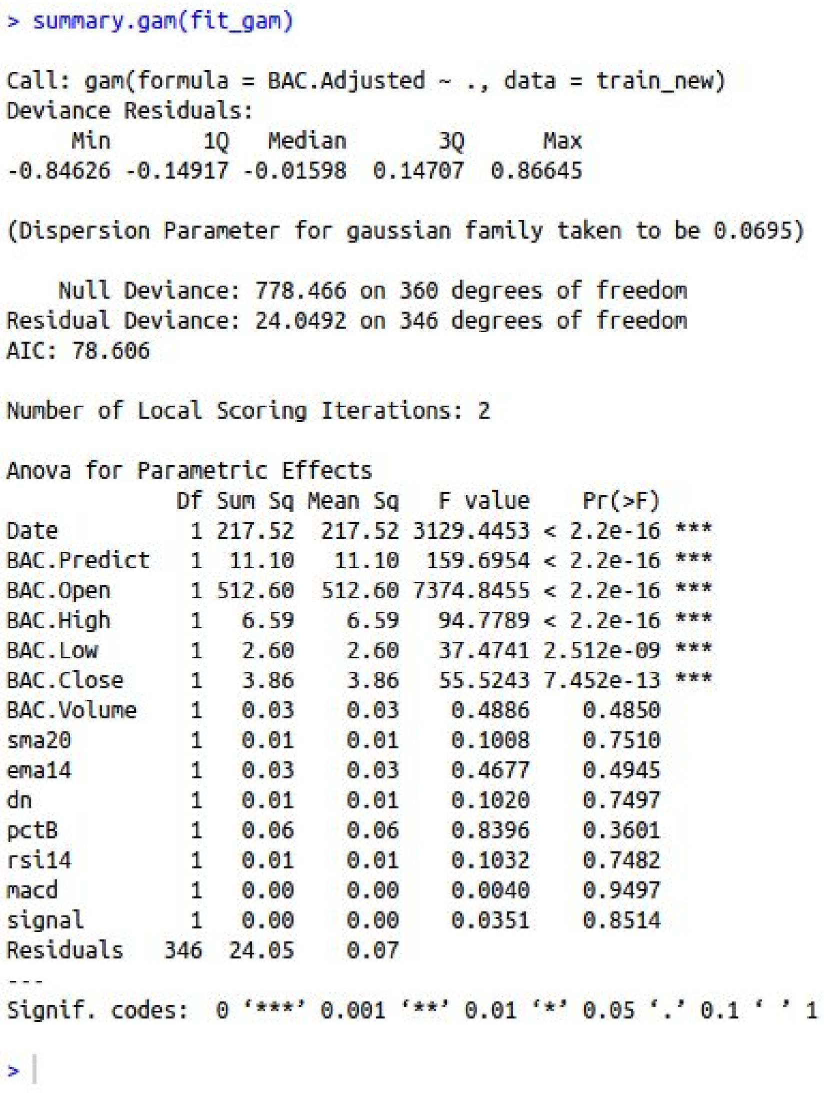
```
> accuracy(pred_spline$y,test_new$ANIP.Adjusted)
                   ME      RMSE       MAE           MPE      MAPE
Test set -0.001187262 0.2133937 0.1404347 -1.917417e-05 0.2396908
> 
```



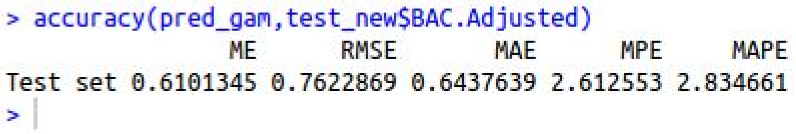
Method 6 : GAM

GAM is an additive modeling technique where the impact of the predictive variables is captured through smooth functions which depending on the underlying patterns in the data. when the model contains nonlinear effects, GAM can provide a regularized and interpretable solution.

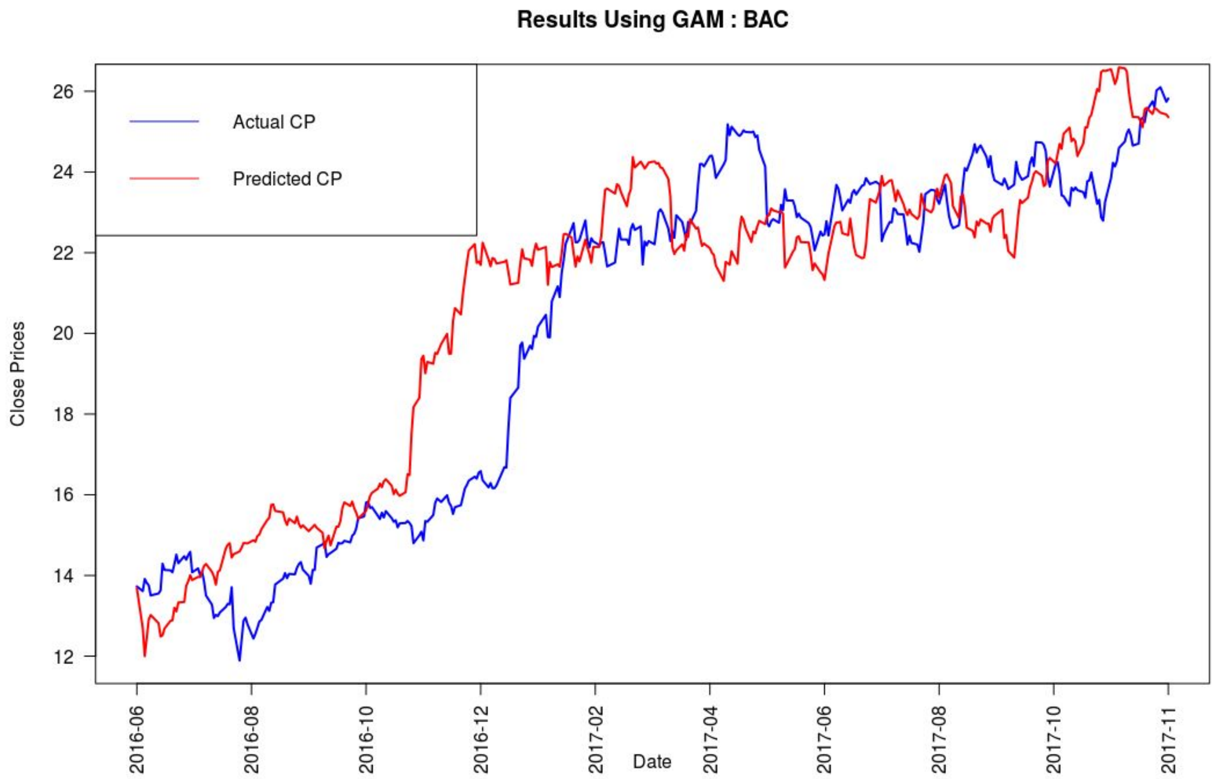
1) Bank of America



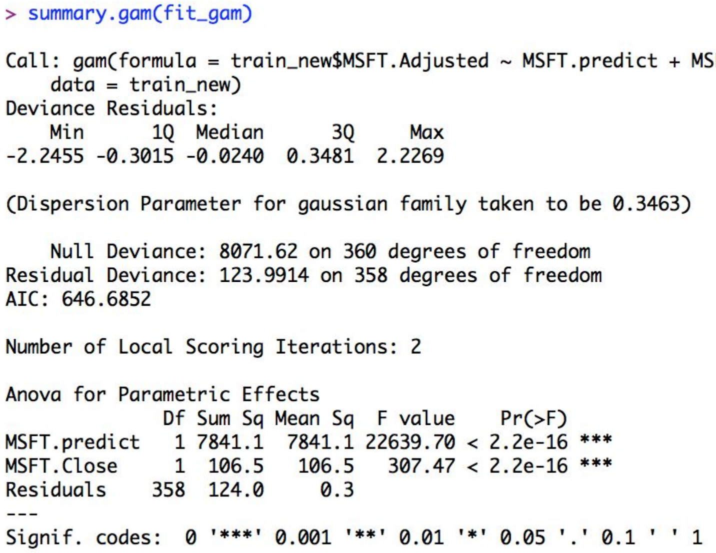
```
> summary.gam(fit_gam)

Call: gam(formula = BAC.Adjusted ~ ., data = train_new)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.84626 -0.14917 -0.01598  0.14707  0.86645

(Dispersion Parameter for gaussian family taken to be 0.0695)

    Null Deviance: 778.466 on 360 degrees of freedom
Residual Deviance: 24.0492 on 346 degrees of freedom
AIC: 78.606

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
             Df Sum Sq Mean Sq   F value    Pr(>F)
Date          1 217.52  217.52 3129.4453 < 2.2e-16 ***
BAC.Predict   1  11.10   11.10  159.6954 < 2.2e-16 ***
BAC.Open      1 512.60  512.60 7374.8455 < 2.2e-16 ***
BAC.High      1   6.59    6.59   94.7789 < 2.2e-16 ***
BAC.Low       1   2.60    2.60   37.4741 2.512e-09 ***
BAC.Close     1   3.86    3.86   55.5243 7.452e-13 ***
BAC.Volume    1   0.03    0.03    0.4886    0.4850
sma20         1   0.01    0.01    0.1008    0.7510
ema14         1   0.03    0.03    0.4677    0.4945
dn            1   0.01    0.01    0.1020    0.7497
pctB          1   0.06    0.06    0.8396    0.3601
rsi14         1   0.01    0.01    0.1032    0.7482
macd          1   0.00    0.00    0.0040    0.9497
signal        1   0.00    0.00    0.0351    0.8514
Residuals   346  24.05    0.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> 
```



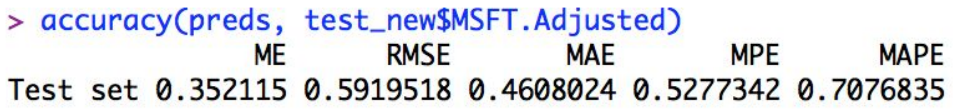
```
> accuracy(pred_gam,test_new$BAC.Adjusted)
                ME      RMSE       MAE      MPE     MAPE
Test set 0.6101345 0.7622869 0.6437639 2.612553 2.834661
> 
```


Results Using GAM : BAC

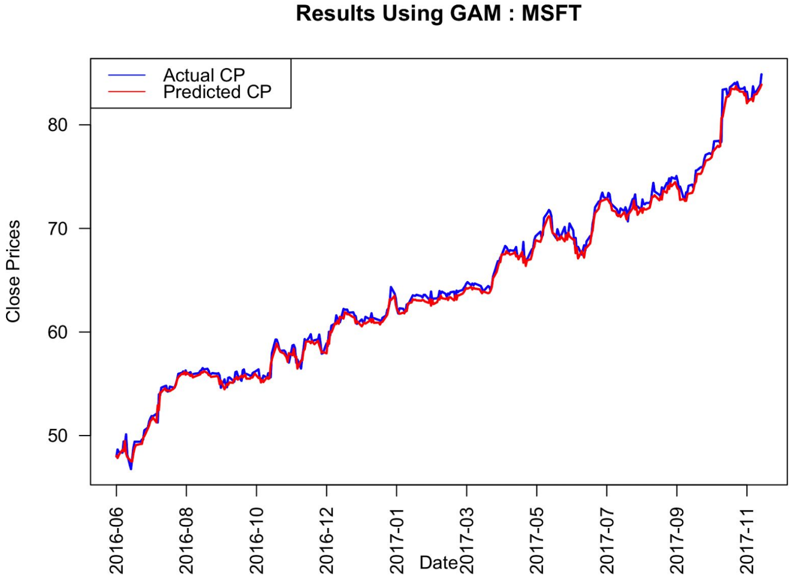
2)Microsoft



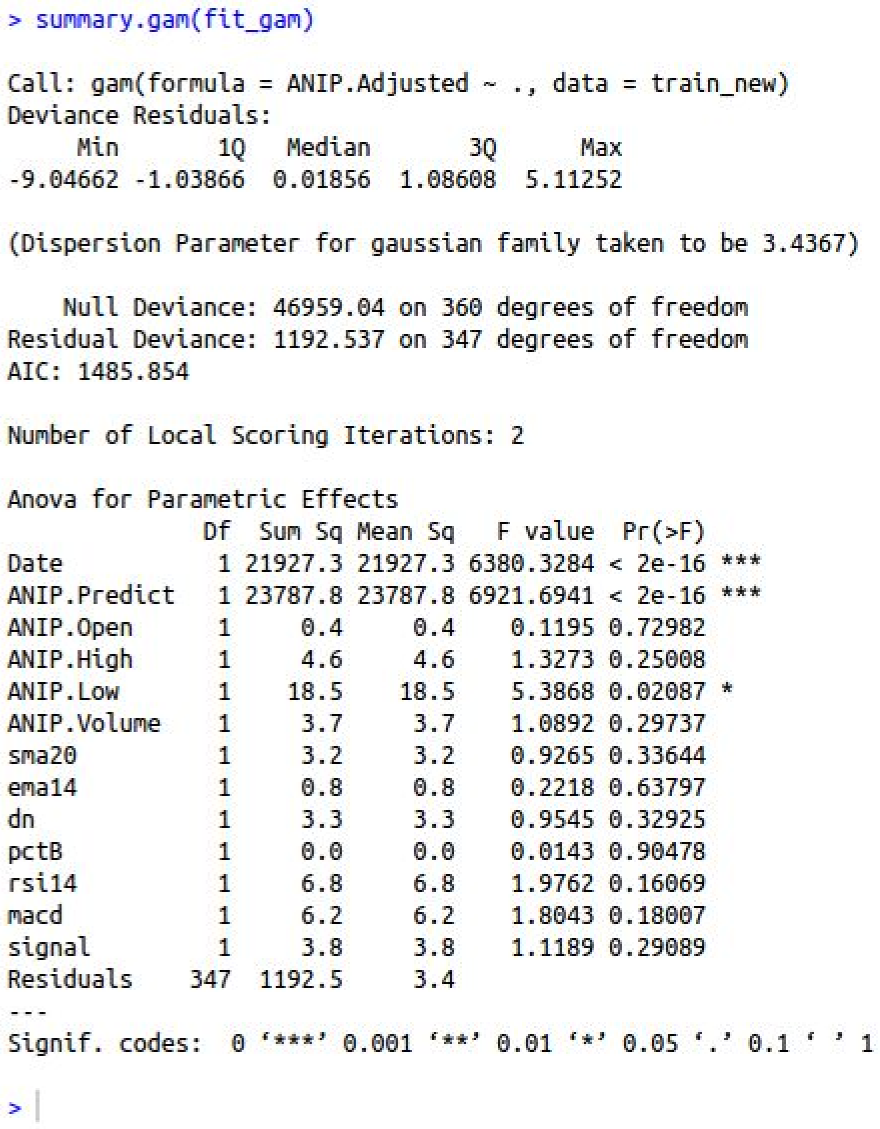
```
> summary.gam(fit_gam)

Call: gam(formula = train_new$MSFT.Adjusted ~ MSFT.predict + MSFT.Close,
    data = train_new)
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.2455 -0.3015 -0.0240  0.3481  2.2269

(Dispersion Parameter for gaussian family taken to be 0.3463)

    Null Deviance: 8071.62 on 360 degrees of freedom
Residual Deviance: 123.9914 on 358 degrees of freedom
AIC: 646.6852

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
              Df Sum Sq Mean Sq  F value    Pr(>F)
MSFT.predict   1 7841.1  7841.1 22639.70 < 2.2e-16 ***
MSFT.Close     1  106.5   106.5   307.47 < 2.2e-16 ***
Residuals    358  124.0     0.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



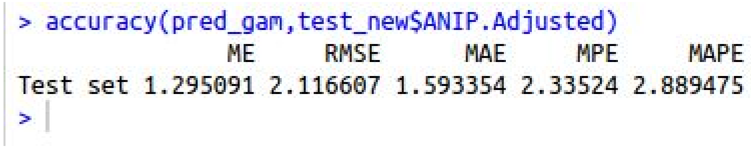
```
> accuracy(preds, test_new$MSFT.Adjusted)
               ME      RMSE       MAE       MPE      MAPE
Test set 0.352115 0.5919518 0.4608024 0.5277342 0.7076835
```



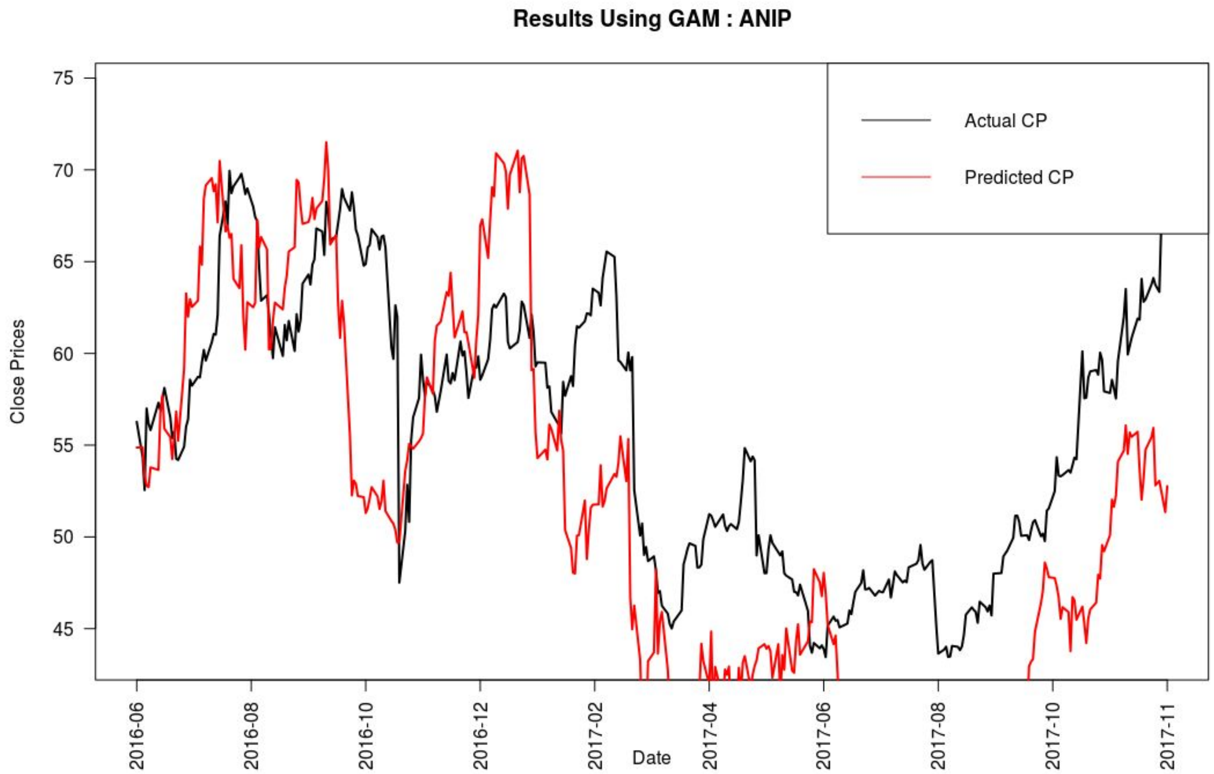
3) ANI Pharmaceuticals



```
> summary.gam(fit_gam)

Call: gam(formula = ANIP.Adjusted ~ ., data = train_new)
Deviance Residuals:
     Min       1Q   Median       3Q      Max 
-9.04662 -1.03866  0.01856  1.08608  5.11252 

(Dispersion Parameter for gaussian family taken to be 3.4367)

    Null Deviance: 46959.04 on 360 degrees of freedom
Residual Deviance: 1192.537 on 347 degrees of freedom
AIC: 1485.854 

Number of Local Scoring Iterations: 2 

Anova for Parametric Effects
              Df  Sum Sq Mean Sq   F value  Pr(>F)    
Date           1 21927.3 21927.3 6380.3284 < 2e-16 ***
ANIP.Predict   1 23787.8 23787.8 6921.6941 < 2e-16 ***
ANIP.Open      1     0.4     0.4    0.1195 0.72982    
ANIP.High      1     4.6     4.6    1.3273 0.25008    
ANIP.Low       1    18.5    18.5    5.3868 0.02087 *  
ANIP.Volume    1     3.7     3.7    1.0892 0.29737    
sma20          1     3.2     3.2    0.9265 0.33644    
ema14          1     0.8     0.8    0.2218 0.63797    
dn             1     3.3     3.3    0.9545 0.32925    
pctB           1     0.0     0.0    0.0143 0.90478    
rsi14          1     6.8     6.8    1.9762 0.16069    
macd           1     6.2     6.2    1.8043 0.18007    
signal         1     3.8     3.8    1.1189 0.29089    
Residuals    347  1192.5     3.4                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> 
```



```
> accuracy(pred_gam,test_new$ANIP.Adjusted)
               ME     RMSE      MAE     MPE     MAPE
Test set 1.295091 2.116607 1.593354 2.33524 2.889475
> 
```

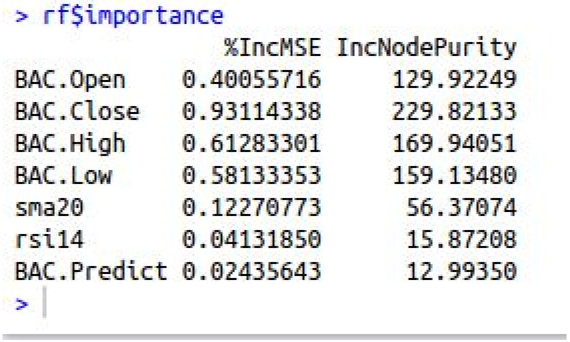


Method 7 : Random Forest

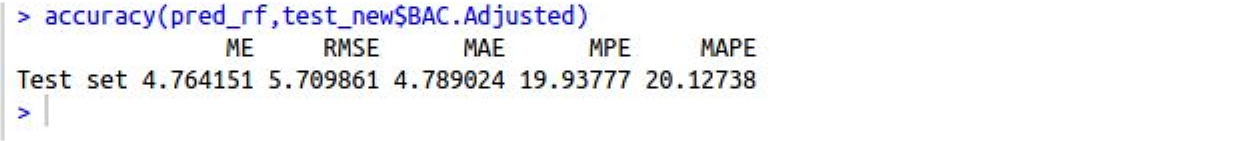
As a learning method, Random Forest constructs a multiple of decision trees and outputs different classes for classification or mean prediction for regression. It considers only a subset of the predictors at each split and lets other predictors have more of a chance except for the strong predictor, thereby making the results less variable and more reliable.

1) Bank of America

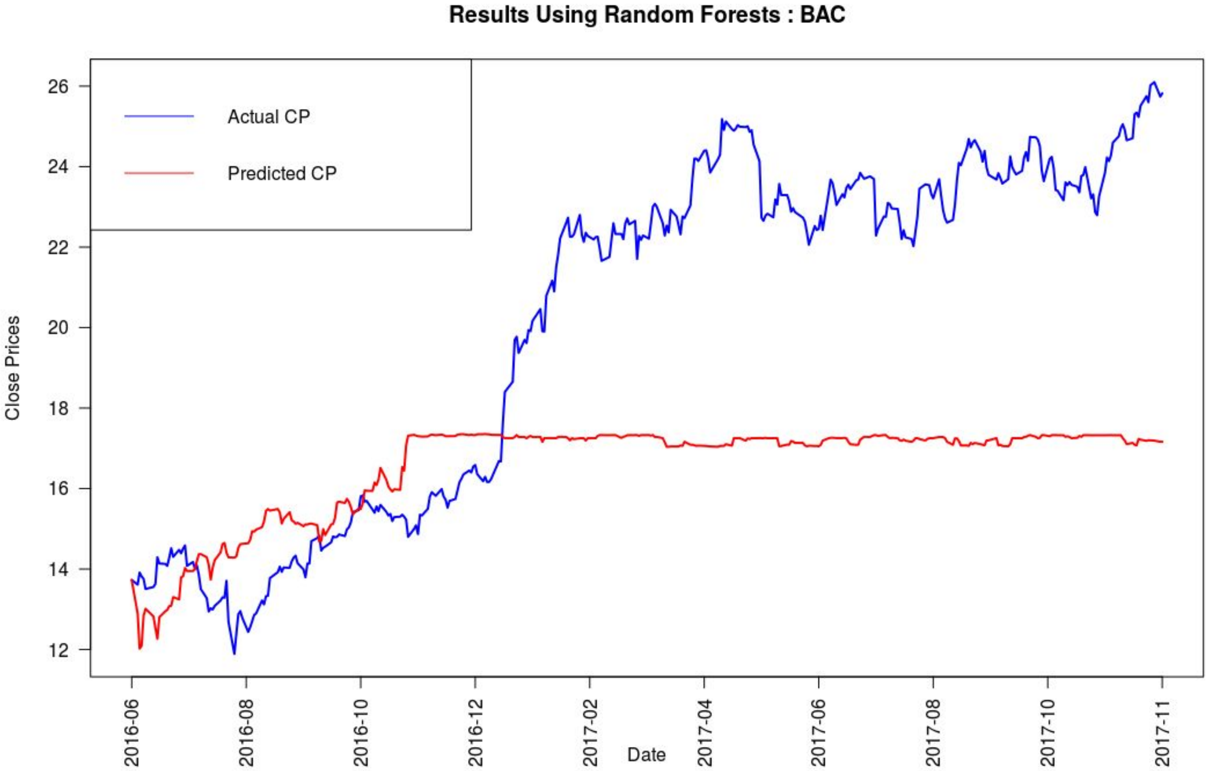
   Below RF importance evaluation for BOFA



```
> rf$importance
                %IncMSE IncNodePurity
BAC.Open    0.40055716     129.92249
BAC.Close   0.93114338     229.82133
BAC.High    0.61283301     169.94051
BAC.Low     0.58133353     159.13480
sma20       0.12270773      56.37074
rsi14       0.04131850      15.87208
BAC.Predict 0.02435643      12.99350
> 
```

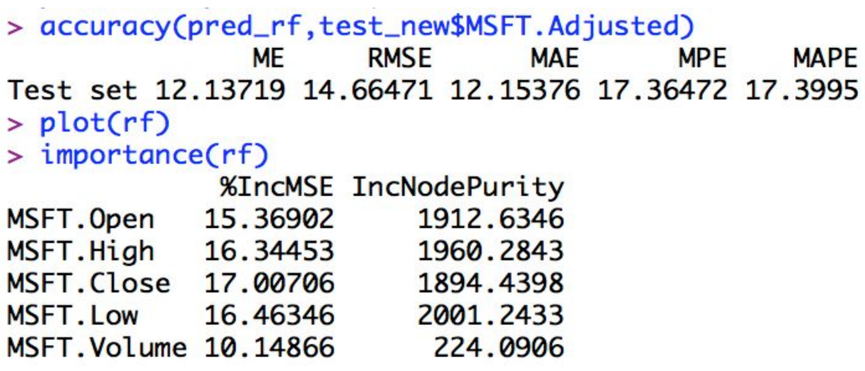


```
> accuracy(pred_rf,test_new$BAC.Adjusted)
               ME     RMSE      MAE      MPE     MAPE
Test set 4.764151 5.709861 4.789024 19.93777 20.12738
> 
```

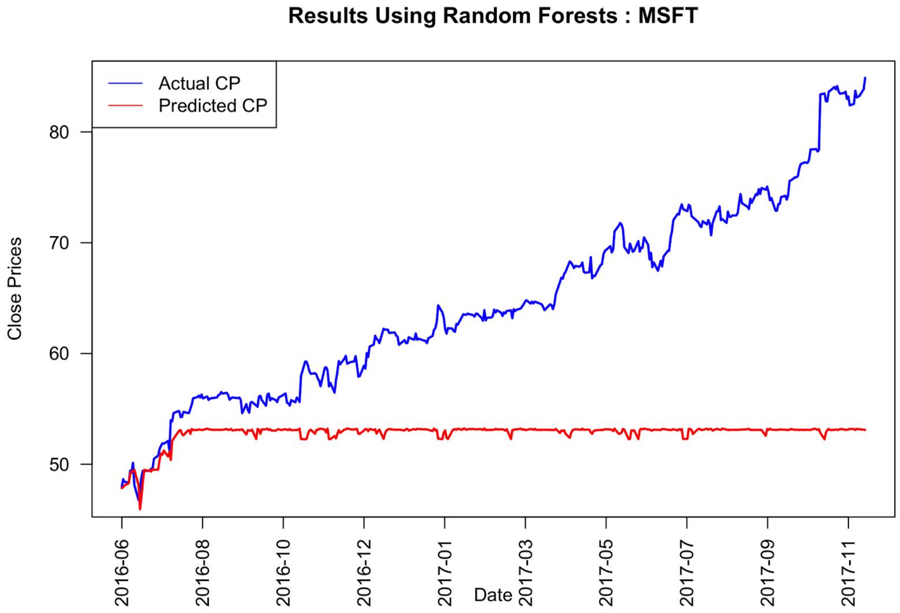


2)Microsoft

Below RF importance evaluation for MICROSOFT.

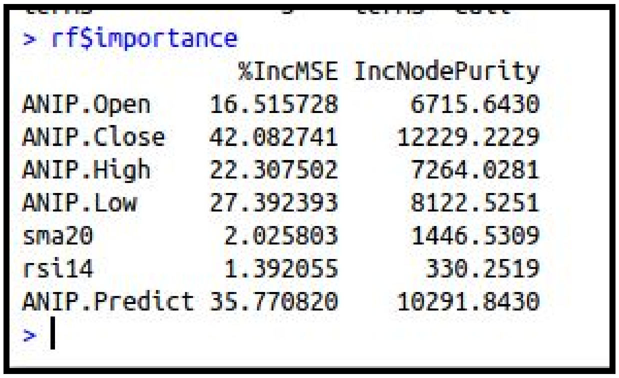


```
> accuracy(pred_rf,test_new$MSFT.Adjusted)
               ME     RMSE      MAE      MPE    MAPE
Test set 12.13719 14.66471 12.15376 17.36472 17.3995
> plot(rf)
> importance(rf)
              %IncMSE IncNodePurity
MSFT.Open    15.36902     1912.6346
MSFT.High    16.34453     1960.2843
MSFT.Close   17.00706     1894.4398
MSFT.Low     16.46346     2001.2433
MSFT.Volume  10.14866      224.0906
```

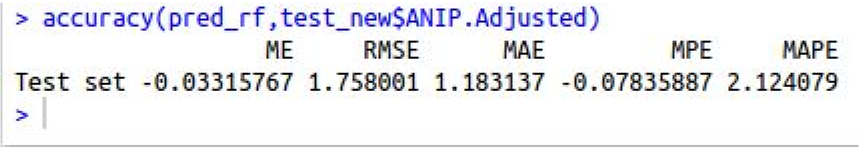


3)ANI Pharmaceuticals

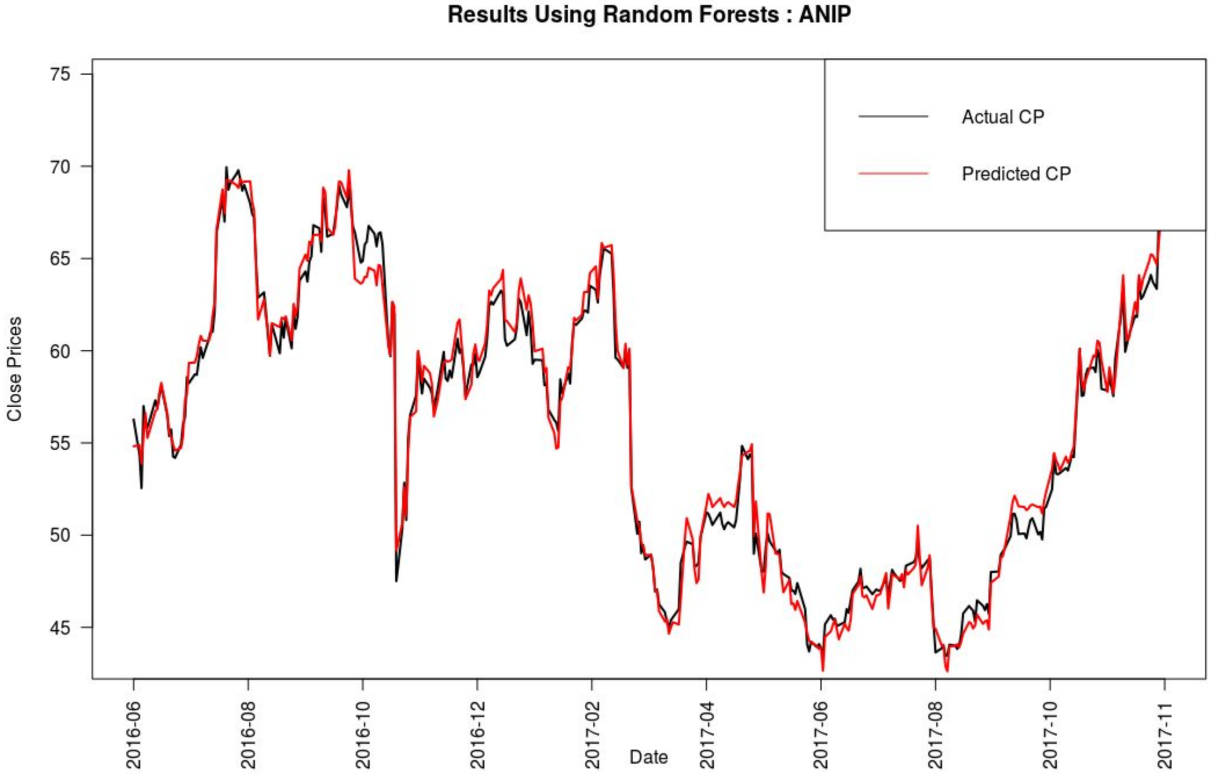
Below RF importance evaluation for ANIP.



```
> rf$importance
               %IncMSE IncNodePurity
ANIP.Open    16.515728     6715.6430
ANIP.Close   42.082741    12229.2229
ANIP.High    22.307502     7264.0281
ANIP.Low     27.392393     8122.5251
sma20         2.025803     1446.5309
rsi14         1.392055      330.2519
ANIP.Predict 35.770820    10291.8430
> 
```



```
> accuracy(pred_rf,test_new$ANIP.Adjusted)
                  ME     RMSE      MAE         MPE     MAPE
Test set -0.03315767 1.758001 1.183137 -0.07835887 2.124079
> 
```



Results Using Random Forests : ANIP

Portfolio Analysis

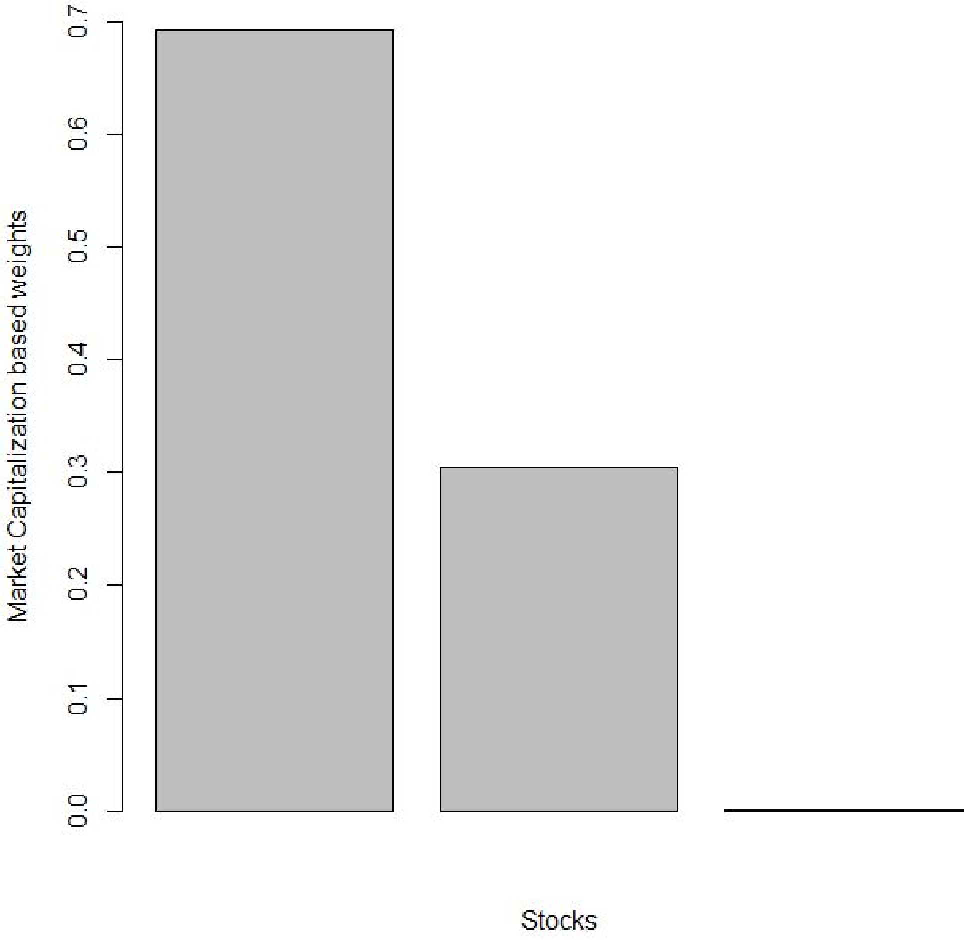
In this section, we shift our focus to the investment aspect of our project. Through our online survey mentioned before, we were able to gather deep insights on how graduate students weigh their investment objectives. Based on our survey,

- 53.8% of the students were willing to invest not more than $500
- 61.5% of the students wanted to invest for long term (Our definition of long term was relative to the time graduate students, on an average, spend in college)
- 80.8% of the students would rely on past performance of the market to make investments
- 59.6% of the students would prefer a low risk investment

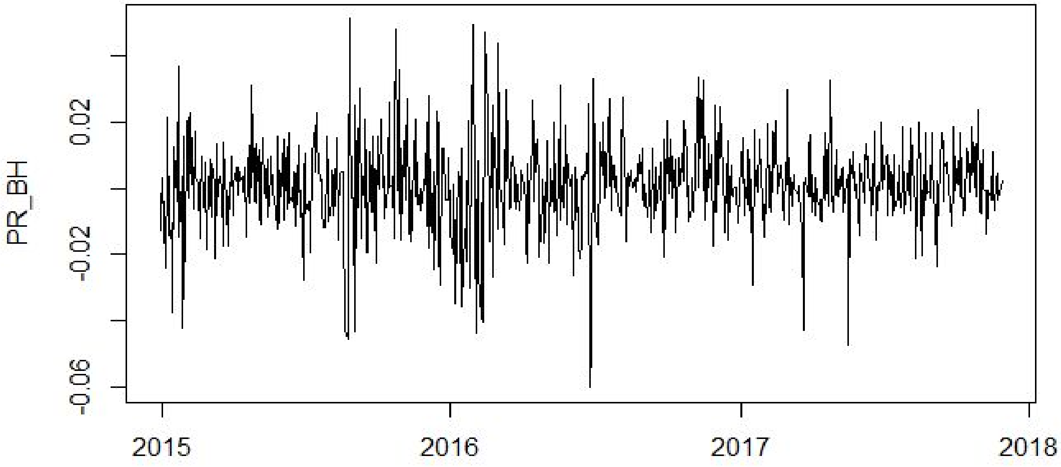
Based on these outcomes, we selected our stocks, adjusted risk tolerance and selected a time horizon of daily as well as monthly rebalancing.

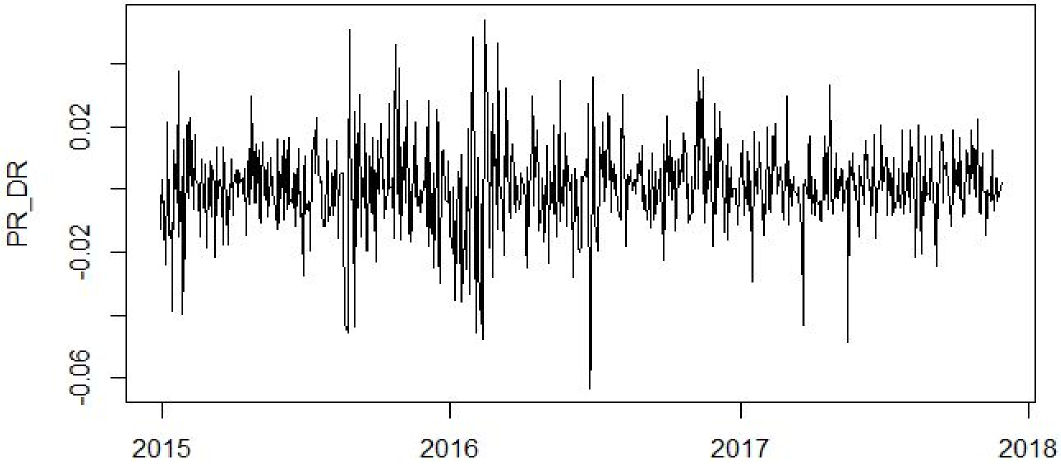
Following were the results obtained from our analysis on stock trends and momentum:

- Portfolio weights based on Market Capitalization used in the portfolio

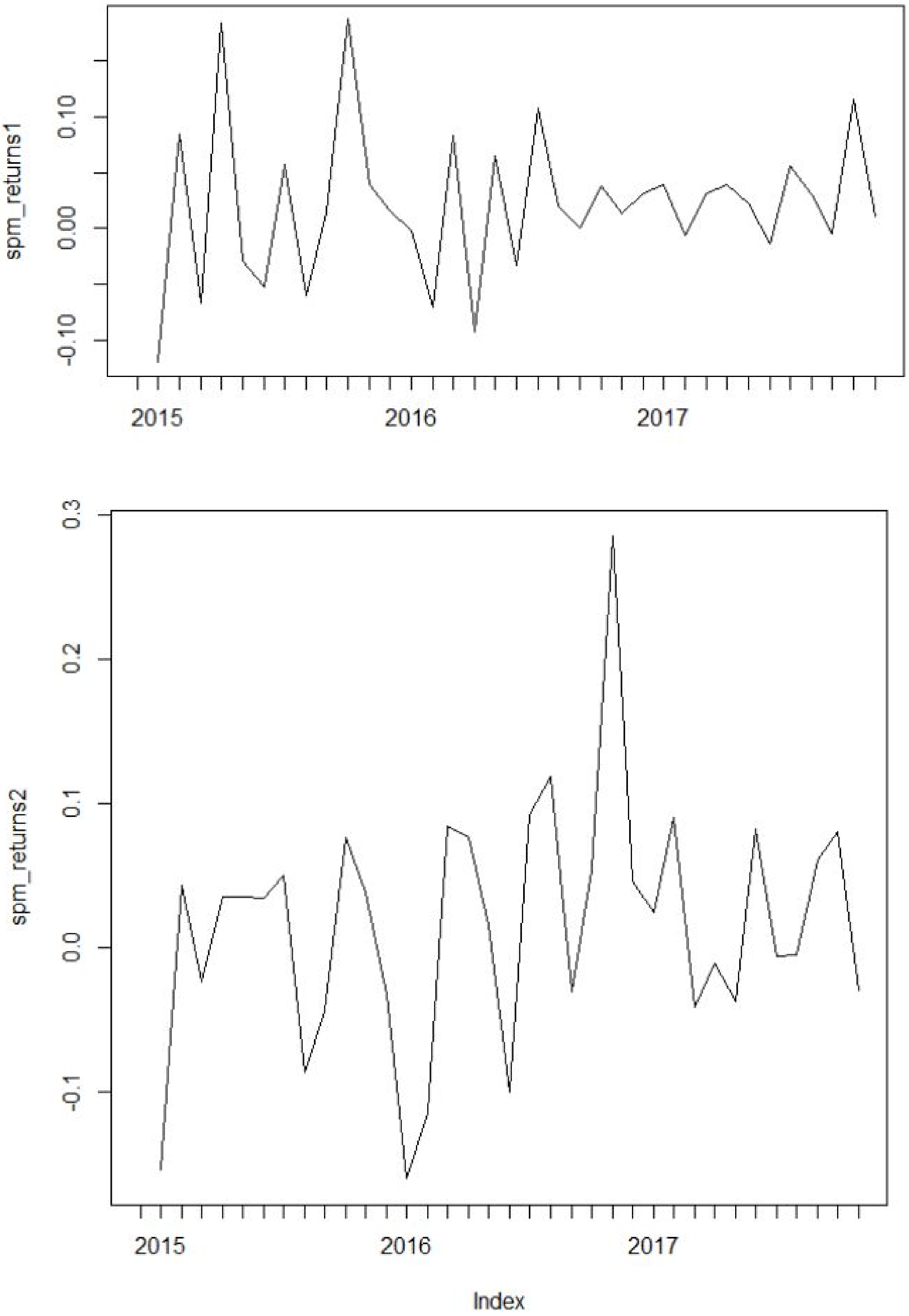


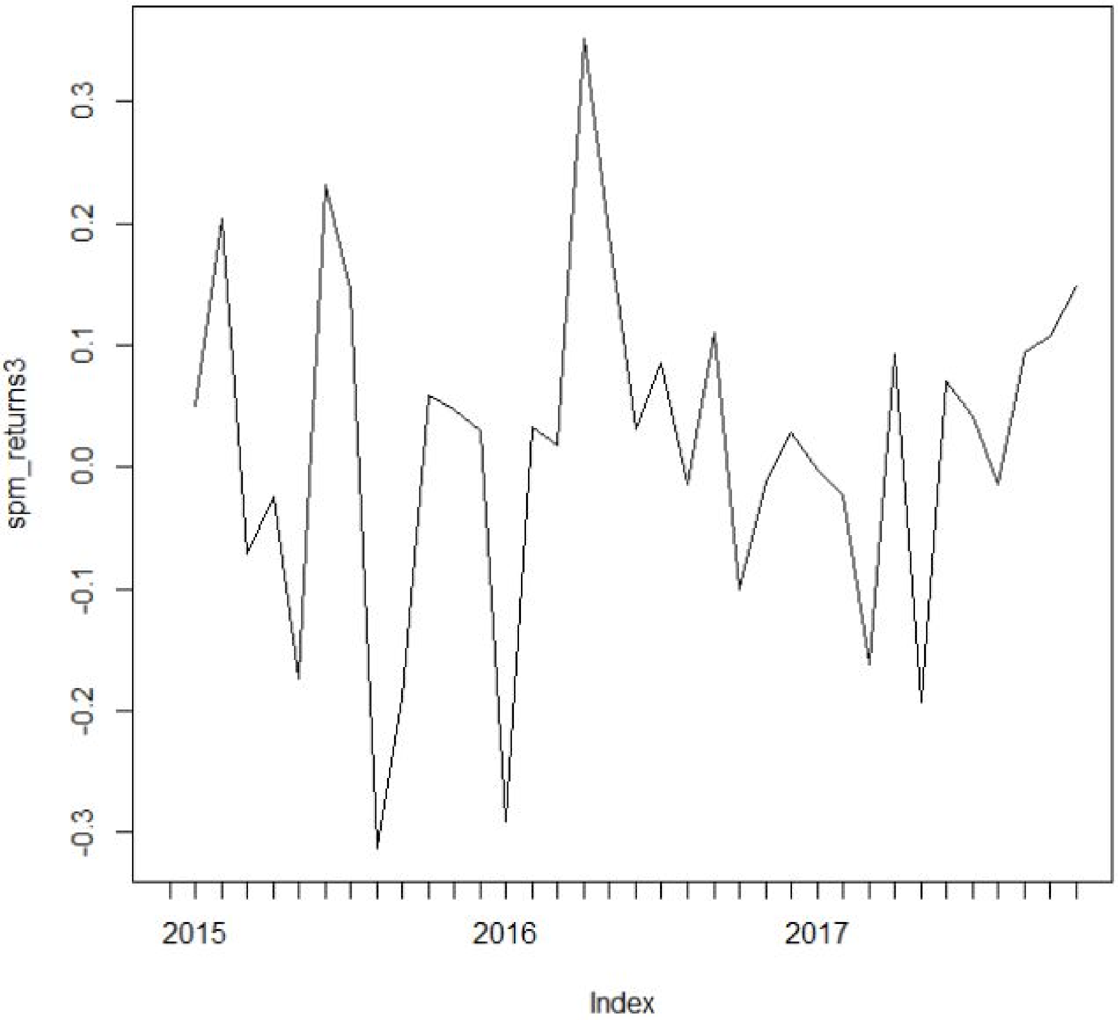
- Time Series Plots of Buy and Hold (BH) and Daily Rebalancing (DR) on our portfolio of 3 Stocks





- Analysis of Monthly Returns for Microsoft, BAC, and ANI Pharmaceuticals


spm_returns1
0.10
0.00
-0.10
2015
2016
2017
spm_returns2
0.3
0.2
0.1
0.0
-0.1
2015
2016
2017
Index

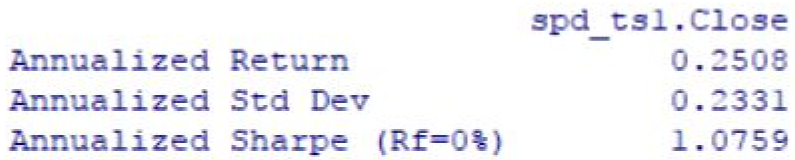


- Sharpe Ratio (Monthly)

Reported Sharpe for MSFT: -5.697845
Reported Sharpe for BAC: -4.599159
Reported Sharpe for ANI: -2.732
Observe above, that as we move from more volatile to less volatile stocks, the sharpe ratio (which is a relative measure from risk free rate, in our case US treasury Bill) decreases (in absolute terms)

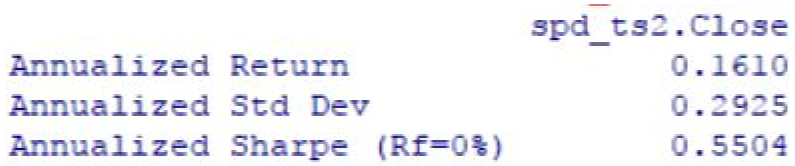
- Reported Annualized Sharpe (With 0 Risk Free Rate)

MSFT:



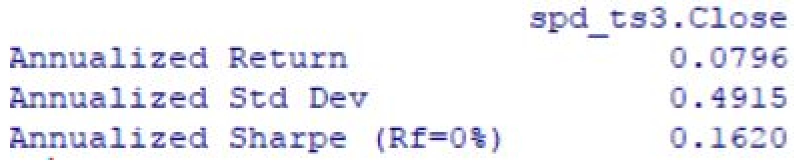
```
                                spd_ts1.Close
Annualized Return                      0.2508
Annualized Std Dev                     0.2331
Annualized Sharpe (Rf=0%)              1.0759
```

BAC:



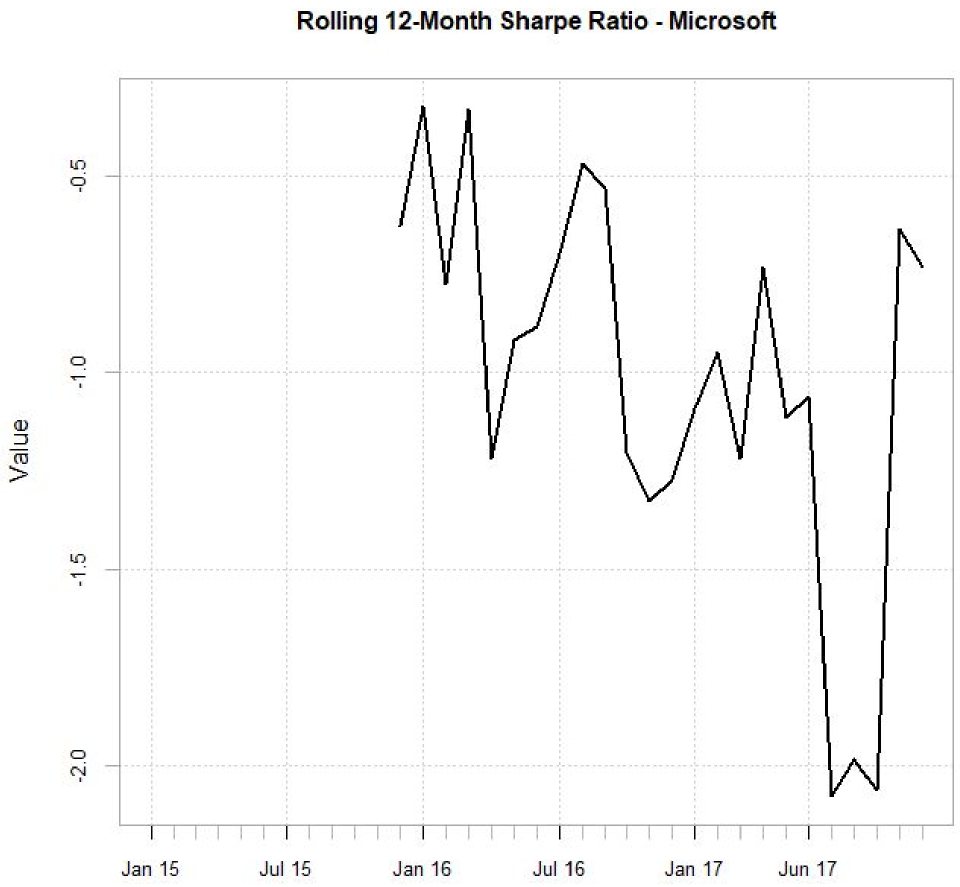
```
                                spd_ts2.Close
Annualized Return                      0.1610
Annualized Std Dev                     0.2925
Annualized Sharpe (Rf=0%)              0.5504
```

ANI:



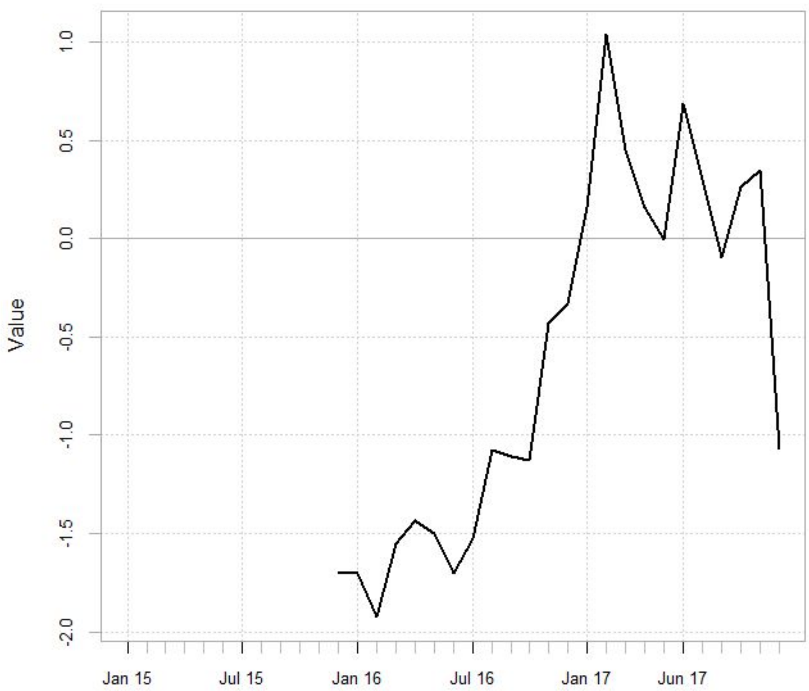
```
                               spd_ts3.Close
Annualized Return                     0.0796
Annualized Std Dev                    0.4915
Annualized Sharpe (Rf=0%)             0.1620
```

- Rolling 12 month estimates to help investors look for yearly dynamic variations

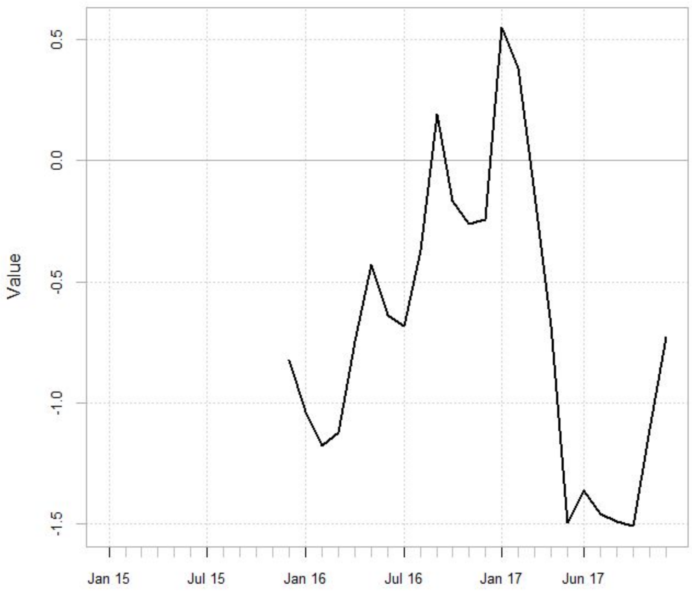


Rolling 12-Month Sharpe Ratio - Microsoft

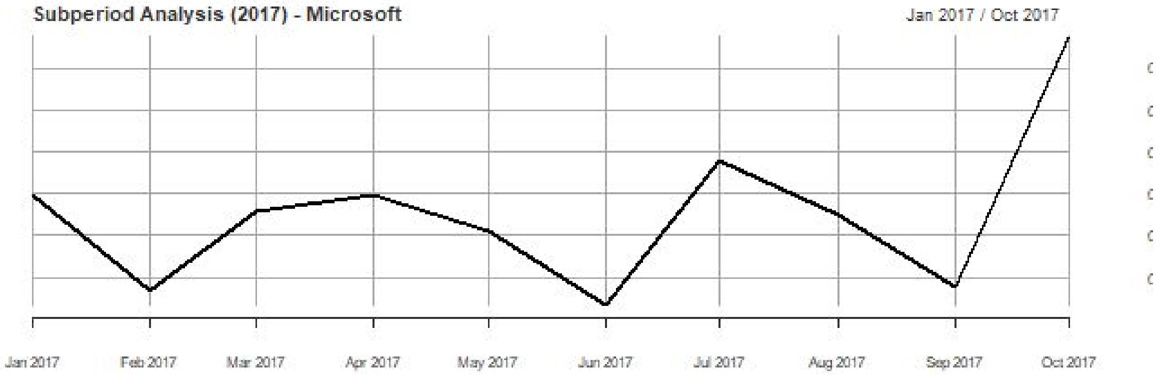
Rolling 12-Month Sharpe Ratio - Bank of America

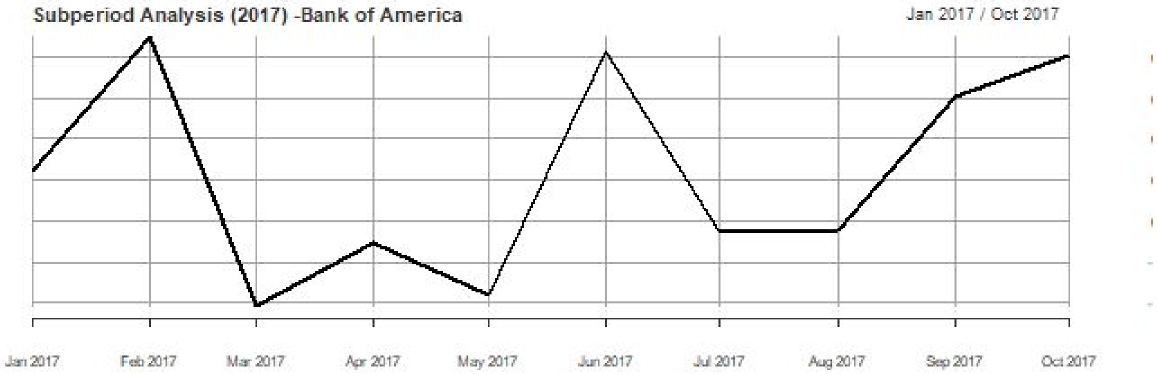


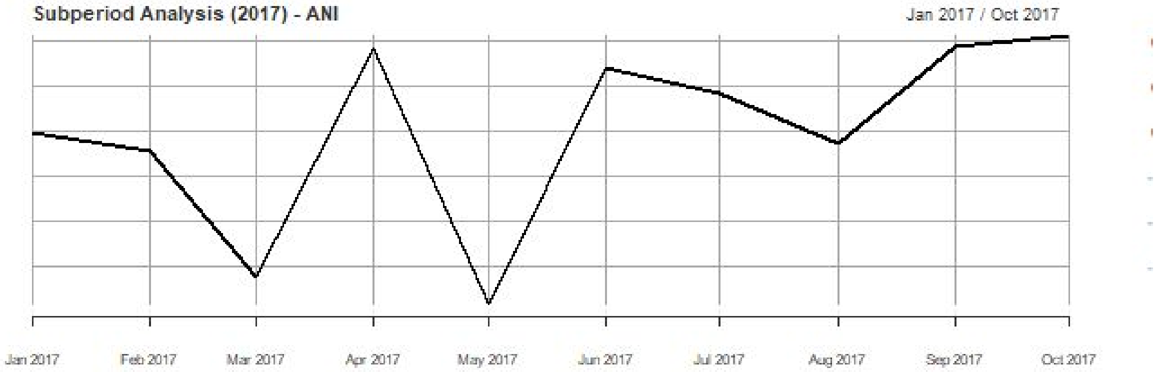
Rolling 12-Month Sharpe Ratio - ANI Pharmaceuticals



- SubPeriod Analysis: Helps figure out how has been the commodity trending in a specific window

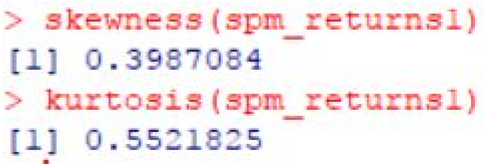






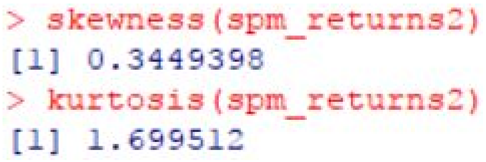
- Skewness and Kurtosis Measures on our commodities

Microsoft:



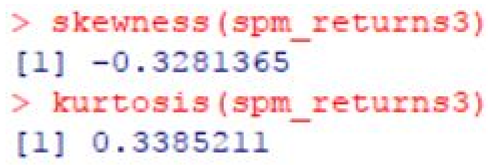
```
> skewness(spm_returns1)
[1] 0.3987084
> kurtosis(spm_returns1)
[1] 0.5521825
```

Bank of America:



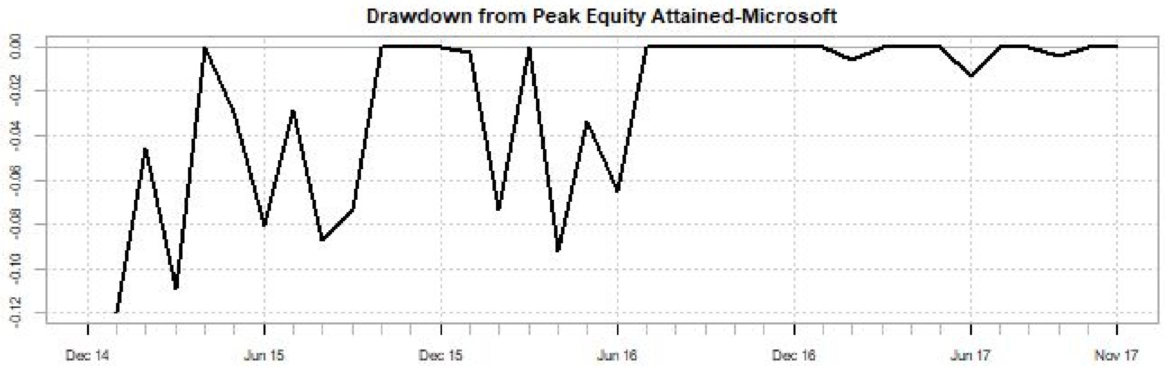
```
> skewness(spm_returns2)
[1] 0.3449398
> kurtosis(spm_returns2)
[1] 1.699512
```

ANI:

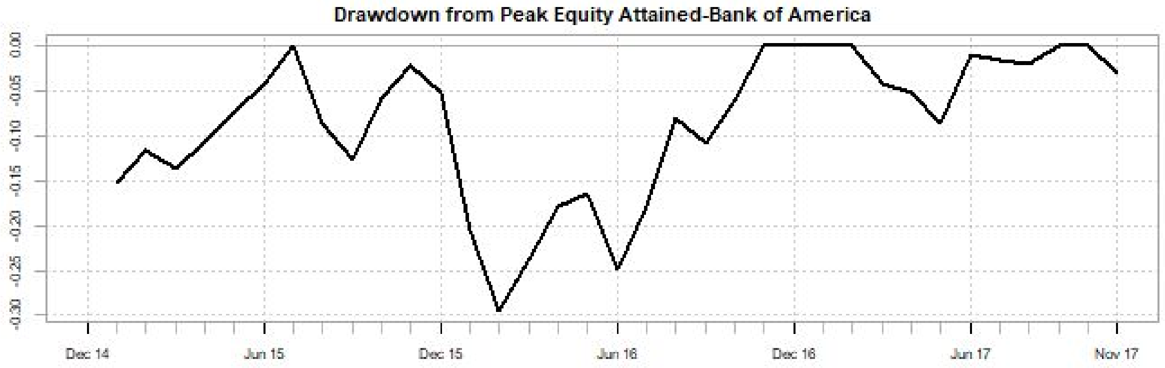


```
> skewness(spm_returns3)
[1] -0.3281365
> kurtosis(spm_returns3)
[1] 0.3385211
```

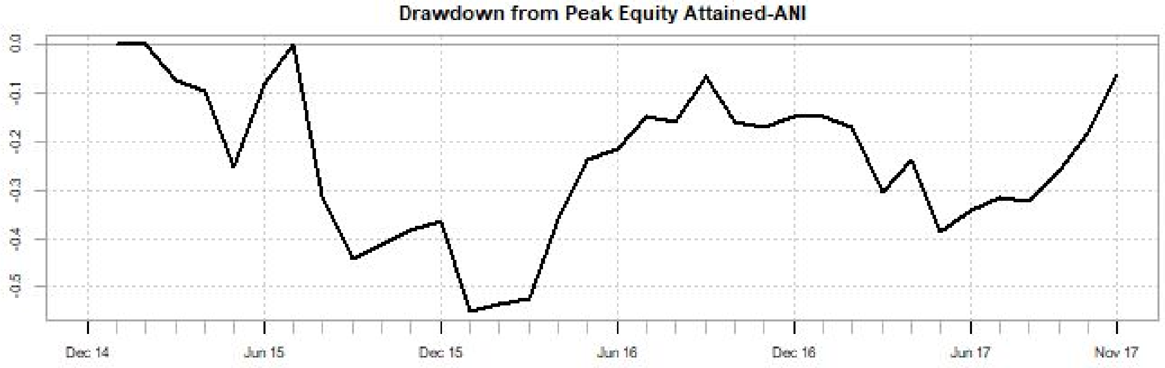
- Drawdowns obtained from Peak equities



Drawdown from Peak Equity Attained-Microsoft

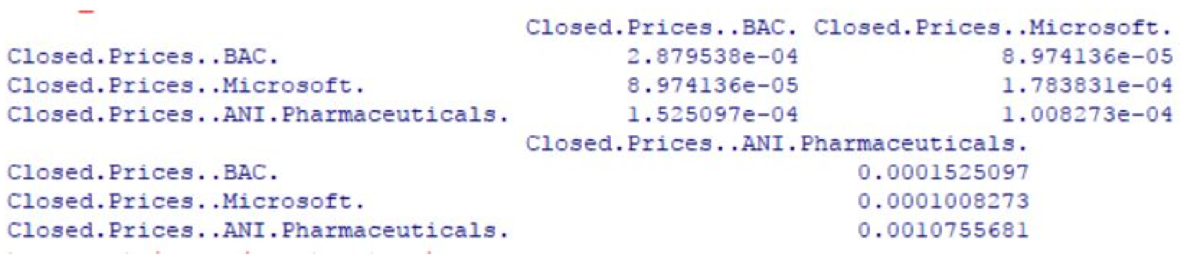


Drawdown from Peak Equity Attained-Bank of America



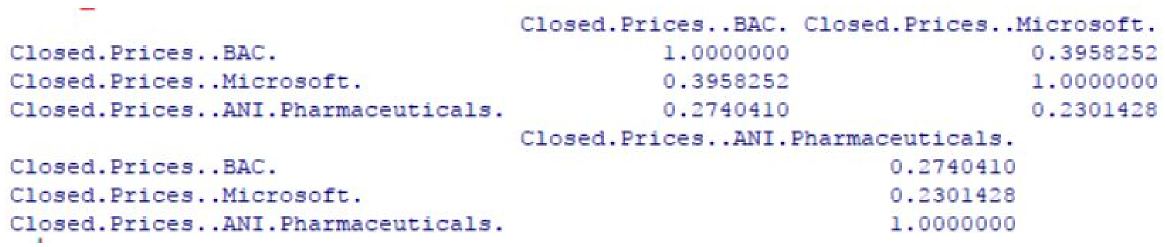
Drawdown from Peak Equity Attained-ANI

- Covariance Matrix between Stocks:



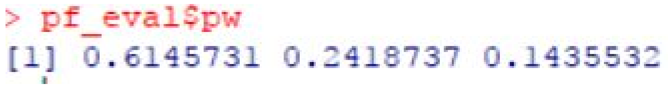
```
                                   Closed.Prices..BAC. Closed.Prices..Microsoft.
Closed.Prices..BAC.                       2.879538e-04              8.974136e-05
Closed.Prices..Microsoft.                 8.974136e-05              1.783831e-04
Closed.Prices..ANI.Pharmaceuticals.       1.525097e-04              1.008273e-04
                                   Closed.Prices..ANI.Pharmaceuticals.
Closed.Prices..BAC.                                       0.0001525097
Closed.Prices..Microsoft.                                 0.0001008273
Closed.Prices..ANI.Pharmaceuticals.                       0.0010755681
```

- Correlation Matrix between stocks:



```
                                   Closed.Prices..BAC. Closed.Prices..Microsoft.
Closed.Prices..BAC.                          1.0000000                 0.3958252
Closed.Prices..Microsoft.                    0.3958252                 1.0000000
Closed.Prices..ANI.Pharmaceuticals.          0.2740410                 0.2301428
                                   Closed.Prices..ANI.Pharmaceuticals.
Closed.Prices..BAC.                                          0.2740410
Closed.Prices..Microsoft.                                    0.2301428
Closed.Prices..ANI.Pharmaceuticals.                          1.0000000
```

- Portfolio Weights before readjusting till 27th November (before forecasting)



```
> pf_eval$pw
[1] 0.6145731 0.2418737 0.1435532
```

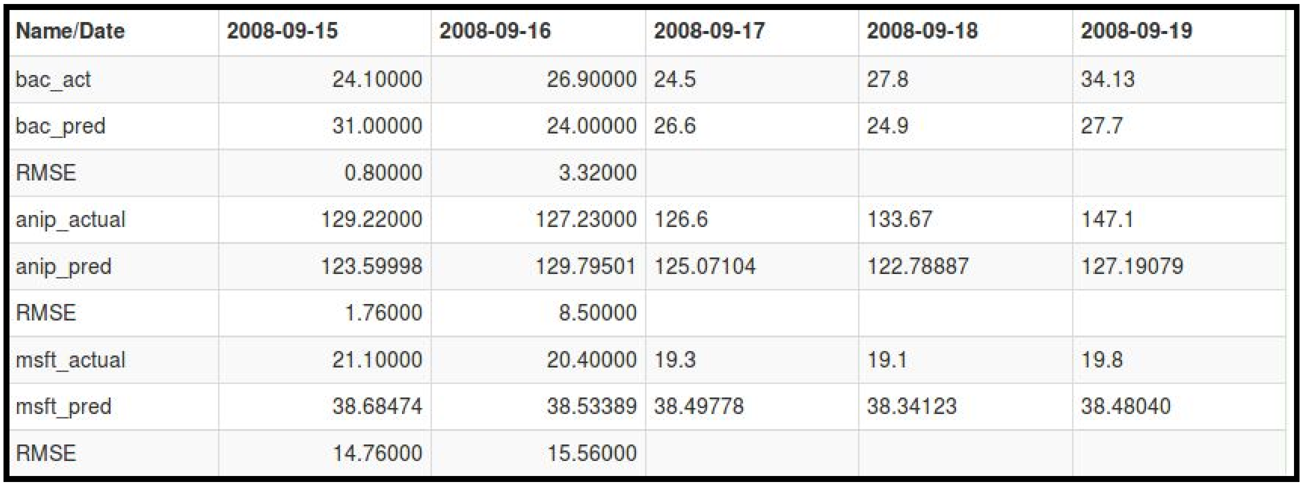
Based on the analysis performed above and the predicted prices obtained, we derive our optimum portfolio weights for daily rebalancing and tested if we were able to beat NASDAQ Index.

Results

Below is the table representing RMSE of all methods we experimented.

Models/Stocks	**BAC (Bank of America)**	**MSFT (Microsoft)**	**ANIP Pharmaceuticals**
SLR	**7.59**	**19.86**	**8.80**
Smooth Splines	**4.55**	**1.64**	**1.65**
MLR	**0.76**	**0.70**	**1.64**
PCA	**7.59**	**19.82**	**8.48**
Random Forest	**5.69**	**5.70**	**1.75**
GAM Splines	**0.76**	**0.59**	**2.11**

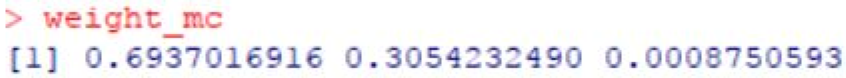
BACKTEST



Name/Date	2008-09-15	2008-09-16	2008-09-17	2008-09-18	2008-09-19
bac_act	24.10000	26.90000	24.5	27.8	34.13
bac_pred	31.00000	24.00000	26.6	24.9	27.7
RMSE	0.80000	3.32000			
anip_actual	129.22000	127.23000	126.6	133.67	147.1
anip_pred	123.59998	129.79501	125.07104	122.78887	127.19079
RMSE	1.76000	8.50000			
msft_actual	21.10000	20.40000	19.3	19.1	19.8
msft_pred	38.68474	38.53389	38.49778	38.34123	38.48040
RMSE	14.76000	15.56000			

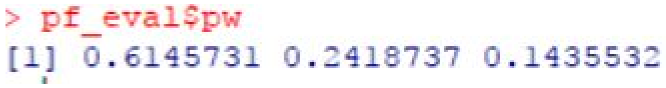
Portfolio Analysis

The results for optimized portfolio shifted from our initial weightage of:



```
> weight_mc
[1] 0.6937016916 0.3054232490 0.0008750593
```

to

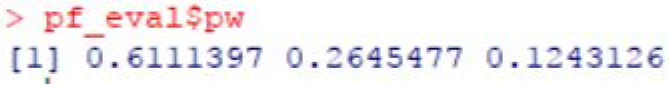


```
> pf_eval$pw
[1] 0.6145731 0.2418737 0.1435532
```

The portfolio weights above are the ones that existed on the 27th of November, i.e. one day prior to prediction made. Let us call them **W_27.** The order of weights is as follows:

Microsoft, Bank of America, and ANI Pharmaceuticals

Based on the predictions obtained for 28th November, we modified the portfolio and the results obtained are as follows:

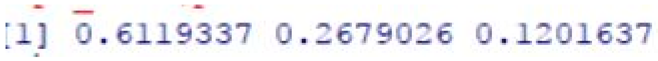


```
> pf_eval$pw
[1] 0.6111397 0.2645477 0.1243126
```

Let us call them **W_P_28.**

Notice above that our weights shifted from W_27 to W_P_28 based on the predictions.

Now let us examine how the portfolio adjusts on the 28th with actual prices.



```
[1] 0.6119337 0.2679026 0.1201637
```

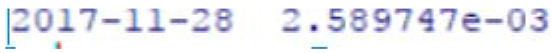
Let us call them **W_A_28.**

The results above lead us to two interesting observations:

- Our portfolio moved in the correct direction, and was slightly off from the actual weights.
- Had we not modified our weights on 27th November, **we would have not gained maximum returns.**

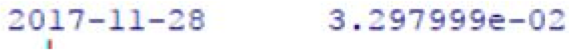
Having said that, a portfolio manager is termed successfully when he is able to “beat the market”. And so, we compared our results with the returns from NASDAQ for 28th November.

Returns from NASDAQ



```
2017-11-28  2.589747e-03
```

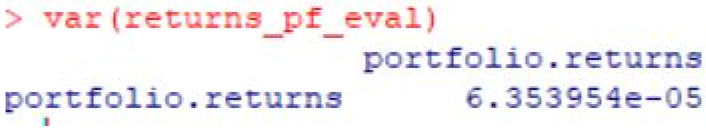
Returns from our portfolio



```
 2017-11-28      3.297999e-02
```

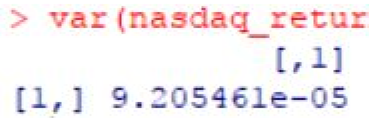
Notice above, that we are able to exceed the returns of NASDAQ.

Moreover, if we compare the variance of our portfolio returns vs that of NASDAQ,
Variance from Portfolio:



```
> var(returns_pf_eval)
                  portfolio.returns
portfolio.returns      6.353954e-05
```

Variance from NASDAQ Returns:



```
 > var(nasdaq_return)
              [,1]
 [1,] 9.205461e-05
```

We observe that our portfolio returns were stable than that of NASDAQ.

Discussion

From the above performed methods to predict stocks, we can note one thing that stock market moves in a very uncertain manner, and prediction of such an uncertainty is very difficult.What matters to buyer is not how prices fared in the past, but how will they perform in future. Not every person has the domain knowledge to understand the intricate details of stock market, hence a machine learning model for novice investors can help in a great way to invest in stock market with ease initially and the model be such that it gives the maximum returns with minimal error possible. In our case, in spite of some error in the prediction, we were able to build a portfolio in the similar direction it would have been developed using the actual close prices. In future we aim to build a much robust and accurate model by the use of advance learning techniques used in Deep Learning like neural networks for a better prediction, using other attributes such as trending news regarding the stock.

Conclusion

Based on our research, we can conclude that simple models are not accurate when it comes to predicting stock prices. With advanced non-linear models and a balanced selection of train-test data proportions, our results showed that GAM showed better results for BAC and MSFT, and MLR and Random Forest worked well with ANIP. However, when we performed out of sample backtesting with data during the 2008 global financial crisis, our models went significantly off during the period 15th September to 19th September, following the collapse of Lehman Brothers on September 15th, 2008. This leads us to an understanding that even though our models fared well with the data we had, our models deviated when tested during a different period. Therefore, to apply this strategy, it would be risky to implement in reality. And therefore, experts are now heading to use deep learning and other advanced techniques to improve accuracy.

For portfolio, our product moved in the correct direction, and had variance that was less than that of NASDAQ 100 returns. We observe that our returns were greater than that of NASDAQ 100. However, we did not take into account friction costs that are associated with investment returns and they have huge impact when it comes to beating the market.

References

1. Asep Juarna, Adi Kuswanto, Mohammad Abdul Mukhyi, and Raden Supriyanto, "Curve Fitting and Stock Price Prediction Using Least Square Method", ICBLCSR'15.
2. Murtaza Roondiwala,, Harshal Patel , Shraddha Varma, "Predicting Stock Prices Using LSTM", IJSR 2017.
3. Jana Cipan, "EVALUATING FORECAST ACCURACY".
4. www.investopedia.com/
5. https://mrjbq7.github.io/ta-lib/
6. www.schroders.com/
7. https://en.wikipedia.org/wiki/Time_series
8. https://www.r-bloggers.com/forecasting-stock-returns-using-arima-model/
9. http://www.tradinggeeks.net/2014/07/technical-analysis-with-r/