Phong Nguyen

[pnguyen014@regis.edu](mailto:pnguyen014@regis.edu)

MSDS 696

Final Project

**Applied Predictive Analytics and Machine Learning in predicting stock outcome.**

**I: Overview**

I’m planning to apply Predictive Analytics technique to analyze and predict the output the data.

The Data source is used SP500 [2] and I picked First Data Corporation to predict the future price. First Data is given the trading code name, FDC.

company to analyze and predict the outcome. The source data is available in yahoo financial section.

Yahoo provided historical data [1] based on the time range. There are two datasets for this project. First data is FDC dataset and second dataset is SP500. These two datasets will be combined and cleaned before starting to analyze and apply some predictive and machine learning to get an outcome. The outcome will show the results of a prediction based on the historical and correlation with the SP500 data.

The outcome will provide the information for the future invest on FDC stock.

**II: Data Sources**

There is one source to collect the newly created data. It’s on Yahoo financial. It’s free of charge. There are two datasets will be used for this. Besides, FDC, there are other variables are added to the dataset. SP500, NASDAQ, DOWJONES, and RUSSELLS 2000

1: <https://finance.yahoo.com/quote/FDC?p=FDC>

2: [https://finance.yahoo.com/quote/%5EGSPC?p=^GSPC](https://finance.yahoo.com/quote/%5EGSPC?p=%5eGSPC)

3. [https://finance.yahoo.com/quote/%5EDJI?p=^DJI](https://finance.yahoo.com/quote/%5EDJI?p=%5eDJI)

4. [https://finance.yahoo.com/quote/%5EIXIC?p=^IXIC](https://finance.yahoo.com/quote/%5EIXIC?p=%5eIXIC)

5. [https://finance.yahoo.com/quote/%5ERUT?p=^RUT](https://finance.yahoo.com/quote/%5ERUT?p=%5eRUT)

**III: Methodologies**

In in project, Predictive Analytics are used to analyze the data. Machine Learning is used to test the data and reinforce the results. There are two predictive analytics that applied to analyze and predict the data.

First, Regression is used to run and analyze. The outcome from regression analysis will show the chance of increasing the FDC stock price.

Second, Machine learning is used to test and predict the data. This method is reinforcing the result from regression analysis.

Finally, Time Series Analysis is used to predict the short term and long term FDC stock price.

**IV: Analysis and Results**

Dataset sample:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | FDC | FDCVOL | SP500 | DOWJONES | NASDAQ | RUSSELL |
| 2015-10-15 | 15.75 | 64971400 | 2023.86 | 17141.75 | 4870.1 | 1162.77 |
| 2015-10-16 | 16 | 11192000 | 2033.11 | 17215.97 | 4886.69 | 1162.31 |
| 2015-10-19 | 15.81 | 3437300 | 2033.66 | 17230.54 | 4905.47 | 1164.3 |
| 2015-10-20 | 15.4 | 5049300 | 2030.77 | 17217.11 | 4880.97 | 1163.27 |
| 2015-10-21 | 15.36 | 15941600 | 2018.94 | 17168.61 | 4840.12 | 1144.95 |
| 2015-10-22 | 16.01 | 6227800 | 2052.51 | 17489.16 | 4920.05 | 1154.52 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 2018-08-13 | 23.88 | 15564700 | 2821.93 | 25187.7 | 7819.71 | 1675.32 |
| 2018-08-14 | 23.93 | 33320800 | 2839.96 | 25299.92 | 7870.89 | 1692.58 |
| 2018-08-15 | 23.76 | 12423600 | 2818.37 | 25162.41 | 7774.12 | 1670.67 |
| 2018-08-16 | 24.11 | 15865600 | 2840.69 | 25558.73 | 7806.52 | 1685.75 |
| 2018-08-17 | 24.31 | 13485100 | 2850.13 | 25669.32 | 7816.33 | 1692.95 |

> MyRegFDC <- lm(UpDateFDC$FDC ~ UpDateFDC$FDCVOL + UpDateFDC$SP500 + UpDateFDC$DOWJONES + UpDateFDC$NASDAQ + UpDateFDC$RUSSELL)

> summary(MyRegFDC)

Call:

lm(formula = UpDateFDC$FDC ~ UpDateFDC$FDCVOL + UpDateFDC$SP500 +

UpDateFDC$DOWJONES + UpDateFDC$NASDAQ + UpDateFDC$RUSSELL)

Residuals:

Min 1Q Median 3Q Max

-4.1470 -0.8307 -0.0317 0.9797 3.8930

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.292e+00 1.818e+00 5.110 4.14e-07 \*\*\*

UpDateFDC$FDCVOL -1.693e-08 1.085e-08 -1.560 0.1193

UpDateFDC$SP500 -6.071e-03 3.578e-03 -1.696 0.0902 .

UpDateFDC$DOWJONES -1.574e-03 2.242e-04 -7.020 5.20e-12 \*\*\*

UpDateFDC$NASDAQ 6.751e-03 4.172e-04 16.181 < 2e-16 \*\*\*

UpDateFDC$RUSSELL 9.949e-03 1.569e-03 6.342 4.05e-10 \*\*\*

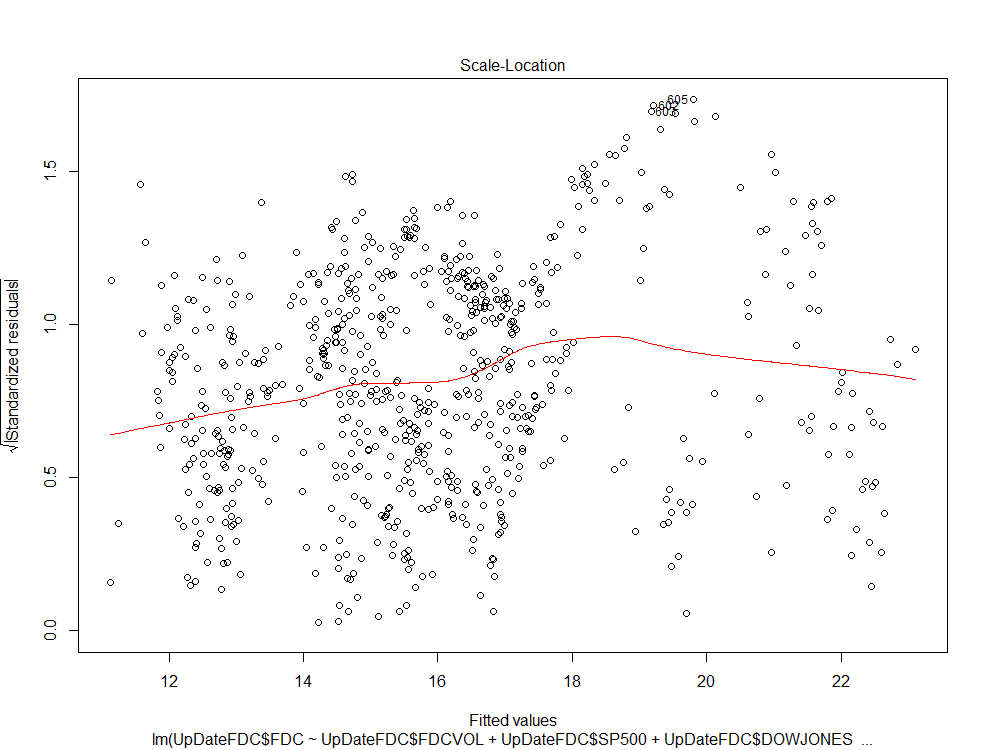
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.385 on 710 degrees of freedom

Multiple R-squared: 0.7683, Adjusted R-squared: 0.7667

F-statistic: 470.9 on 5 and 710 DF, p-value: < 2.2e-16



Linear regression shows 77% chance of increasing FDC stock price.

Next, Machine learning is used to test the data and compare the result from linear regression analysis.

> set.seed(50)

> SplitFDC <- sample(seq\_len(nrow(UpDateFDC)), size = floor(0.75 \* nrow(UpDateFDC)))

> trainData <- UpDateFDC[SplitFDC, ]

> testData <- UpDateFDC[-SplitFDC, ]

> PredictionFDCmodelUpDate <- lm(trainData$FDC ~ trainData$SP500 + trainData$DOWJONES + trainData$NASDAQ + trainData$RUSSELL + trainData$FDCVOL)

> summary(PredictionFDCmodelUpDate)

Call:

lm(formula = trainData$FDC ~ trainData$SP500 + trainData$DOWJONES +

trainData$NASDAQ + trainData$RUSSELL + trainData$FDCVOL)

Residuals:

Min 1Q Median 3Q Max

-4.1377 -0.8385 -0.0451 0.9485 3.9481

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.047e+01 2.086e+00 5.018 7.14e-07 \*\*\*

trainData$SP500 -8.159e-03 4.099e-03 -1.991 0.047 \*

trainData$DOWJONES -1.407e-03 2.619e-04 -5.372 1.17e-07 \*\*\*

trainData$NASDAQ 6.896e-03 4.720e-04 14.609 < 2e-16 \*\*\*

trainData$RUSSELL 9.531e-03 1.773e-03 5.375 1.15e-07 \*\*\*

trainData$FDCVOL -1.822e-08 1.350e-08 -1.350 0.178

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.369 on 531 degrees of freedom

Multiple R-squared: 0.772, Adjusted R-squared: 0.7699

F-statistic: 359.6 on 5 and 531 DF, p-value: < 2.2e-16

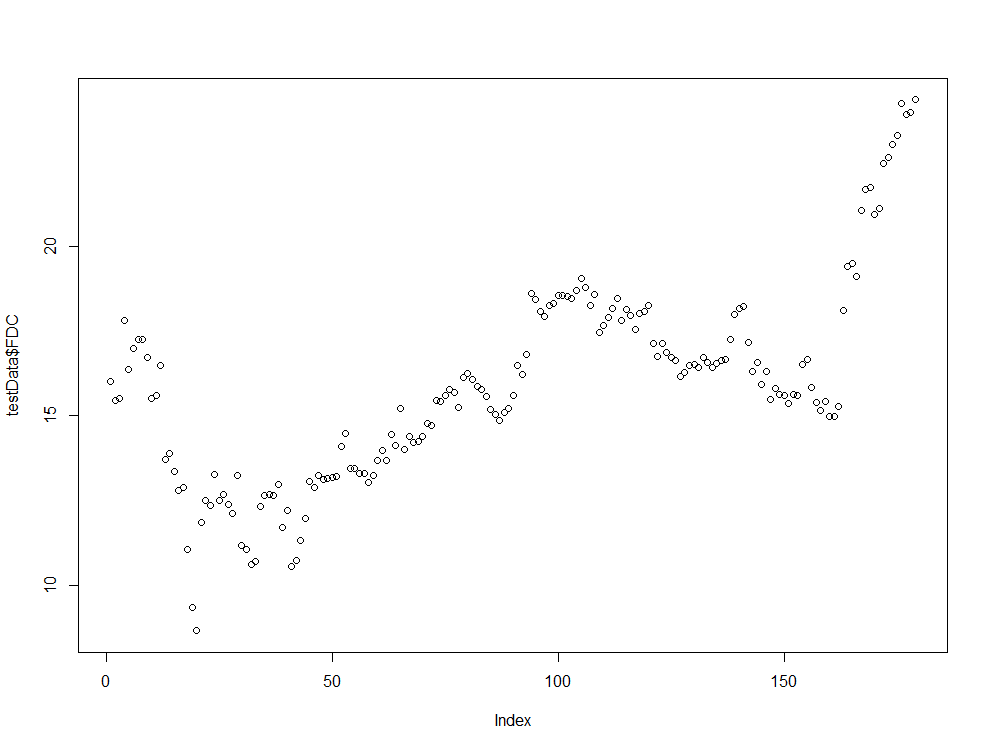
The outcome is 77% chance of increasing FDC stock price. Linear regression analysis and machine learning, the chance of increasing FDC stock price is the same.

Next, Machine learning predicts the FDC stock price.

> UpDatePredictionFDC <- predict(PredictionFDCmodelUpDate, newdata = testData)

Warning message:

'newdata' had 179 rows but variables found have 537 rows



The result from this group tell us that the highest price is closer to 25 dollars.

Final prediction is predicting short term and long term.

> UpDateFDC1$Date = as.Date(UpDateFDC1$Date)

> ggplot(UpDateFDC1,aes(UpDateFDC1$Date, UpDateFDC1$FDC)) + geom\_line() + scale\_x\_date('month')+ ylab("FDC Stock Price")+xlab("")

> TsFdc1 = ts(UpDateFDC1[, c('FDC')])

> UpDateFDC1$Fdccln = tsclean(TsFdc1)

> ggplot() + geom\_line(data = UpDateFDC1, aes(x = UpDateFDC1$Date, y = UpDateFDC1$FDC)) + ylab('Cleaned FDC STock Price')

> UpDateFDC1$Fdc\_ma = ma(UpDateFDC1$Fdccln, order=7)

> UpDateFDC1$Fdc\_ma30 = ma(UpDateFDC1$Fdc\_ma, order=30)

> ggplot() + geom\_line(data = UpDateFDC1, aes(x = UpDateFDC1$Date, y = UpDateFDC1$Fdc\_ma, colour = "Weekly Moving Average"))

Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.

Warning message:

Removed 6 rows containing missing values (geom\_path).

> ggplot() + geom\_line(data = UpDateFDC1, aes(x = UpDateFDC1$Date, y = UpDateFDC1$Fdc\_ma30, colour = "Weekly Moving Average")) + ylab('Fdc Price')

Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.

Warning message:

Removed 36 rows containing missing values (geom\_path).

> UpDateFdc\_ma = ts(na.omit(UpDateFDC1$Fdc\_ma), frequency=30)

> decomp = stl(UpDateFdc\_ma, s.window="periodic")

> deseasonal\_UpDateFDCprice <- seasadj(decomp)

> plot(decomp)

> fitfdc<-auto.arima(deseasonal\_UpDateFDCprice, seasonal=FALSE)

> tsdisplay1(residuals(fitfdc), lag.max=45, main='(1,1,1) Model Residuals')

Error in tsdisplay1(residuals(fitfdc), lag.max = 45, main = "(1,1,1) Model Residuals") :

could not find function "tsdisplay1"

> tsdisplay(residuals(fitfdc), lag.max=45, main='(1,1,1) Model Residuals')

> fitfdc3 = arima(deseasonal\_UpDateFDCprice)

> fitfdc3

Call:

arima(x = deseasonal\_UpDateFDCprice)

Coefficients:

intercept

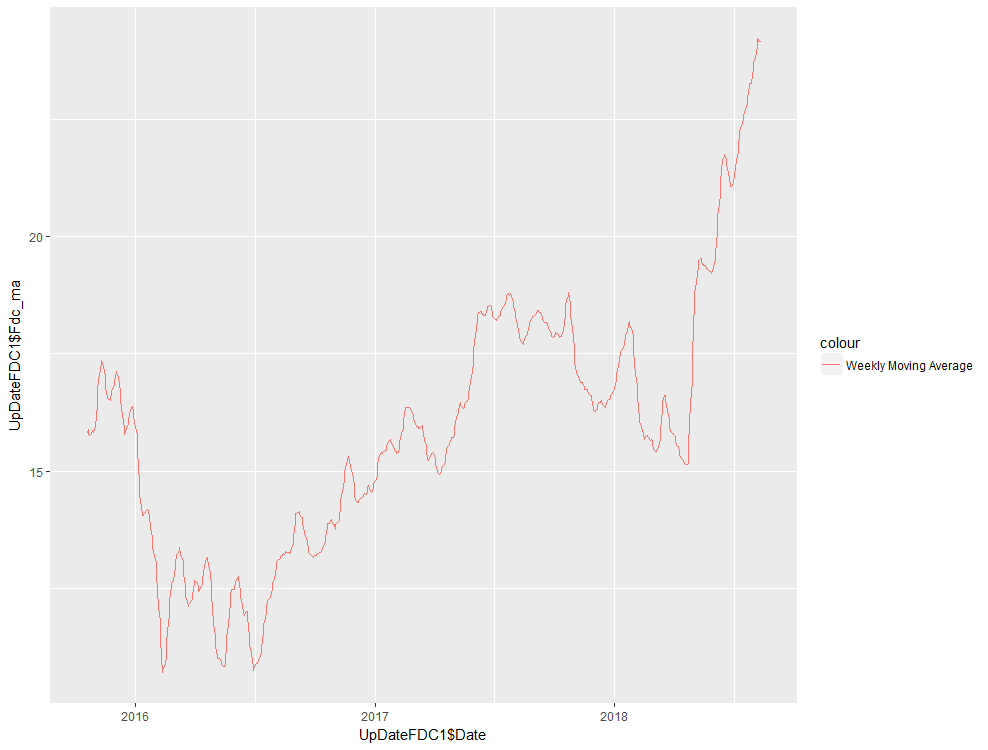
15.8888

s.e. 0.1047

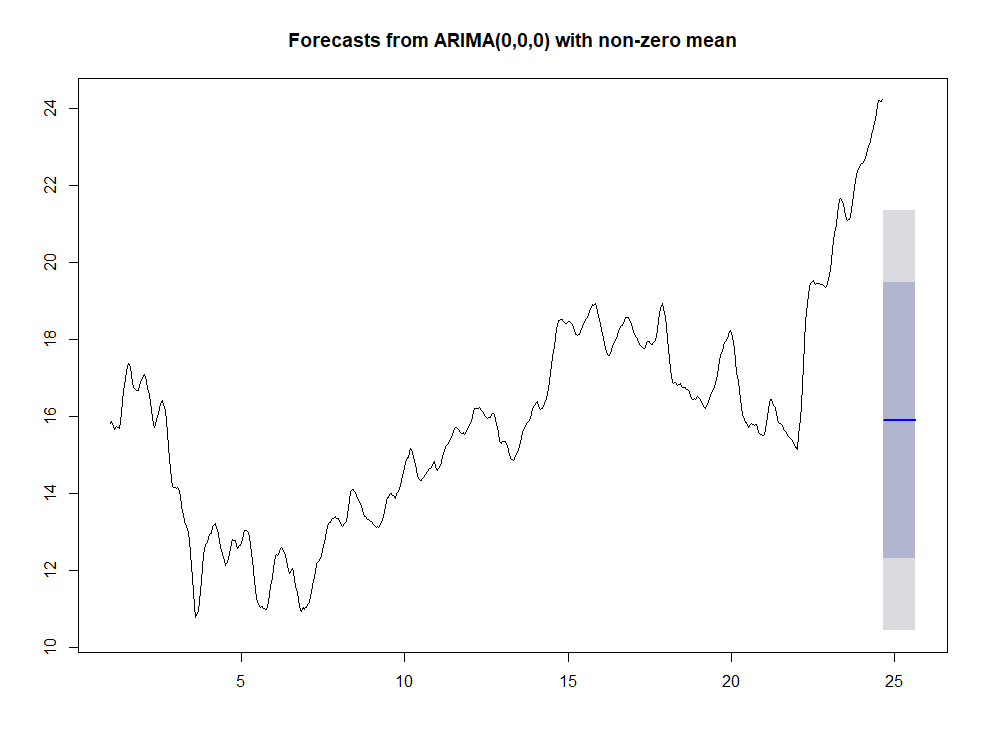
sigma^2 estimated as 7.781: log likelihood = -1735.82, aic = 3475.63

> UpDateFdcForecast <- forecast(fitfdc3, h=30)

> plot(UpDateFdcForecast)



Short term, FDC stock is increasing.



Long term, the FDC stock will fall down around 19 dollars and 12 dollars. This time series analysis predicts the median stock FDC price will be around 16 dollars.

**IV: Conclusions and Challenge**

**FDC** stock price in short term will look very good. FDC stock price in the last 3 years is 30% increase. In the last one year, FDC stock has very high return vale.

Predict the stock price is very difficult. It’s the biggest challenge stock trading professional. This project could produce more accuracy result if oil price and gold price variables are included in the dataset but Oil data and gold price data are hard to get for free. Gold and oil have very importance the correlations to the market.