LINEAR REGRESSION

Introduction

There are various factors that needs to be considered when discovering or predicting the price of a stock. In this paper, some of these factors are tested for the price prediction. Volume, Spread, Number of trades and Depth are taken into consideration as predictor variables. To predict the value of price, these predictor variables are included in a multiple linear regression and tested for their significance and relationship with the price.

Data Preparation and Clearance

The data used on this project is an hourly stock data of Apple Inc. (AAPL) from 01/01/2014 to 31/12/2019. It is downloaded from Reuters and it includes the sections Price, Volume, Number of Trades, Depth and Spread values for each hour (row) of data.

Initial observation by plotting the data sections revealed some of the corrupt data that needed to be fixed:

- The data was including some negative Spread values as well as some really high spread values(ex:500) that couldn't be true. Thus, these outliers were eliminated using z-score method.

  **See figure 1 for initial spread values, figure 2 for after-elimination Spread values*

- Some rows were including 'NaN' values, that was probably missing in the original data. These were eliminated.

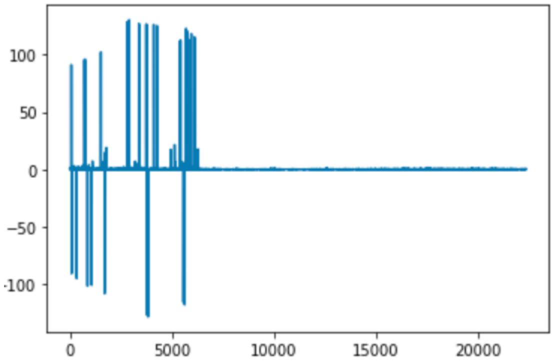


**Figure 1*

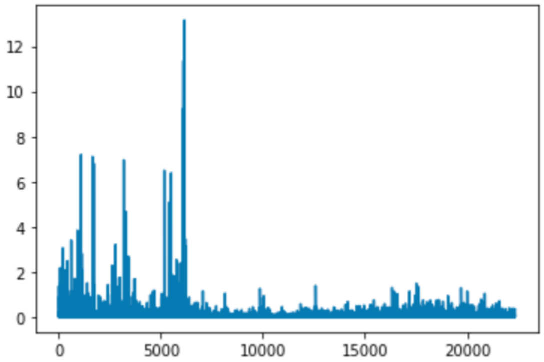
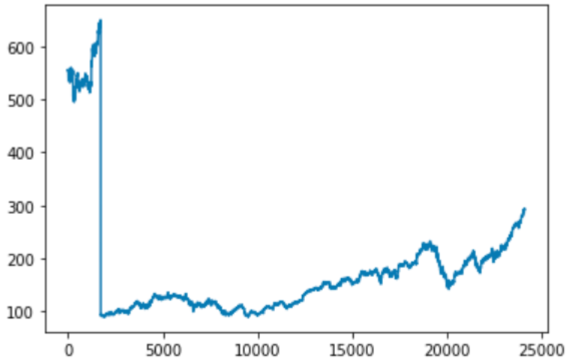


**Figure 2*

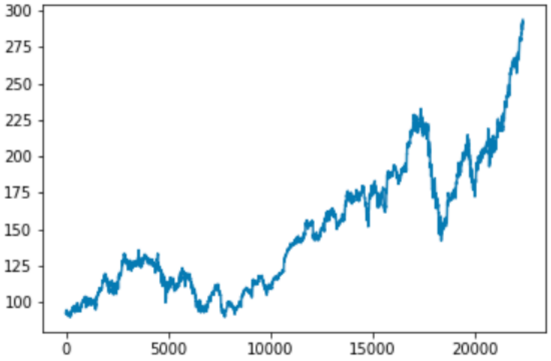
- On 6th of June, 2014, Apple Inc. has performed a stock split[1]. Thus, the stock's price appears to be dropping from $645.7 to $92.5 in just one trading hour. To perform a better analysis and prediction, the data before the stock split is not used in this research. *

**One could say it was possible to divide the old prices by seven to equalize the data, but it would be a wrong assumption due to the 'volume'. If the prices were cheaper at the beginning, there would be more individual investors contributing in the AAPL market.*

Following figures are showing the Price graph before and after the elimination:



Before the elimination



After the elimination

After the data is cleared out for the multiple linear regression, a normalization has been performed on both the dependent and predictor variables. Thus, the values' size would not affect the coefficients and relations. Volume's magnitude would be much larger than Spread without the normalization.

Research Method: Multiple Linear Regression Model

Multiple Linear Regression is a widely used model to understand different variables' effect on a dependent variable. The models most important use is to understand if there is a correlation between the explanatory variables and the dependent variable.

The model is preferable in many Finance related areas because of its ability to explain the significance and to predict relying on the data. The mathematical model of MLR is as follows:

Regression Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$

Prediction Equation: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ...+ b_k x_k$

y = Dependent Variable
x_i = Explanatory Variables
β_0 = y-intercept
β_i = Corresponding Correlation Coefficient
ε = Random Error

In this research, Number of trades, Depth, Volume and Spread are chosen as the predictor (explanatory) variables while the Price is the dependent variable. In the following tables mean, standard deviation and median values can be found for all of the variables.

Predictor Variables:

	Mean	3933.20
Number of Trades	Standard Deviation	4315.41
	Median	3183.00

	Mean	111581.54
Depth	Standard Deviation	106359.17
	Median	102776.00

	Mean	922792.26
Volume	Standard Deviation	1127384.7
	Median	698530.50

	Mean	0.07564
Spread	Standard Deviation	0.28407
	Median	0.02000

Dependent Variable:

	Mean	141.41
Price	Standard Deviation	42.507
	Median	126.13

The regression result table obtained from python is in the following:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.016
Model:                            OLS   Adj. R-squared:                  0.016
Method:                 Least Squares   F-statistic:                     57.59
Date:                Thu, 03 Jun 2020   Prob (F-statistic):           2.74e-48
Time:                        01:42:22   Log-Likelihood:                -20140.
No. Observations:               14274   AIC:                         4.029e+04
Df Residuals:                   14269   BIC:                         4.033e+04
Df Model:                           4
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------
x1             0.0449      0.016      2.823      0.005       0.014       0.076
x2            -0.1361      0.014     -9.405      0.000      -0.164      -0.108
x3             0.0706      0.010      7.174      0.000       0.051       0.090
x4            -0.0715      0.008     -8.412      0.000      -0.088      -0.055
const      -1.743e-16      0.008    -2.1e-14      1.000      -0.016       0.016
==============================================================================
Omnibus:                     1810.098   Durbin-Watson:                    .023
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                 .304
Skew:                           1.020   Prob(JB):                         0.00
Kurtosis:                       3.405   Cond. No.                         3.68
==============================================================================
```

```
Mean Absolute Error: 0.8333458972280601
Mean Squared Error: 0.9921084571434668
Root Mean Squared Error: 0.9960464131472322
```

According to the multiple linear regression performed above, all predictor variables are significant to chosen alpha = 0.05 and should be included in the model. The y-intercept is 0 as its P-value is 1. The model explains the 1.6% of the change in price.

Findings

Interpreting Coefficients:

Number of trades has a positive coefficient which means the price is in a direct relationship with number of trades. One standard deviation shock to Number of trades increases the Price by about $1.91.

Depth has an inverse relationship with Price, where one standard deviation shock decreases the Price by $5.79.

Volume is directly related with Price: as the volume increases the Price tends to increase according to the model above. One standard deviation shock to Volume increases the Price by $3.00.

Spread has a negative coefficient, thus there is an inverse relationship between Price and Spread. One standard deviation to Spread lowers the Price by $3.04.

Looking at the results all together, we understand that the market depth alone affects the price more than volume and the number of trades combined. As J. Brogaard, T. Hendershott, and R. Riordan explains their findings on their paper[2], limit orders' effect on the price is higher than the number of trades executed in the price discovery. The data worked with on this project is hourly data, thus expected HFT effect is rather lower compared to more frequent data. But the findings of this model and the research still manages to match on this point.

If the stock's volume is high, it means the stock's market is liquid, thus the spread tends to be lower. On the other hand; if the spread is high, it means that most probably the stock is illiquid and the volume is low. On each case the volume and the spread have to have an inverse relationship between each other. The model constructed above affirms this inverse relationship as they have different signs in front of their coefficients. One can also observe the absolute values of coefficients between the Spread and Volume are very close to each other.

DECISION TREE REGRESSION

Background and Process

I have performed the decision tree analysis for both the data I have gathered from Reuters and to the data I have calculated through python myself. The data calculated consists of the RSI, Simple moving average, exponential moving average and MACD-Signal.

RSI is relative strength index which is used to see if the market is oversold or overbought. MACD is used by traders to see if there is any convergence/divergence in different spanned moving averages of the same stock. MACD usually used with Exponential moving averages. I have also added a signal line to understand the convergence even beforehand.

Then, I have took the difference of MACD and Signal values to observe when the subtraction changes its sign. In other words, when there is a convergence in the MACD-Signal lines.

References

La Monica, Paul R. “Apple Stock Will Be 'Cheaper' on Monday.” *CNN Business*, CNN, 9 June 2104, 9.32 AM, money.cnn.com/2014/06/06/investing/apple-stock-split/index.html.

Brogaard, Jonathan, et al. “Price Discovery without Trading: Evidence from Limit Orders.” *The Journal of Finance*, vol. 74, no. 4, 2019, pp. 1621–1658., doi:10.1111/jofi.12769.