**WQD7005 DATA MINING**

**SEMESTER 2 SESSION 2018/2019**

**Name: Lim Kaomin,Leslie**

**Matric: WQD180076**

**Introduction**

This project focuses on using machine learning algorithms to help predict stock price movements. Predicting stock prices has always been difficult as many factors are involved such as monetary interest rates, political stability of a country, internal company developments, and many more. Predicting the stock trend is somewhat important for companies and investors as they want to invest in stocks that can bring them high returns and ensure their capital. With the emergence of machine learning techniques, stock price prediction could be greatly enhanced and serve as a strategy for investors to make smart investment decisions.

**Objective**

The objectives of the project are as follows:

1. Gain insights into the qualitative and quantitative data
2. Use machine learning techniques for stocks prediction
3. Compare the machine learning techniques and evaluate them

**Methodology**

Diagram 1: Workflow of project

**Data Collection/Pre-processing**

The stocks info are crawled from this website, <https://www.thestar.com.my/business/marketwatch/stocks/?qcounter=> using python.

The news headlines are crawled from this website, <https://www.klsescreener.com/v2/news> using python as well.

Both the data are then stored in Xampp as SQL file and are combined later in the Excel file as CSV format. This project only focuses on the banking stocks, thus the rest of the stocks that are unrelated were filtered out from the CSV file. The stocks data are crawled from the perod of 1st March 2019 to 29th April 2019.

**Data Visualization**

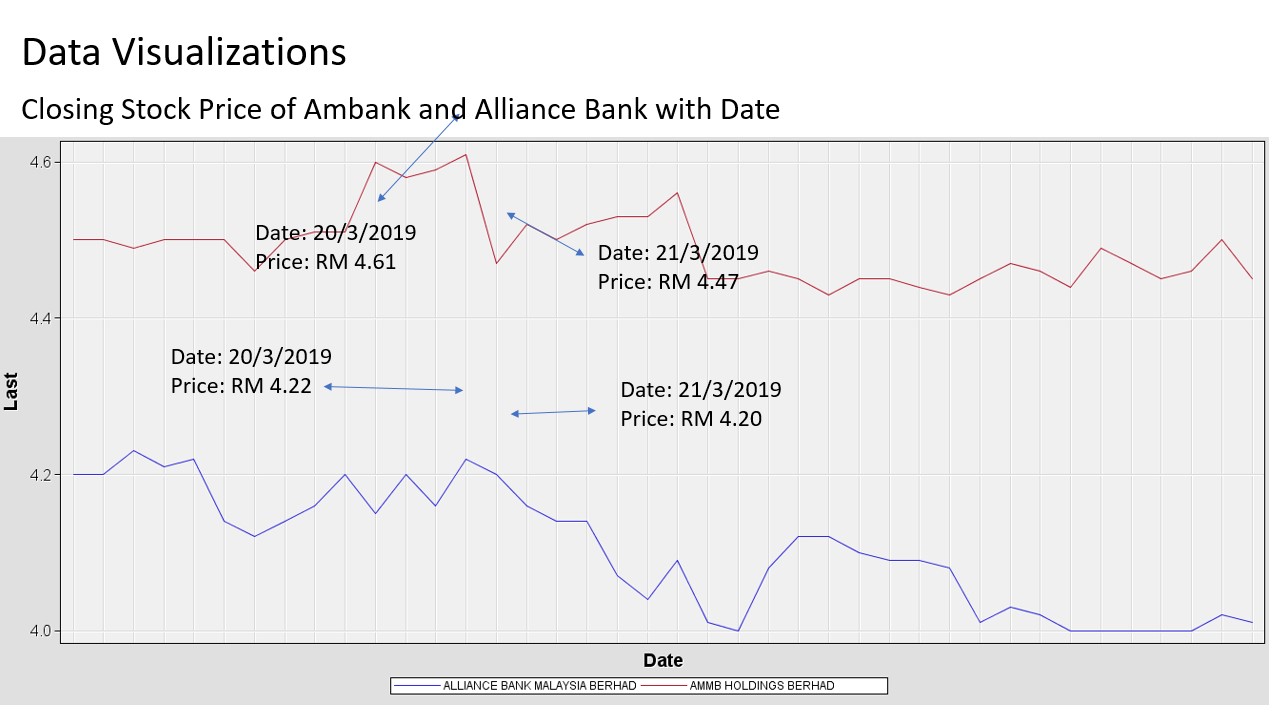


Diagram 2: Closing Stock Price Line Graph

It can be observed that both stock prices for Ambank and Alliance Bank almost showed a similar trend. An example would be from the 20th March to 21st March, Ambank stock prices declined from RM 4.61 TO RM 4.47. Alliance Bank also showed price decline from RM4.22 to RM 4.20 albeit a smaller decline.

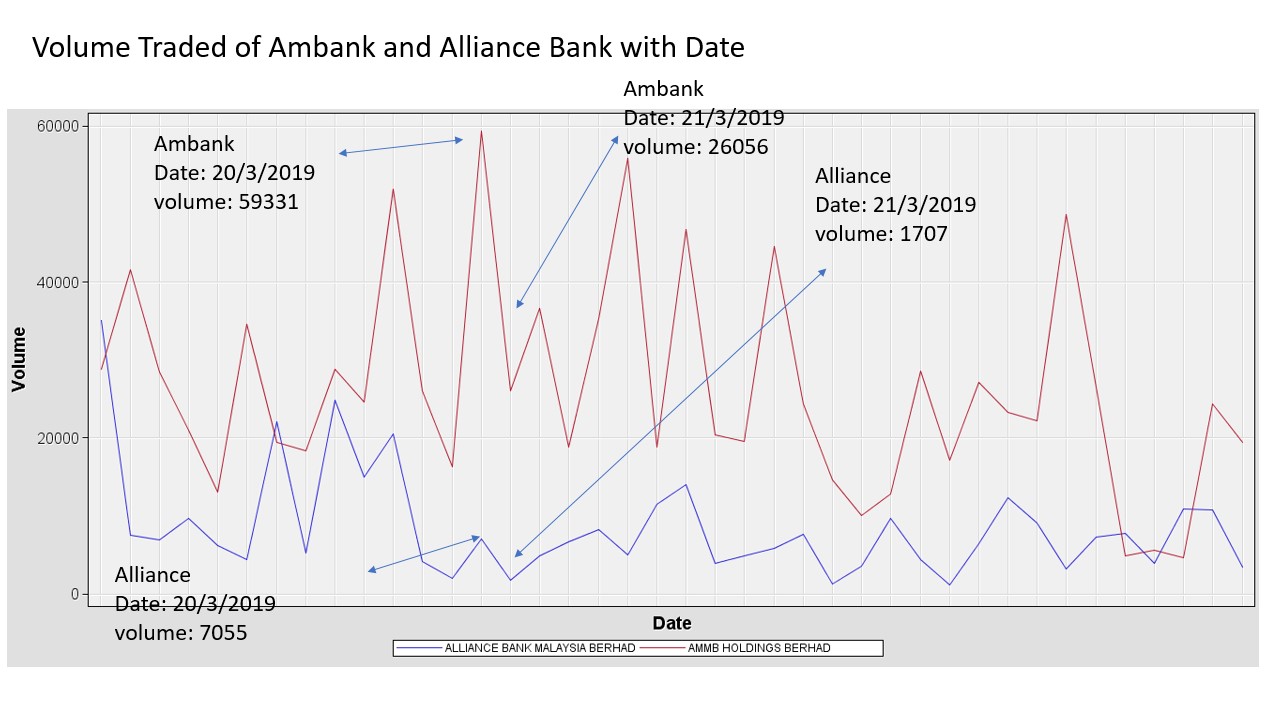


Diagram 3: Volume of Stocks Traded Line Graph

The graph shows the volume of stocks traded for Ambank and Alliance Bank respectively. Both banks showed variations as Ambank’s stock volumes traded were much volatile and constantly changing at a daily basis. Sometimes the stock volume traded rocket sky high while on other days the volume took a dip. At even one point, towards the end of the month April, the volume of Ambank’s stocks were lower than that of Alliance Bank. On the other hand, Alliance Bank initial stocks volume traded were high but gradually drop and showed a slow pace of pickup towards the beginning of the month April and gaining momentum throughout the rest of the period amid the volatility.

By comparing Diagram 2 and 3, it can be observed that both graphs exhibit similar trends. Showing that when prices are down, volume of traded stocks are also down and vice versa. This make sense as when there are demand for stocks, prices go up and vice versa.

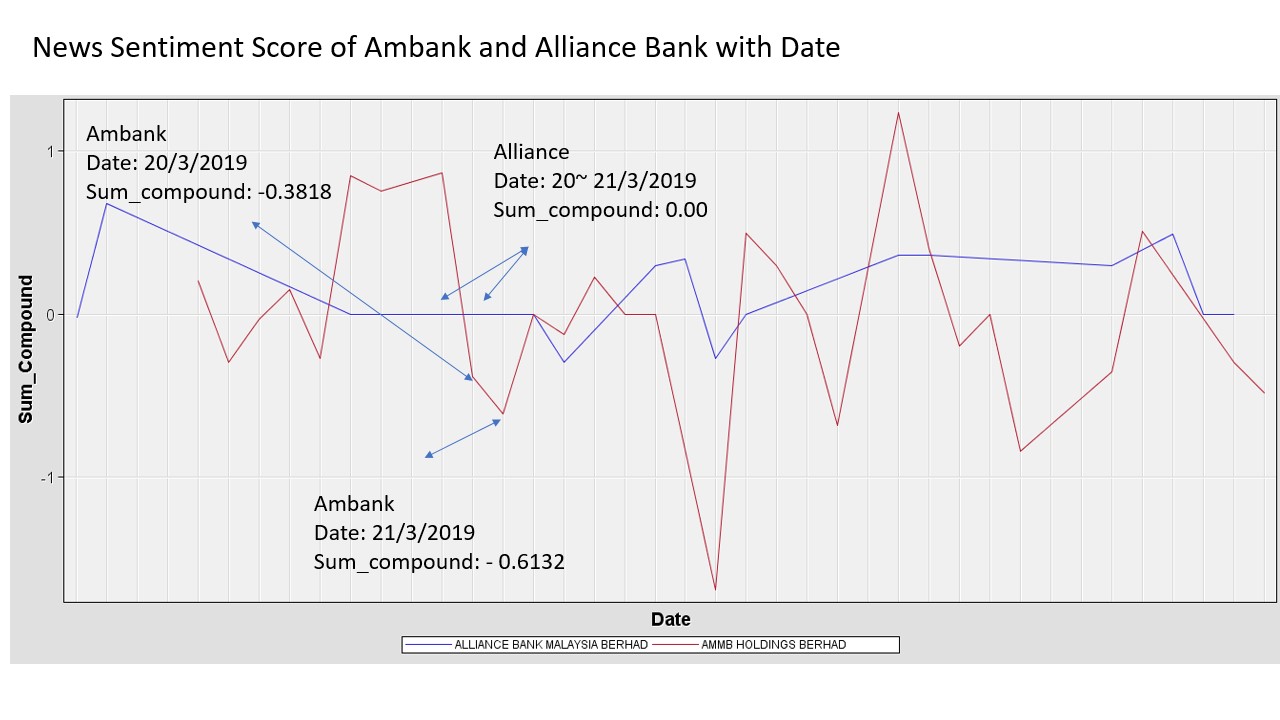


Diagram 4: Sentiment Score Line Graph

The graph shows the sentiment analysis score for both bank throughout the period. It shows that Ambank’s sentiment score in the news has a more share of negative news compared to Alliance Bank. Alliance Bank has been relatively positive on the news headlines with most of its sentiment score being positive throughout the period.

When a comparison is made for Diagram 4 with Diagram 2 or 3, not much can be derived as how news sentiment score could affect the stock prices or volume stocks traded.

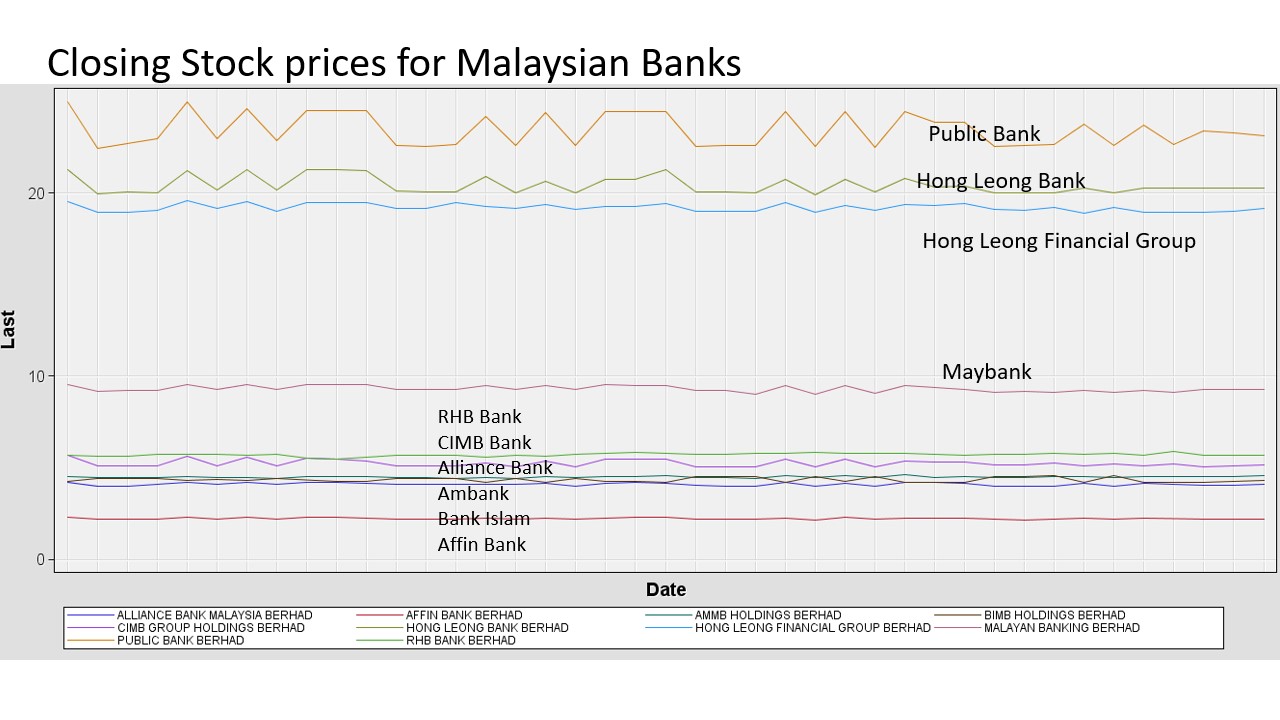
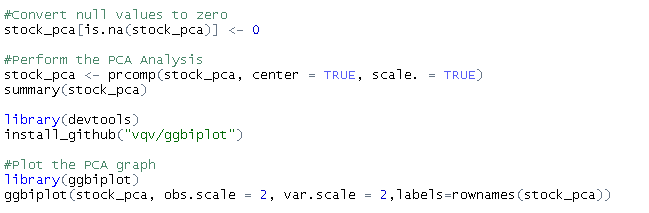


Diagram 5: Closing Stock Prices for Malaysian Banks

Diagram 5 show the trend for closing price for Malaysian banks. All of the bank’s closing stock price show almost similar pattern.

Principal Component Analysis (PCA)



The code above show the steps the do the Principal Component Analysis. The PCA is done to determine any correlation between the attributes.

The attributes correlation are also plotted in a graph using the ‘ggbiplot’ package.

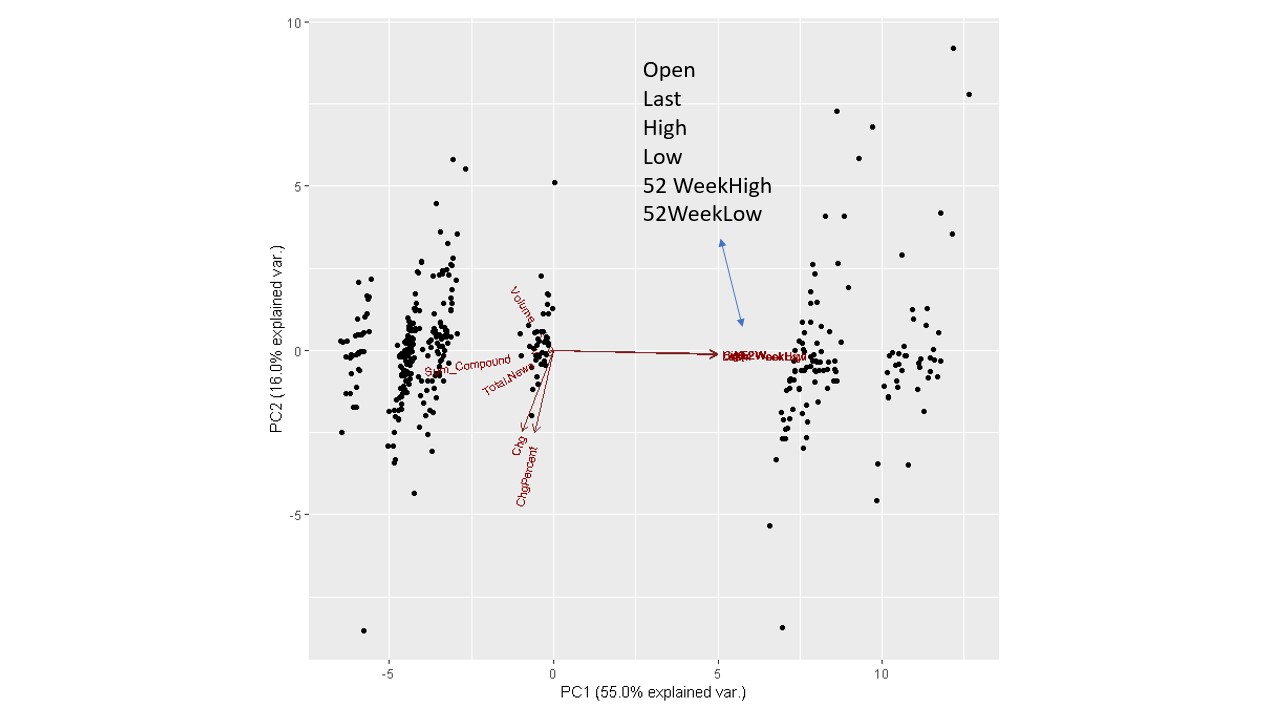
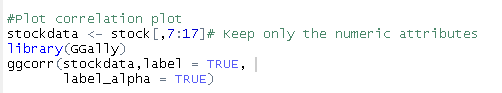


Diagram 6: Principal Component Plot of Attributes

From the Diagram 6, it can be seen that the sum\_compound, total\_news, volume, change and change percentage are quite close to each other. Suggesting that there might be certain correlation between these attributes. While the other attributes such as Open, Last and so on are at a distant but closely link to each other.

Correlation Plot



The code for plotting the correlation plot is shown above

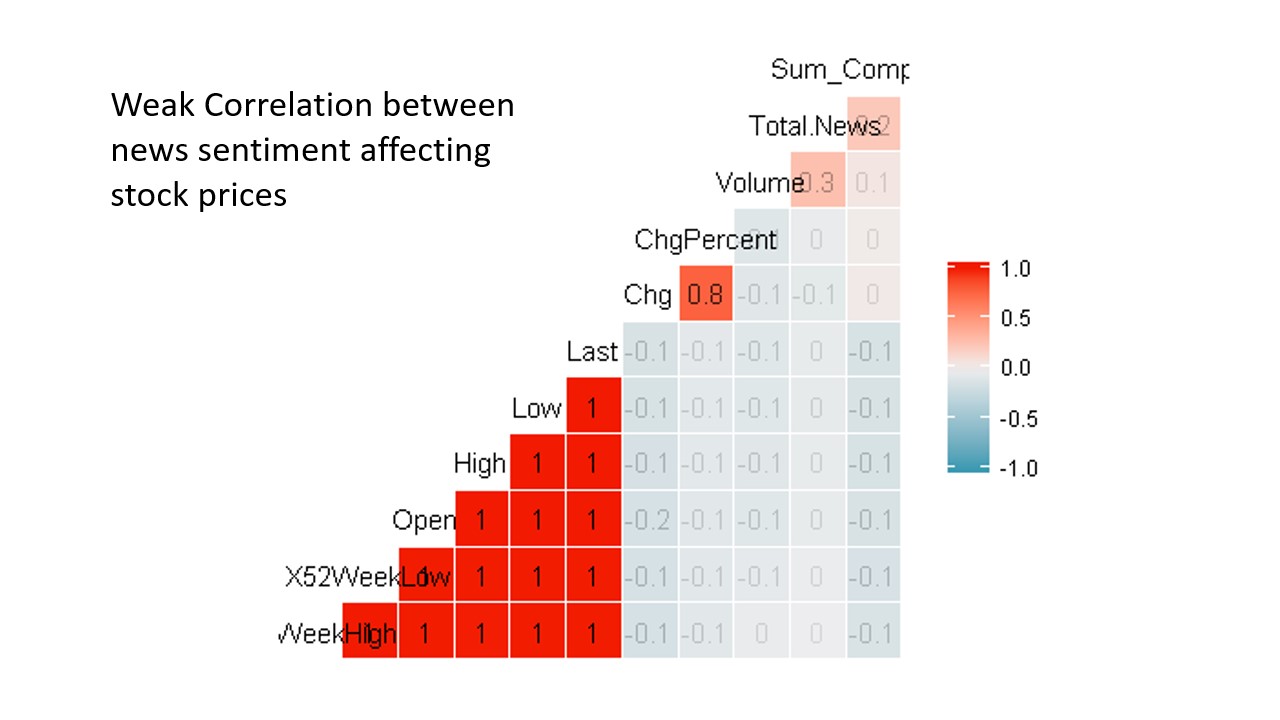


Diagram 7: Correlation Plot

From the correlation plot, it shows that there is a weak correlation on how news sentiment score (sum\_compound) could affect the stock closing prices. The correlation of 1 or -1 indicating highly correlated while a correlation of 0 indicating no correlation. In this case, the correlation between sum\_compound and last price is -0.1 which is a weak correlation score. Thus concluding that news headlines have little effect on the closing stock prices.

**Modelling**

For this project, 4 machine learning techniques will be used, namely Logistic Regression, Naïve Bayes, Decision Tree and Random Forest. The machine learning models were selected because this project aims to classify stock price movement, split into 2 categorical classes, namely Increase and No Change/Decrease.

The 'last price', 'price change' and 'percentage of price change' attributes are removed from the dataset as this attributes provides hints to the stock price movements, the target variable is the 'PriceLabel'.​

**Evaluation**

After modelling, it will be useful to evaluate the model to determine the model performance. By plotting the ROC Curve, the performance for the 4 machine learning models can be observed.

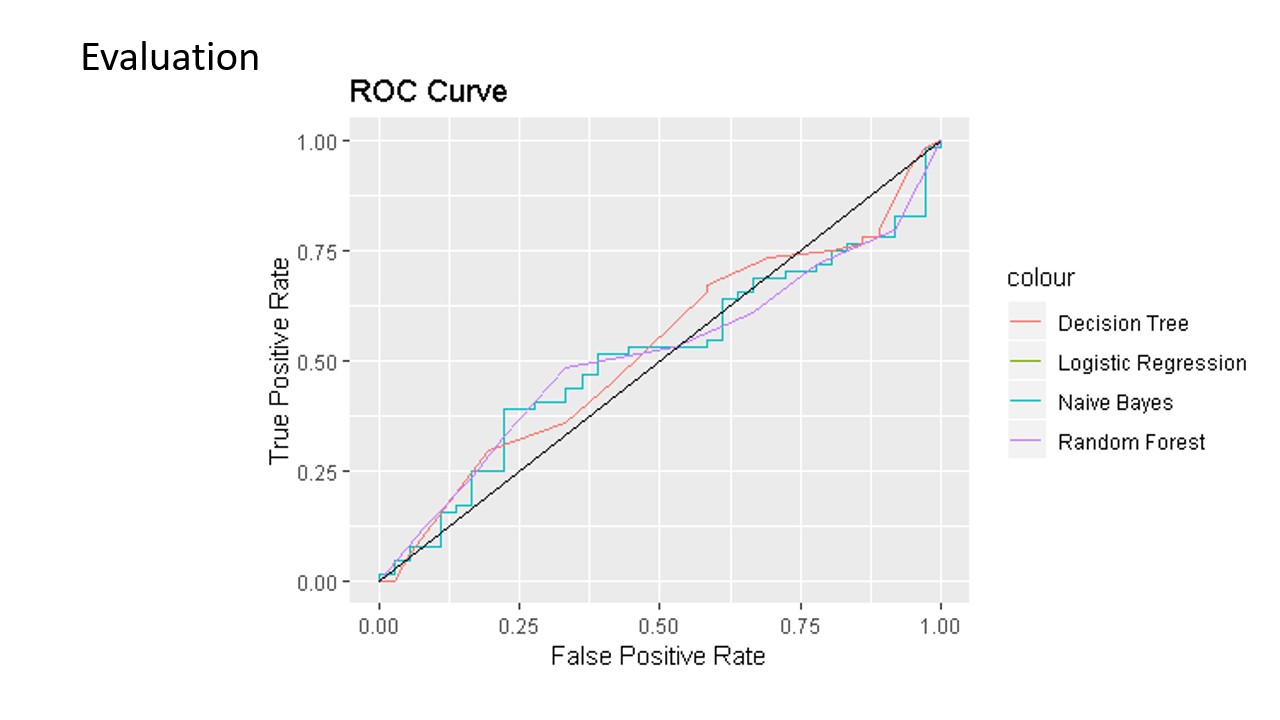


Diagram 8: ROC Curve

|  |  |  |
| --- | --- | --- |
| **Model​** | **Accuracy​** | **Area Under Curve​** |
| Logistic Regression​ | 46%​ | 0.5128​ |
| Naïve Bayes​ | 54%​ | 0.4918​ |
| Decision Tree​ | 58%​ | 0.4755​ |
| Random Forest​ | 57%​ | 0.5130​ |

Diagram 9: Machine Learning Performance Comparison Table

From Diagram 8, all the models showed the AUC performance close to 0.5. When AUC is approximately 0.5, model has no discrimination capacity to distinguish between positive class and negative class. This showed that all 4 models have a weak capacity to differentiate the True Positives from the False Positives. By observing Diagram 9, it can be seen that all 4 models has an AUC of close to 0.5. Only Logistic Regression and Random Forest showed slightly better performance among all 4 models.

Diagram 9 also shows the accuracy for each models, Logistic Regression registered the lowest accuracy of only 46% while Decision Tree registered the highest at 58%. Ironically, Decision Tree AUC showed the lowest at 0.4755.

**Discussion**

From the visualization to the correlation plot, news headlines does not seem to affect the stock prices. However, there could be many factor which affect the stock prices thus by just using news headlines as indicators is insufficient. Furthermore, more data pre-processing needs to be done for crawling news headlines, to determine the type of news headlines which truly has effect on the stocks prices. To achieve this, an AI algorithm needs to be developed to be able to screen any positive or negative news of the stocks online.

For the modelling and evaluation part, by taking into account more factors such as rating agencies score, the model might be able to improve the predictability of the stock price movements. The attributes in the dataset are insufficient to predict stock price movements, thus more data sources can be sourced from other sources such as Bank Negara evaluation and rating agencies to enhance model performance.

**Conclusion**

Machine Learning has been widely used to predict stocks movement. More research needs to be done in order to enhance the model accuracy but this technology would prove useful as more robust algorithms are developed.