Report

Milestone 1

Acquisition of data

The stock price data was extracted from The Star Online and Investing.com. Some of the news data was extracted from klse.i3investor.com. The main packages used to extract the stock data and stock news was BeautifulSoup. After scrapping all the data, we have done some processing to manage the data into Dataframe for later analysis. The code can be found in github (Web\_Scrapping\_v1.ipynb and Stock\_News\_Scraping.ipynb)

Milestone 2

Management of data

The purpose of this milestone is to demonstrate we are able to import our data in a proper data warehouse or a database. For that, I have used to Cloudera VM and started the HUE services. In HUE, we have HIVE service available. First, we upload our csv data into HDFS from local using HDFS command. The uploading process can also be done using the interface provided by the HUE. Next, I used to table creation facilities to create Hive table. The table was able to be query. For more details, please refer to the video Milestone2.

Milestone 3

Processing of data

The purpose of this milestone is to demonstrate how to do covariance of the stock data. It was done in Jupyter notebook. The stocks that were selected in this example is AHEALTH, CIMB, GENTING, MYEG and TOPGLOV. The covariance can be done using the function cov(). I also calculated the portfolio variance and portfolio risk. Each stock was assigned with a weightage of 20%. Then, the portfolio variance was calculated by the dot product of the annualise covariance matrix and the weightage matrix. While the portfolio risk was calculated by finding the standard deviation of the stock price. The second part of the code was to show the PCA analysis. The PCA was able to extract the main component for the open, high, low and last price of the data. The code can be found in github (Covariance matrix of stock data.ipynb)

Milestone 4

Interpretation of data

The purpose of this milestone is to demonstrate if the stock news was correlated to the stock price. I was able to show that some factors were important as it move the stock price. For example, if a stock was upgraded by institution or analyst, the stock normally will reacted to that news and the price shut up. The upgraded can be interpreted as a positive sign that the company earning result will be better in near future. Some stock was sensitive to the global market. For example, Petronas to some extend affected by the crude oil price from US. In Malaysia, I found that EPF is the one that supporting the market and the acquire or disposal of the share will affect the stock price as it hold a lot of share for most of the blue chips company. The code can be found in github (Stock Interpretation v2.ipynb)

Milestone 5

Communication of insight of data

The purpose of this milestone is to use machine learning to predict the stock price. I used decision tree and random forest to predict the whether the percentage change is positive or negative for the day. For this milestone, I have manually get more historical data from Investing.com as the website allow user to extract historical data easily with much hassles. In the example, I have used Maxis as example. The result shown that decision tree generate a better accuracy than random forest. The code can be found in github (Stock market prediction using ML.ipynb)

Milestone 6

Provide a user with recommendations around optimal actions to achieve investment objectives such as profits and return on investment.

I have generated much more features as compared to the previous milestone. The features generated as based on the market technical indicator. I have used momentum based, volume based and volatility based indicator. Example of momentum based are ADX, APO and CCI. Example of volume based are OBV and ADOSC. Example of volatily based are ATR. There was a package called TA-Lib that provide all the technical indicator that can be called in Python. The code can be found in github (Stock market prediction using ML\_v1.ipynb)