Faculty of Computer Science and Information Technology

Master of Data Science
Semeter I

Data Mining

WQD7005

Project Title

Investment recommendation in Malaysia stock based on a data mining approach

Khashayar Namsehchi
WQD170034

Contents

Introduction

Investing in stocks is one of the areas of research and hot topics for analysis. There were various theories and methods used to study and analyze stocks, influential factors, and effects. In addition, there are different views on suitable investment strategies. The aim of this project is to identify and integrate a data set with several representations, with which you can use machine learning methods to recommend potential reserves for investment, One's investment decision making generally influenced by different types of behavior and psychological (Ricciardi & Simon, 2000). To achieve their financial goals in investing, investors may need good investment planning. Good investment planning involves making the right decisions among investors. They must choose the right investments and manage resources for various types of investments in order to make a profit, while at the same time, investment risks can be avoided. Malaysia's capital market, also known as Bursa Malaysia, provides various investment products such as stocks, bonds, warrants, mutual funds, etc.

Data mining is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. On the other hand, during this data mining assignment the researcher use the multi domain sources of data such as Twitter[1], Financial data[2], Stock data[3], and Google trend[4] to achieve the best result for this assignment. Therefore, in investigating across the Twitter the researcher uses text analysis for sentiment approach to find the positive or negative tweets across the desire company.

1- Twitter 2- Thestar 3- i3investor 4- Google Trends

Dataset Information

As part of the six stages of the individual work, The researcher looked at stock data for 3 months from the Market Watch portal at “thestar” Malaysia and scanned the opening, closing, maximum and minimum prices, as well as the trading volume for all share since August 2019 to October 2019. From this web crawl task in Millstone 1, we collected information on 1500 stocks, for a total of 7500 data points. In addition, we reviewed the latest financial information from the i3investor portal. Tweets were scanned using the Tweepy package. Finally, data is downloaded from Google Trends, which list and scale the number of requests from stock companies.

No.	Dataset	Description
1	Stocks	Crawled opening, closing, high and low prices as well as trade volume Source: The Star online, Market Watch
2	Financial	Crawled PBT, EPS, Revenue, dividend and many more Source: i3investor
4	Tweets	Crawled Twitter messages, date and author Source: Twitter
5	Google Trends	Search queries relating to company on Google. *Source: Google Trends*

1- Tweepy

METHODOLOGY

The basic rule of investing is portfolio diversification to include several investments in one portfolio. If one investment loses money, other investments may offset or make higher returns. There are different views on this theory. Warren Buffett said that "diversification is a defense against ignorance. It is pointless if you know what you are doing." In fact, he is against the theory of diversification, and it will do more good if you focus on the portfolio. Traditional portfolio theory says that maximizing the number of investments to a maximum limit can minimize the risk of the entire portfolio. They recommended from 20 to 30 stocks to achieve an optimal portfolio size, in which the risks of diversification are minimized, while not compensating for the benefits of diversification, that is, portfolio profitability is not sacrificed (Wan & Wang, 2010). In accordance with this study, the researcher will investigation on the multi machine learning approach to find best fit algorithm for this study that comes from model compression of SAS enterprise miner.

Average of return and volatility of stock

- Average of return

  Stock returns are calculated based on changes in daily stock prices. In particular, the logarithmic return formula is used. The advantage of using logarithmic differences is that this difference can be interpreted as a percentage change in stock but does not depend on the denominator of the share.

$$AverageStockreturn = \frac{\sum \log(closedpricei) - \log(closedpricei - 1)}{Days}$$

- Average of volatility

  Volatility is a statistical measure of the dispersion of returns for a given security or market index. In most cases, the higher the volatility, the higher the risk of stocks.

$$AverageVolatilityreturn = \sqrt{\frac{\sum(dailyreturn - mean(return))^2}{Days}}$$

Price to Earnings (P/E) ratio

Having identified the various risk categories for stocks, the next step is to find out which stock is worth investing in. The price-earnings ratio (P / E) is the most widely used as a factor for investing. It reflects the amount that investors are willing to pay for shares in exchange for profit. The higher the P / E ratio, the higher the investor's expectations for earnings. A high P / E ratio may also indicate a revalued share, as a result of which the share price is much higher relative to its profit, for example. The stock price is $ 100 and earnings per share are $ 2, which gives a P / E ratio of 50, the writer recommended that investors compare the P / E ratio and the growth of the company (Swan, 2018).

Sentiment (opinion) analysis

Sentiment analysis is an arrangement to extracts the feelings, opinions, thoughts, and behavior of people from a text data by Natural Language Processing (NLP) methods (Danneman & Heimann, 2014). In addition, sentiment analysis also known as Opinion mining, with a stress on the problem of text classification.

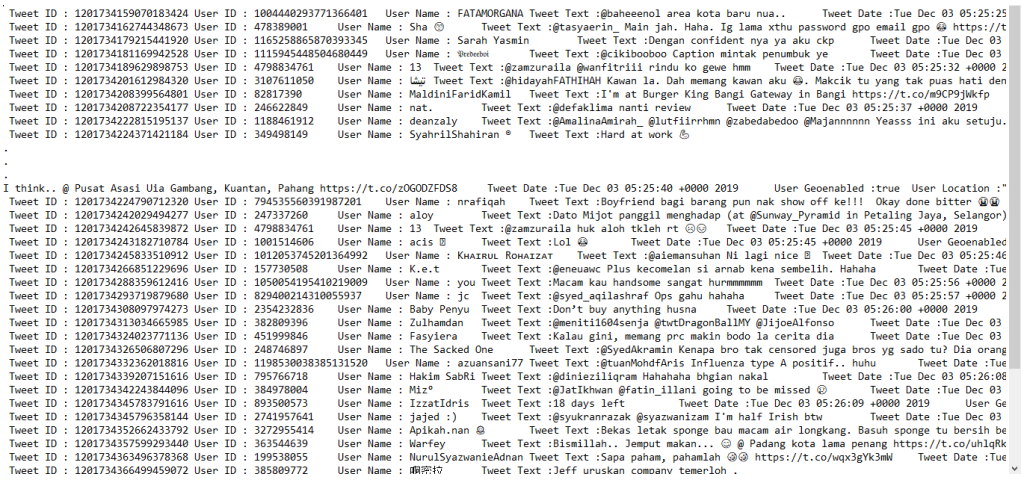
Negative market sentiment has a significant effect on stock prices (Allen, McAleer, & Singh, 2019). Investors will act immediately and sell their positions if losses are expected down. To better understand the effect of sentiment, this project explores the news and sentiment on Twitter about the company and how it affects stock prices. In addition, interest in the company will be measured based on the number of search queries on the Internet using data on Google trends.

Result and Milestone overview

Milestone I. Data acquisition

During this research the researcher using open source Python[1] version 3.7.4 via Jupyter[2] version 6.0.1 to prepare maximum reproducibility and using powerful libraries such as Pandas[3], NumPy[4], BeautifulSoup[5], NLTK[6], TextBlob[7], Tweepy[8], Datetime[9], Json[10] and etcetera for this study.

During this stage, researcher should be collecting the data from multiple sources, for achieve this target the researcher needs to sign as developer in Twitter's developer area and get authorization cardinal such as the API's code and API's secret that be use within Twitter API for communication with Twitter contents. On the other hand, should investigate the stock and finance website to find the website design pattern to extract the data from their websites. Also, use the latitude, longitude in t in the part of coding for retrieve tweets from Malaysia, in this case collected data and their Metadata in JSON format.



Sample of raw Twitter dada

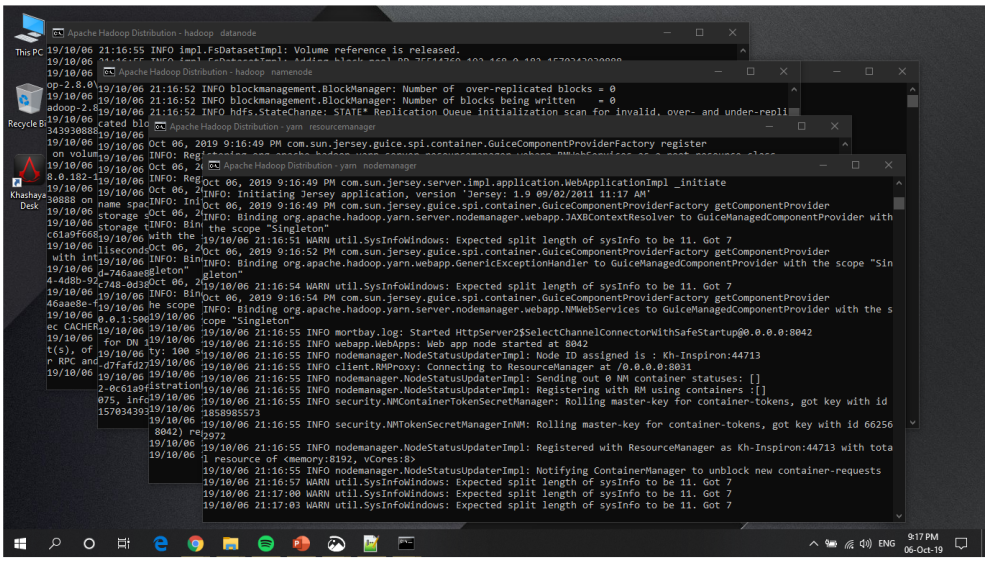
1- Python 2- Jupyter 3- Pandas 4- Numpy
5-Beautful Soup 6- Nltk 7- Textblob 8- Tweepy
9- Datetime 10- Json

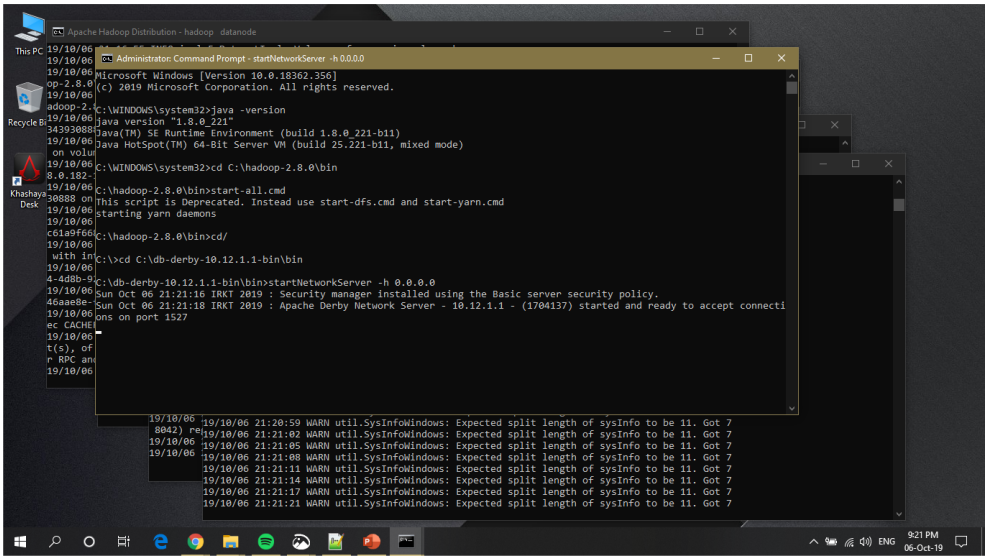
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
64	AAX-WA OR AIRASIA	28-08-19	AAX-WA	Main	0.07	0.07	0.075	0.065	1979800	-0.005	-0.06667	0.3928	Positive	Down	
65	AAX-WA OR AIRASIA	01-08-19	AAX-WA	Main	0.075	0.065	0.075	0.065	3782100	0.005	0.071429	0.18308	Positive	Up	
66	AAX-WA OR AIRASIA	04-08-19	AAX-WA	Main	0.065	0.07	0.07	0.065	272900	-0.005	-0.07143	0.1204	Positive	Down	
67	AAX-WA OR AIRASIA	05-08-19	AAX-WA	Main	0.065	0.07	0.075	0.065	1864200	0	0	0.100475	Positive	Unchanged	
68	AAX-WA OR AIRASIA	06-08-19	AAX-WA	Main	0.07	0.065	0.07	0.065	127800	0.005	0.076923	-0.12254	Negative	Up	
69	AAX-WA OR AIRASIA	08-08-19	AAX-WA	Main	0.07	0.07	0.07	0.065	1402800	0	0	0.29188	Positive	Unchanged	
70	AAX-WA OR AIRASIA	11-08-19	AAX-WA	Main	0.075	0.07	0.075	0.07	1413000	0.005	0.071429	-0.296	Negative	Up	
71	AAX-WA OR AIRASIA	12-08-19	AAX-WA	Main	0.07	0.075	0.075	0.07	1368100	-0.005	-0.06667	0.166411	Positive	Down	
72	AAX-WA OR AIRASIA	14-08-19	AAX-WA	Main	0.075	0.07	0.08	0.07	5996600	0.005	0.071429	0.687967	Positive	Up	
73	AAX-WA OR AIRASIA	15-08-19	AAX-WA	Main	0.07	0.075	0.08	0.07	4273400	-0.005	-0.06667	0.5745	Positive	Down	
74	AAX-WA OR AIRASIA	18-08-19	AAX-WA	Main	0.07	0.075	0.075	0.07	1694800	0	0	-0.5106	Negative	Unchanged	
75	AAX-WA OR AIRASIA	19-08-19	AAX-WA	Main	0.075	0.075	0.075	0.07	791900	0.005	0.071429	0.1702	Positive	Up	
76	AAX-WA OR AIRASIA	20-08-19	AAX-WA	Main	0.07	0.065	0.075	0.065	649800	-0.005	-0.06667	-0.0148	Negative	Down	
77	AAX-WA OR AIRASIA	21-08-19	AAX-WA	Main	0.065	0.07	0.07	0.065	3314200	-0.005	-0.07143	0.228617	Positive	Down	
78	AAX-WA OR AIRASIA	22-08-19	AAX-WA	Main	0.065	0.065	0.07	0.065	502600	0	0	-0.10758	Negative	Unchanged	
79	AAX-WA OR AIRASIA	25-08-19	AAX-WA	Main	0.065	0.065	0.065	0.065	336800	0	0	0.06835	Positive	Unchanged	
80	AAX-WA OR AIRASIA	26-08-19	AAX-WA	Main	0.07	0.065	0.07	0.065	883000	0.005	0.076923	0.00645	Positive	Up	
81	AAX-WA OR AIRASIA	27-08-19	AAX-WA	Main	0.065	0.07	0.07	0.065	1454000	-0.005	-0.07143	0.33354	Positive	Down	
82	AAX-WA OR AIRASIA	28-08-19	AAX-WA	Main	0.07	0.065	0.07	0.065	981100	0.005	0.076923	0	Neutral	Up	
83	AAX-WA OR AIRASIA	29-09-19	AAX-WA	Main	0.065	0.07	0.07	0.065	367200	-0.005	-0.07143	-0.296	Negative	Down	
84	AAX-WA OR AIRASIA	03-09-19	AAX-WA	Main	0.065	0.065	0.07	0.065	554200	0	0	0	Neutral	Unchanged	
85	AAX-WA OR AIRASIA	04-09-19	AAX-WA	Main	0.075	0.065	0.075	0.065	2650700	0.01	0.153846	-0.04526	Negative	Up	

clean and join all the sources of data

Milestone II. Hadoop installation and Data transform

According at this stage the researcher installs the Hadoop and all the requirement data in standalone to preparing data transformation to that.

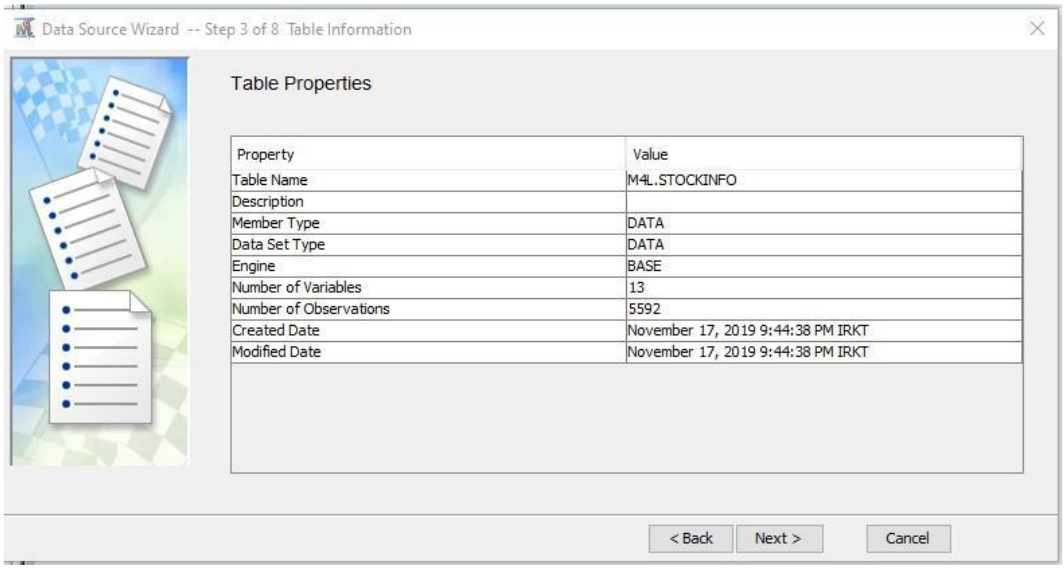




SAS Enterprise miner

According to this stage the researcher starts working with SAS Enterprise miner to working with the data and find the pattern of data and finally achieved to the knowledge discovery.

After transfer all the data from Hadoop and converting to the SAS data type with helps the SAS Enterprise Guide, the researcher can import the data set to the SAS software.



Clustering the Malaysia Stock

Hierarchical agglomeration clustering (HAC) was performed on filtered 450 stocks. The values of the correlation matrix are used as an indicator of similarity. 5 clusters were considered the best cutting point. Clusters received through the HAC are then mapped to profitability and Risk values. A regression line is also set for each cluster to view the projection / trend line of the respective clusters. As shown in Figure 4.1, we can see that cluster 3 is a high-risk cluster, as a result of which stock returns sharply increase with increasing risk. Cluster 1 is a low-risk cluster in which stock returns are risk lower compared to other clusters. Cluster 5 is considered a medium risk cluster, resulting in a moderate return with increasing risk.

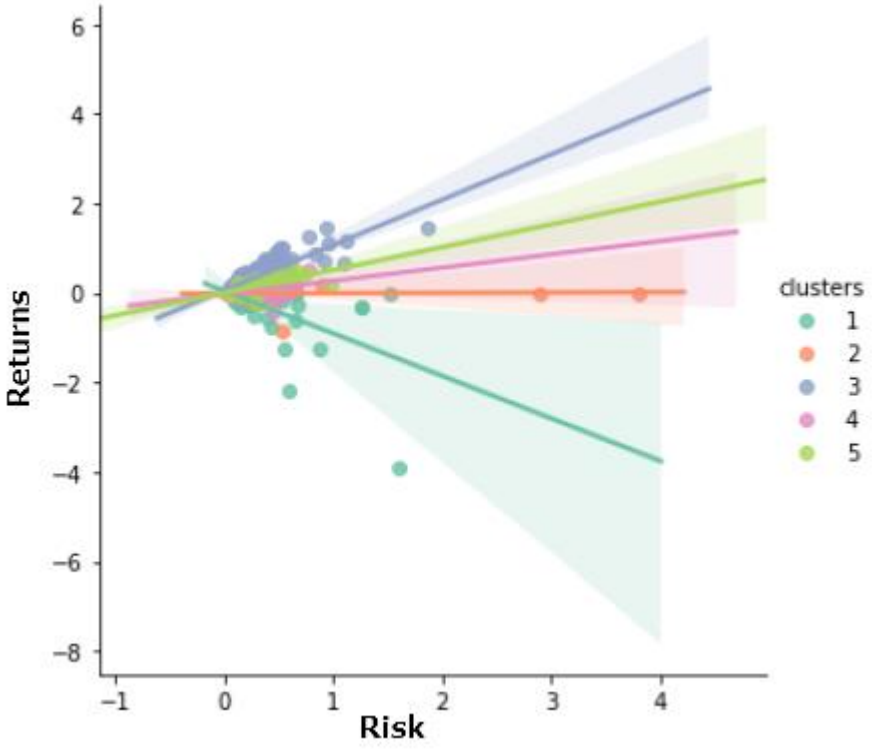


Figure 4.1

Determine Malaysia stock value

The latest financial information of the companies was scanned. However, this does not contain information for all stocks. Further filtering should have been carried out. A low-risk cluster is reduced from 50 to 30 shares, a low-risk cluster is reduced from 65 to 32 shares, and a high-risk cluster is reduced from 180 to 90 shares. The P / E ratio was calculated for all three clusters by dividing stock prices by earnings per share (earnings per share). This is then compared with annualized growth (YoY) and company earnings before taxes (pbt). The key step is filtering undervalued and revalued stocks. At the same time, we can be sure that we are investing in valuable or promising shares. To do this, filtering is based on the rules, in accordance with which only shares with a positive rate of return, a growth rate of at least -1% and a P / E ratio of 0 to 3 were selected (Figure 4.2).

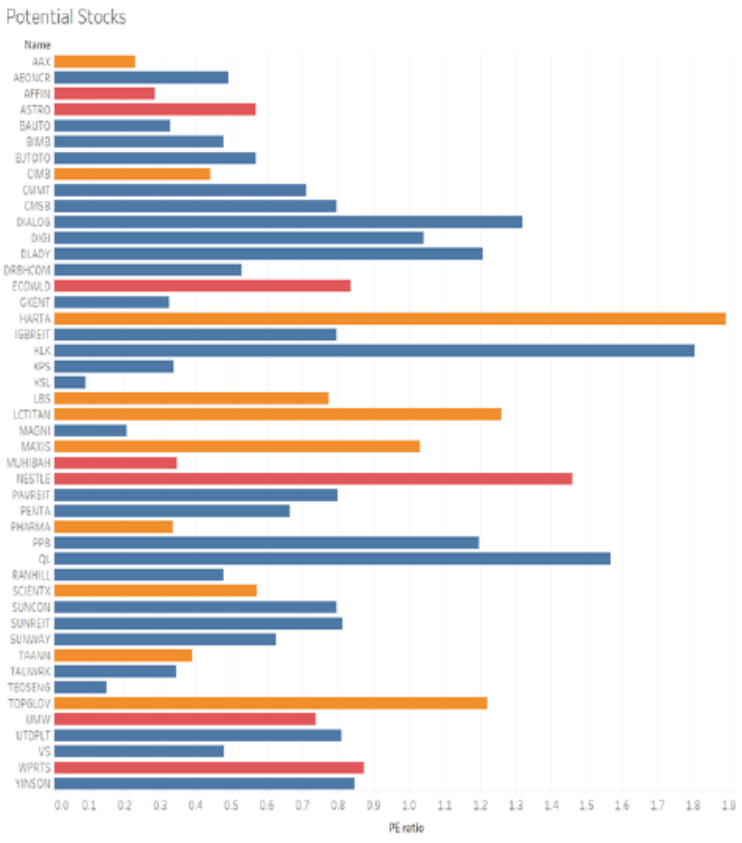


Figure 4.2

According to the above result Four stocks were picked from the list, namely AirAsia X Bhd, MAXIS Bhd, NESTLE (Malaysia) Bhd and Aeon Credit Service (M) Bhd in view that they comprise many tweets during the 3 months period as well as representing the different risk categories and business sectors.

Company	Sector	Risk category
AirAsia X Bhd	Aviation	Low risk
MAXIS Bhd	Telco	Low risk
NESTLE (Malaysia) Bhd	Consumer	Medium risk
Aeon Credit Service (M) Bhd	Finance	High risk

Google Trends stocks overview

According to (Figure 4.3) Google Trends allows to view the number of searches for a specific word on Google. Results may be an indicator of current demand for company services. This is an experimental approach to find out if a significant increase in search volume is associated with an increase or decrease in stock prices.

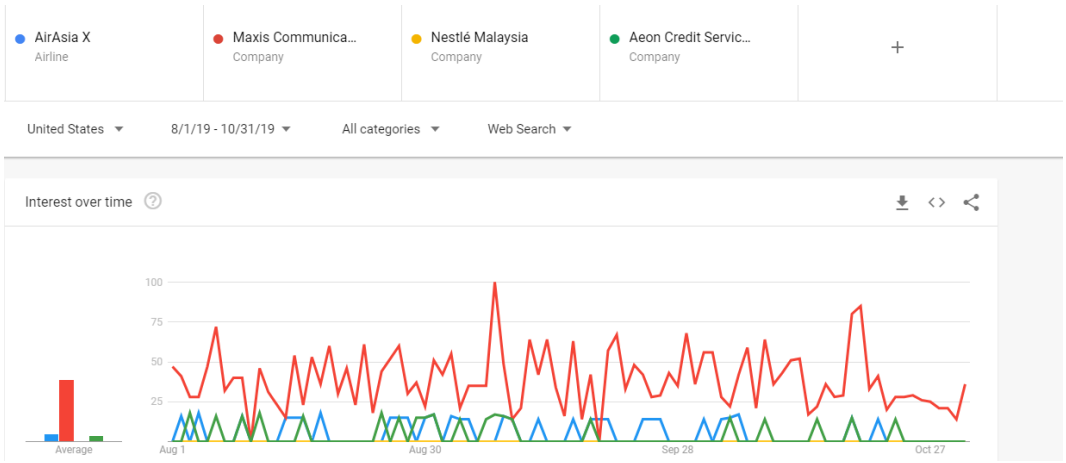


Figure 4.3

Machine learning approach of Malaysia stock

The machine learning classification model requires a target variable. Despite this, the task of creating a target variable for our time series inventory data is not easy. The original idea was to create buy / not buy classes based on stock prices. However, this would lead to a bias classification model. Further, determining a specific purchase date is very subjective. Therefore, while ensuring a fair model, the machine learning classification model will study the detection of changes in stock returns based on sentiment and interest. The target variable is created based on changes in stock returns. Three classes were chosen, namely: down (decrease in profitability), up (increase in profitability) and none (unchanged). The variables used for this model are mood ratings for news, mood categories for news, mood ratings for tweets, mood categories for tweets and interests (data from Google trends). Stock data are excluded to reduce model bias. This is due to the fact that stock returns are correlated with stock prices and trading volume.

The researcher during this study used the various machine learning models to such as Decision Tree, Regression, High Performance (HP) Neural, High Performance (HP) Forest and High Performance (HP) Cluster to comparison and find the best machine learning approach base on SAS Enterprise miner (Figure 4.4).

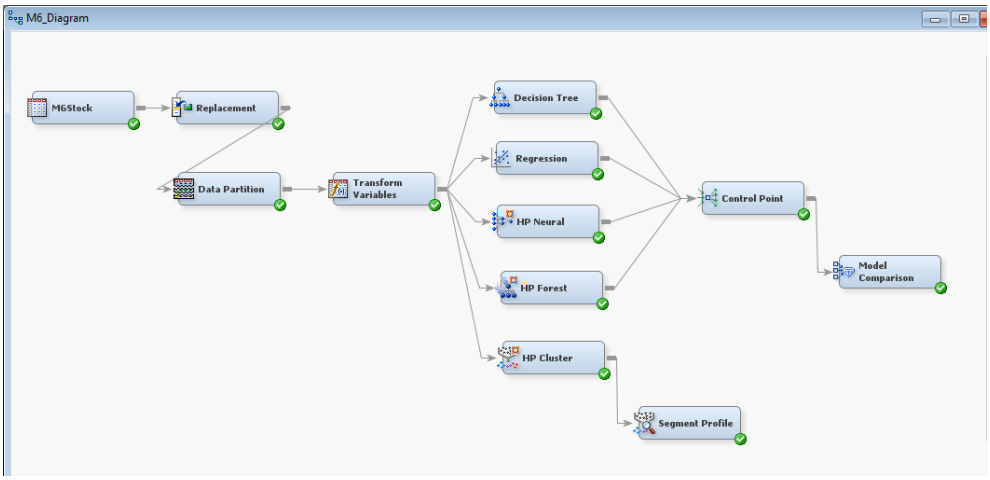
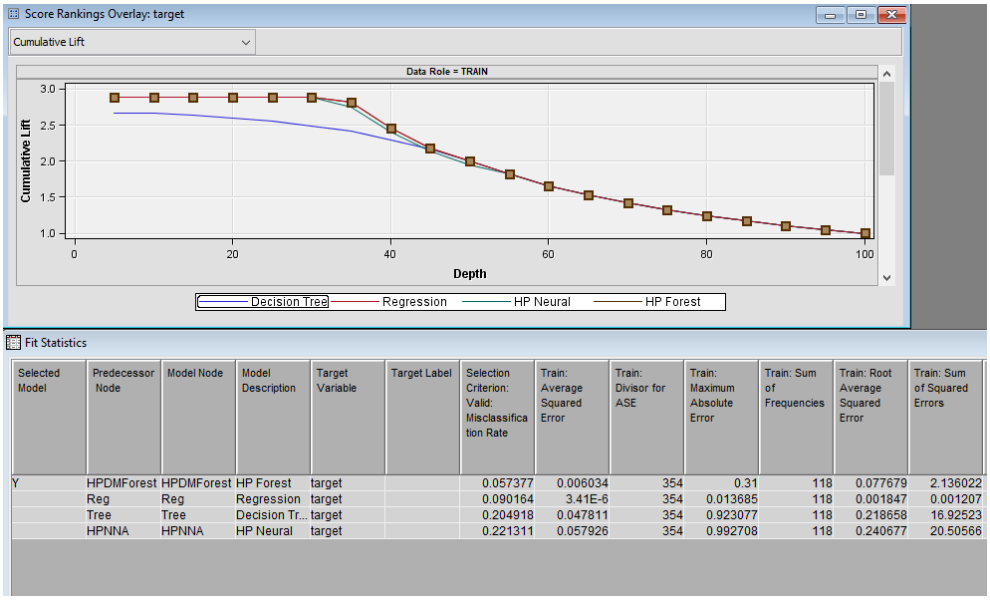


Figure 4.4

In conclusion, based on the SAS model comparison the High Performance (HP) Forest (Figure 4.5) is the best suite model for this study.



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors
Y	HPDMForest	HPDMForest	HP Forest	target		0.057377	0.006034	354	0.31	118	0.077679	2.136022
	Reg	Reg	Regression	target		0.090164	3.41E-6	354	0.013685	118	0.001847	0.001207
	Tree	Tree	Decision Tr...	target		0.204918	0.047811	354	0.923077	118	0.218658	16.92523
	HPNNA	HPNNA	HP Neural	target		0.221311	0.057926	354	0.992708	118	0.240677	20.50566

Figure 4.5

Conclusion

In conclusion, the key findings of this project are diversified portfolio that includes stocks of various risk categories is a likely option. This can be determined using the clustering method without supervision, using data on the yield and volatility of stock prices. Secondly, revalued and undervalued stocks should be filtered out to focus investments on potential stocks with good value. This can be achieved by comparing the P / E ratio with the growth of the company. Thirdly, the Google Trends search that affect the profitability of stocks and should be monitored throughout the entire investment period. Twitter sentiment needs to be explored and expanded beyond the scope of this project in order to properly study its effect on stock prices. last word, It is worth noting that the decline in stock returns can be better predicted based on the sentiment (opinion) analysis and interests of the market due to the fear of investors to lose money.

References

Hargreaves, C., Dixit, P., & Solanki, A. (2013). Stock portfolio selection using data mining approach. *IOSR Journal of Engineering, 3*(11), 42-48.

Ricciardi, V., & Simon, H. (2000). What is behavioral finance? *Business, Education & Technology Journal, 2*(2), 1-9.

Swan, A. (2018). Five Powerful Ways To See If A Stock Is Overvalued. Retrieved from https://www.forbes.com/

Wan, J., & Wang, R. (2010). Empirical research on critical success factors of agile software process improvement. *Journal of Software Engineering and Applications, 3*(12), 1131.

Anaconda, I. (2019). Anaconda, Inc. Retrieved June 16, 2019, from Anaconda: https://www.anaconda.com/distribution/

Bo Pang, Lillian Lee. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 1-135. doi:10.1561/1500000011

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press