UNIVERSITY OF Southampton

MANG6554 Advanced Analytics

Individual Coursework

Student ID:31302815

Word Count:2723

Part A

Question 1.1

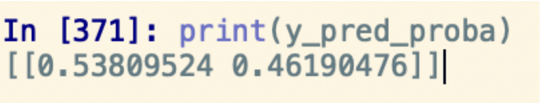
A Bayesian probabilistic model can be used to build Naive Bayes reasoning, which provides a posterior probability to $P(Y = yj|X = xi)$. (Berrar, 2018) According to Bayes' theorem, the Posterior probability can be obtained: $P(yj|xi) = \frac{P(xi|yj)*P(yj)}{P(xi)}$, which $P(xi|yj)$ refer to the likelihood, $P(yj)$ refer to class Prior Probability, and $P(xi)$ refer to Predictor Prior Probability.

From the given dataset, let Label = 1 present customer who purchases phone, Label = 0 present customer who did not purchase the phone, while Platform = 1 means Android Platform user. The probability of customer purchase, given it is an Android user can be listed in the following:

$$P(Label = 1|Platform = 1) = \frac{P(Platform = 1|Label = 1) * P(Label = 1)}{P(Platform = 1)}$$

Filter Pandas DataFrame data via python to count the needed number. Among 1000 people, 420 Android platform users and 506 customers purchased the phone. Also, while a total of 506 customers purchase the phone, 194 people are using the Android platform. After calculation, the probability is 0.4619.

As all the data belong to discrete features, the categorical Naive Bayes classifier is selected to solve the question. By applying statistical package, from "sklearn.naive_bayes" import "CategoricalNB", then using "CategoricalNB"(set alpha=0) and fit Naive Bayes classifier according to X, y, where X = platform and y = label. Using ".predict_proba" to verify the probability is 0.46190476 (Figure 1).



```
In [371]: print(y_pred_proba)
[[0.53809524 0.46190476]]
```

Figure1-The probability of P(Label=1|Platform=1)

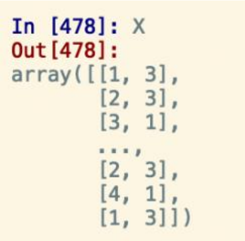
Question 1.2

From the given dataset, Prior_purchase =3 refer to a customer purchase last purchase between 6-12months. The equation that a customer purchases, given it is an Android user and had a purchase in “between 6 and 12 months” can be solved in the following:

$$P(Label = 1|\ Platform = 1, Prior\ Purchase = 3) = \frac{P(Platform = 1, Prior\ Purchase = 3|Label = 1) * P(Label = 1)}{P(Platform = 1, Prior\ Purchase = 3)}$$

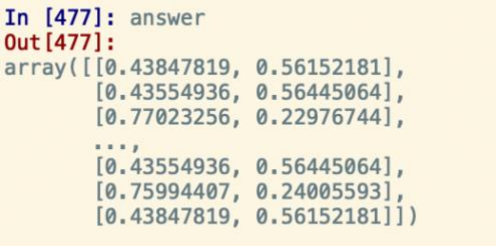
In accordance with Law of total probability, P(Platform =1,Prior Purchase =3) equal to P(Label = 1)*P(Platform =1|Label = 1)*P(Prior Purchase = 3|Label = 1)+P(Label = 0)*P(Platform = 1|Label = 1)*P(Prior Purchase = 3|Label = 1),which is 0.18571778. Moreover, since Naive Bayes classifier assume that events are independent, the equation P(Platform = 1,Prior Purchase = 3|Label = 1)*P(Label = 1) can be extend to P(Platform = 1|Label = 1)*P(Prior Purchase = 3|Label = 1)*P(Label = 1), which is 0.10428458. Therefore, the result that calculating though python will be around 0.5615.

Similarly, the categorical Naive Bayes classifier is chosen since all of the data are discrete variables. Using the statistics package, from "sklearn.naive bayes" importing "CategoricalNB," then apply "CategoricalNB" with setting alpha = 0 to fulfill the requirement, and fit Categorical Naive Bayes classifier X, y via ".fit", where X = platform, Prior Purchase and y = label. As the result, given Platform = 1 and Prior Purchase = 3 (Figure2), the probability of Label = 1 is 0.56152181 (Figure3) using ".predict proba".



```
In [478]: X
Out[478]:
array([[1, 3],
       [2, 3],
       [3, 1],
       ...,
       [2, 3],
       [4, 1],
       [1, 3]])
```

Figure 2-Array for Platform and Prior_Purchase



```
In [477]: answer
Out[477]:
array([[0.43847819, 0.56152181],
       [0.43554936, 0.56445064],
       [0.77023256, 0.22976744],
       ...,
       [0.43554936, 0.56445064],
       [0.75994407, 0.24005593],
       [0.43847819, 0.56152181]])
```

Figure 3 - *The probability of P(Label= 1| Platform=1,Prior Purchase=3)*

Question 1.3

Naïve Bayes is not an optimal solution as a predictive model. It is often misunderstood that the Naive Bayes model has the lowest error rate compared to other classification methods.

This is because the naive Bayes model assumes that the attributes are independent of each other when the result class is given. In practice, this assumption is often incorrect. The classification result is poor when the correlation is high. Figure 4 shows that the variables in each category (label, platform, prior purchase, premium, and redeem card) are related.

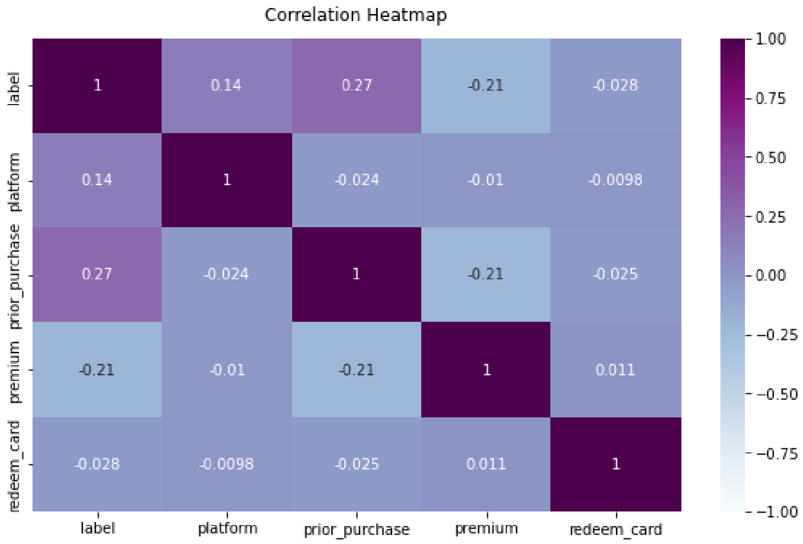
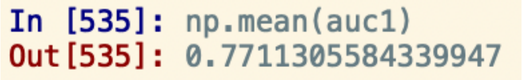


Figure 4- Correlation Heatmap

Therefore, take the Random Forest algorithm for comparison as it is suitable for dealing with categorical data. Ali et al. (2014) suggested that Random Forest can minimize the correlation between trees via randomly picking features, which enhances prediction power and further improves efficiency. Other benefits of the Random Forest include handling the problem of overfitting and the automated generation of variable significance and accuracy. Furthermore, using random sampling has more accurate forecasts.

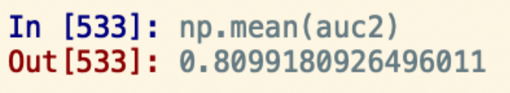
The comparison of Naïve Bayes and Random Forest for a predictive model is presented. To get the AUC, “cross_val_score” from “sklearn.model_selection” is selected to estimate the expected accuracy of the model on training data and test for goodness of fit. The results indicate that the Random Forest classification for a predictive model has achieved the highest accuracy at 0.8099(Figure 6) compared to Naïve Bayes classification methods at 0.7711(Figure 5).

As a result, it can be concluded that the Random Forest achieves higher classification performance than Naïve Bayes, and it is accurate and efficient in the large dataset.



```
In [535]: np.mean(auc1)
Out[535]: 0.7711305584339947
```

Figure 5 - Accuracy of Naïve Bayes classification



```
In [533]: np.mean(auc2)
Out[533]: 0.8099180926496011
```

Figure 6 - Accuracy of Random Forest classification

Part B

The Business Innovation Development Plan is based is based on a journal article, Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. The source is in the International Journal of Information Management. The AJG ranking for this journal is 2, which shows the journal is known and reliable in the field.

- **The target challenges**

In modern society, many businesses and researchers are interested in analyzing social media data because this data contains essential information, including consumer suggestions, comments, and ratings on products and services. Nevertheless, social networks and how they are utilized are continually developing, so the challenge's complexity is continuously increasing.

The specific challenge to target is collecting reviews and ratings from TripAdvisor, booking.com, Google Reviews, and Google trend and applying techniques related to aspect-based sentiment analysis methods such as data crawler, data preprocessing, sentiment-sensitive tree construction, convolution tree kernel classification, aspect extraction and category detection to analysis reviews and rating those data and leveraging software to produce the visualization (Chang, Ku, and Chen, 2019). As a result, by this business plan, recommendations such as how to improve service, product quality, promotion, and business reputation can be suggested to the whole hospitality business, including Food and Beverage Industry, Lodging Industry, Recreation Industry, Travel and Tourism Industry, Meetings and Events Industry.

- **Data collection**

The data types chosen to collect are the rating and reviews of hospitality businesses such as reputed hotels, restaurants, or amusement parks on TripAdvisor, booking.com, and Google Reviews. Those reviews are unstructured data as they do not have a specific format. Furthermore, the structure data such as customer rating, customer profiles, and travel type can also be collected on that website. They are also searching for relevant information on Google trends to do the visualization. TripAdvisor is a global interactive travel forum website. The reviewers can provide ratings in several aspects, such as the Overall Score, Location, Cleanliness, Service, Rooms, Sleep Quality, and Value. The rating is from score one(unsatisfied) to five(satisfied). Booking.com is one of the world's major travel e-commerce companies, mainly providing global accommodation booking services. In addition to the reviews, people can provide a rating from a score of 1 to 10 in several aspects, such as the Overall Score, Staff, Cleanliness, Comfort, Value for money, and Location. As for Google Reviews, people provide reviews and overall scores from one to five stars. After gathering data from those platforms via a data crawler and further processing data cleaning to ensure data quality (Chang, Ku, and Chen, 2019).

- **Techniques to analyse the collected data**

The five main components of the **aspect-based sentiment analysis method**:

Data processing: After collecting data by data crawler, the review data was subjected to some basic preprocessing to set a dataset with those reviews and sentences. First, determine the tone of reviews, defined as 1–2(TripAdvisor, Google),1-6(Booking.com) score rating for negative reviews, and a 3–5(TripAdvisor, Google),7-10(Booking.com) score rating for positive reviews. Second, remove reviews without a labeled aspect. Third, lowercase words and then delete punctuation, stop words three times, and words that appear fewer than five times in the corpus. Finally, stem each word to its root (Chang, Ku and Chen, 2019).

Sentiment sensitive tree (SST) construction：represent the reviews with the SST structure, which is a constituent tree (CT) of review titles enhanced with operations: "decoration", "extension", and "pruning". CT of review titles is effective in sentiment classification because it can model the syntactic structure of the text that affects the sentiment classification performance. SST Decoration, rather than words in verbs or adjectives, finds the characteristics of emotion closely related to a polarity of sentiment in an SST to acquire a clear sentimental expression. SST Extension, connect extra sentences to richer context details if the review title is too short of providing sentiment information. SST Pruning, pruning the redundant elements to maintain the effectiveness of sentiment classification (Chang, Ku and Chen, 2019).

Convolution tree kernel classification: to resolve sentence similarity, the convolution tree kernel (CTK) is selected to apply in a support vector machine (SVM). SVM is a supervised learning method that uses the principle of statistical risk minimization to estimate the hyperplane of a classification. The basic concept is to find a decision boundary to maximize the margins between the two classes so that they can be perfectly separated. Apply CTK to develop a classification for each structural type, identifying sentiment in reviews (Chang, Ku and Chen, 2019).

Aspect category detection: through the attributes such as travel type, customer's profile, and location of reviews, detect the aspect that can influences sentiment (Chang, Ku and Chen, 2019).

Data visualization: using Tableau, apply visual analytics techniques to investigate the extracted data, e.g., ratings, aspect-sentiment, and traveller categories.

- **Data storing**

Non-relational data like social media is ideal for storing in the data lake. After data collection, keep the data in its original format as a data lake capable of holding all forms of data, including structured, semi-structured, and unstructured data, without defining data structures, schema, or transformations (Amazon Web Services, 2022). It also allows for the efficient import of a large amount of data, enabling the acquisition of vast volumes of data from several sources, including TripAdvisor, Booking.com, and Google Reviews. Subsequently, capture the data needed for analysis via SQL (for structured data) and NoSQL (for unstructured data).

- **Novelty and potential business insight from the project plan**

This project plan has some novelty and potential business insight. First, this comprehensive model demonstrates how to manage review and rating data consistently across platforms. Because each platform has its own set of reviews and scoring criteria, integrating data and developing a framework for collecting and processing heterogeneous data and further extracting and classifying aspect-level information. Finally, make a visual representation of data from TripAdvisor, Booking.com, Google Reviews, and Google Trends. The outcomes of the analysis could then be used to improve hospitality services and uncover marketing opportunities.

- **Potential risk and limitations**

Those travel platforms contain a wealth of word-of-mouth Information that may influence the behaviour of other reviewers. Consumers can be biased in writing reviews due to online herding and self-selection bias, causing misleading when analysis (Lee, Xie, Besharat and Tan, 2017). Gerrath and Usrey (2021) stated that influencers are likely to influence the direction of the entire review. Fake reviews may also cause errors in analysis. For instance, merchants may ask guests to leave a positive message to receive free redemption for goods or services. The above factors that may lead to review distortion can be considered, and more designs can be explored in the future.

Part C

Question 3.1

The Momentum strategy is about the Moving Average Crossover method. The Google stock is selected to develop the trending strategy. Initially, make two separate Simple Moving Averages (SMA) of a time series with different lookback periods. When moving averages cross, buy and sell signals will be generated based on historical market price data. The short moving average is set for 20 days, and the long moving average is set for 130 days. In figure 7, purchase the stock in the short moving average line across the long moving average line in February 2020. Hold it until March 2020 and sell the stock since the short moving average line falls below the long moving average line. There are two buying signals and one selling signal issued in the graph. It is worth noting that there is a significant drop in the stock around March 2020. The stock markets worldwide are affected by the circuit breaker of the U.S. stock market during the covid period (Coronavirus: U.S. stocks see worst fall since 1987, 2020).

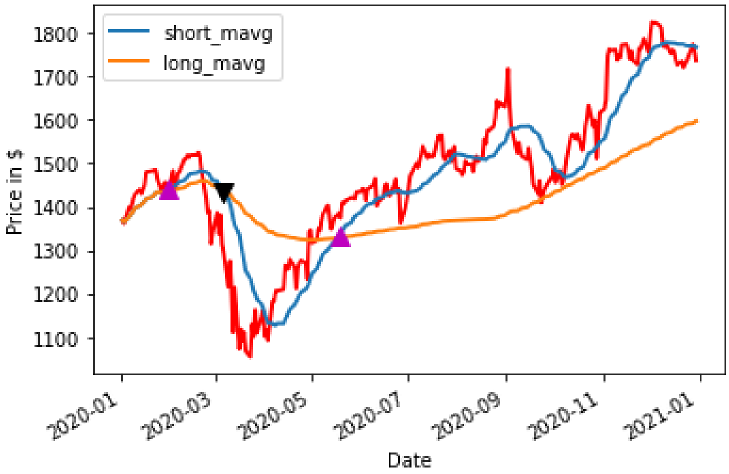


Figure 7 - Simple Moving Average (20days & 130days)

Question 3.2

Applying Backtest to validate the trading strategy. Backtest is a simulation based on historical data to test its performance. It can compute the sum of profits and losses generated by this strategy (Bailey et al., 2016). Investors allotted initial capital of $100,000 to buy google stock in Feb 2020, but the stock value plunged in March, so they lost around $10,000. Subsequently, open the position around May 2020 open the position. From the Backtest can see that there is a total return of $30,000 for the annual year, the rate of return is 30%.

The pros of the strategy are that moving average analysis is relatively simple, allowing beginners to clearly understand the current price trend. Moreover, investors can observe the price trend of the stock, regardless of the accidental changes in the stock price, so that the timing of buying and selling can be automatically selected. Besides, the moving average can automatically select the signals of "buy" and "sell", reducing the risk level. Using the moving average as a buy or sell signal can often yield a good return on investment.

The cons of the strategy are that there has a chance that the total return might not reach the maximum profit. For example, in Figure 8, an investor should purchase the stock at every low point (such as March 2020 and October 2020) and sells it at every high point (such as

September 2020 and January 2021). Moving averages can only allow investors to purchase at a relatively low point and sell at a relatively high point. Furthermore, when transaction costs are factored into returns, the strategy is impractical since it cannot produce positive net profits (de Souza. Et al., 2018).

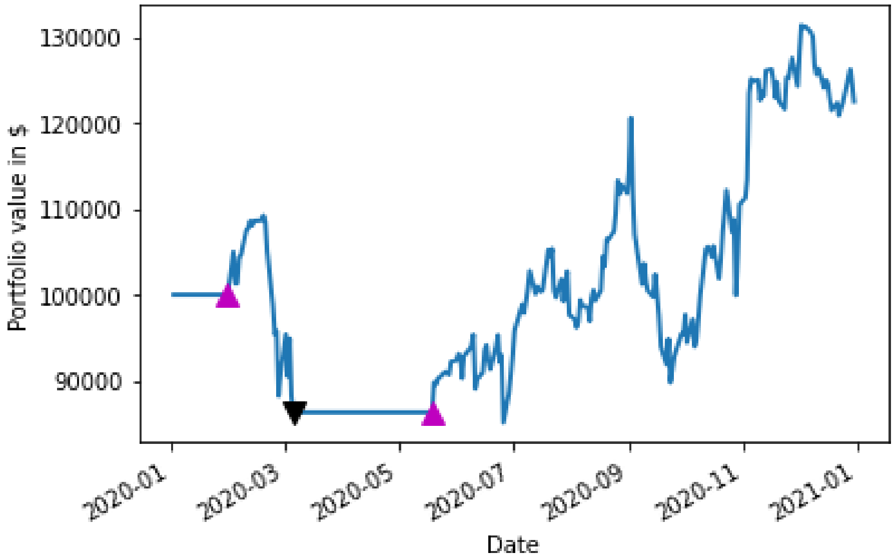
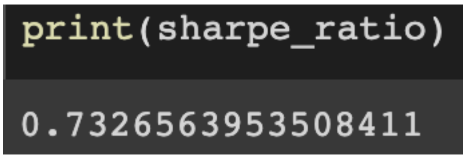


Figure 8 – The result of Backtesting

Question 3.3

To evaluate the strategy, annualized Sharpe ratio to see if it is profitable. Sharpe ratio is a performance measure, used to calculate the return on risk of the backtested strategy (Bailey et al., 2016). The Sharpe ratio is approximately 0.73265(Figure 9), indicate that the operational risk of the stock is greater than the rate of return.



```
print(sharpe_ratio)
0.7326563953508411
```

Figure 9 - Sharpe ratio

To Optimize the strategy, adjust the short moving and long moving windows to 5 days and 20 days because the period is captured only for one year. As it can be seen, a complete transaction contains two reverse transactions, there are six times of complete transactions within one year, which shows better results.

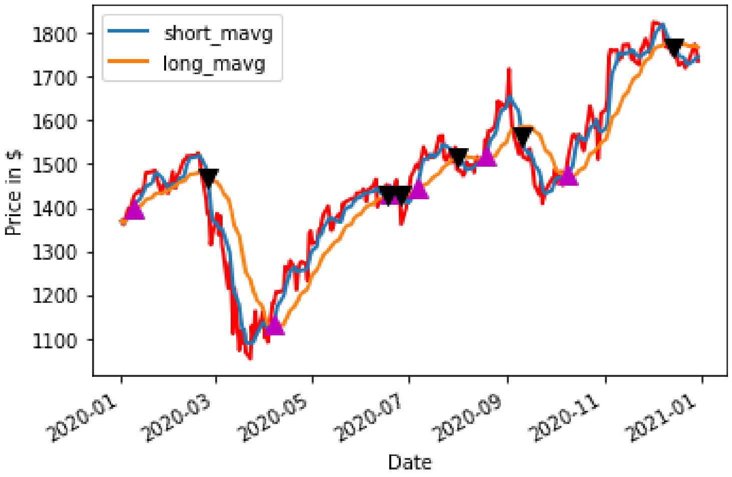


Figure 10 - Simple Moving Average (5days & 20days)

The result of backtesting (Figure 11) shows a total return of $40,000 from the time the stock was purchased in January 2020 to the time it was sold in January 2021. The rate of return on the stock investment is 40%, which is relatively high. Compared to the original strategy, as the range of short and long MA are shorter, the number of transactions in the result is more frequent, and the overall return on investment is also higher.

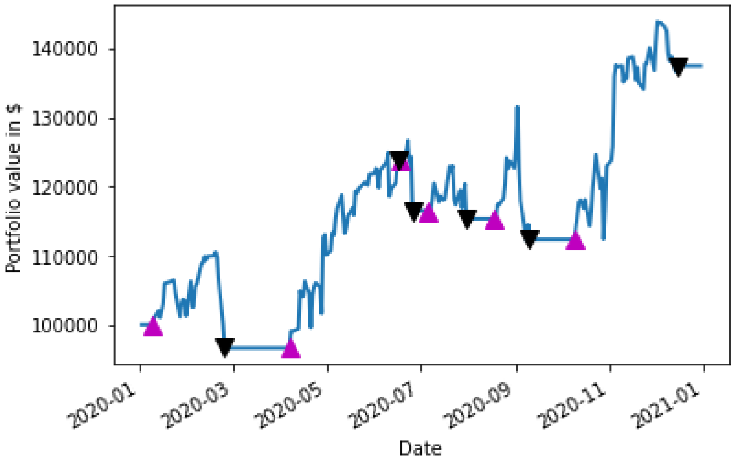
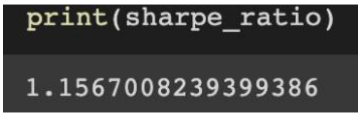


Figure 11- The result of Backtesting (after adjusting)

The Sharpe ratio is approximately 1.1567(Figure 9), indicating that the improved strategy is more profitable than buying and holding the position.



```
print(sharpe_ratio)

1.1567008239399386
```

Figure 12- Sharpe ratio (after adjusting)

Consequently, calculate the max drawdown(orange line) and the daily drawdown(blue line) in the past window days. A maximum drawdown (MDD) is the maximum greatest loss from a peak to a trough of a portfolio before a new peak is attained. (Hayes, 2021)

Based on the drawdown in Figure 13 can see how much an investor will lose the money until getting the profit. The drawdown in Figure 13 shows the amount of money investment can lose until earning profits. A drawdown of 0 indicates that the position has not lost any money. The maximum drawdown of -0.3 indicates that the greatest money that an investor will lose is 30%.

Another option to improve the strategy is to "Take Profit, Stop Loss," which means that when reaching a given percentage of positive return, take the profit, and set the security mechanism to close the position when losing a certain percentage of money to prevent further losses. (CHEN, 2021) This is suitable for investors who want capital preservation.

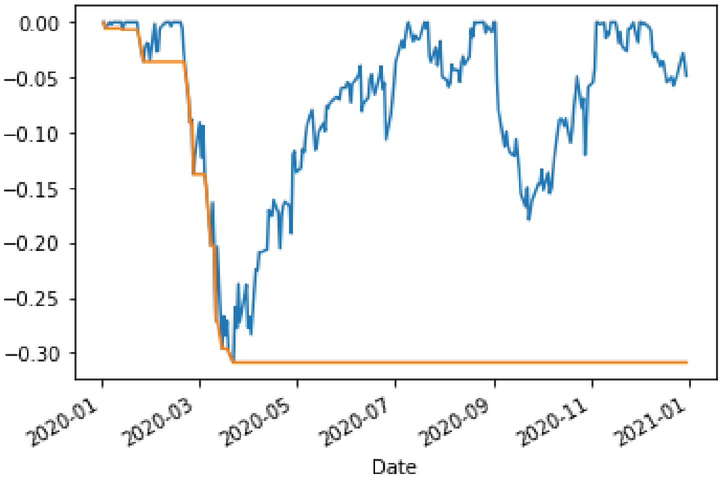


Figure 13-Maximum Drawdown

Reference

1. Ali, J., Khan, R., Ahmad, N. and Maqsood, I., (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), p.272.
2. Amazon Web Services, Inc. (2022). *What is a data lake?*. [online] Available at: <https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/?nc1=h_ls> [Accessed 24 May 2022].
3. Bailey, D.H., Borwein, J., Lopez de Prado, M. and Zhu, Q.J., (2016). The probability of backtest overfitting. *Journal of Computational Finance, forthcoming*.
4. Berrar, D., (2018). Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 403.
5. BBC News. (2020). Coronavirus: US stocks see worst fall since 1987. [online] Available at: <https://www.bbc.co.uk/news/business-51903195> [Accessed 27 May 2022].
6. CHEN, J., 2021. Take-Profit Order - T/P. [online] Investopedia. Available at: <https://www.investopedia.com/terms/t/take-profitorder.asp> [Accessed 27 May 2022].
7. Chang, Y.C., Ku, C.H. and Chen, C.H., (2019). Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. International Journal of Information Management, 48, pp.263-279.
8. de Souza, M.J.S., Ramos, D.G.F., Pena, M.G., Sobreiro, V.A. and Kimura, H., (2018). Examination of the profitability of technical analysis based on moving average strategies in BRICS. *Financial Innovation*, *4*(1), pp.1-18.
9. Gerrath, M.H. and Usrey, B., (2021). The impact of influencer motives and commonness perceptions on follower reactions toward incentivized reviews. *International Journal of Research in Marketing*, *38*(3), pp.531-548.

10. Hayes, A., (2021). *Maximum Drawdown (MDD) Definition*. [online] Investopedia. Available at: <https://www.investopedia.com/terms/m/maximum-drawdown-mdd.asp> [Accessed 27 May 2022].

11. Lee, Y., Xie, K., Besharat, A. and Tan, Y., (2017). Management Responses to Online WOM: Helpful or Detrimental?. *SSRN Electronic Journal*,.