Table of Contents

[Introduction 2](#_Toc90664357)

[Clustering and Classification 2](#_Toc90664358)

[Time Series Analysis 2](#_Toc90664359)

[Association Analysis 2](#_Toc90664360)

[Portfolio Optimization 2](#_Toc90664361)

[Argument 3](#_Toc90664362)

[Clustering and Classification 3](#_Toc90664363)

[Model Evaluation 3](#_Toc90664364)

[Classification 5](#_Toc90664365)

[Time Series Analysis 5](#_Toc90664366)

[Model performance 9](#_Toc90664367)

[Challenges of Stock Forecasting Since the outbreak of COVID-19 11](#_Toc90664368)

[Association Analysis 11](#_Toc90664369)

[Portfolio Optimization 12](#_Toc90664370)

[Methods For Detection of Sensitive Information Using Machine Learning 14](#_Toc90664371)

[Conclusion 15](#_Toc90664372)

[Summary of the Findings 15](#_Toc90664373)

[References 15](#_Toc90664374)

# Introduction

The problems in the current work include:

### Clustering and Classification

For the clustering problem, the objective is to assign clusters to a selected dataset using the optimal method. After implementing and selecting the best model as well as assigning clusters to the original data, a classification model will be defined and trained to predict the resulting clusters.

### Time Series Analysis

Time series analysis involves examining the characteristics of a given time series and implementing at least two time-series models to predict 10 future steps for the selected time series problem.

### Association Analysis

The objective of association analysis is to develop association rules between item consumption and period of consumption. This involves generating antecedents (*the item that precedes an outcome*) and consequents (*the result that follows the antecedent*).

### Portfolio Optimization

Portfolio optimization involves optimizing for the Return on Investment for a selected dataset using a proposed algorithm i.e., efficient frontier.

# Argument

## Clustering and Classification

As we had established earlier on, the objective of the clustering analysis is to implement the most optimal model with which to assign groups to the data based on the underlying characteristics of the data. In this work, we chose to implement two clustering models including k-means and k-means minibatch algorithms. The motivation for implementing the two models was to enable us to evaluate the performance of the models after which we would select the optimally performing model.

#### Data

Data used in this study is the *Online News Popularity* dataset which summarizes a heterogeneous set of features about Mashable articles in two years.

### Model Evaluation

Evaluation of the clustering algorithms was conducted using the silhouette score which measures the goodness of a clustering technique using a score ranging from -1 to +1 (Bhardwaj, 2020). Before comparing the performance of the two models, we implemented several k-means and k means minibatch models to determine the optimal value of k for each of the models.

Figures 1 and 2 below show the optimal values for k for each model.

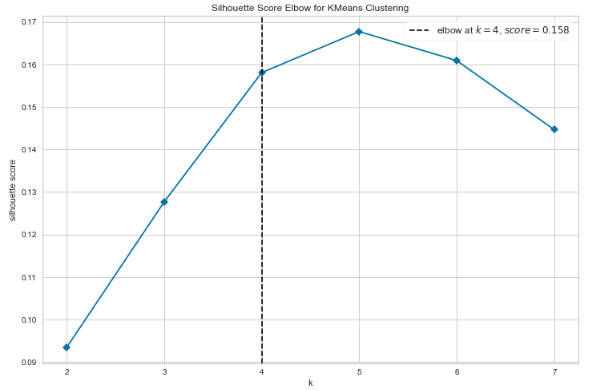


Figure : K-means optimal value of k

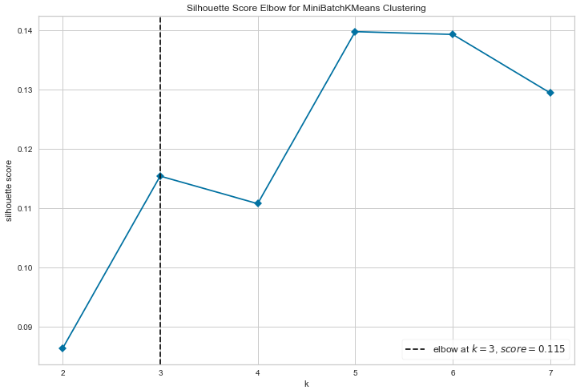
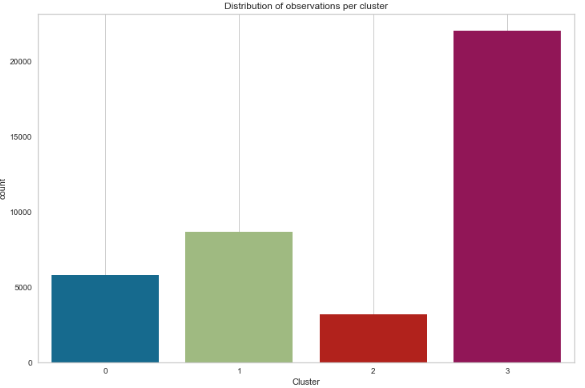


Figure : the optimal value of k for the: Minibatch k-means model

For the k-means model, the optimal value of k was 4 while for the minibatch the optimal value was 3. We fitted each model with the underlying optimal value of k and evaluated how each model performed. The k-means model attained the best silhouette score of 0.54336 while the minibatch k-means model attained a score of 0.53985. Figure 3 shows the distribution of the assigned clusters.

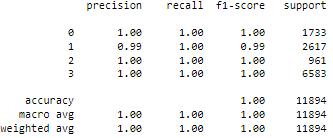


Figure

### Classification

We fitted a random forest model to attempt a prediction of the clusters to which a given observation belongs. the model as noted below had a classification accuracy of 100% (*see table 1*) indicating that the model would assign groups to new observations with a 100% accuracy.

Table : Classification performance

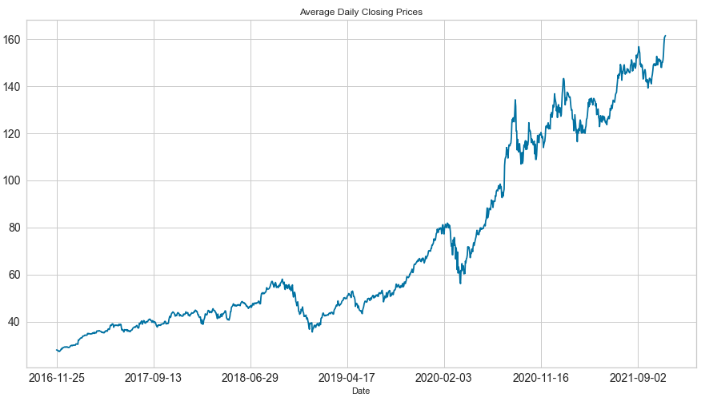


## Time Series Analysis

The closing price for Apple stocks’ data was selected for this problem which includes observations from the daily stock trading of Apple (APPL).

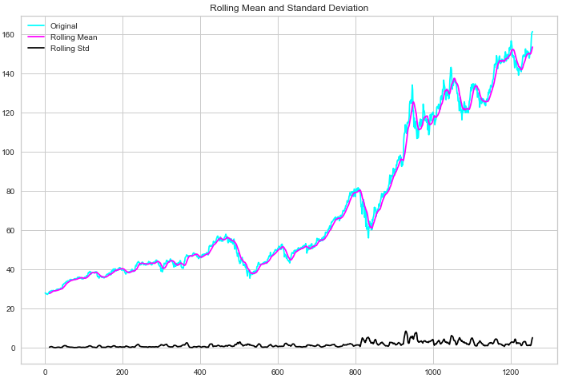
#### Time series characteristics

Figure 4 below shows the distribution of the average daily closing price of the stock.



Figure

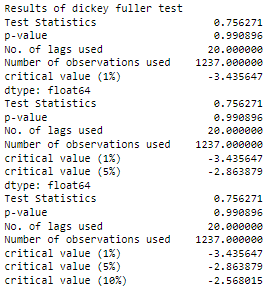
From figure 5 we realize that while the rolling mean is increasing, the standard deviation is relatively constant.



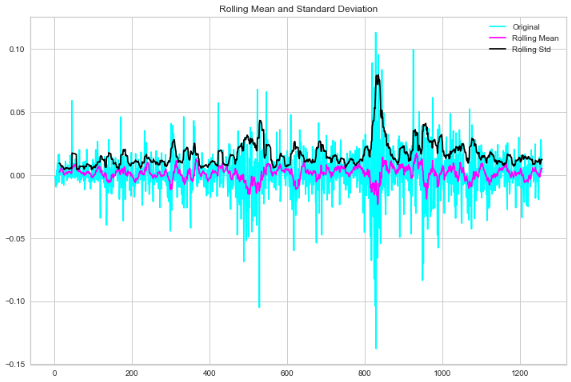
Figure

Using a Dickey-Fuller Test (*see table 2*) we noted that the time series was non-stationary hence differenced and detrended the data to make it suitable for modeling.

Table

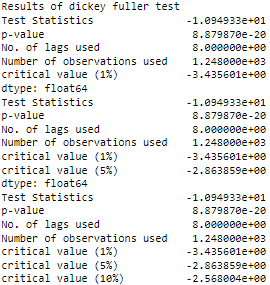


The plot for the final time series as well as the results from the Dickey-Fuller are given in figure 6 and table 3 below.



Figure

Table



Test Statistic as noted above is less than the Critical Value and the p-value is less than 5%. Therefore, we can be confident that the trend is almost removed hence the data is suitable for analysis.

#### Modeling

Both Auto Regressive Moving Average (ARIMA) and ARIMA and exponential smoothing models were proposed. The models were evaluated using the Akaike Information Criterion (AIC).

### Model performance

Tables 4 and 5 below provide an overview of the performance of both the ARIMA and Exponential Smoothing models.

#### ARIMA

An auto Arima approach was adopted to help determine the best ARIMA model whuch nin this case resulted to a SARIMAX model.

Table : ARIMA model diagnostics

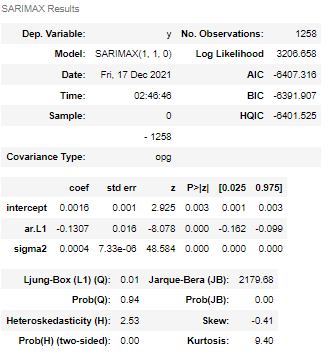
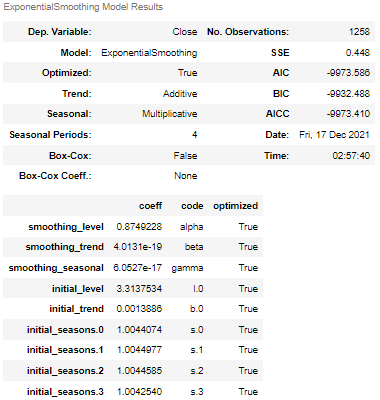
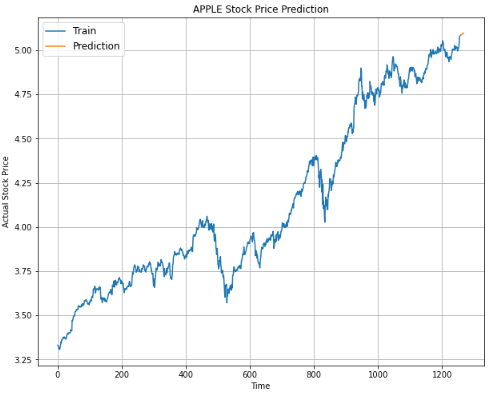


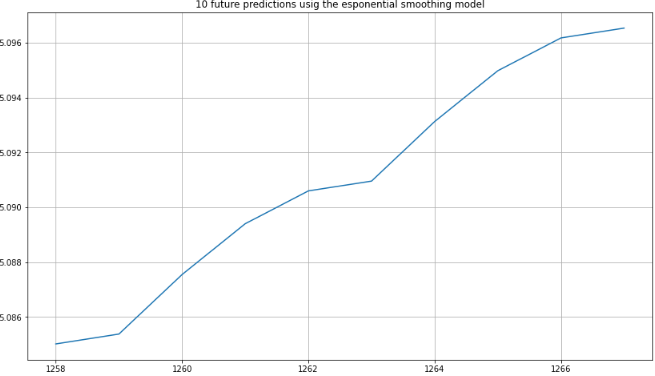
Table : Exponential Smoothing



We note that the Exponential Smoothing model has a lower AIC hence the best prediction model. Figures 7 and 8 show the distribution of the stock over time and the forecasts of the future 10 periods respectively.



Figure



Figure

In general, from figure 8 we note that the closing price of APPL’s stock is expected to increase over the subsequent 10 periods.

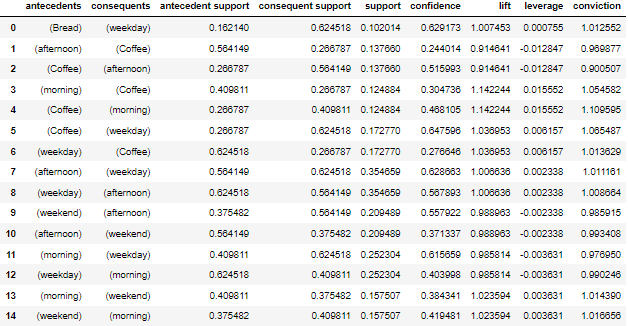
### Challenges of Stock Forecasting Since the outbreak of COVID-19

According to (Hong, et al., 2021), COVID-19 has had significant implications on the performance of the stock market. To a large extent, since the outbreak of COVID-19, the market has been affected by the instability of both stock return predictability and price volatility during the early days following the outbreak of COVID-19. Ideally, instability on stock return predictability affects the dependence of forecasting methods due to a decline in the available information when making decisions (Buldyrev, et al., 2021).

## Questtion 3: Association Analysis

An apriori algorithm was adopted for the association analysis which as we had noted earlier was designed to generate association rules. We generated 15 rules from the data as shown in table 1. From table 1 below, the antecedent is the preceding cause for the consequent effect. For instance, in rule 1, if an individual purchases bread, then it is most likely a weekend and for rule 2, if it is afternoon, then an individual is most likely going to order coffee.

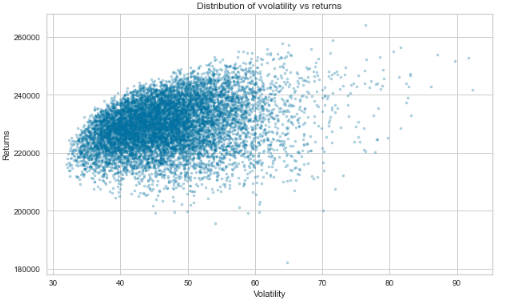
Table : Generated association rules



From table 6 above we note that morning and coffee have the highest life i.e., 1.1422 meaning that coffee is likely to be taken in the morning.

## Question 4: Portfolio Optimization

An efficient frontier algorithm in modern portfolio theory which is the set of optimal portfolios that offer the highest expected return for the defined level of risk was adopted for the current problem (Schulmerich, 2013; Malik, 2019). The maximization in the efficient frontier algorithm is conducted using Sharpe Ratio. We replaced observations that have 0 entries with a very small number to avoid division by zero during experimentation. The following figure shows the relationship between returns and volatility.

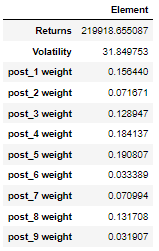


Figure

From figure 9, an increase in volatility corresponds to an increase in returns since a risk increase corresponds with an increase in the reward.

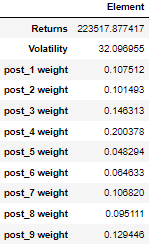
Table 7 below shows the initially expected return of the portfolio.

Table



From table 1 we note that *post\_1, post\_3, post\_4, post\_5, and post\_8* have the highest weights. To optimize the portfolio, the Sharpe ratio. Table 8 shows the optimized portfolio.

Table



From table 8, one would get the maximum returns if they “invested” based on the distribution of the weights given in table 8. As noted in table 8, investing 10.75% in post\_1, 10.1493% on post\_2, 14.6313% on post\_3, 20.0378% on post\_4, 4.8294% on post\_5, 6.4633% on post\_6, 10.6820% on post 7, 9.5111% on post\_8 as well as 12.9446% on post\_9 would lead to better returns.

## Question 5: Methods For Detection of Sensitive Information Using Machine Learning

Given the growth, the widespread adoption of digital data, the process of protection as well as detection of sensitive information is a paramount issue for both businesses and individuals (Huang, 2021) since the cost of disclosure of sensitive data is high.

According to (Xu, et al., 2019), methods for detecting sensitive information can be categorized into 2 approaches i.e., sensitive word matching and traditional machine learning methods. Traditional methods include CNN and RNN which (Huang, 2021) observes as insufficient compared to SVM. Through supervised machine learning methods, sensitive information can be predicted with relatively good accuracy (Valera & Shah, 2018). Besides, usage of unsupervised methods such as clustering and association rule mining to generate understandable insights regarding the patterns in the data.

# Conclusion

## Summary of the Findings

Overall, we noted that:

1. Based on the silhouette scores of the k-means and MiniBatch K-means models that the k-means had a better score (0.5433630416718219) compared to that of DBSCAN (0.5398523687294093) indicating that the k-means model performs better hence will be used for the final cluster allocation.
2. The Random Forest model can predict with 100% certainty the group that a given observation belongs to.
3. Overall, the ARIMA model predicts a continuous increase of the APPL stock over the next 10 trading days while the SARIMAX model predicts an increase in the first day which is ideally the peak followed by a drop with a subsequent slight increase after which the closing price of the stock is set to be constant.
4. Up to 15 rules can be generated from the bread basket data
5. The portfolio had a variance of i.e., 5.6662 with a minimum return of 224441.276243.
6. After optimizing, the portfolio had a return of 224441.276243.

# References

Bhardwaj, A., 2020. *Silhouette Coefficient.* [Online]   
Available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c#:~:text=Silhouette%20Coefficient%20or%20silhouette%20score%20is%20a%20metric%20used%20to,each%20other%20and%20clearly%20distinguished.&text=a%3D%20average% [Accessed 17 December 2021].

Buldyrev, S., Flori, A. & Pammolli, F., 2021. Market instability and the size-variance relationship. *Sci Rep,* 11(2021), p. 5737.

Hong, H., Bian, Z. & Lee, C., 2021. COVID-19 and instability of stock market performance: evidence from the U.S. *Financ Innov,* 12(2021), p. 7.

Huang, Z., 2021. *Sensitive Information Detection Using HMM&SVM.* New York, NY, USA, Association for Computing Machinery.

Malik, F., 2019. *Understanding Efficient Frontier.* [Online]   
Available at: https://towardsdatascience.com/understanding-efficient-frontier-46f1a429d526  
[Accessed 17 December 2021].

Schulmerich, M., 2013. The Efficient Frontier in Modern Portfolio theory. *Investment Management,* 1(6), pp. 27-31.

Valera, M. & Shah, R., 2018. *Survey of Sensitive Information Detection Techniques: The need and usefulness of Machine Learning Techniques,* s.l.: InfoSec.

Xu, G., C. Qi, H., Yu, S. & Xu, C. Z. Y., 2019. *Detecting Sensitive Information of Unstructured Text Using Convolutional Neural Network.* s.l., international Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC).