**An Analytical Way To Get More Customers After COVID-19**

Tan Li Tung

Electrical and Electronics Department

Universiti Teknologi PETRONAS

[li\_17002803@utp.edu.my](mailto:li_17002803@utp.edu.my)

+6011-10868940

**n Analytical Way To Get More Customers After COVID-19** is an exploratory analysis to find insights in the current customer database. This document will visualise the data using plots and graphs. Techniques like Simple Linear Regression, Multiple Linear Regression, Logistic Regression, Navie Bayes, Linear Discriminant Analysis and Quadratic Discriminant Analysis will be carried out for predictive analysis. This report would include the use of Principal Component Analysis and Singular Vector Decomposition as the unsupervised learning on the country dataset.

**1 Introduction to The Study**

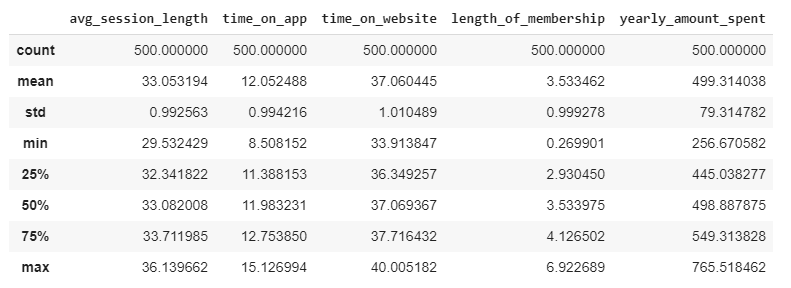
Due to the recent COVID-19 outbreak, lots of businesses lost their customers. The businesses needs to prepare for the opening of their business so that they can serve their customers better. They also needs to acquire new customers and retain the existing customers so that the business can survive the pandemic. In this study, a dataset contains 500 records of customers dataset is used for analysis. The dataset includes the average session length, time on app, time on website, length of membership and also yearly amount spent by the customers. The goal of the study, the yearly amount spent by the customers is set to be the target variable for prediction. The **relationship of the other variables and the target variable should be examine** and **policies to keep the customer are to be proposed.** All the codes are run in Google Colab which can be found at https://colab.research.google.com/drive/1Bj4sz-yz61W897xDiZyAqkVPlTCtKkCz?usp=sharing

**2 Method and Algorithm Used**

**2.1 Exploratory Data Analysis (EDA)**

**Feature Selection.** Before the data is used further, the data needs to be pre-processed. The Email, Address and Avatar attributes are dropped as it does not affect the yearly amount spent of a customers.

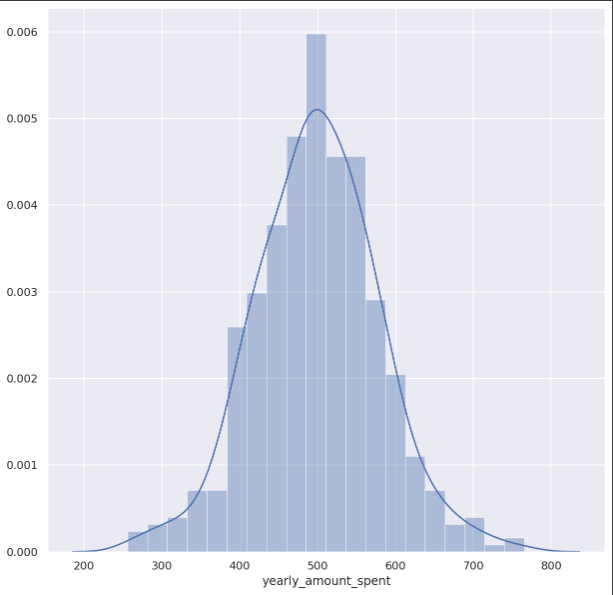
**Description of Data.** Some of the information can be gotten from the description of data and it is shown as below in Figure 1.



**Figure 1: Description of Data**

In Figure 1, it is shown than this dataset has no null values and it can be considered clean. Therefore the pro-processing of data becomes more easy. The minumum and maximum values in the dataset is also reasonable which means that there are no outliers.

**Distribution of Attributes and Target Variable.**

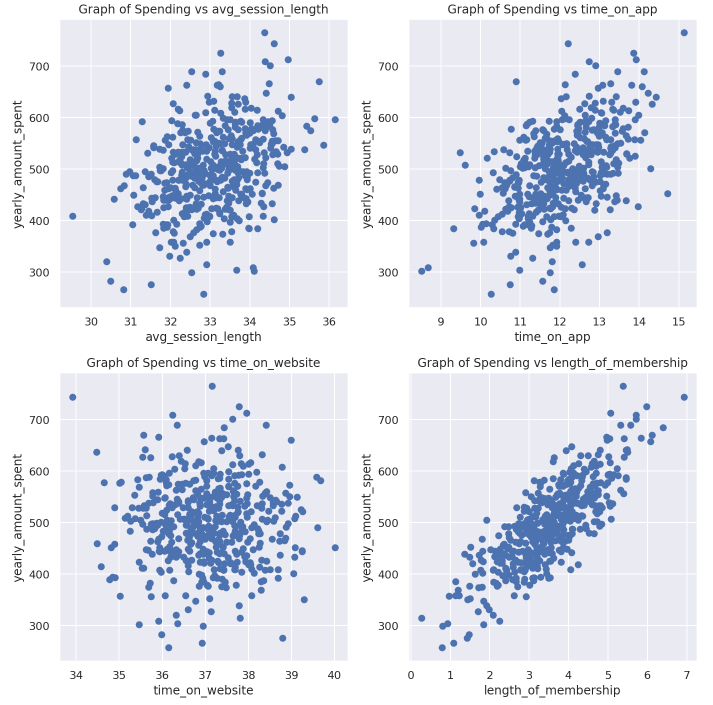


**Figure 2: Distribution of Attributes and Target VariableNumbering and Running Heads**

As shown in Figure 1, all the attributes and target variable is almost normally distributed. This makes the data more explainable.

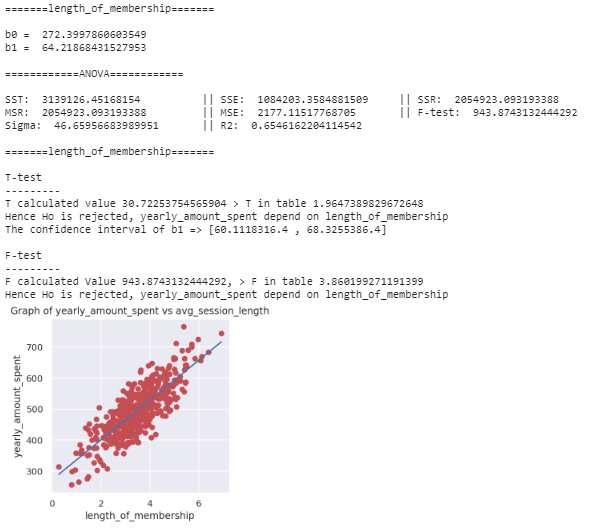
**2.2 Simple Linear Regression (SLR)**

The business are interested to determine whether each of the attributes are highly related to the yearly amount spent. This can be achieved by using SLR. The realtionship between each attribute and the target variable are plotted in Figure 3 below.

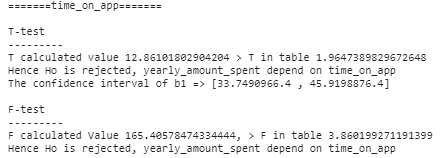
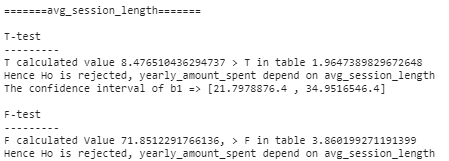


**Figure 3: Relationship Between Attributes and Target Variable**

In Figure 3, it is clearly shown that the length of membership has a linear relationship with the yearly amount spent. This means that the longer the membership, the more the member spent. However, we still need to test the hypothesis to know that whether the the target variable depends on the attributes. The result of SLR on each attribute are shown below.



**Figure 4(a)**



**Figure 4(b) Figure 4(c)**



**Figure 4(d)**

**Figure 4: SLR Results on Each Attribute**

As the result of SLR suggest, the length of membership is 65.46% linearly related to the yearly amount spent. The aerage session length and time on app is not highly related to the yearly amount spent but both of these attributes are still satistically significant as shown in Figure 4(b) and 4(c). The yearly amount spent has no dependencies on the time on website as shown in Figure 4(d). Through SLR, we found out that the customers needs to be retained so that they spent more. Since the yearly amount spent does not depends on the time on website, the business can consider the switch the website service to the app as most customers spent their time on app.

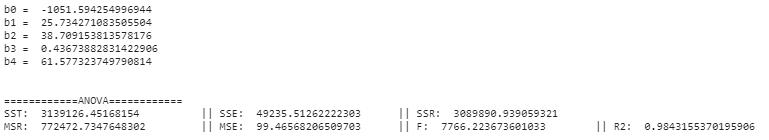
Although the SLR can give a prediction accuracy up to 65.46% throug the equation:

*(yearly amount spent) = 272.400 + 64.219 \* (length of membership)*

However the accuracy is not that high to make accurate predictions. Therefore, ther is a need to perform other techniques to improve the accuracy of the predictions.

**2.3 Multiple Linear Regression (MLR)**

MLR is similar to SLR but it consideres all of the atrributes. Through MLR, the results are as below.



**Figure 5: Result of MLR**

In the result above, we can see that the R2 score is 98.43%. This indicates that the MLR model can perform predictions better. The equation can be written as:

*(yearly amount spent) = -1051.59 +25.73 \* (average session length) + 38.71 \* (time on app) +*

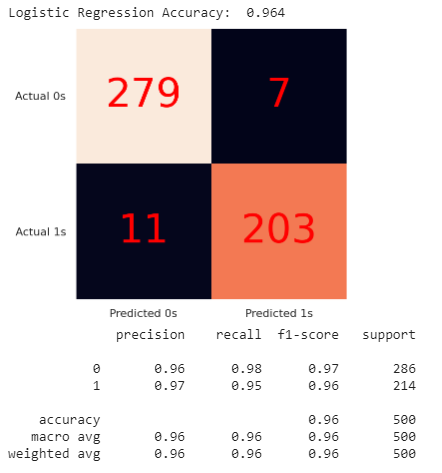
*0.44 \* (time on website) + 61.58 \* (length of membership)*

The MLR model also have a lower mean sqaured error (MSE) and mean squared regression (MSR) as compared to the SLR model using only the length of membership. Both SLR and MLR models can predict the yearly amount spent of the customers. The prediction can be generalized into categories to make it more easy to predict.

**2.4 Logistic Regression**

Logistic Regression is used to solve a classification problem. Therefore, the target variable needs to be changed to classes before logistic regression can be used. In the study, the target variable is splitted into 2 categories which are 0 (low) and 1 (high) where 0 are the yearly amount spent less than the midpoint of yearly amount spent while 1 represent yearly amount spent that are at least the midpoint of the yearly amount spent.

The logistic regression predicts the class of a new data based on the probabilities of it belongs to a class. The results are shown as below. The function used to determine the probability of a data belonging to a particular class is called a logistic function.



As seen in Figure 6, the model correctly predict 482 data out of 500 data. The model provides an accuracy of 96.4% which is high enough to be used for predictions. However, more models still can be used to predict the class of yearly amount spent from the attributes.

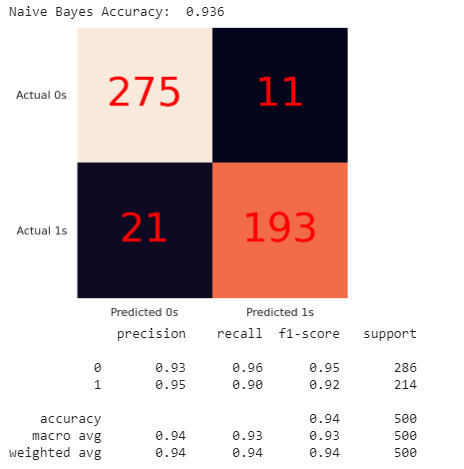
The Logistic Regression can be performed easily using the ScikitLearn library. By default, the Logistic Regression model in ScikitLearn used the logistic function (solver algorithm) of Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm (L-BFGS).

The L-BFGS algorithm approximate the Hessian matrix using updates specified by gradient evaluations. This algorithm is chosen as compared to others as it is efficient and it consume lesser amount of memory.

**Figure 6: Results of Logistic Regression**

**2.5 Naive Bayes Classfier**

Naive Bayes is a classifier working based on the Bayes’ Thoerem. The algorithm assumes that the attributes are independent of each other. In real life, this is nearly impossible but the Naive Bayes classifier works suprisingly well. The model are able to predict the class with a high accuracy.

As shown in Figure 7, the Naive Bayes model correctly predict 486 data out of 500 data. In other words, the Naive Bayes model predicts the output with an accuracy of 93.6%. This model have a high accuracy but it is not as good as the Logistic Regression model.

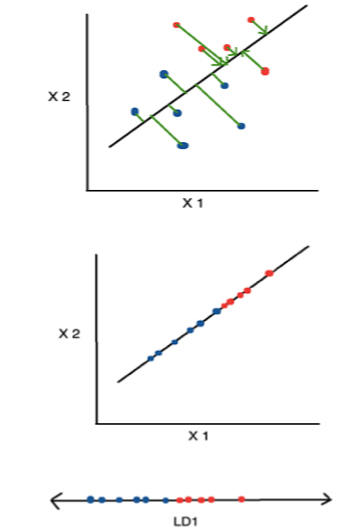
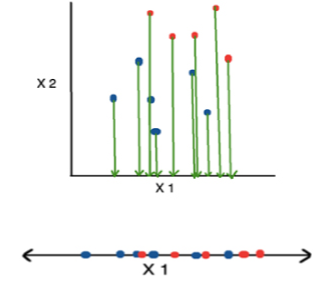
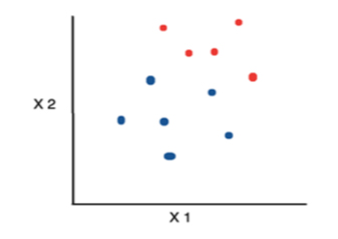
This is because the model assumes that all the attributes are normally (Gaussian) distributed. However, as shown in Section 2.1, Fıgure 2, the distribution is not exact normal. This is the reason that the Navie Bayes Classifier has a lower accuracy as compared to the Logistic Regression that uses a logistic function to calculate probabilities.

Naive Bayes classifier can be constructed easily as it does not require a lot of training. Due to the assumption made earlier, the model is insensitive towards the irrelevant attributes, which makes the model accurate.

**Figure 7: Results of Naive Bayes Classifier**

**2.6 Linear Discriminant Analysis (LDA)**

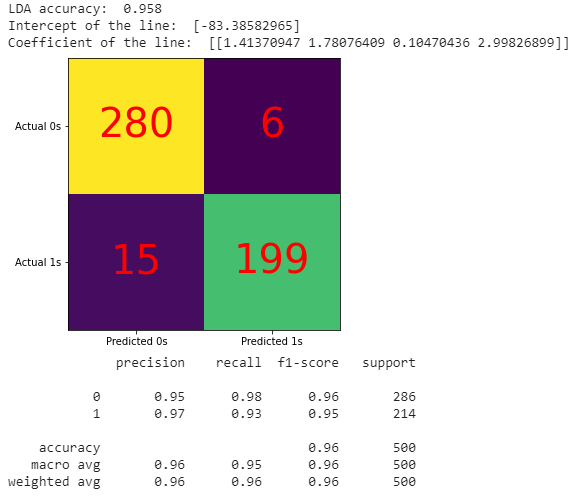
LDA is an method used to reduce the dimensiona of the dataset. At the same time, it also have the ability to perform classification. The general idea of LDA is to find a line that can best separate the classes.



**Figure 8(a) Figure 8(b) Figure 8(c)**

**Figure 8(a): Sample Dataset, 8(b): Data Projected to x-axis, 8(c): Data Projected to LD1**

Consider a dataset as shown in Figure 8(a). To differentiate the classes, the data can be projected to the a-axis as shown in Figure 8(b). However, there are lots of overlapping in the output, therefore a linear line with the best saparability needs to be found. The line in Figure 8(c) has the best separability in this case. The data are projected to the linear line LD1 which the data points are found to be easily separable.

Base on the confusion matrix shown in Figure 9, the LDA model yields an accuracy of 95.8% and the linear line can be observed through the intercept and also coefficient which can be written as:

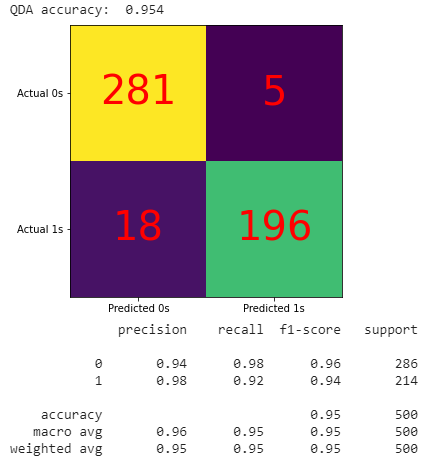
*LD1 = -83.39 + 1.41 \* (average session length) + 1.78 \* (time on app) + 0.1 \* (time on website) + 2.98 \* (length of membership)*

The line is found to have a greater lean towards the length of membership, which shows that length of membership containes most information that can be the most significant attribute for the classification. However, The line that separate the classes may not always be a linear line. Therefore, QDA can be carried out to examine different relationship between the attributes and target variable.

**Fıgure 9: Results of LDA**

**2.7 Quadratic Discriminant Analysis (QDA)**

QDA shares the same concept as LDA. However, the line that can fit make the best separability might not always be a linear line. Therefore, QDA finds a line in quadratic that have the best separability of data.



**Figure 10: Results of QDA**

In Figure 10, it is observed that the QDA model provides an accuracy of 95.4%. The accuracy is lower than the LDA model. This happens due to the data is related to the target variable linearly instead of quadractically.

**2.8 Conclusion of Supervised Learning**

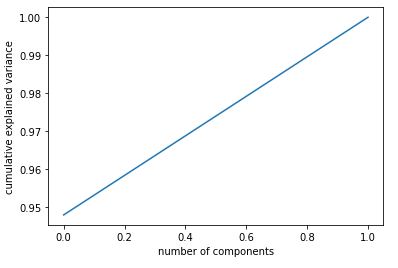
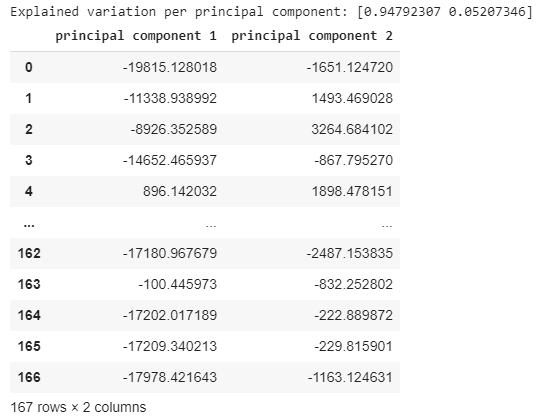
As 6 techniques are performed with 2 regression and 4 classification, the business can now get some insights from the analysis. The customers generally spent more when they have a longer length of membership. Therefore, the business can send out gifts to their existing customer to retain them longer. The analysis also suggest that the time on website does not affect the yearly amount spent. Therefore, the business can considering move the website fully to the app. The MLR model can be used to predict the yearly amount spent of the customers and the Logistic Regression model can be used to predict classes of the customers. If a customer is predicted to be in the high yearly amount spent class, the business can consider to upgrade their membership so that they can retain the customers by rewarding them.

**3 Unsupervised Learning on Country Dataset**

**3.1 Principal Component Analysis (PCA)**

PCA is a dimensionality reduction technique that preserve most of the information. The principal components (PC) represents the direction of the data that explains the maximum amount of variance. PCn always have a greater variance than the PC(n+1), for instance, PC1 has a greater variance than PC2.

In the country dataset, there are 167 records from 167 different countries. Therefore, the country column is dropped as it does not play a significant role in the clustering in PCA.

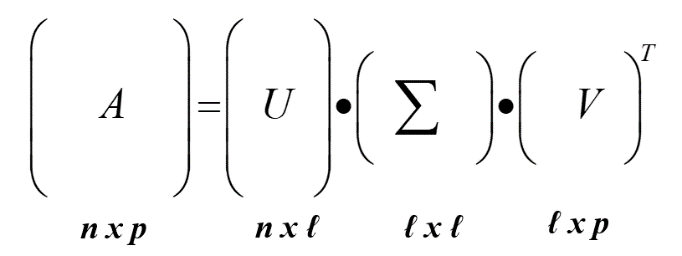


**Figure 11: Output of PCA**

As shown in Figure 11, when the number of components reaches 1, the cummulative explained variance also reahces 1. This means that the PCA has a high performance. The result also shows that 94.79% of information is hold by PC1 and 5.20% information is hold by PC2. This means that if a 4-dimension data is projected into 2-dimension data, there are only 0.01% of information loss. This means that the PCA wourks perfectly fine on the country dataset.

**3.2 Singular Vector Decomposition (SVD)**

Every table data can be represented by a matrix. With the help of SVD, any matrix can be decomposed into 3 matrices.



**Figure 12: SVD Concept**

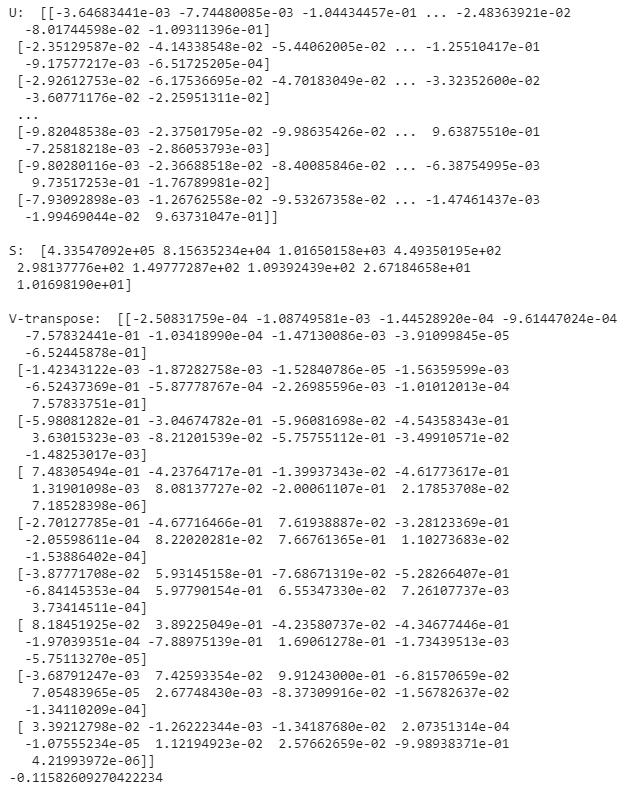
A is the matrix that needs to be decomposed

U is the left singular vector matrix

∑ is the singular values

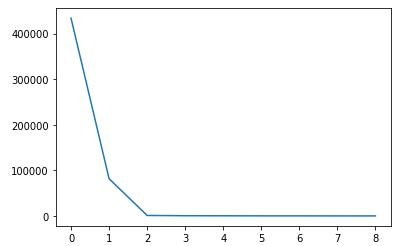
VT (Transpose of V) is the right singular vector matrix.

SVD can be used to decompose the country data. The result are shown as below.



**Figure 13: The U, S, V Matrix After Decomposition**

As shown in Figure 13, the country dataset is decomposed into 3 components. In the S matrix, we can see that there are sigma points. These point shows that how close all the points are close to that axis. In our case, the sigma points are as below.



**Figure 12: Sigma Points**

**3.3 Conclusion of Unsupervised Learning**

There are 2 most popular algorithm used for unsupervised learning, which are PCA and SVD. Both of the algorithms are performed on the country dataset and the clusters can be shown after performing them. PCA and SVD are generally helpful in performing clustering that can find the common pattern in the dataset. Although the patterns can be deduced, but it still requires experts to decide whether the pattern are useful or not.