MSC IN BUSINESS ANALYTICS

Module:

Data Mining- B9BA103

Assignment Title:

CA 1 – Supervised(Marketing) & Unsupervised Learning(Drink)

Submitted to:

Mr. Kunwar Madan

Submitted By:

Lukose Roy Pannapara (10529675)

Contents

PART 1

1. Introduction (Bank Marketing)

The dataset contains data pertaining to a marketing campaign carried out by a bank with its existing clients in order to identify and convince potential customers to subscribe to their term deposit scheme. Each row in the data represents information pertaining to a client and there are a total of 45211 client information rows distributed among 17 columns respectively. This report presents a detailed analysis and predictive model by identifying the various factors that could result in clients to subscribing to term deposit of the bank.

2. Data Preparation & Analysis

- **Importing the dataset**

  The pandas library is used to import the marketing.csv dataset.

- **Descriptive analysis**

  The descriptive analytics on the dataset states that there are 45211 rows and 17 columns in the dataset of which 10 columns contain categorical values and 7 columns contain continuous data.

- **Converting categorical data into numerical values**

  The dataset contains two types of data viz. integer and object.

  The 10 object or categorical data were converted to numerical values using label encoding method for further analysis and processing.

- **Dependent variable data distribution**

  'Subscribed' column in the dataset gives us an overview of number of clients who have subscribed and rejected the term deposit scheme. The seaborn library bar plot is used to visualize the distribution of 'yes' and 'no'. The bar plot suggests that somewhere around 5000 customers subscribed out of 45211 contacted in the whole campaign which is only **11-12% success rate**.

<matplotlib.axes._subplots.AxesSubplot at 0x211c8ab4198>

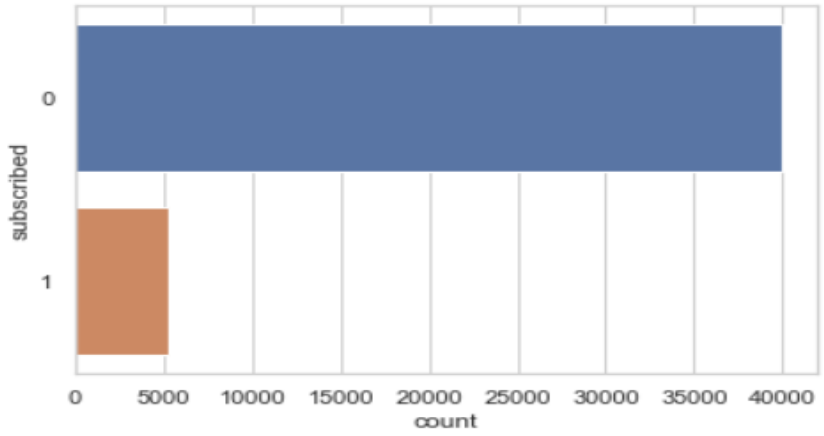


Figure 1

- **Independent variable analysis (IV vs DV)**

  Box & Bar plot analysis was carried out to visualize the distribution of each independent variable with respect to the target variable.

  The insights obtained from these analysis are as follows:

  (i) The below box plot illustrates that more number of rejections were from customers contacted via ‘unknown’ and ‘cellular’ medium.

<matplotlib.axes._subplots.AxesSubplot at 0x211c8e841d0>

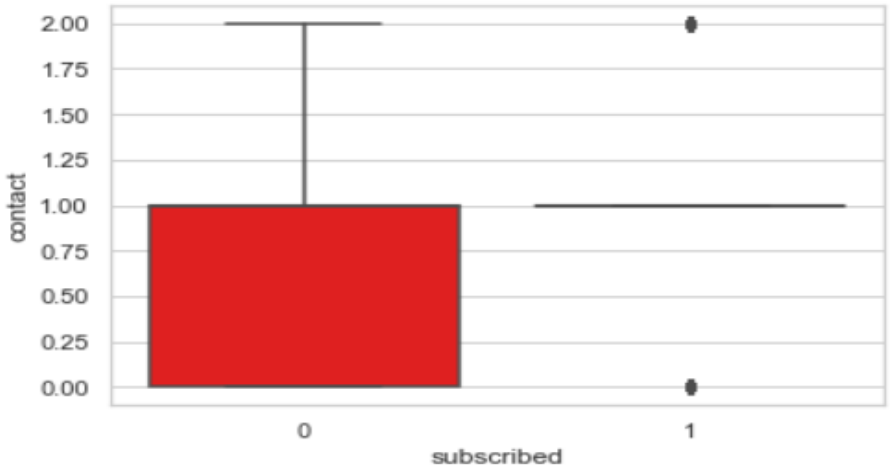


Figure 2

  (ii) The comparative study for month and subscription suggests that higher number of people subscribed when contacted between April & August.

```
<matplotlib.axes._subplots.AxesSubplot at 0x211ca2302b0>
```

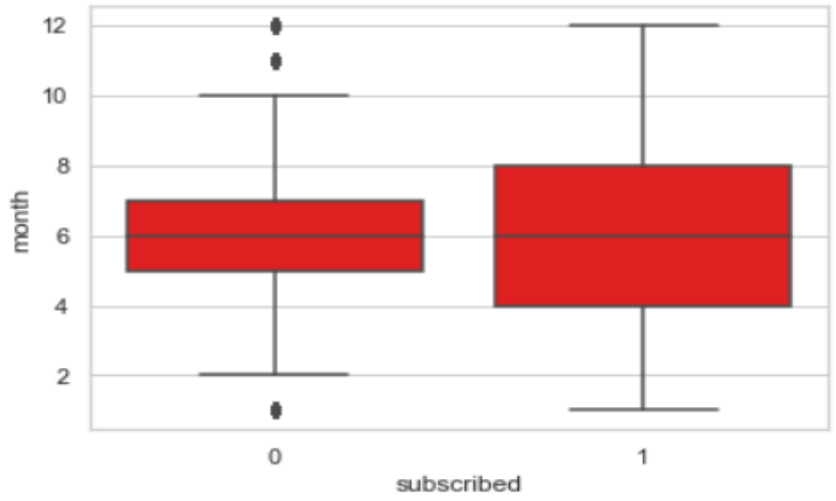


Figure 3

(iii) The duration with subscription bar plot suggests that majority of who subscribed have conversed for a duration of more than **200+ secs**.

```
<matplotlib.axes._subplots.AxesSubplot at 0x2295227acc8>
```

Figure 4

- **Correlation and Causation**

  A high **negative correlation** was observed on the heat map between '**poutcome'** and '**pdays'** of **-0.86**. Number of days passed after the client was last contacted from previous campaign could influence the outcome of campaign in the form of success or failure. Hence this suggests that a high causation is also possible. Also, 'poutcome' has a negative correlation around -0.49 with 'previous'.

  Due to high correlation + causation, it was decided that 'poutcome' must be dropped.

- **Dividing dataset to train and test sets**

  The data was first divided into label (target variable) and feature set as first step to predictive analytics. The feature set contains 15 columns and label set contains the 'subscribed' column.

  Normalization was performed in order to rescale the numeric values in the range 0 (mean) and 1 (variance).

  The data was divided in the ratio **75:25** with training set containing around **33908** rows and test set with **11303** values.

- **SMOTE**

  To avoid the overfitting problem, synthetic data was developed based on feature similarities of existing data.

```
Number of observations in each class before oversampling (training data):
 0    29928
1     3980
Name: subscribed, dtype: int64
Number of observations in each class after oversampling (training data):
 1    29928
0    29928
Name: subscribed, dtype: int64
```

3. Model Building / Testing & Evaluation

- In the random forest classifier, grid search cross validation method was carried out to evaluate combinations [50, 100, 150, 200, 250, 300] to identify optimal values in order to be used for hyperparameters. The scoring was set to '**recall'** with the idea to minimize false negatives and cv to 5.

- **It is important to minimize the error rate in the prediction model where the client has subscribed to the bank deposit scheme, but the model predicts it the other way round. This would not only impact the customer base but also cause major financial loss to the bank.**

- The n-estimators obtained from mean cross-validated score of best estimator is **200**. The n-estimators indicates the number of trees in the forest.

- Upon building the model with tuned parameter of 200, it is observed that the most **significant variables are of the sequence duration, month, campaign, balance, day, age etc.**

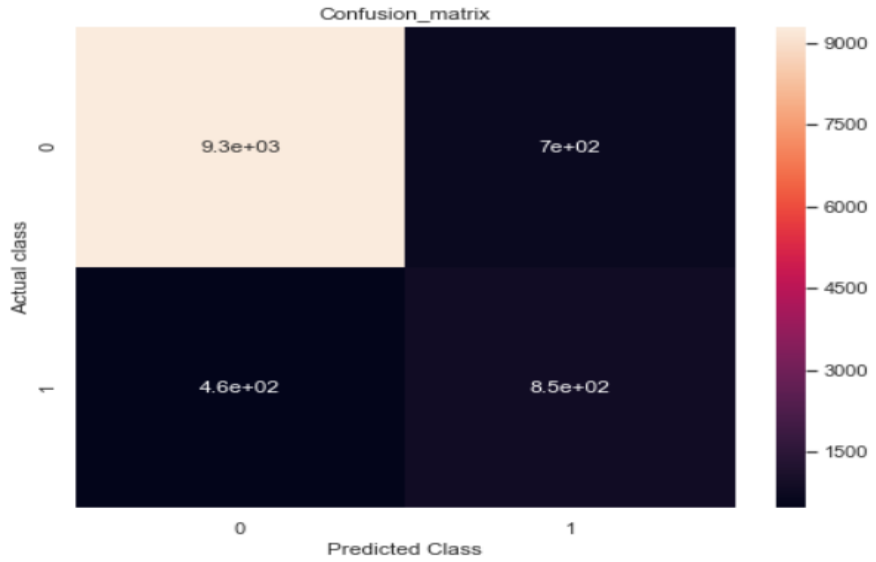
4. Final Classification Model

- The confusion matrix returned a **FN of 499** and a **TP of 810 for the entire dataset**. In order to minimize the false negatives, various set of feature selections were done and the top 4 models are illustrated as shown below:

Labels	True Positive	True Negative	False Positive	False Negative		
all	810	9315	679	499	11303	
7	810	9254	740	499	11303	
9	826	9297	697	483	11303	2nd Best Model
11	847	9291	703	462	11303	**Best Model**

Figure 5

- However, the **final model** with a feature set of **11 out of 17** variables produced the best results with a **FN of 462 and TP of 847**. Each of this variables are inter-related to each other and could help in setting up a good marketing roadmap. **The final set of significant factors are as follows: duration, month, campaign, balance, day, age, contact, job, pdays, previous, housing.**




```
Confusion matrix:
 [[9291  703]
 [ 462  847]]
TP:  847
TN:  9291
FP:  703
FN:  462
```

Figure 6

5. Recommendation

The initial analysis and model prediction suggests that the Bank Marketing team must take into account the following measures:

- The **duration** of communication must be more than 200 seconds as observed from the previous historical data analysis (Figure 4). It is also worth noticing that a customer contacted more than 3 times is less likely to subscribe during a **campaign**. Hence, it is important to make a maximum of 2-3 client calls with more than 200 secs of conversation. It is also imperative to note that the customers must be **contacted** via telephone (Figure3).

- A high success rate of subscription was observed during the **months** of April - August. The marketing team must also focus on contacting clients between October – January (Figure3). Similarly, any contact made in between 8th & 22nd **day** of the months produced a higher success rate.

- People with an average annual salary(**balance**) of more than 1250 euros or more and an **age** group of 30 – 50 are more likely to subscribe to this term deposit scheme. Similarly, people with existing housing **loan** are less likely to subscribe. The box plot also suggests that people of all **job** profiles have subscribed to this scheme, but the bar plot gives a clear overview that, majority of subscriptions were obtained from people belonging to entrepreneur, housemaid, management, retired, self-employed, services category.

*****The underlined bold characters represents the significant variables of the best model.**

PART 2

6. Introduction (Drink) – TSNE & K-Means

The dataset contains data pertaining to an alcoholic drink along with its physiochemical properties carried out by a company in order to identify natural clusters with this dataset. Each row in the data represents information pertaining to a batch of produced drink and there are a total of 4899 rows distributed among 11 columns respectively. This report presents a detailed analysis of the natural clusters with unsupervised learning algorithms and a possible target variable in order to perform predictive analytics.

7. Data Preparation & Analysis

- **Importing the dataset**

  The pandas library is used to import the Drink.csv dataset.

- **Descriptive analysis**

  The descriptive analytics on the dataset states that there are 4898 rows and 11 columns with continuous values. The dataset contains only float type of data.

- **Correlation and Causation**

  A high **negative correlation** was observed on the heat map between '**density'** and '**alcohol'** of **-0.78** and a high **positive correlation** of **0.84** between **density** and **residual sugar**. Similarly, the correlation of **0.53** between **density** and **total sulfur dioxide**. Since alcohol and/or residual sugar could increase the density of a drink, a **causation** is assumed and 'density' was dropped.

  Due to high correlation and a possible causation, it was decided that 'density' must be dropped.

- **Creating Subsets from Dataset**

  An initial TSNE plot on the dataset (Subset0) created natural clusters based on 4 variables viz. Residual Sugar, Free Sulphur Dioxide, Total Sulphur Dioxide and Alcohol.

  The subsets were created based on setting one of the above mentioned column as a categorical variable in all of the subsets.

The first set of 4 subsets were tested by converting residual sugar into a categorical data by dividing with its mean value. **Residual sugar was chosen on the basis of high standard deviation observed. This indicates that the data is well scattered by +/- 5 from mean.**

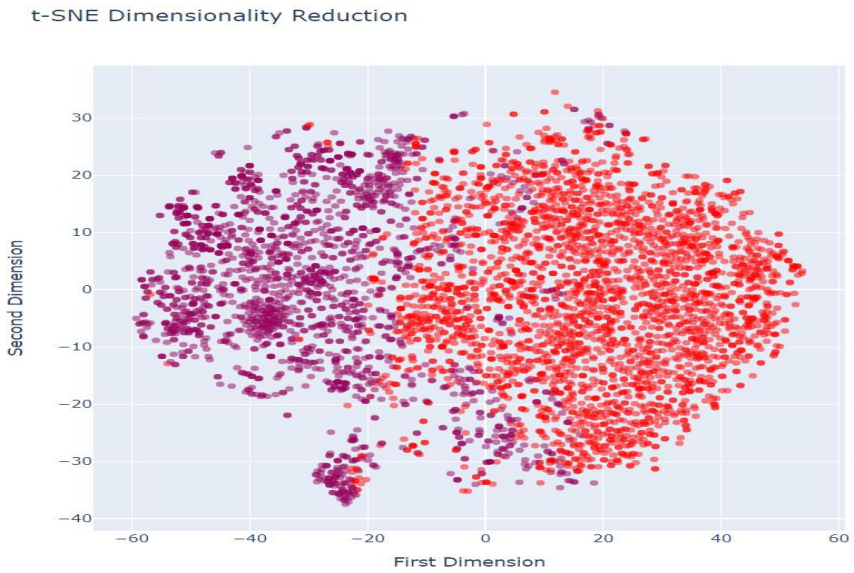


Figure 7: Whole Dataset Visualization - Subset0

8. K – Means Implementation

- K-Means is an unsupervised machine learning algorithm used to **label data** into specifically defined groups.
- The elbow plot method was implemented to find the number of probable clusters within each of these subsets. **The elbow plot tells us the optimal number of clusters in k-means clustering.**
- Subset 1, 2, 4 and 5 produced an elbow plot of 2 , whereas subset 3 produced an elbow plot of 5.
- The K-Means model was fit into each of these subsets with the value of number of clusters and K-means cluster centers are obtained. These K-means labels can be used cluster/plot TSNE visualizations.

9. Why TSNE?

- PCA (dimensionality reduction method) was performed on each of these subsets before implementing TSNE.
- Subset 1,2,3 and 5 produced a **total variance** of **0.63** (poor), **0.87** (good), **0.56** (poor) and **0.81**(good) respectively. But, neither of the PCA subsets could produce a distinguishable plot in 2-dimension.
- TSNE, another dimensionality reduction algorithm was used due to **non-linear data** in the dataset, whereas PCA is believed to work well with linear set of data.

10. TSNE Implementation

- The Subset 1 (**'fixed acidity', 'volatile acidity', 'pH', 'alcohol', 'residual sugar'**) could produce clusters on the basis of **residual sugar**. The **perplexity** is set to **80** and **iterations** to **2000**.

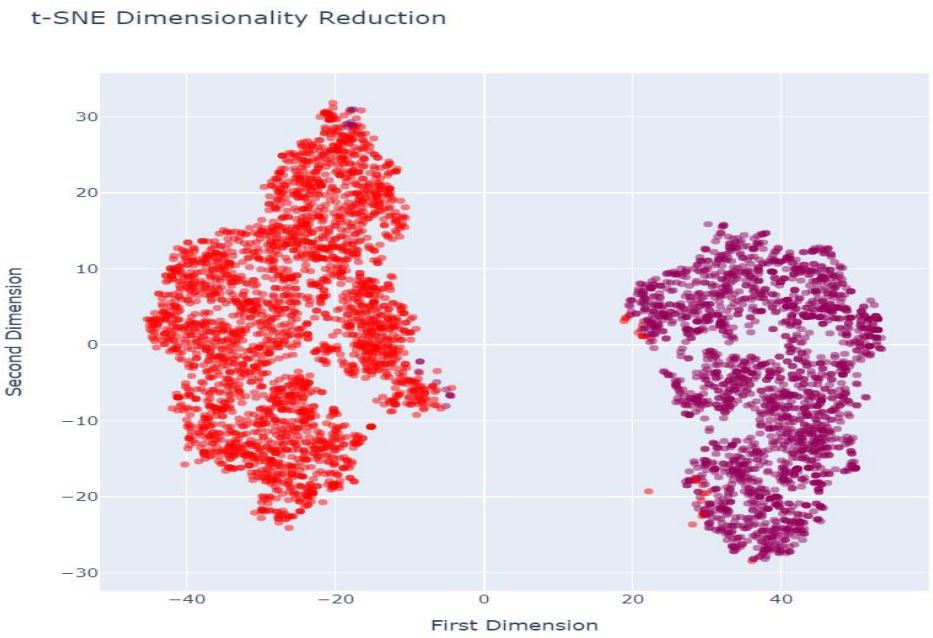


Figure 8: Subset1

- The Subset 2 (**'residual sugar', 'free sulfur dioxide', 'total sulfur dioxide'**) could produce clusters on the basis of **residual sugar**. The **perplexity** was set to **130** and **iterations** to 2000.

Figure 9:Subset2

- The Subset 3 (**'residual sugar', 'sulphates', 'citric acid', 'chlorides'**) produced clusters on the basis of **residual sugar**. The **perplexity** was set to **220** and **iterations** to **2000**. Multiple sub clusters can be seen in the below visualization which are well separated by residual sugar.

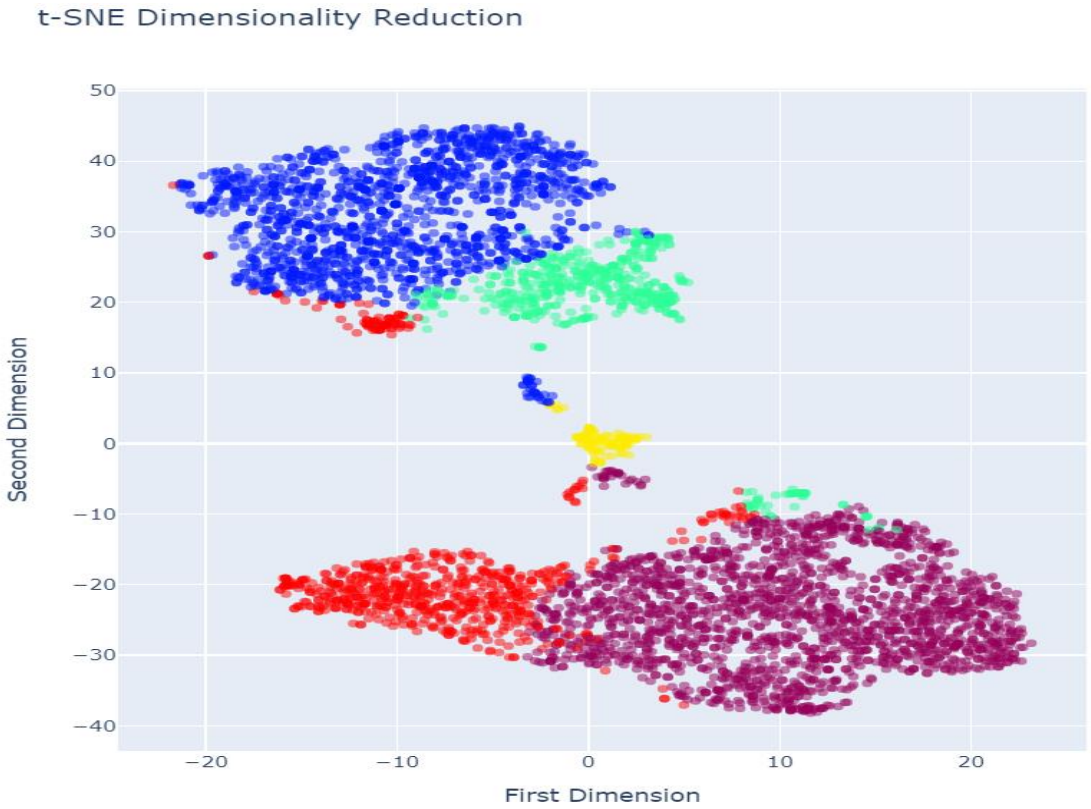


Figure 10: Subset3

- Subset 4 is the **entire dataset** which is separated on the basis of **residual sugar**. The **perplexity** was set to **110** and **iterations** to **2000**. There are multiple sub-clusters within the 2 major clusters.

Figure 11

- The Subset 5 ('fixed acidity', 'alcohol', 'residual sugar') could produce 2 clusters on the basis of residual sugar. The **perplexity** is set to **220** and **iterations** to **2000**.

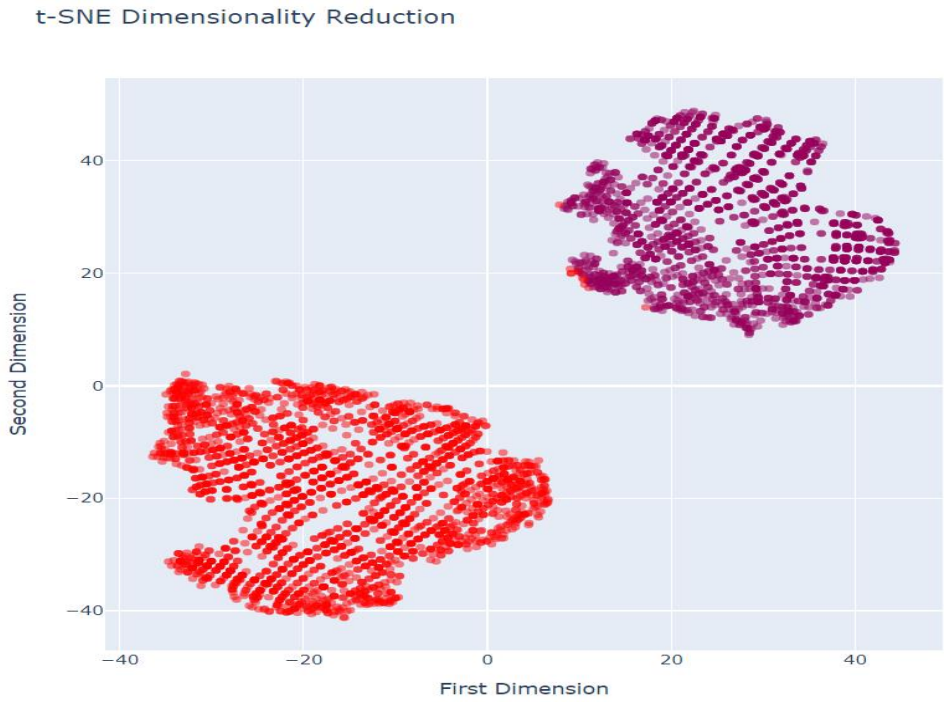


Figure 12: Subset5

11. Conclusion

From the above observation of subsets, it can be concluded that **residual sugar** plays a significant role in creating clusters and hence can be used as a **target variable for predictive analytics** from the drink dataset. **Residual sugar helps us determine if fermentation is over and whether stabilization is needed, or helps to classify the type of alcoholic drink(Sweet or Dry).**