Revanth Janapriyan

**MACHINE LEARNING**

Contents

[Part 1: Machine Learning Models 2](#_Toc125410522)

[**EXECUTIVE SUMMARY** 2](#_Toc125410523)

[**INRODUCTION** 2](#_Toc125410524)

[**Data description** 2](#_Toc125410525)

[1.1 Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset. 3](#_Toc125410526)

[**Data sample** 3](#_Toc125410527)

[**EXPLORATORY DATA ANALYSIS** 3](#_Toc125410528)

[1.2 Split the data into train and test in the ratio 70:30. Is scaling necessary or not? 11](#_Toc125410529)

[1.3 Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance: 11](#_Toc125410530)

[1.4 Which model performs the best? 22](#_Toc125410531)

[1.5 What are your business insights? 22](#_Toc125410532)

[Text Mining 24](#_Toc125410533)

[**INRODUCTION** 24](#_Toc125410534)

[2.1 Pick out the Deal (Dependent Variable) and Description columns into a separate data frame. 24](#_Toc125410535)

[2.2 Create two corpora, one for those who secured a Deal, the other for those who did not secure a deal. 24](#_Toc125410536)

[2.3 The following exercise is to be done for both the corpora: 25](#_Toc125410537)

[a) Find the number of characters for both the corpuses. 25](#_Toc125410538)

[2.4 Refer to both the word clouds. What do you infer? 27](#_Toc125410539)

[2.5 Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis? 27](#_Toc125410540)

# Part 1: Machine Learning Models

## **executive summary**

A transport company is in discussions with ABC Consulting company for providing transport for their employees. For this purpose, tasks are carried out in order to understand how do the employees of ABC Consulting prefer to commute presently (between home and office), based on the parameters like age, salary, work experience etc. given in the data set ‘Transport.csv’, data analysis and models have been built to predict the preferred mode of transport. The project involves building several Machine Learning models and comparing them so that the best model can be finalised.

## **InRODUCTION**

The objective is to build various Machine Learning models on this data set and based on the accuracy metrics decide which model is to be finalised for finally predicting the mode of transport chosen by the employee.

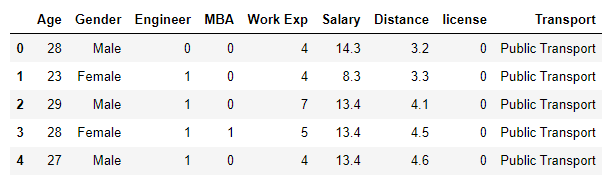
## **Data description**

* **Age:** Age of the Employee in Years.
* **Gender:** Gender of the Employee.
* **Engineer:** For Engineer =1, Non Engineer =0.
* **MBA:** For MBA =1, Non MBA =0.
* **Work Exp:** Experience in years.
* **Salary:** Salary in Lakhs per Annum.
* **Distance:** Distance in Kms from Home to Office.
* **license:** If Employee has Driving Licence -1, If not, then 0.
* **Transport:** Mode of Transport.

### 1.1 Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.

## **Data samplE**

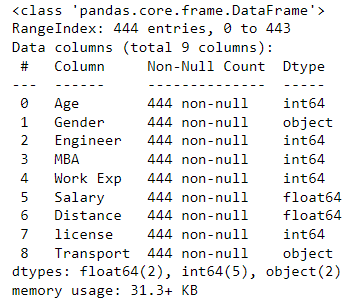
The below table shows the variables and their first five values.



**Table1.1**

There are 9 features available in the given dataset.

## **EXPLORATORY DATA ANALYSIS**

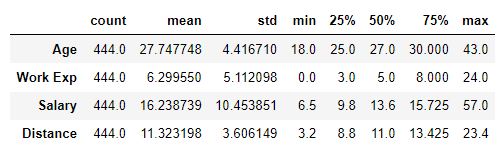


**Table1.2**

The table 1.2, shows all the datatypes of the variables and also the number of non-null datapoints present in the given dataset.

Descriptive Analysis

Below table shows the descriptive stats of all the numerical variables present in the dataset.



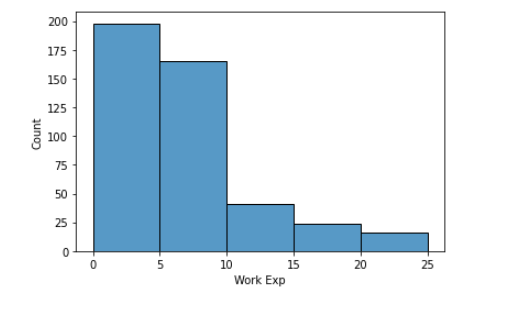
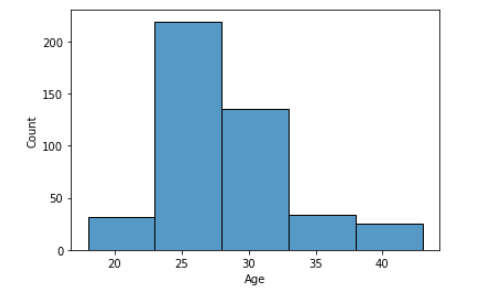
**Table1.3**

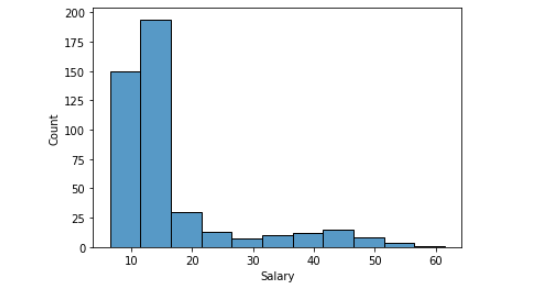
Various measures such as average, minimum value, maximum value, standard deviation, Q1, Q2 and Q3 of each numerical variable is provided in the above table. Below are some of them:

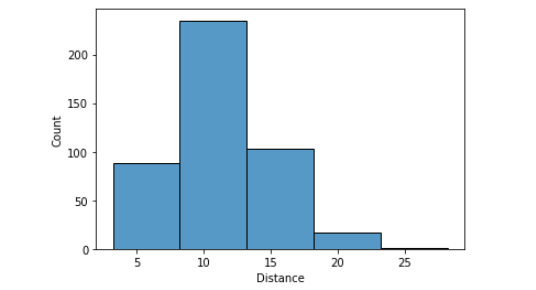
* The average age of the employees is 28, while the minimum and maximum age being 18 and 43 respectively.
* Average distance travelled by a employee is 11 kilometres.
* Similarly, we can find the values for work experience and salary variables in the above table.

Visualization

Now, let us visualize and see the spread of each numeric variables present in the dataset.





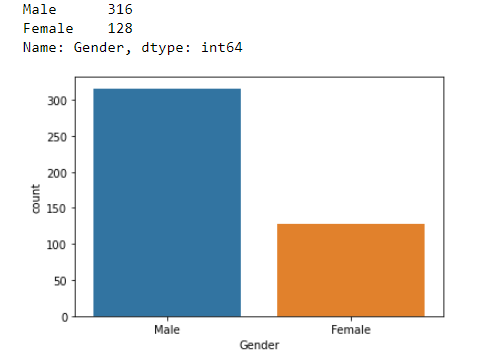


**Fig 1.1**

The above figure shows the histograms of all the numeric features available in the dataset.

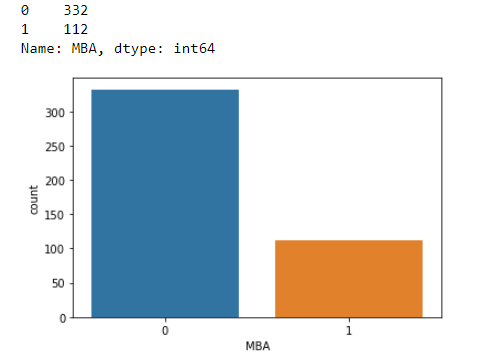
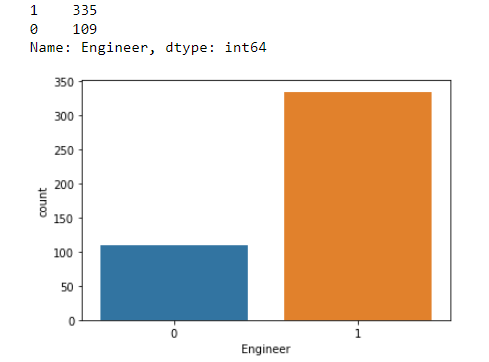
* Most the employees fall under the age group – (22-33).
* Employees with 0-5 years work experience are high in counts than other groups.
* More than 300 employees fall under 7L – 18L salary per annum.
* Only few employees travel more than 20 kms.

Now, let us see the univariate, bivariate and multivariate analysis of the variables.

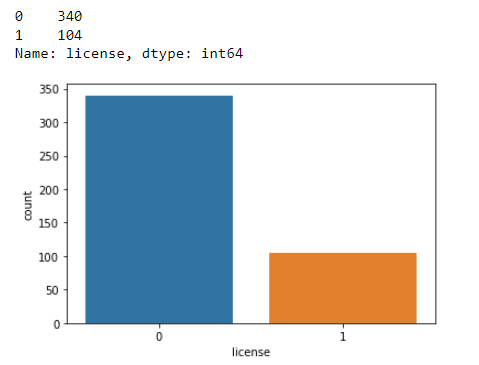
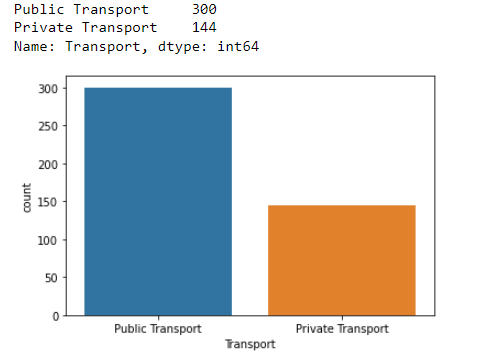


**Fig 1.2**

The Figure 1.2, provides us with the information about the male and female employees in working the company. Male employees are more in numbers than female.

  
**Fig 1.3**

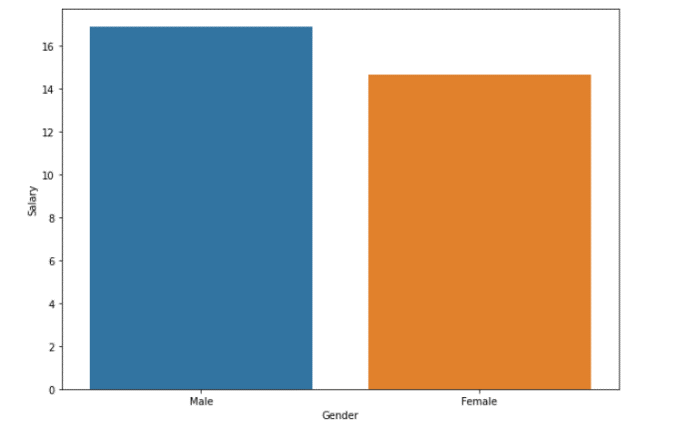
The above figure shows the number of engineers and MBA graduates working in the company. It can be seen that almost same number of engineers and MBA graduates are present in the company.

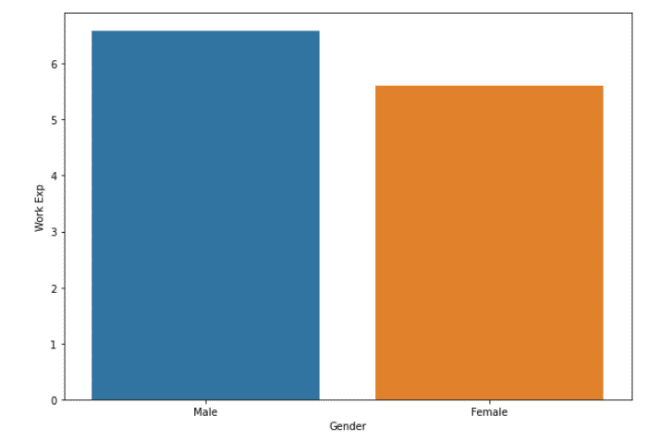
  
**Fig 1.4**

It can be interpreted from the figure 1.4, that most of the employees opt for public transport and most of them don’t have license. It may be the reason that many of the employees are choosing public transport since they don’t have license. Yet, it is too early to conclude that. So, Let’s proceed with data exploration and see what it brings on.

The Bivariate analysis of the features present in the data is done and has been visualized for better understanding here.

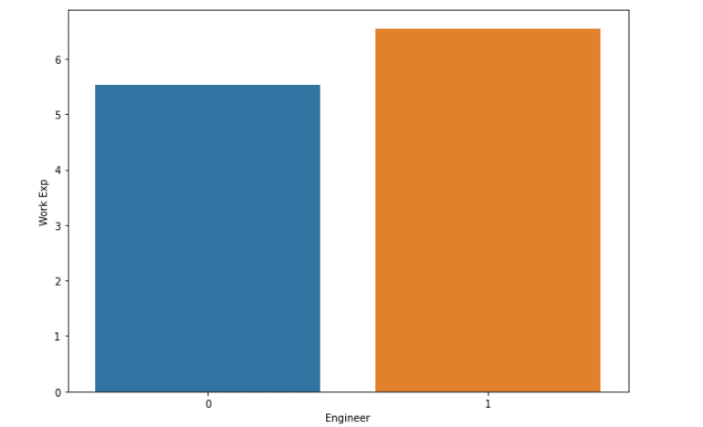
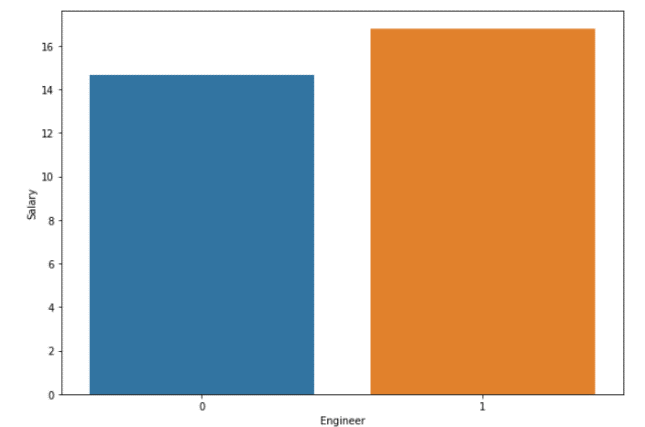
* First let us look at the Salary and Work experience of employees based on gender using the fig 1.5. Both in the salary and work experience category Male employees dominate. Also, we can see that since the work experience of male employees are higher than female employees it is a reason for higher salary.





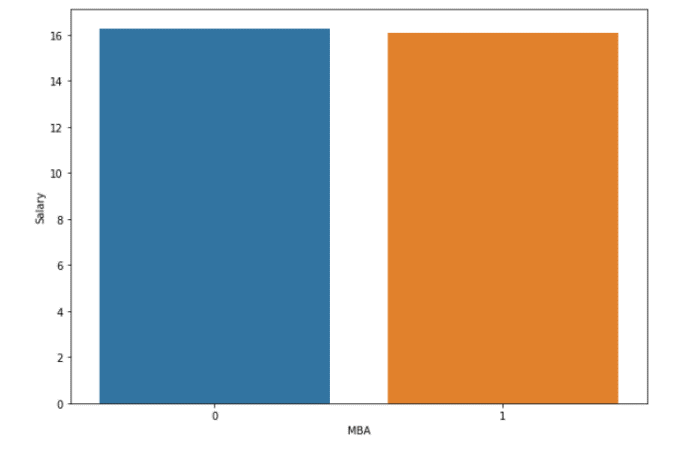
**Fig 1.5**

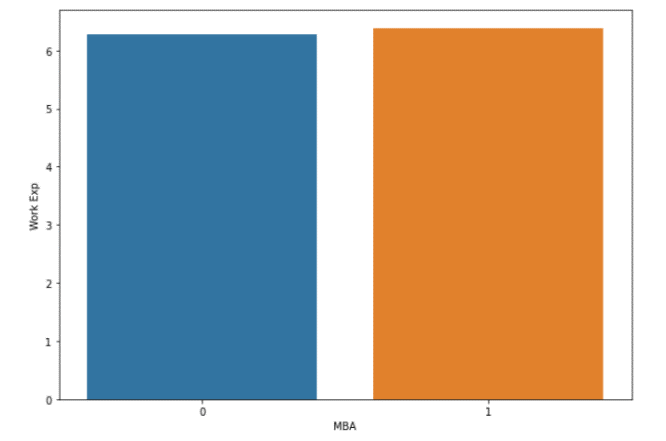
* Now, let’s look at the salary and work experience of engineers using fig 1.6.Here, we can see that Salary and Work experience of employees who are engineers is higher than what employees who are not engineer.



**Fig 1.6**

* Similar, let us look at the salary and work experience of employees who are MBA graduate with the help of below figure. The work Experience and salary of the employees with MBA are almost identical to that of the employees who are engineers.

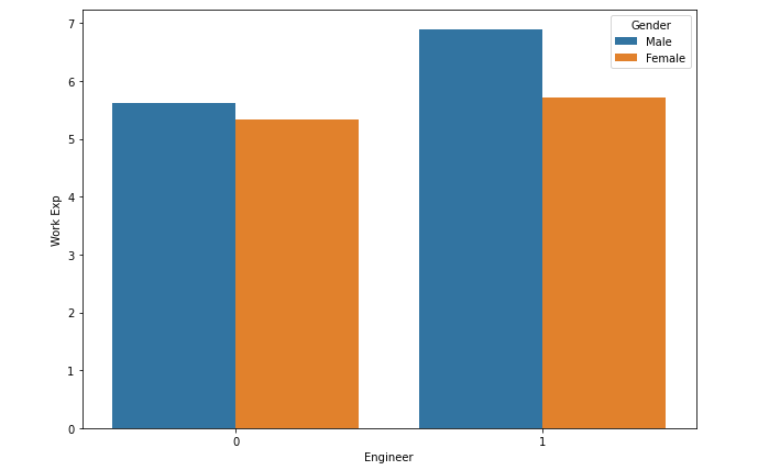




**Fig 1.7**

Multivariate Analysis has also been carried out with the features present in the dataset and let’s see what it reveals.

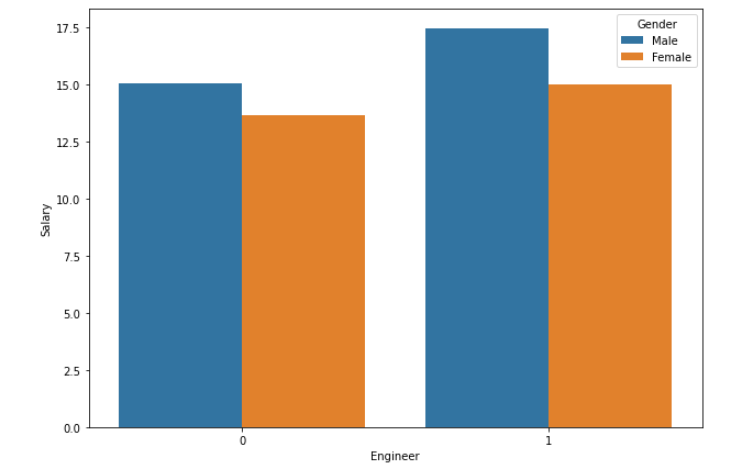
* Below figure shows the work experience of Engineers along with gender classification.



**Fig 1.8**

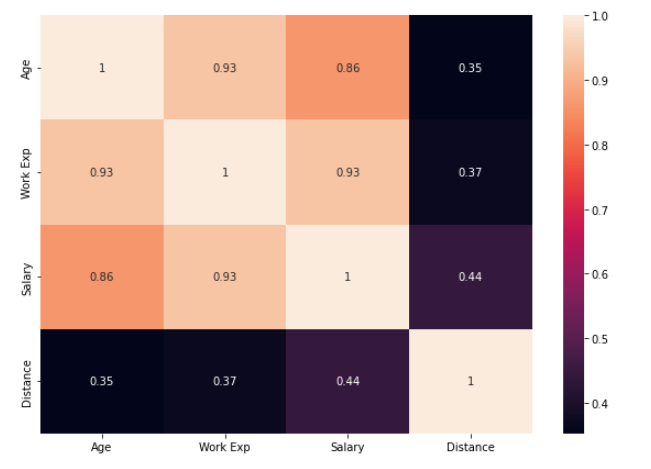
We can clearly that most of the experienced engineers are Male and also most of the non-engineer employees are also male.

* Below bar graph provides us with the information that male engineer are getting paid higher than the female engineers. As, we have already seen that the work experience of male is higher than that of the female, it might be a reason for the salary as well.



**Fig 1.8**

Below figure is the correlation heatmap of the numeric feature.



**Fig 1.9**

* We can conclude from the heatmap, that Age and Work Experience are highly correlated.
* The least correlated features are Age and Distance. Irrespective of the age the employees are travelling the average distance between work place and home every day.

### 1.2 Split the data into train and test in the ratio 70:30. Is scaling necessary or not?

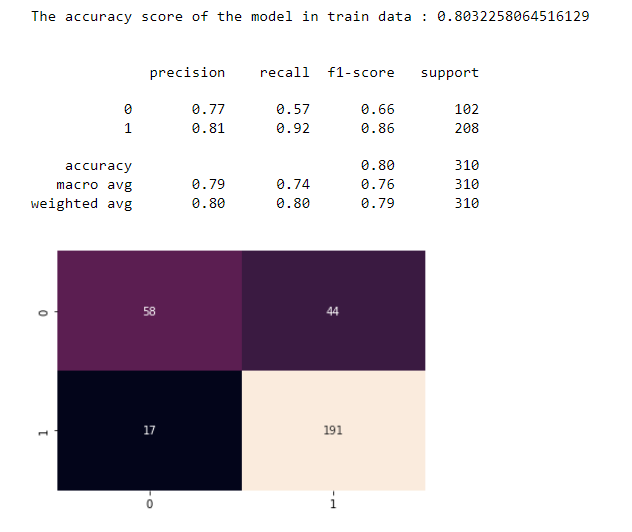
* The whole dataset has been split into train dataset and test dataset in such a way that 70% of the data is in train dataset and 30% in test dataset in order to train and test the performance the machine learning models.
* Scaling has been done for the data, since each feature is measured on different quantity. Thus, we generalize each datapoints and reduce the distance between each of them, which helps in improving the performance of the machine learning models.

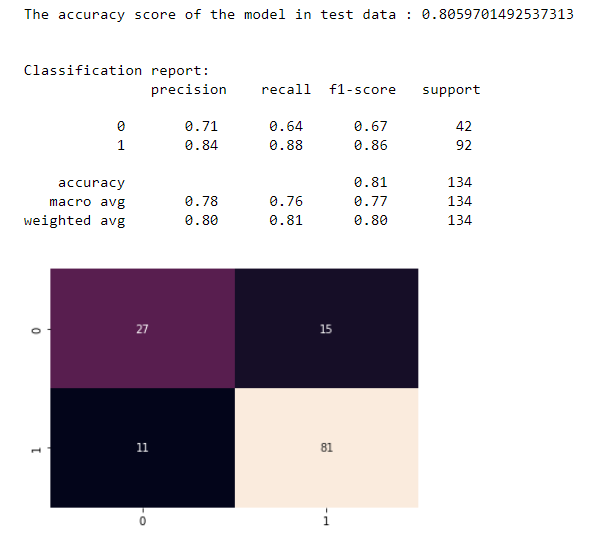
### 1.3 Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance:

With all the Exploratory data analysis and the pre-processing of the data been carried out, its time to build machine learning models check their performance to choose the best one. The performance of the model on train and test data is analysed using various metrics such as accuracy score, confusion matrix and auc value and auc-roc curve.

a. Logistic Regression Model

First model built and tested is the Logistic Regression model. Both training and testing dataset have been used to train and predict the outcome.

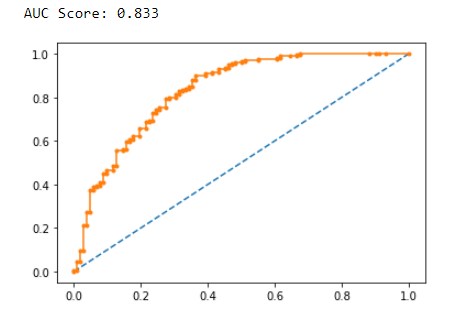


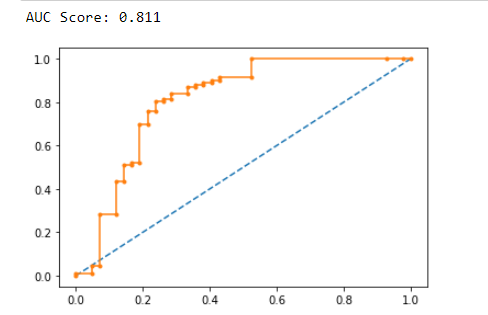


**Fig 1.10**

Figure above shows various metrics such as accuracy, precision, recall, f1-score and confusion matrix of logistic model on both training and test data. It can be seen that the logistic model performs similar on both train and test data.

Now, let us look at the auc score and auc-roc curve of the model for training and test data using fig 1.11.





Train Test

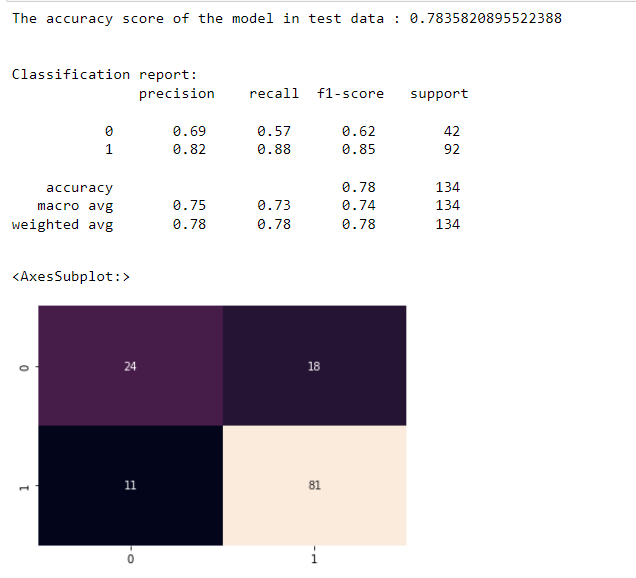
**Fig 1.11**

The figure above shows that the auc score of the test data is almost identical to that of train data.

b. Linear Discriminant Analysis

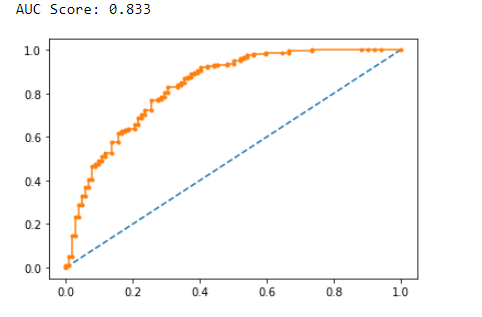
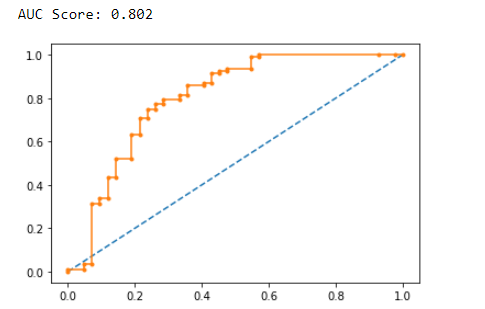


**Fig 1.12**

****

**Fig 1.12**

Various metrics of the Linear Discriminant analysis model is shown in the figure 1.12 above. We can clearly see a drop in the accuracy of the model on test data.

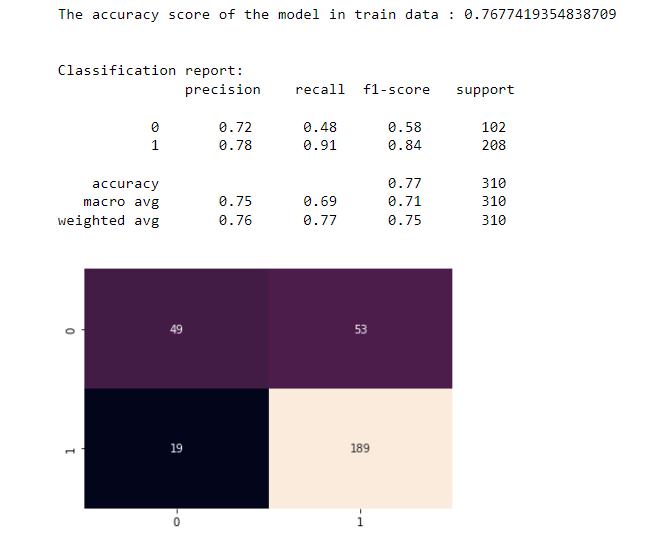
  
 Train Test

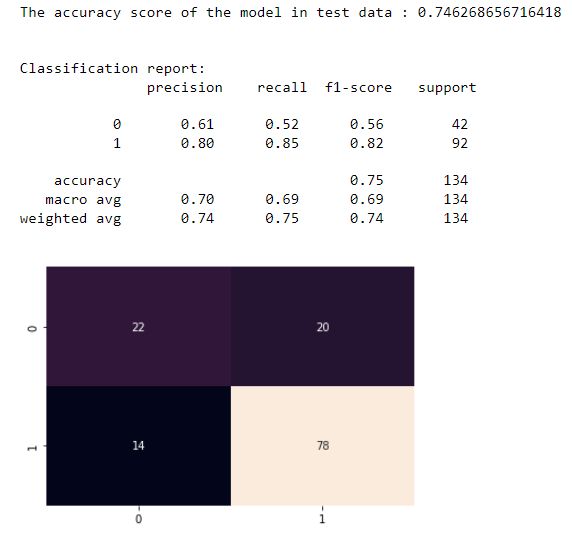
**Fig 1.13**

Just as we have seen in Fig 1.12, here the auc score has also dropped because of the accuracy of the model on test data.

c. Decision Tree Classifier – CART model

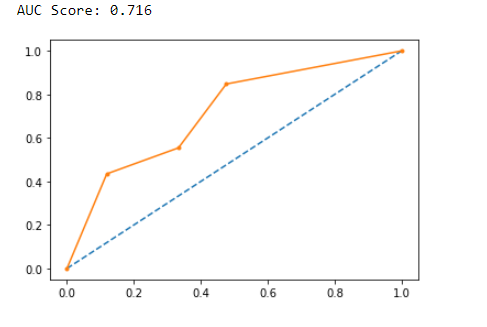
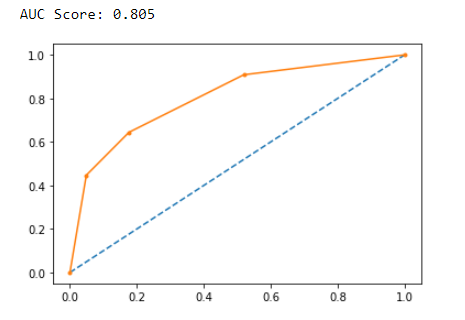
Here, the decision tree classifier built has been built and fed with train and test data to see if it performs better than the other two models.





**Fig 1.14**

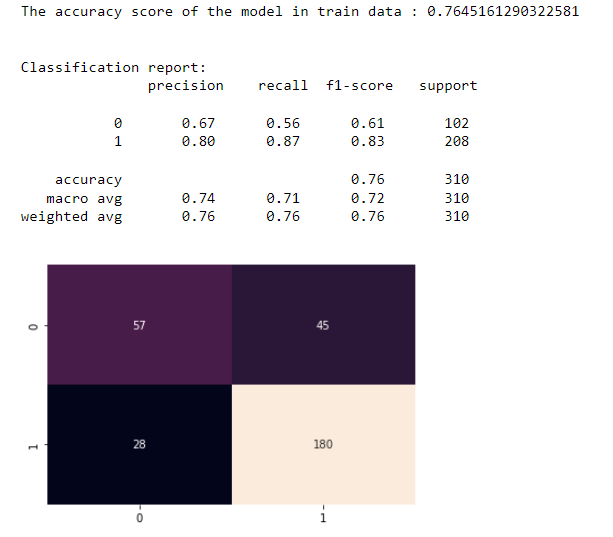
Upon on testing, the decision tree classifier model is no better than the previous model and hence is accuracy is low which resulted in low auc score (refer fig 1.15).



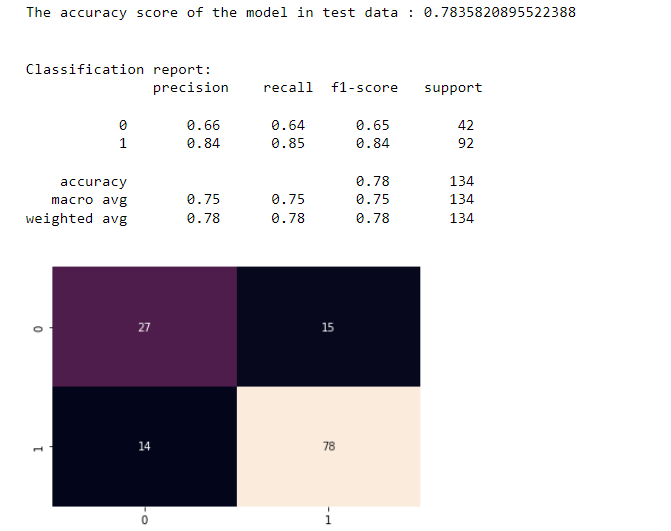
**Fig 1.15**

d. Naïve Bayes model

The Naïve Bayes model’s performance metrics on train and test data is shown fig 1.16.

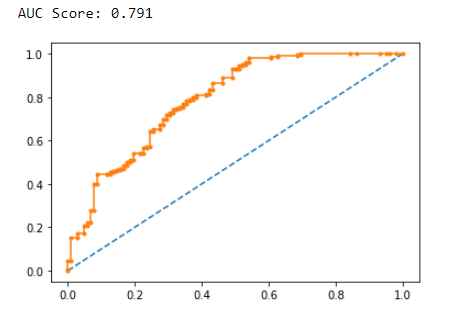


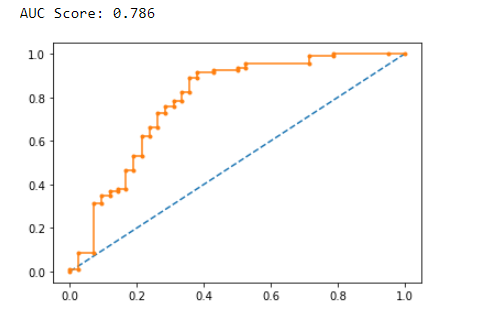
**Fig 1.16**



**Fig 1.16**

The naïve bayes model accuracy on test data is lower than accuracy of train data.



  
**Fig 1.17**

As the accuracy of the model of is low so is the auc score of the model is low.

e. KNN Model

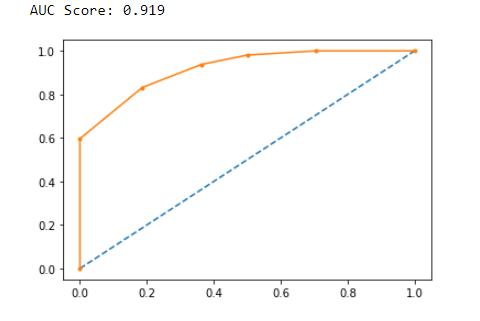
The Knn model performance metrics on train and test data is shown fig 1.18.

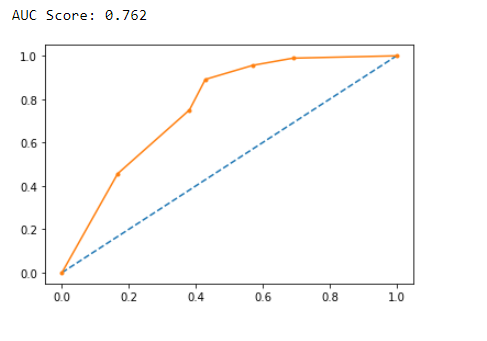




**Fig 1.18**

The performance of the knn of the model is also very low on test data just like naïve bayes model.





**Fig 1.19**

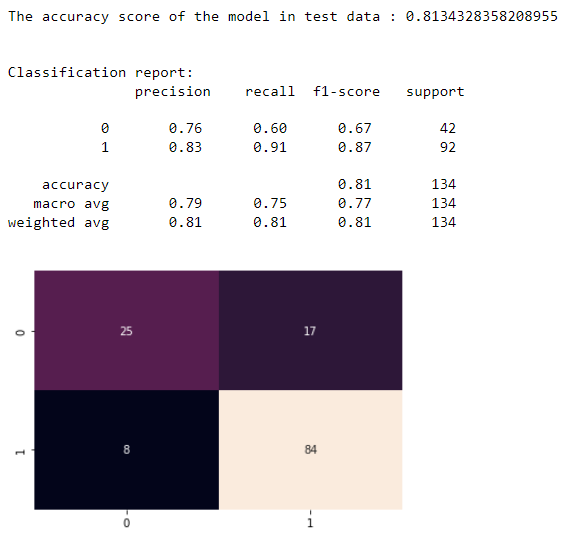
The auc score of the test data is low even though the auc score on train data.

f. Random Forest Model

The Random forest model has been build and fed with train data and tested using the test data. Figure 1.20, show the accuracy, confusion matrix and other metrics of the random forest model’s performance on train and test data.

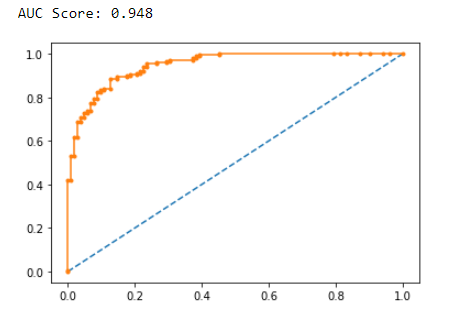
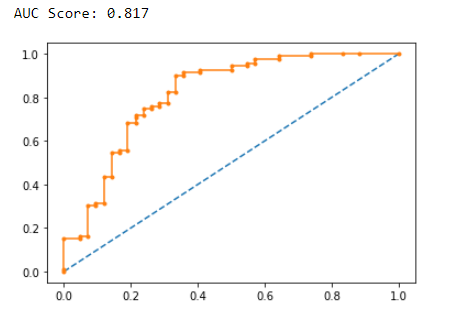


**Fig 1.20**

****

**Fig 1.20**

We can see that accuracy of the random forest model is equally better on both train and test data when compared to other machine learning models we have built so far.

  
**Fig 1.21**

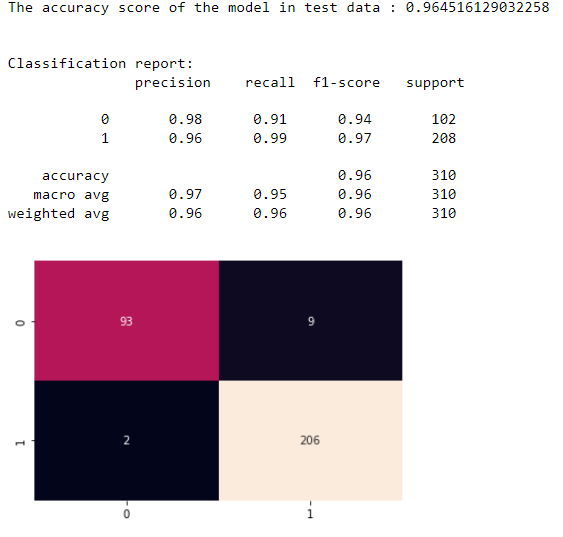
The performance of random forest model is clearly better as we saw in the fig1.20 is also reflected in the auc-roc curve in fig1.21.

g. Boosting Classifier Model using Gradient boost.

Finally, let’s build boosting classifier model using gradient boost , compare and see which of these machine learning models performs well. Thus, choosing the best model for prediction.

First, let’s see the accuracy, precision, recall, f1-score and confusion matrix of the model on both train and test data using the figure below.

Train



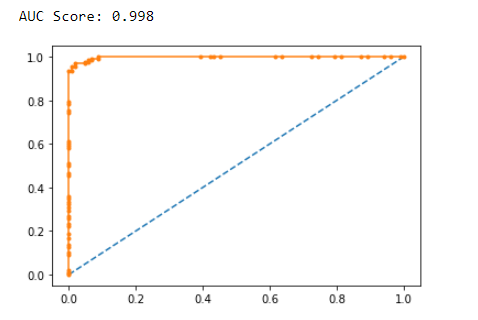
**Fig 1.22**

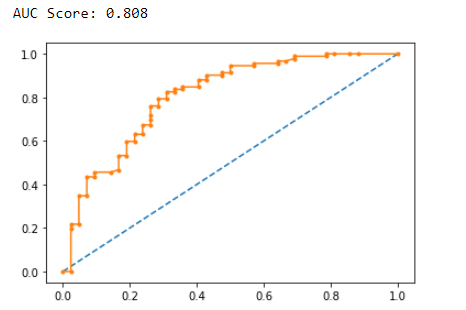
Test



The fig 1.22 provides us with required metrics to analyse the performance of the model on both train and test data. It looks similar to the rest of the model that it performs well on train data but the performance drops on test data.

The auc score and auc-roc curve in the figure below also displays that it under performs with the test data.



  
**Fig 1.23**

### 1.4 Which model performs the best?

After building and training and testing various machine learning models, we compared each models using metrics such as accuracy score, precision, recall, f1-score, support, confusion matrix, auc score and auc-roc curve it can be seen that most the models build performs similar. Yet, the Random forest classifier and the Logistic Regression models outperformed the rest the model in all the performance metrics. Based on this, we can either use Random Forest classifier or Logistic Regression Model.

### 1.5 What are your business insights?

Even though, we have built various machine learning models to predict the preferred mode of transport of the employee, the data provided is not adequate enough to accurate do the prediction.

We need to collect more data other their designation such as reason for not having license, what they like about the mode of transport they are using now, also surveys can be conducted to know whether they are comfortable with their current mode of transport or will be looking to change for something better. Thus, we can get more business insight and help them in a better way.

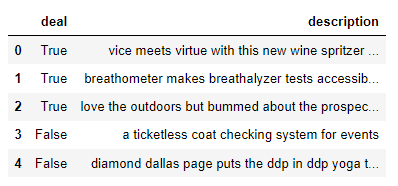
# Text Mining

## **InRODUCTION**

In this section shark tank episodes is used for text mining, creating world cloud and predicting things about the sharks.

### 2.1 Pick out the Deal (Dependent Variable) and Description columns into a separate data frame.

From the given shark tank dataset, both the deal (dependent variable) and Description variable are taken out and are clubbed to create a new data frame. Below is the newly created data frame’s first few rows.

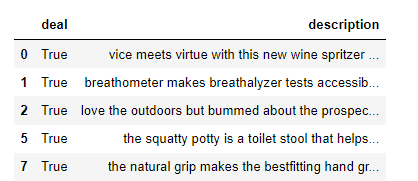


**Table 2.1**

### 2.2 Create two corpora, one for those who secured a Deal, the other for those who did not secure a deal.

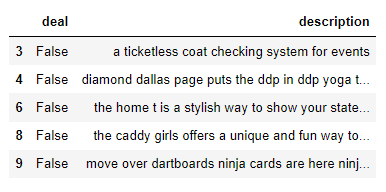
The newly dataframe is used to create two corpora, one for those who secured a deal and other for those who did not secure a deal.

Table 2.2 shows the first few rows of the corpora of the those who secured a deal.



**Table 2.2**

Similarly, Table 2.3 shows the first few rows of the corpora of the those who have not secured a deal.



**Table 2.3**

### 2.3 The following exercise is to be done for both the corpora:

### a) Find the number of characters for both the corpuses.

* Number of characters in corpora of those who secured deal is 62360.
* Number of characters in corpora of those who not secured deal is 45997.

b) Remove Stop Words from the corpora. (Words like ‘also’, ‘made’, ‘makes’, ‘like’, ‘this’, ‘even’ and ‘company’ are to be removed).

The stop words such as also, made, makes, like, this, even and company are removed by taking necessary step.

c) What were the top 3 most frequently occurring words in both corpuses (after removing stop words)?

The top 3 most frequently occurring words in both corpuses are given below. They are:

* Corpus – Deal Secured:

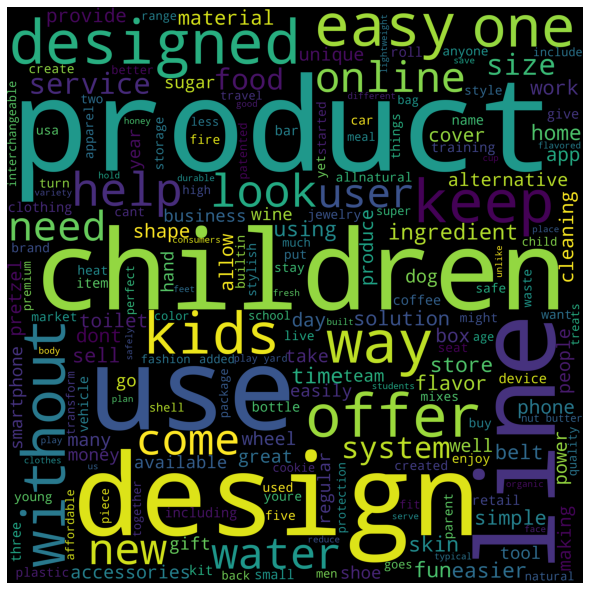
1. Products
2. Designed
3. Without

* Corpus -Deal Not Secured:
  + 1. Water
    2. Use
    3. System

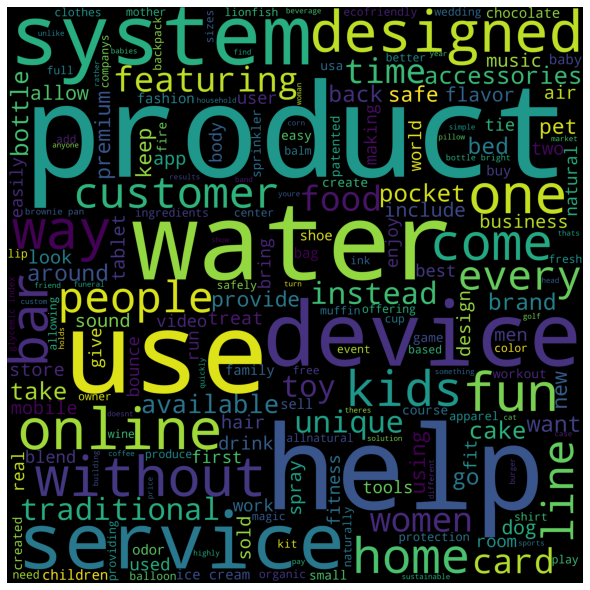
d) Plot the Word Cloud for both the corpora.

Below are the word clouds of both the corpora:

Corpus – Deal Secured



Corpus – Deal Not Secured



### 2.4 Refer to both the word clouds. What do you infer?

Upon referring to both the word cloud, we can infer that words like product, use, design are most commonly occurring words in both the corpora.

### 2.5 Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis?

Based on the word clouds of both the corpuses, we cannot conclude that the entrepreneurs

who introduced devices are less likely to secure a deal. Since, there is no clear evidence that backs up this hypothesis in both the word clouds. We can see the words such as children, use, offer, kids, easy one in the word cloud of those who secured a deal. This leads us towards a idea that the entrepreneurs who knew their end customer and designed easy to use product secured more deal than the rest.