Alternative Project-Machine Learning-29th Nov,2022.

Submitted by,

Deepa.K

Table of Contents

List of Figures:

Problem 1:

You work for an office transport company. You are in discussions with ABC Consulting company for providing transport for their employees. For this purpose, you are tasked with understanding 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’, you are required to predict the preferred mode of transport. The project requires you to build several Machine Learning models and compare them so that the model can be finalised.

Data Dictionary:

- 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.

Solution:

- The dataset has 444 entries with 9 columns in it.
- No missing values and no duplicate values present.
- Integer,float and object data types are present in the dataset.
- Outliers are present and they are treated.
- Summary statistic dataset is shown below.

	Age	Engineer	MBA	Work Exp	Salary	Distance	license
count	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000
mean	27.747748	0.754505	0.252252	6.299550	16.238739	11.323198	0.234234
std	4.416710	0.430866	0.434795	5.112098	10.453851	3.606149	0.423997
min	18.000000	0.000000	0.000000	0.000000	6.500000	3.200000	0.000000
25%	25.000000	1.000000	0.000000	3.000000	9.800000	8.800000	0.000000
50%	27.000000	1.000000	0.000000	5.000000	13.600000	11.000000	0.000000
75%	30.000000	1.000000	1.000000	8.000000	15.725000	13.425000	0.000000
max	43.000000	1.000000	1.000000	24.000000	57.000000	23.400000	1.000000

Figure 1: Summary dataset

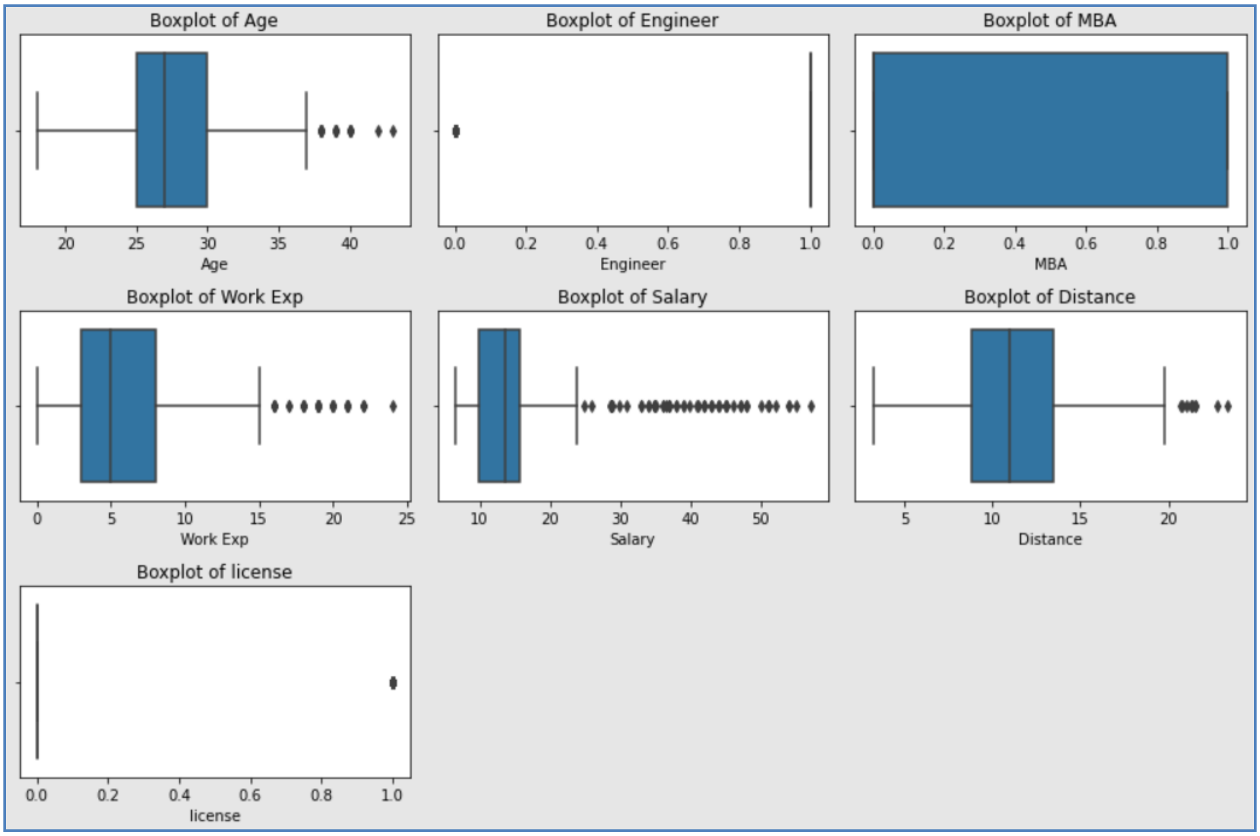

Figure 2:Boxplot

The above boxplot have outliers in almost all the entries and for the categorical variables the range varies from 0 to 1.

The outliers are removed and visualized as shown below.

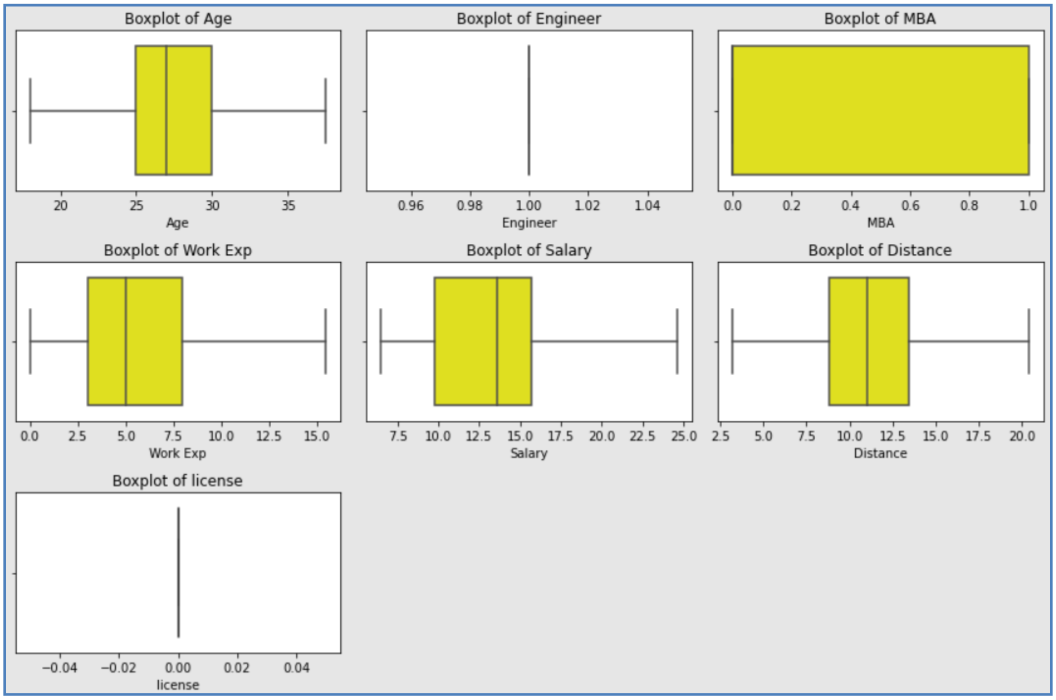

Figure 3:No outlier

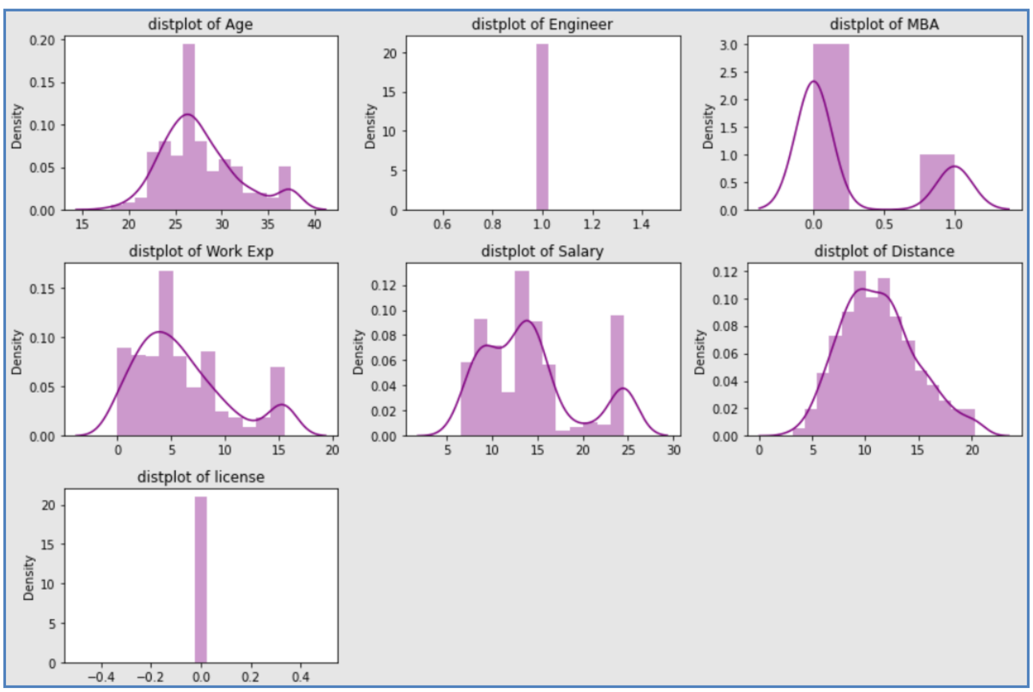

Figure 4:Distplot

There is a normal distribution seen among all the variables and no distribution found for "Engineer" and "License"categories.

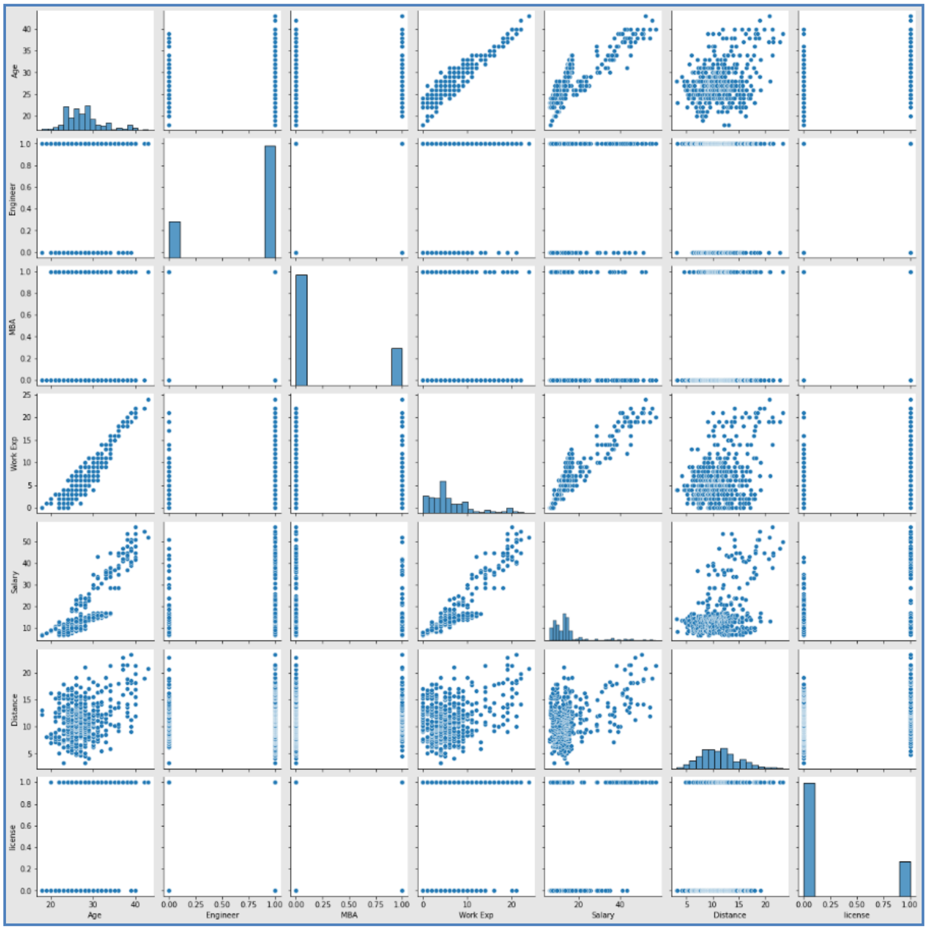

Figure 5:Pairplot

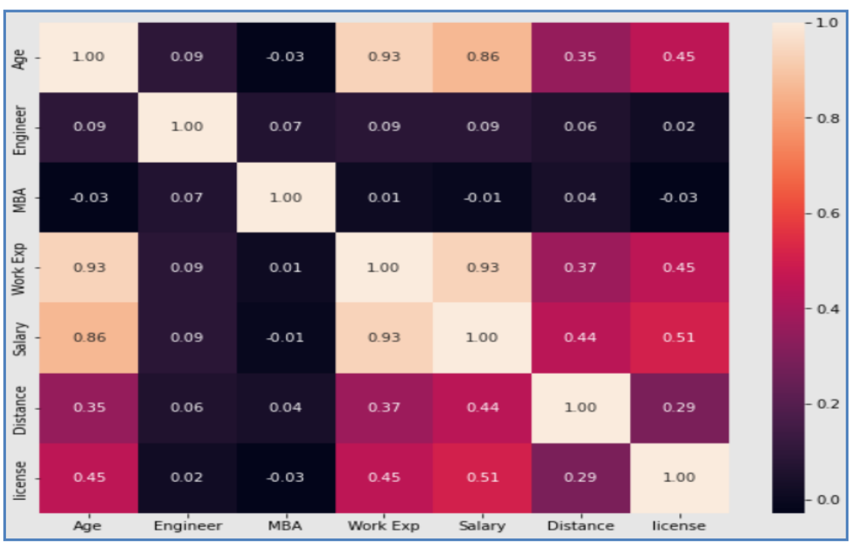

Figure 6:Heatmap

There is a strong positive correlation between age and Work experience(93%) and weakest correlation is between Age and MBA(-0.03) also between License and MBA.

The lighter shade along the diagonal is a strong positive correlation and the darker shades represent weak negative correlation.

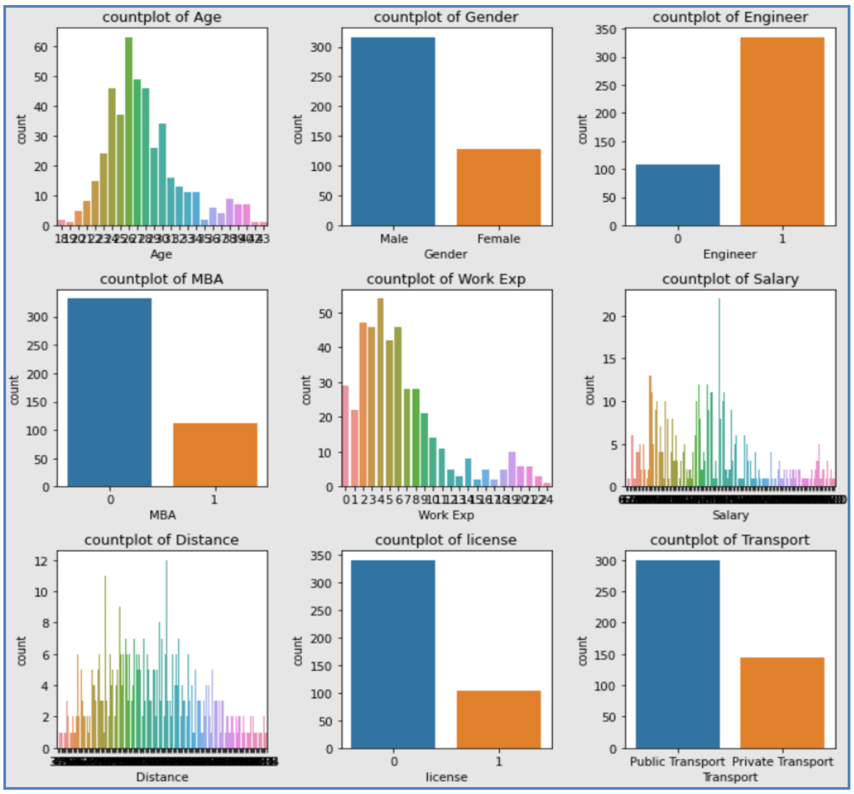

Figure 7:Countplot

The above countplot shows the distribution of attributes with their values.

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

Solution:

Data Split: X_train, X_test, y_train, y_test.

Train set: 70% data

Test set: 30% data

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license
201	29	1	0	0	5	15.9	10.5	0
386	27	1	1	1	6	12.9	15.6	0
329	27	1	1	0	6	12.9	13.3	0
249	23	1	1	0	0	6.9	11.7	0
349	30	1	1	0	7	14.9	14.0	0

Figure 8:X_Train

Scaling is necessary for converting the features to change the values of numerical or categorical variable to follow a common scale.

Here Zscore scaling is used to make the dataset precisely range from +9 to -9 values.

Z-score scaling helps in standardizing the values in same scale and using this technique helps us understand the number of standard deviations above and below the mean that each value falls.

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
count	4.440000e+02	4.440000e+02	4.440000e+02	4.440000e+02	4.440000e+02	4.440000e+02	4.440000e+02	4.440000e+02	4.440000e+02
mean	1.470295e-16	-1.220245e-16	1.797861e-16	-4.400884e-17	-1.030207e-16	3.440691e-16	-7.621531e-16	-1.500301e-17	-4.510906e-16
std	1.001128e+00	1.001128e+00	1.001128e+00	1.001128e+00	1.001128e+00	1.001128e+00	1.001128e+00	1.001128e+00	1.001128e+00
min	-2.209505e+00	-1.571226e+00	-1.753110e+00	-5.808179e-01	-1.233673e+00	-9.326442e-01	-2.255137e+00	-5.530663e-01	-1.443376e+00
25%	-6.228272e-01	-1.571226e+00	5.704149e-01	-5.808179e-01	-6.461675e-01	-6.166150e-01	-7.004825e-01	-5.530663e-01	-1.443376e+00
50%	-1.694907e-01	6.364458e-01	5.704149e-01	-5.808179e-01	-2.544974e-01	-2.527026e-01	-8.972528e-02	-5.530663e-01	6.928203e-01
75%	5.105141e-01	6.364458e-01	5.704149e-01	1.721710e+00	3.330078e-01	-4.919892e-02	5.834957e-01	-5.530663e-01	6.928203e-01
max	3.457201e+00	6.364458e-01	5.704149e-01	1.721710e+00	3.466369e+00	3.903560e+00	3.352724e+00	1.808101e+00	6.928203e-01

Figure 9:Zscore

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.

Solution:

a.Logistic Regression:

ROC curves of Logistic Regression(Train &Test):

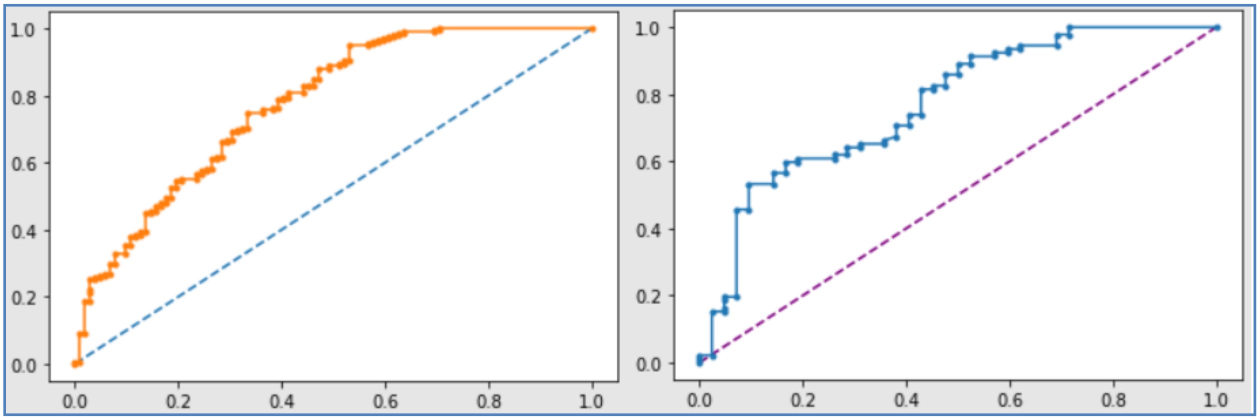

Figure 10:LR-ROC

	precision	recall	f1-score	support
0	0.79	0.47	0.59	102
1	0.78	0.94	0.85	208
accuracy			0.78	310
macro avg	0.79	0.70	0.72	310
weighted avg	0.78	0.78	0.77	310

	precision	recall	f1-score	support
0	0.66	0.50	0.57	42
1	0.79	0.88	0.84	92
accuracy			0.76	134
macro avg	0.73	0.69	0.70	134
weighted avg	0.75	0.76	0.75	134

Figure 11:LR-Metrics

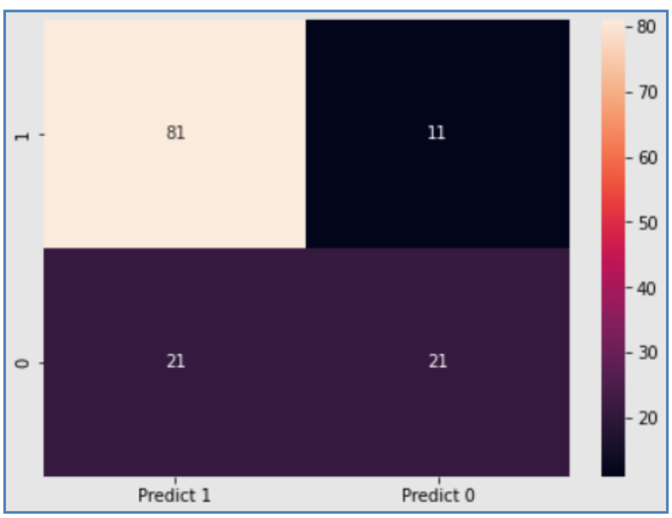

Figure 12:Confusion matrix

AUC(Train):0.776

AUC(Test):0.773

b.Linear Discriminant Analysis(LDA):

ROC curves of LDA(Train &Test):

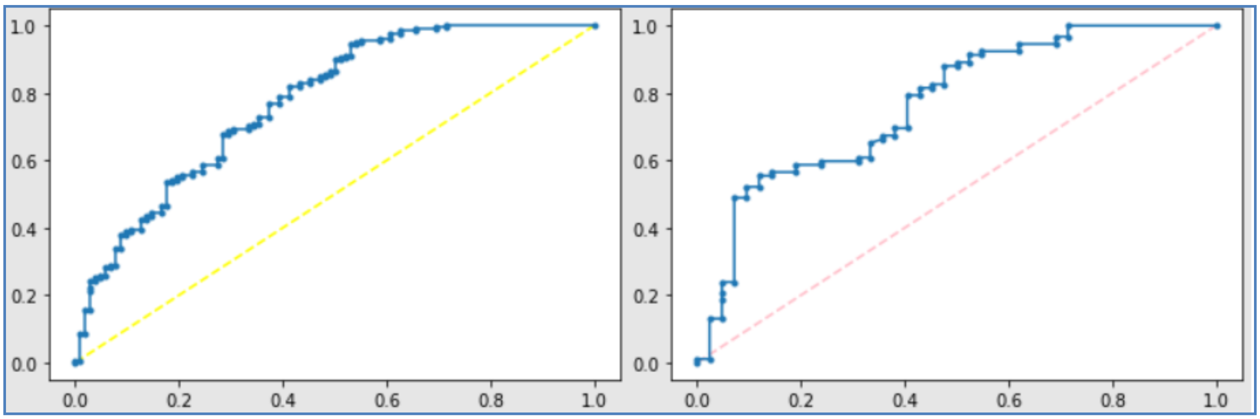

Figure 13:LDA-ROC

	precision	recall	f1-score	support
0	0.47	0.81	0.60	59
1	0.95	0.78	0.86	251
accuracy			0.79	310
macro avg	0.71	0.80	0.73	310
weighted avg	0.86	0.79	0.81	310

	precision	recall	f1-score	support
0	0.50	0.68	0.58	31
1	0.89	0.80	0.84	103
accuracy			0.77	134
macro avg	0.70	0.74	0.71	134
weighted avg	0.80	0.77	0.78	134

Figure 14: LDA-Metrics

Confusion matrix: (X_train)

[[48 11]
 [54 197]]

Confusion matrix: (X_test)

[[21 10]
 [21 82]]

AUC(Train):0.776

AUC(Test):0.771

c.Decision Tree(DT) CART Model:

ROC curves of DT(Train &Test):

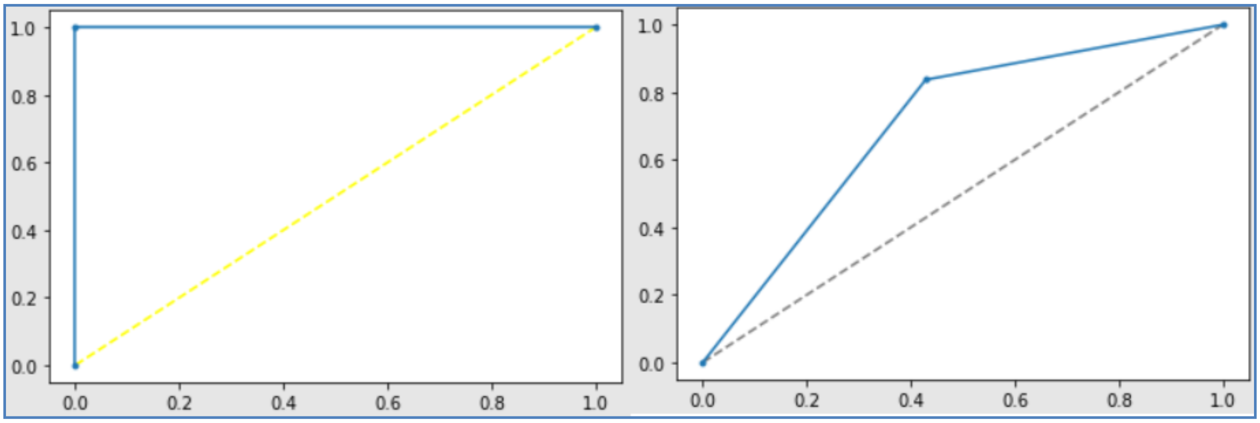

Figure 15:DT-ROC

	precision	recall	f1-score	support
0	1.00	1.00	1.00	102
1	1.00	1.00	1.00	208
accuracy			1.00	310
macro avg	1.00	1.00	1.00	310
weighted avg	1.00	1.00	1.00	310

	precision	recall	f1-score	support
0	0.62	0.57	0.59	42
1	0.81	0.84	0.82	92
accuracy			0.75	134
macro avg	0.71	0.70	0.71	134
weighted avg	0.75	0.75	0.75	134

Figure 16: DT-Metrics

Confusion matrix: (X_train)

[[102 0]
 [0 208]]

Confusion matrix: (X_test)

[[24 18]
 [14 78]]

AUC(Train):1.00

AUC(Test):0.710

d.Naive Bayes Model(NB):

ROC curves of NB(Train &Test):

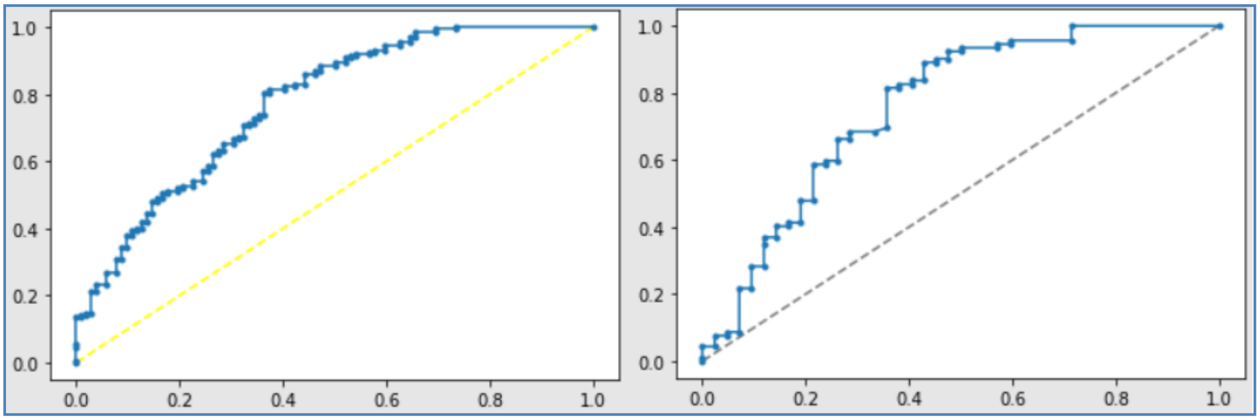

Figure 17:NB-ROC

	precision	recall	f1-score	support
0	0.79	0.40	0.53	102
1	0.76	0.95	0.85	208
accuracy			0.77	310
macro avg	0.78	0.67	0.69	310
weighted avg	0.77	0.77	0.74	310

	precision	recall	f1-score	support
0	0.78	0.43	0.55	42
1	0.78	0.95	0.86	92
accuracy			0.78	134
macro avg	0.78	0.69	0.71	134
weighted avg	0.78	0.78	0.76	134

Figure 18:NB-Metrics

Confusion matrix: (X_train)

[[41 61]
 [11 197]]

Confusion matrix: (X_test)

[[18 24]
 [5 87]]

AUC(Train):0.775

AUC(Test):0.762

e.KNN Model:

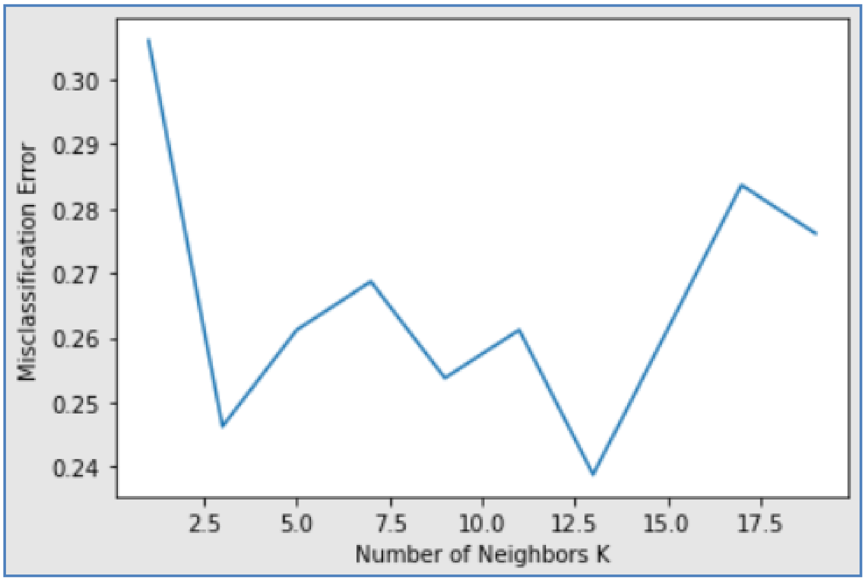

Figure 19:KNN

	precision	recall	f1-score	support
0	0.82	0.55	0.66	102
1	0.81	0.94	0.87	208
accuracy			0.81	310
macro avg	0.82	0.75	0.76	310
weighted avg	0.81	0.81	0.80	310

	precision	recall	f1-score	support
0	0.65	0.52	0.58	42
1	0.80	0.87	0.83	92
accuracy			0.76	134
macro avg	0.72	0.70	0.71	134
weighted avg	0.75	0.76	0.75	134

Figure 20:KNN Metrics

Confusion matrix: (X_train)

[[56 46]
 [12 196]]

Confusion matrix: (X_test)

[[22 20]
 [12 80]]

AUC(Train):0.812

AUC(Test):0.761

The difference between train and test set accuracies is 5% which is a valid model.

f.Random Forest Model(RF):

ROC curves of RF(Train &Test):

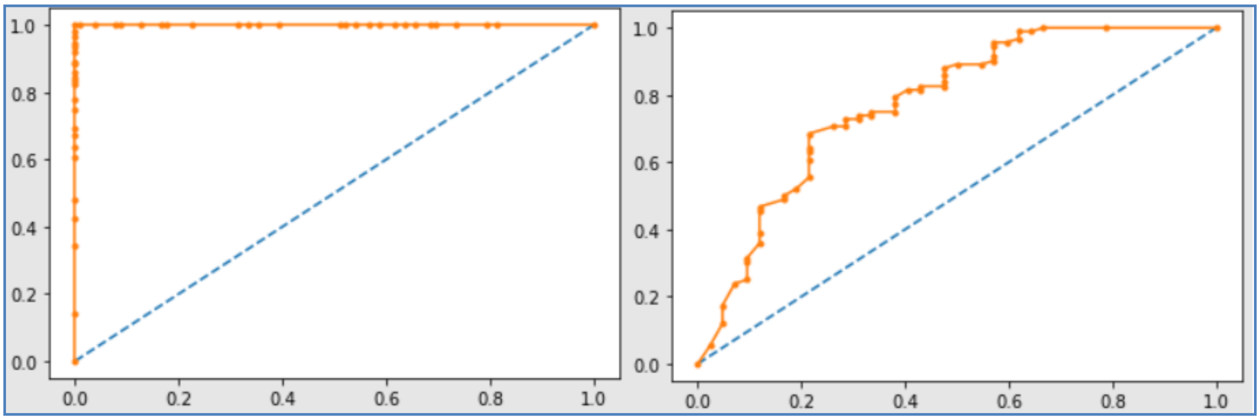

Figure 21:RF-ROC

	precision	recall	f1-score	support
0	1.00	1.00	1.00	102
1	1.00	1.00	1.00	208
accuracy			1.00	310
macro avg	1.00	1.00	1.00	310
weighted avg	1.00	1.00	1.00	310

	precision	recall	f1-score	support
0	0.67	0.52	0.59	42
1	0.80	0.88	0.84	92
accuracy			0.77	134
macro avg	0.73	0.70	0.71	134
weighted avg	0.76	0.77	0.76	134

Figure 22:RF-Metrics

Confusion matrix: (X_train)

[[102 0]
 [0 208]]

Confusion matrix: (X_test)

[[22 20]
 [11 81]]

AUC(Train):1.00

AUC(Test):0.779

g.Boosting classifier model using gradient boost:

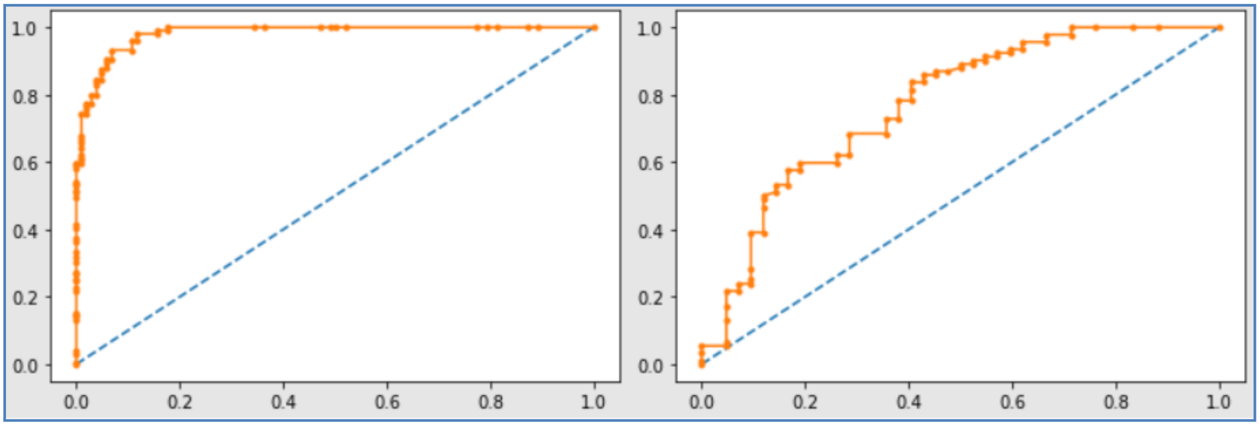

Figure 23:GBCL-ROC

	precision	recall	f1-score	support
0	1.00	0.82	0.90	102
1	0.92	1.00	0.96	208
accuracy			0.94	310
macro avg	0.96	0.91	0.93	310
weighted avg	0.95	0.94	0.94	310

	precision	recall	f1-score	support
0	0.68	0.50	0.58	42
1	0.80	0.89	0.84	92
accuracy			0.77	134
macro avg	0.74	0.70	0.71	134
weighted avg	0.76	0.77	0.76	134

Figure 24: GBCL-Metrics

Confusion matrix: (X_train)

[[84 18]
 [0 208]]

Confusion matrix: (X_test)

[[21 21]
 [10 82]]

AUC(Train):0.982

AUC(Test):0.772

1.4: Which model performs the best?

Solution:

Let's look at the performance of all the models on the Train and Test Dataset.

Recall refers to the percentage of total relevant results correctly classified by the algorithm and hence we will compare Recall of class "1" for all models.

Recall @ 1	Train Dataset	Test Dataset
Logistic Regression	0.94	0.88
LDA	0.78	0.80
Decision Tree	1.00	0.84
Naïve Bayes	0.95	0.95
KNN(@K=13)	0.94	0.87
Random Forest	1.00	0.88
Gradient Boosting	1.00	0.89

Table 1:Model Values

Model which have not performed well on the train data set also have not performed well on the test data set. However Decision Tree, Random Forest and Gradient boosting classifier which had a 100% score on the train data set have shown a poor result on the test data set. Hence a clear case of overfitting.

So the best model is Gradient Boosting classifier model.

1.5: What are your business insights?

Solution:

- Decision Tree,Random forest and Gradient boosting classifier performs best.
- The model performance heavily depends on the type of input data and distributions.
- Model building is an iterative process and performance can be improved by using feature engineering, feature extraction and hyper parameter tuning.
- Hence there are more chances of choosing private transport as per recall 1 of each model.

Problem 2:

A dataset of Shark Tank episodes is made available. It contains 495 entrepreneurs making their pitch to the VC sharks. You will ONLY use "Description" column for the initial text mining exercise.

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

Solution:

	deal	description
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
5	True	One of the first entrepreneurs to pitch on Sha...

Figure 25:Dataframe

The basic preprocessing such as viewing info,summary statistics,checking data types,dropping unnecessary columns are done.

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

Solution:

	deal	description
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
5	True	One of the first entrepreneurs to pitch on Sha...
6	False	A mixed martial arts clothing line looking to ...
7	False	Attach Noted is a detachable "arm" that holds ...
8	False	A safety device for seatbelts. It prevents the...
12	True	A line of books written to help children find ...
16	True	Coverplay is a slipcover for children's play y...

Figure 26:Deal&Description

	deal	description
count	204	204
unique	1	203
top	True	Echo Valley Meats is a retail, online gift cat...
freq	204	2

Figure 27:True deal

	deal	description
count	183	183
unique	1	182
top	False	Premium wine sold by the glass in individually...
freq	183	2

Figure 28:False deal

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

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

Solution:

- True corpus: 50302
- False corpus: 34899

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

Solution:

Stop words are frequently occurring words that do not add value to the analysis,hence should be removed.NLTK package has an in built list of 179 'stopwords'.We use this list to remove any occurrence of such words.

Comma and punctuations are removed.

d).Plot the Word Cloud for both the corpora.

Solution:

Figure 29:True corpora

Figure 30:False corpora

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

Solution:

The 'secured a deal' wordcloud contains words such as 'one', 'design' , 'free' ,'children' ,'offer', 'easy' ,'online','use' .These indicate that Deals aimed towards catering to the children, which provided offers or a free sample/product.

The 'Did not secure a deal' wordcloud contains words such as 'one', 'designed' , 'help' ,'device' ,'bottle', 'premium' ,'use' .These indicate that Deals with a mediocre design, less suited to solve/help a problem.

Words such as 'one', 'designed' ,'system' and 'use' have a higher weight in both these wordclouds.This indicates that either these were not the defining factors to whether a deal is made or not.

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?

Solution:

- The word 'device' is not easily found in the 'secured a deal' wordcloud while it is easily spotted in 'not secured a deal' wordcloud.
- This indicates that the word 'device' occurred frequently when a deal was rejected hence implying the statement given in the question is true.

************************THANK YOU*******************************