Question 1- Machine learning models

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 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.

Question 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-

-Reading of the dataset is done. Head of dataset

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28	Male	0	0	4	14.3	3.2	0	Public Transport
1	23	Female	1	0	4	8.3	3.3	0	Public Transport
2	29	Male	1	0	7	13.4	4.1	0	Public Transport
3	28	Female	1	1	5	13.4	4.5	0	Public Transport
4	27	Male	1	0	4	13.4	4.6	0	Public Transport

Independent variables- Age, Gender, Engineer, MBA, Work exp, Salary, Distance, License.
Dependent variable-Transport.

- Summary of dataset-

RangeIndex: 444 entries, 0 to 443
Data columns (total 9 columns):

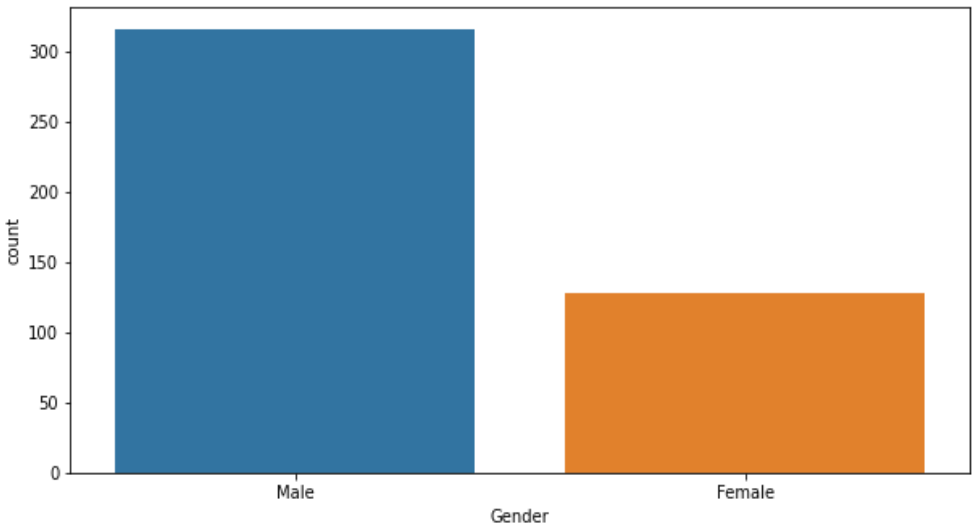
#	Column	Non-Null Count	Dtype
0	Age	444 non-null	int64
1	Gender	444 non-null	object
2	Engineer	444 non-null	int64
3	MBA	444 non-null	int64
4	Work Exp	444 non-null	int64
5	Salary	444 non-null	float64
6	Distance	444 non-null	float64
7	license	444 non-null	int64
8	Transport	444 non-null	object

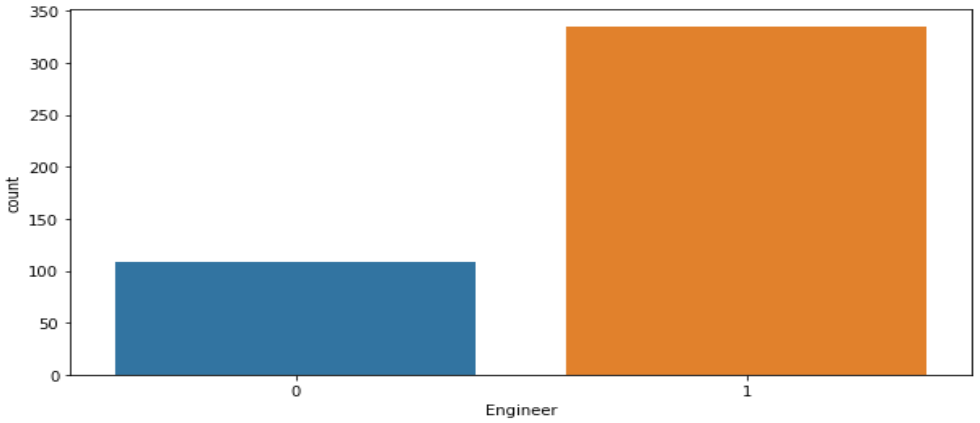
dtypes: float64(2), int64(5), object(2)
memory usage: 31.3+ KB

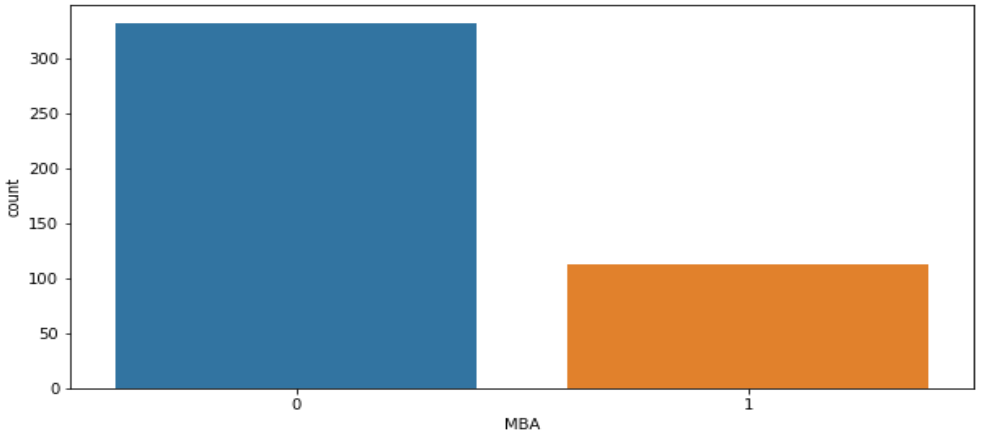
Inference- There are 9 columns and 444 rows from 0 to 443

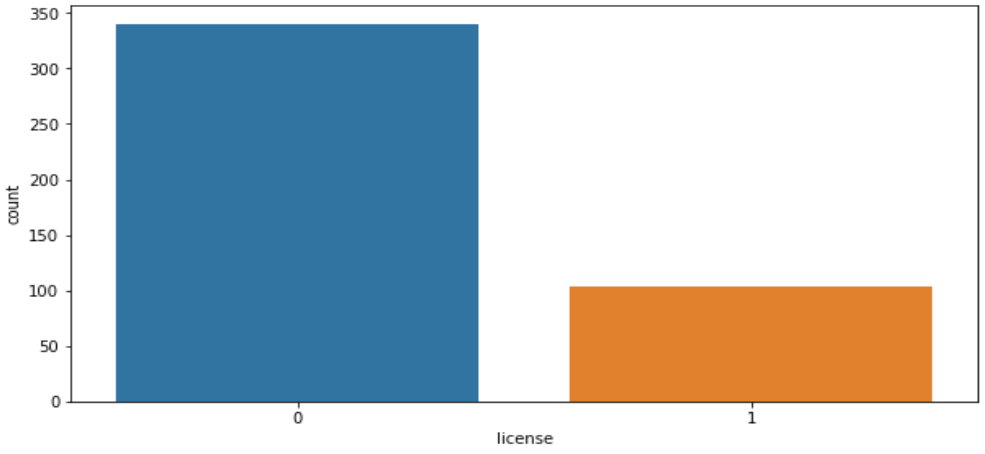
- Object variables- Gender, Transport
- Numerical variables- Age,Engineer ,MBA,Work exp, Salary, Distance,License.
- There are no null values.
- There are no duplicate values.

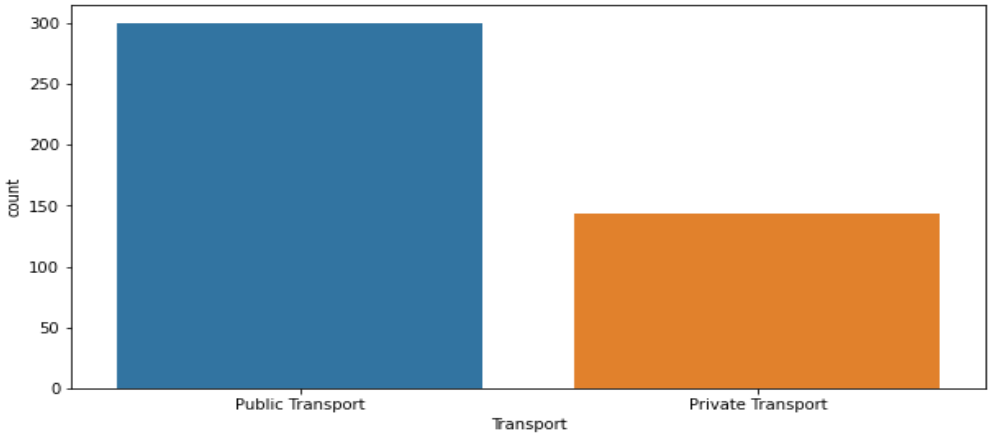
- <u>Univariate Analysis-</u>
    1) Categorical variables- Count plot has been used to visualize categorical variables




count
0
50
100
150
200
250
300
350
0
1
Engineer


count
0
50
100
150
200
250
300
0
1
MBA

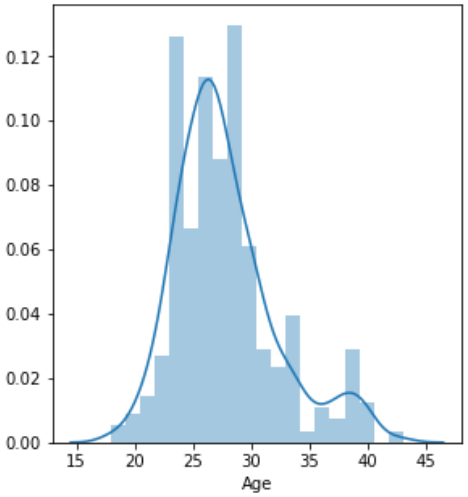

count
0
50
100
150
200
250
300
350
0
1
license

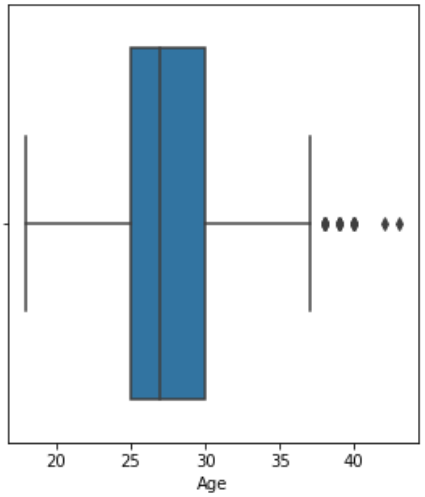

count
0
50
100
150
200
250
300
Public Transport
Private Transport
Transport

Inference - Male employees are more than female employees.
- Employees with an engineering degree are more than without an engineer degree.
- Employees with no mba degree are more than with mba degree.
- Most of the employees don't have driving licenses.
-Most of the employees are using public transport.

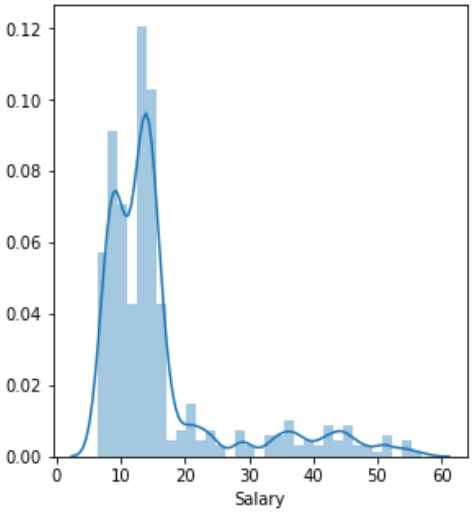
2)Continuous variables- Distribution plot is used to see the distribution of data and box plot is used to visualize any outliers.

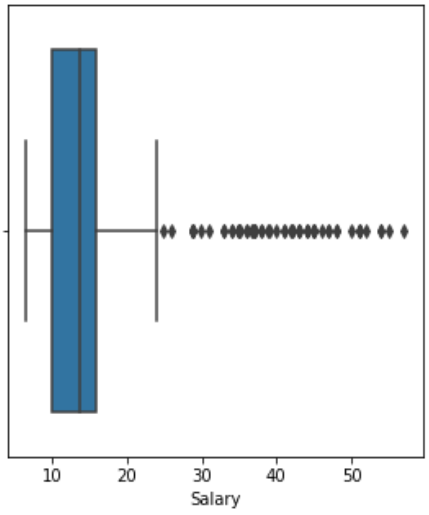
Age
Skewness : 0.9552759761192868
kurtosis : 0.9388711328850645



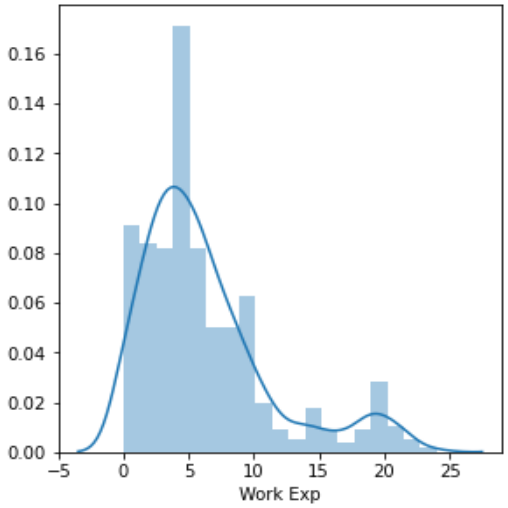


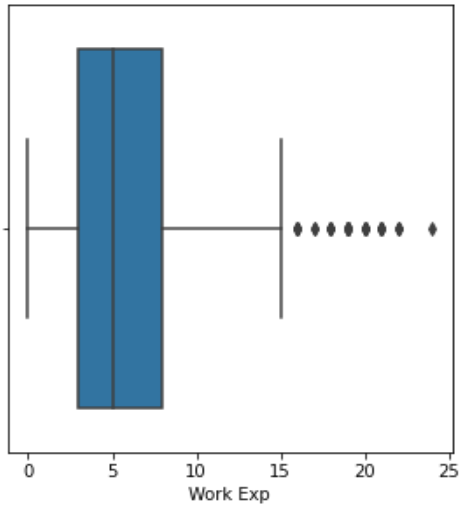
Salary
Skewness : 2.0445329291548857
kurtosis : 3.479376931422676



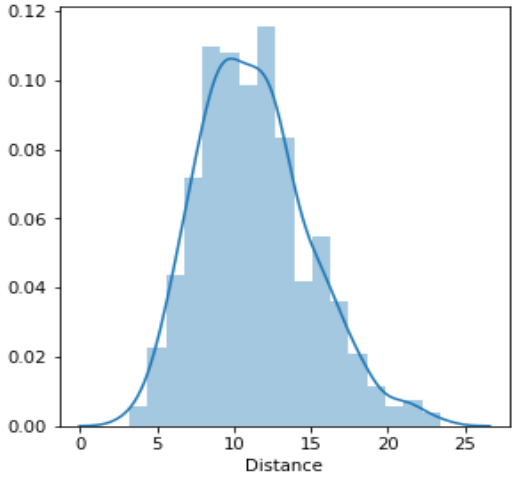


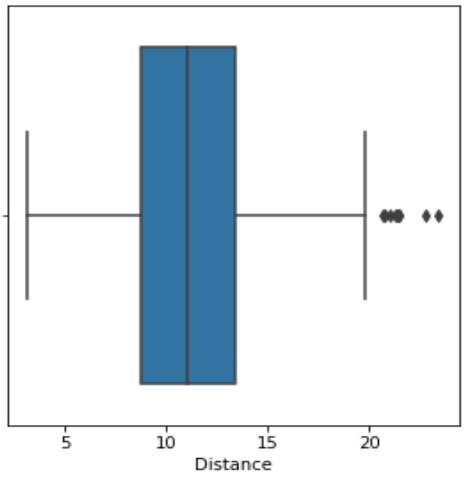
Work Exp
Skewness : 1.3528403114201042
kurtosis : 1.4785733585322038





Distance
Skewness : 0.5398513071476282
kurtosis : 0.19146510525849614





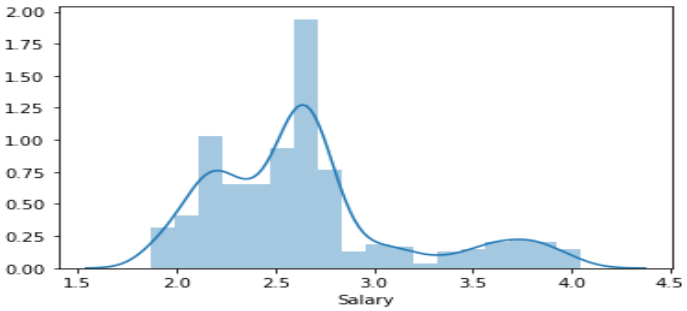
Inference

- All variables are normally distributed except salary which is positively skewed.
- All variables are having outliers on their higher side.

Correction of Skewness-

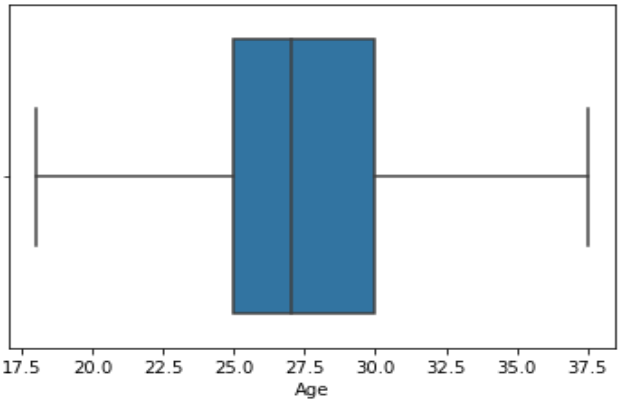
-Correction of skewness of salary variable is done by using log conversion.

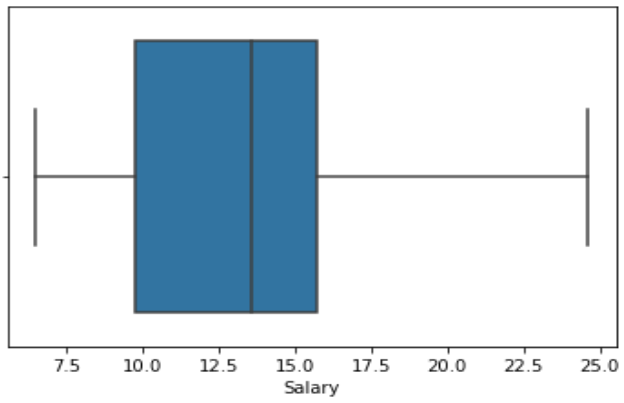
-Plot after correction of skewness-

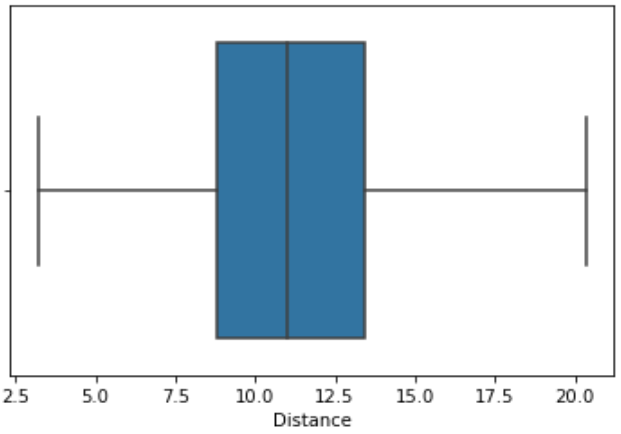


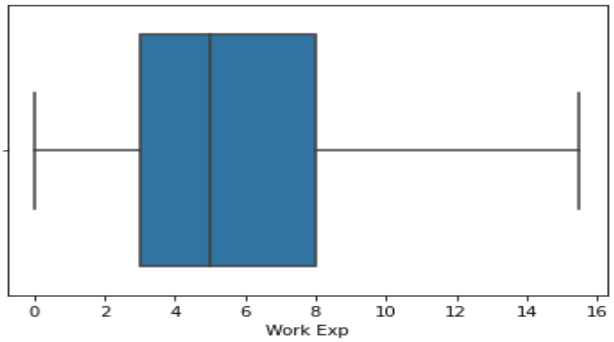
Correction of outliers-

- Correction of outliers is done by using inter quartile range.
- plots after correction of outliers-





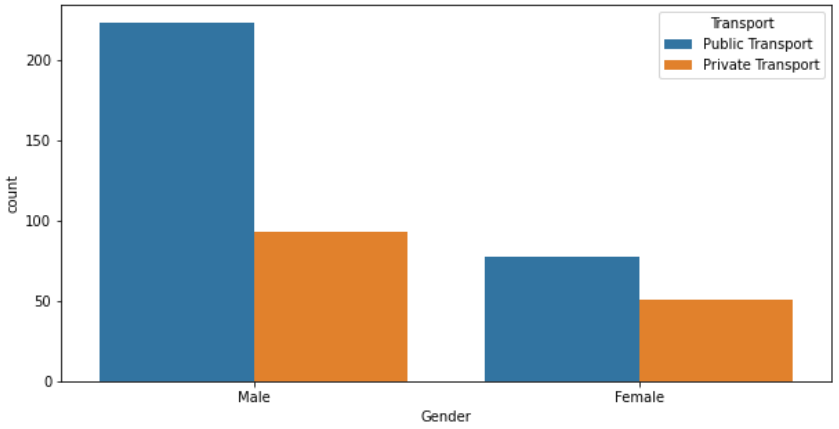


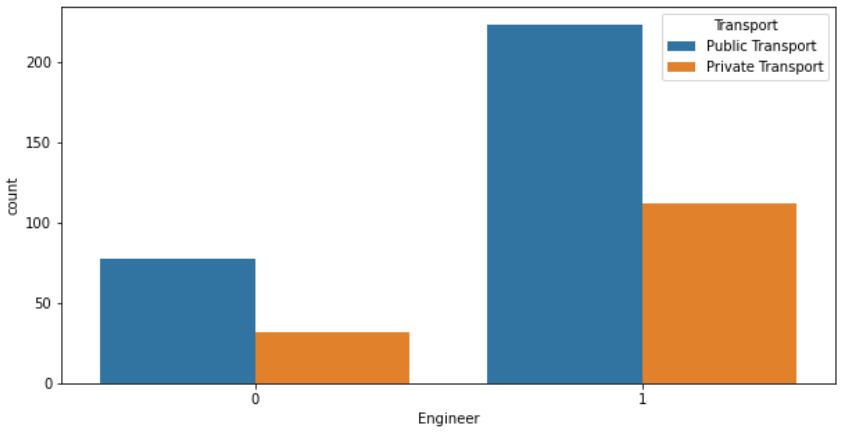


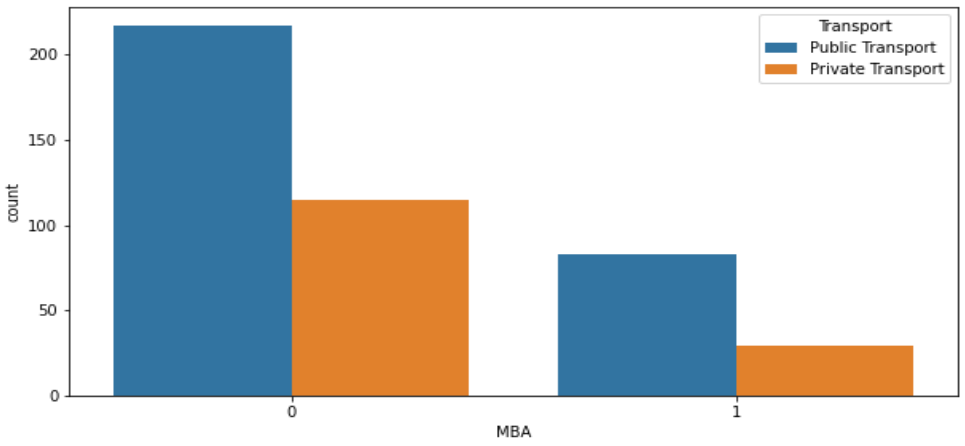
2_Bivariate Analysis-

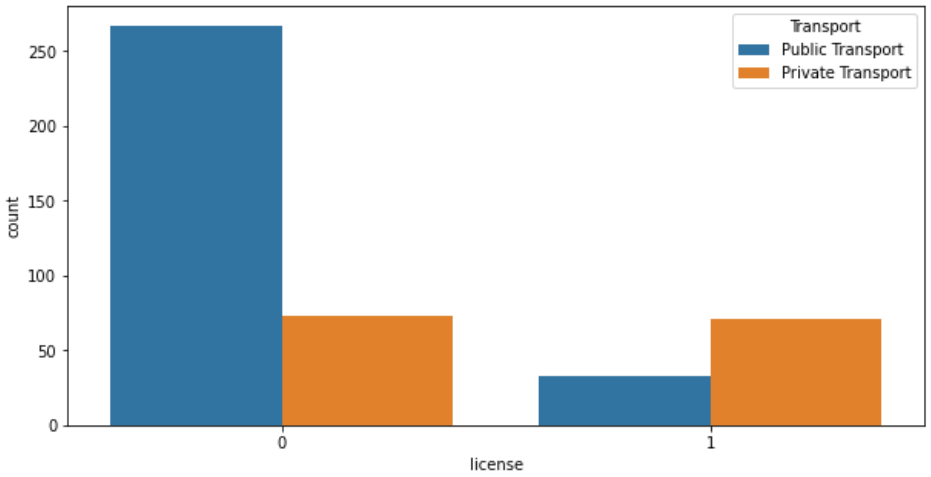
1)Categorical variables-

- Count plot is used by keeping independent variables on the x axis and dependent variables as hue.
- Plots of categorical variables are-









Inference- More males and females are using public transport

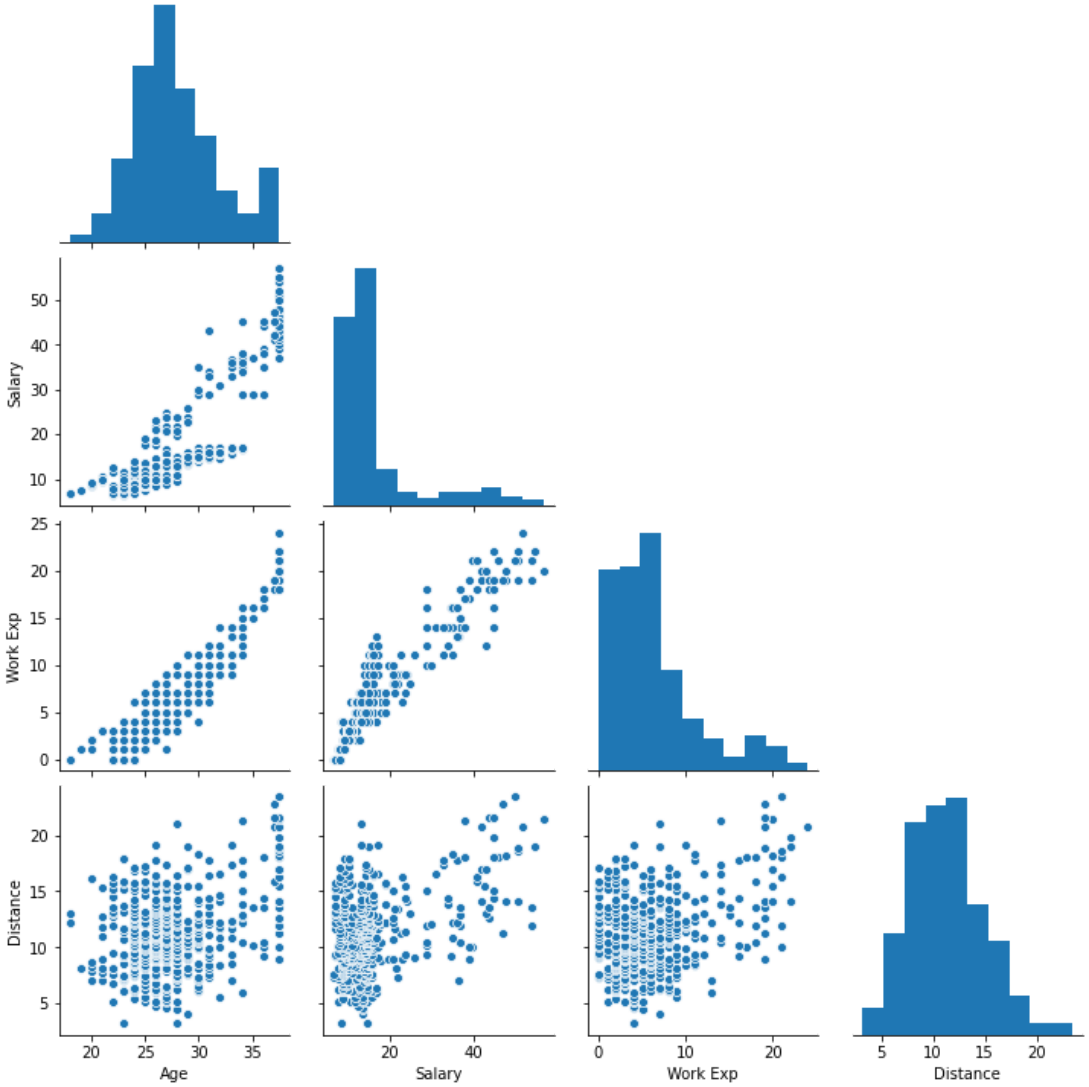
- More employees having engineer degrees and not having engineer degrees are using public transport.

-More employees with and without MBA degrees are using public transport.

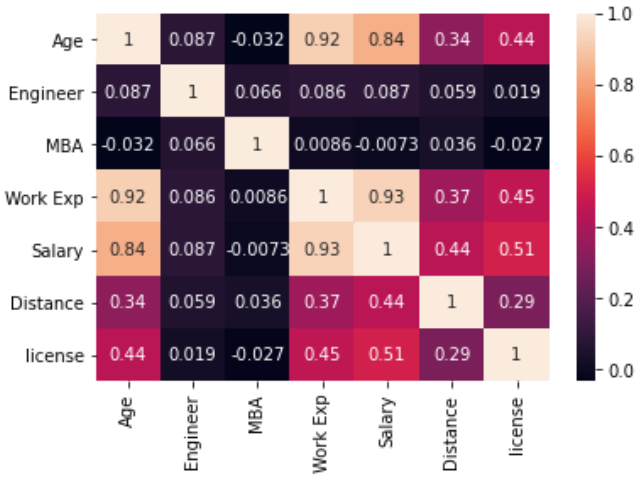
- More employees with no driving license are using public transport and more employees with driving licenses are using private transport.

2)Continuous variables- Pairplot and Heatmap are used to see any correlation between different continuous variables.

Pairplot



Heatmap



Inference- Salary and age are having positive correlation.
- Work experience and age are having positive correlation.
- Salary and work experience are having positive correlation.

Question 1.2 -

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

Solution-
-70% of data is split into a train set and 30% into a test set by using the train test split function from sklearn.
- Scaling is not necessary for this data as values of all the variables are almost in the same range. So there will be no discrimination based on the values while using different algorithms.

Question 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.:
a. Logistic Regression Model
b. Linear Discriminant Analysis
c. Decision Tree Classifier – CART model
d. Naïve Bayes Model
e. KNN Model
f. Random Forest Model
g. Boosting Classifier Model using Gradient boost.

Solution-

1) Logistic Regression-Logistic regression model works by assigning probabilities to different classes to which a query point is likely to belong.To do so , it learns from the training set a vector of weight and bias.Each weight is assigned to one input variable. To classify a query point,the classifier takes the weight sum of features and bias to represent the evidence of the query point belonging to the class of interest.

$$Z=WX+B$$

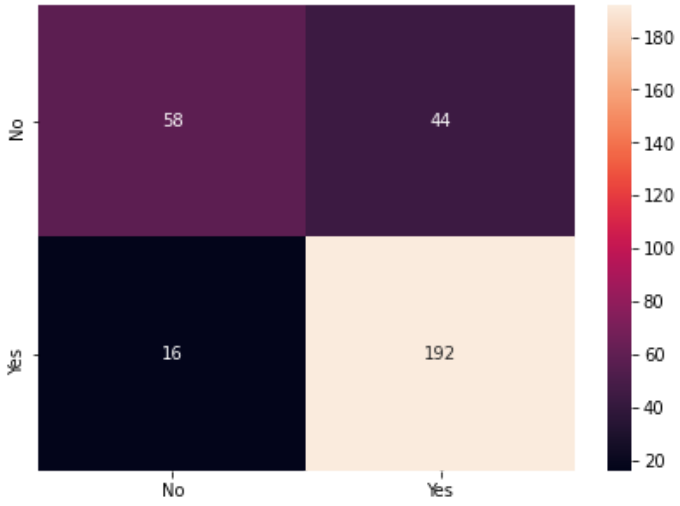
W = Weight
X = input variable
B= Bias

To transform the z value into probability, Z is passed through the sigmoid function - The algorithm uses a cross entropy loss function to find optimal weight and bias across the entire data set put together. - Here we have made a logistic regression model by using OVR(one vs rest) scheme and regularized it by using the L2 method with ‘ibfgs’ solver.
- Logistic regression is imported from sklearn.linear_model and then train data is fit into the model.

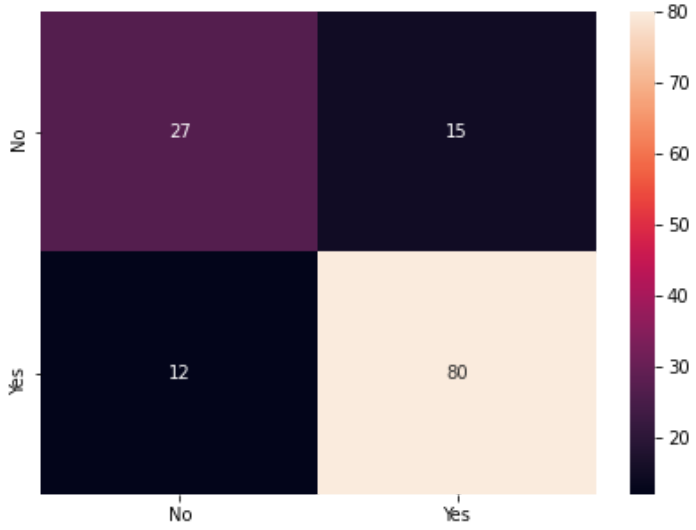
	Train set	Test set
Accuracy	81%	80%
AUC Score	83%	81%

Confusion matrix-

Train set

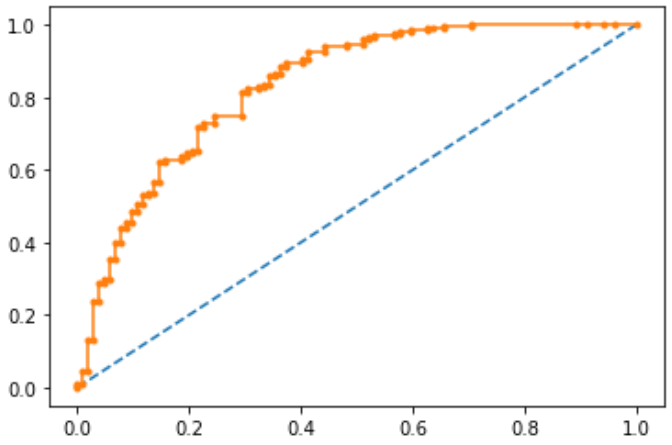


Test set

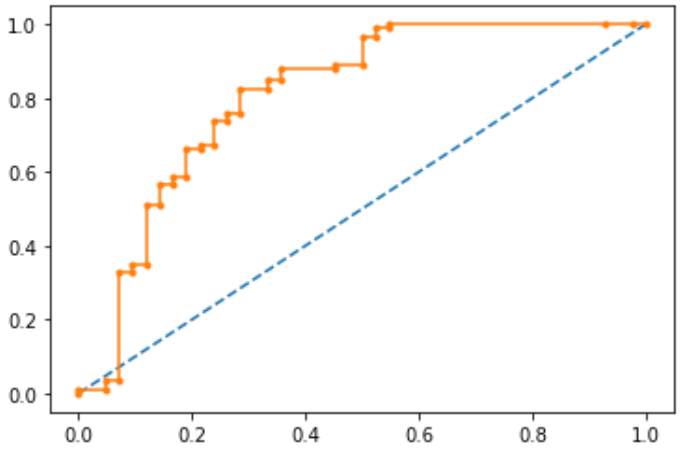


ROC-AUC Curve-

Train set



Test set



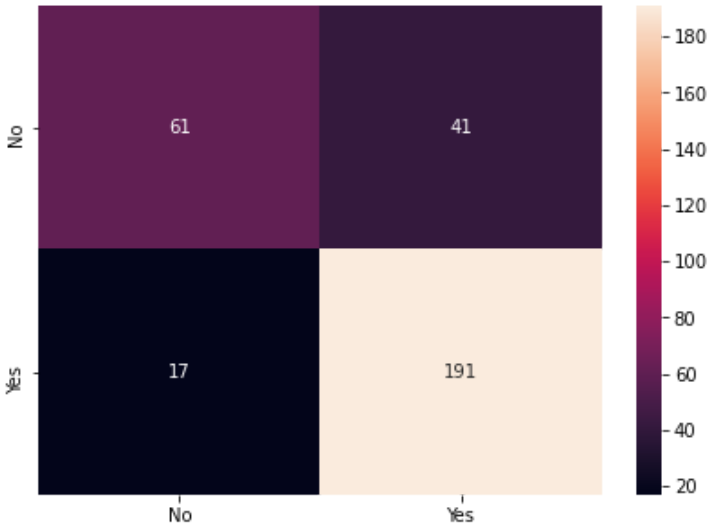
2) Linear Discriminant Analysis- LDA model uses Bayes' Theorem to estimate probabilities. They make predictions upon the probability that a new input dataset belongs to each class. The class which has the highest probability is considered as the output class and then the LDA makes a prediction

- LDA model is imported from sklearn library and then training data is fitted into the model for training the model.

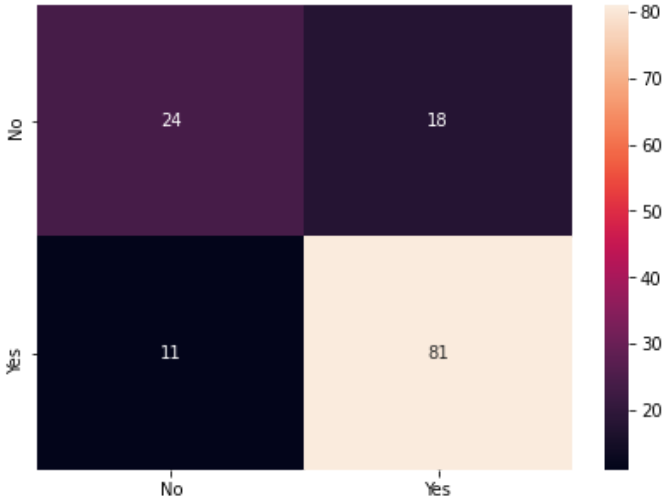
	Train set	Test set
Accuracy	81%	78%
AUC Score	83%	80%

Confusion matrix-

Train set

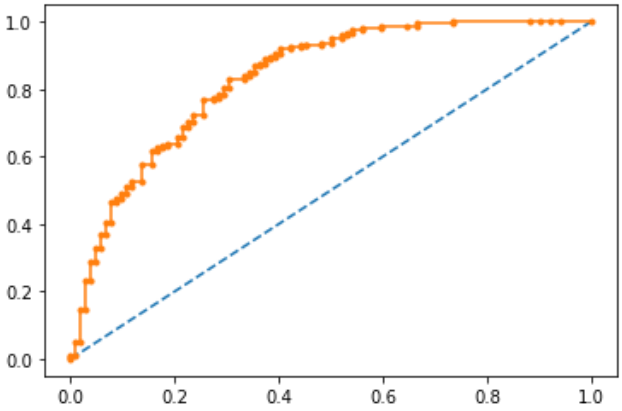


Test set

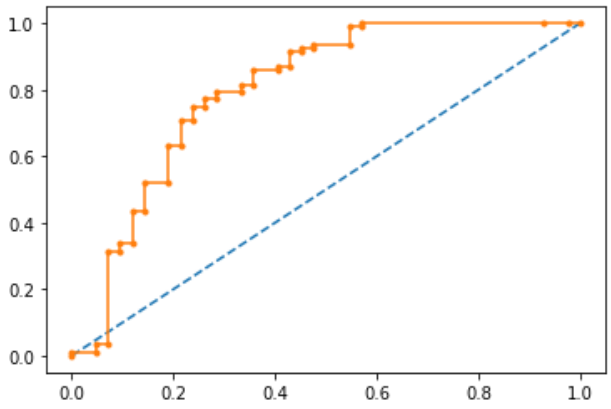


ROC- AUC Curve-

Train set



Test set



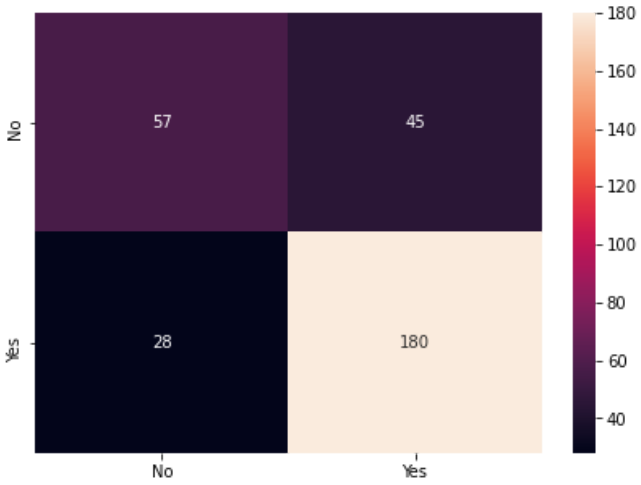
3)Naive Bayes model- It is a probabilistic model based on Bayes theorem. It is called naive due to the assumption the features in the data set are mutually independent.It estimates conditional probability which is the probability that something will happen, given that something else has already occurred.

- GaussianNB is imported from the sklearn library and then train set is fitted into the model and then tested on test set.

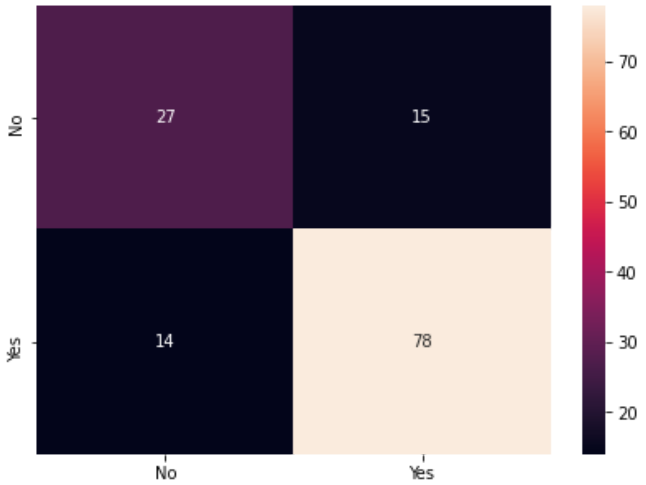
	Train set	Test set
Accuracy	76%	78%
AUC Score	79%	78.7%

Confusion matrix-

Train set

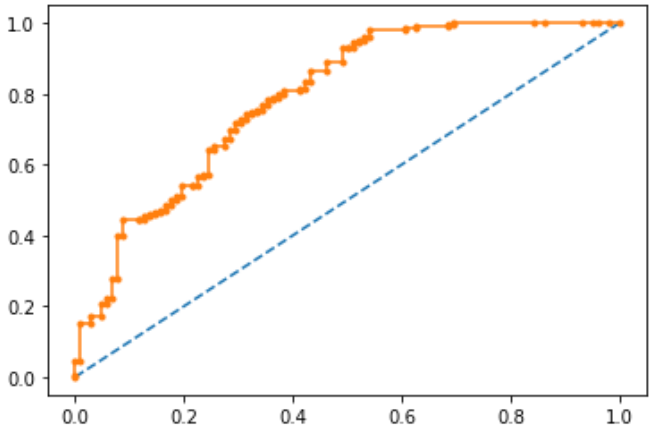


Test set

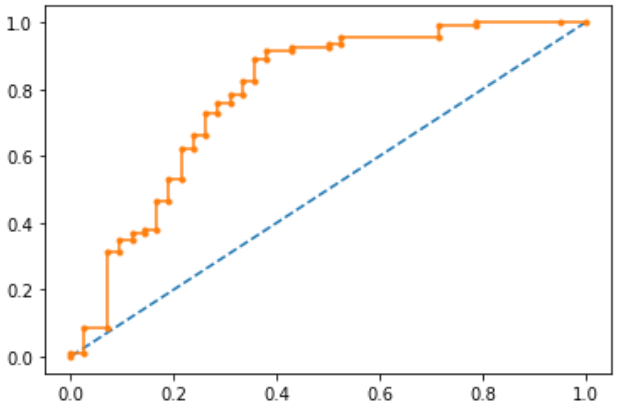


ROC-AUC Curve-

Train set



Test set



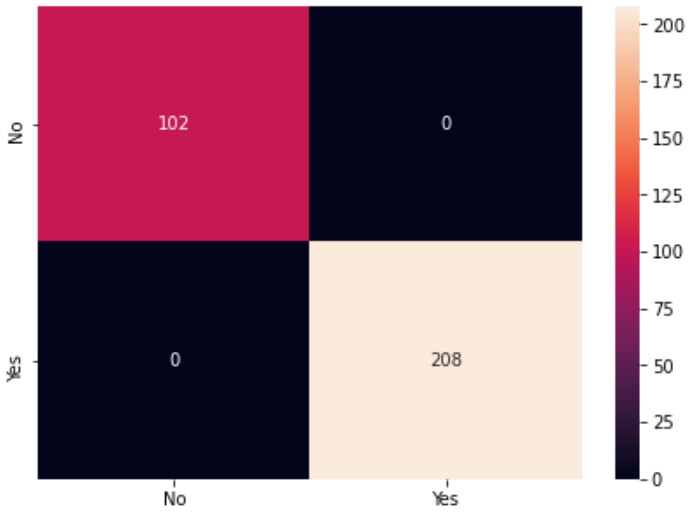
4)KNN model- The KNN algorithm uses the approach to find the nearest neighbor using distance between the query point and all other points which is known as Brute force. Most commonly distance is measured by using Euclidean distances.

- K Neighbor Classifier is imported from sklearn.neighbors library and the model is made by using 5 nearest neighbors . Train set is fitted into the model and the model is tested on the test set.

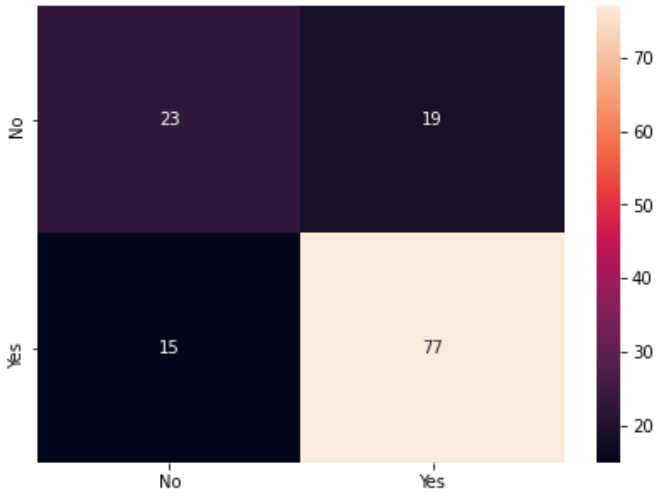
	Train set	Test set
Accuracy	100%	75%
AUC Score	100%	80.4%

Confusion matrix-

Train set

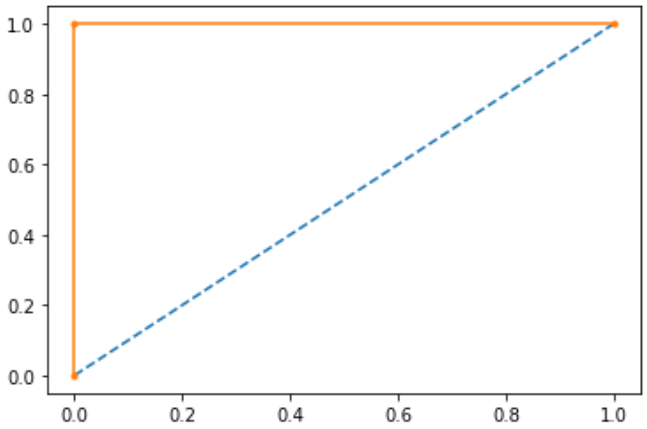


Test set

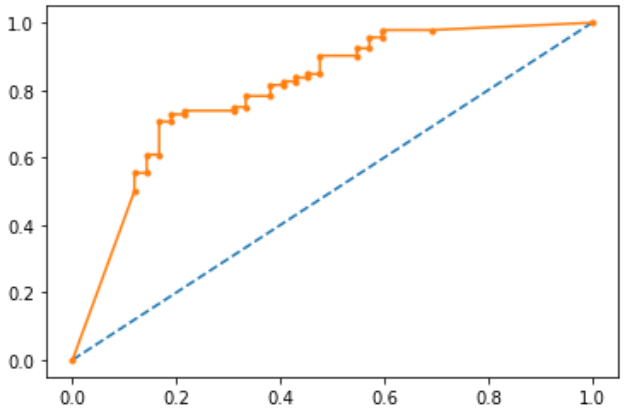


ROC-AUC Curve-

Train set



Test set

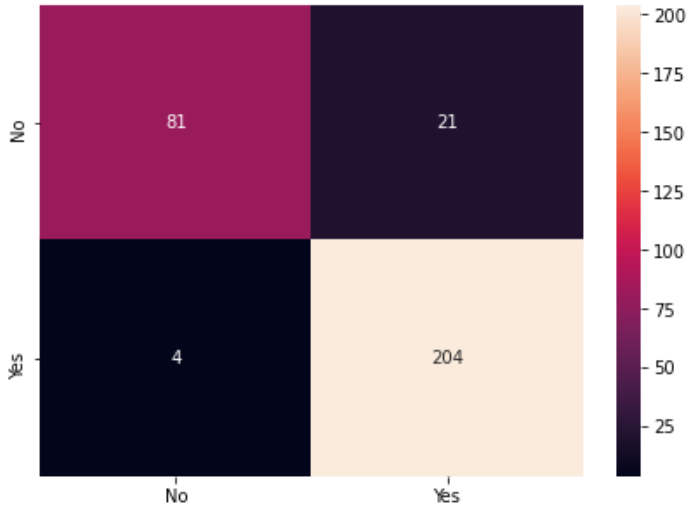


5) Gradient Boosting-Boosting trains a large number of "weak" learners in sequence. A weak learner is a simple model that is only slightly better than random.
-Miss-classified data weights are increased for training the next model. So training has to be done in sequence.
-Boosting then combines all the weak learners into a single strong learner.
-Each learner is fit on a modified version of original data (original data is replaced with the x values and residuals from previous learner
-By fitting new models to the residuals, the overall learner gradually improves in areas where residuals are initially high.

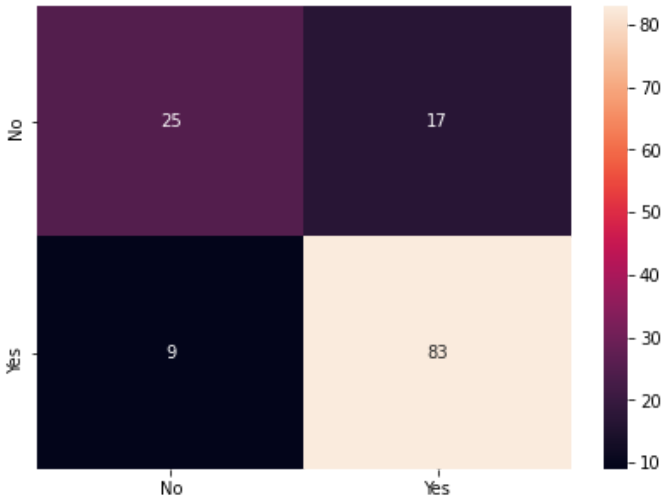
	Train set	Test set
Accuracy	92%	81%
AUC Score	98%	81%

Confusion matrix-

Train set

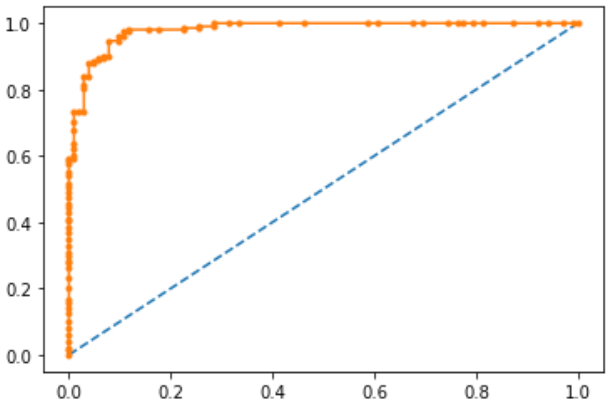


Test set

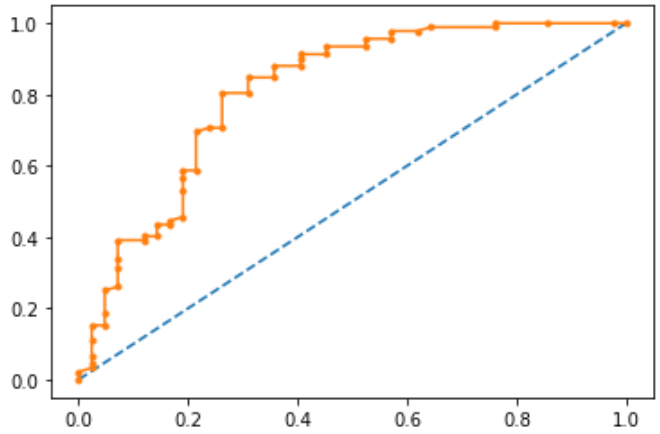


ROC-AUC Curve-

Train set



Test set



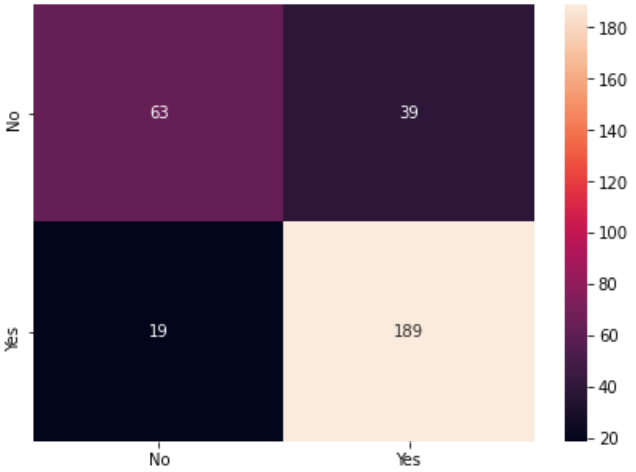
6)CART Model- It is a predictive model which explains how the outcome of a target variable can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable.

- cart model has been created by using DecisionTreeClassifier from sklearn by using 'Gini' as a criterion. Decision tree is created by using the Gender variable as a root node.
  - As the tree is overgrown pruning is required to prevent overfitting of the dataset.
  -pruning is done by using gridsearch.Prunning is done by using following best gridsearch parameters-
    Max_depth- 8
    Min_samples_leaf- 25
    Min_samples_split- 40

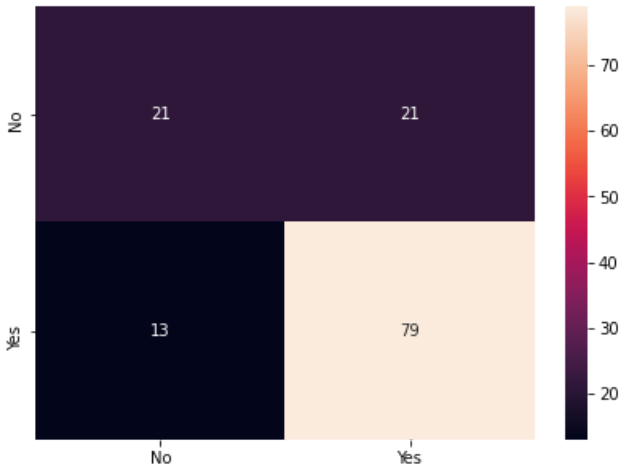
	Train set	Test set
Accuracy	81%	75%
AUC Score	87%	78%

Confusion matrix-

Train set

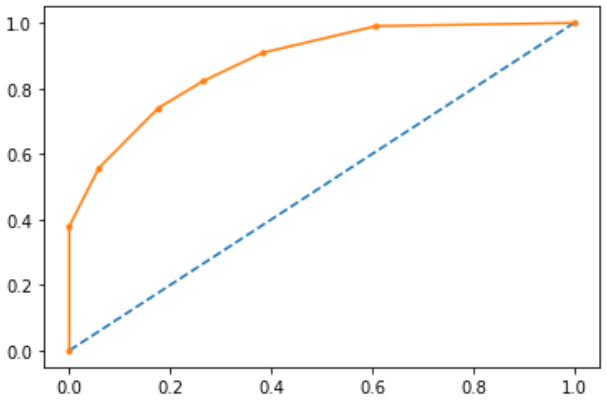


Test set

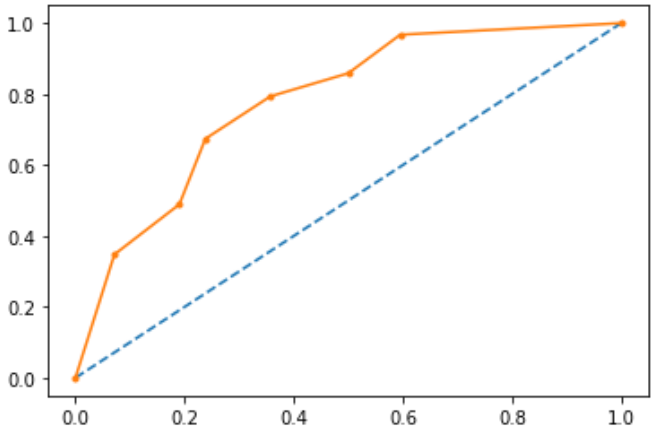


ROC-AUC Curve-

Train set



Test set



7) Random forest model- A random forest is a supervised machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees.

-Random forest model is formed by using RandomForestClassifier from sklearn.

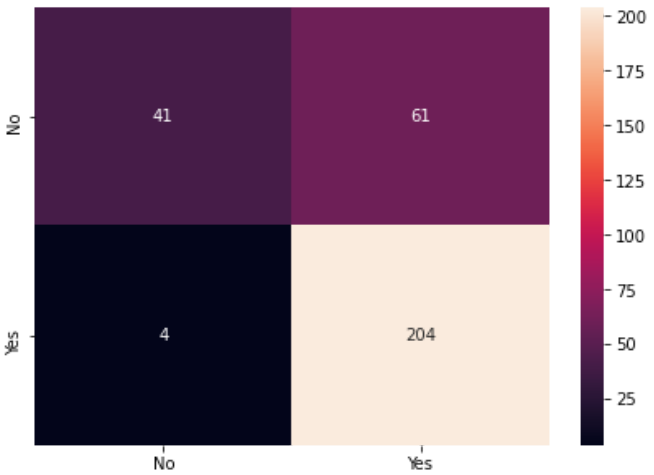
-Best gridsearch parameters for random forest model are-

Max_depth- 8
Max_features- 3
Min_samples_leaf- 30
Min_samples_split- 100
N_estimators- 101

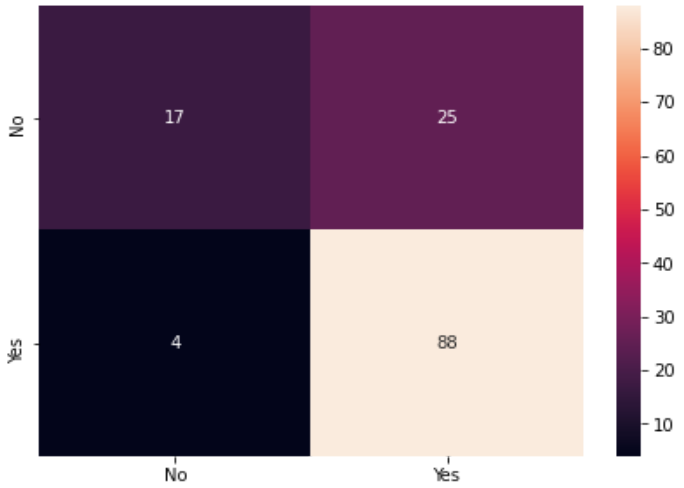
	Train set	Test set
Accuracy	79%	78%
AUC Score	85%	79%

Confusion matrix-

Train set

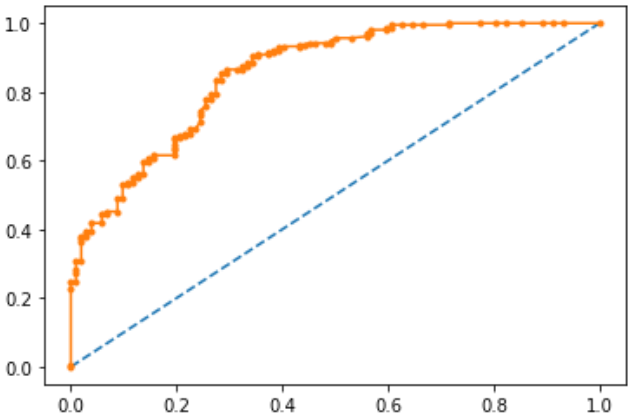


Test set

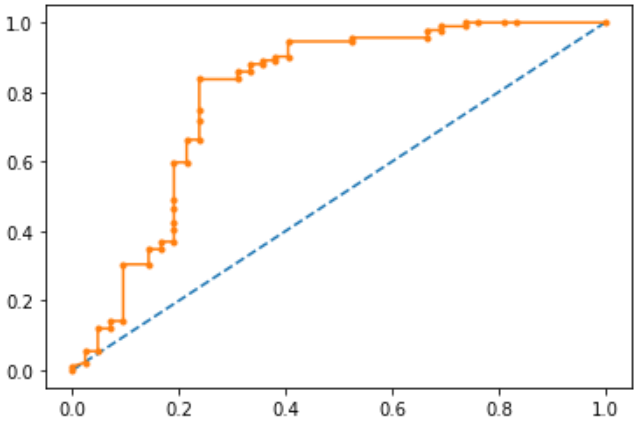


ROC-AUC Curve-

Train set



Test set



Question 1.4-

Which model performs the best?

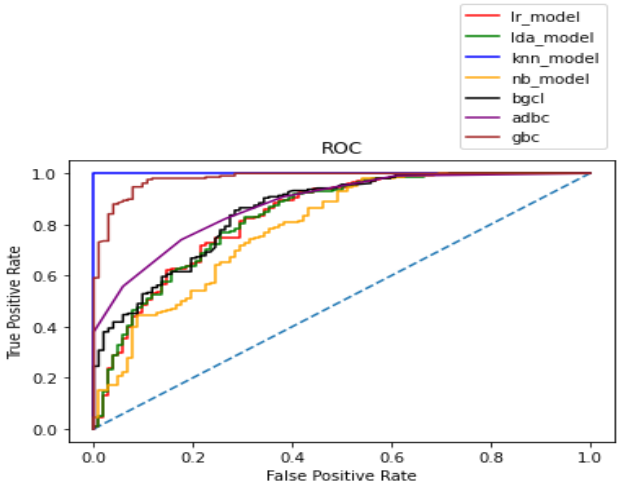
Solution-

- A summary of Accuracy and AUC Score of different models on train and test set-

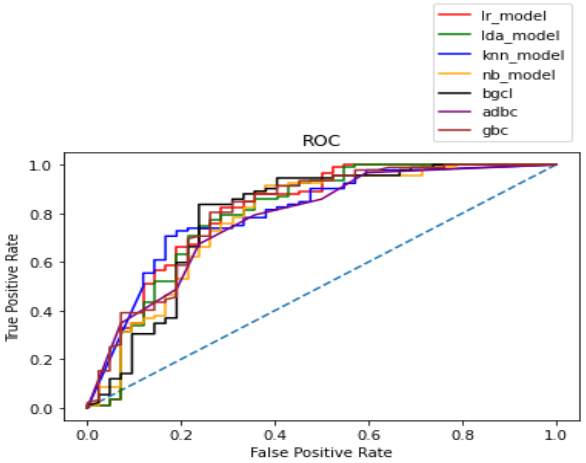
models	Accuracy		AUC Score	
	Train set	Test set	Train set	Test set
Logistic regression	81%	80%	83%	81%
Linear discriminant analysis	81%	78%	83%	80%
Naive bayes	76%	78%	79%	78%
KNN	100%	75%	100%	80%
Gradient Boosting	92%	81%	98%	81%
Decision tree	81%	75%	87%	78%
Random forest	80%	78%	86%	77%

- A summary of ROC-AUC curve of train and test set-

Train set



Test set



Inference- By looking at Accuracy , AUC Score and ROC-AUC Curve we can see that Gradient boosting model is working best on test data.

Question 1.5-

What are your business insights?

Solution- As Gradient boosting model is working best on test data, we would recommend them to use this model for future reference.

- As maximum staff are using public transport we recommend the company to use large vehicles like bus or tempo for transport of employees because it will be more efficient than small vehicles like cars.

Question 2- Text mining

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

Question 2.1-

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

Solution-

Top 5 Head of data set-

	deal	description
0	False	Bluetooth device implant for your ear.
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...

Dependent variable- Deal
Independent variable- Description

-Basic summary of data set-

```
RangeIndex: 495 entries, 0 to 494
Data columns (total 2 columns):
#   Column       Non-Null Count    Dtype
---  ------          --------------         -----
 0   deal            495 non-null       bool
 1   description  495 non-null    object
    dtypes: bool(1), object(1)
    memory usage: 4.5+ KB
```

Inference- There are 495 rows ranging from 0 to 494
- There are 2 columns.Deal is boolean type and description is object type
- There are no null values.

Question 2.2-

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

Solution-

- From nltk word_tokenization is imported.
- Made a function by name of tokenization_w which takes the data frame and made 2 corpora-

1) True_df_corpora- companies got the deal.
2) False_df_corpora- companies didn't get the deal.

Question 2.3-

The following exercise is to be done for both the corpora:
a) Find the number of characters for both the corpuses.
b) Remove Stop Words from the corpora. (Words like ‘also’, ‘made’, ‘makes’, ‘like’, ‘this’, ‘even’ and ‘company’ are to be removed)
c) What were the top 3 most frequently occurring words in both corpuses (after removing stop words)?
d) Plot the Word Cloud for both the corpora.

Solution-

2.3.a) True_df_corpora- 11801 characters are there.
False_df_corpora- 8721 characters are there.

2.3.b) From nltk.corpus stopwords are imported and all stop words and punctuations are removed and both corpus are saved by name of true_df_without_sw and false_df_without_sw.
sw(stopwords)

2.3.c) Top 3 most frequently occurring words after removing stop words are-

i) True_df_without_sw-
    Also - 42 times
    Makes - 32 times
    Made - 32 times
ii) False_df_without_sw-
    Made - 41 times
    Company - 25 times
    Also - 20 times

2.3d) Wordcloud is imported and a show_wordcloud function is made which will take corpora as an argument and plot the wordcloud.

True_df_without_wf-

False_df_without_sw

Question 2.4-

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

Solution-

i) By looking at wordcloud of true_df_without_sw we can infer that entrepreneurs with ideas related to children, kids, products,giving something free have secured the deal.
ii) By looking at wordcloud of false_df_without_sw we can infer that entrepreneurs with ideas related to device,system,and customers have not secured the deal.

Question 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-

_Yes, by looking at wordcloud of false_df_without_sw we can infer that entrepreneurs who introduced devices are less likely to secure a deal.