MACHINE LEARNING PROJECT REPORT

-

Pooja Gujja

Table of Contents

Problem Statement:

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.

Following is a guideline for developing a solution:

Data Ingestion:
1. Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.
2. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

Data Preparation:
1. Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).

Modeling:
1. Apply Logistic Regression .
2. Apply KNN Model. Interpret the results.
3. Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.
4. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

Inference:
1. Based on these predictions, what are the insights?

Solution:

Checking the first 5 rows of the data:

To start with the analysis, let's look at the sample data and perform basic checks.

	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

Checking the dimensions of data:

```
<class 'pandas.core.frame.DataFrame'>
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
```

Inferences:

1. The dataset has a total of eight independent variables and one target variable
2. Shape (dimension) of the Dataset is (444, 9)
3. There are NO NULL values present in the dataset
4. There are NO Duplicate values present in the dataset
5. There are a total of 9 variables and 444 records

Data types of the variables:

```
Age             int64
Gender         object
Engineer        int64
MBA             int64
Work Exp        int64
Salary        float64
Distance      float64
license         int64
Transport      object
dtype: object
```

Exploratory Data Analysis:

Univariate Analysis:

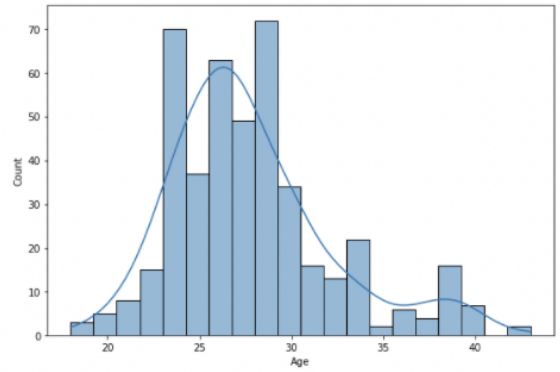
To perform Univariate analysis on continuous variables, let us start with looking at the summary statistics of the dataset.

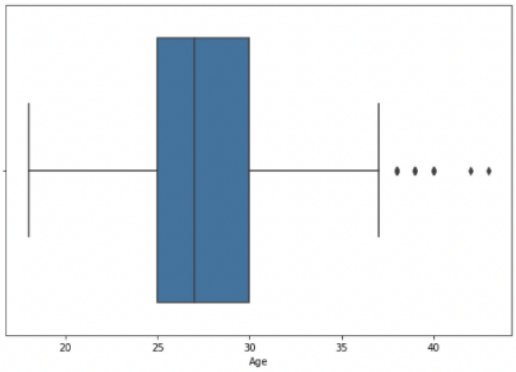
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	444.0	NaN	NaN	NaN	27.747748	4.41671	18.0	25.0	27.0	30.0	43.0
Gender	444	2	Male	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Engineer	444.0	NaN	NaN	NaN	0.754505	0.430866	0.0	1.0	1.0	1.0	1.0
MBA	444.0	NaN	NaN	NaN	0.252252	0.434795	0.0	0.0	0.0	1.0	1.0
Work Exp	444.0	NaN	NaN	NaN	6.29955	5.112098	0.0	3.0	5.0	8.0	24.0
Salary	444.0	NaN	NaN	NaN	16.238739	10.453851	6.5	9.8	13.6	15.725	57.0
Distance	444.0	NaN	NaN	NaN	11.323198	3.606149	3.2	8.8	11.0	13.425	23.4
license	444.0	NaN	NaN	NaN	0.234234	0.423997	0.0	0.0	0.0	0.0	1.0
Transport	444	2	Public Transport	300	NaN	NaN	NaN	NaN	NaN	NaN	NaN

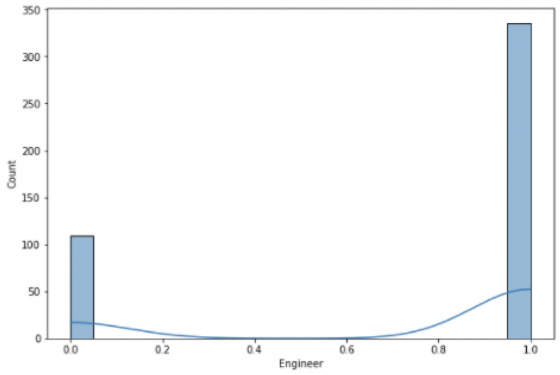
From above table:

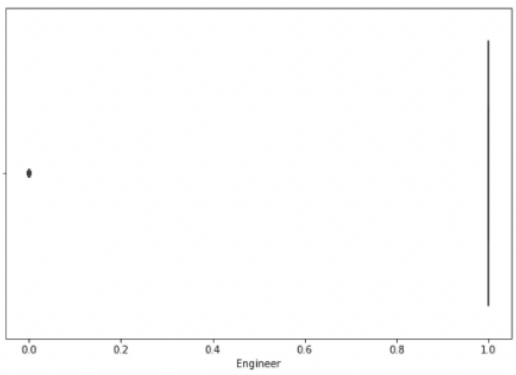
- We can see that the Public Transport has more frequency in the summary statistics of the dataset
- Mean is high for Age compared to other variables

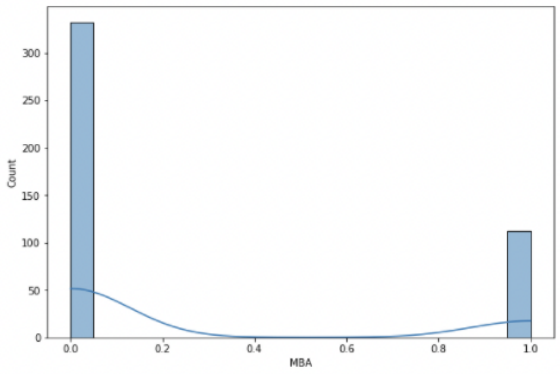
Plotting Box plots and hist plots:

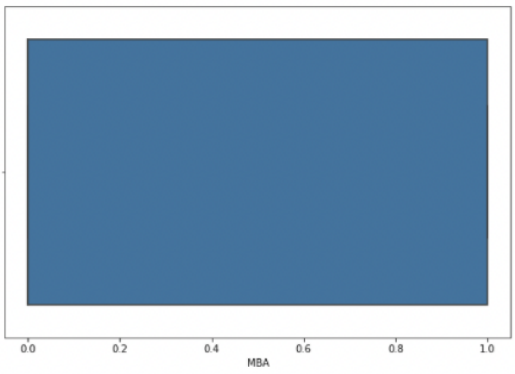


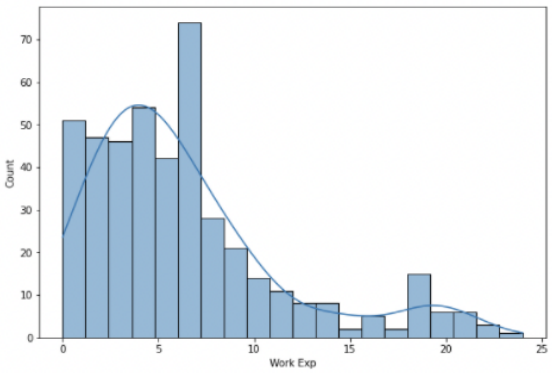


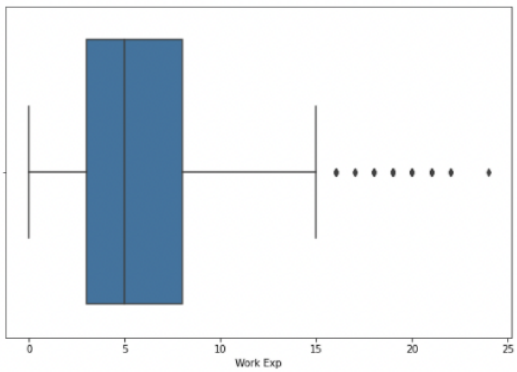


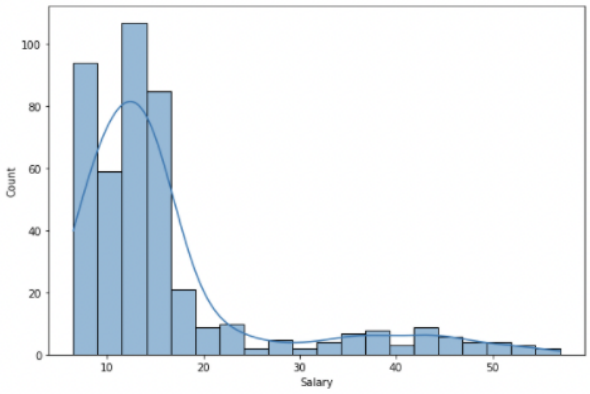


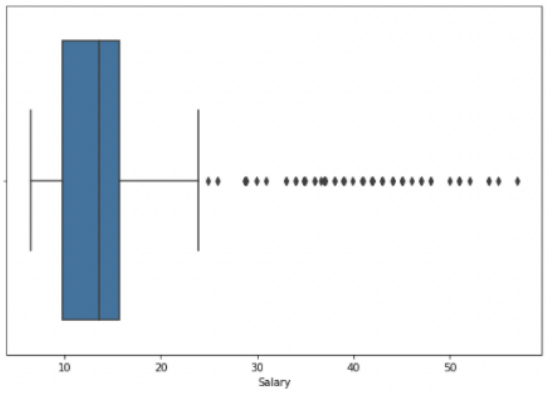


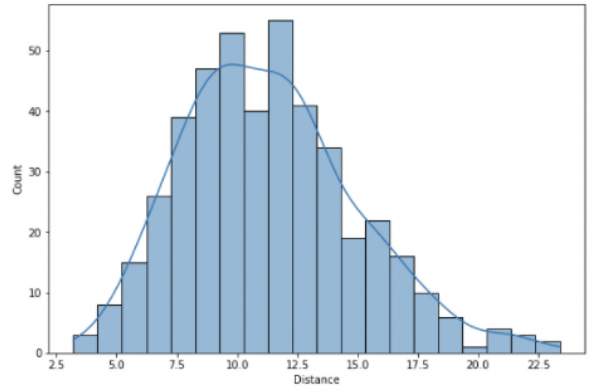


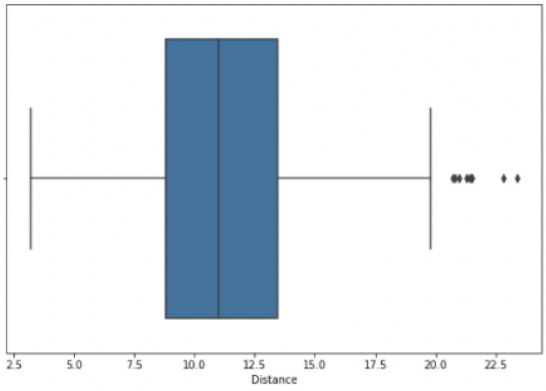


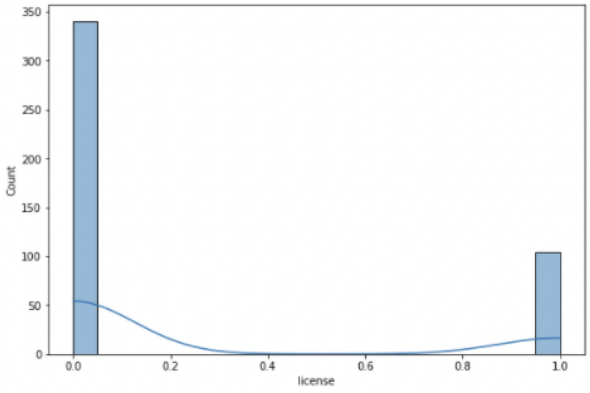


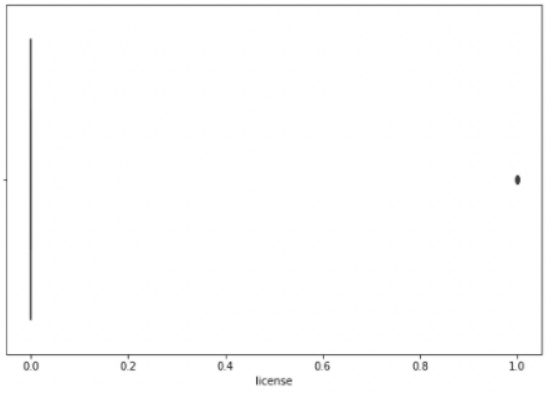

Count
Salary


Salary


Count
Distance


Distance


Count
license


license

Percentage of Outliers:

	Outlier%
Age	5.630631
Distance	2.027027
Engineer	24.549550
Gender	0.000000
MBA	0.000000
Salary	13.288288
Transport	0.000000
Work Exp	8.558559
license	23.423423

Kurtosis and Skewness in the Dataset:

	Skewness
Age	0.955276
Engineer	-1.186708
MBA	1.144763
Work Exp	1.352840
Salary	2.044533
Distance	0.539851
license	1.259293

	Kurtosis
Age	0.938871
Engineer	-0.594422
MBA	-0.692657
Work Exp	1.478573
Salary	3.479377
Distance	0.191465
license	-0.416075

Inferences on Univariate Analysis:

1. Engineer and License columns have outlier values with a Percentage of 24.54 and 23.42 which is very high above 20%
2. However, if we look at the dataset, the two columns are binary with only 0's and 1's. Hence, it is showing as outliers as mean is around either 0 or 1. Therefore it is not required to treat these 2 columns for outliers
3. We are not treating outliers as there might be an exceptional case where these values are possible for the business in certain cases.
4. As per the dataset, none of the values are so unrealistic. Hence, I am considering not to treat outliers for my analysis

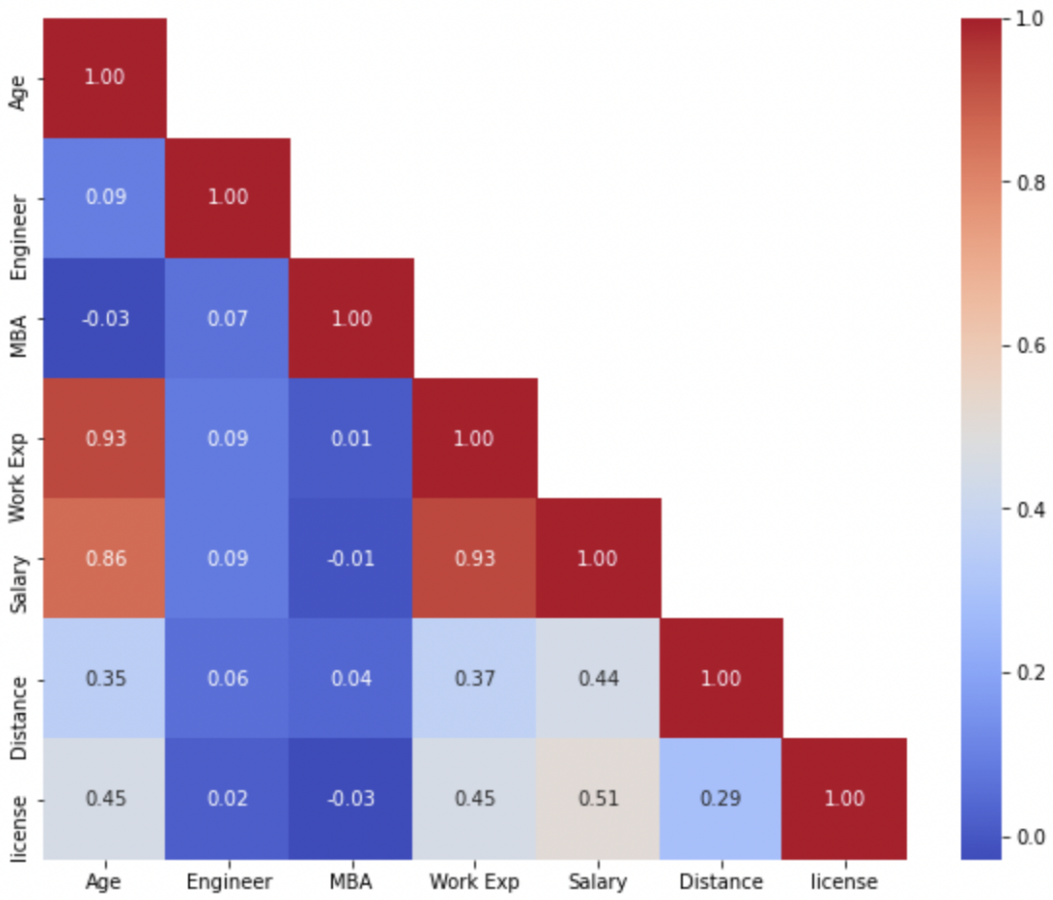
5. From the Box plots, we can conclude that there are outliers present in Age, Workexp, Salary, Distance variables.
6. Variable Age and Distance seems to be normally distributed and variable Salary is right skewed.
7. A standard normal distribution has kurtosis of 3. Kurtosis values for all variables are not very high, hence near to normal distribution
8. There is positive kurtosis seen in one variable Salary which is > 3 which means more of the values are located in the tails of the distribution rather than around the mean.

Multivariate analysis:

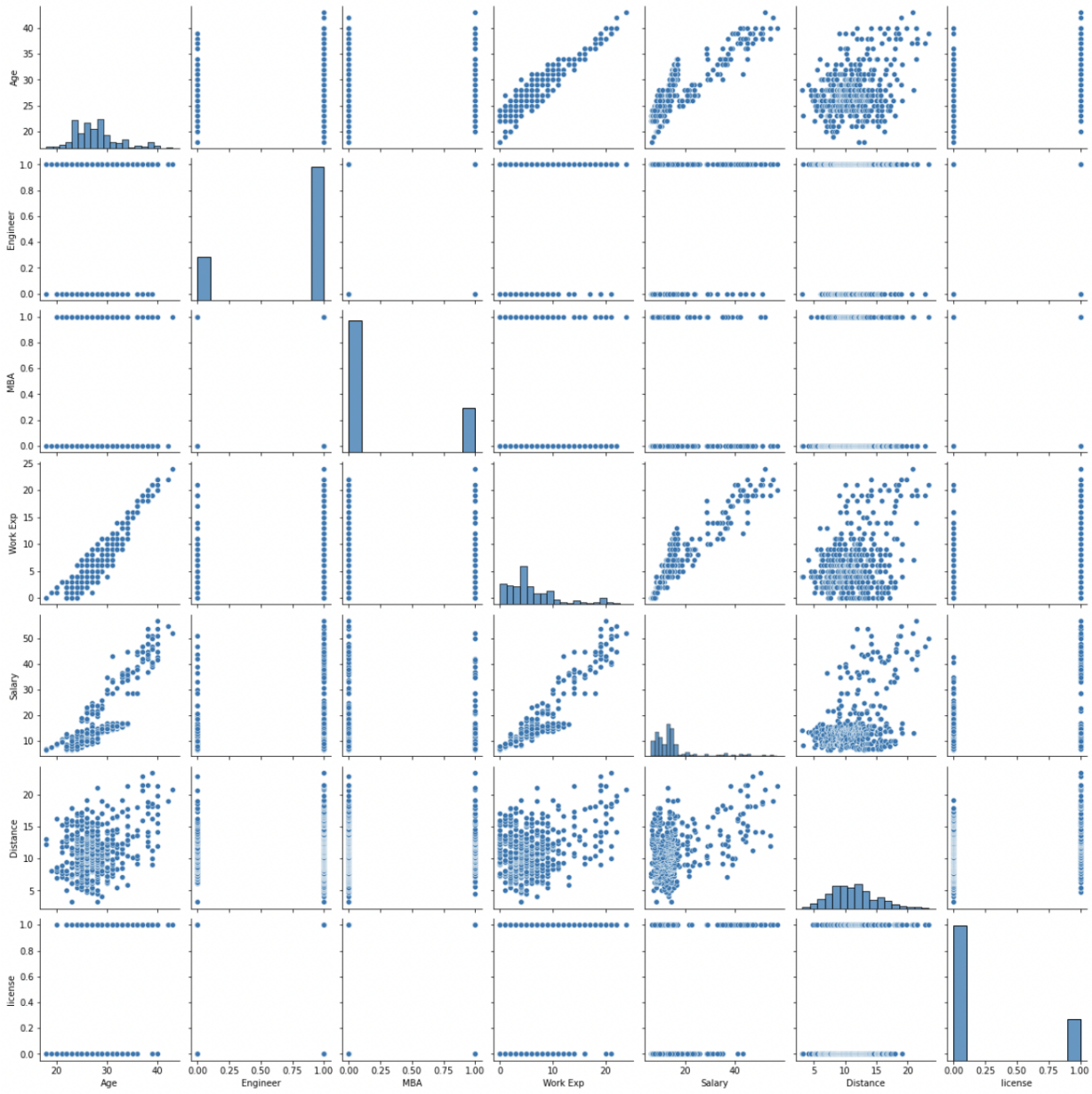
Analyzing the relationship among continuous variables by using Pair plot and Correlation Heatmap.

Checking for multicollinearity:

Heatmap:



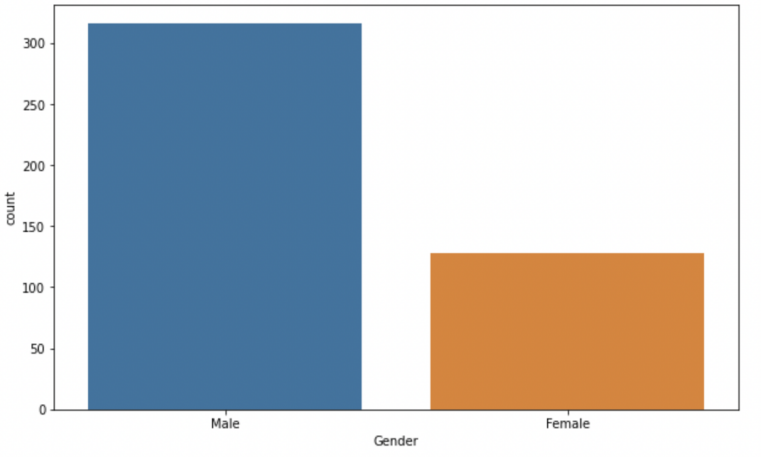
Pair Plot:

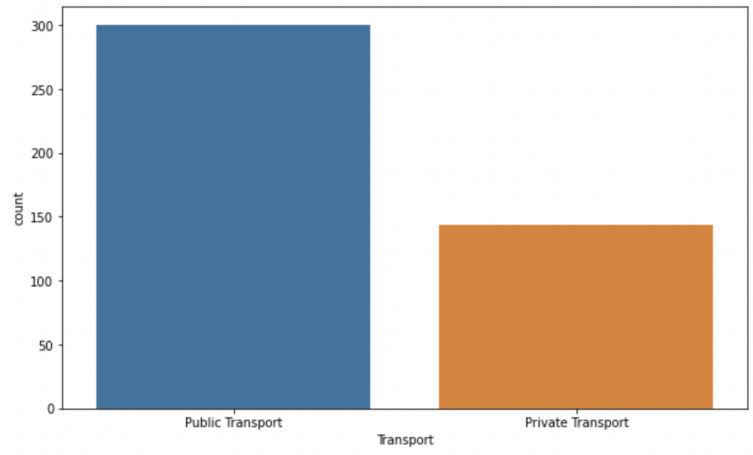


Inferences on Multivariate Analysis:

1. Workexp and Salary variables are highly positively correlated with Age variable
2. Workexp is highly correlated with Salary
3. There is no high Negative correlation for any of the variables
4. From pairplot, we see that as the Age increases the workexp is also increasing showing a positive relationship
5. Few of the variables have no correlation with each other, i.e., datapoints are spread wide across(ex: Workexp and Distance, Age and Distance). There might be small amount of correlation but there is no particular trend
6. Few of the variables are binary with 0's and 1's. Hence the distribution has a particular pattern of all 0's and 1's

EDA for Categorical Variables:





Inference:

- From the above plot, we can state that there are more number of male than female
- People who opt for public transport are more in number than private transport

Convert all objects to categorical codes:

```
feature: Gender
['Male', 'Female']
Categories (2, object): ['Female', 'Male']
[1 0]


feature: Transport
['Public Transport', 'Private Transport']
Categories (2, object): ['Private Transport', 'Public Transport']
[1 0]
```

Inferences:

- We see that Gender and Transport are object type. Hence, we are converting all objects to categorical codes
- Female : 1
- Male : 0
- Private Transport : 1
- Public Transport : 0
- Datatypes are automatically changed to int for Gender and Transport

Inferences:

- Extracting target column into separate vectors for training and test set.
- We have extracted target column into y variable and all the other variables into X.
- As KNN is a distance based model, it is suggested to scale the data.
- This is suggested because the distance calculation done in KNN uses feature values. When one of the feature values are large than other, that feature will dominate the distance, hence the outcome of the KNN.
- We are scaling the data after splitting into train and test.
- Scaling is performed only on the x variables leaving the target column.

Building the Logistic Regression Models - Descriptive approach:

Model 1: Using all the variables

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	435
Method:	MLE	**Df Model:**	8
Date:	Fri, 04 Mar 2022	**Pseudo R-squ.:**	0.3049
Time:	16:42:58	**Log-Likelihood:**	-194.45
converged:	True	**LL-Null:**	-279.76
Covariance Type:	nonrobust	**LLR p-value:**	9.551e-33

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.8340	1.788	-0.466	0.641	-4.339	2.671
Age	0.2083	0.077	2.706	0.007	0.057	0.359
Gender	1.2810	0.288	4.441	0.000	0.716	1.846
Engineer	-0.1543	0.296	-0.521	0.603	-0.735	0.427
MBA	0.5601	0.314	1.782	0.075	-0.056	1.176
Workexp	-0.1005	0.100	-1.001	0.317	-0.297	0.096
Salary	-0.0805	0.040	-2.003	0.045	-0.159	-0.002
Distance	-0.2248	0.043	-5.290	0.000	-0.308	-0.142
license	-2.0463	0.334	-6.135	0.000	-2.700	-1.393

VIF Values:

Age VIF = 7.89
Gender VIF = 1.07
Engineer VIF = 1.02
MBA VIF = 1.03
Workexp VIF = 15.74
Salary VIF = 8.87
Distance VIF = 1.28
license VIF = 1.45

Inferences from Model 1 of LR:

- Adjusted Pseudo R^2 value is 0.27.
- 27% of variance is explained by including all the variables
- Coefficient of license is greater than all other variables. License variable explains a high variance in the target variable.
- After calculating the vif values to check for multicollinearity, we can see that Workexp has high multicollinearity of 15.74(considering threshold as >5)
- In Model 2, Dropping variable 'workexp' as it has high vif

Model 2: Dropping variable ‘workexp’

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	436
Method:	MLE	**Df Model:**	7
Date:	Fri, 04 Mar 2022	**Pseudo R-squ.:**	0.3032
Time:	16:43:08	**Log-Likelihood:**	-194.95
converged:	True	**LL-Null:**	-279.76
Covariance Type:	nonrobust	**LLR p-value:**	3.014e-33

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.3889	1.317	0.295	0.768	-2.193	2.971
Age	0.1561	0.057	2.749	0.006	0.045	0.267
Gender	1.2805	0.287	4.454	0.000	0.717	1.844
Engineer	-0.1459	0.295	-0.494	0.622	-0.725	0.433
MBA	0.5265	0.311	1.690	0.091	-0.084	1.137
Salary	-0.1107	0.028	-4.000	0.000	-0.165	-0.056
Distance	-0.2193	0.042	-5.222	0.000	-0.302	-0.137
license	-2.0088	0.330	-6.089	0.000	-2.655	-1.362

VIF Values:

Age VIF = 3.89
Gender VIF = 1.07
Engineer VIF = 1.02
MBA VIF = 1.02
Salary VIF = 4.46
Distance VIF = 1.26
license VIF = 1.43

Inferences from Model 2 of LR:

- Adjusted Pseudo R^2 value is 0.27.
- 27% of variance is explained after dropping workexp variable
- It is clear that workexp has a high multicollinearity with other variables. After removing workexp, same amount of variance is explained by all the other variables
- Hence, workexp has no significant use in the prediction of target variable
- From the vif values, we can see that all the variables meet the threshold of 5.
- We can see that p-values are high for Engineer and MBA of 0.6 and 0.09 respectively
- As the p-value is greater than 0.05 for Engineer, we are dropping the variable in model 3

Model 3: Dropping Engineer variable

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	437
Method:	MLE	**Df Model:**	6
Date:	Fri, 04 Mar 2022	**Pseudo R-squ.:**	0.3027
Time:	16:43:16	**Log-Likelihood:**	-195.07
converged:	True	**LL-Null:**	-279.76
Covariance Type:	nonrobust	**LLR p-value:**	6.095e-34

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.3196	1.310	0.244	0.807	-2.248	2.887
Age	0.1547	0.057	2.730	0.006	0.044	0.266
Gender	1.2758	0.287	4.446	0.000	0.713	1.838
MBA	0.5115	0.309	1.653	0.098	-0.095	1.118
Salary	-0.1105	0.028	-4.001	0.000	-0.165	-0.056
Distance	-0.2192	0.042	-5.217	0.000	-0.302	-0.137
license	-2.0004	0.329	-6.077	0.000	-2.646	-1.355

VIF Values:

Age VIF = 3.89
Gender VIF = 1.07
MBA VIF = 1.02
Salary VIF = 4.46
Distance VIF = 1.26
license VIF = 1.43

Inferences from Model 3 of LR:

- Adjusted Pseudo R^2 value is 0.28.
- 28% of variance is explained after dropping engineer variable which is slightly higher than model 2
- p-value of MBA is 0.09
- As the p-value is greater than 0.05 for MBA, we are dropping the variable in model 4

Model 4: Dropping MBA variable

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	438
Method:	MLE	**Df Model:**	5
Date:	Fri, 04 Mar 2022	**Pseudo R-squ.:**	0.2977
Time:	16:43:22	**Log-Likelihood:**	-196.49
converged:	True	**LL-Null:**	-279.76
Covariance Type:	nonrobust	**LLR p-value:**	3.988e-34

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.5268	1.297	0.406	0.685	-2.016	3.069
Age	0.1478	0.056	2.636	0.008	0.038	0.258
Gender	1.3271	0.285	4.650	0.000	0.768	1.886
Salary	-0.1060	0.027	-3.917	0.000	-0.159	-0.053
Distance	-0.2190	0.042	-5.197	0.000	-0.302	-0.136
license	-2.0216	0.329	-6.138	0.000	-2.667	-1.376

Model Evaluation:

	model_name	model_perf	variables
0	model 1	0.276350	All variables
1	model 2	0.278139	Drop workexp
2	model 3	0.281274	Drop Engineer
3	model 4	0.279781	Drop MBA

Inferences from Model 4 of LR:

- Adjusted Pseudo R^2 value is 0.27
- 27.9% of variance is explained after dropping MBA variable which is almost as same as model 3
- Approximately 28% is explained by less number of variables (i.e., 5 variables) in Model 4
- Comparing all the models, we can see that model 4 is performing better
- Model 4 is better compared to all other models because 28% of variance is explained by 5 variables
- In all other models, similar variance is explained compared to model 4. However, number of variables used are more.

Predictive Approach:

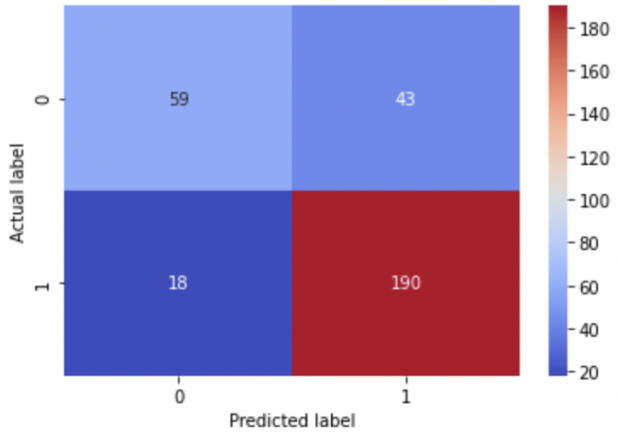
Evaluating models based on classification metrics of choice - Model 4

- Accuracy for model 4 on train data: 0.8032258064516129
- Accuracy for model 4 on test data: 0.8283582089552238

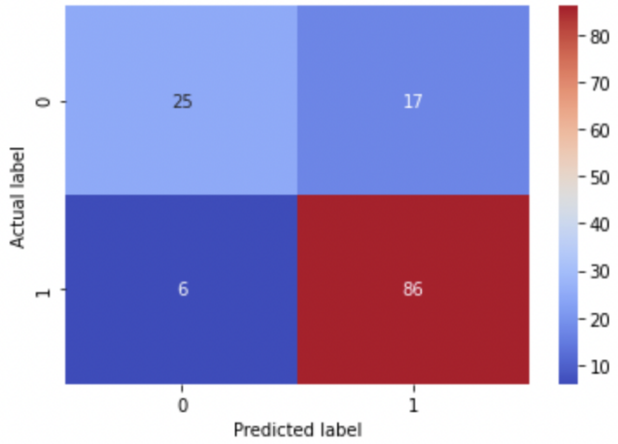
Confusion matrix and classification report:

Model 4:

Confusion matrix for train data:



Confusion matrix for test data:



```
classification report for train data:
              precision    recall  f1-score   support

           0       0.77      0.58      0.66       102
           1       0.82      0.91      0.86       208

    accuracy                           0.80       310
   macro avg       0.79      0.75      0.76       310
weighted avg       0.80      0.80      0.80       310
```

Inference on train data for LR model 4 :

For predicting people who opt for Public Transport:

- Precision (77%) – 77% of prediction of people who will opt for public transport are correct
- Recall (58%) – 58% of employees who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport:

- Precision (82%) – 82% of prediction of people who will opt for private transport are correct
- Recall (91%) – 91% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 80% of total predictions are correct

```
classification report for test data:
              precision    recall  f1-score   support

           0       0.81      0.60      0.68        42
           1       0.83      0.93      0.88        92

    accuracy                           0.83       134
   macro avg       0.82      0.77      0.78       134
weighted avg       0.83      0.83      0.82       134
```

Inference on test data for LR model 4:

For predicting people who opt for Public Transport:

- Precision (81%) – 81% of prediction of people who will opt for public transport are correct
- Recall (60%) – 60% of employees who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport:

- Precision (83%) – 83% of prediction of people who will opt for private transport are correct
- Recall (93%) – 93% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 83% of total predictions are correct

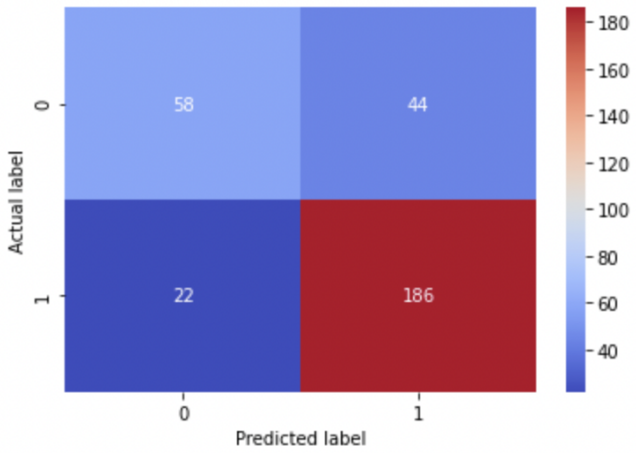
Evaluating models based on classification metrics of choice - Model 1

- Accuracy including all variables for train data: 0.7870967741935484
- Accuracy including all variables for test data: 0.8059701492537313

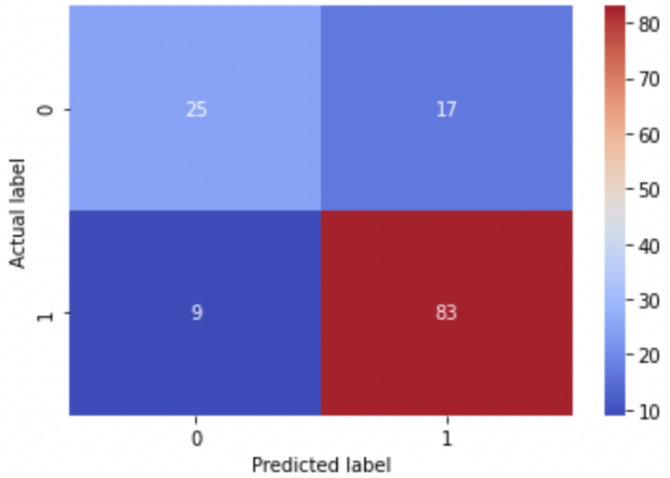
Confusion matrix and classification report:

Model 1:

Confusion matrix for train data:



Confusion matrix for test data:



```
Classification report Train set :
              precision    recall  f1-score   support

           0       0.72      0.57      0.64       102
           1       0.81      0.89      0.85       208

    accuracy                           0.79       310
   macro avg       0.77      0.73      0.74       310
weighted avg       0.78      0.79      0.78       310
```

Inference on train data for LR model 1 :

For predicting people who opt for Public Transport:

- Precision (72%) – 72% of prediction of people who will opt for public transport are correct
- Recall (57%) – 57% of employees who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport:

- Precision (81%) – 81% of prediction of people who will opt for private transport are correct
- Recall (89%) – 89% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 79% of total predictions are correct

```
Classification report Test set :
              precision    recall  f1-score   support

           0       0.74      0.60      0.66        42
           1       0.83      0.90      0.86        92

    accuracy                           0.81       134
   macro avg       0.78      0.75      0.76       134
weighted avg       0.80      0.81      0.80       134
```

Inference on test data for LR model 1 :

For predicting people who opt for Public Transport:

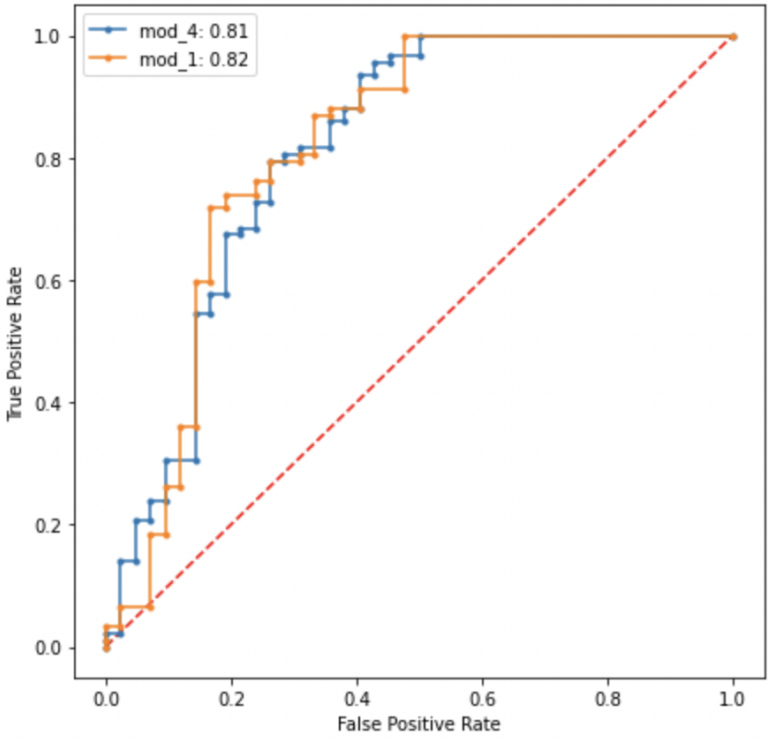
- Precision (74%) – 74% of prediction of people who will opt for public transport are correct
- Recall (60%) – 60% of employees who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport:

- Precision (83%) – 83% of prediction of people who will opt for private transport are correct
- Recall (90%) – 90% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 83% of total predictions are correct

ROC AUC Curve:



Inferences:

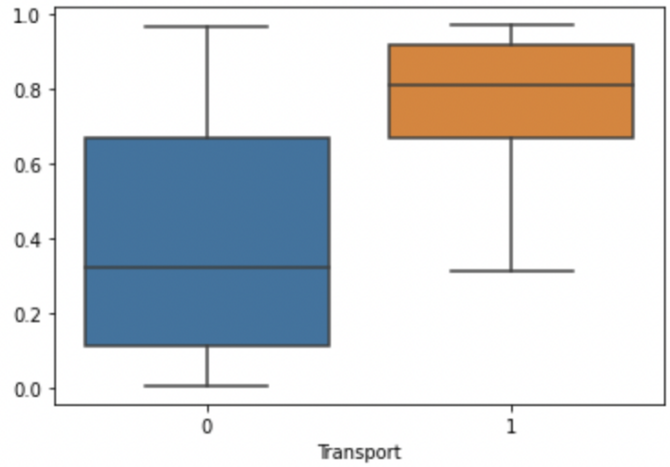
- The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 –FPR).
- Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- To compare different classifiers, it can be useful to summarize the performance of each classifier into a single measure. One common approach is to calculate the area under the ROC curve. It is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.
- A classifier with high AUC can occasionally score worse in a specific region than another classifier with lower AUC. However, in practice, the AUC performs well as a general measure of predictive accuracy.
- In the above AUC for both the models is high and nearly same. There is not much difference in the area under the curve for model 4.
- Though model 1 has a bit high AUC, we can choose to consider model 4 also as 0.81 is a good number.

Hyper Parameter Tuning for Logistic Regression :

Tuned Hyperparameters : {'C': 429981696.0, 'penalty': 'l1', 'solver': 'liblinear'}

Accuracy: 0.8059701492537313

Let us see the spread of the probability of the people using public and private transport for test data:



- From the above boxplot, we need to decide on one such value of a cut-off which will give us the most reasonable descriptive power of the model. A value between 0.6 and 0.7 seems reasonable
- Probability of of people choosing public transport ranges untill 0.6

From Logistic Regression Model (Feature Analysis)

- Age and Gender variables which have a positive influence of person opting for private transportation
- As age increases, probability of a person choosing private transportation also increases.
- We can see that the coefficients of all the parameters are quite low which may not be the perfect implementation in production
- As all the values of all performance parameters is not high, company should look at considering some more features which can explain high variance in the target variable in order to improve the model performance.

KNN Model:

Train Accuracy is : 0.8419354838709677

Test Accuracy is : 0.7835820895522388

Train ROC-AUC score is : 0.9224877450980392

Test ROC-AUC score is : 0.7496118012422359

Confusion matrix for train set :
[[64 38]
[11 197]]

Confusion matrix for test set :
[[23 19]
[10 82]]

Classification report Train set :

	precision	recall	f1-score	support
0	0.85	0.63	0.72	102
1	0.84	0.95	0.89	208
accuracy			0.84	310
macro avg	0.85	0.79	0.81	310
weighted avg	0.84	0.84	0.83	310

Inference on train data for KNN:

For predicting people who opt for Public Transport(0):

- Precision (85%) – 85% of prediction of people who will opt for public transport are correct
- Recall (63%) – 63% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (84%) – 84% of prediction of people who will opt for private transport are correct
- Recall (95%) – 95% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 84% of total predictions are correct

```
Classification report Test set :
              precision    recall  f1-score   support

           0       0.70      0.55      0.61        42
           1       0.81      0.89      0.85        92

    accuracy                           0.78       134
   macro avg       0.75      0.72      0.73       134
weighted avg       0.78      0.78      0.78       134
```

Inference on test data for KNN:

For predicting people who opt for Public Transport(0):

- Precision (70%) – 70% of prediction of people who will opt for public transport are correct
- Recall (55%) – 55% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (81%) – 81% of prediction of people who will opt for private transport are correct
- Recall (89%) – 89% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 78% of total predictions are correct

Hyper Parameter Tuning for KNN Classifier:

Train Accuracy is : 0.8258064516129032

Test Accuracy is : 0.7910447761194029

Train ROC-AUC score is : 0.8801611990950227

Test ROC-AUC score is : 0.776268115942029

Confusion matrix for train set :
[[55 47]
[7 201]]

Confusion matrix for test set :
[[20 22]
[6 86]]

```
Classification report Train set :
              precision    recall  f1-score   support

           0       0.89      0.54      0.67       102
           1       0.81      0.97      0.88       208

    accuracy                           0.83       310
   macro avg       0.85      0.75      0.78       310
weighted avg       0.84      0.83      0.81       310
```

Inference on train data for KNN after tuning:

For predicting people who opt for Public Transport(0):

- Precision (89%) – 89% of prediction of people who will opt for public transport are correct
- Recall (54%) – 54% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (81%) – 81% of prediction of people who will opt for private transport are correct
- Recall (97%) – 97% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 83% of total predictions are correct

```
Classification report Test set :
              precision    recall  f1-score   support

           0       0.77      0.48      0.59        42
           1       0.80      0.93      0.86        92

    accuracy                           0.79       134
   macro avg       0.78      0.71      0.72       134
weighted avg       0.79      0.79      0.77       134
```

Inference on test data for KNN after tuning:

For predicting people who opt for Public Transport(0):

- Precision (77%) – 77% of prediction of people who will opt for public transport are correct
- Recall (48%) – 48% of people who will opt for public transport are correctly predicted

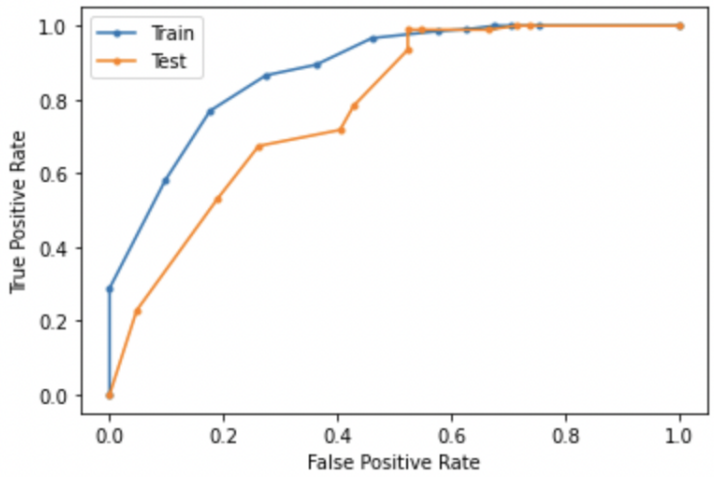
For predicting people who opt for Private Transport(1):

- Precision (80%) – 80% of prediction of people who will opt for private transport are correct
- Recall (93%) – 93% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 79% of total predictions are correct

ROC AUC Curve:

```
KNN Train: ROC AUC=0.880
KNN Test: ROC AUC=0.776
```



Boosting Classifier:

Train Accuracy is : 0.967741935483871

Test Accuracy is : 0.7686567164179104

Train ROC-AUC score is : 0.9979732277526395

Test ROC-AUC score is : 0.8118530020703933

Confusion matrix for train set :
[[93 9]
[1 207]]

Confusion matrix for test set :
[[25 17]
[14 78]]

```
Classification report Train set :
              precision    recall  f1-score   support

           0       0.99      0.91      0.95       102
           1       0.96      1.00      0.98       208

    accuracy                           0.97       310
   macro avg       0.97      0.95      0.96       310
weighted avg       0.97      0.97      0.97       310
```

Inference on train data for Boosting:

For predicting people who opt for Public Transport(0):

- Precision (99%) – 99% of prediction of people who will opt for public transport are correct
- Recall (91%) – 91% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (96%) – 96% of prediction of people who will opt for private transport are correct
- Recall (100%) – 100% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 97% of total predictions are correct

```
Classification report Test set :
              precision    recall  f1-score   support

           0       0.64      0.60      0.62        42
           1       0.82      0.85      0.83        92

    accuracy                           0.77       134
   macro avg       0.73      0.72      0.73       134
weighted avg       0.76      0.77      0.77       134
```

Inference on test data for Boosting:

For predicting people who opt for Public Transport(0):

- Precision (64%) – 64% of prediction of people who will opt for public transport are correct
- Recall (60%) – 60% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (82%) – 82% of prediction of people who will opt for private transport are correct
- Recall (85%) – 85% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 77% of total predictions are correct

Hyper Parameter Tuning of Boosting Classifier:

Train Accuracy is : 0.9096774193548387

Test Accuracy is : 0.8507462686567164

Train ROC-AUC score is : 0.9844221342383108

Test ROC-AUC score is : 0.8148291925465838

Confusion matrix for train set :
[[75 27]
[1 207]]

Confusion matrix for test set :
[[27 15]
[5 87]]

Classification report Train set :

	precision	recall	f1-score	support
0	0.99	0.74	0.84	102
1	0.88	1.00	0.94	208
accuracy			0.91	310
macro avg	0.94	0.87	0.89	310
weighted avg	0.92	0.91	0.91	310

Inference on train data for Boosting after tuning:

For predicting people who opt for Public Transport(0):

- Precision (99%) – 99% of prediction of people who will opt for public transport are correct
- Recall (74%) – 74% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (88%) – 88% of prediction of people who will opt for private transport are correct
- Recall (100%) – 100% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 91% of total predictions are correct

```
Classification report Test set :
              precision    recall  f1-score   support

           0       0.84      0.64      0.73        42
           1       0.85      0.95      0.90        92

    accuracy                           0.85       134
   macro avg       0.85      0.79      0.81       134
weighted avg       0.85      0.85      0.84       134
```

Inference on test data for Boosting after tuning:

For predicting people who opt for Public Transport(0):

- Precision (84%) – 84% of prediction of people who will opt for public transport are correct
- Recall (64%) – 64% of people who will opt for public transport are correctly predicted

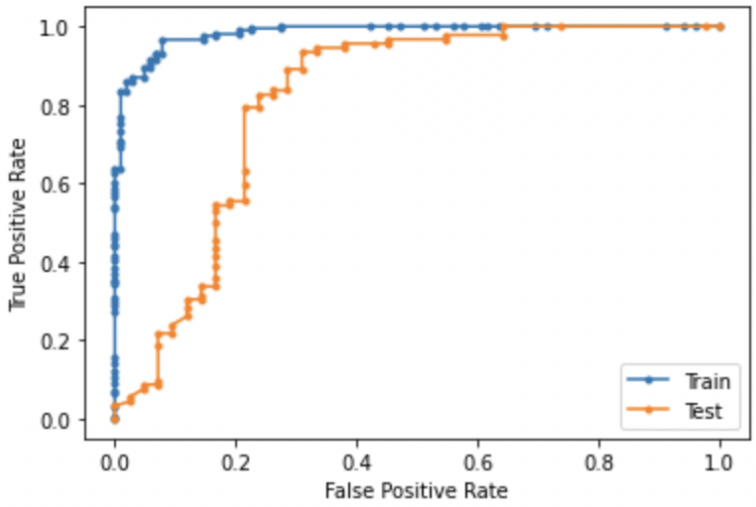
For predicting people who opt for Private Transport(1):

- Precision (85%) – 85% of prediction of people who will opt for private transport are correct
- Recall (95%) – 95% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 85% of total predictions are correct

ROC AUC Curve:

```
Boosting Classifier Train: ROC AUC=0.984
Boosting Classifier Test: ROC AUC=0.815
```



Bagging Classifier:

Train Accuracy is : 0.9838709677419355

Test Accuracy is : 0.7910447761194029

Train ROC-AUC score is : 0.9989866138763197

Test ROC-AUC score is : 0.8197463768115942

Confusion matrix for train set :
[[99 3]
[2 206]]

Confusion matrix for test set :
[[30 12]
[16 76]]

Classification report Train set :

	precision	recall	f1-score	support
0	0.98	0.97	0.98	102
1	0.99	0.99	0.99	208
accuracy			0.98	310
macro avg	0.98	0.98	0.98	310
weighted avg	0.98	0.98	0.98	310

Inference on train data for Bagging:

For predicting people who opt for Public Transport(0):

- Precision (98%) – 98% of prediction of people who will opt for public transport are correct
- Recall (97%) – 97% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (99%) – 99% of prediction of people who will opt for private transport are correct
- Recall (99%) – 99% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 98% of total predictions are correct

```
Classification report Test set :
              precision    recall  f1-score   support

           0       0.65      0.71      0.68        42
           1       0.86      0.83      0.84        92

    accuracy                           0.79       134
   macro avg       0.76      0.77      0.76       134
weighted avg       0.80      0.79      0.79       134
```

Inference on test data for Bagging:

For predicting people who opt for Public Transport(0):

- Precision (65%) – 65% of prediction of people who will opt for public transport are correct
- Recall (71%) – 71% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (86%) – 86% of prediction of people who will opt for private transport are correct
- Recall (83%) – 83% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 79% of total predictions are correct

Hyper Parameter Tuning for Bagging Classifier:

Train Accuracy is : 0.9741935483870968

Test Accuracy is : 0.8208955223880597

Train ROC-AUC score is : 0.9998585972850679

Test ROC-AUC score is : 0.8479554865424431

Confusion matrix for train set :
[[94 8]
[0 208]]

Confusion matrix for test set :
[[27 15]
[9 83]]

```
Classification report Train set :
              precision    recall  f1-score   support

           0       1.00      0.92      0.96       102
           1       0.96      1.00      0.98       208

    accuracy                           0.97       310
   macro avg       0.98      0.96      0.97       310
weighted avg       0.98      0.97      0.97       310
```

Inference on train data for Bagging after tuning:

For predicting people who opt for Public Transport(0):

- Precision (100%) – 100% of prediction of people who will opt for public transport are correct
- Recall (92%) – 92% of people who will opt for public transport are correctly predicted

For predicting people who opt for Private Transport(1):

- Precision (96%) – 96% of prediction of people who will opt for private transport are correct
- Recall (100%) – 100% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 97% of total predictions are correct

```
Classification report Test set :
              precision    recall  f1-score   support

           0       0.75      0.64      0.69        42
           1       0.85      0.90      0.87        92

    accuracy                           0.82       134
   macro avg       0.80      0.77      0.78       134
weighted avg       0.82      0.82      0.82       134
```

Inference on test data for Bagging:

For predicting people who opt for Public Transport(0):

- Precision (75%) – 75% of prediction of people who will opt for public transport are correct
- Recall (64%) – 64% of people who will opt for public transport are correctly predicted

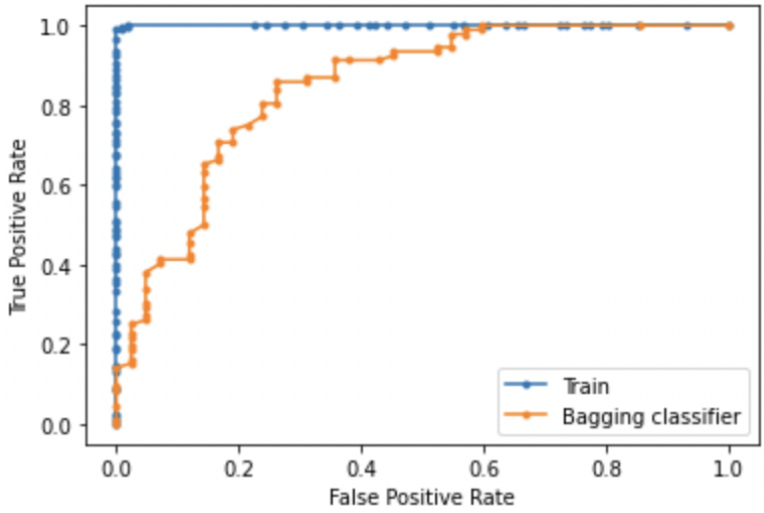
For predicting people who opt for Private Transport(1):

- Precision (85%) – 85% of prediction of people who will opt for private transport are correct
- Recall (90%) – 90% of people who will opt for private transport are correctly predicted

** Overall accuracy of the model – 82% of total predictions are correct

ROC AUC Curve:

```
Bagging Classifier Train: ROC AUC=1.000
Bagging classifier test: ROC AUC=0.848
```

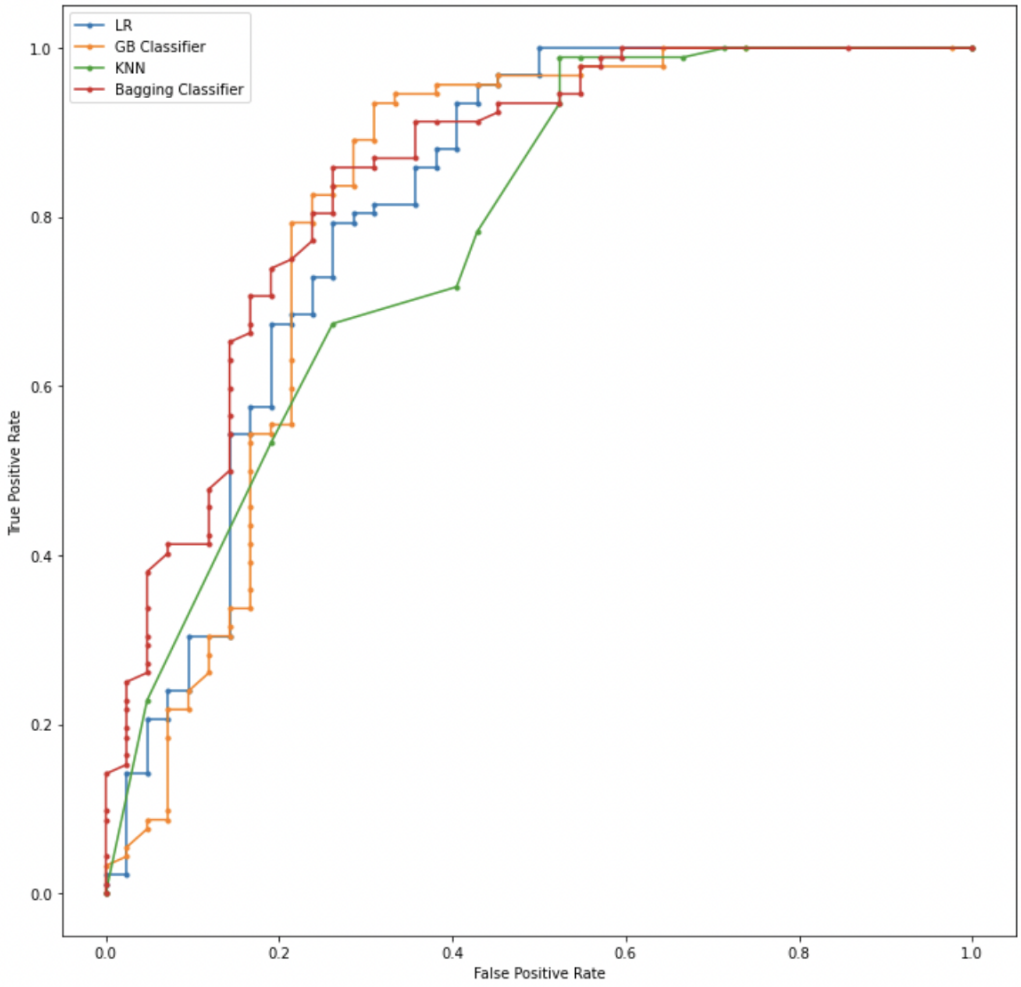


MODEL COMPARISON:

	Model Names	Train Accuracy	Test Accuracy	Train AUC ROC	Test AUC ROC
0	Log Reg model 4	0.803226	0.828358	0.827253	0.814700
1	KNN	0.841935	0.783582	0.922488	0.749612
2	KNN Hyper tuning	0.825806	0.791045	0.880161	0.776268
3	Boosting	0.967742	0.768657	0.997973	0.811853
4	Boosting Hyper Tuning	0.909677	0.850746	0.984422	0.814829
5	Bagging	0.983871	0.791045	0.998987	0.819746
6	Bagging Hyper Tuning	0.974194	0.820896	0.999859	0.847955

Final Conclusion:

```
Logistic Regression: ROC AUC=0.81
Boosting Classifier: ROC AUC=0.815
KNN Classifier: ROC AUC=0.776
Bagging Classifier: ROC AUC=0.848
```



Inferences from the Classification reports:

- If the company is looking for a model which can help them to find the preferred mode of transport,and thus helping them to choose a model with higher precision makes sense. As in this correctness of predicting the mode of transport as private, people opting for private transport is more important such that the company can provide transportation facilities to the ones which are opting private transport. Lower the percentage means that the company lost money on the respective lead as the model predicts it as public transport even though they are a prospective private transport.
- If the company is looking to acquire more number of people irrespective of cost of acquisition than recall makes more sense. Here they want to maximise prediction of employee opting for private transportation.
- In this case, both precision and recall are important for which the company may concentrate on expansion of their services or their profit acquisition.

Inferences from ROC AUC:

- The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 –FPR).
- Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- To compare different classifiers, it can be useful to summarize the performance of each classifier into a single measure. One common approach is to calculate the area under the ROC curve. It is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.
- A classifier with high AUC can occasionally score worse in a specific region than another classifier with lower AUC. However, in practice, the AUC performs well as a general measure of predictive accuracy.

Conclusion:

- Train and Test accuracies of Boosting are 96% and 76% respectively. It clearly seems to be overfitting.
- After hyper tuning the parameters, the overfitting issue is solved and the train and test accuracies came closer for Boosting technique i.e., 90% and 85%
- After hyper tuning, KNN accuracy scores are 82% and 79% which seems to be performing well on both training and test data.
- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.
- The ultimate goal is to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results.
- However, for bagging there is only a slight change in the accuracy scores after hyper tuning the parameters. Overfitting is reduced very slightly.
- As per the model comparison, we see that Boosting after hyper tuning the parameters have good accuracy scores on both training and test dataset.
- For class 1, Precision(85%) and Recall(95%) are also high for Boosting after tuning the model. It can be used to get good predictions of people who opt for private transport.
- From the ROC AUC graph, we can see that AUC for Boosting and Logistic are almost same. For Bagging, AUC is a bit high and KNN has less AUC when compared to all other models. Higher the AUC, the better the model.
- Company can introduce more features that have a positive correlation which can explain more variance in the target variable. Both boosting and bagging can be performed and can be compared to get the best optimized model.