Project – Machine Learning

Submitted By – Akashatra Sharma

Contents

List Of Figures

Executive Summary

There were basically one Dataset provided which gives us a lot of information. The data set was named as "Transport" and it consists of the attributes of 444 Employees that tells how does he/she commute from home to office. In the dataset, we performed different machine learning techniques and build different models in order to get better understanding and give business implications regarding the data set.

Introduction

The purpose of this assignment was to explore the data set. For that, we'll do different inferential & statistical operations in order to get the most of out the data and help in building a very good model for the ABC Consulting Company.

Starting with the data set, we had gone through it and the briefing of the data set was as follows :

- Dataset 'Transport' consists of Age, Salary and others features of 444 entries and using that data we'll find out the best effective predictors that will determine the mode of transport used by Employees to commute on the basis of these features.

Data Dictionary

The data dictionary is mainly for the understanding of meaning of columns provided in the data set .

1. **Transport**
   It was as follows :

Data Dictionary :

1. Age - Age of the Employee in Years
2. Gender - Gender of the Employee
3. Engineer - For Engineer = 1, Non Engineer = 0
4. MBA - For MBA = 1, Non MBA = 0
5. Work Exp - Experience In Years
6. Salary - Salary in Lakhs per annum
7. Distance - Distance in kms from Home to Office
8. license - If Employees has Driving license = 1, If not, then 0
9. Transport - Mode of Transport(Target Variable)

Figure 1 : Data Dictionary

Data Description

Description of data set was as follows :

1. **Transport**

- Age : Continuous Data from 18.0 to 43.0
- Gender : Categorical Data featuring Male, Female
- Engineer : Discrete Data (For Engineer = 1, For Non Engineer = 0)
- MBA : Discrete Data (For MBA = 1, For Non MBA = 0)
- Work Exp : Continuous Data from 0 to 24.0
- Salary : Continuous Data from 6.5 LPA to 57.0 LPA
- Distance : Continuous Data from 3.2 Kms to 23.4 Kms
- license : Discrete Data (With Driving license = 1 , with no license = 0)
- Transport : Categorical Data featuring Private Transport, Public Transport

Datasets

1. **Transport**

Here were the first five observations of this data set :

	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

Figure 2 : First Five Observations Of Dataset

Data Analysis

- ## Data Types

  - **Transport**
    The data types of variables in the data set were :

```
<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
```

Figure 3 : Data Types

Interpretations :

1. There were 5 integer variables, 2 float variables and 2 object variables.
2. Transport variables was the target variable as per the question.
3. As per the info function, we inferred that there were no null values present in the data set.

- ## Descriptive Statistics
  - **Transport**
    For finding descriptive statistics, we split the data into two categories named as cat(for categorical variables) & num (for numerical variable)

	count	mean	std	min	25%	50%	75%	max
Age	444.0	27.747748	4.416710	18.0	25.0	27.0	30.000	43.0
Engineer	444.0	0.754505	0.430866	0.0	1.0	1.0	1.000	1.0
MBA	444.0	0.252252	0.434795	0.0	0.0	0.0	1.000	1.0
Work Exp	444.0	6.299550	5.112098	0.0	3.0	5.0	8.000	24.0
Salary	444.0	16.238739	10.453851	6.5	9.8	13.6	15.725	57.0
Distance	444.0	11.323198	3.606149	3.2	8.8	11.0	13.425	23.4
license	444.0	0.234234	0.423997	0.0	0.0	0.0	0.000	1.0

Figure 4 : Descriptive Statistics for Num

	count	unique	top	freq
Gender	444	2	Male	316
Transport	444	2	Public Transport	300

Figure 5 : Descriptive Statistics for Cat

Interpretations :

- We inferred that ‘Age’ & ‘Distance’ had their mean and median almost equal which indicated that there was very less skewness exists in their distribution.
- 'Transport' had highest frequency of Public Transport with 300 out of total 444.
- We observed that Employees mostly working in ABC Consulting Company were mostly Male with a count of 316 out of total 444 Employees.

Checking For Null Values

1. Transport

```
Age          0
Gender       0
Engineer     0
MBA          0
Work Exp     0
Salary       0
Distance     0
license      0
Transport    0
dtype: int64
```

Figure 6 : Output For Null Value Check

```
<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
```

Figure 7 : Null Value Cross Confirmation

Interpretations :

- There were zero null values present in the data set.
- Also, we noted that the shape/ dimensions of data set is (444, 9) which means that there were 444 entries and 9 columns in the data set.

Problem – Transport

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 finalized.

Problem 1.1 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.

At first we loaded the Data Dictionary for understanding of column name for data set "Cubic Zirconia".

The Data Dictionary is as follows :

Data Dictionary :

1. Age - Age of the Employee in Years
2. Gender - Gender of the Employee
3. Engineer - For Engineer = 1, Non Engineer = 0
4. MBA - For MBA = 1, Non MBA = 0
5. Work Exp - Experience In Years
6. Salary - Salary in Lakhs per annum
7. Distance - Distance in kms from Home to Office
8. license - If Employees has Driving license = 1, If not, then 0
9. Transport - Mode of Transport(Target Variable)

Figure 8 : Data Dictionary – Transport

After loading the data dictionary and dataset itself, we then did descriptive statistics on the data. It's shown above along with null value check.

There were zero null values present in the dataset.

Checking For Duplicates

We checked for any duplicated values in the dataset. If these duplicated were to be found then they must be treated otherwise they will impact while building model.

```
Number of duplicate rows = 0
```

Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport

Figure 9 : Check For Duplicate Data

Interpretations :

- Hence, there were no duplicates present in the dataset.

Checking for Anomalies (Bad Data)

We checked for unknown or special characters present in the dataset that may impact the outcome of our assignment. Here were the results :

```
Gender : 0
Transport : 0
```

Figure 10 : Check for Anomalies - Categorical Variables

```
Age : 0
Engineer : 0
MBA : 0
Work Exp : 0
Salary : 0
Distance : 0
license : 0
```

Figure 11 : Check For Anomalies - Numerical Variables

As per the insights, no anomalies were found in the dataset.

Skewness

We checked the skewness value for all variables present in the dataset. Here it was as follows :

```
Age          0.955276
Gender       0.937952
Engineer    -1.186708
MBA          1.144763
Work_Exp     1.352840
Salary       2.044533
Distance     0.539851
license      1.259293
Transport   -0.753102
dtype: float64
```

Figure 12 : Skewness Value

Interpretations :

- We inferred that only Engineer & Transport variable were negatively (left) skewed. Rest all variables were positively (right) skewed.
- Salary had maximum value for skewness which indicated its tail is at the right end side of the distribution.

Now, before moving to the EDA, we did some necessary changes that were required in the dataset.

- We transformed 'Transport' data type from 'object' to 'categorical' using astype function because it was good for the model building.
- Also, we did rename the 'Work Exp' variable and removed the spacing between them so that minors errors won't disturb during model building and in the EDA process.
- Also, the Independent variable 'Gender' had been converted to integer format (ordinal format) using replace function.

```
<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    int64
 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    category
dtypes: category(1), float64(2), int64(6)
memory usage: 28.4 KB
```

Figure 13 : Data Info After Transforming Transport Data Type

EDA

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. It is of various types such as univariate, bi-variate and multi-variate.
After getting a brief understanding of what is EDA, we did analyze the data and here it is what we had found :

Univariate Analysis

For Univariate analysis, we plotted a Distribution plot and a Boxplot for each column provided in the data set .

The Distribution plot was used for univariant set of observations and visualizes it through a histogram i.e. only one observation and hence we choose one particular column of the dataset.

And, Boxplot was used as a measure of how well the data is distributed in a data set. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data and also shows us whether there are outliers or not.

Here were the Plots for **Transport** data set :

1. **Age**

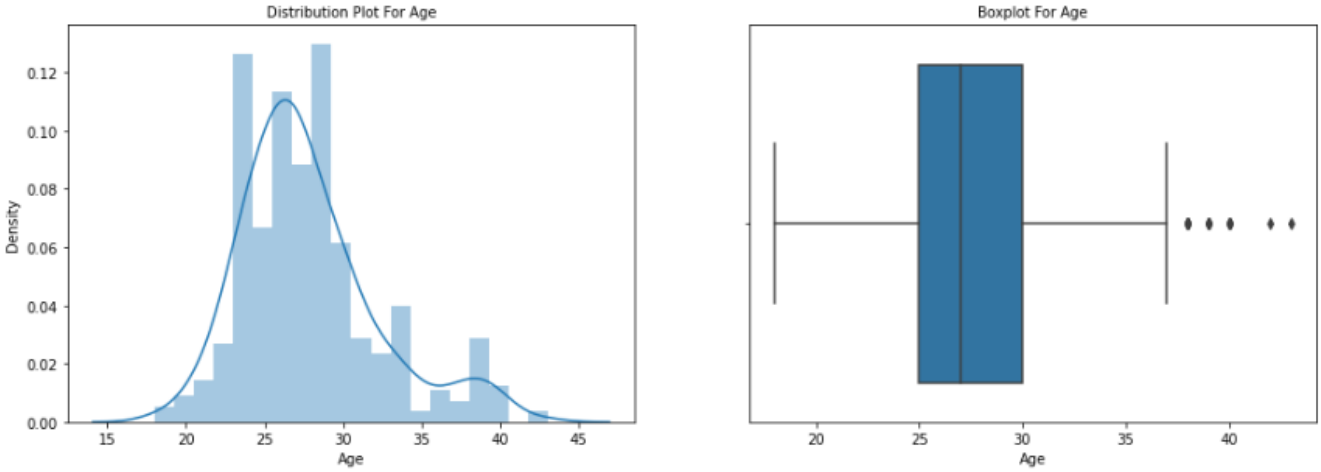

Figure 14 : Dist Plot & Boxplot – Age

The Distplot tells us that the graph is positively (right) skewed as its tail was at the right side of the distribution. Skewness value can be cross checked from the descriptive analysis section.

The 'Age' variable showed extreme values at the right side of boxplot but they cannot be considered as outliers.

2. **Engineer**

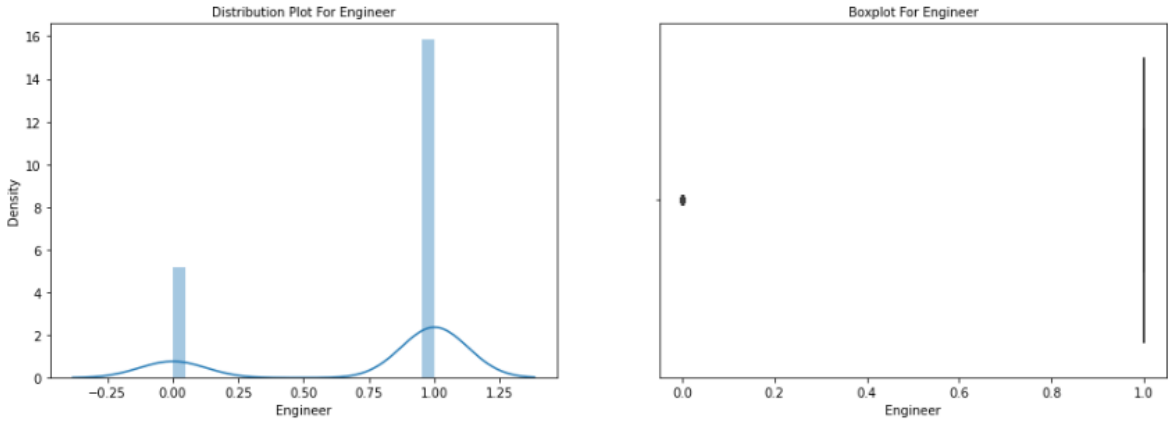

Figure 15 : Dist Plot & Boxplot – Engineer

The boxplot told us that there were many extreme values which exceed the lower limit but these values were not outliers.

It was because these extreme values were checked thoroughly and no absurd or irrelevant values were not present there. This Distplot indicated that it was negatively left skewed.

3. MBA

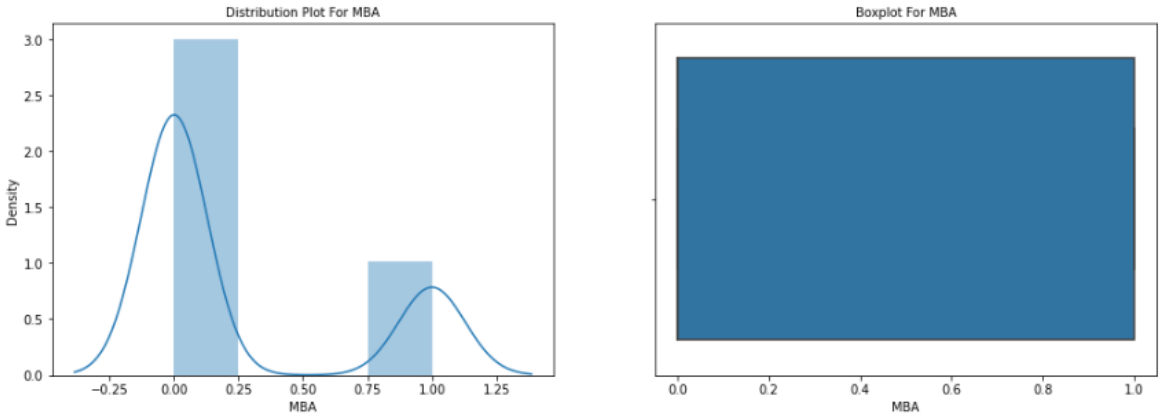

Figure 16 : Dist Plot & Boxplot - MBA

The boxplot of 'MBA' told us that there were no extreme values at either of the ends. The boxplot indicated that 'MBA' variable was positively right skewed.

The 'MBA' variable showed no outliers as per the Boxplot which means all values were laying inside the range from minimum to maximum.

4. Work_Exp

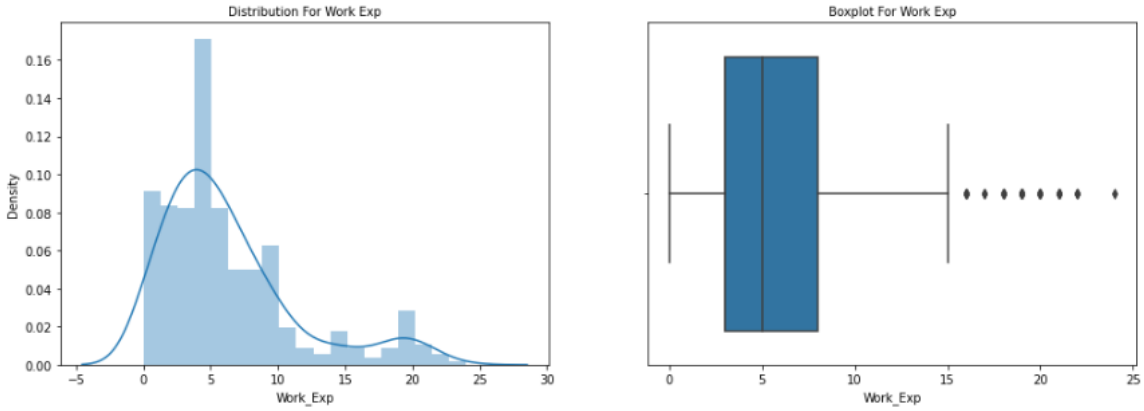

Figure 17 : Dist Plot & Boxplot - Work_Exp

The boxplot told us that there were many extreme values which exceed the upper limit but these values were not outliers. Work Experience increases with age and as the Employees continues to work, his experience also increases.

It was because these extreme values were checked thoroughly and no absurd or irrelevant values were not present there. This Distplot indicated that it was positively right skewed.

5. Salary

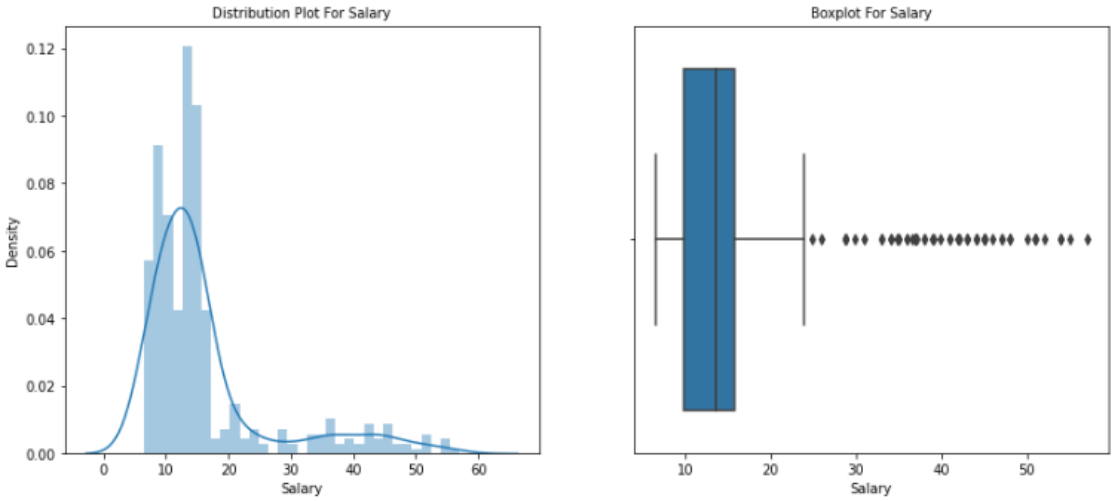

Figure 18 : Dist Plot & Boxplot - Salary

The boxplot told us that there were many extreme values which exceed the upper limit but these values were not outliers. Salary varies on variety of factors and it can be high and low as per them so these extreme values cannot be considered as outliers.

It was because these extreme values were checked thoroughly and no absurd or irrelevant values were not present there. This Distplot indicated that it was positively right skewed.

6. Distance

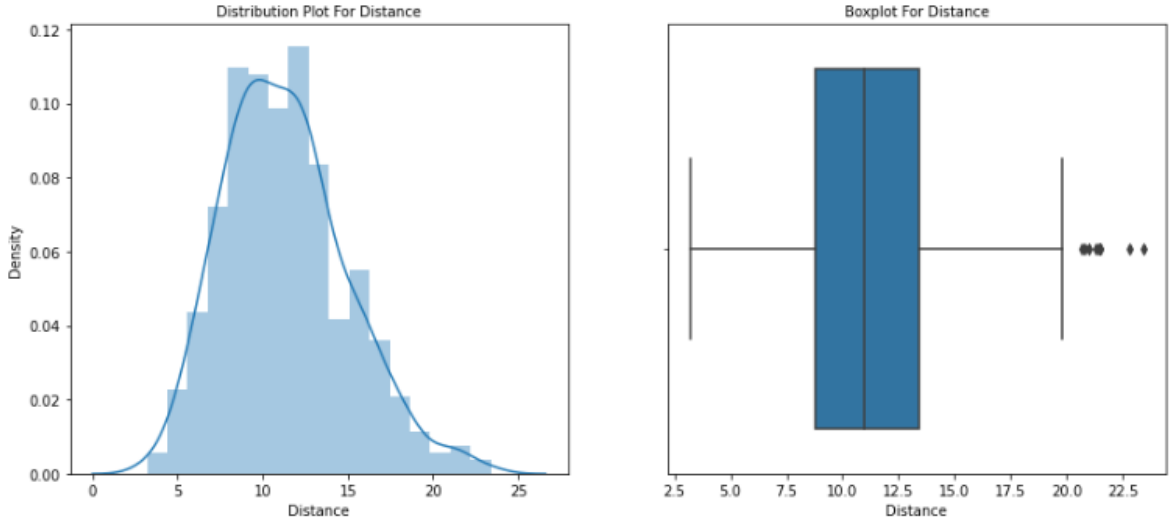

Figure 19: Dist Plot & Boxplot - Distance

As per the Distplot, we got to know that 'Distance' was almost a normally distribution (neither left nor right skewed) but as we saw to the right side of the Distplot, we saw there were values present in high amount which breaks the normal distribution curve but it was close to normal distribution. We can cross confirm this using the skewness value present in the descriptive analysis.

Hence, as per the skewness value, it was considered as positively (right) skewed as its tail was at the right side of the distribution.

7. license

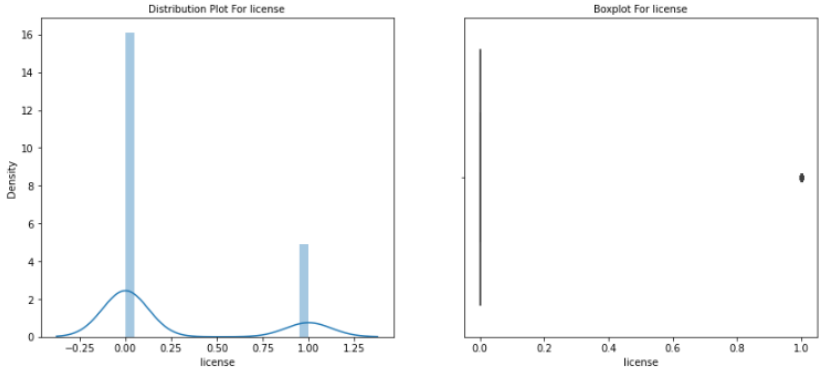

Figure 20 : Dist Plot & Boxplot - license

The boxplot told us that there were many extreme values which exceed the upper limit but these values were not outliers.

It was because these extreme values were checked thoroughly and no absurd or irrelevant values were not present there. This Distplot indicated that it was positively right skewed. We can cross check with the skewness value present in the descriptive analysis.

Bi-Variate Analysis :

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

There are many types of bivariate analysis such as scatter plot**,** regression analysis, correlation matrix analysis and much more.

Boxplot Between Target variable and Independent Variable

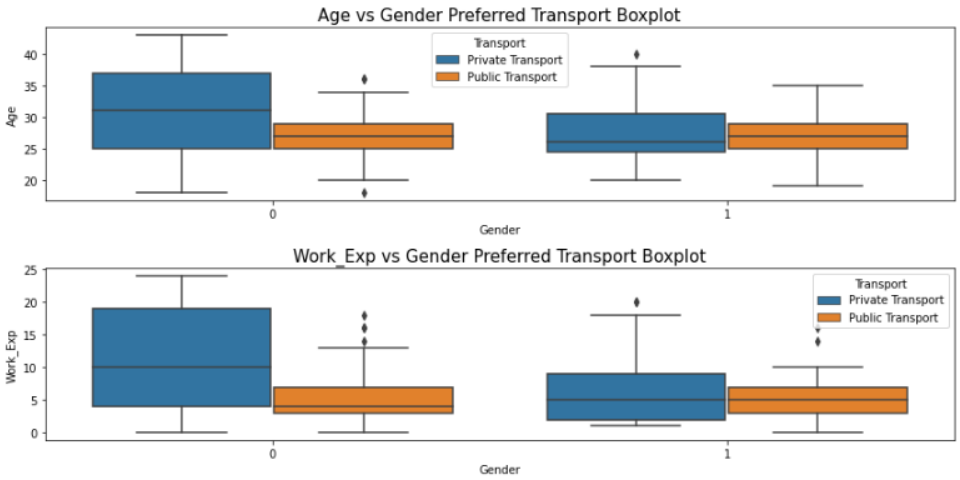

Figure 21 : Age vs Gender Boxplot & Work_Exp vs Gender Boxplot

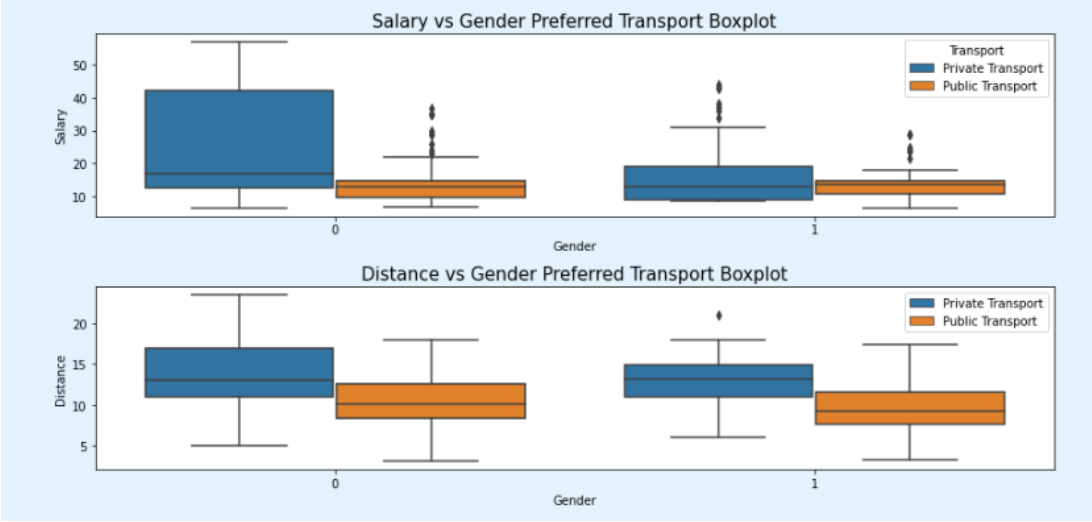

Figure 22 : Salary vs Gender & Distance vs Gender Boxplot

Stripper Plot

We first created a stripper plot with target variable 'Transport' on the X axis and plotted these as follows :

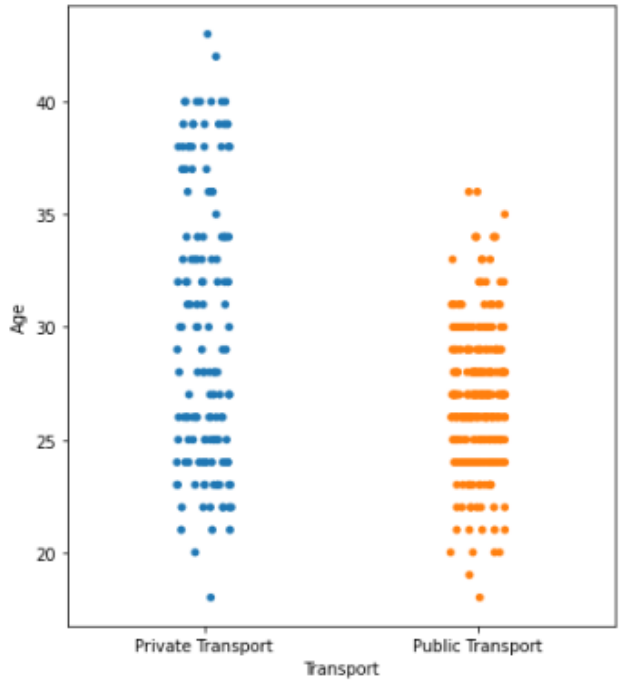

Figure 23 : Stripper Plot - Age vs Transport

Interpretations :

- This tells us that Employees with age greater than 35 tends touse Private Transport.
- Moreover, it also tells us that, young Employees(Age < 35) prefers to use public transport as a mode of vehicle.

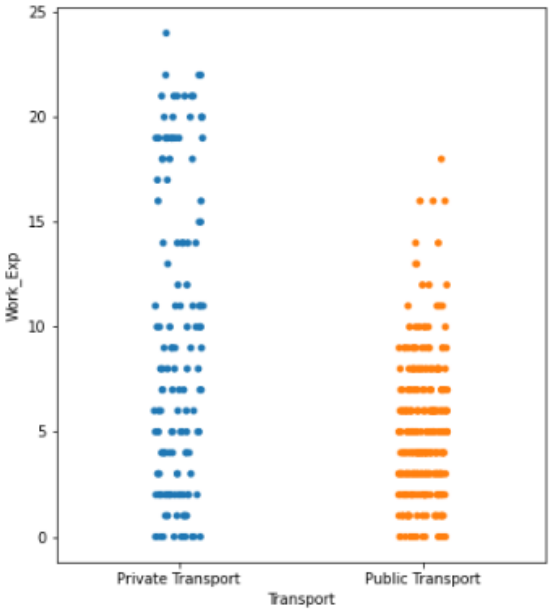

Figure 24 : Stripper Plot – Work_Exp vs Transport

Interpretations :

- It is inferred that Employees with work experience greater than 15 years tends to use Private Transport more as compared to Employees with work experience less than 15 years.
- Moreover, Employees with work experience less than 15 years tends to use Public Transport more.

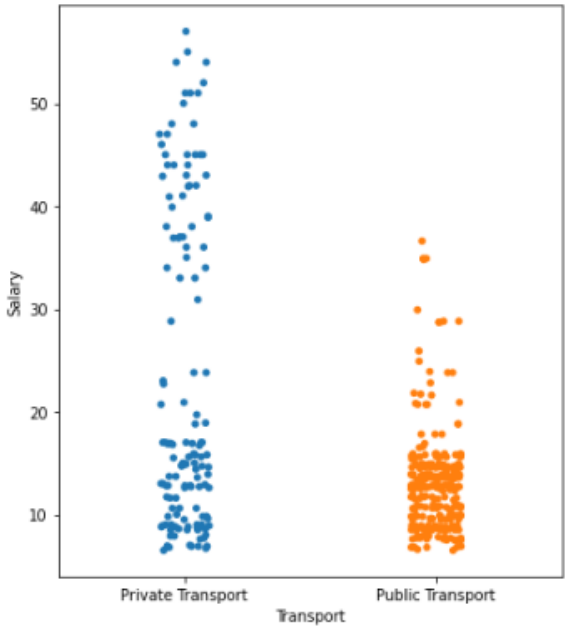

Figure 25 : Stripper Plot - Salary vs Transport

Interpretations :

- We inferred that Employees with salary higher than 30 Lakhs per annum use Private Transport .

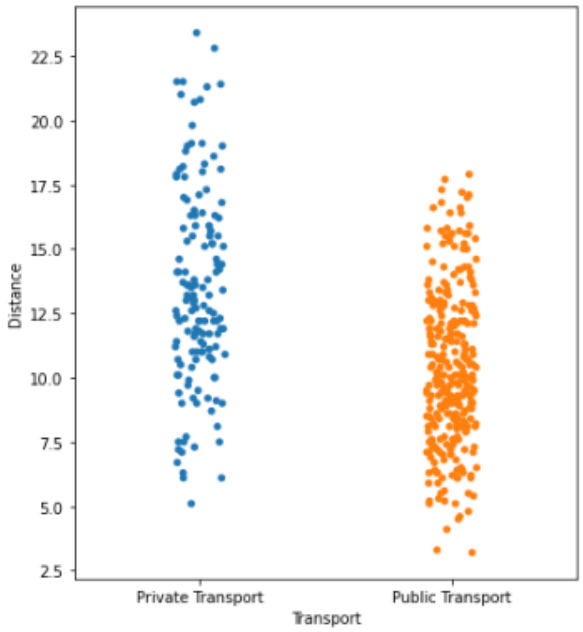

Figure 26 : Stripper Plot - Distance vs Transport

Interpretations :

- We noticed that when longer distance needed to be travelled then Employees took Private Transport. Moreover, the number of employees with private transport were less as compared to number of employees using public transport.
- While , a majority of Employees tends to use Public transport when the distance travelled was 17.5 or below .

Correlation Matrix :

For this data, we did the correlation matrix and find out many insights from it. It is as follows :

	Age	Gender	Engineer	MBA	Work_Exp	Salary	Distance	license
Age	1.000	-0.099	0.092	-0.029	0.932	0.861	0.353	0.452
Gender	-0.099	1.000	-0.018	-0.095	-0.086	-0.096	-0.054	-0.235
Engineer	0.092	-0.018	1.000	0.066	0.086	0.087	0.059	0.019
MBA	-0.029	-0.095	0.066	1.000	0.009	-0.007	0.036	-0.027
Work_Exp	0.932	-0.086	0.086	0.009	1.000	0.932	0.373	0.453
Salary	0.861	-0.096	0.087	-0.007	0.932	1.000	0.442	0.508
Distance	0.353	-0.054	0.059	0.036	0.373	0.442	1.000	0.290
license	0.452	-0.235	0.019	-0.027	0.453	0.508	0.290	1.000

Figure 27 : Correlation Matrix

Interpretations :

- 'Work Exp' and 'Age' had highest correlation between them which is generally obvious as more the age , experience also increases for a working person. Also, 'Work Exp' and 'Salary' had the equal

correlation value which indicated that more the work experience , the more the salary of an Employee.

- Just after that, 'Salary' & 'Age' has second highest correlation between them which meant that age is quite a big factor for salary of an Employee.
- Other than these mostly rest of the variables don't have that much of correlation value as per the plot.

Heatmap :

A **Correlation Heatmap** is a rectangular representation of data and it repeats the same data description twice because the categories are repeated on both axis for computing analysis. Hence, the same result is obtained twice. A correlation heatmap that presents data only once without repetition that is categories are correlated only once is known as a **Triangle Correlation Heatmap**.

Since data is symmetric across the diagonal from left-top to right bottom the idea of obtaining a triangle correlation heatmap is to remove data above it so that it is depicted only once. The elements on the diagonal are the parts where categories of the same type correlate.

The Heatmap was as follows :

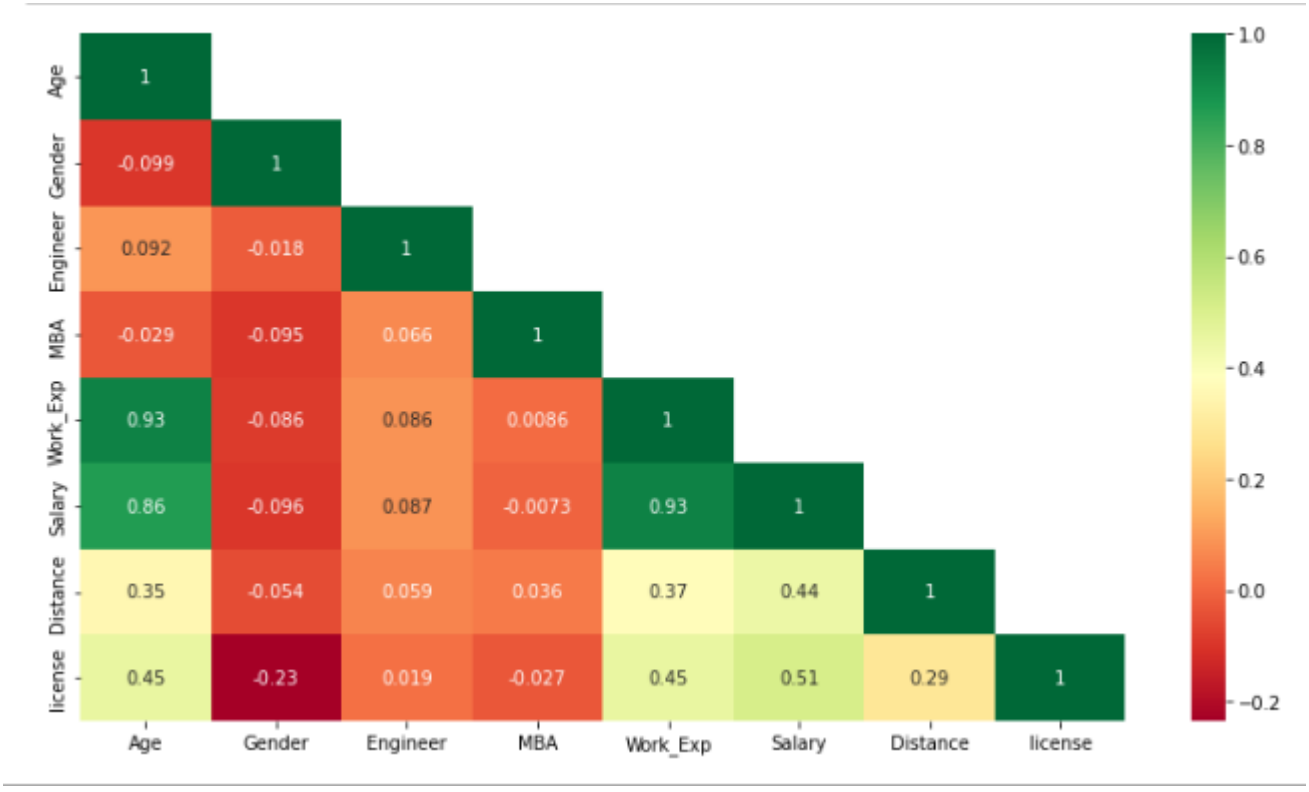

Figure 28 : Heatmap

Multivariate Analysis :

- **Multivariate analysis** (**MVA**) is a Statistical procedure for analysis of data involving more than one type of measurement or observation. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables .

Pairplot :

- **Pairplot** function allows the users to create an axis grid via which each numerical variable stored in data is shared across the X- and Y-axis in the structure of columns and rows. We can create the Scatter plots in order to display the pairwise relationships in addition to the distribution plot displaying the data distribution in the column diagonally.
- The **pairplot()** function can also be used to showcase the subset of variables, or we can plot different types of variables on rows and columns.
- The Pairplot of the following data set was as follows :

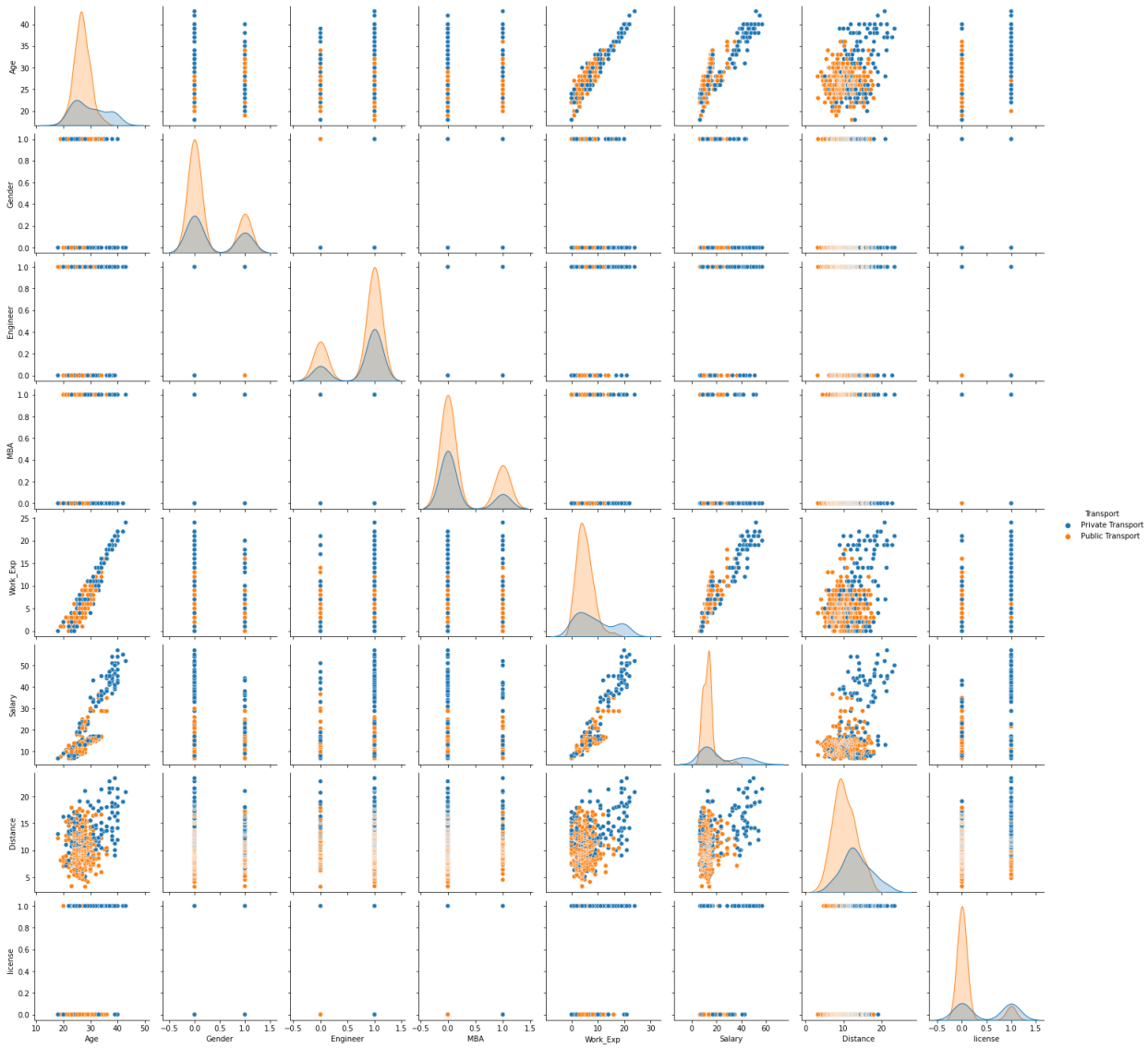

Figure 29 : Pairplot – Transport

Interpretations :

- 'Work Exp' and 'Age' had strong linear positive relationship between them which is generally obvious as more the age , experience also increases for a working person. Also, 'Work Exp' and 'Salary' had the similar strong linear positive relation which indicated that more the work experience , the more the salary of an Employee.
- Just after that, 'Salary' & 'Age' has second strong positive relation between them which meant that age is quite a big factor for salary of an Employee.
- When we looked towards the diagonal then we noticed that ‘Work_Exp’, ‘Salary’, ‘Distance’, ‘license’ do not overall completely with hue as target variable. This meant that these could be useful in predicting the Transport mode for Employees.
- Other than these mostly rest of the variables don't have that much of correlation value as per the plot.

Variance Check

We checked the variance for the dataset and the output was as shown below :

```
Age          19.507331
Gender        0.205641
Engineer      0.185646
MBA           0.189047
Work_Exp     26.133544
Salary      109.283011
Distance     13.004314
license       0.179773
dtype: float64
```

Figure 30 : Check Variance

Interpretations :

- By looking at the variances of the following columns, Engineer, MBA and license had almost zero variance meaning that they will have zero influence on the classification model.
- We also inferred that Salary had the highest variance followed by Work Exp so it means that they both will a large variance on the model.

Checking For Outliers

We plotted a boxplot consisting all variables of the dataset in order to find the outliers.

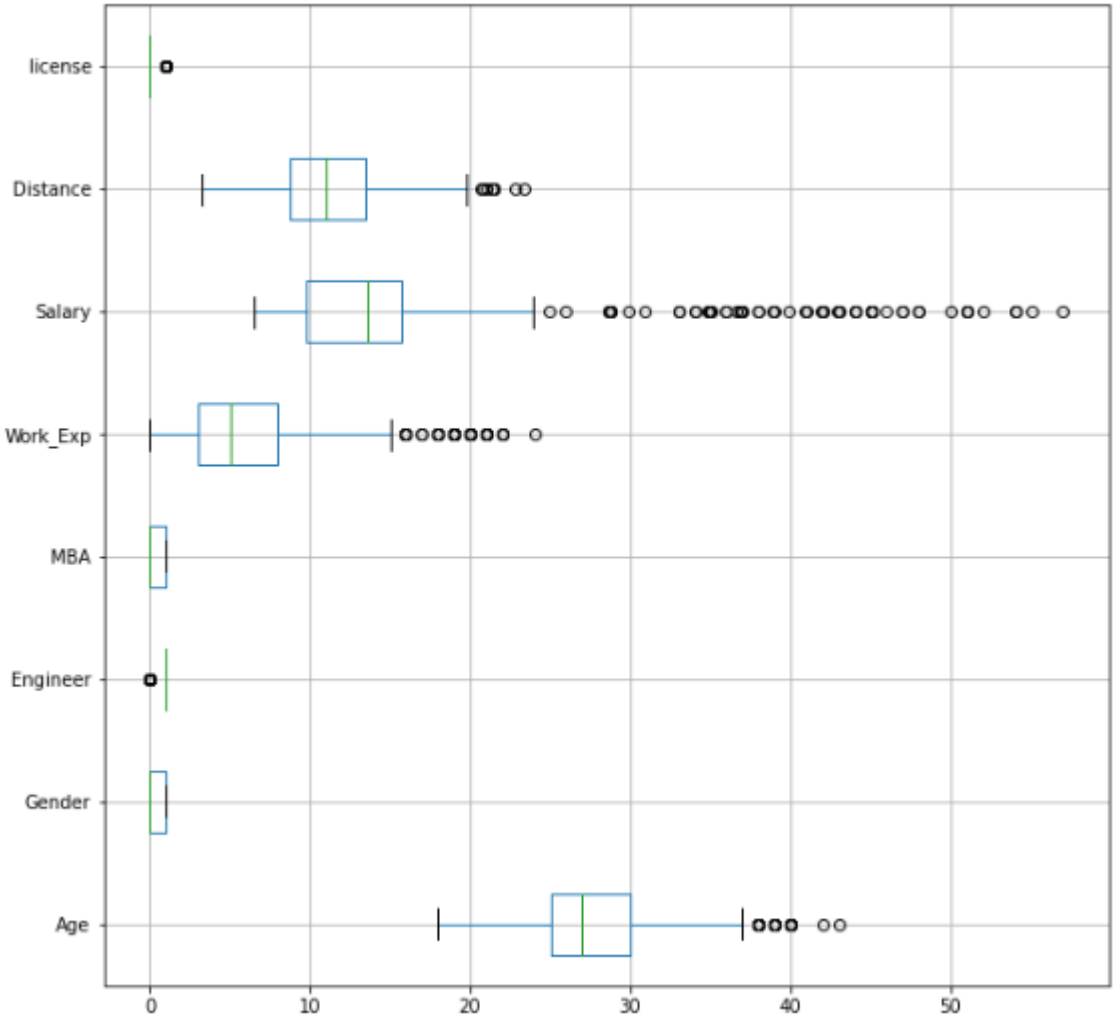

Figure 31 : Check For Outliers

Interpretations :

- There were no extreme values present in all the variables except for MBA and Gender.
- Maximum number of extreme values were present in the Salary variable but these were not considered as outliers because Salary is based on the 'Work Exp' and 'Age' as shown by the correlation plot.
- 'Engineer' column represent some extreme value but actually it cannot be considered as an outlier as this variable represents the which Employee was Engineer or not.

Conclusion

- We noted that only extreme values were present in the dataset. And those values were not considered as Outliers.
- As inferred from the boxplot, there were no outliers present in the dataset.
- So, no treatment of outliers was required in the dataset.

Is Scaling Necessary ?

- **Yes**, **Scaling** is necessary for KNN.
- **Scaling** is a necessity when using Distance-based models such as KNN etc. Scaling can be done on continuous and ordinal variables.
- It is because, if the scale of features is very different then normalization is required. This is because the distance calculation done in KNN uses feature values. When the one feature values are large than other, that feature will dominate the distance hence the outcome of the KNN.

Encode The Data

The 'Transport' variable was converted from 'object' type to 'categorical' type. 'Gender' variable has been transformed earlier but still we did encoding. Hence, it is okay to build our model.

We used Label Encoder to encode ‘Transport’ variable (Target variable) and transformed it .

Here was the dataframe , after applying Label Encoding.

It was as follows :

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

Figure 32 : Dataset After Label Encoding

Logistic Regression

We need to apply logistic regression model as per the question.

At start, we builded a base model of logistic regression using statsmodel library and Sklearn library. Then we fit that base model and checked the summary of the model.

Base Model

It was as follows :

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	435
Method:	MLE	**Df Model:**	8
Date:	Thu, 18 Aug 2022	**Pseudo R-squ.:**	0.3049
Time:	01:15:34	**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	0.4470	1.791	0.250	0.803	-3.063	3.957
Age	0.2083	0.077	2.706	0.007	0.057	0.359
Gender	-1.2810	0.288	-4.441	0.000	-1.846	-0.716
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
Work_Exp	-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

Figure 33 : Logistic Regression - Base Model

After that we checked for multicollinearity in the predictor variable(independent variables) using Variance Inflation Factor(VIF) . We used a user defined function for finding VIF.

The VIF Values for base model were as follows :

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

Figure 34 : VIF For Base Model

Interpretations :

- As per the logistic regression base model, we noticed that 'Engineer' & 'Work_Exp' had maximum p value but as we check for multicollinearity using variance inflation factor(VIF), we got to know that 'Work_Exp' & 'Salary' had highest VIF value.

- Our first priority to build a better model was by removing multicollinearity, so for that we decided to remove 'Work_Exp' variable and build another model as it had highest VIF among the rest.

2nd Iteration Model

We removed 'Work_Exp' and then build a new model using the same process.

The 2nd Iteration Model was as follows :

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	436
Method:	MLE	**Df Model:**	7
Date:	Thu, 18 Aug 2022	**Pseudo R-squ.:**	0.3032
Time:	01:15:34	**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	1.6694	1.324	1.261	0.207	-0.925	4.264
Age	0.1561	0.057	2.749	0.006	0.045	0.267
Gender	-1.2805	0.287	-4.454	0.000	-1.844	-0.717
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

Figure 35 : Logistic Regression - 2nd Iteration Model

The VIF Values for 2nd Iteration model were as follows :

```
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
```

Figure 36 : VIF - 2nd Iteration Model

Interpretations :

- As we removed 'Work_Exp' variable, there had been an improvement in model but still a lot of improvement/tuning can be done in the model.

- The p value for 'Engineer was very high as compared to others so for a better regularized model, we must remove it.

3rd Iteration Model

We removed ‘Engineer’ variable as per the interpretations made in the 3rd Iteration Model.

The 3rd Iteration Model summary was as follows :

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	437
Method:	MLE	**Df Model:**	6
Date:	Thu, 18 Aug 2022	**Pseudo R-squ.:**	0.3027
Time:	01:15:34	**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	1.5954	1.315	1.214	0.225	-0.981	4.172
Age	0.1547	0.057	2.730	0.006	0.044	0.266
Gender	-1.2758	0.287	-4.446	0.000	-1.838	-0.713
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

Figure 37 : Logistic Regression - 3rd Iteration Model

The VIF Values for 3rd Iteration model were as follows :

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

Figure 38: VIF Values – 3rd Iteration Model

Interpretations :

- The p value for everyone is now less than 0.05 as per the significance value.
- But asa we checked VIF value for this 3rd Iteration model, we noticed that the VIF value for 'Salary' and 'Age' were high.

- So, we needed to remove 'Age' as it had high VIF and also it was not a big factor as per the correlation with Transport.

4th Iteration Model

We removed 'Age' variable as per the interpretations made in the 4th Iteration Model.

The 4TH Iteration Model summary was as follows :

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	438
Method:	MLE	**Df Model:**	5
Date:	Thu, 18 Aug 2022	**Pseudo R-squ.:**	0.2890
Time:	01:15:35	**Log-Likelihood:**	-198.91
converged:	True	**LL-Null:**	-279.76
Covariance Type:	nonrobust	**LLR p-value:**	4.308e-33

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.9342	0.555	8.892	0.000	3.847	6.022
Gender	-1.2804	0.284	-4.507	0.000	-1.837	-0.724
MBA	0.4560	0.306	1.488	0.137	-0.145	1.057
Salary	-0.0515	0.016	-3.179	0.001	-0.083	-0.020
Distance	-0.2202	0.041	-5.317	0.000	-0.301	-0.139
license	-1.8981	0.324	-5.861	0.000	-2.533	-1.263

Figure 39 : Logistic Regression - 4th Iteration Model

The VIF Values for 4th Iteration model were as follows :

```
Gender  VIF =  1.07
MBA  VIF =  1.01
Salary  VIF =  1.55
Distance  VIF =  1.26
license  VIF =  1.43
```

Figure 40 : VIF Values - 4th Iteration Model

Interpretations :

- We observed that as we removed 'Age' variable, the model become quite good . All VIF values for variables were within the range.
- The p value for only 'MBA' was greater than 0.05 , so we only need to remove it in order to make this model best model.

5th Iteration Model

We removed 'MBA' variable and builded a new model using the same process.

The 5th Iteration Model summary was as follows :

Logit Regression Results

Dep. Variable:	Transport	**No. Observations:**	444
Model:	Logit	**Df Residuals:**	439
Method:	MLE	**Df Model:**	4
Date:	Thu, 18 Aug 2022	**Pseudo R-squ.:**	0.2849
Time:	01:15:35	**Log-Likelihood:**	-200.05
converged:	True	**LL-Null:**	-279.76
Covariance Type:	nonrobust	**LLR p-value:**	1.960e-33

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	5.0360	0.555	9.080	0.000	3.949	6.123
Gender	-1.3261	0.283	-4.690	0.000	-1.880	-0.772
Salary	-0.0501	0.016	-3.149	0.002	-0.081	-0.019
Distance	-0.2199	0.042	-5.288	0.000	-0.301	-0.138
license	-1.9231	0.323	-5.945	0.000	-2.557	-1.289

Figure 41 : Logistic Regression - 5th Iteration Model

The VIF Values for 5th Iteration model were as follows :

```
Gender  VIF =  1.06
Salary  VIF =  1.55
Distance  VIF =  1.25
license  VIF =  1.42
```

Figure 42 : VIF - 5th Iteration Model

Interpretations :

- Now, every variable were having p value less than 0.05 and the VIF value for every variable were good enough and in the range also.
- After all these changes, p value and VIF values seemed to be in a correct range and thus the model was looking very good.

Thus, we plotted a distribution plot based on the 5th Iteration Model. It was as follows :

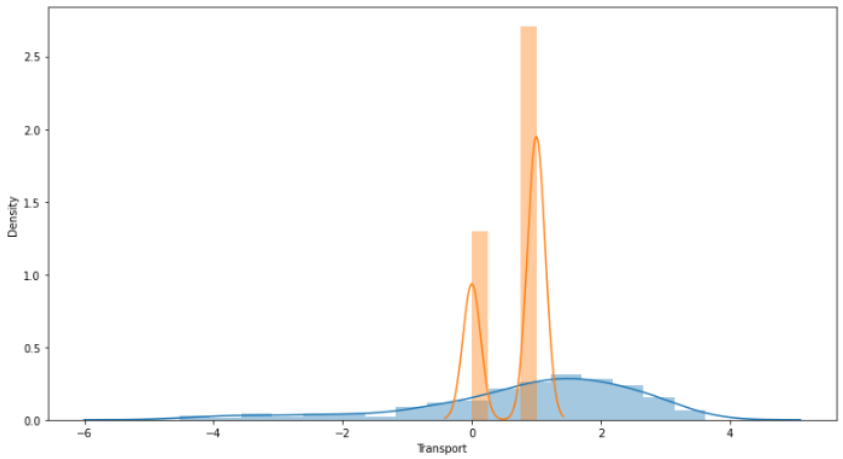

Figure 43 : Dist Plot - 5th Iteration Model

Conclusion :

- As per the insights and making appropriate changes into the model, the **5th Iteration model** was considered to be the **Best Model** from all the models.
- The p value, VIF value and factors of all the attributes were appropriate and this model looks pretty good model.

After this, we used Sklearn library to Train-Test Split on the 5th Iteration Model.
We splitted X and y into training and test set in 70:30 ratio with random_state = 1 as required by the question.
Thus after splitting, we predict function to check the score for both train & test set.

```
Train Accuracy Score of the Final Model: 83 %
Test Accuracy Score of the Final Model: 82 %
```

Figure 44 : Train & Test Score - 5th Iteration Model

The Confusion Matrix was plot for the 5th Iteration Model. It was as follows :

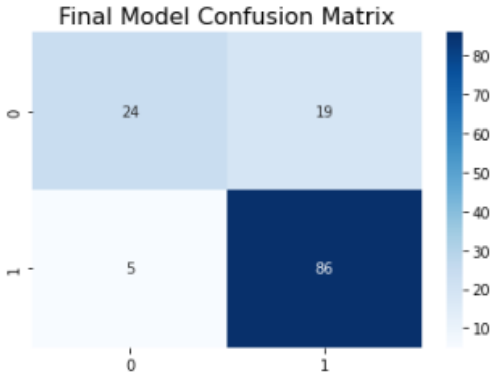

Figure 45 : Confusion Matrix - 5th Iteration Model/Final Model

The classification report for Final Model was as follows :

Final Model Classification Report

	precision	recall	f1-score	support
0	0.83	0.56	0.67	43
1	0.82	0.95	0.88	91
accuracy			0.82	134
macro avg	0.82	0.75	0.77	134
weighted avg	0.82	0.82	0.81	134

Figure 46 : Final Model Classification Report

At last, we plotted the ROC Curve and found ROC-AUC score . It was as follows :

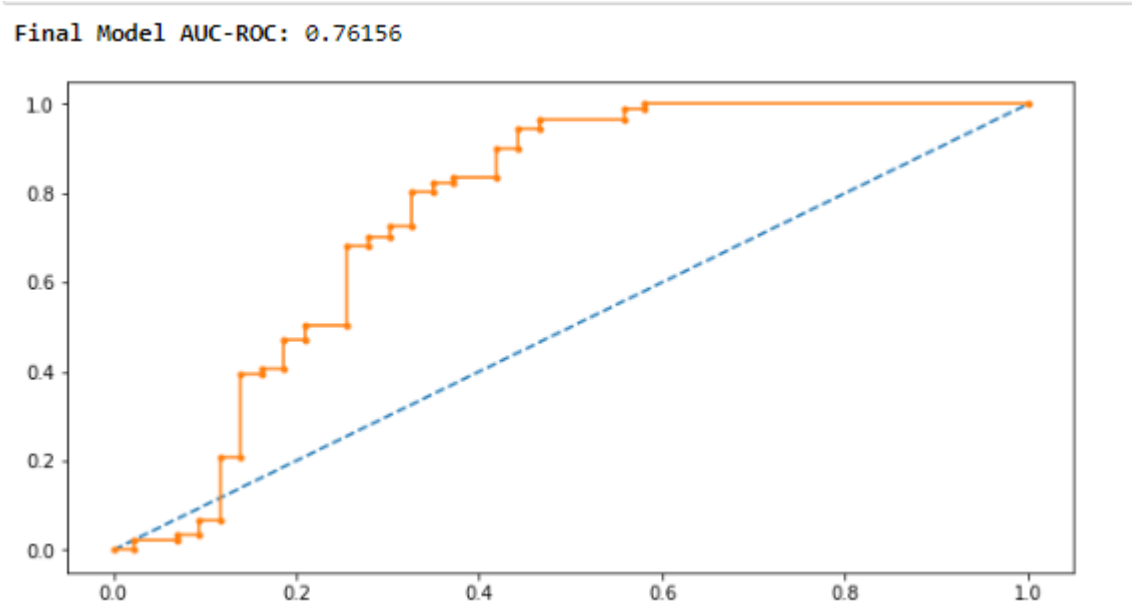

Figure 47 : ROC Curve - Logistic Regression Final Model

Conclusion that 5th Iteration that is Best Model

Transport = (Gender x [-1.3261]) + (Salary x [-0.0501]) + (Distance x [-0.2199]) + (license x [-1.9231])

- The above equation represents the best predictors used in finding out the target variable which is 'Transport' in order to determine whether an employee opted Private Transport or Public Transport.
- AUC for target variable showed us that 76% area came under the curve which was good.
- As per the ROC curve, the curve was going to the other side at the beginning but as we proceeded further , the curve got to the correct side and a proper ROC curve was build up.
- After considering all the factors, predictors, plots & ROC curve, we came to a decision that 5TH Iteration Model was the **Best Model** .
- Hence, **5th Iteration Model** was the **Best Model for Logistic Regression**.

KNN Model

We used **Train-Test Split** and split X and y into training and test set in 70:30 ratio with random_state = 1 as required by the question.

KNN Base Model

The train and test set shape were as follows :

```
X_train: (310, 8)
X_test: (134, 8)
y_train: 310
y_test: 134
```

Figure 48 : KNN Model - Train & Test Data Shape

Then we used standard scalar for scaling the dataset and then build the model.

After that , we fit the model . Moreover, using the predict function we found the accuracy score for train and test set as well as we plotted the confusion matrix for the same.

```
Train Accuracy is: 86 %
Test Accuracy is: 81 %
Train ROC-AUC score is: 93 %
Test ROC-AUC score is: 84 %

Confusion matrix for train set:
 [[ 65  36]
 [  6 203]]

Confusion matrix for test set:
 [[22 21]
 [ 4 87]]
```

Figure 49 : KNN Base Model Confusion Matrix

Interpretations :

- We inferred that the train and test accuracy for the KNN Model were close to each other but the model could be further regularized using certain parameters.
- The confusion matrix of train set indicated that out of total training values 203 values provided were true positive meaning that were classified correctly and the 65 were also classified as true negative.
- As we see through the confusion matrix of test set, we inferred that out of total training values 87 were classified as true positive and 22 were classified as true negative.

After that we build a classification report for the same base model. It was as follows :

Classification Report Train set

	precision	recall	f1-score	support
0	0.92	0.64	0.76	101
1	0.85	0.97	0.91	209
accuracy			0.86	310
macro avg	0.88	0.81	0.83	310
weighted avg	0.87	0.86	0.86	310

Figure 50 : Classification Report Train Set - KNN Base Model

Classification Report Test set

	precision	recall	f1-score	support
0	0.85	0.51	0.64	43
1	0.81	0.96	0.87	91
accuracy			0.81	134
macro avg	0.83	0.73	0.76	134
weighted avg	0.82	0.81	0.80	134

Figure 51 : Classification Model Test Set - KNN Base Model

Interpretations :

- From the classification report of train & test data, we inferred that f1 score for class '1'(Public Transport) were 91% & 87% which were good. But as we saw the f1 score for class '0'(Private Transport), the f1 score values 76% & 64% which were not that good.
- Thus in order to improve the model, hyper tuning must be done using the Grid Search CV which will help in building the best KNN model.

Scaling the dataset / Tuning Using Grid Search CV

We used k =3 for best KNN model and build that model accordingly.

Then we used the same process and build the model using certain parameters used by Grid Search CV. The accuracy and confusion matrix for both train and test set were found out.

Grid Search CV Best KNN Model

The Grid Search CV model output for train & test set were as follows :

```
Train Accuracy is: 83 %
Test Accuracy is: 81 %
Train ROC-AUC score is: 89 %
Test ROC-AUC score is: 84 %
Confusion matrix for train set:
 [[ 55  46]
 [  7 202]]
Confusion matrix for test set:
 [[21 22]
 [ 3 88]]
```

Figure 52 : Accuracy, ROC-AUC Score & Confusion Matrix For KNN Best Model

Interpretations :

- We inferred that accuracy for train & test data for this KNN Model was 83% & 81% which were almost equal.
- This meant that the KNN Model build was quite good. It was cross confirmed through the ROC-AUC score for both train and test set.
- The confusion matrix for train set classified that out of total training values ,202 were found to be true positive and 55 were found to be true negative which meant they were correctly classified by the model.
- As we go through the confusion matrix for test set, we classified that out of total test values, 88 were found to be true positive and 21 were found to be true negative which indicated that they were correctly predicted by the model.

After that we build a classification report for the same Grid Search CV best model. It was as follows :

Classification Report Train set

	precision	recall	f1-score	support
0	0.89	0.54	0.67	101
1	0.81	0.97	0.88	209
accuracy			0.83	310
macro avg	0.85	0.76	0.78	310
weighted avg	0.84	0.83	0.82	310

Figure 53 : Classification Report Train Set - Best KNN Model

Classification Report Test set

	precision	recall	f1-score	support
0	0.88	0.49	0.63	43
1	0.80	0.97	0.88	91
accuracy			0.81	134
macro avg	0.84	0.73	0.75	134
weighted avg	0.82	0.81	0.80	134

Figure 54 : Classification Report Test Set - Best KNN Model

Interpretations :

- We compared the classification report for both train and test set and found out that f1 score values for class'0' (Private Transport) were 67% & 63 %. And the f1 score for class '1' (Public Transport) were 88% & 88% .
- This indicated that the classification report for train & test set were almost identical and thus our model predicts very good.

At last, we plotted the ROC Curve and found ROC-AUC score. It was as follows :

```
KNN Train: ROC-AUC = 0.887
KNN Test: ROC-AUC = 0.844
```

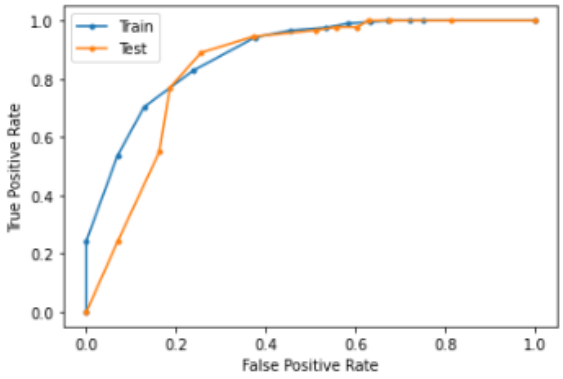

Figure 55 : ROC Curve - Grid Search CV Best KNN Model

Interpretations :

- We compared the ROC Curve for both train and test set and found out that they were some difference at the beginning but after that as the curve goes ahead, the ROC curve becomes identical indicating that train and test set evaluate very accurately.
- It can be confirmed by checking the ROC-AUC value for train and test set. For train set, the value was 88.7% and for test set, the value was 84.4% which were close to each other in terms of value.

Boosting

We used Gradient Boosting Classifier with random_state = 1 and then fit in into the model.

We know that Boosting is a linear sequential process, where next or upcoming model tries to minimize the errors made by previous model in prediction.

So, by using this method, we build a model and found out the accuracy score for both train and test set. In addition we also, evaluate the ROC-AUC score for both the sets and also plotted the confusion matrix for checking the count of True Negative, True Positive, False Negative & False Positive.

Base Model

It was as follows :

```
Train Accuracy is: 97 %
Test Accuracy is: 82 %
Train ROC-AUC score is: 100 %
Test ROC-AUC score is: 87 %
Confusion matrix for train set:
 [[ 92   9]
 [  1 208]]
Confusion matrix for test set:
 [[26 17]
 [ 7 84]]
```

Figure 56 : Boosting Base Model

Interpretations :

- The Train and Test accuracy after applying boosting technique were found to be 97% & 82% .
- The ROC-AUC score for train & test set were found to be 100% & 87%. This was a classic case of overfitted model.
- The confusion matrix tells us the values we got from out both train and test sets.

After that we build a classification report for the same boosting base model. It was as follows :

Classification Report Train set

	precision	recall	f1-score	support
0	0.99	0.91	0.95	101
1	0.96	1.00	0.98	209
accuracy			0.97	310
macro avg	0.97	0.95	0.96	310
weighted avg	0.97	0.97	0.97	310

Figure 57 : Classification Report Train Set - Boosting Base Model

Classification report Test set

	precision	recall	f1-score	support
0	0.79	0.60	0.68	43
1	0.83	0.92	0.87	91
accuracy			0.82	134
macro avg	0.81	0.76	0.78	134
weighted avg	0.82	0.82	0.81	134

Figure 58 : Classification Report Test Set - Boosting Base Model

Interpretations :

- We compared both classification reports of train & test data and got to understand that f1 score for class '0'(Private Transport) was 95% & 68% . And for class '1'(Public Transport) , f1 score values were 98% & 87% .
- As we observed the model accuracy as well as ROC-AUC curve, the model was found to be overfitted.
- Thus, the model must be hyper-tuned using Grid Search CV and build a new better model.

Boosting Technique Using Grid Search CV

Using best params , we used certain parameters such as max depth, min_sample_split etc and build a new model using Grid Search CV and checked its performance.

So, by using this method, we build a model and found out the accuracy score for both train and test set. In addition we also, evaluate the ROC-AUC score for both the sets and also plotted the confusion matrix for checking the count of True Negative, True Positive, False Negative & False Positive.

```
Train Accuracy is: 87 %
Test Accuracy is: 80 %
Train ROC-AUC score is: 96 %
Test ROC-AUC score is: 84 %
Confusion matrix for train set:
 [[ 61  40]
 [  0 209]]
Confusion matrix for test set:
 [[19 24]
 [ 3 88]]
```

Figure 59 : Accuracy, ROC-AUC Score & Confusion Matrix For Boosting Best Model

Interpretations :

- We inferred that accuracy for train & test data for this Boosting Model was 87% & 80% which were near to each other and better than the base model.
- This meant that the Boosting Grid Search CV Model build was quite good. It was cross confirmed through the ROC-AUC score for both train and test set as they better than the base model.

Classification Report Train set

	precision	recall	f1-score	support
0	1.00	0.60	0.75	101
1	0.84	1.00	0.91	209
accuracy			0.87	310
macro avg	0.92	0.80	0.83	310
weighted avg	0.89	0.87	0.86	310

Figure 60 : Classification Report Train Set - Boosting Best Model

Classification Report Test set

	precision	recall	f1-score	support
0	0.86	0.44	0.58	43
1	0.79	0.97	0.87	91
accuracy			0.80	134
macro avg	0.82	0.70	0.73	134
weighted avg	0.81	0.80	0.78	134

Figure 61 : Classification Report Test Set - Boosting Best Model

Interpretations :

- We compared the classification report for both train and test set and found out that f1 score values for class'0' (Private Transport) were 75% & 58%. And the f1 score for class '1' (Public Transport) were 91% & 87% .

- This indicated that the classification report for train & test set were good and thus our model build was a good one.

At last, we plotted the ROC Curve and found ROC-AUC score for this best model. It was as follows :

```
Boosting Classifier Train: ROC-AUC = 0.957
Boosting Classifier Test: ROC-AUC = 0.840
```

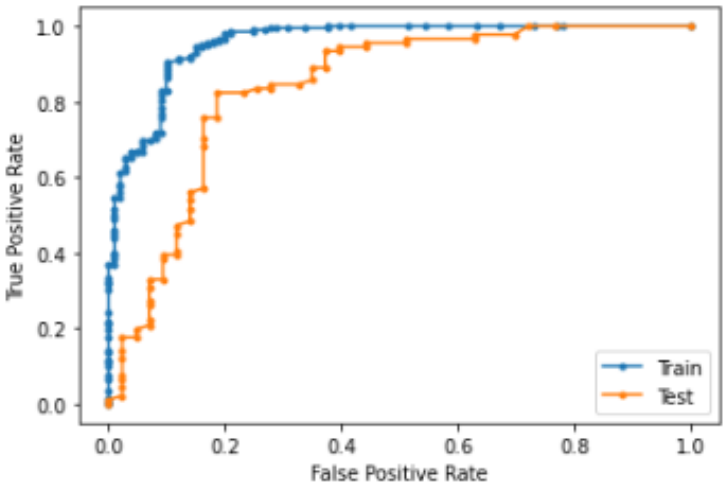

Figure 62 : ROC-AUC Score and ROC Curve - Boosting Best Model

Interpretations :

- We compared the ROC Curve for both train and test set and found out that they were some difference at the beginning but after that as the curve goes ahead, the ROC curve becomes close to each other indicating that train and test set evaluate just average.
- It can be confirmed by checking the ROC-AUC value for train and test set. For train set, the value was 88.7% and for test set, the value was 84.4% which were close to each other in terms of value.

Bagging

We used Bagging Classifier with random_state = 1 and then fit in into the model.

We know that, bagging is a technique of merging the outputs of various models to get a final result.

So, by using this method, we build a Bagging model and found out the accuracy score for both train and test set. In addition we also, evaluate the ROC-AUC score for both the sets and also plotted the confusion matrix for checking the count of True Negative, True Positive, False Negative & False Positive.

Base Model

```
Train Accuracy is: 100 %
Test Accuracy is: 80 %
Train ROC-AUC score is: 100 %
Test ROC-AUC score is: 84 %
Confusion Matrix for train set:
 [[101   0]
 [  0 209]]
Confusion Matrix for test set:
 [[27 16]
 [11 80]]
```

Figure 63 : Accuracy score, ROC-AUC Curve & Confusion Matrix - Bagging Base Model

Interpretations :

- The Train and Test accuracy after applying bagging ensemble technique were found to be 100% & 80% . This meant that the base model is overfitted.
- The ROC-AUC score for train & test set were found to be 100% & 84%. This was a classic case of overfitted model.
- The confusion matrix tells us the values we got from out both train and test sets.

After that , we checked the classification report for the same bagging base model. It was as follows :

Classification Report Train set

	precision	recall	f1-score	support
0	1.00	1.00	1.00	101
1	1.00	1.00	1.00	209
accuracy			1.00	310
macro avg	1.00	1.00	1.00	310
weighted avg	1.00	1.00	1.00	310

Figure 64 : Classification Report Train Set - Bagging Base Model

Classification report Test set

	precision	recall	f1-score	support
0	0.71	0.63	0.67	43
1	0.83	0.88	0.86	91
accuracy			0.80	134
macro avg	0.77	0.75	0.76	134
weighted avg	0.79	0.80	0.79	134

Figure 65 : Classification Report Test Set - Bagging Base Model

Interpretations :

- We compared both classification reports of train & test data and got to understand that f1 score for class '0'(Private Transport) was 100% & 67% . And for class '1'(Public Transport) , f1 score values were 100% & 67% .
- As we observed the model accuracy as well as ROC-AUC curve, the model was found to be overfitted.
- Thus, the model must be hyper-tuned using Grid Search CV and build a new better model.

Bagging Technique Using Grid Search CV

Using best params , we used certain parameters such as max depth, min_sample_split etc and build a new model using Grid Search CV and checked its performance.

So, by using this method, we build a model and found out the accuracy score for both train and test set. In addition we also, evaluate the ROC-AUC score for both the sets and also plotted the confusion matrix for checking the count of True Negative, True Positive, False Negative & False Positive.

```
Train Accuracy is: 91 %
Test Accuracy is: 78 %
Train ROC-AUC score is: 99 %
Test ROC-AUC score is: 84 %
Confusion matrix for train set:
 [[ 76  25]
 [  4 205]]
Confusion matrix for test set:
 [[18 25]
 [ 4 87]]
```

Figure 66 : Accuracy Score, ROC-AUC score, Confusion Matrix - Bagging Best Model

Then afterwards, we checked a classification report for the same Grid Search CV Bagging Best Model. It was as follows :

Classification Report Train set

	precision	recall	f1-score	support
0	0.95	0.75	0.84	101
1	0.89	0.98	0.93	209
accuracy			0.91	310
macro avg	0.92	0.87	0.89	310
weighted avg	0.91	0.91	0.90	310

Figure 67 : Classification Report Train Set - Bagging Best Model

Classification Report Test set

	precision	recall	f1-score	support
0	0.82	0.42	0.55	43
1	0.78	0.96	0.86	91
accuracy			0.78	134
macro avg	0.80	0.69	0.71	134
weighted avg	0.79	0.78	0.76	134

Figure 68 : Classification Report Test Set - Bagging Best Model

Interpretations :

- We compared the classification report for both train and test set and found out that f1 score values for class'0' (Private Transport) were 84% & 55%. And the f1 score for class '1' (Public Transport) were 93% & 86% .
- This indicated that the classification report for train & test set were good and thus our model build was a good one.

At last, we plotted the ROC Curve and found ROC-AUC score for this best model. It was as follows :

```
Bagging Classifier Train: ROC-AUC = 0.990
Bagging classifier test: ROC-AUC = 0.845
```

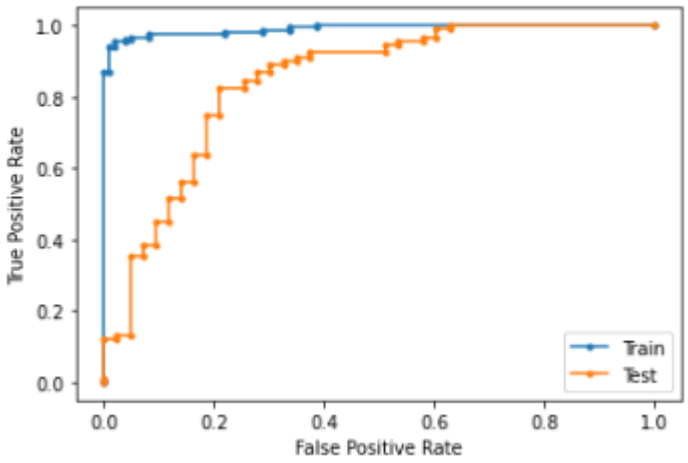

Figure 69 : ROC-SUC Score, ROC Curve - Bagging Best Model

Interpretations :

- We compared the ROC Curve for both train and test set and found out that they were some difference between them as the ROC Curve for train was almost perpendicular but the ROC Curve for test was not similar to that. Although , it was better than the base model, but still quite a development/better model could be build by using some certain parameters.
- It can be confirmed by checking the ROC-AUC value for train and test set. For train set, the value was 99.0% and for test set, the value was 84.5% which were not close to each other in terms of value.

Model Comparison

All the models were compared to each other in order to find out which model was best. The best among these would be used by ABC Company for predicting their Employees commune.

So, we created a new dataframe and put all model score values in it and compared them with each other.

Here is the new dataframe as follows :

	Logistic Regression Test	KNN Train	KNN Test	Boosting Train	Boosting Test	Bagging Train	Bagging Test
Accuracy	82.0	83.0	81.0	87.0	80.0	91.0	78.0
Precision	82.0	81.0	80.0	84.0	79.0	89.0	78.0
Recall	95.0	97.0	97.0	100.0	97.0	98.0	96.0
F1 Score	88.0	88.0	88.0	91.0	87.0	93.0	86.0
AUC	76.0	89.0	84.0	96.0	84.0	99.0	84.0

Figure 70 : Model Comparison DataFrame

Interpretations :

- We compared all the models, especially the best models build during the assignment.
- If we looked at the accuracy for all the models, we got to know that KNN Best model had a very accurate evaluation for train and test set. While rest other models do not have that similar outcomes.
- Since f1 score is the balance for both precision and recall, we checked and compared that for all the models. We inferred that Logistic Regression , KNN train & test model were having same values for f1 score. Also, f1 score for boosting model for train & test were very high meaning that the evaluate quite good.
- The AUC curved is said to be area under curve which shows the comparison between the train and test set. The higher the value of AUC , better is the model.
- From AUC Curve, we found that only for KNN best model, train and test set the value were best as compared to other models.

Conclusions :

- By looking at all the factors, KNN Model can be considered as most optimized model.
- It can be used by the ABC Company to predict the Employees commune of their respective company.

Assignment Insights :

As we all know, KNN classifier is different from other probabilistic classifiers where the model comprises a learning step of computing probabilities from a training sample and use them for future prediction of a test sample.
In KNN there is no learning step involved instead the dataset is stored in memory and is used to classify the test query on the fly. KNN is also known as Lazy learner as it does not create a model using training

set in advance. It is one of the simplest methods of classification. In KNN, the term `k' is a parameter which refers to the number of nearest neighbors.

Coming back to the conclusions, the insights of this assignment were :

- We used k =3 for building the best KNN model using Grid Search CV. KNN is based on distance metric and for this model, we used 'Euclidean', 'Minkowski', 'Manhattan' as the metric and then fit that into the model and runned it.

- We inferred that accuracy for train & test data for this KNN Model was 83% & 81% which were almost equal.
- This meant that the KNN Model build was quite good. It was cross confirmed through the ROC-AUC score for both train and test set.

- The confusion matrix for train set classified that out of total training values ,202 were found to be true positive and 55 were found to be true negative which meant they were correctly classified by the model.
- As we go through the confusion matrix for test set, we classified that out of total test values, 88 were found to be true positive and 21 were found to be true negative which indicated that they were correctly predicted by the model.

- We compared the classification report for both train and test set and found out that f1 score values for class '0' (Private Transport) were 67% & 63 %. And the f1 score for class '1' (Public Transport) were 88% & 88% .

- This indicated that the classification report for train & test set were almost identical and thus our model predicts very good.

- We compared the ROC Curve for both train and test set and found out that they were some difference at the beginning but after that as the curve goes ahead, the ROC curve becomes identical indicating that train and test set evaluate very accurately.

- It can be confirmed by checking the ROC-AUC value for train and test set. For train set, the value was 88.7% and for test set, the value was 84.4% which were close to each other in terms of value.

In the end, all these things keeping into mind, we can surely say that **KNN Grid Search CV model** was said to be the **Best Model** among all**.**

Thus **KNN Best Model** can be used by the ABC Consulting Company and evaluate of their Employees i.e. which one commune using Private Transport or Public Transport.