Machine Learning 2 Extended Project - Business Report

By: Aaryani Kadiyala

PGP-Data Science and Business Analytics

(PGPDSBA.O.JAN24.A)

Table of Contents

List of Figures:

Problem 1:

Context:

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.

Objective

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

Data Dictionary

Age: Age of the Employee in Years

Gender: Gender of the Employee

Engineer: For Engineer =1 , Non Engineer =0

MBA: For MBA =1 , Non-MBA =0

Work Exp: Experience in years

Salary: Salary in Lakhs per Annum

Distance: Distance in km from Home to Office

license: If Employee has Driving Licence -1, If not, then 0

Transport: Mode of Transport

Solution:-

	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 1: First 5 rows Dataset Sample

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
439	40	Male	1	0	20	57.0	21.4	1	Private Transport
440	38	Male	1	0	19	44.0	21.5	1	Private Transport
441	37	Male	1	0	19	45.0	21.5	1	Private Transport
442	37	Male	0	0	19	47.0	22.8	1	Private Transport
443	39	Male	1	1	21	50.0	23.4	1	Private Transport

Figure 2: Last 5 rows Dataset Sample

- The data contains 444 rows and 9 columns.
- Data consists of 2 object, 5 int64 & 2 float64 data type columns.
- 2 object type columns are Transport & gender.

```
<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 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

Figure 4: Statistical summary of the dataset

- There are no null values in the dataset.
- There are no duplicate values in the dataset.
- The mean and median for the only integer column ‘age’ is almost same indicating the column is normally distributed.
- ‘Transport’ has two unique values Public Transport & Private Transport, which is also a dependent variable. ‘gender’ has two unique values male and female.

```
Gender
Male      316
Female    128
Name: count, dtype: int64
--------------------------------------------------
Transport
Public Transport     300
Private Transport    144
Name: count, dtype: int64
--------------------------------------------------
```

Figure 5: Unique categories of categorical variables

Univariant, Bivariant & Multivariant Analysis: -

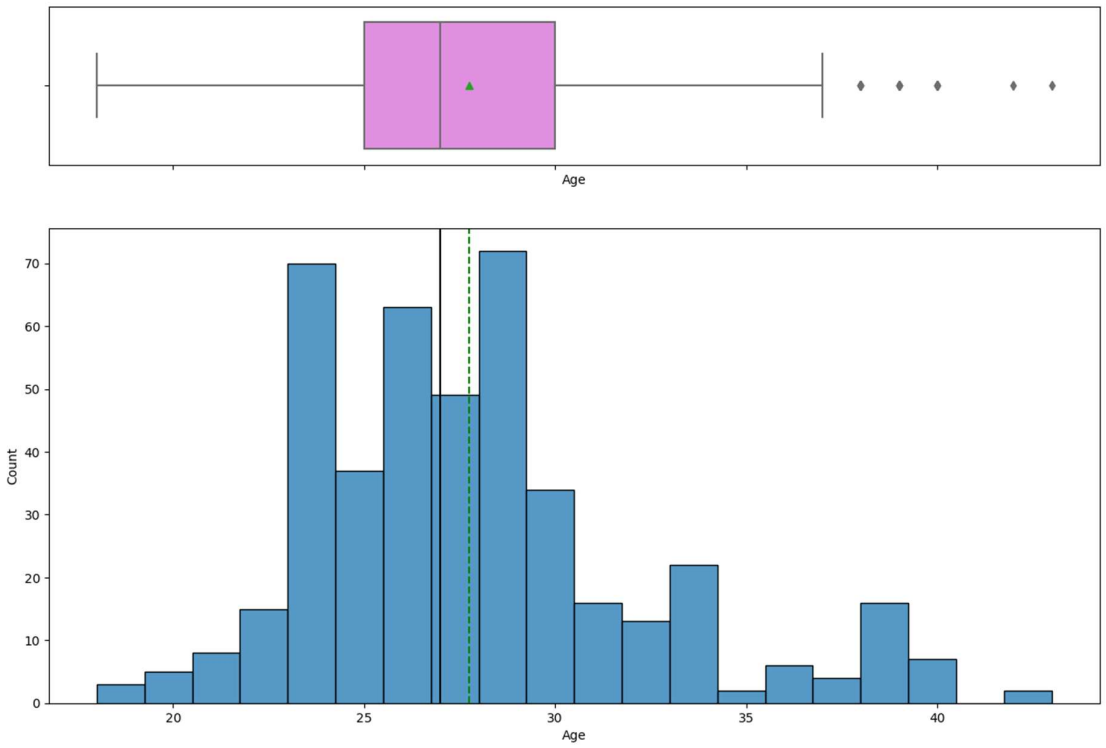

Figure 6: Univariant analysis of Age

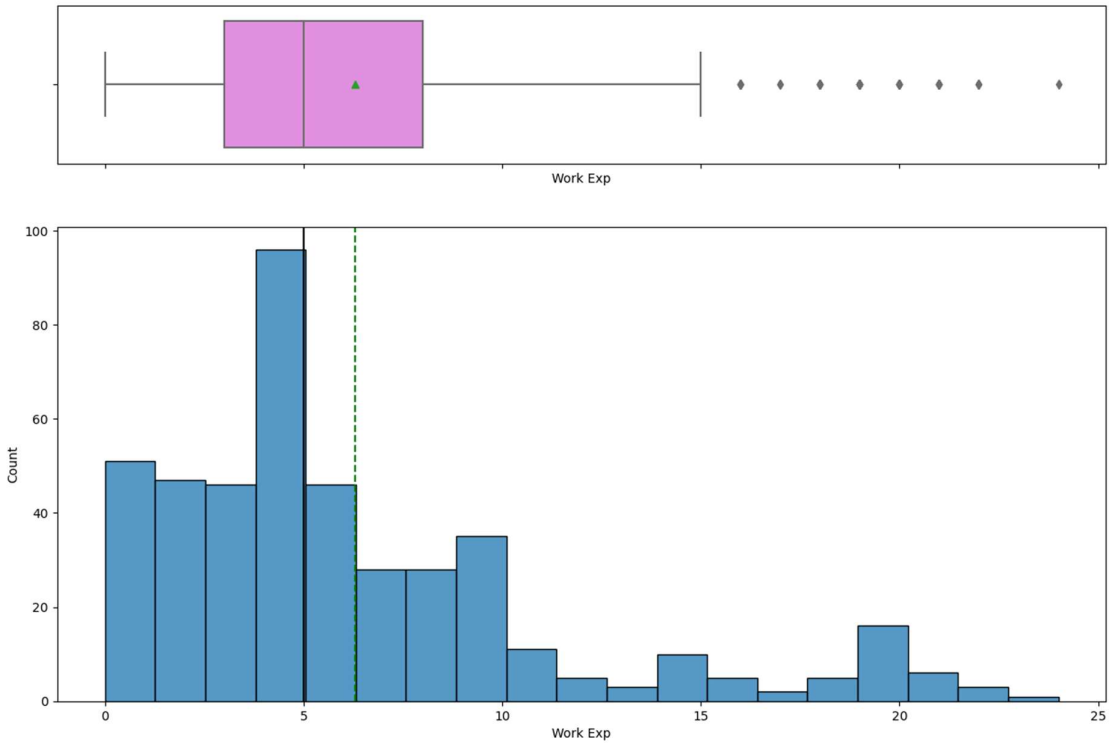

Figure 7: Univariant analysis of Work Experience

From the above analysis it can summarized as follows,

- Both Age & work experience are right skewed.

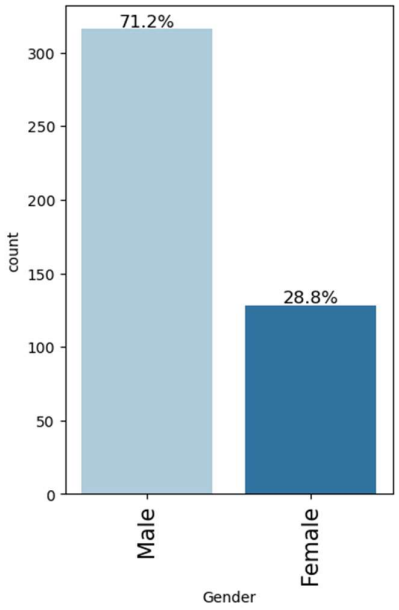

Figure 8: Bar Plot of the Gender

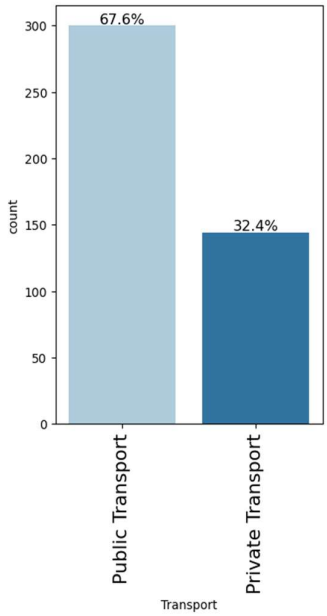

Figure 9: Bar Plot of the Transport

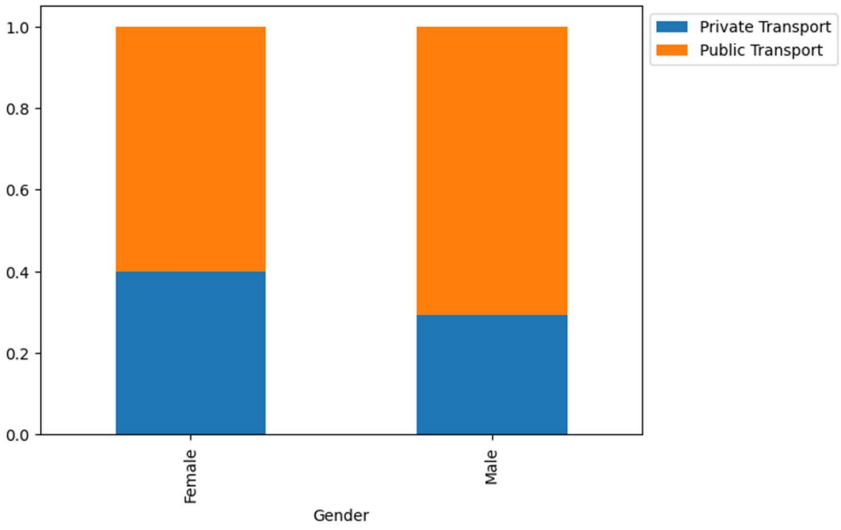

Figure 10: Bar Plot between Transport and Gender

Transport	Private Transport	Public Transport	All
Gender			
All	144	300	444
Male	93	223	316
Female	51	77	128

Figure 11: Transport and Gender in numbers

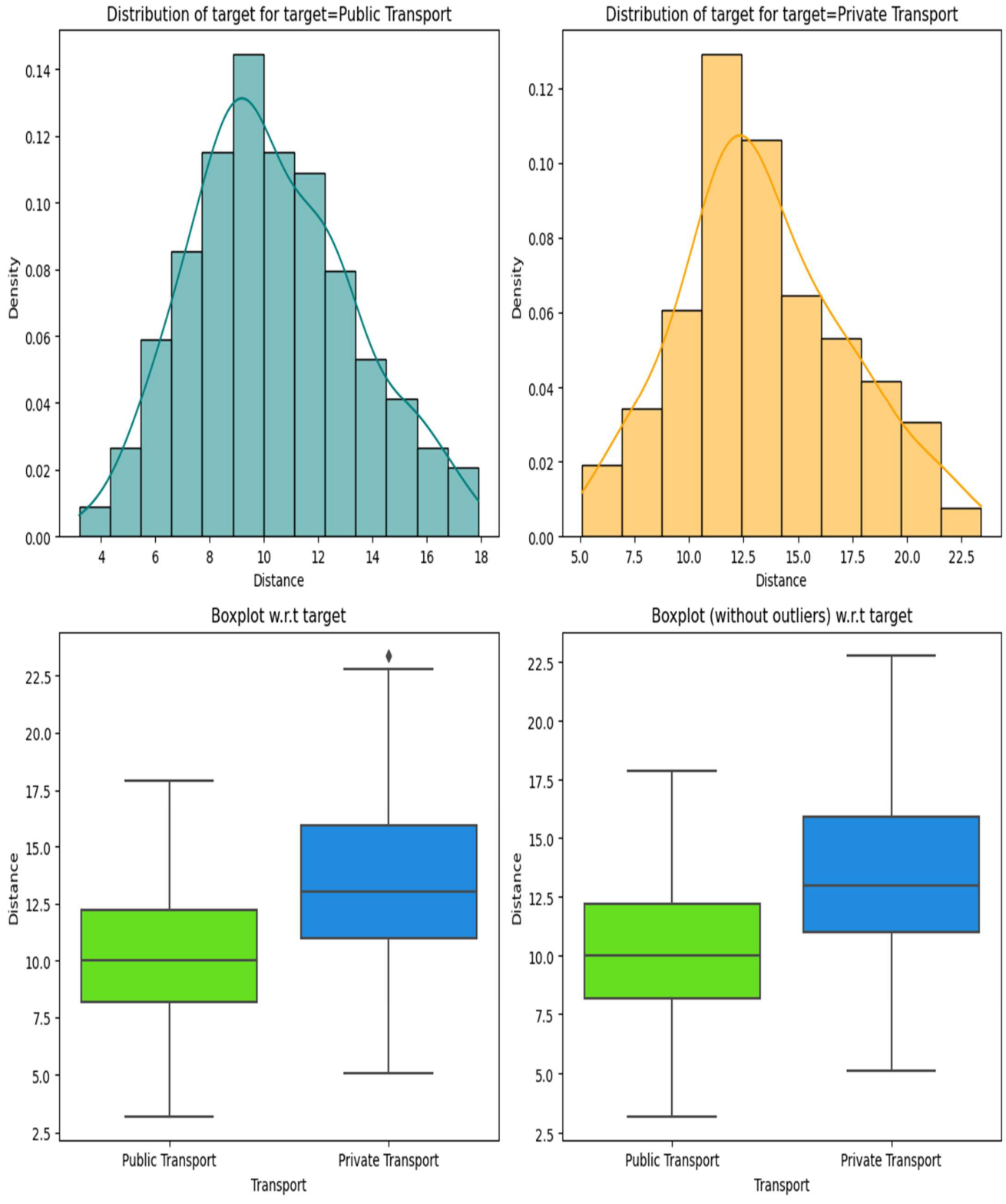

Figure 12: Prediction with distance w.r.t Transport

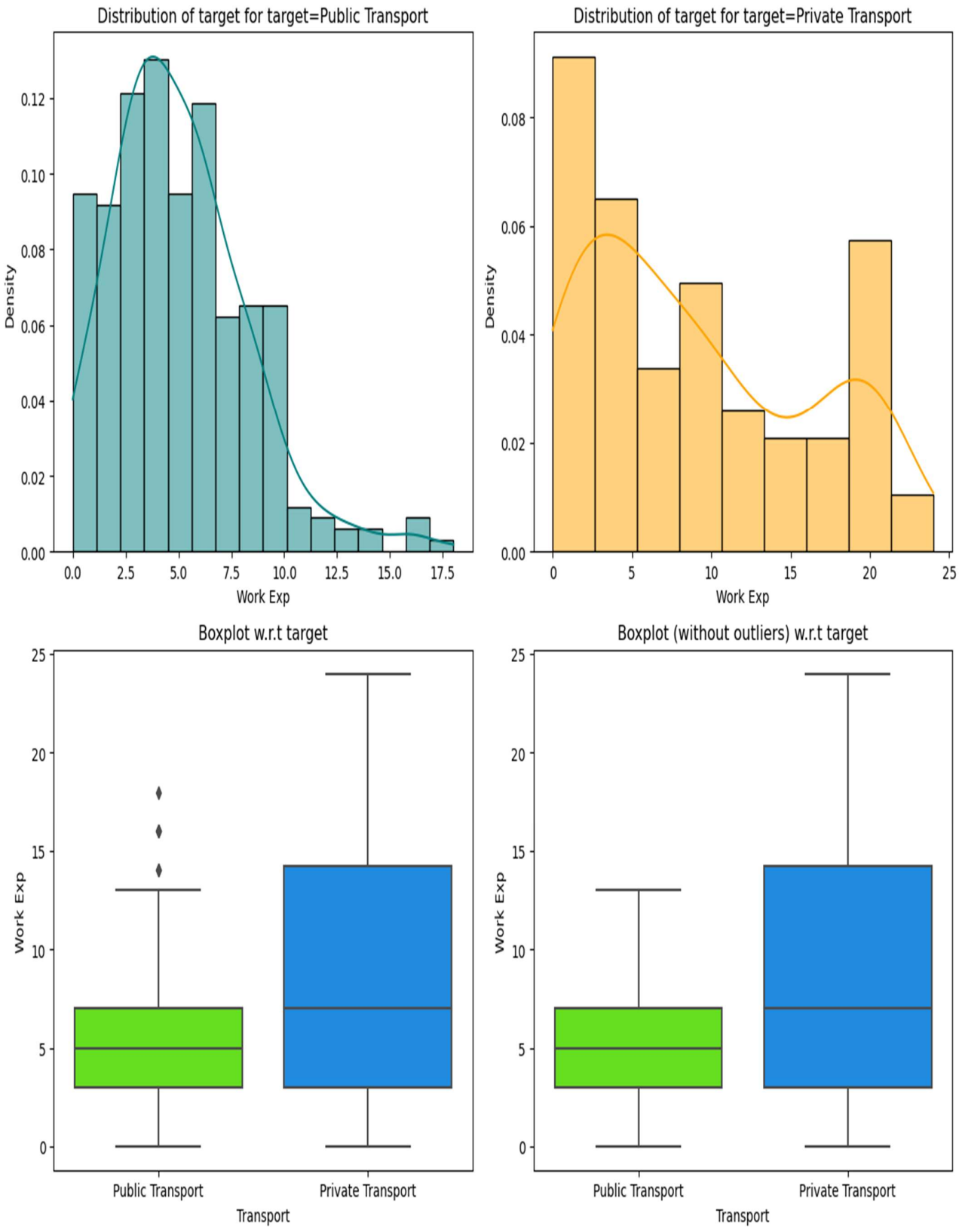

Figure 13: Prediction with Work Experience w.r.t Transport

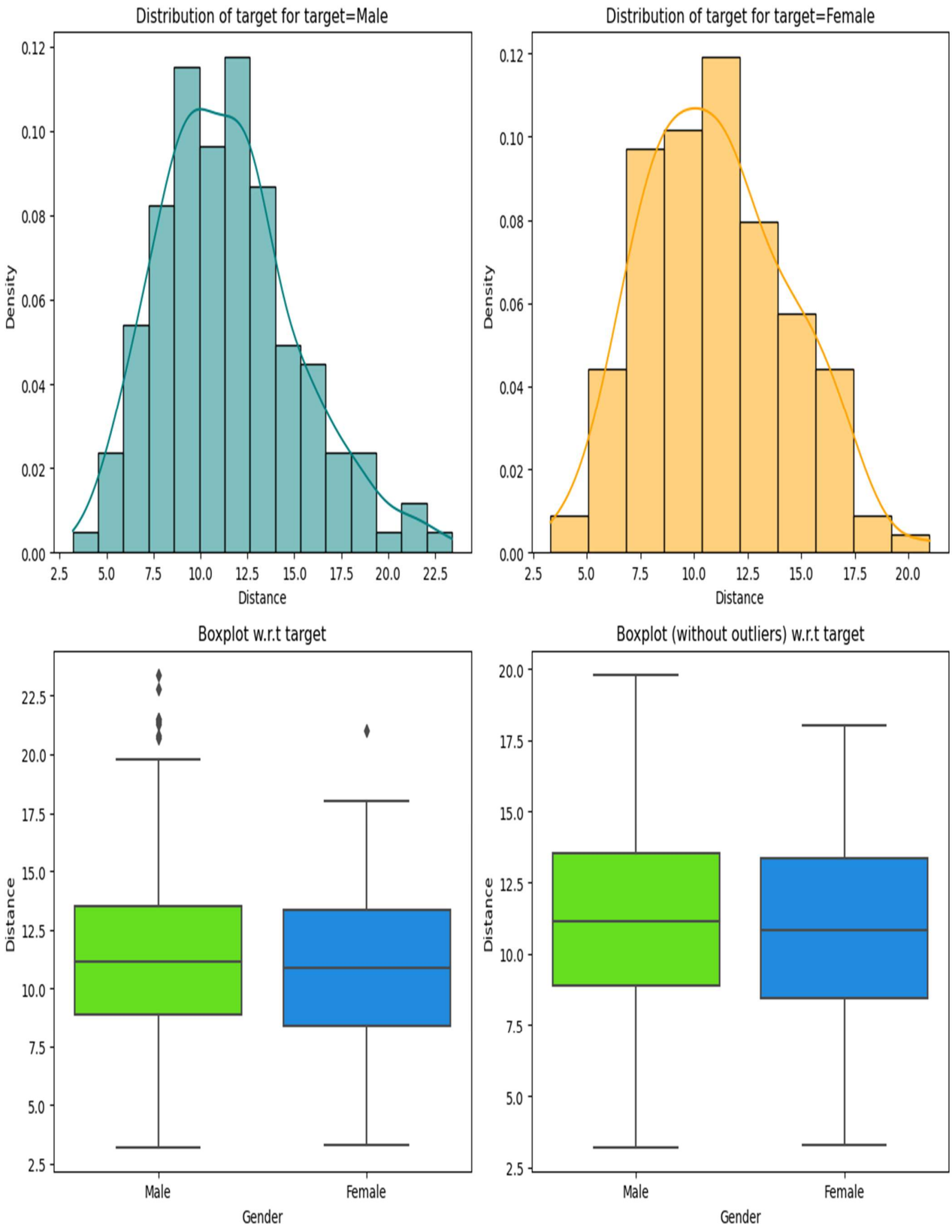

Figure 14: Prediction with Distance w.r.t Gender

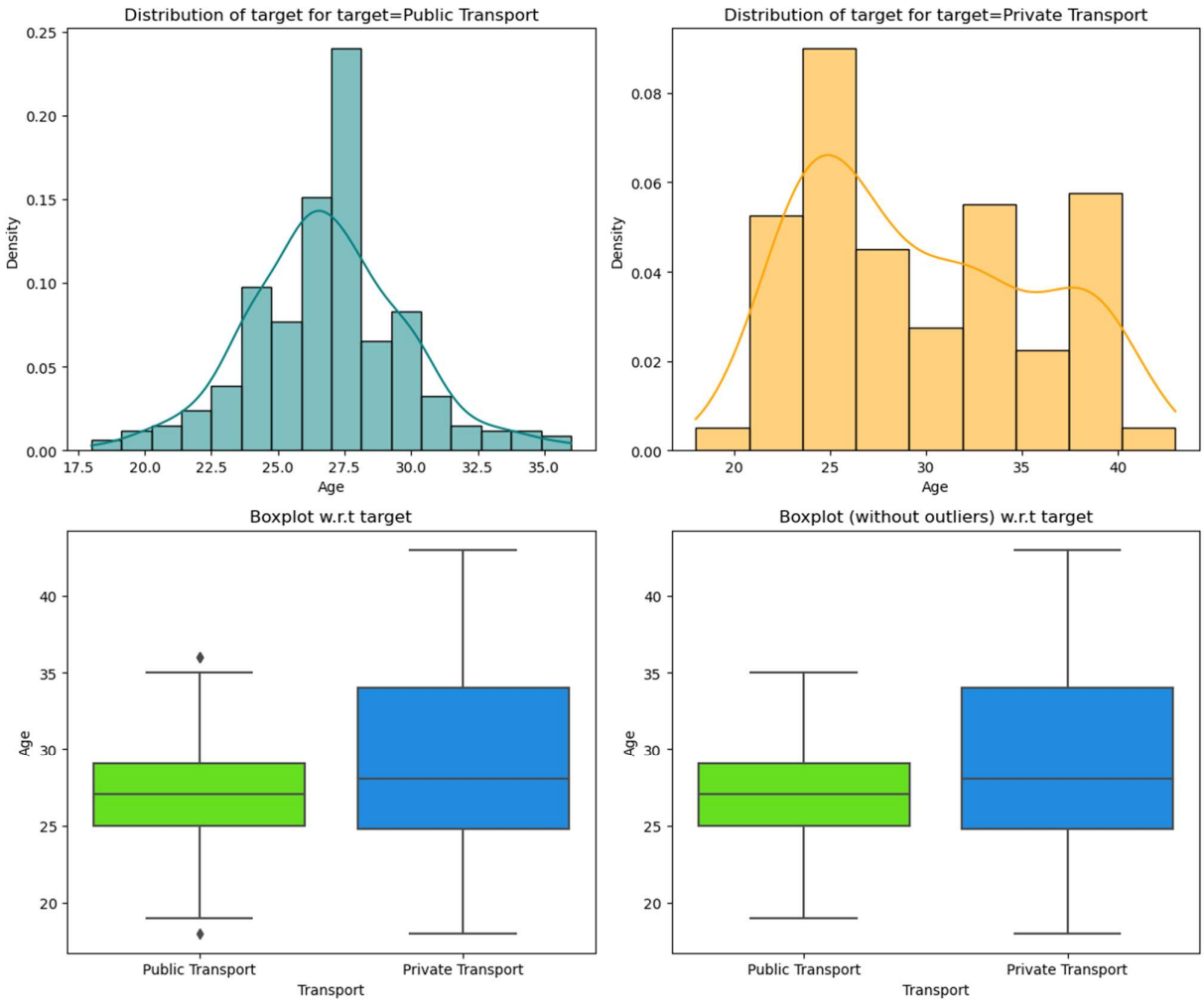

Figure 15: Prediction with Age w.r.t Transport

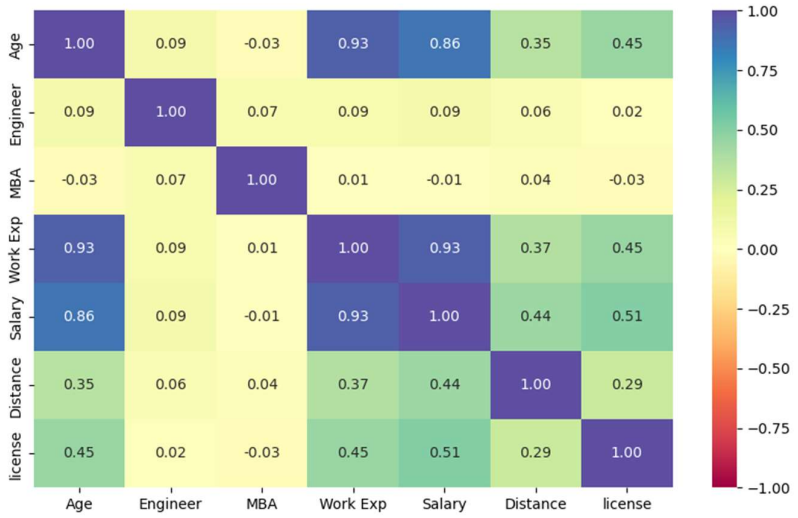

Figure 16: Correlation of the dataset of the dataset

From the above analysis it can summarized as follows,

- Most of employees prefer public transport than private transport.
- Male prefer public transport more compared to female employees.
- The preference for Private transport increases as the age and work experience increases.
- The preference for Public transport is more with less working experience employees.

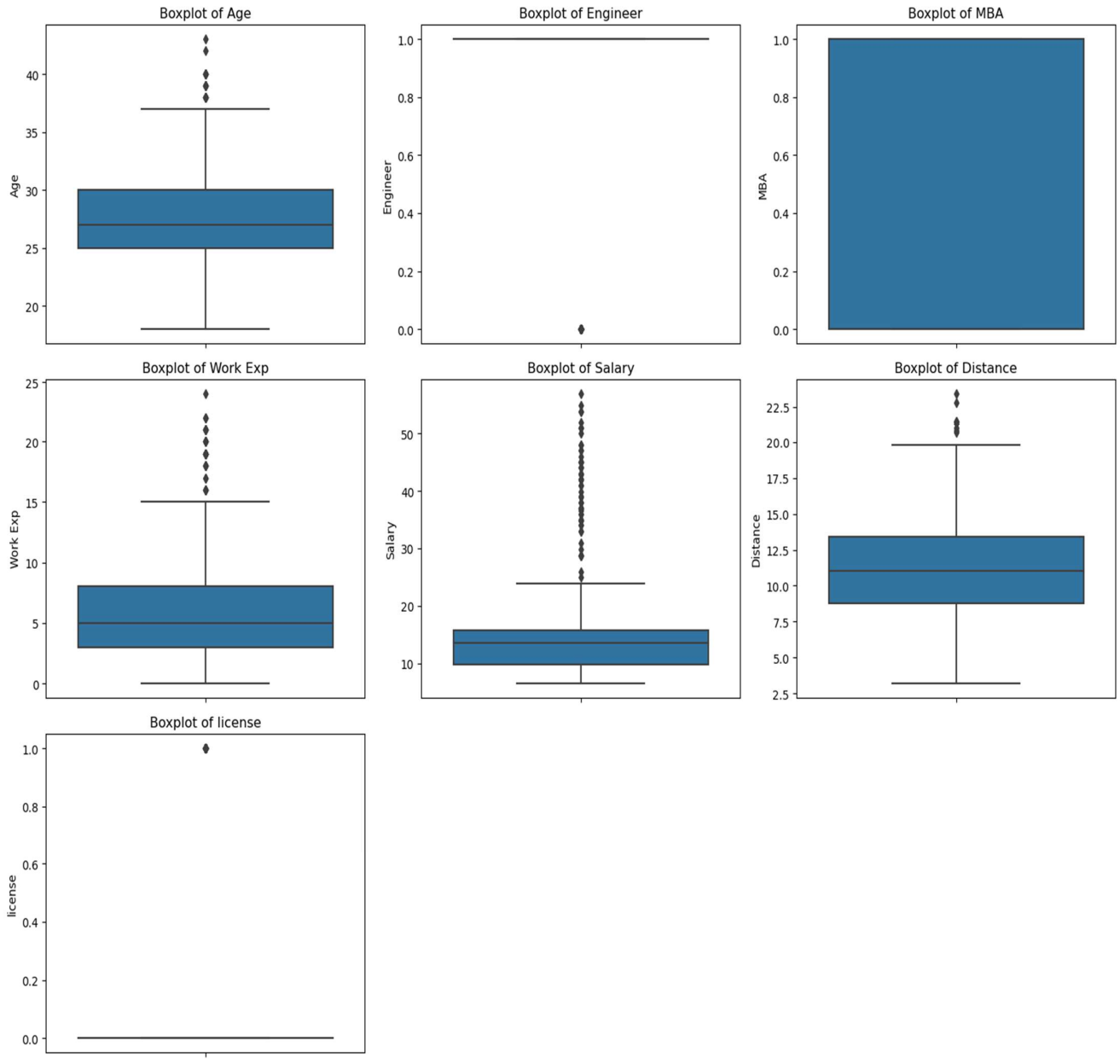

Figure 17: Before Outlier Treatment

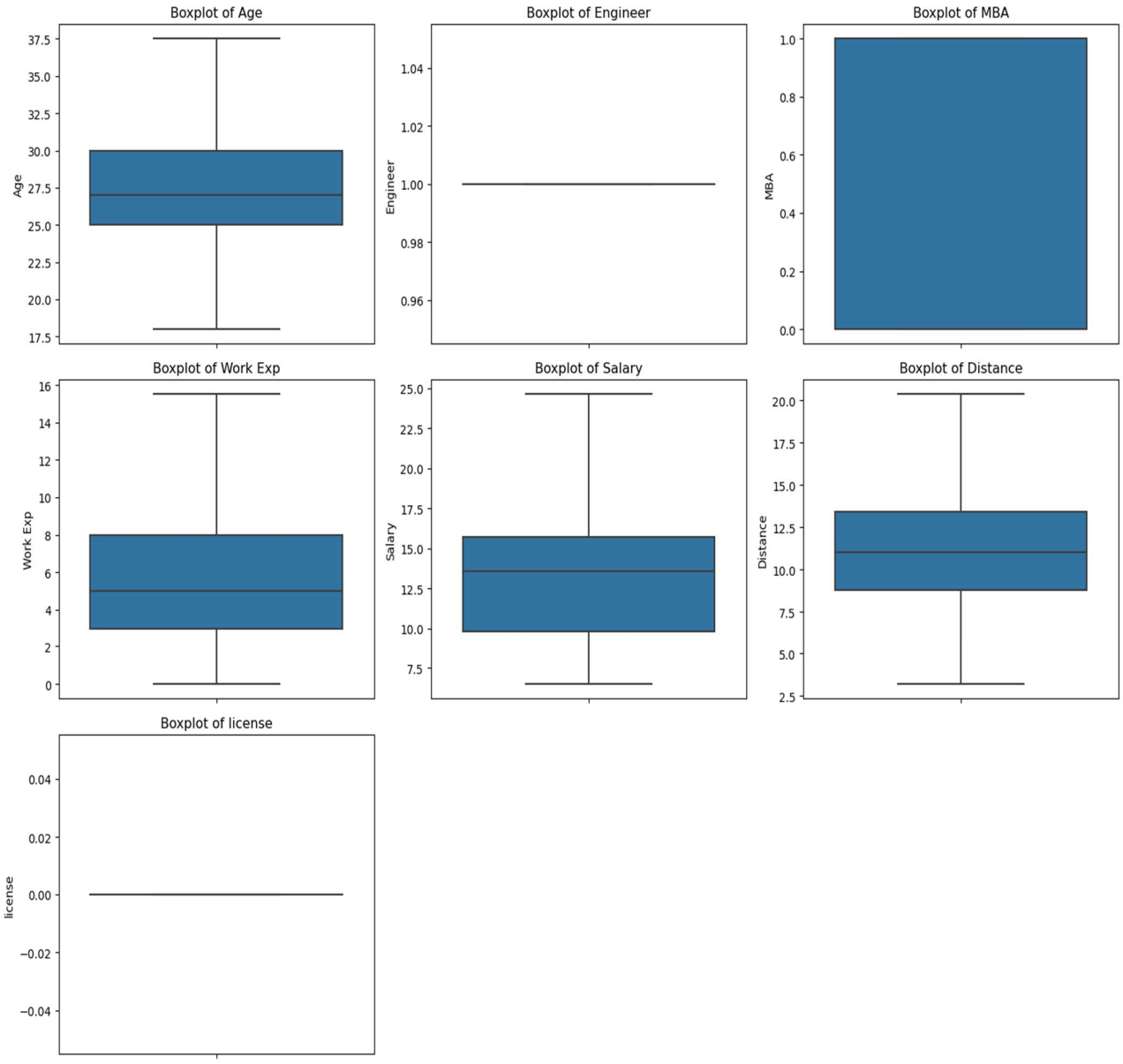

Figure 18: After Outlier Treatment

Model Building, Model Performance evaluation & Model Performance improvement: -

As there are two categorical variables ie, Gender & Transport, they are encoded in the following way,

0 for Female & 1 for Male in gender and 0 for Private transport & 1 for Public transport.

'Transport' variable has taken has a y variable (dependent variable) and all other variables are taken has x variable (independent variable).

The dataset has been split into training & testing dataset with 70:30 ratio. 70% as training dataset & 30% as testing dataset.

```
Shape of Training set :  (310, 8)
Shape of test set :  (134, 8)
Percentage of classes in training set:
Transport
1    0.674194
0    0.325806
Name: proportion, dtype: float64
Percentage of classes in test set:
Transport
1    0.679104
0    0.320896
Name: proportion, dtype: float64
```

Figure 19: Shape of training & testing dataset

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license
296	30.0	1	1.0	1.0	8.0	14.7000	12.6	0.0
404	31.0	1	1.0	0.0	8.0	15.9000	16.4	0.0
84	30.0	0	1.0	0.0	8.0	14.6000	8.1	0.0
424	37.5	1	1.0	0.0	15.5	24.6125	18.1	0.0
432	33.0	1	1.0	1.0	10.0	17.0000	19.1	0.0

Figure 20: Training Dataset

152	26.0	1	1.0	0.0	2.0	9.6000	9.5	0.0
6	28.0	1	1.0	0.0	5.0	14.4000	5.1	0.0
109	24.0	1	1.0	0.0	6.0	12.7000	8.7	0.0
434	37.5	1	1.0	0.0	15.5	24.6125	19.8	0.0
367	32.0	0	1.0	1.0	10.0	15.8000	14.6	0.0

Figure 21: Testing Dataset

Models: -

Bagging Classifier: -

	Accuracy	Recall	Precision	F1
0	0.993548	0.990431	1.0	0.995192

Figure 22: Classification Report of Training Dataset

	Accuracy	Recall	Precision	F1
0	0.783582	0.857143	0.829787	0.843243

Figure 23: Classification Report of Testing Dataset

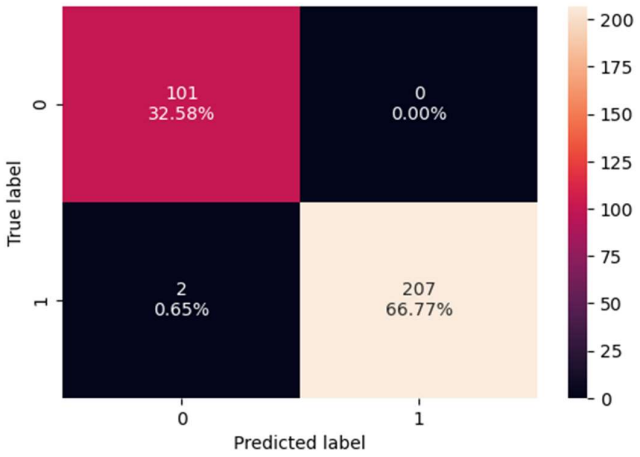

Figure 24: Confusion Matrix of training Data Set

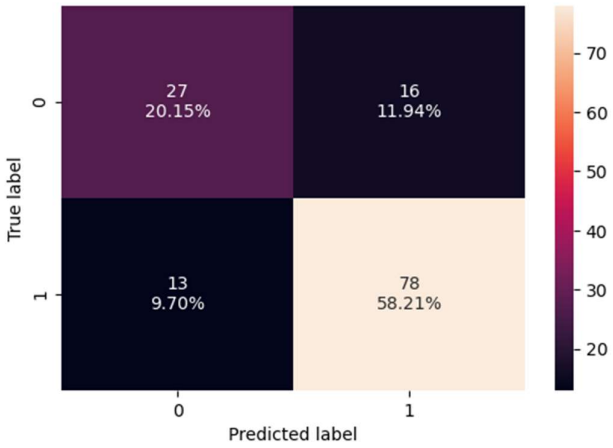

Figure 25: Confusion Matrix of testing Data Set

Inference from Train data:

- 207 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 101 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 2 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 0 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 78 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 27 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 16 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 13 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Hyperparameter Tuning – Bagging Classifier:

	Accuracy	Recall	Precision	F1
0	0.974194	0.995215	0.967442	0.981132

Figure 26: Classification Report of Training Dataset

	Accuracy	Recall	Precision	F1
0	0.798507	0.934066	0.801887	0.862944

Figure 27: Classification Report of Testing Dataset

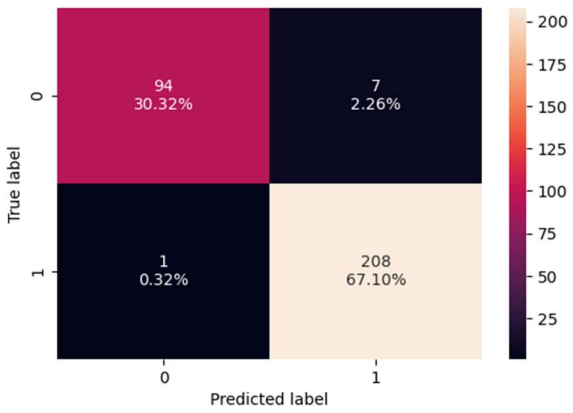

Figure 28: Confusion Matrix of training Data Set

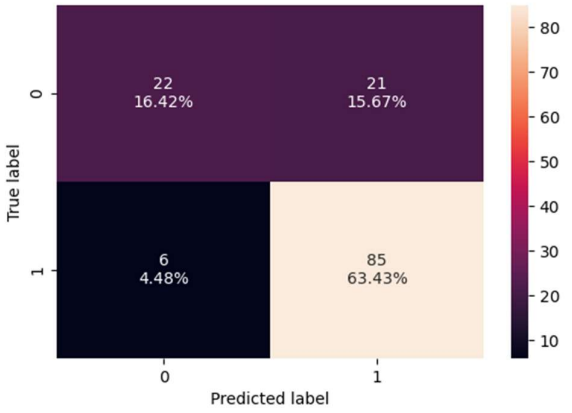

Figure 29: Confusion Matrix of testing Data Set

Inference from Train data:

- 208 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 94 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 7 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 1 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 85 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 22 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 21 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 6 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Random Forest:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Figure 30: Classification Report of Training Dataset

	Accuracy	Recall	Precision	F1
0	0.80597	0.923077	0.815534	0.865979

Figure 31: Classification Report of Testing Dataset

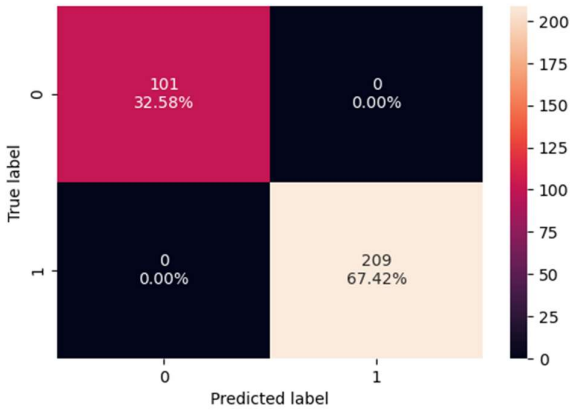

Figure 32: Confusion Matrix of training Data Set

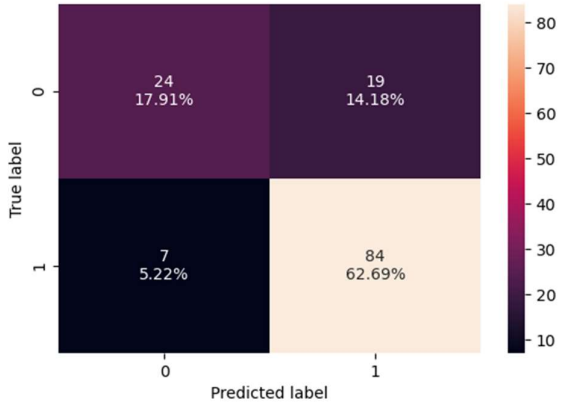

Figure 33: Confusion Matrix of testing Data Set

Inference from Train data:

- 209 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 101 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 0 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 0 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 84 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 24 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport

- 19 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 7 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Hyperparameter Tuning – Random Forest: -

	Accuracy	Recall	Precision	F1
0	0.987097	0.990431	0.990431	0.990431

Figure 34: Classification Report of Training Dataset

	Accuracy	Recall	Precision	F1
0	0.798507	0.934066	0.801887	0.862944

Figure 35: Classification Report of Testing Dataset

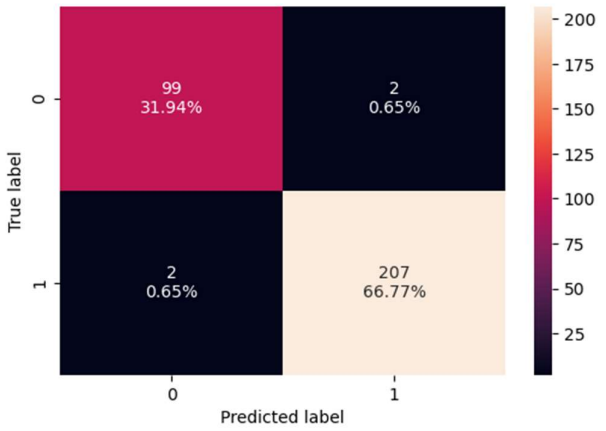

Figure 36: Confusion Matrix of training Data Set

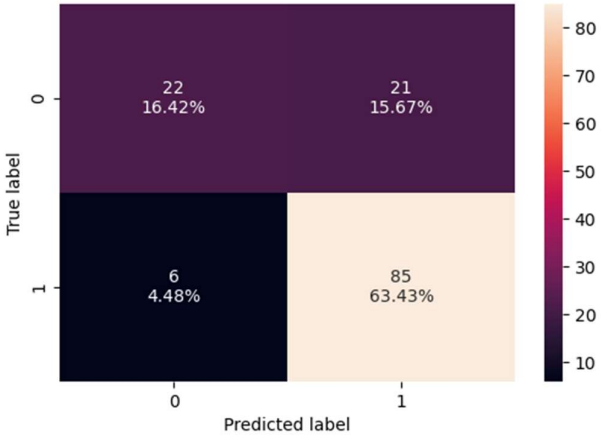

Figure 37: Confusion Matrix of testing Data Set

Inference from Train data:

- 207 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 99 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 2 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 2 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 85 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 22 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 21 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 6 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

AdaBoost Classifier: -

	Accuracy	Recall	Precision	F1
0	0.854839	0.952153	0.850427	0.89842

Figure 38: Classification Report of Training Dataset

	Accuracy	Recall	Precision	F1
0	0.776119	0.901099	0.796117	0.845361

Figure 39: Classification Report of Testing Dataset

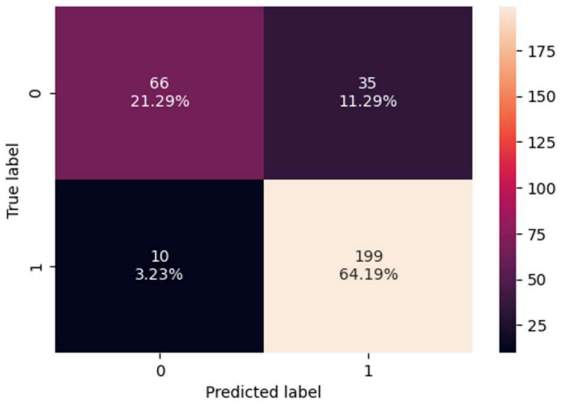

Figure 40: Confusion Matrix of training Data Set

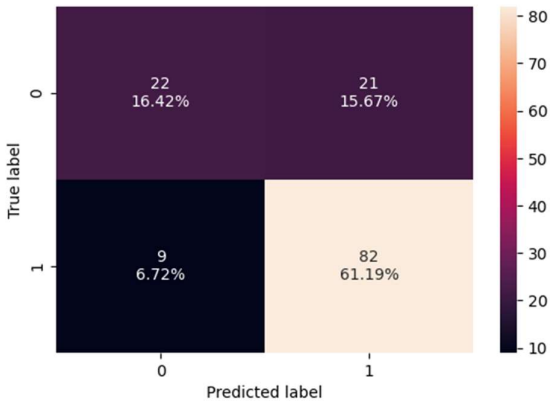

Figure 41: Confusion Matrix of testing Data Set

Inference from Train data:

- 199 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport

- 66 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 35 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 10 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 82 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 22 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 21 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 9 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Hyperparameter Tuning – AdaBoost Classifier: -

	Accuracy	Recall	Precision	F1
0	0.816129	0.976077	0.796875	0.877419

Figure 42: Classification Report of Training Dataset

	Accuracy	Recall	Precision	F1
0	0.768657	0.934066	0.772727	0.845771

Figure 43: Classification Report of Testing Dataset

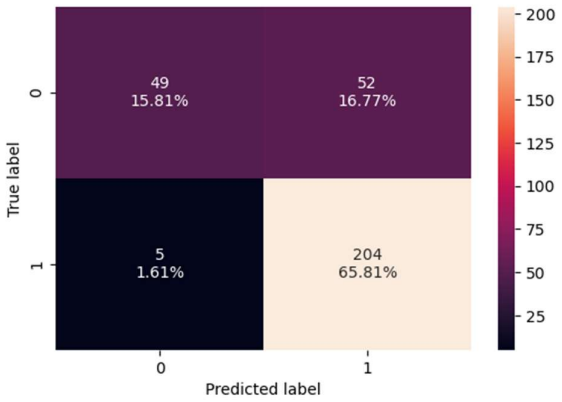

Figure 44: Confusion Matrix of training Data Set

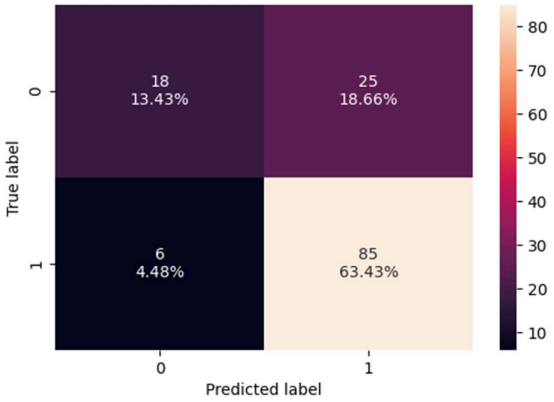

Figure 45: Confusion Matrix of testing Data Set

Inference from Train data:

- 204 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 52 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 49 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 5 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Inference from Test data:

- 85 is True Positive; this denotes cases where the actual calss of the data point and the predicted is same; transport for public transport is predicted as public transport
- 25 is True Negative; this denotes cases where the actual calss of the data point and the predicted is same; transport for private transport is predicted as private transport
- 18 is False Positive; this denotes cases where actual class was negative (0) but predicted as positive (1)
- 6 is False Negative; this denotes cases where actual class was positive (1) but predicted as negative (0)

Model Comparison:

Bagging Classifier, Tuned Bagging Classifier, Random Forest, Tuned Random Forest Adaboost Classifier & Tuned Adaboost Classifier models are thoroughly explained in the before sections. We are here to compare 6the all 4 models and identify which make more sense with respect you predicting dependent variable (Transport).

Training performance comparison:

	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier
Accuracy	0.993548	0.974194	1.0	0.987097	0.854839	0.816129
Recall	0.990431	0.995215	1.0	0.990431	0.952153	0.976077
Precision	1.000000	0.967442	1.0	0.990431	0.850427	0.796875
F1	0.995192	0.981132	1.0	0.990431	0.898420	0.877419

Figure 46 : Model Comparison for training data

Training performance comparison:

	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier
Accuracy	0.783582	0.798507	0.805970	0.798507	0.776119	0.768657
Recall	0.857143	0.934066	0.923077	0.934066	0.901099	0.934066
Precision	0.829787	0.801887	0.815534	0.801887	0.796117	0.772727
F1	0.843243	0.862944	0.865979	0.862944	0.845361	0.845771

Figure 47 : Model Comparison for testing data

From the above two model comparisions it helps us to understand how each models came out with the important component like AUC, Accuracy, precision, recall, f1-score. Bagging Classifier, Tuned Bagging Classifier, Random Forest, Tuned Random Forest Adaboost Classifier & Tuned Adaboost Classifier performed on a same level predicting the dependent variable, but when it is compared with Random Forest it shows performance in both train and test data.

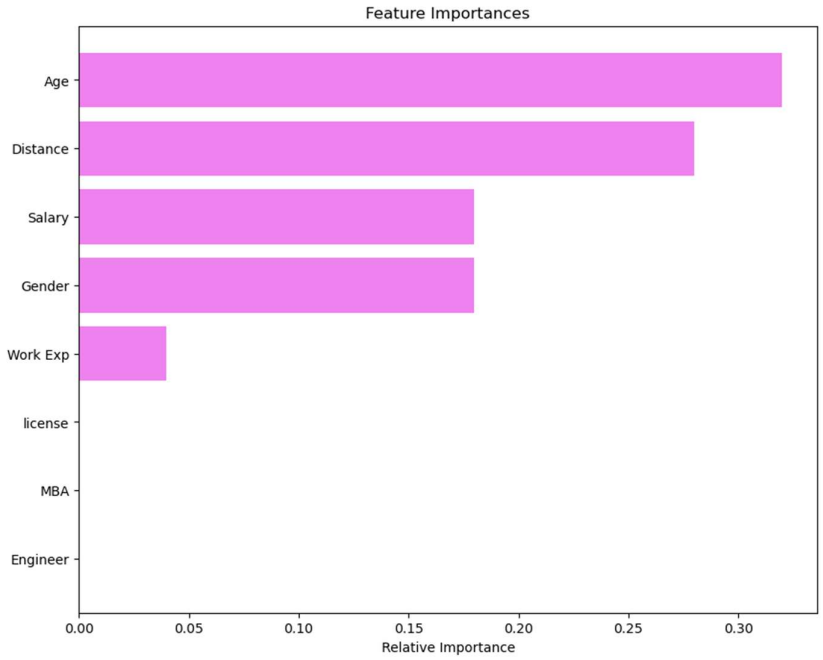

Figure 48 : Important features of final model

Observations

Comparing all the performance measure, Random Forest model is performing the best. Although there are some other models which is performing almost same as that of Random Forest. But AdaBoost Classifier is very consistent when train and test results are compared with each other. Along with other parameters such as Recall value, Precision etc, those results were pretty good is this model.

- Over 60% of workers say they would rather use public transportation than their own vehicle.
- When it comes to public transportation, men prefer it over women.
- As one gets older and has more work experience, the preference for private transportation grows.
- It is more common for employees with less job experience to favor public transportation.

Figure 46 shows that "Age" is the most crucial factor in determining the chosen mode of transportation. This makes sense because experience increases with age. Thus, it can be said that workers in higher positions have a preference for private transportation, whereas those with less and intermediate experience have a preference for public transportation.

Problem 2: -

Context

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

Solution:-

Figure 49: First 5 rows of the data

Figure 50: Last 5 rows of the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 495 entries, 0 to 494
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   deal                   495 non-null    bool
 1   description            495 non-null    object
 2   episode                495 non-null    int64
 3   category               495 non-null    object
 4   entrepreneurs          423 non-null    object
 5   location               495 non-null    object
 6   website                457 non-null    object
 7   askedFor               495 non-null    int64
 8   exchangeForStake       495 non-null    int64
 9   valuation              495 non-null    int64
 10  season                 495 non-null    int64
 11  shark1                 495 non-null    object
 12  shark2                 495 non-null    object
 13  shark3                 495 non-null    object
 14  shark4                 495 non-null    object
 15  shark5                 495 non-null    object
 16  title                  495 non-null    object
 17  episode-season         495 non-null    object
 18  Multiple Entreprenuers 495 non-null    bool
dtypes: bool(2), int64(5), object(12)
memory usage: 66.8+ KB
```

Figure 51: Data types of the dataset

- There are null values and duplicates and dropped the missing values & duplicates.

	deal	description
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
5	True	One of the first entrepreneurs to pitch on Sha...
...	...	...
490	True	Zoom Interiors is a virtual service for interi...
491	True	Spikeball started out as a casual outdoors gam...
492	True	Shark Wheel is out to literally reinvent the w...
493	False	Adriana Montano wants to open the first Cat Ca...
494	True	Sway Motorsports makes a three-wheeled, all-el...

387 rows × 2 columns

Figure 52: Deal & Description in separate dataset

- Dropped the column "deal" because it is not needed for additional analysis and divided the corpor into secured and non-secured groups.

	description
1	Retail and wholesale pie factory with two reta...
2	Ava the Elephant is a godsend for frazzled par...
5	One of the first entrepreneurs to pitch on Sha...
12	A line of books written to help children find ...
16	Coverplay is a slipcover for children's play y...

Figure 53: Deals Secured

	description
3	Organizing, packing, and moving services deliv...
4	Interactive media centers for healthcare waiti...
6	A mixed martial arts clothing line looking to ...
7	Attach Noted is a detachable "arm" that holds ...
8	A safety device for seatbelts. It prevents the...

Figure 54: Deals Not Secured

```
True Corpus 50302
False Corpus 34899
```

Figure 55: No. of characters in each corpus

- Removal of hhtp links from bot secured & not secured deals

	description	char_count
1	Retail and wholesale pie factory with two reta...	73
2	Ava the Elephant is a godsend for frazzled par...	244
5	One of the first entrepreneurs to pitch on Sha...	365
12	A line of books written to help children find ...	57
16	Coverplay is a slipcover for children's play y...	722

	description	char_count
3	Organizing, packing, and moving services deliv...	68
4	Interactive media centers for healthcare waiti...	112
6	A mixed martial arts clothing line looking to ...	110
7	Attach Noted is a detachable "arm" that holds ...	91
8	A safety device for seatbelts. It prevents the...	111

Figure 56: After removing http links

- After De-contraction of words

	description	char_count
1	Retail and wholesale pie factory with two reta...	73
2	Ava the Elephant is a godsend for frazzled par...	244
5	One of the first entrepreneurs to pitch on Sha...	365
12	A line of books written to help children find ...	57
16	Coverplay is a slipcover for children's play y...	722

Figure 57: After words de-contraction of deals secured

	description	char_count
3	Organizing, packing, and moving services deliv...	68
4	Interactive media centers for healthcare waiti...	112
6	A mixed martial arts clothing line looking to ...	110
7	Attach Noted is a detachable "arm" that holds ...	91
8	A safety device for seatbelts. It prevents the...	111

Figure 58: After words de-contraction of deals not secured

- After removal of numbers

	description	char_count	char
1	Retail and wholesale pie factory with two reta...	73	73
2	Ava the Elephant is a godsend for frazzled par...	244	244
5	One of the first entrepreneurs to pitch on Sha...	365	362
12	A line of books written to help children find ...	57	57
16	Coverplay is a slipcover for children's play y...	722	723

Figure 59: After removing numbers of deals secured

	description	char_count	char
3	Organizing, packing, and moving services deliv...	68	68
4	Interactive media centers for healthcare waiti...	112	112
6	A mixed martial arts clothing line looking to ...	110	110
7	Attach Noted is a detachable "arm" that holds ...	91	91
8	A safety device for seatbelts. It prevents the...	111	111

Figure 60: After removing numbers of deals not secured

- After Tokenization of words

	description	char_count	char	char_token
1	[Retail and wholesale pie factory with two ret...	73	73	77
2	[Ava the Elephant is a godsend for frazzled pa...	244	244	251
5	[One of the first entrepreneurs to pitch on Sh...	365	362	369
12	[A line of books written to help children find...	57	57	61
16	[Coverplay is a slipcover for children's play ...	722	723	745

Figure 61: After tokenizing of deals secured

	description	char_count	char	char_token
3	[Organizing, packing, and moving services deli...	68	68	72
4	[Interactive media centers for healthcare wait...	112	112	116
6	[A mixed martial arts clothing line looking to...	110	110	114
7	[Attach Noted is a detachable "arm" that holds...	91	91	95
8	[A safety device for seatbelts., It prevents t...	111	111	118

Figure 62: After tokenizing of deals not secured

- After converting all words into lowercase.

	description	char_count	char	char_token	char_lower
1	[retail and wholesale pie factory with two ret...	73	73	77	77
2	[ava the elephant is a godsend for frazzled pa...	244	244	251	251
5	[one of the first entrepreneurs to pitch on sh...	365	362	369	369
12	[a line of books written to help children find...	57	57	61	61
16	[coverplay is a slipcover for children's play ...	722	723	745	745

Figure 63: After converting into lower case of deals secured

	description	char_count	char	char_token	char_lower
3	[organizing, packing, and moving services deli...	68	68	72	72
4	[interactive media centers for healthcare wait...	112	112	116	116
6	[a mixed martial arts clothing line looking to...	110	110	114	114
7	[attach noted is a detachable "arm" that holds...	91	91	95	95
8	[a safety device for seatbelts., it prevents t...	111	111	118	118

Figure 64: After converting into lower case of deals not secured

- After removing all the punctuation marks

	description	char_count	char	char_token	char_lower	char_punc
1	[retail and wholesale pie factory with two ret...	73	73	77	77	76
2	[ava the elephant is a godsend for frazzled pa...	244	244	251	251	249
5	[one of the first entrepreneurs to pitch on sh...	365	362	369	369	356
12	[a line of books written to help children find...	57	57	61	61	60
16	[coverplay is a slipcover for childrens play y...	722	723	745	745	726

Figure 65: After removing punctuations of deals secured

	description	char_count	char	char_token	char_lower	char_punc
3	[organizing packing and moving services delive...	68	68	72	72	69
4	[interactive media centers for healthcare wait...	112	112	116	116	115
6	[a mixed martial arts clothing line looking to...	110	110	114	114	112
7	[attach noted is a detachable arm that holds p...	91	91	95	95	91
8	[a safety device for seatbelts, it prevents th...	111	111	118	118	116

Figure 66: After removing punctuations of deals not secured

- After removing stopwords from the data

	description	char_count	char	char_token	char_lower	char_punc	char_stop
1	[retail and wholesale pie factory with two ret...	73	73	77	77	76	76
2	[ava the elephant is a godsend for frazzled pa...	244	244	251	251	249	249
5	[one of the first entrepreneurs to pitch on sh...	365	362	369	369	356	356
12	[a line of books written to help children find...	57	57	61	61	60	60
16	[coverplay is a slipcover for childrens play y...	722	723	745	745	726	726

Figure 67: After removing stopwords of deals secured

	description	char_count	char	char_token	char_lower	char_punc	char_stop
3	[organizing packing and moving services delive...	68	68	72	72	69	69
4	[interactive media centers for healthcare wait...	112	112	116	116	115	115
6	[a mixed martial arts clothing line looking to...	110	110	114	114	112	112
7	[attach noted is a detachable arm that holds p...	91	91	95	95	91	91
8	[a safety device for seatbelts, it prevents th...	111	111	118	118	116	116

Figure 68: After removing stopwords of deals not secured

- After lemmatization

	description	char_count	char	char_token	char_lower	char_punc	char_stop	char_lemm
1	[retail and wholesale pie factory with two ret...	73	73	77	77	76	76	76
2	[ava the elephant is a godsend for frazzled pa...	244	244	251	251	249	249	249
5	[one of the first entrepreneurs to pitch on sh...	365	362	369	369	356	356	356
12	[a line of books written to help children find...	57	57	61	61	60	60	60
16	[coverplay is a slipcover for childrens play y...	722	723	745	745	726	726	726

Figure 69: After lemmatization of deals secured

- After Normalization

	description	char_count	char	char_token	char_lower	char_punc	char_stop	char_lemm	char_norm
1	retail and wholesale pie factory with two reta...	73	73	77	77	76	76	76	72
2	ava the elephant is a godsend for frazzled par...	244	244	251	251	249	249	249	242
5	one of the first entrepreneurs to pitch on sha...	365	362	369	369	356	356	356	349
12	a line of books written to help children find ...	57	57	61	61	60	60	60	56
16	coverplay is a slipcover for childrens play ya...	722	723	745	745	726	726	726	704

Figure 70: After Normalization of deals secured

Word clouds: -

Figure 71: Wordcloud of deals secured

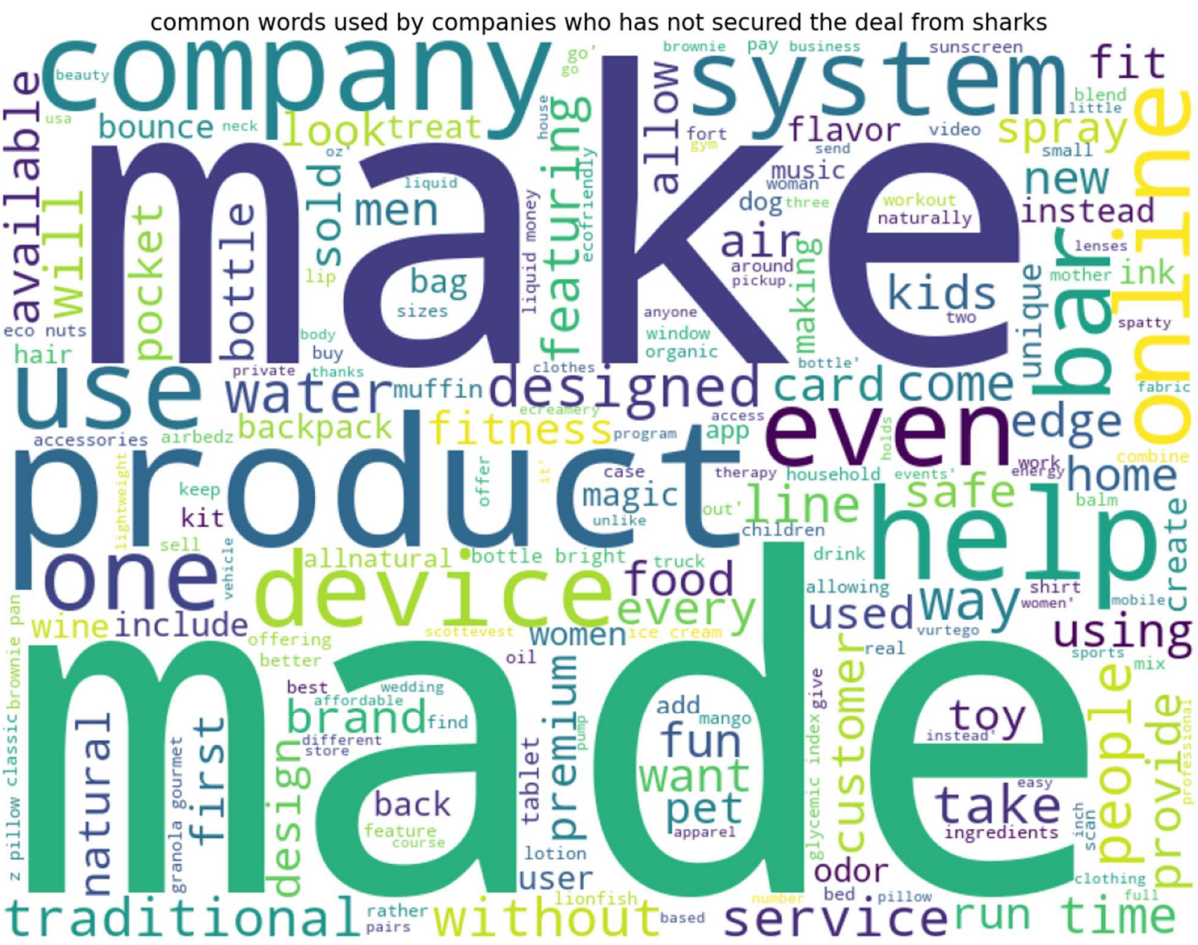

Figure 72: Wordcloud of deals not secured