MACHINE LEARNING BUSINESS REPORT

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

Part 1: Machine Learning Models

You work for an office transport company. You are in discussions with ABC Consulting company for providing transport for their employees. For this purpose, you are tasked with understanding how do the employees of ABC Consulting prefer to commute presently (between home and office). Based on the parameters like age, salary, work experience etc. given in the data set 'Transport.csv', you are required to predict the preferred mode of transport. The project requires you to build several Machine Learning models and compare them so that the model can be finalised.

Column name	Description
Age	Age of the Employee in Years
Gender	Gender of the Employee
Engineer	For Engineer =1 , Non Engineer =0
MBA	For MBA =1 , Non MBA =0
Work Exp	Experience in years
Salary	Salary in Lakhs per Annum
Distance	Distance in Kms from Home to Office
license	If Employee has Driving Licence -1, If not, then 0
Transport	Mode of Transport

Table 1.1: Data Dictionary

Question 1.1: Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.

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
...	...	...	...	...	...	...	...	...	...
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

Table 1.2: Transport Dataset

#	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

Table 1.3:Transport Dataset Info

Columns	Null values present
Age	0
Gender	0
Engineer	0
MBA	0
Work Exp	0
Salary	0
Distance	0
license	0
Transport	0

Table 1.4: Null value check

Observations

- no. of rows: 444
- no. of columns: 9
- From Table 1.3 we Observe No missing Values in the dataset
- From Table 1.4 we Observe No Null values present
- Categorical columns = ['Gender', 'Transport']
- Numerical Columns = ['Age', 'Engineer', 'MBA', 'Work Exp', 'Salary', 'Distance', 'license']
- Let’s try to test whether any categorical attribute contains a “?” in it or not. At times there exists “?” or ” ” in place of missing values. Using the below code snippet we are going to test whether adult_df data frame consists of categorical variables with values as “?”.
- we see Gender : 0, Transport : 0 indicating no “?” or ” ”
- there are no duplicate Rows present in the given Dataset.

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

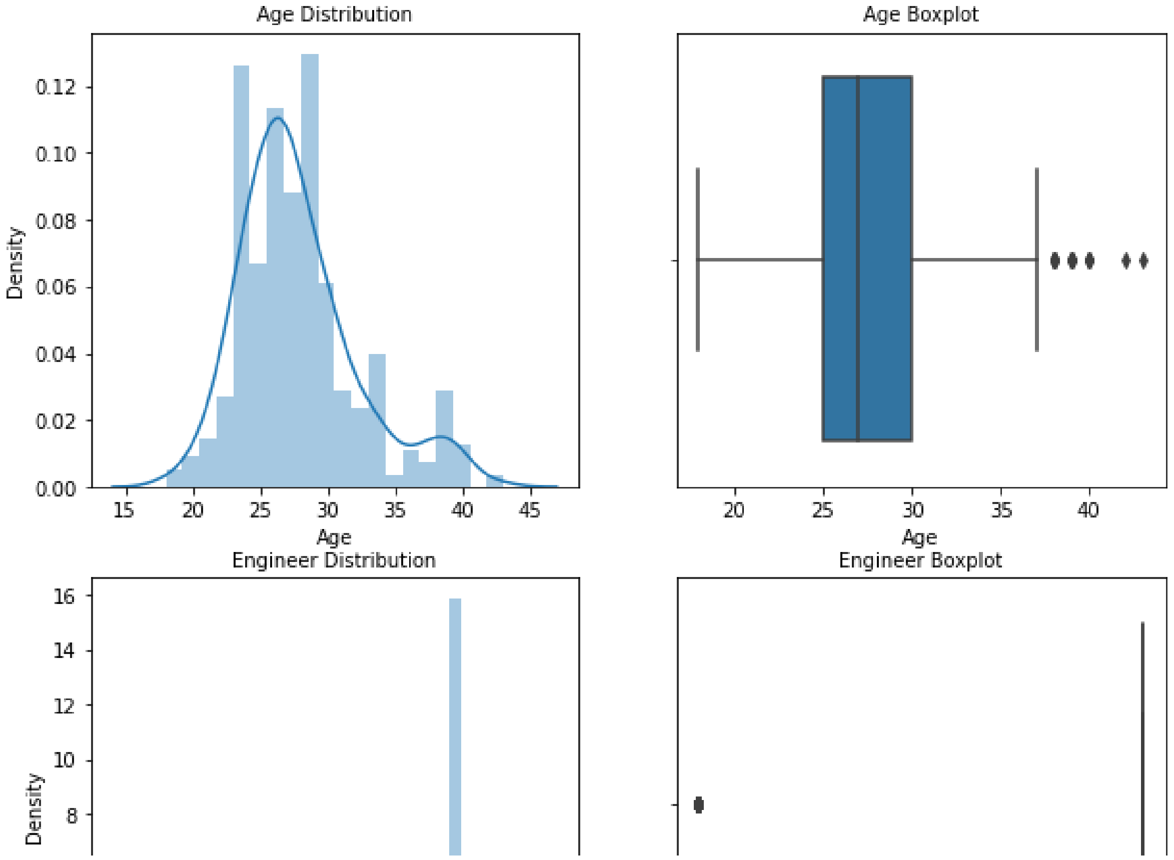
count	mean	std	min	25%	50%	75%	max	
MBA	444.0	0.252252	0.434795	0.0	0.0	0.0	1.000	1.0
Work Exp	444.0	6.299550	5.112098	0.0	3.0	5.0	8.000	24.0
Salary	444.0	16.238739	10.453851	6.5	9.8	13.6	15.725	57.0
Distance	444.0	11.323198	3.606149	3.2	8.8	11.0	13.425	23.4
license	444.0	0.234234	0.423997	0.0	0.0	0.0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

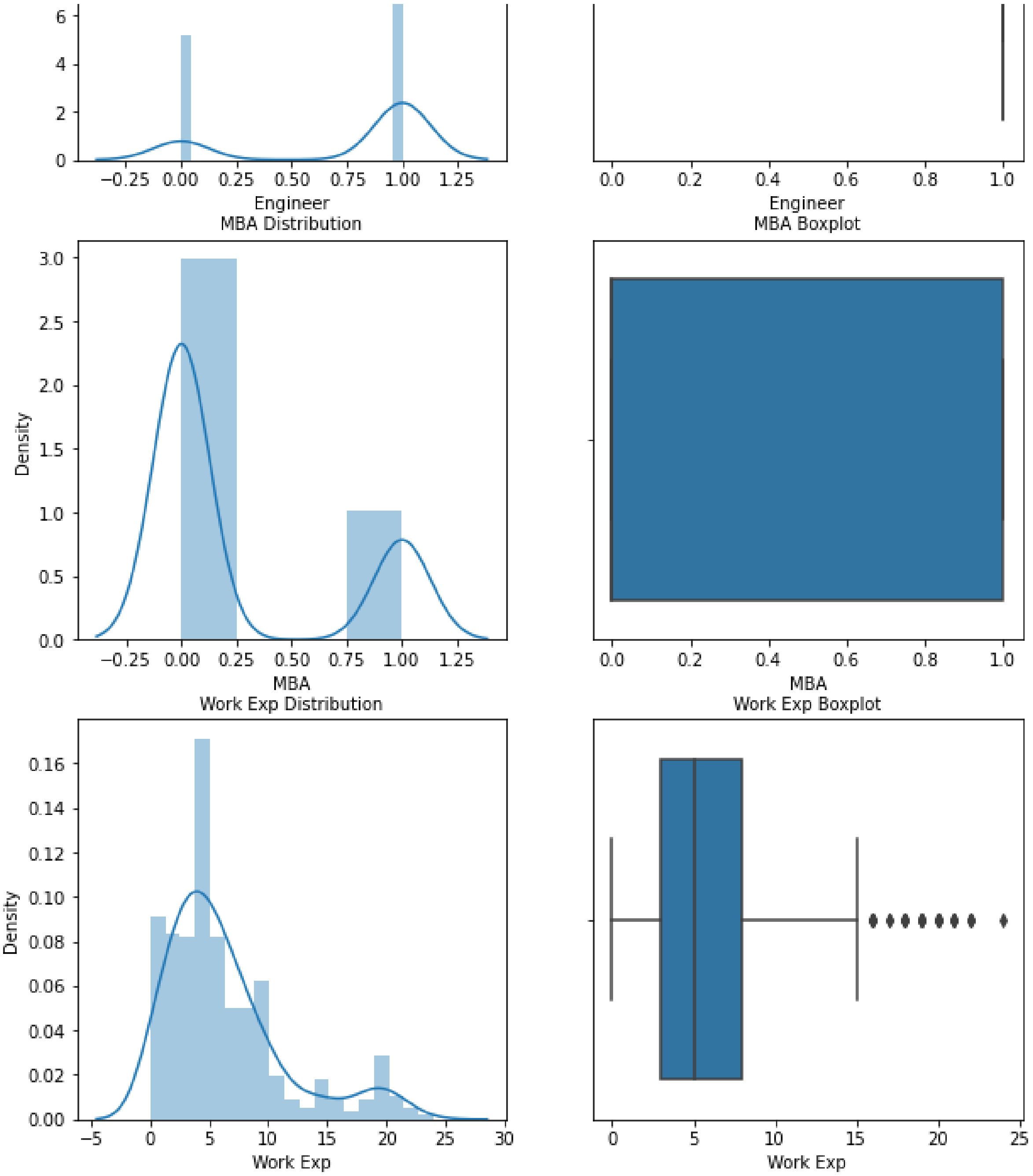
count	unique	top	freq	
Gender	444	2	Male	316
Transport	444	2	Public Transport	300

Table 1.5: Dataset Description

- From the above table we can see that the Average age is 28
- In the given data set, 75% are Engineers and 25% are MBA graduates
- Average work experience of 6 years with minimum years of exp being 0 and max years of exp being 24
- Average Salary earned is 16.23 Lakhs per annum of which 6.5 lakhs per annum being the least and 57 lakhs per annum being the maximum
- Average Distance travelled form Home to Office is 11.32 KM. Minimum distance covered by an emploee is 3.2 KM and maximum distance covered is 23.4 KM
- In the given Dataset Male to female ratio is high. There are 316 Male employees and 128 Female Employees
- We can also see that 300 employees use Public transport and rest 144 use Private transport.

Univariate Analysis




6
4
2
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Engineer
0.0
0.2
0.4
0.6
0.8
1.0
Engineer
MBA Distribution
MBA Boxplot
Density
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
MBA
0.0
0.2
0.4
0.6
0.8
1.0
MBA
Work Exp Distribution
Work Exp Boxplot
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−5
0
5
10
15
20
25
30
Work Exp
0
5
10
15
20
25
Work Exp

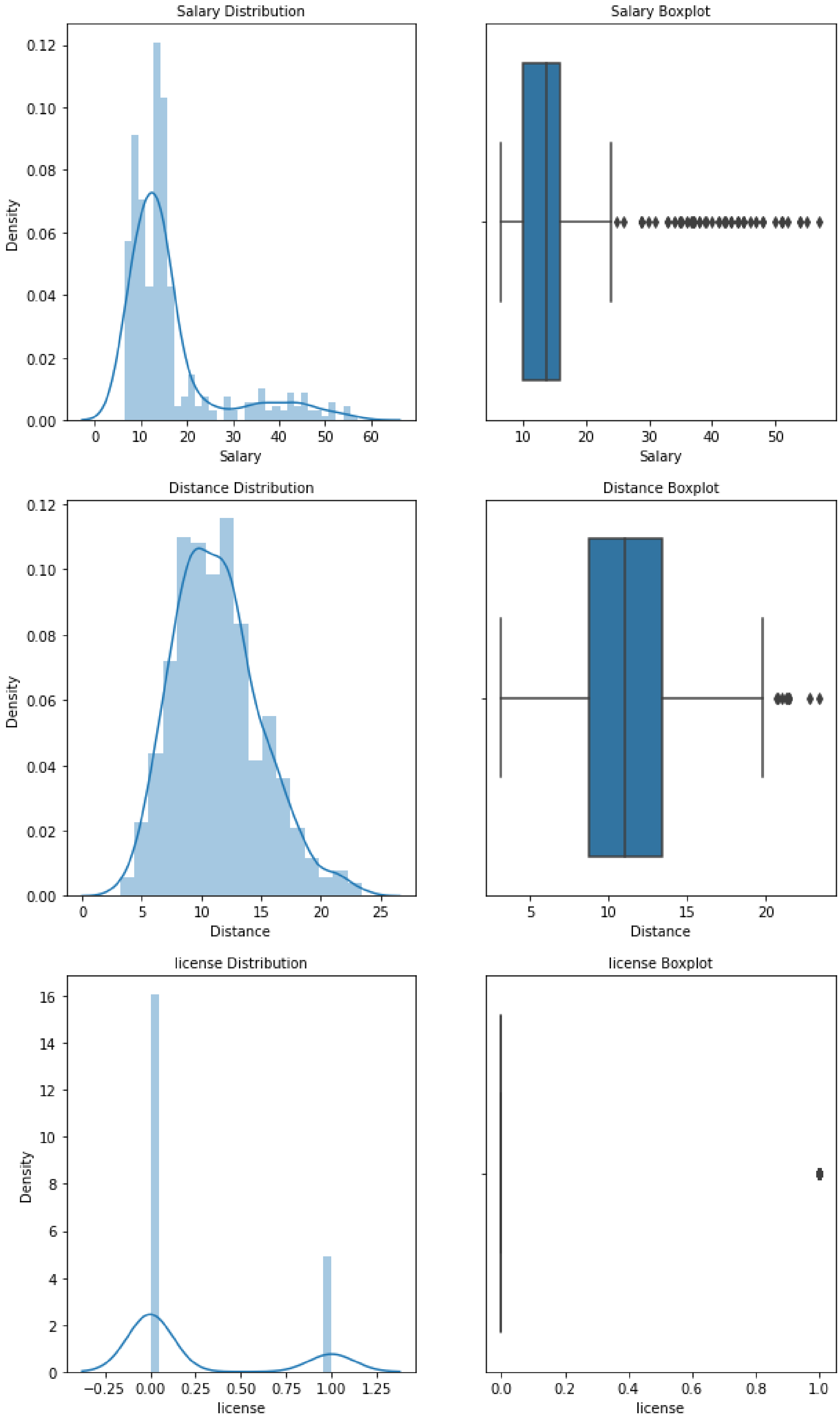

Salary Distribution
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
10
20
30
40
50
60
Salary
Salary Boxplot
10
20
30
40
50
Salary
Distance Distribution
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
5
10
15
20
25
Distance
Distance Boxplot
5
10
15
20
Distance
license Distribution
Density
0
2
4
6
8
10
12
14
16
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
license
license Boxplot
0.0
0.2
0.4
0.6
0.8
1.0
license

Fig 1.1: Univariate analysis on dataset showing Distplot and Histplot of all Numerical Columns

- We see that Age, Work Experience, Salary columns are Right Skewed. Distance travelled has a Normal Distribution.
- All the above 4 columns mentioned have Outliers which will be treated later.

Bivariate Analysis

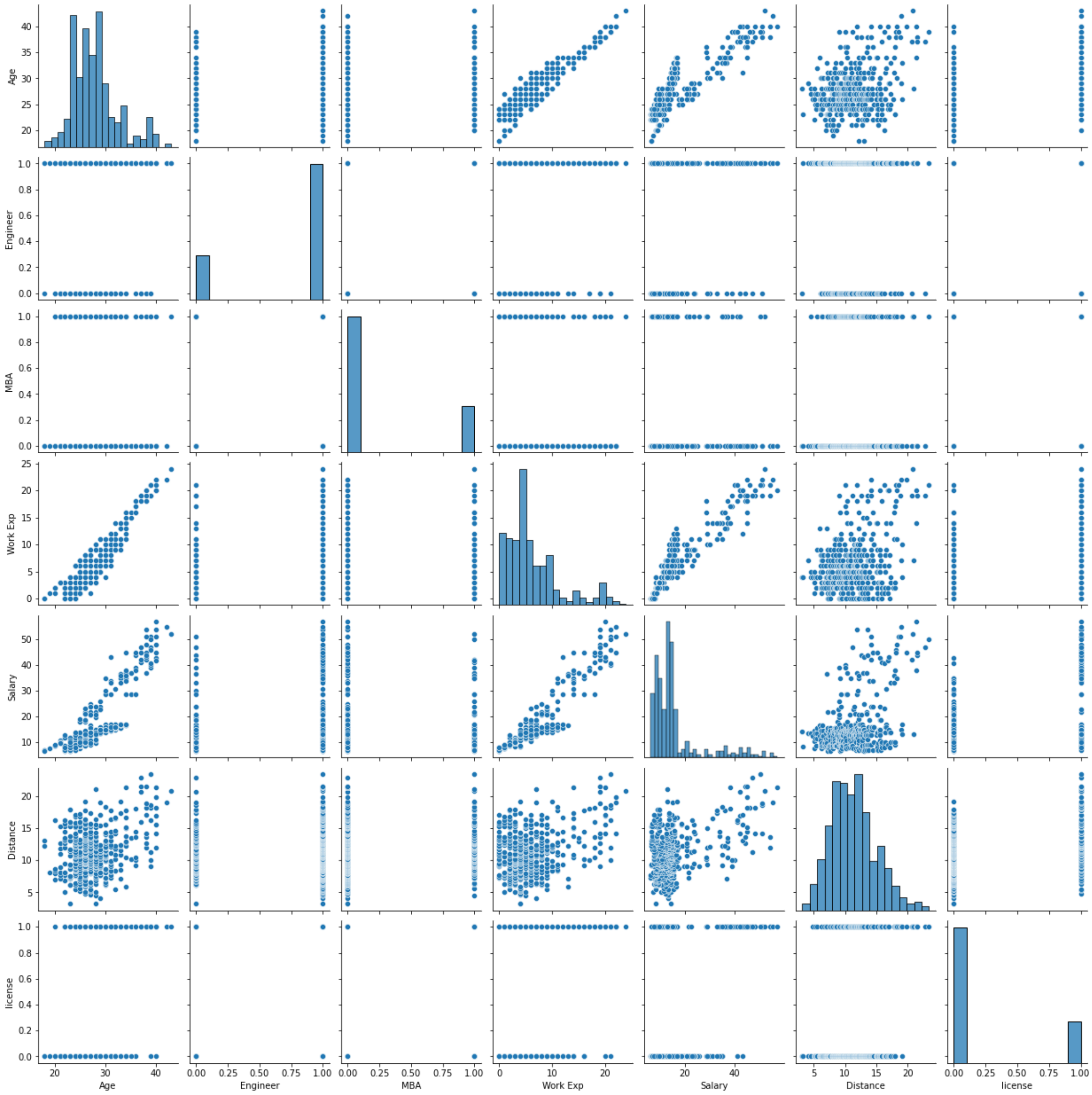


Fig 1.2: Pairplot of Dataset

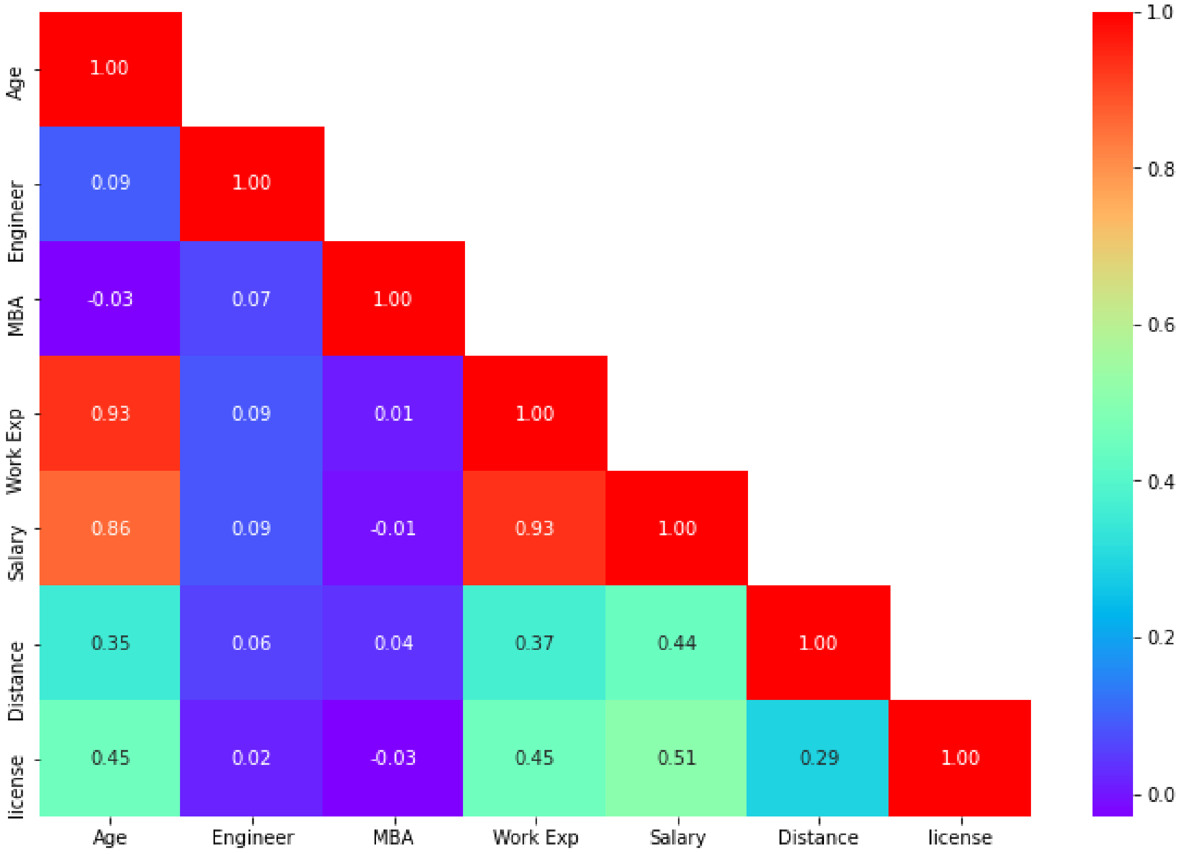


Fig 1.3: Heatmap showing Correlation within Dataset

- We can see that Age is highly Correlated ot Work experience and Salary. Salary is also Highly Correlated to Work expreience
- We see that Engineer column and MBA column have the least correlation with other columns.

Outlier Check

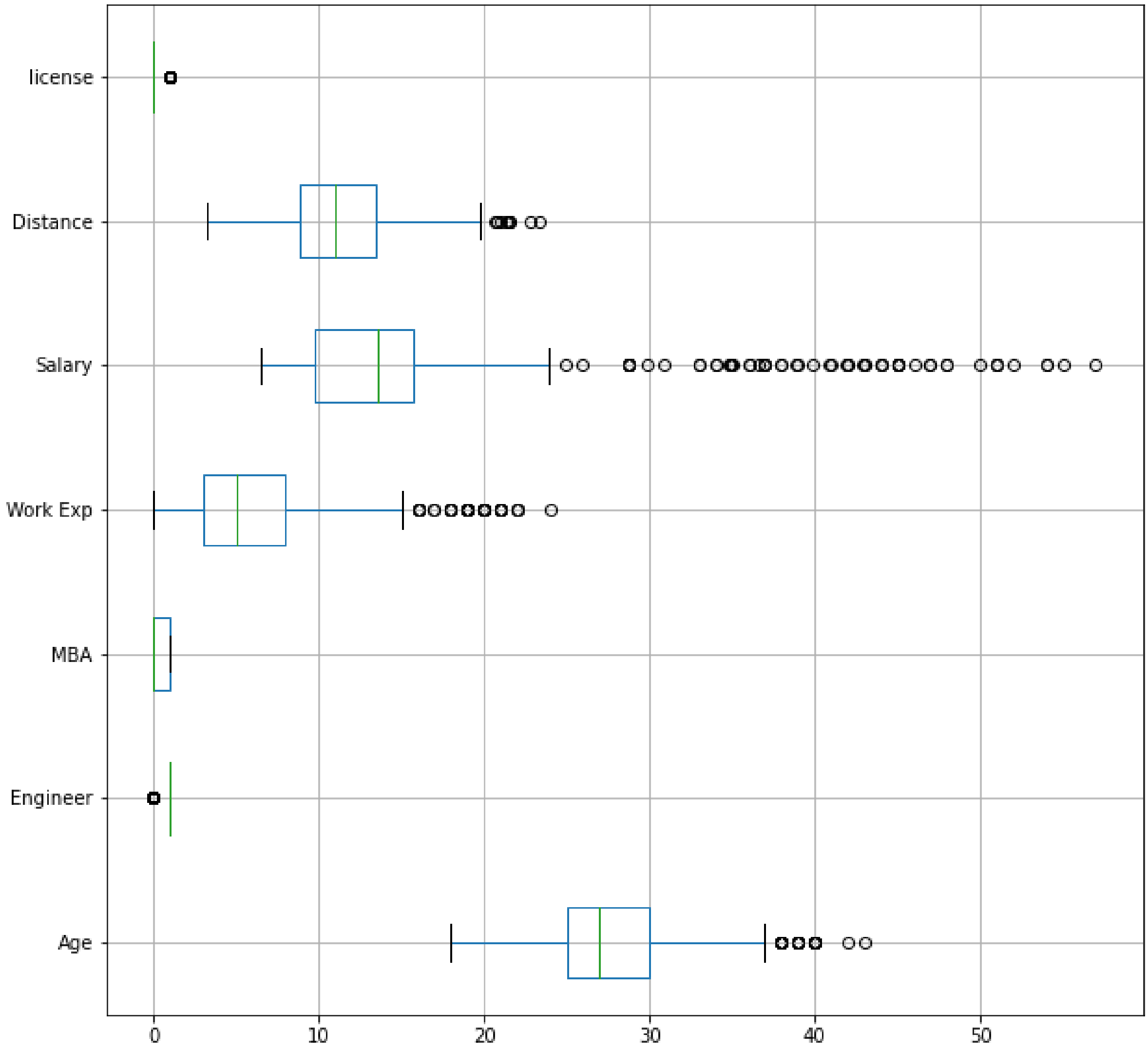


Fig 1.4: Outlier check on Dataset prior to treating it

- We see that Age, Salary , Work Exp & Distance have Outliers present in them.
- In Gaussian Naive Bayes, outliers will affect the shape of the Gaussian distribution and have the usual effects on the mean etc. So depending on our use case, it makes sense to remove outlier .

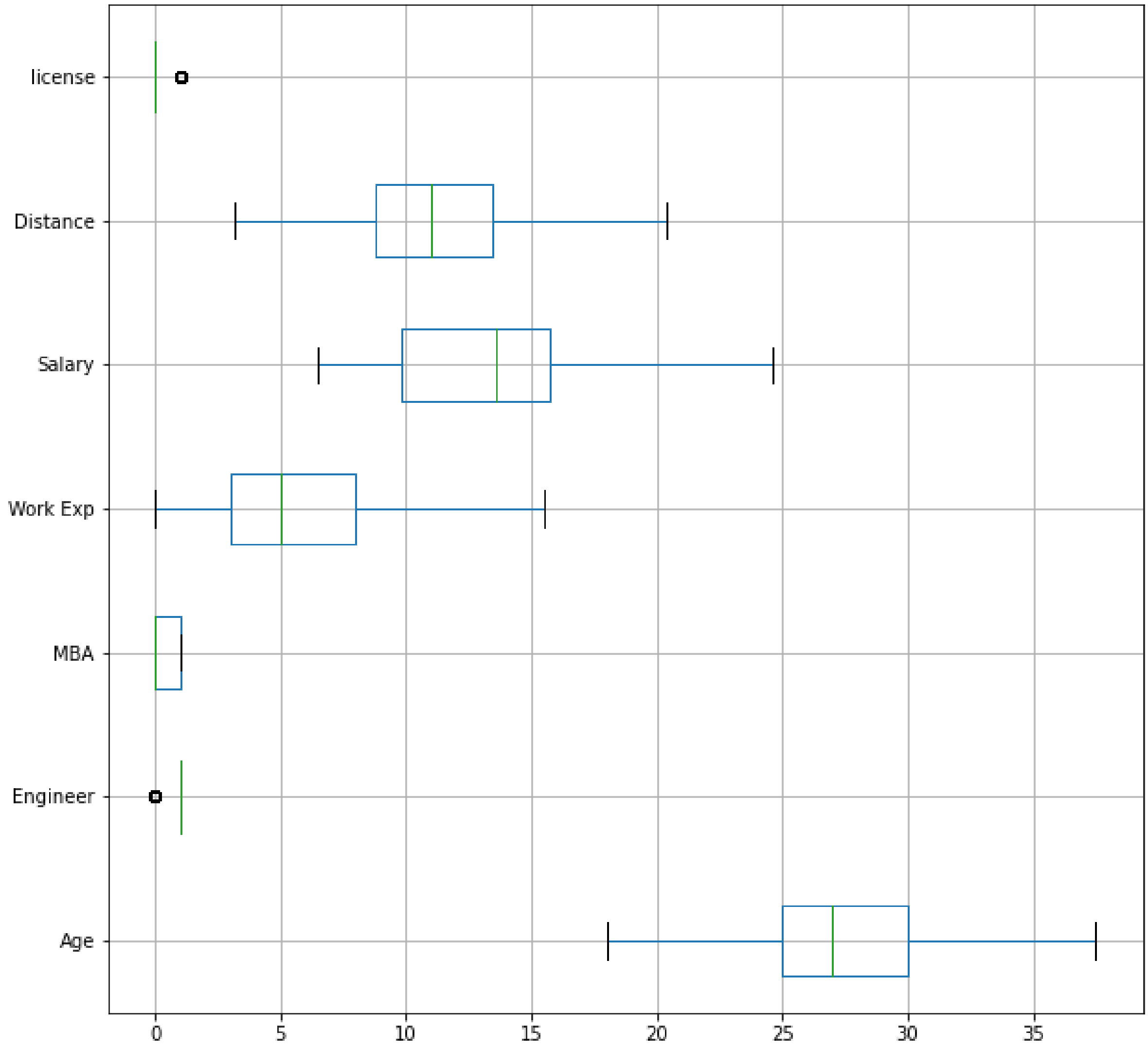


Fig 1.5: Outlier check on Dataset post treating it

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

Scaling

- Scaling is necessary in this case as Dataset has features with different "weights".
- Scaling the variables as continuous variables have different weightage using min-max technique

Age	Engineer	MBA	Work Exp	Salary	Distance	license	Gender_Male	Transport_Public Transport	
0	0.512821	0.0	0.0	0.258065	0.430642	0.000000	0.0	1	1
1	0.256410	1.0	0.0	0.258065	0.099379	0.005827	0.0	0	1
2	0.564103	1.0	0.0	0.451613	0.380952	0.052440	0.0	1	1
3	0.512821	1.0	1.0	0.322581	0.380952	0.075747	0.0	0	1
4	0.461538	1.0	0.0	0.258065	0.380952	0.081573	0.0	1	1

Table 1.6: Scaled Dataset

Data is Split in Training and Testing in the Ratio of 70:30

Question 1.3: Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance.:

a. Logistic Regression Model

b. Linear Discriminant Analysis

c. Decision Tree Classifier – CART model

d. Naïve Bayes Model

e. KNN Model

f. Random Forest Model

g. Boosting Classifier Model using Gradient boost.

Logistic Regression Model

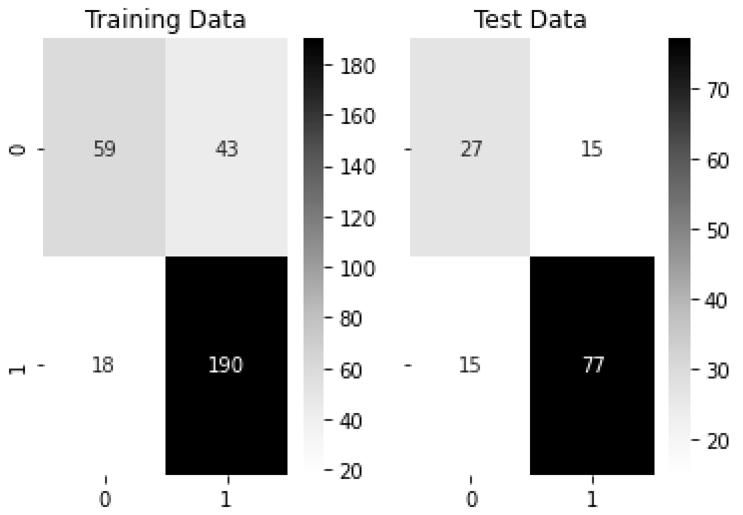


Fig 1.6: Confusion Matrix of Train and Test Dataset for Logistic Regression Model

Classification Report on Training Data for Logistic Regression Model

```
              precision    recall  f1-score   support

           0       0.77      0.58      0.66       102
           1       0.82      0.91      0.86       208

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

Classification Report on Training Data for Logistic Regression Model

```
              precision    recall  f1-score   support

           0       0.64      0.64      0.64        42
           1       0.84      0.84      0.84        92

    accuracy                           0.78       134
   macro avg       0.74      0.74      0.74       134
weighted avg       0.78      0.78      0.78       134
```

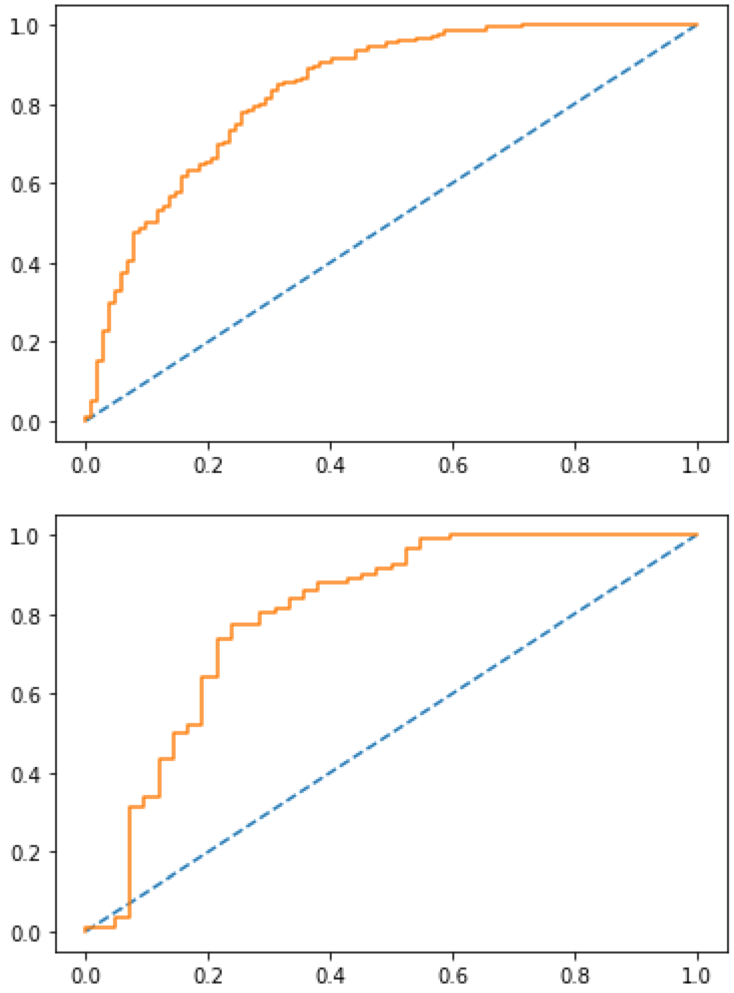


Fig 1.7: ROC Curve of Train and Test Dataset for Logistic Regression Model

LDA Model

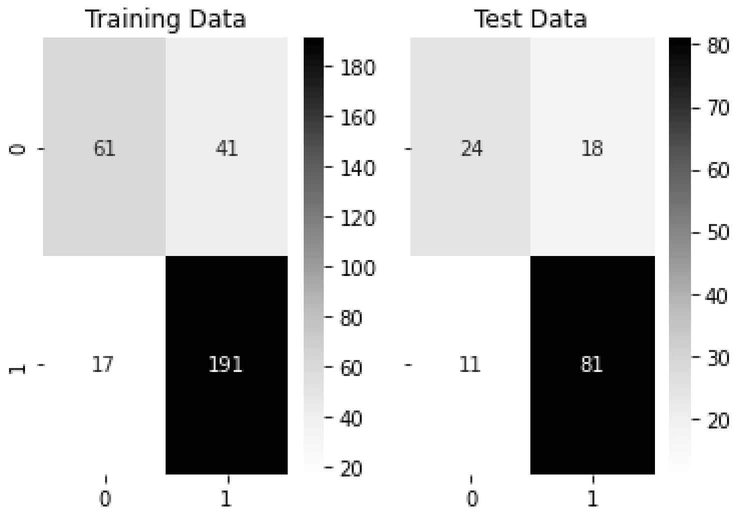


Fig 1.8: Confusion Matrix of Train and Test Dataset for LDA Model

Classification Report on Training Data for LDA Model

```
              precision    recall  f1-score   support

           0       0.78      0.60      0.68       102
           1       0.82      0.92      0.87       208

    accuracy                           0.81       310
   macro avg       0.80      0.76      0.77       310
weighted avg       0.81      0.81      0.81       310
```

Classification Report on Testing Data for LDA Model

```
              precision    recall  f1-score   support

           0       0.69      0.57      0.62        42
           1       0.82      0.88      0.85        92

    accuracy                           0.78       134
   macro avg       0.75      0.73      0.74       134
weighted avg       0.78      0.78      0.78       134
```

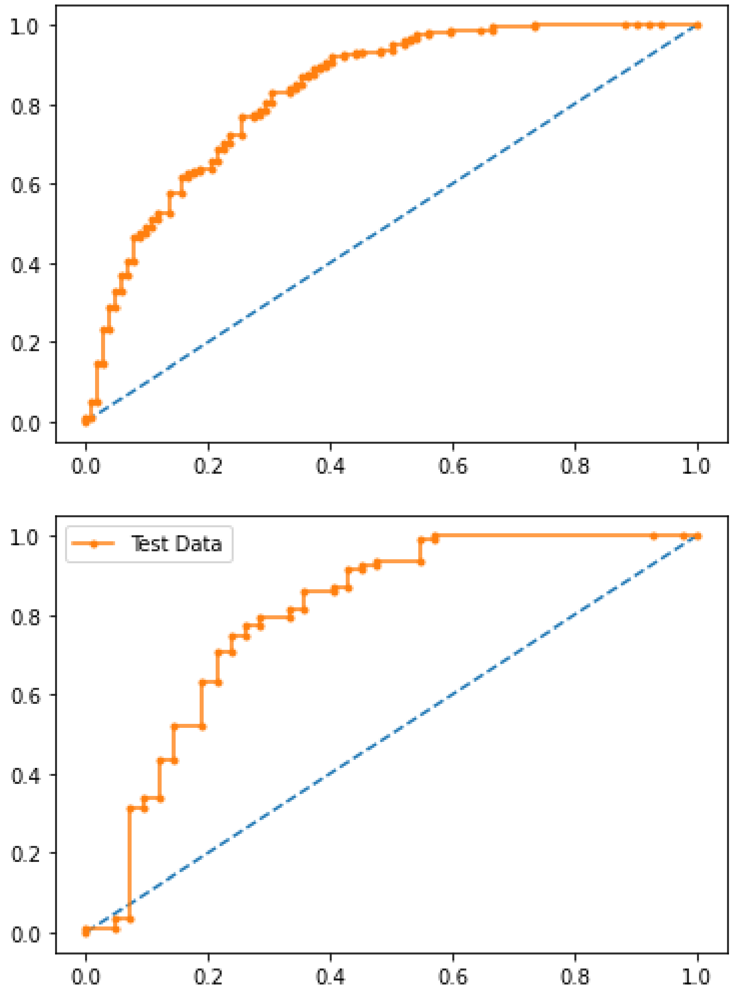


Fig 1.9: ROC Curve of Train and Test Dataset for LDA Model

Decision Tree Classifier – CART model

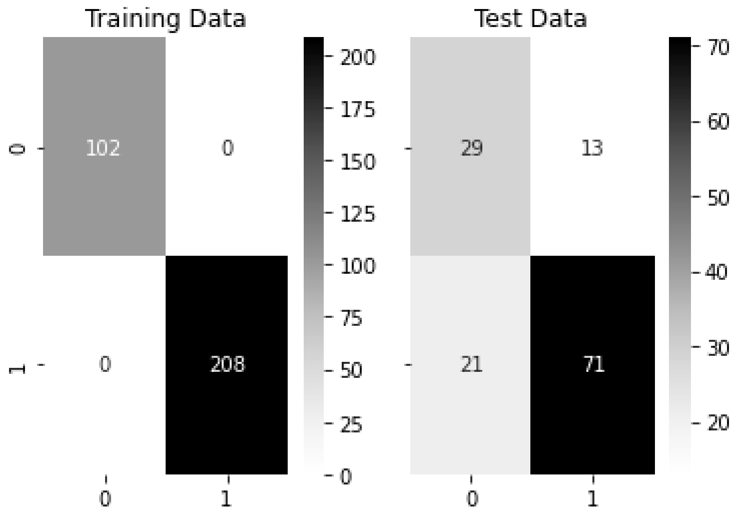


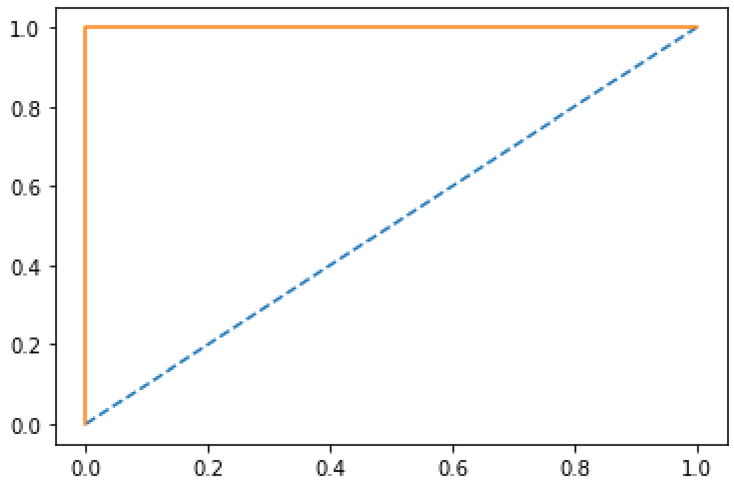
Fig 1.10: Confusion Matrix of Train and Test Dataset for CART model

Classification Report on Training Data for CART Model

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       102
           1       1.00      1.00      1.00       208

    accuracy                           1.00       310
   macro avg       1.00      1.00      1.00       310
weighted avg       1.00      1.00      1.00       310
```

Classification Report on Testing Data for CART Model

```
              precision    recall  f1-score   support

           0       0.58      0.69      0.63        42
           1       0.85      0.77      0.81        92

    accuracy                           0.75       134
   macro avg       0.71      0.73      0.72       134
weighted avg       0.76      0.75      0.75       134
```



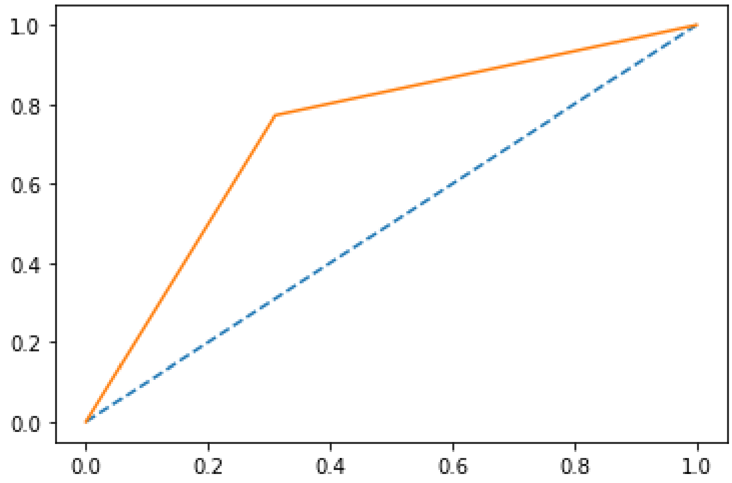


Fig 1.11: ROC Curve of Train and Test Dataset for CART Model

Naïve Bayes Model

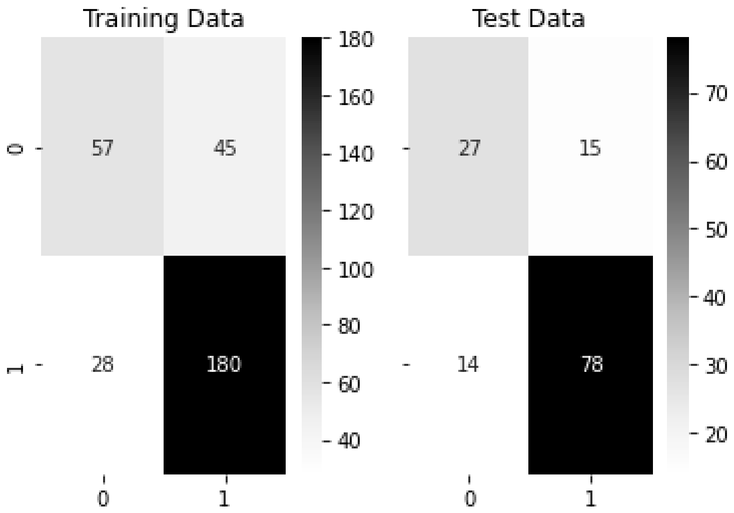


Fig 1.12: Confusion Matrix of Train and Test Dataset for Naïve Bayes Model

Classification Report on Training Data for Naive Bayes Model

```
              precision    recall  f1-score   support

           0       0.67      0.56      0.61       102
           1       0.80      0.87      0.83       208

    accuracy                           0.76       310
   macro avg       0.74      0.71      0.72       310
weighted avg       0.76      0.76      0.76       310
```

Classification Report on Testing Data for Naive Bayes Model

```
              precision    recall  f1-score   support

           0       0.66      0.64      0.65        42
           1       0.84      0.85      0.84        92

    accuracy                           0.78       134
```

```
   macro avg       0.75      0.75      0.75       134
weighted avg       0.78      0.78      0.78       134
```

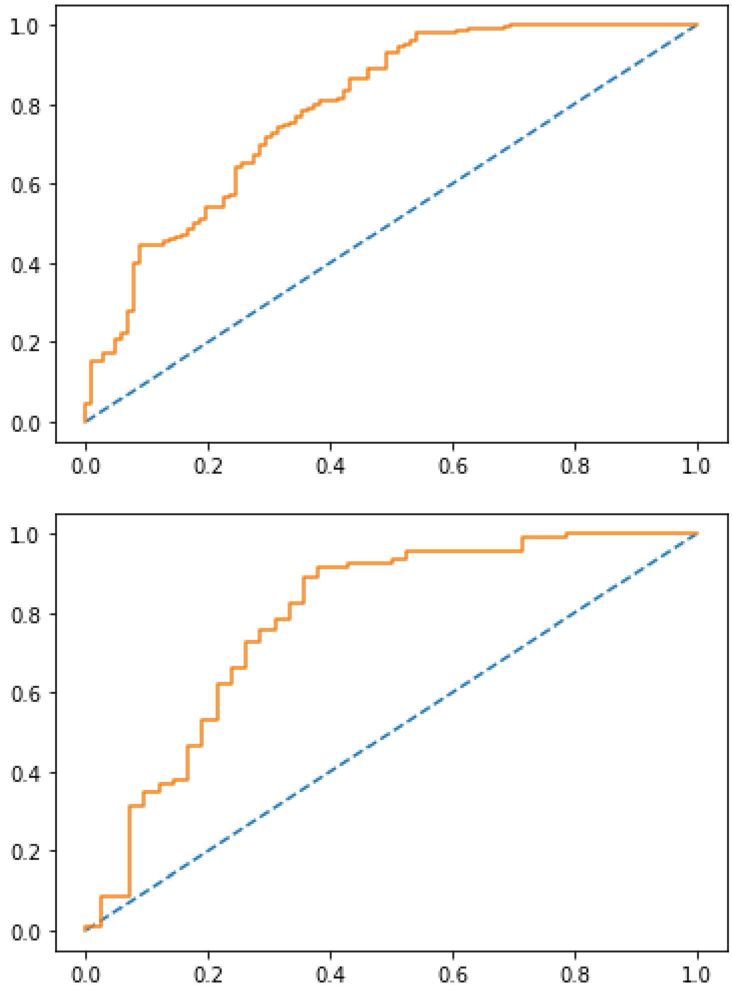


Fig 1.13: ROC Curve of Train and Test Dataset for Naive Bayes Model

KNN Model

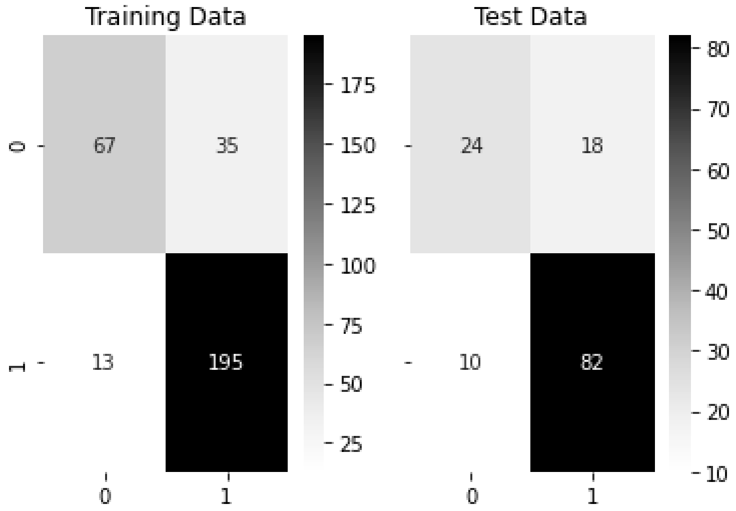


Fig 1.14: Confusion Matrix of Train and Test Dataset for KNN Model

Classification Report on Training Data for KNN Model

```
              precision    recall  f1-score   support
```

```
           0       0.84      0.66      0.74       102
           1       0.85      0.94      0.89       208

    accuracy                           0.85       310
   macro avg       0.84      0.80      0.81       310
weighted avg       0.84      0.85      0.84       310
```

Classification Report on Testing Data for KNN Model

```
              precision    recall  f1-score   support

           0       0.71      0.57      0.63        42
           1       0.82      0.89      0.85        92

    accuracy                           0.79       134
   macro avg       0.76      0.73      0.74       134
weighted avg       0.78      0.79      0.78       134
```

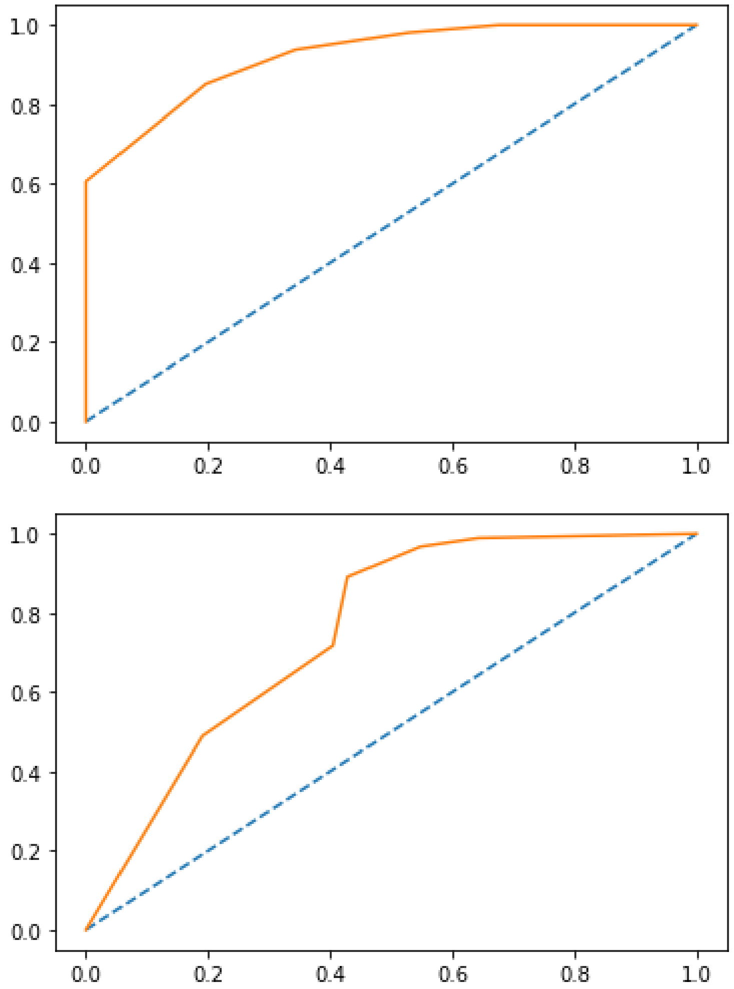


Fig 1.15: ROC Curve of Train and Test Dataset for KNN Model

Random Forest Model

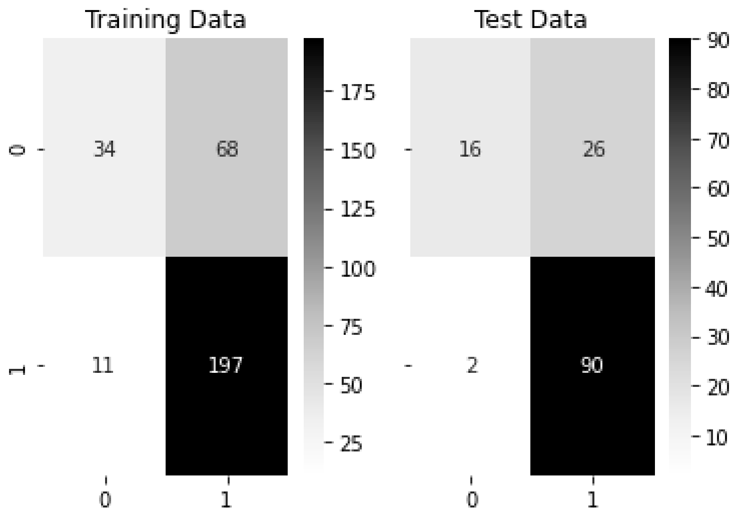


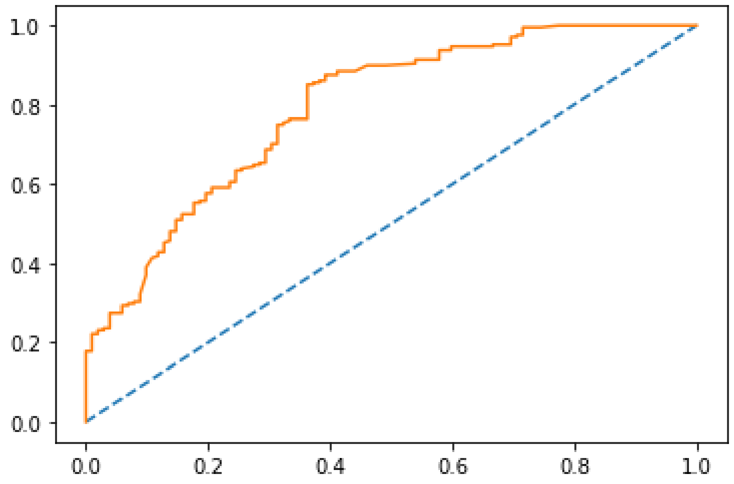
Fig 1.16: Confusion Matrix of Train and Test Dataset for Random Forest Model

Classification Report on Training Data for Random Forest Model

```
              precision    recall  f1-score   support

           0       0.76      0.33      0.46       102
           1       0.74      0.95      0.83       208

    accuracy                           0.75       310
   macro avg       0.75      0.64      0.65       310
weighted avg       0.75      0.75      0.71       310
```

Classification Report on Testing Data for Random Forest Model

```
              precision    recall  f1-score   support

           0       0.89      0.38      0.53        42
           1       0.78      0.98      0.87        92

    accuracy                           0.79       134
   macro avg       0.83      0.68      0.70       134
weighted avg       0.81      0.79      0.76       134
```



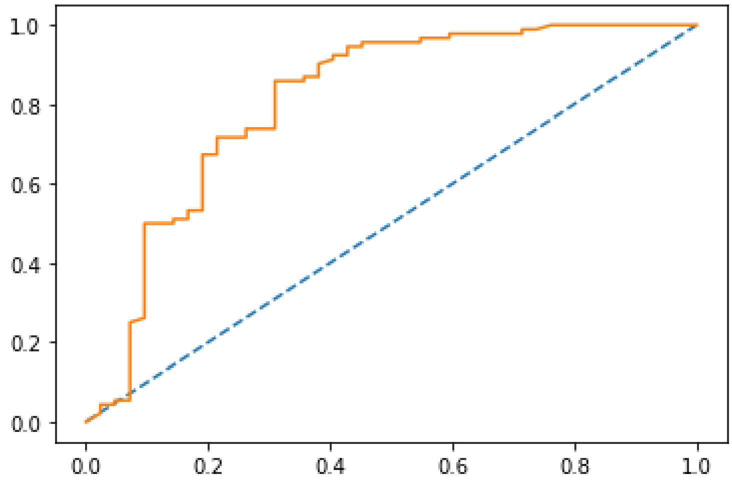


Fig 1.17: ROC Curve of Train and Test Dataset for Random Forest Model

Boosting Classifier Model using Gradient boost.

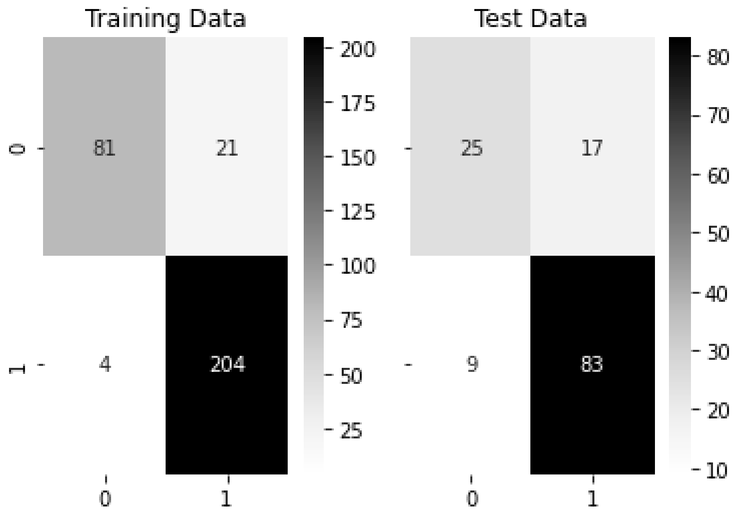


Fig 1.18: Confusion Matrix of Train and Test Dataset for Boosting Classifier Model

Classification Report on Training Data for Boosting Classifier Model

```
              precision    recall  f1-score   support

           0       0.95      0.79      0.87       102
           1       0.91      0.98      0.94       208

    accuracy                           0.92       310
   macro avg       0.93      0.89      0.90       310
weighted avg       0.92      0.92      0.92       310
```

Classification Report on Testing Data for Boosting Classifier Model

```
              precision    recall  f1-score   support

           0       0.74      0.60      0.66        42
           1       0.83      0.90      0.86        92

    accuracy                           0.81       134
```

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

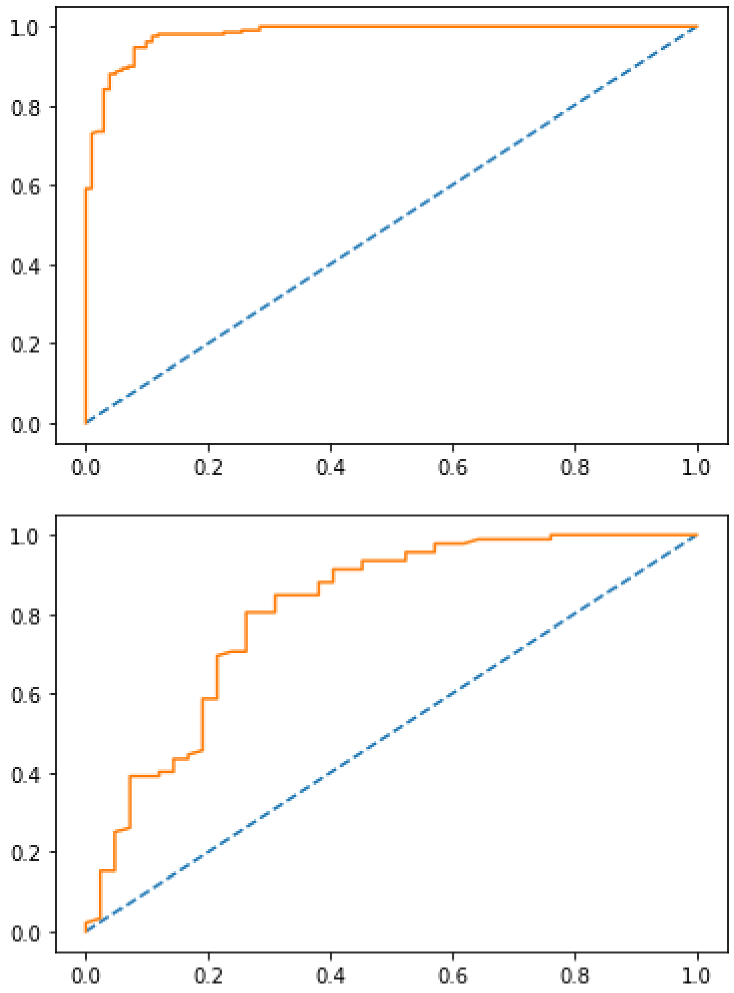
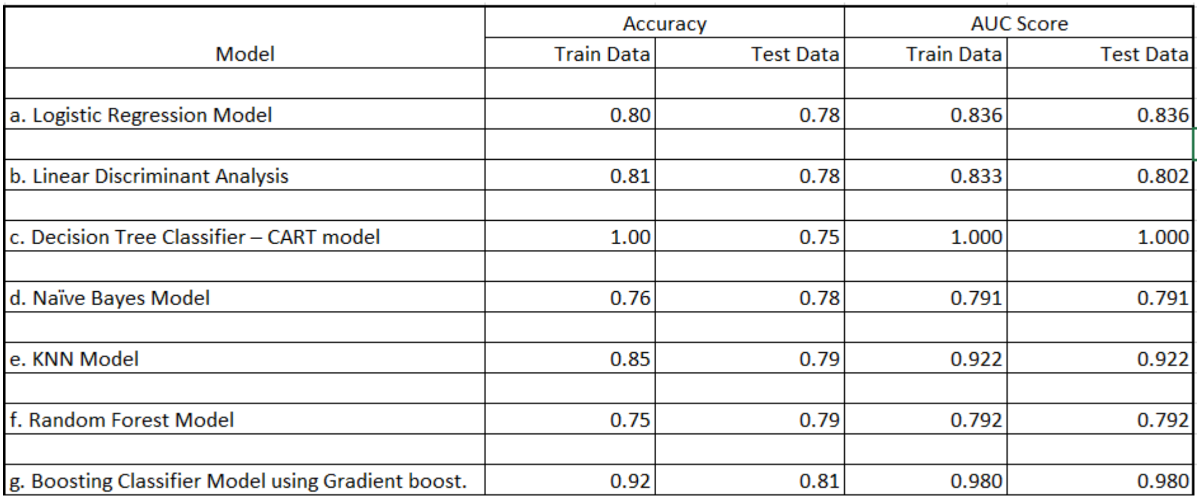


Fig 1.19: ROC Curve of Train and Test Dataset for Boosting Classifier Model

Question 1.4: Which model performs the best?



Model	Accuracy		AUC Score	
	Train Data	Test Data	Train Data	Test Data
a. Logistic Regression Model	0.80	0.78	0.836	0.836
b. Linear Discriminant Analysis	0.81	0.78	0.833	0.802
c. Decision Tree Classifier – CART model	1.00	0.75	1.000	1.000
d. Naïve Bayes Model	0.76	0.78	0.791	0.791
e. KNN Model	0.85	0.79	0.922	0.922
f. Random Forest Model	0.75	0.79	0.792	0.792
g. Boosting Classifier Model using Gradient boost.	0.92	0.81	0.980	0.980

- From the Above Table we can conclude that Boosting Classifier Model using Gradient Boost is the Best Model. Boosting Classifier Model has an AUC score of 0.98 on both Train and Test data set. Accuracy on train data is nearly 92% and on test data is 81%.
- But The difference in Accuracy between Train data and test Data is more than 10% which could result in poor results.

- 2nd Best Model is the KNN Model with AUC score of 0.922 on both Train and test data set.
- Accuracy of the KNN Model on Train data is 0.85 and on test Data is 0.79
- Worst that can be consider in the above scenario is the CART model.

Question 1.5: What are your business insights?

- Majority of the employees recorded in the data set prefer Public Transport.
- Age, Salary, Distance travelled play a major Role in Mode of Transport. Generally if the Distance between Home and Office is greater than 20 KM, Employees prefer Private Transport.
- A better model can be built if Enterprise provides more Employee records. More the records better the Accuracy and Prediction.
- With the given Dataset and Based on Accuracy of Prediction, Boosting Classifer Model can be equipped to predict what mode of transport (Public or Private) Employee use.

Part 2: Text Mining

A dataset of Shark Tank episodes is made available. It contains 495 entrepreneurs making their pitch to the VC sharks.

You will ONLY use “Description” column for the initial text mining exercise.

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

	deal	description	episode	category	entrepreneurs	location	website
0	False	Bluetooth device implant for your ear.	1	Novelties	Darrin Johnson	St. Paul, MN	NaN
1	True	Retail and wholesale pie factory with two reta...	1	Specialty Food	Tod Wilson	Somerset, NJ	http://whybake.com/
2	True	Ava the Elephant is a godsend for frazzled par...	1	Baby and Child Care	Tiffany Krumins	Atlanta, GA	http://www.avatheelephant.com/
3	False	Organizing, packing, and moving services deliv...	1	Consumer Services	Nick Friedman, Omar Soliman	Tampa, FL	http://collegehunkshaulingjunk.com/

	deal	description	episode	category	entrepreneurs	location	website
4	False	Interactive media centers for healthcare waiti...	1	Consumer Services	Kevin Flannery	Cary, NC	http://www.wispots.com/

Table 2.1: Shark Tank dataset

#	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

Table 2.2: Shark tank Dataset datatype Info

- We see that entrepreneurs and website have missing data
- We will append "No entrepreneurs mentioned" for entrepreneurs column and "No_website_mentioned" for Website column

#	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

#	Column	Non-Null Count	Dtype
4	entrepreneurs	495 non-null	object
5	location	495 non-null	object
6	website	495 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

Table 2.3: Shark tank Dataset datatype Info Updated

	deal	description
0	False	Bluetooth device implant for your ear.
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
...	...	...
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...

Table 2.4: Separate Dataframe containing Deal and Description

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

	deal	description
0	True	Retail and wholesale pie factory with two reta...
1	True	Ava the Elephant is a godsend for frazzled par...

	deal	description
2	True	One of the first entrepreneurs to pitch on Sha...
3	True	An educational record label and publishing hou...
4	True	A battery-operated cooking device that siphons...
...	...	...
246	True	SynDaver Labs makes synthetic body parts for u...
247	True	Zoom Interiors is a virtual service for interi...
248	True	Spikeball started out as a casual outdoors gam...
249	True	Shark Wheel is out to literally reinvent the w...
250	True	Sway Motorsports makes a three-wheeled, all-el...

Table 2.5: Corpora - Secured a Deal

	deal	description
0	False	Bluetooth device implant for your ear.
1	False	Organizing, packing, and moving services deliv...
2	False	Interactive media centers for healthcare waiti...
3	False	A mixed martial arts clothing line looking to ...
4	False	Attach Noted is a detachable "arm" that holds ...
...	...	...
239	False	Buck Mason makes high-quality men's clothing i...
240	False	Frameri answers the question, "Why aren't your...
241	False	The Paleo Diet Bar is a nutrition bar that is ...
242	False	Sunscreen Mist adds another point of access fo...
243	False	Adriana Montano wants to open the first Cat Ca...

Table 2.6: Corpora - Not secured a Deal

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

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

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

c) What were the top 3 most frequently occurring words in both corpuses (after removing stop words)?

d) Plot the Word Cloud for both the corpora.

Deal = True Corpus

	description	char_count

	description	char_count
0	Retail and wholesale pie factory with two reta...	73
1	Ava the Elephant is a godsend for frazzled par...	244
2	One of the first entrepreneurs to pitch on Sha...	365
3	An educational record label and publishing hou...	122
4	A battery-operated cooking device that siphons...	117

Table 2.7: Character Count for Deal Secured Corpus

- Total Number of Characters = 64060
- 1st 5 Words after Stop words were removed = ['Retail', 'wholesale', 'pie', 'factory', 'two']
- Top 3 most frequently occurring words are 'The': 79, 'A': 64, 'make': 25

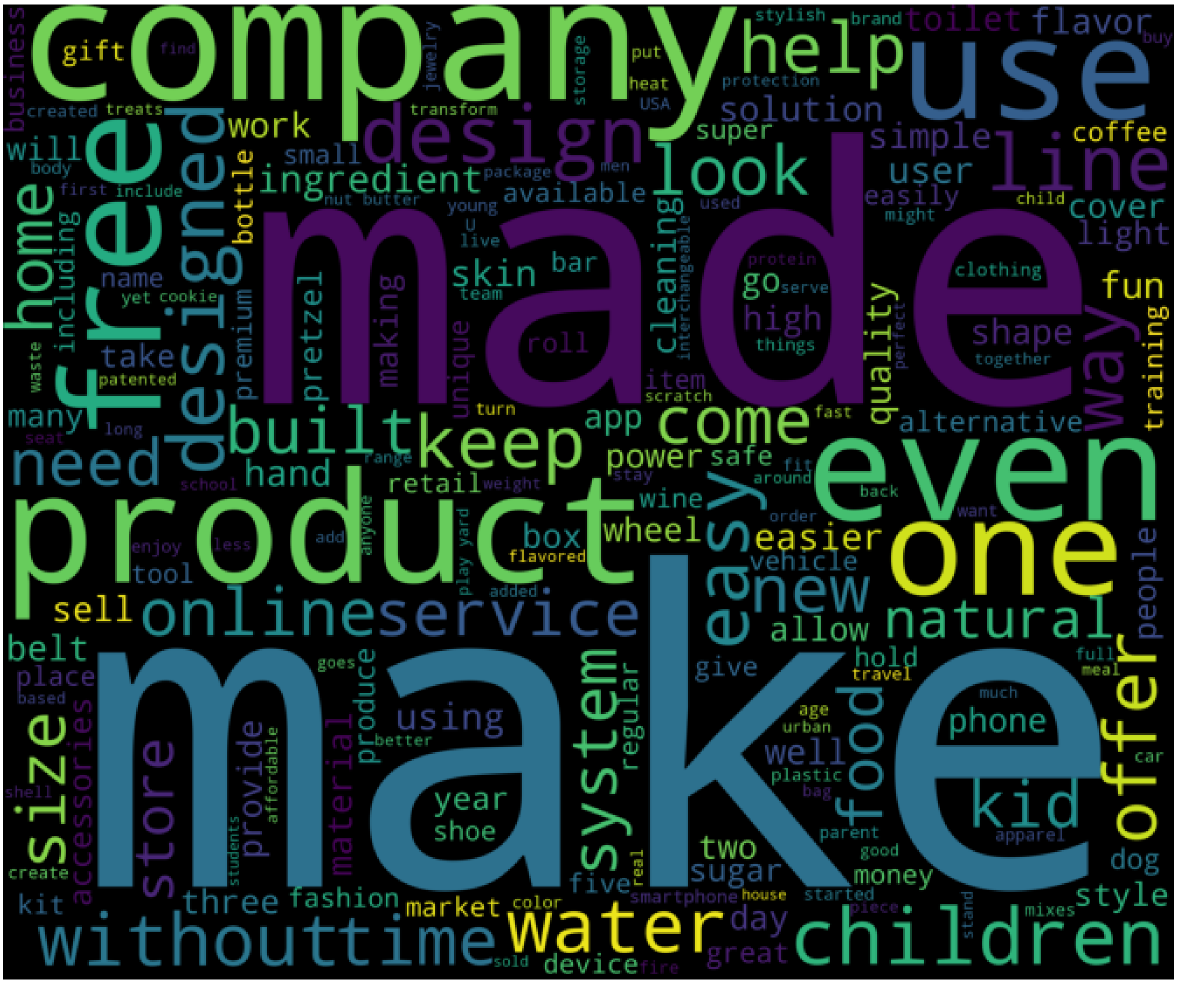


Fig 2.1: Word Cloud for Deal Secured Corpus

Deal = False Corpus

	description	char_count
0	Bluetooth device implant for your ear.	38
1	Organizing, packing, and moving services deliv...	68
2	Interactive media centers for healthcare waiti...	112

	description	char_count
3	A mixed martial arts clothing line looking to ...	110
4	Attach Noted is a detachable "arm" that holds ...	91

Table 2.7: Character Count for Deal Not Secured Corpus

- Total Number of Characters = 47184
- 1st 5 Words after Stop words were removed = ['Bluetooth', 'device', 'implant', 'ear.', 'Organizing,']
- Top 3 most frequently occurring words are 'A': 76, 'The': 54, 'An': 19

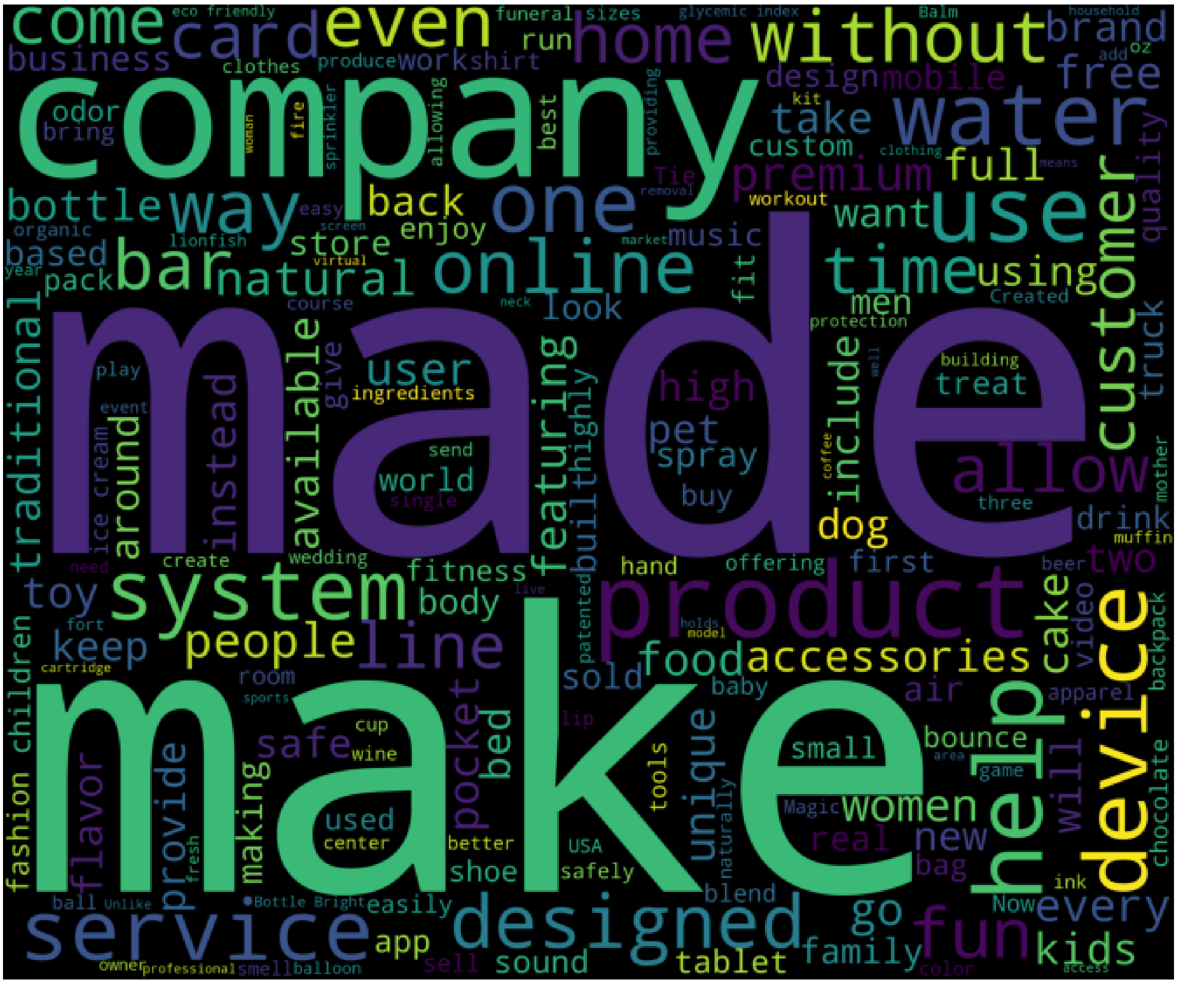


Fig 2.2: Word Cloud for Deal Not Secured Corpus

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

- From both the Word clouds, we observe that the words like "make, made, Company, Product" are common indicating that those words have repeated multiple times.
- From Deal Secured Corpus, Words like Online Service, Designed, System, offer, Free, Easy are highlighted which could part of the reason why they could secure deal.
- From Deal Not Secured Corpus, Service, without, people, device, traditional, Food, fitness are highlighted.

Question 2.5: Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis?

- Based on Word Cloud it looks like Entrepreneurs who couldnt bag a deal most likely used the word "Device" which could be a reason but not entirely sure.
- There could be other reason that could have subjugated 'A No Deal' for the Entrepreneurs.
- Looking at the Word Cloud, Entrepreneurs who focus on traditional, people, Clothes, Fun, Food, Services couldn't bag a deal.
- A detailed Analysis should provide us more information that can be used to bag a deal on Shark tank.

THE END