PREDICTIVE MODELLING PROJECT

SUBMITTED BY – SHWETA TRIPATHI

List Of Tables

List of Figures

Problem 1: Linear Regression

Problem Statement: You are a part of an investing firm and your work is to do research about these 759 firms. You are provided with the dataset containing the sales and other attributes of these 759 firms. Predict the sales of these firms on the bases of the details given in the dataset so as to help your company in investing consciously. Also, provide them with 5 attributes that are most important.

1.1 Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis.

Solution:

Importing the dataset in jupyter notebook and then reading the dataset as follows:-

The top 5 rows of the dataset can be obtained using the head () function which is shown below

	Unnamed: 0	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	0	826.995050	161.603986	10	382.078247	2.306000	no	11.049511	1625.453755	80.27
1	1	407.753973	122.101012	2	0.000000	1.860000	no	0.844187	243.117082	59.02
2	2	8407.845588	6221.144614	138	3296.700439	49.659005	yes	5.205257	25865.233800	47.70
3	3	451.000010	266.899987	1	83.540161	3.071000	no	0.305221	63.024630	26.88
4	4	174.927981	140.124004	2	14.233637	1.947000	no	1.063300	67.406408	49.46

Table:-1

The bottom 5 rows of the dataset can be obtained by using the tail () function

	Unnamed: 0	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
754	754	1253.900196	708.299935	32	412.936157	22.100002	1	0.697454	267.119487	33.50
755	755	171.821025	73.666008	1	0.037735	1.684000	0	NaN	228.475701	46.41
756	756	202.726967	123.926991	13	74.861099	1.460000	0	5.229723	580.430741	42.25
757	757	785.687944	138.780992	6	0.621750	2.900000	1	1.625398	309.938651	61.39
758	758	22.701999	14.244999	5	18.574360	0.197000	0	2.213070	18.940140	7.50

Table:-2

The information of the dataset can be obtained by using the info() function.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 759 entries, 0 to 758
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Unnamed: 0     759 non-null    int64
 1   sales          759 non-null    float64
 2   capital        759 non-null    float64
 3   patents        759 non-null    int64
 4   randd          759 non-null    float64
 5   employment     759 non-null    float64
 6   sp500          759 non-null    object
 7   tobinq         738 non-null    float64
 8   value          759 non-null    float64
 9   institutions   759 non-null    float64
dtypes: float64(7), int64(2), object(1)
memory usage: 59.4+ KB
```

Table:- 3

The above table represent the information of the dataset such as the dataset contains total 759 rows, in which total 7 is float64, 2 is int64 and 1 object.

The number of rows and columns from the dataset can be obtained by using the shape function. Thus, the dataset contains total 759 rows and 10 columns.

```
The number of rows in dataframe is 759
The number of rows in dataframe is 10
```

Describing the dataset:-

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	759.0	379.000000	219.248717	0.000000	189.500000	379.000000	568.500000	758.000000
sales	759.0	2689.705158	8722.060124	0.138000	122.920000	448.577082	1822.547366	135696.788200
capital	759.0	1977.747498	6466.704896	0.057000	52.650501	202.179023	1075.790020	93625.200560
patents	759.0	25.831357	97.259577	0.000000	1.000000	3.000000	11.500000	1220.000000
randd	759.0	439.938074	2007.397588	0.000000	4.628262	36.864136	143.253403	30425.255860
employment	759.0	14.164519	43.321443	0.006000	0.927500	2.924000	10.050001	710.799925
sp500	759.0	0.285903	0.452141	0.000000	0.000000	0.000000	1.000000	1.000000
tobinq	738.0	2.794910	3.366591	0.119001	1.018783	1.680303	3.139309	20.000000
value	759.0	2732.734750	7071.072362	1.971053	103.593946	410.793529	2054.160386	95191.591160
institutions	759.0	43.020540	21.685586	0.000000	25.395000	44.110000	60.510000	90.150000

Table:-4

The number of duplicate rows in the dataset is:-

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

Thus, no duplicate rows present in the dataset.

Checking the no of missing values using is null function:-

```
tobinq          21
Unnamed: 0       0
sales            0
capital          0
patents          0
randd            0
employment       0
sp500            0
value            0
institutions     0
dtype: int64
```

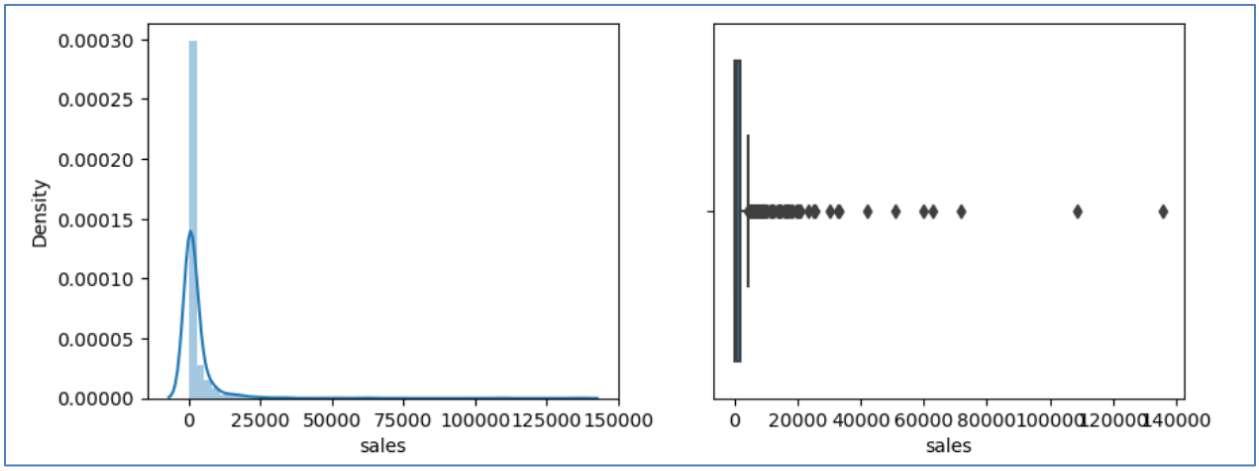
Since tobinq has 21 missing values present in the dataset. Thus, the missing values are removed using fill na () function as follows:-

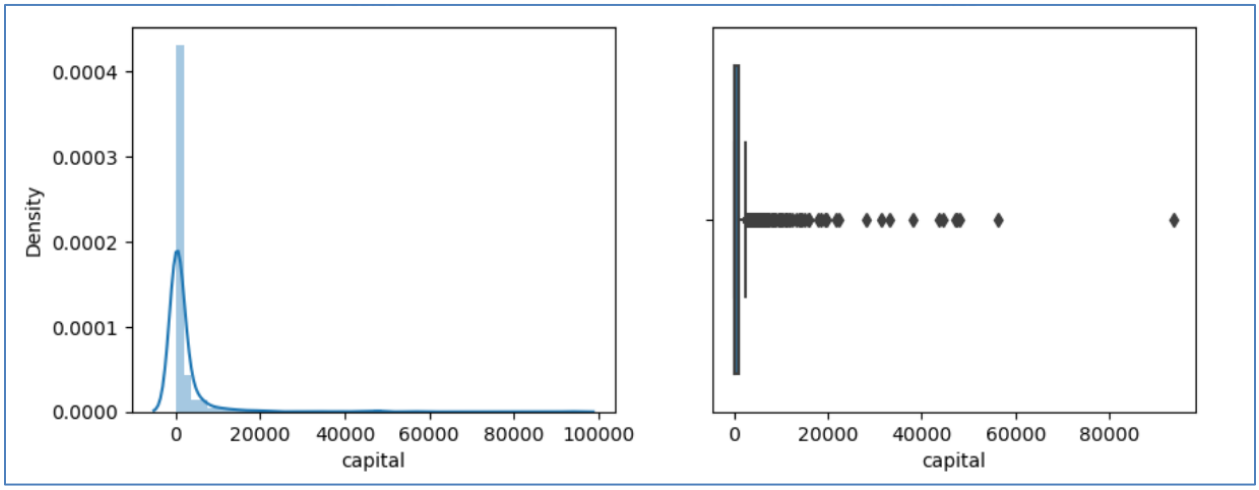
```
Unnamed: 0      0
sales           0
capital         0
patents         0
randd           0
employment      0
sp500           0
tobinq          0
value           0
institutions    0
dtype: int64
```

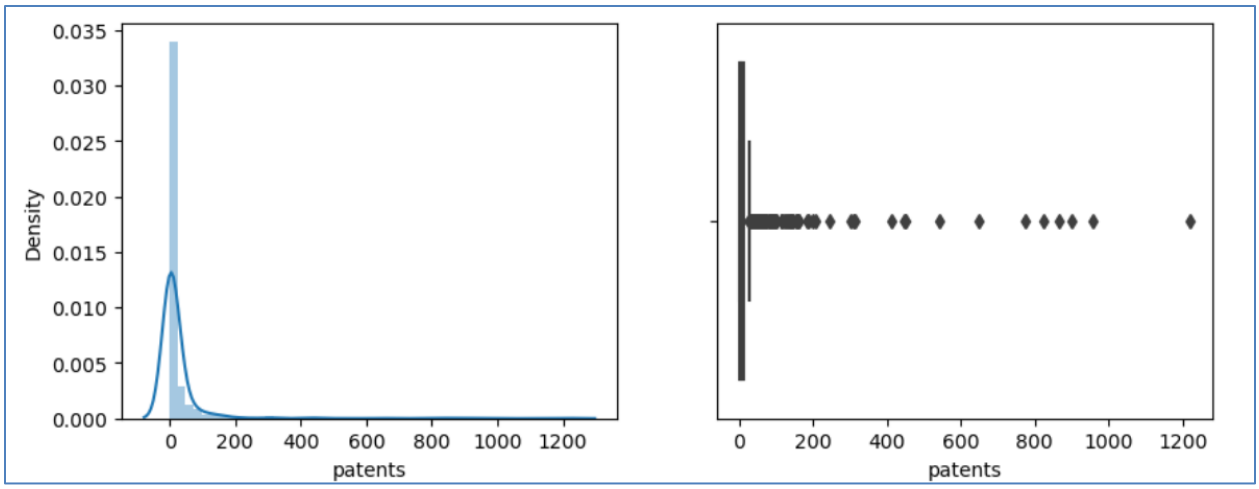
Now, no missing values present in the dataset.

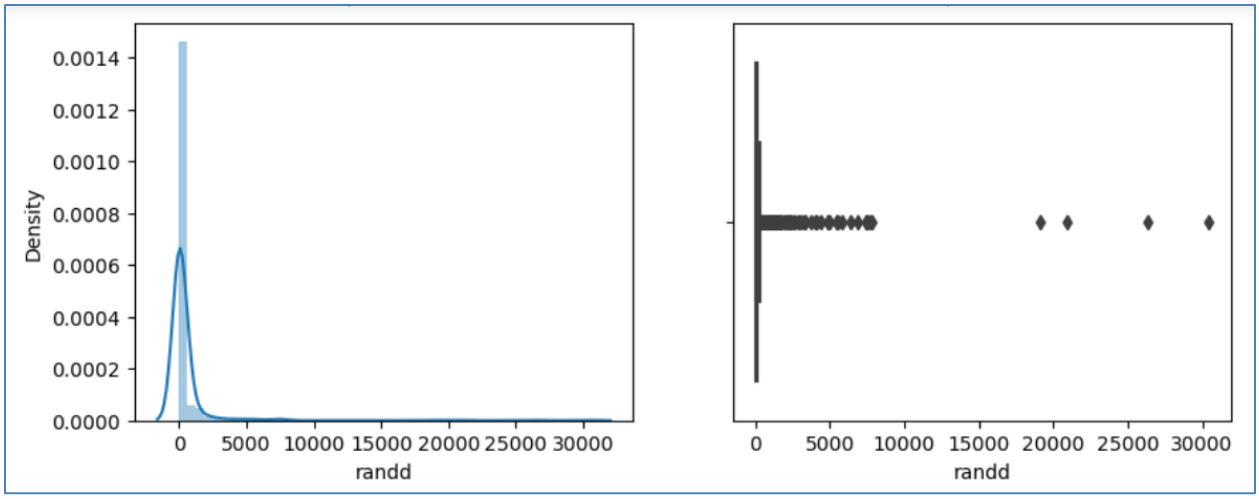
Univariate analysis:-

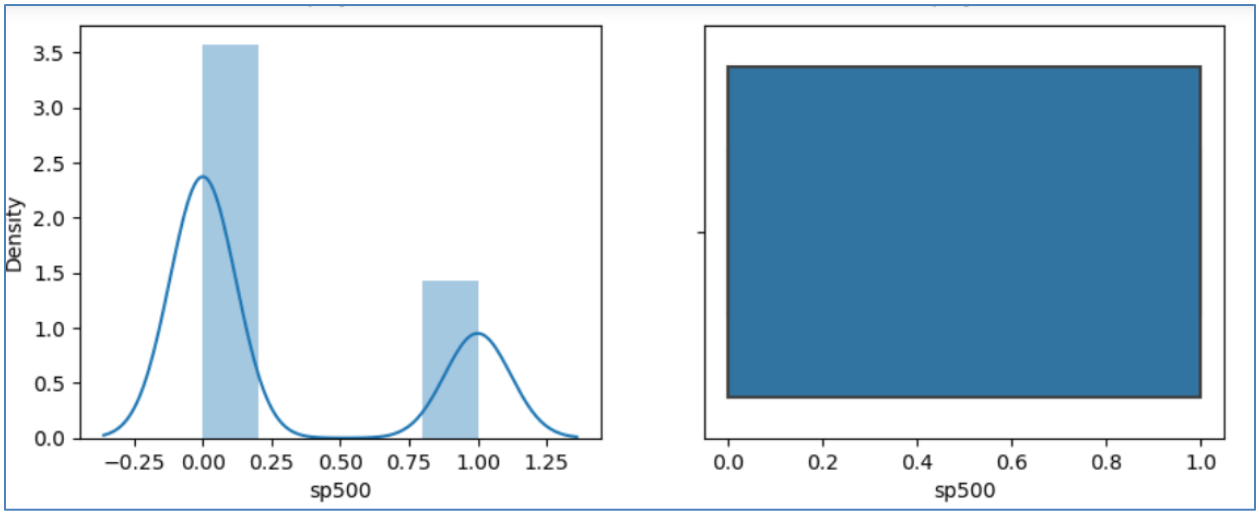
Univariate analysis of the dataset is shown below:-

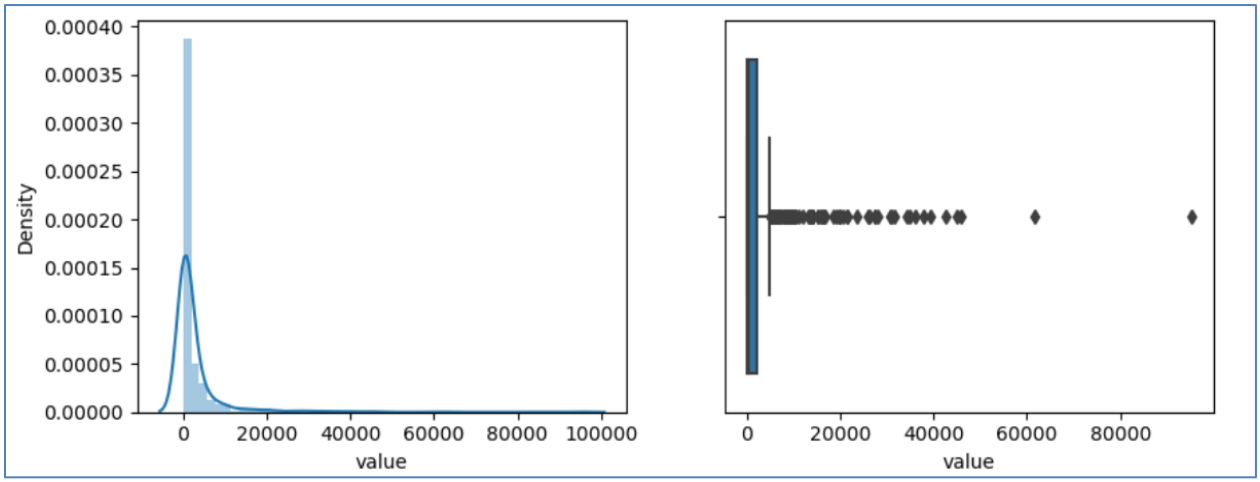







Density
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
5000
10000
15000
20000
25000
30000
randd
0
5000
10000
15000
20000
25000
30000
randd


Density
3.5
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
sp500
0.0
0.2
0.4
0.6
0.8
1.0
sp500


Density
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
20000
40000
60000
80000
100000
value
0
20000
40000
60000
80000
value

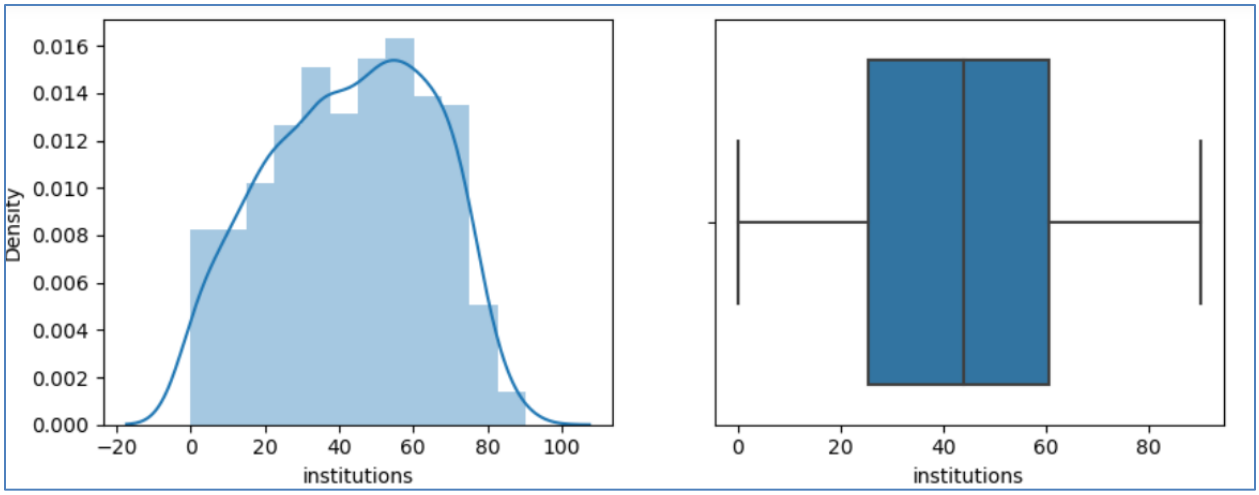


Figure:-1

From the above univariate analysis we have observed that in most of the cases the data is normally distributed. A normal distribution curve is also known as the bell curve.

Checking the outliers:-

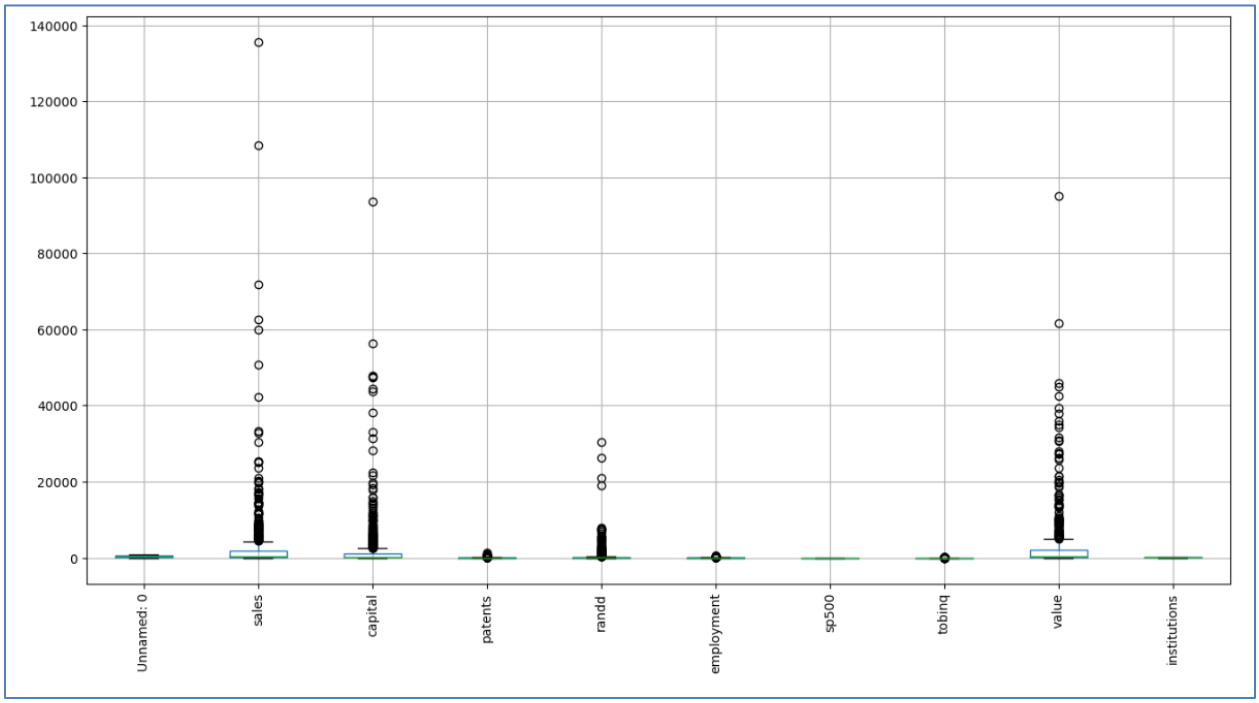


Figure:-2

Removing the outliers we will get:-

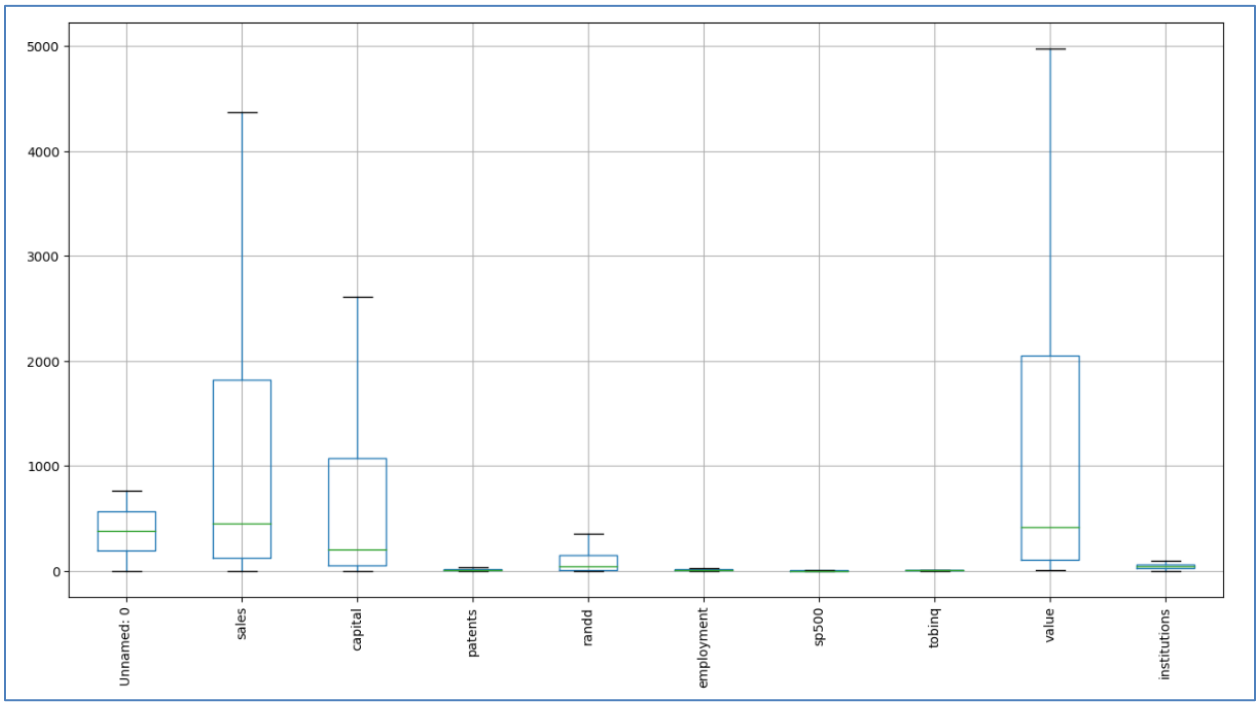


Figure:-3

Bivariate Analysis

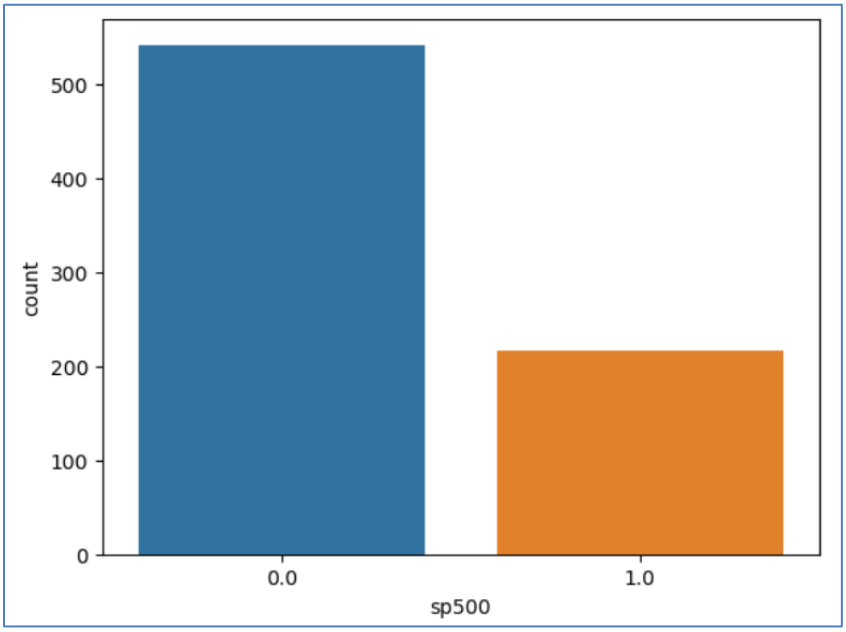


Figure:-4

Thus the number of candidate having membership of firms in the S&P 500 index is less compared to the number of candidates who do not have membership of the firms in the S&P 500 index.

Pair Plot:-

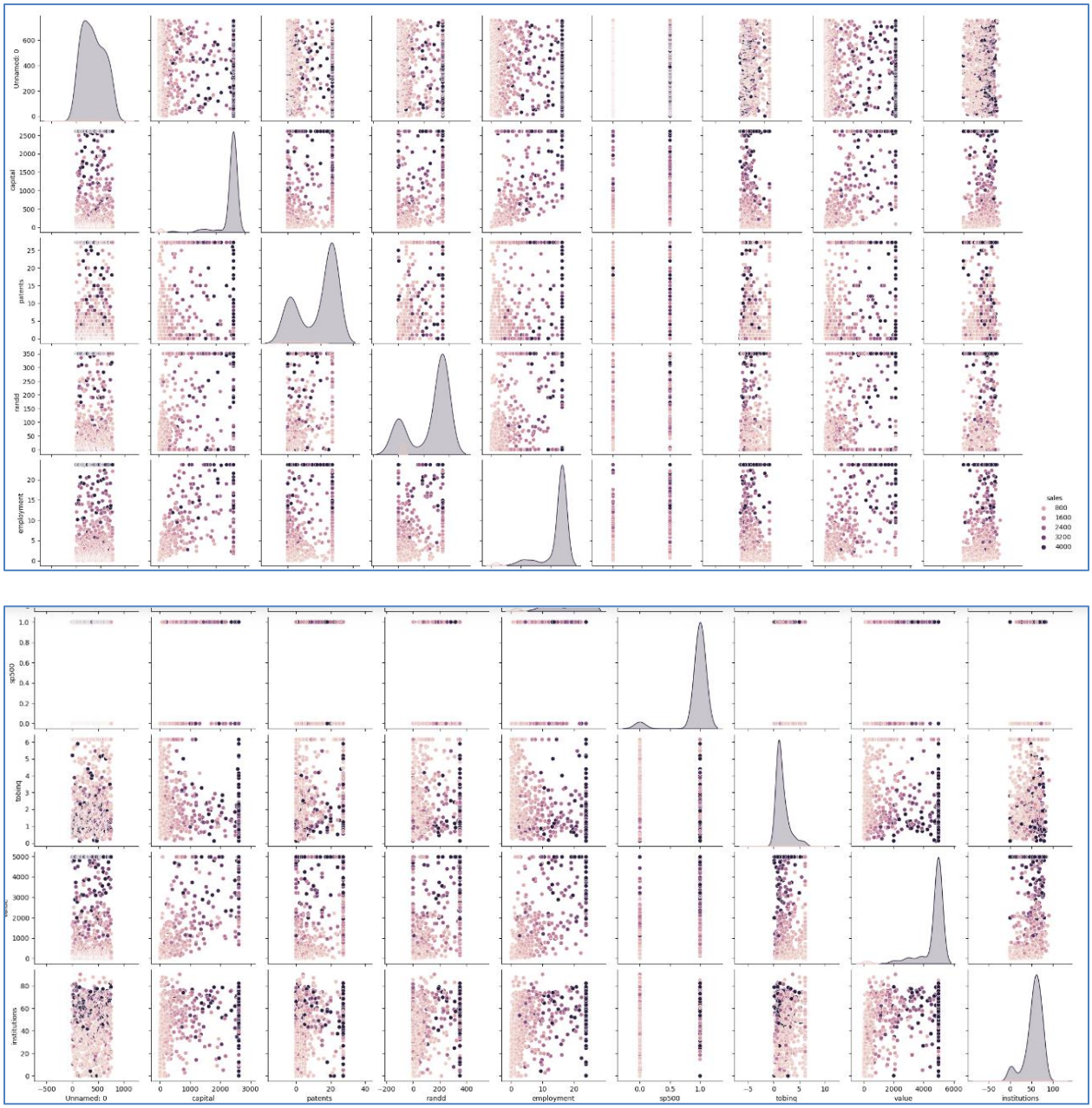


Figure:-5

A pair plot is used to compare relationships across multiple variables. The above pair plot shows the multi- collinearity

Heat Map:-

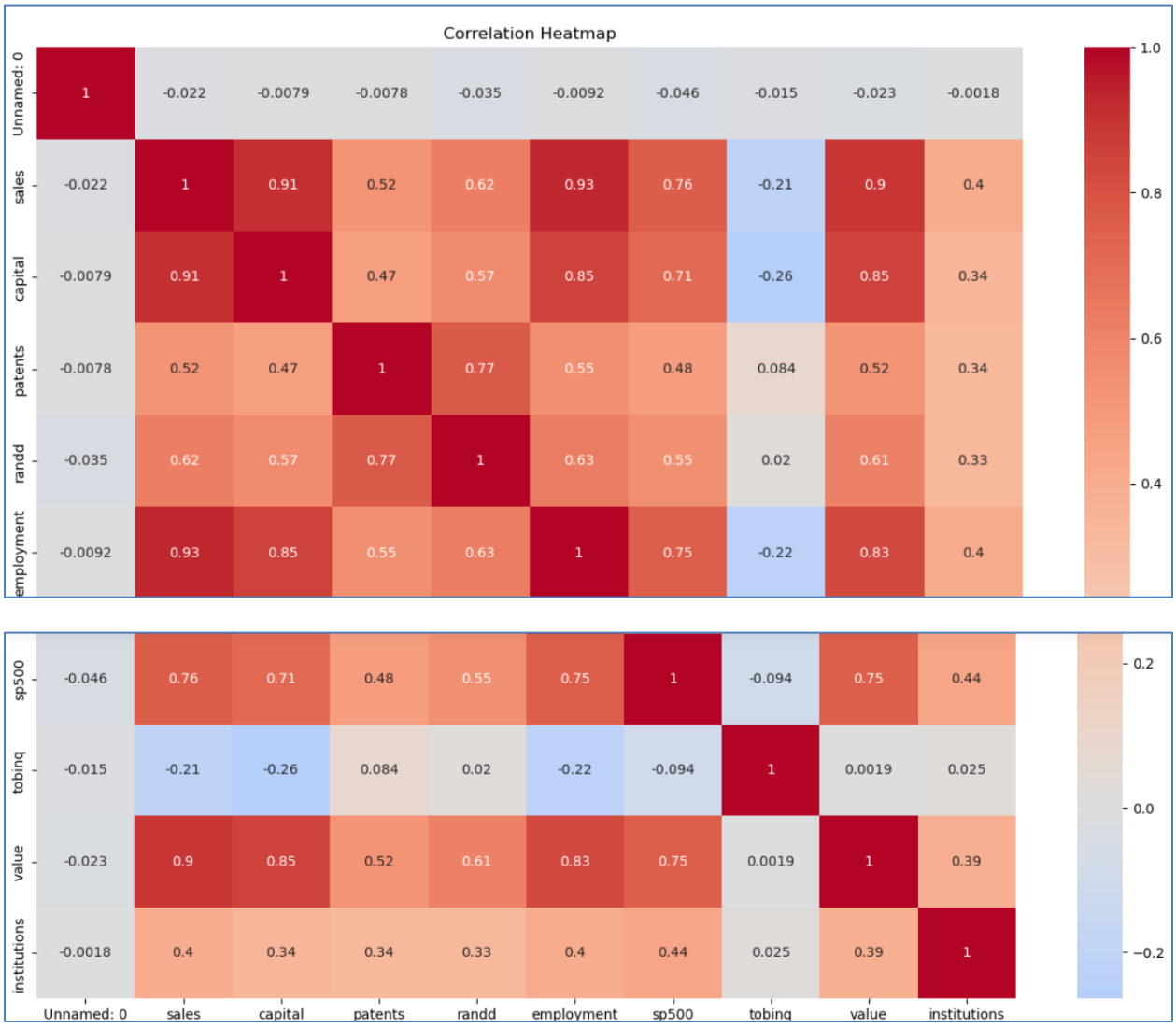


Figure:-6

A heat map is a graphical representation of data where values in a matrix are represented as colors. The red color shows the represents the values of a variable against itself. From the above graph it is clear that employment and capital is highly correlated with sales whereas tobinq is least correlated.

1.2 Impute null values if present? Do you think scaling is necessary in this case? (8 marks)

Since tobinq has 21 missing values present in the dataset. Thus, the missing values are removed using fillna function as follows:-

```
Unnamed: 0      0
sales           0
capital         0
patents         0
randd           0
employment      0
sp500           0
tobinq          0
value           0
institutions    0
dtype: int64
```

Now, no missing values present in the dataset.

Scaling is necessary thus after scaling the data is

	Unnamed: 0	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	-1.729770	-0.267788	-0.591504	0.221152	1.979986	-0.564800	-0.632747	2.221566	0.142598	1.718839
1	-1.725206	-0.542217	-0.632706	-0.583181	-0.782879	-0.619331	-0.632747	-0.864319	-0.645807	0.738279
2	-1.720642	2.052715	1.962722	1.955496	1.979986	2.055116	1.580410	1.670452	2.055843	0.215929
3	-1.716078	-0.513909	-0.481679	-0.683723	-0.125658	-0.471265	-0.632747	-1.177581	-0.748521	-0.744789
4	-1.711514	-0.694622	-0.613908	-0.583181	-0.670901	-0.608694	-0.632747	-0.736965	-0.746022	0.297142
...	...	...	...	...	...	...	...	...	...	...
754	1.711514	0.011658	-0.021294	1.955496	1.979986	1.855361	1.580410	-0.949605	-0.632117	-0.439316
755	1.716078	-0.696656	-0.683224	-0.683723	-0.782582	-0.640850	-0.632747	0.269493	-0.654157	0.156403
756	1.720642	-0.676425	-0.630802	0.522777	-0.193937	-0.668238	-0.632747	1.684672	-0.453422	-0.035556
757	1.725206	-0.294827	-0.615309	-0.181015	-0.777987	-0.492173	1.580410	-0.410259	-0.607695	0.847640
758	1.729770	-0.794267	-0.745201	-0.281556	-0.636752	-0.822662	-0.632747	-0.068688	-0.773664	-1.639059

759 rows × 10 columns

Table:-5

1.3 Encode the data (having string values) for Modeling. Data Split: Split the data into test and train (70:30). Apply linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R square, RMSE.

After splitting into 70:30 ratios. The top 5 rows of the 70% data are as follows:-

	Unnamed: 0	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	-1.729770	-0.591504	0.221152	1.979986	-0.564800	-0.632747	2.221566	0.142598	1.718839
1	-1.725206	-0.632706	-0.583181	-0.782879	-0.619331	-0.632747	-0.864319	-0.645807	0.738279
2	-1.720642	1.962722	1.955496	1.979986	2.055116	1.580410	1.670452	2.055843	0.215929
3	-1.716078	-0.481679	-0.683723	-0.125658	-0.471265	-0.632747	-1.177581	-0.748521	-0.744789
4	-1.711514	-0.613908	-0.583181	-0.670901	-0.608694	-0.632747	-0.736965	-0.746022	0.297142

Table:-6

And the 30% of the data is as follows

```
0   -0.267788
1   -0.542217
2    2.052715
3   -0.513909
4   -0.694622
Name: sales, dtype: float64
```

The VIF values are as following:-

```
VIF values:

const           1.003914
Unnamed: 0      1.005452
capital         5.689439
patents         2.662589
randd           2.951166
employment      5.277221
sp500           3.057323
tobinq          1.443364
value           6.727279
institutions    1.287788
dtype: float64
```

Now removing all the values for which P>|t|>0.5. The model summary is as follows:-

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     1267.
Date:                Mon, 19 Feb 2024   Prob (F-statistic):          3.31e-308
Time:                        15:21:33   Log-Likelihood:                -36.107
No. Observations:                 531   AIC:                             86.21
Df Residuals:                     524   BIC:                             116.1
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0027      0.011      0.238      0.812      -0.020       0.025
capital        0.2573      0.026      9.770      0.000       0.206       0.309
randd          0.0335      0.015      2.203      0.028       0.004       0.063
employment     0.4167      0.025     16.491      0.000       0.367       0.466
sp500          0.0471      0.019      2.444      0.015       0.009       0.085
tobinq        -0.0478      0.014     -3.523      0.000      -0.074      -0.021
value          0.2802      0.029      9.570      0.000       0.223       0.338
==============================================================================
Omnibus:                      183.971   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1237.923
Skew:                           1.347   Prob(JB):                    1.54e-269
Kurtosis:                       9.978   Cond. No.                         6.40
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table:-7

Now fitting the Linear regression in this model:-

```
▾ LinearRegression
LinearRegression()
```

Linear Regression value for R^2 and RMSE for training and test data is as follows:-

```
Training RMSE: 394.57620904349244
Test RMSE: 399.39086236931183
Training R^2: 0.9358926020714278
Test R^2: 0.9242631315733321
```

The model appears to perform well on both the training and test data, as indicated by the relatively low root mean squared error (RMSE) values and high R-squared values. A low RMSE indicates that the model's predictions are close to the actual values, while high R-squared values suggest that a large proportion of the variance in the target variable is explained by the predictors

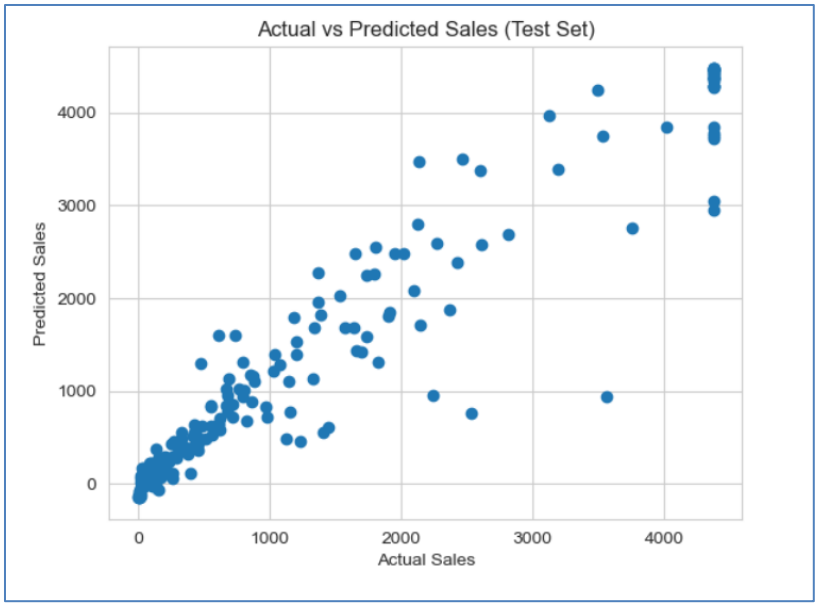


Figure:-7

ROC curve for training data:-

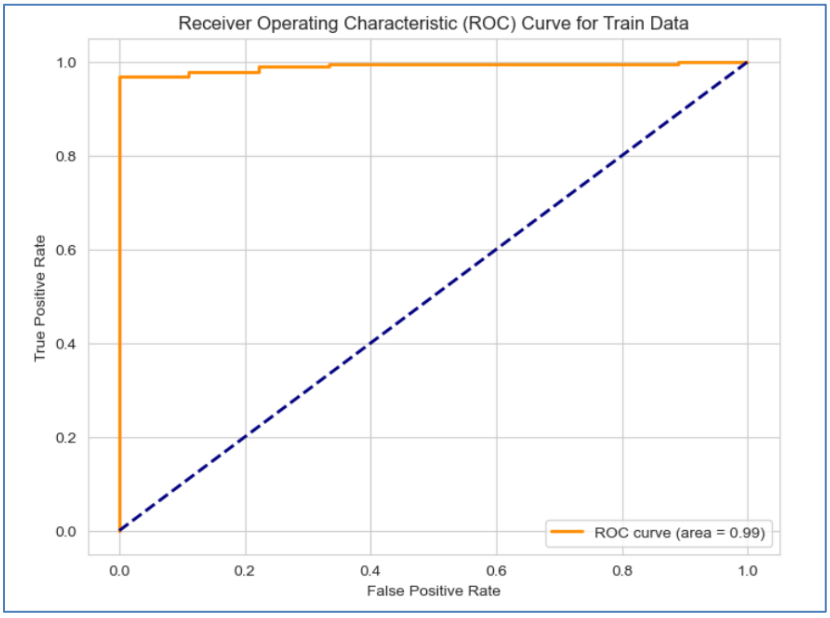


Figure:-8

ROC Curve for test data is as follows :-

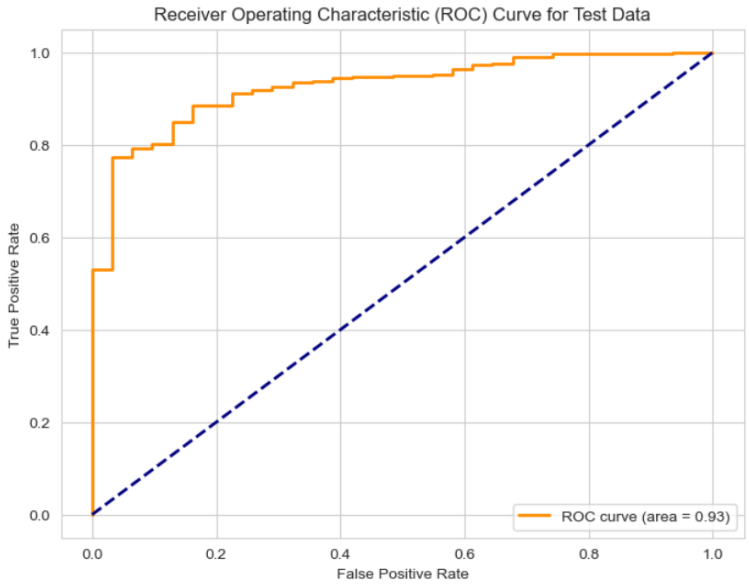


Figure:-9

2.4 Inference: Based on these predictions, what are the insights and recommendations?

1) As employment, capital investment, and value have a relatively stronger influence on sales thus businesses should focus their resources and strategies in these areas to increase sales.

2) A one-unit increase in capital is associated with a 0.2573 unit increase in sales thus capital investment is good.

3) Capital, randd, employment, sp500, tobinq, and value have 95% confidence level. Thus they have high impact on sales.

Problem 2: Logistic Regression and LDA

Problem Statement: You are hired by Government to do analysis on car crashes. You are provided details of car crashes, among which some people survived and some didn't. You have to help the government in predicting whether a person will survive or not on the basis of the information given in the data set so as to provide insights that will help government to make stronger laws for car manufacturers to ensure safety measures. Also, find out the important factors on the basis of which you made your predictions.

2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

Reading the dataset:-

	Unnamed: 0	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
0	0	55+	27.078	Not_Survived	none	none	1	m	32	1997	1987.0	unavail	driver	0	4.0	2:13:2
1	1	25-39	89.627	Not_Survived	airbag	belted	0	f	54	1997	1994.0	nodeploy	driver	0	4.0	2:17:1
2	2	55+	27.078	Not_Survived	none	belted	1	m	67	1997	1992.0	unavail	driver	0	4.0	2:79:1
3	3	55+	27.078	Not_Survived	none	belted	1	f	64	1997	1992.0	unavail	pass	0	4.0	2:79:1
4	4	55+	13.374	Not_Survived	none	none	1	m	23	1997	1986.0	unavail	driver	0	4.0	4:58:1

Table:-8

Bottom 5 dataset is as follows:-

	Unnamed: 0	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
11212	11212	2	3179.688	1	0	1	1	1	17	2002	1985.0	2	1	0	0.0	5679
11213	11213	1	71.228	1	1	1	1	1	54	2002	2002.0	3	1	0	2.0	5681
11214	11214	1	10.474	1	1	1	1	0	27	2002	1990.0	1	1	1	3.0	5686
11215	11215	2	10.474	1	1	1	1	0	18	2002	1999.0	1	1	1	0.0	5687
11216	11216	2	10.474	1	1	1	1	1	17	2002	1999.0	1	0	1	0.0	5687

Table:-9

The no of duplicate rows:-

```
Number of duplicate rows = 0
(11217, 16)
```

Thus, no duplicate rows present in the dataset.

The information of the dataset is as follows:-

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11217 entries, 0 to 11216
Data columns (total 16 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   11217 non-null  int64
 1   dvcat        11217 non-null  object
 2   weight       11217 non-null  float64
 3   Survived     11217 non-null  object
 4   airbag       11217 non-null  object
 5   seatbelt     11217 non-null  object
 6   frontal      11217 non-null  int64
 7   sex          11217 non-null  object
 8   ageOFocc     11217 non-null  int64
 9   yearacc      11217 non-null  int64
 10  yearVeh      11217 non-null  float64
 11  abcat        11217 non-null  object
 12  occRole      11217 non-null  object
 13  deploy       11217 non-null  int64
 14  injSeverity  11140 non-null  float64
 15  caseid       11217 non-null  object
dtypes: float64(3), int64(5), object(8)
memory usage: 1.4+ MB
```

Table: - 10

Thus above dataset contains 11217 values and 15 columns in which 3 are float type , 5 are integer type and 8 are object type.

The null values in the dataset are as follows:-

```
Unnamed: 0      0
dvcat           0
weight          0
Survived        0
airbag          0
seatbelt        0
frontal         0
sex             0
ageOFocc        0
yearacc         0
yearVeh         0
abcat           0
occRole         0
deploy          0
injSeverity    77
caseid          0
dtype: int64
```

From the above dataset we can conclude that total 77 null values are present in injSeverity column.

The description of the dataset is as follows:-

	Unnamed: 0	weight	frontal	ageOFocc	yearacc	yearVeh	deploy	injSeverity
count	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000	11140.000000
mean	5608.000000	431.405309	0.644022	37.427654	2001.103236	1994.177944	0.389141	1.825583
std	3238.213319	1406.202941	0.478830	18.192429	1.056805	5.658704	0.487577	1.378535
min	0.000000	0.000000	0.000000	16.000000	1997.000000	1953.000000	0.000000	0.000000
25%	2804.000000	28.292000	0.000000	22.000000	2001.000000	1991.000000	0.000000	1.000000
50%	5608.000000	82.195000	1.000000	33.000000	2001.000000	1995.000000	0.000000	2.000000
75%	8412.000000	324.056000	1.000000	48.000000	2002.000000	1999.000000	1.000000	3.000000
max	11216.000000	31694.040000	1.000000	97.000000	2002.000000	2003.000000	1.000000	5.000000

Table:-11

Removing the missing values:-

```
Unnamed: 0     0
dvcat          0
weight         0
Survived       0
airbag         0
seatbelt       0
frontal        0
sex            0
ageOFocc       0
yearacc        0
yearVeh        0
abcat          0
occRole        0
deploy         0
injSeverity    0
caseid         0
dtype: int64
```

The null values are removed using using fillna function mean values. Thus, the dataset do not contain any null values

UNIVARIATE ANALYSIS:-

The univariate analysis of the numerical dataset is as follows:-

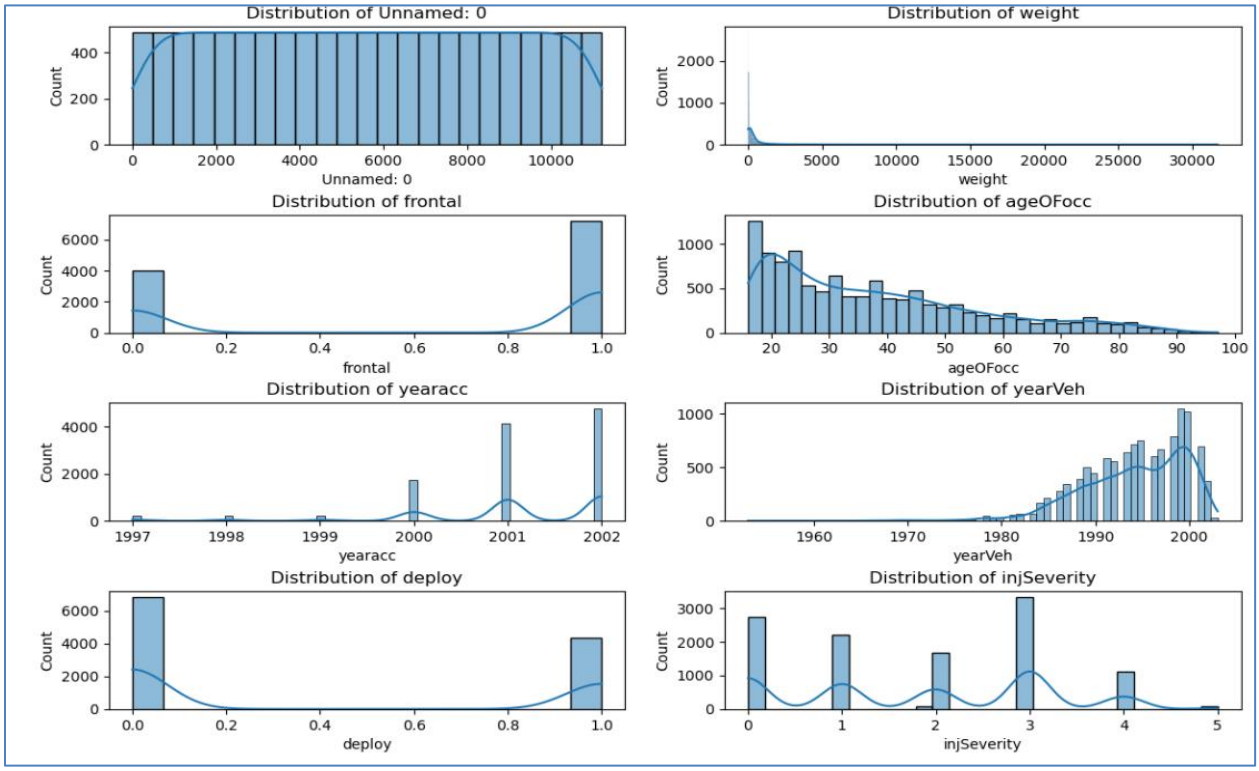


Figure:-10

The univariate analysis of the categorical dataset is as follows:-

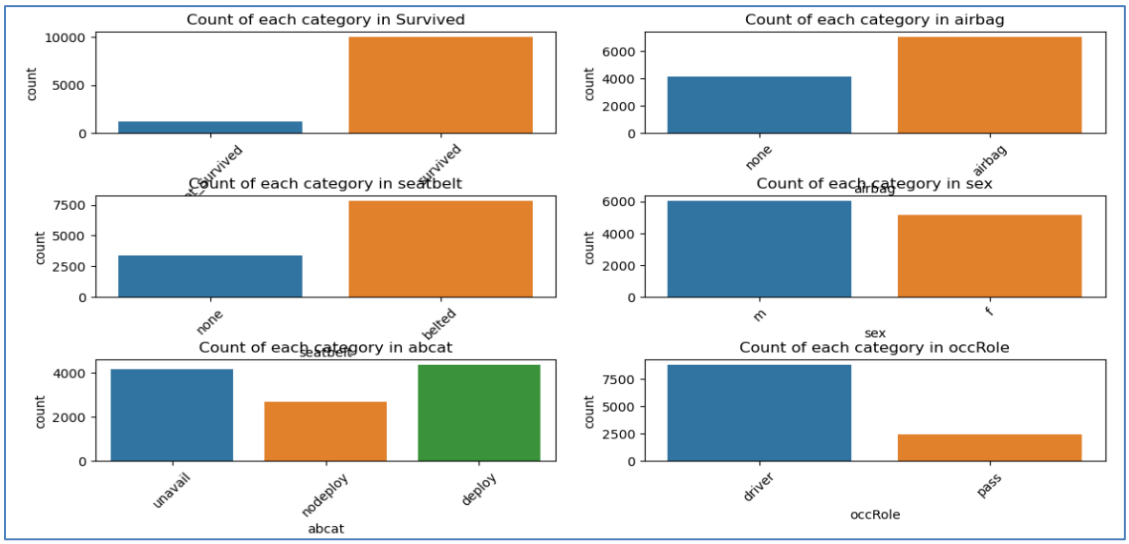


Figure:-11

Bivariate Analysis:-

The pair plot of the dataset is as follows:-

Figure:-12

Above pair plot, also known as a scatterplot matrix, is a grid of scatterplots showing the relationship between each pair of variables in a dataset

Heat Map:-

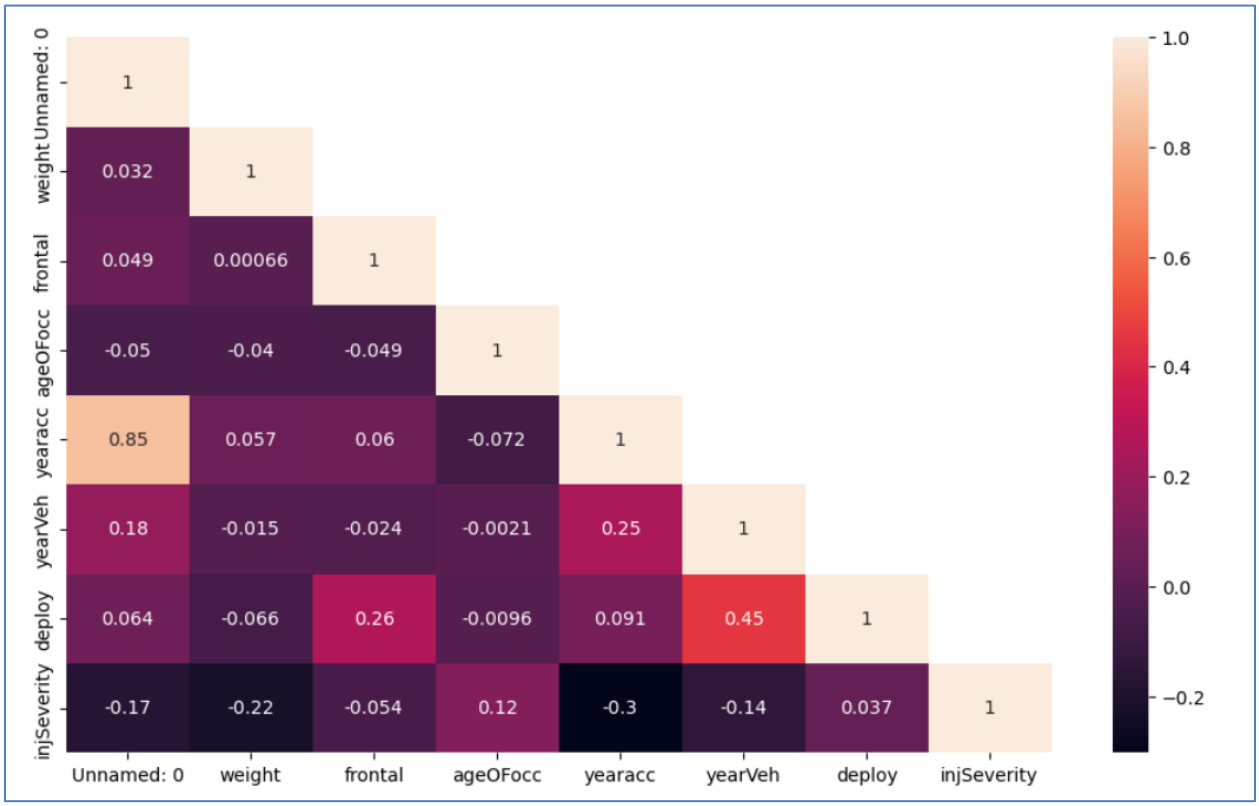


Figure:-13

Above heat map shows the correlation among the variables yearacc is highly correlated as its value is positive and color is light whereas negative value and dark color represent least correlation.

2.2 Encode the data (having string values) for Modeling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA

Encoding the data and converting the string values into numerical values the data will look like:-

	Unnamed: 0	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
0	0	4	27.078	0	0	0	1	1	32	1997	1987.0	2	1	0	4.0	1000
1	1	2	89.627	0	1	1	0	0	54	1997	1994.0	3	1	0	4.0	1010
2	2	4	27.078	0	0	1	1	1	67	1997	1992.0	2	1	0	4.0	1099
3	3	4	27.078	0	0	1	1	0	64	1997	1992.0	2	0	0	4.0	1099
4	4	4	13.374	0	0	0	1	1	23	1997	1986.0	2	1	0	4.0	3160

Table:-12

Now splitting the x and y into training and test dataset in the ratio 70:30. The total y train count and test count is as follows:-

```
1    0.89479
0    0.10521
Name: Survived, dtype: float64
```

```
1    0.894831
0    0.105169
Name: Survived, dtype: float64
```

Logistic Regression Model

Applying logistic regression on the model as follows :-

```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg',
                   verbose=True)
```

Predicting on Training and Test dataset

	0	1
0	0.024287	0.975713
1	0.001489	0.998511
2	0.001065	0.998935
3	0.000037	0.999963
4	0.017250	0.982750

Accuracy of the Training dataset:-

```
0.9816583874665648
```

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Compare both the models and write inferences, which model is best/optimized.

The ROC Curve of train Dataset for Logistic regression is as follows:-

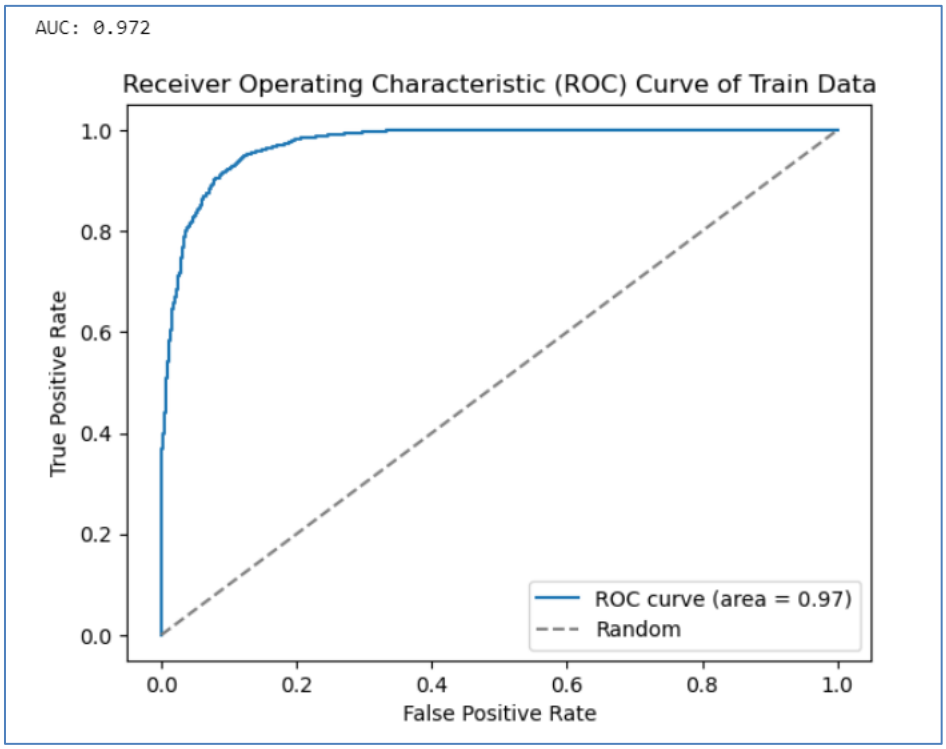


Figure:-14

Roc curve for test dataset for Logistic regression is as follows :-

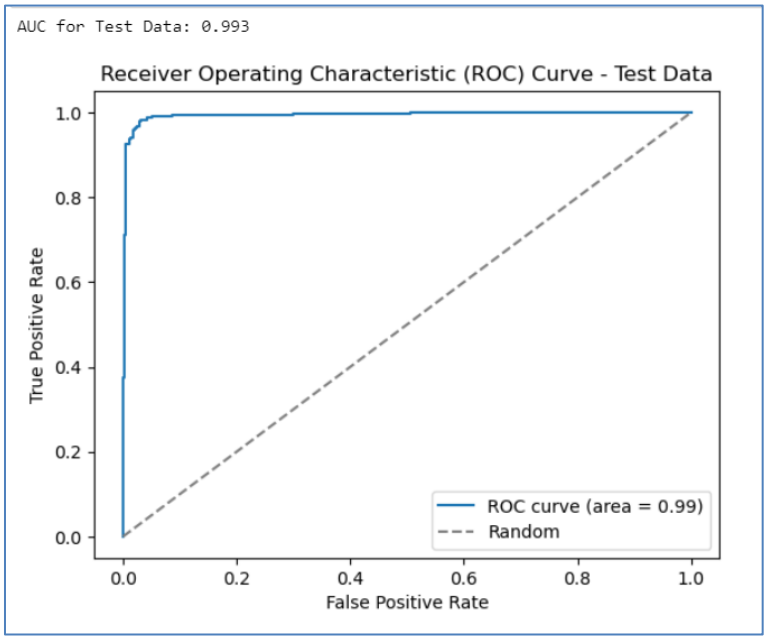


Figure:-15

Accuracy of the test dataset for logistic regression:-

```
0.9836601307189542
```

LDA Model

Confusion matrix of the train and test dataset for LDA is as follows:-

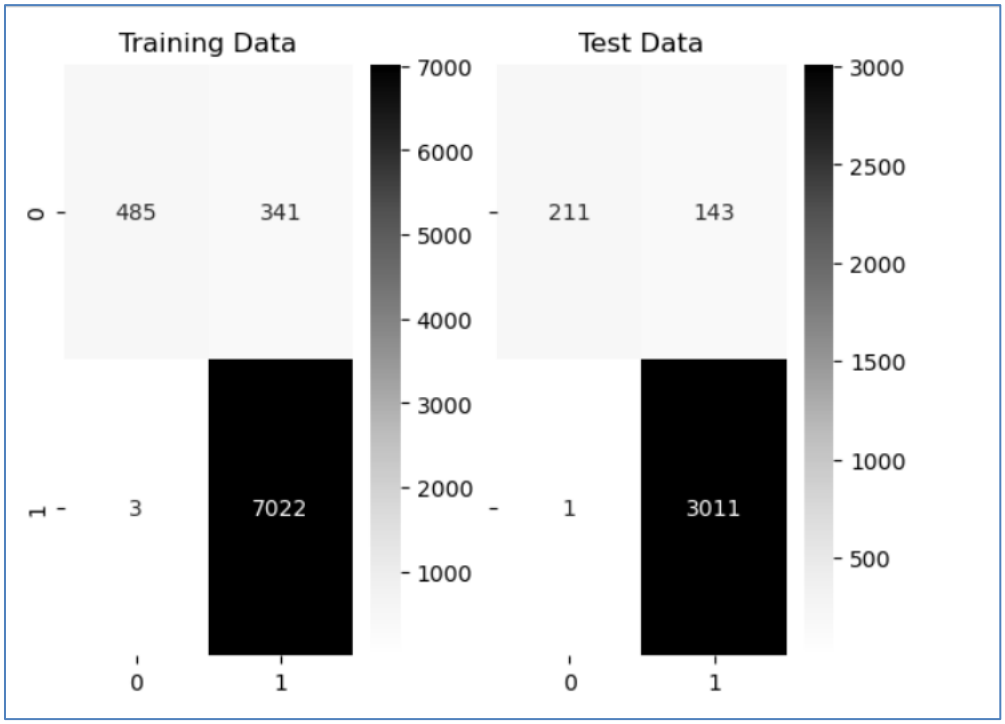


Figure:-16

From the above confusion matrix of the training and test data we can conclude that TP and TN values are very high as compared to FN and FP.

This indicates that the model is performing well in correctly identifying both the positive and negative instances. High TP and TN values suggest that the model is effectively capturing instances of both classes, while low FN and FP values indicate that the model is making fewer errors in misclassifying instances.

Classification report of the training and test dataset for LDA is as follows:-

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.99	0.59	0.74	826
1	0.95	1.00	0.98	7025
accuracy			0.96	7851
macro avg	0.97	0.79	0.86	7851
weighted avg	0.96	0.96	0.95	7851

Classification Report of the test data:

	precision	recall	f1-score	support
0	1.00	0.60	0.75	354
1	0.95	1.00	0.98	3012
accuracy			0.96	3366
macro avg	0.97	0.80	0.86	3366
weighted avg	0.96	0.96	0.95	3366

Table:-13

The above classification report for training and test data shows the precision, recall f1-score and support values of the training and test data. Overall, the model demonstrates robustness and generalization capabilities, achieving an accuracy of 96% on both training and test data, suggesting it's well-optimized and effectively captures patterns in the data.

ROC curve for training dataset for LDA is as follows:-

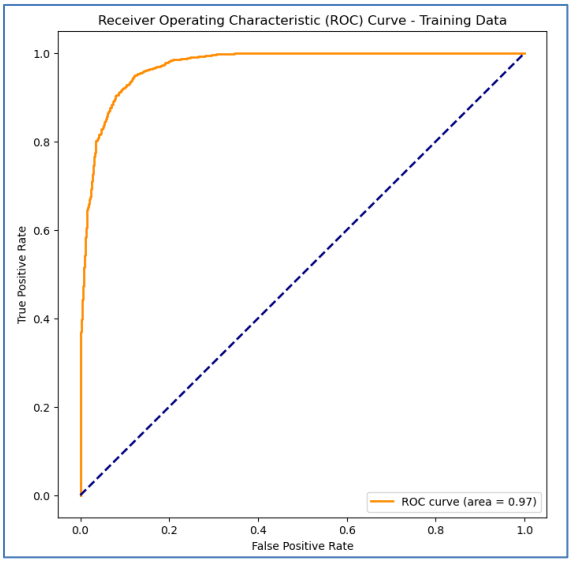


Figure:-17

ROC curve for test dataset for LDA is as follows:-

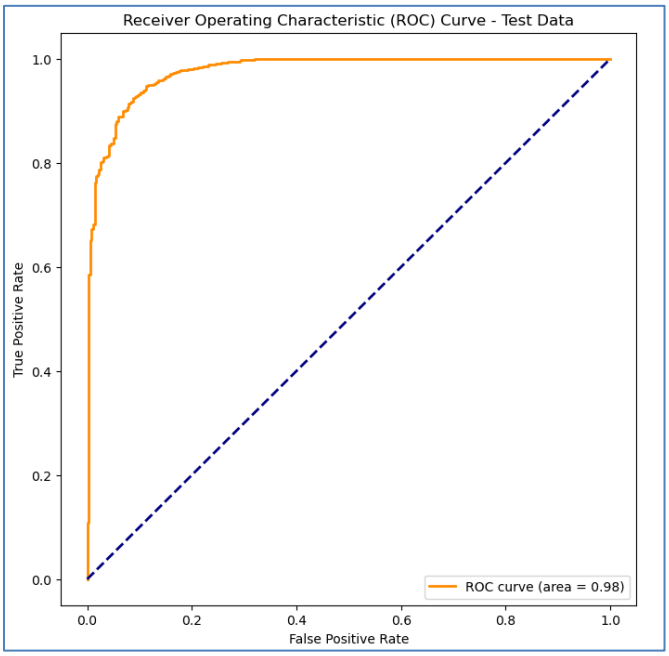


Figure:-18

ROC curve for train and test dataset is as follows:-

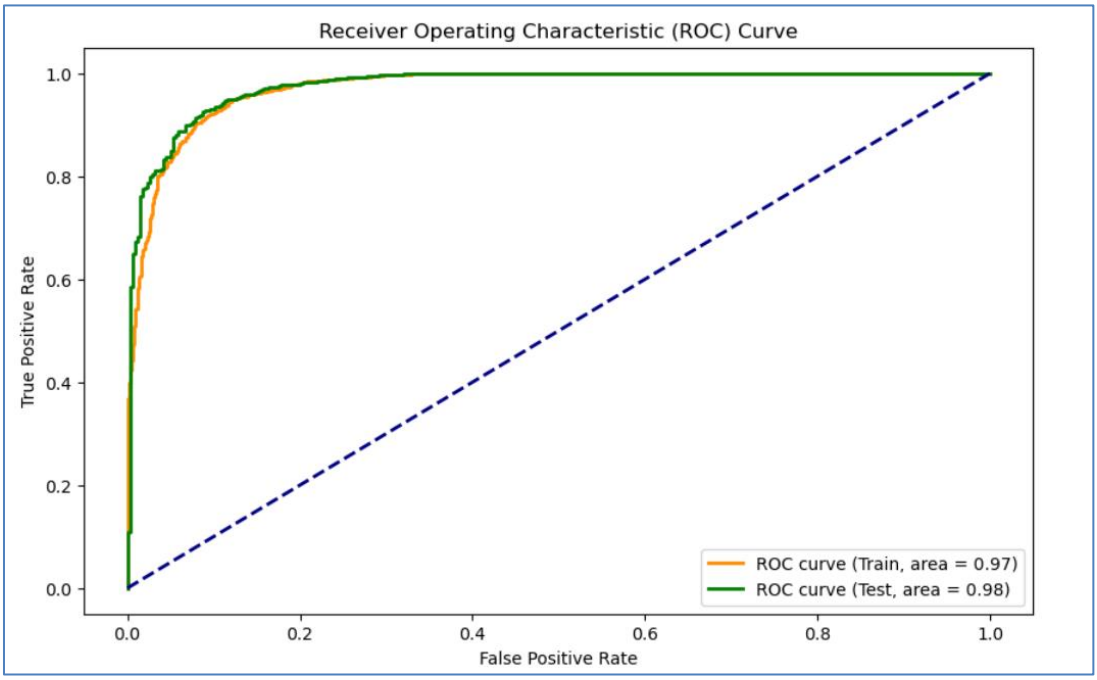


Figure:-19

2.4 Inference: Based on these predictions, what are the insights and recommendations?

INFERENCES:-

1) For logistic regression model AUC score for train and test dataset is as 0.97 and 0.99 which is higher whereas AUC scores for LDA for both the train and test are 0.97 and 0.98. Thus, Logistic regression proves to be better option.
2) Both models show high precision, recall, and F1-score values for the positive class (1), indicating strong performance in correctly identifying instances of the positive class. However, LDA appears to have lower recall for the negative class (0) compared to logistic regression, as evidenced by the lower precision value.
3) The logistic regression model is preferred for classification tasks in this scenario due to its higher AUC scores and better generalization performance.