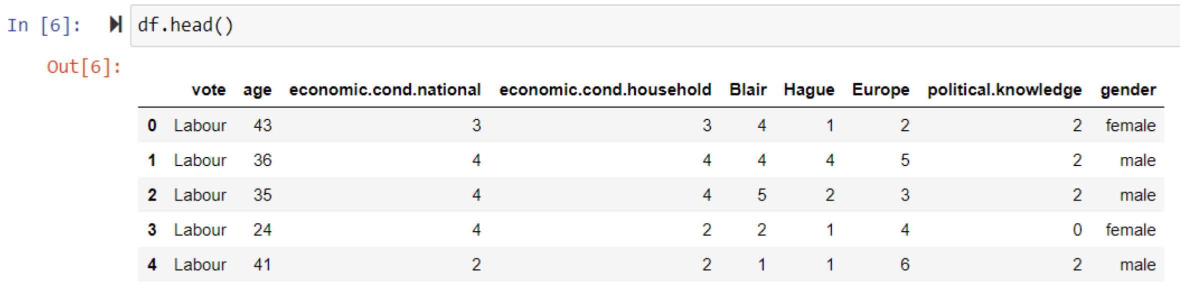
Project Machine Learning.

Name: Ajay Patwari

1. Read the dataset. Do the descriptive statistics and do null value condition check. Write an inference on it. (5 Marks)

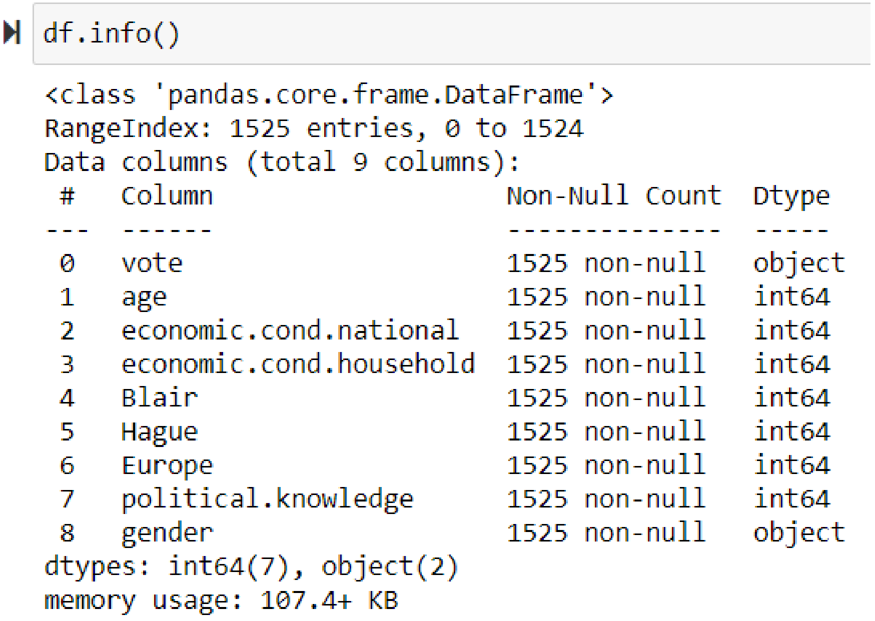


```
In [6]:  df.head()
Out[6]:
```

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	Labour	43	3	3	4	1	2	2	female
1	Labour	36	4	4	4	4	5	2	male
2	Labour	35	4	4	5	2	3	2	male
3	Labour	24	4	2	2	1	4	0	female
4	Labour	41	2	2	1	1	6	2	male

There are 1525 Rows and 9 Columns.

Except “Vote” and “Gender” all the columns are integer datatype.



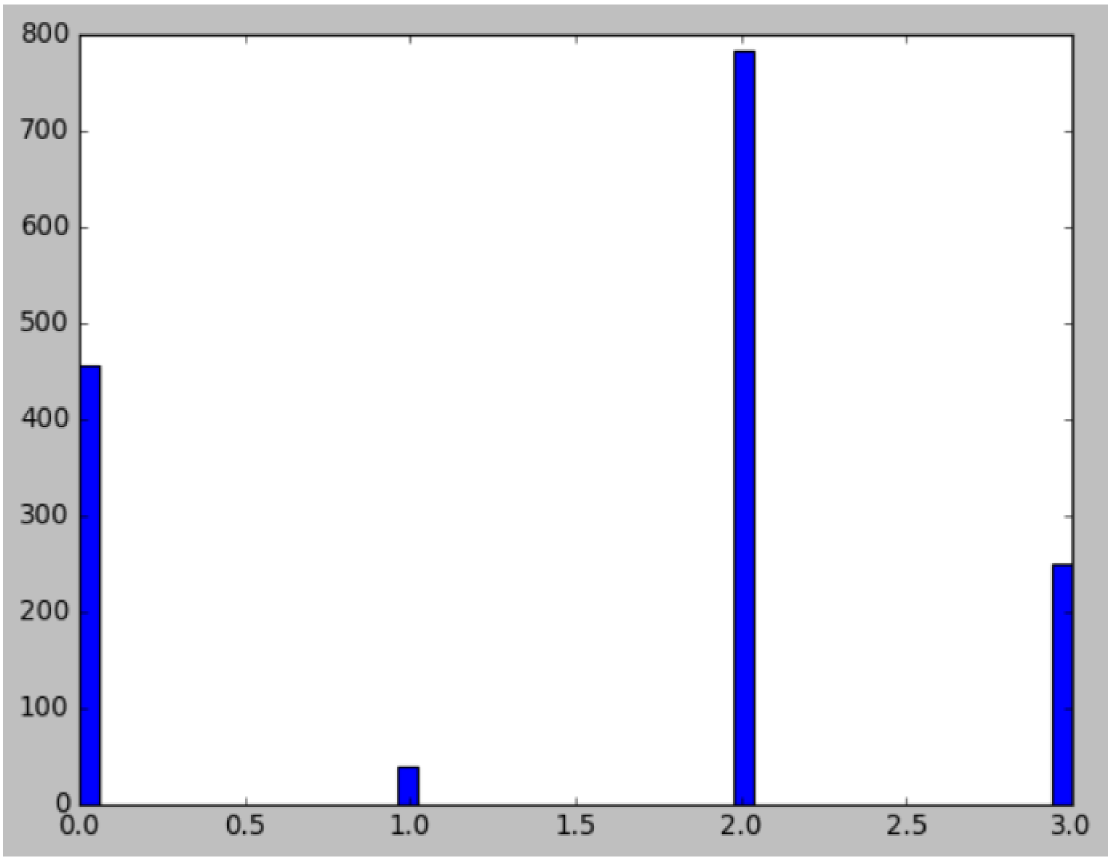
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   vote                     1525 non-null   object
 1   age                      1525 non-null   int64 
 2   economic.cond.national   1525 non-null   int64 
 3   economic.cond.household  1525 non-null   int64 
 4   Blair                    1525 non-null   int64 
 5   Hague                    1525 non-null   int64 
 6   Europe                   1525 non-null   int64 
 7   political.knowledge      1525 non-null   int64 
 8   gender                   1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

There are no missing values present in the data. I’ve checked for the outliers as well and there are no outliers present in the data.

There are 713 male voters and 812 Female voters.

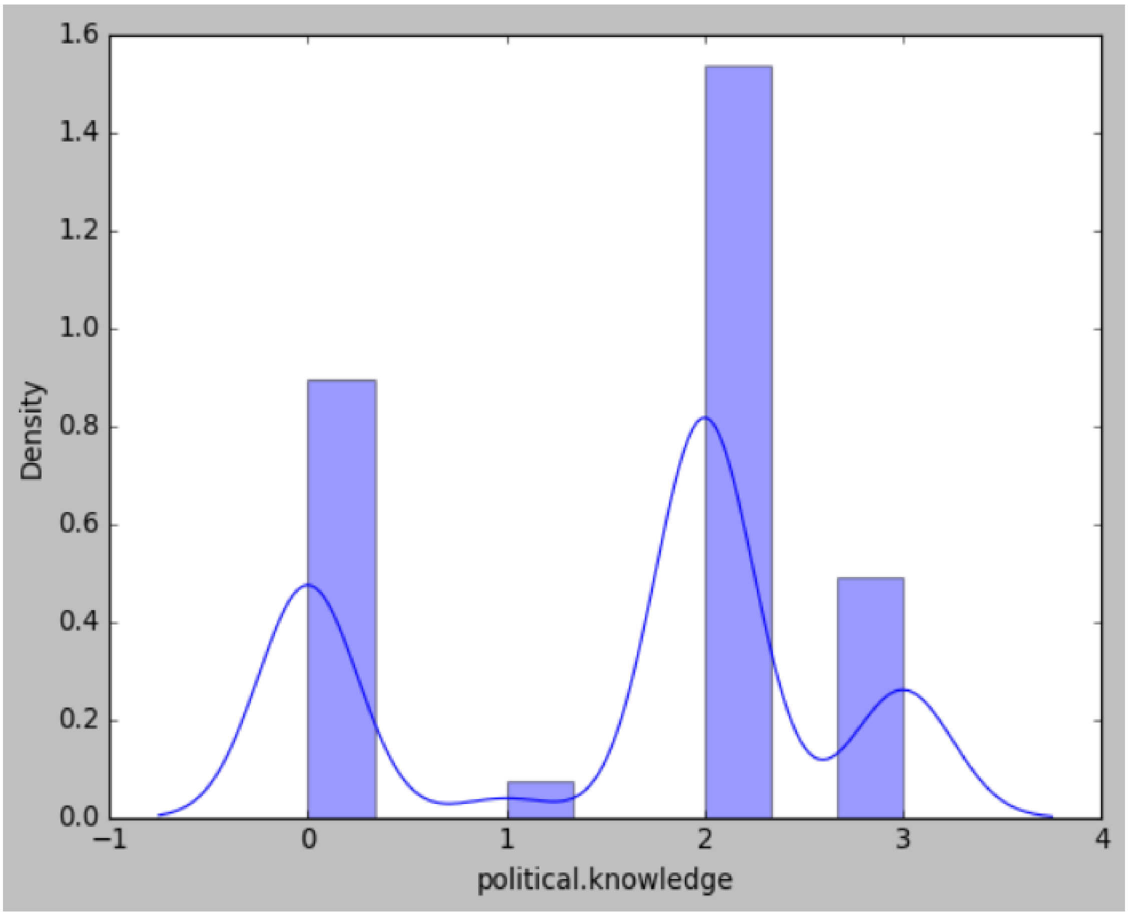
2. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

Below are the screenshots of the univariate analysis done on column name “Political Knowledge”



The above is the histogram figure for the selected column which shows the knowledge of party position in European integration. Here 3.0 indicated Eurosceptic sentiment.

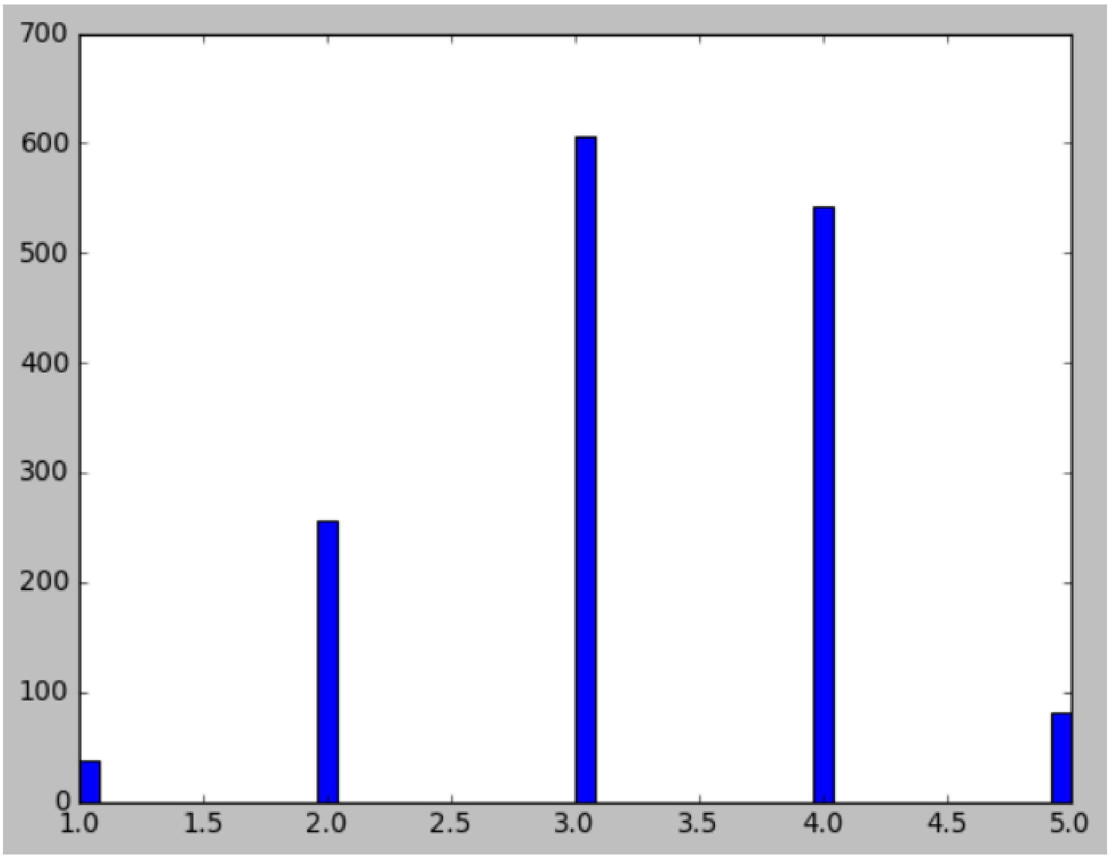
The next figure explains the distribution of the column “Political Knowledge”



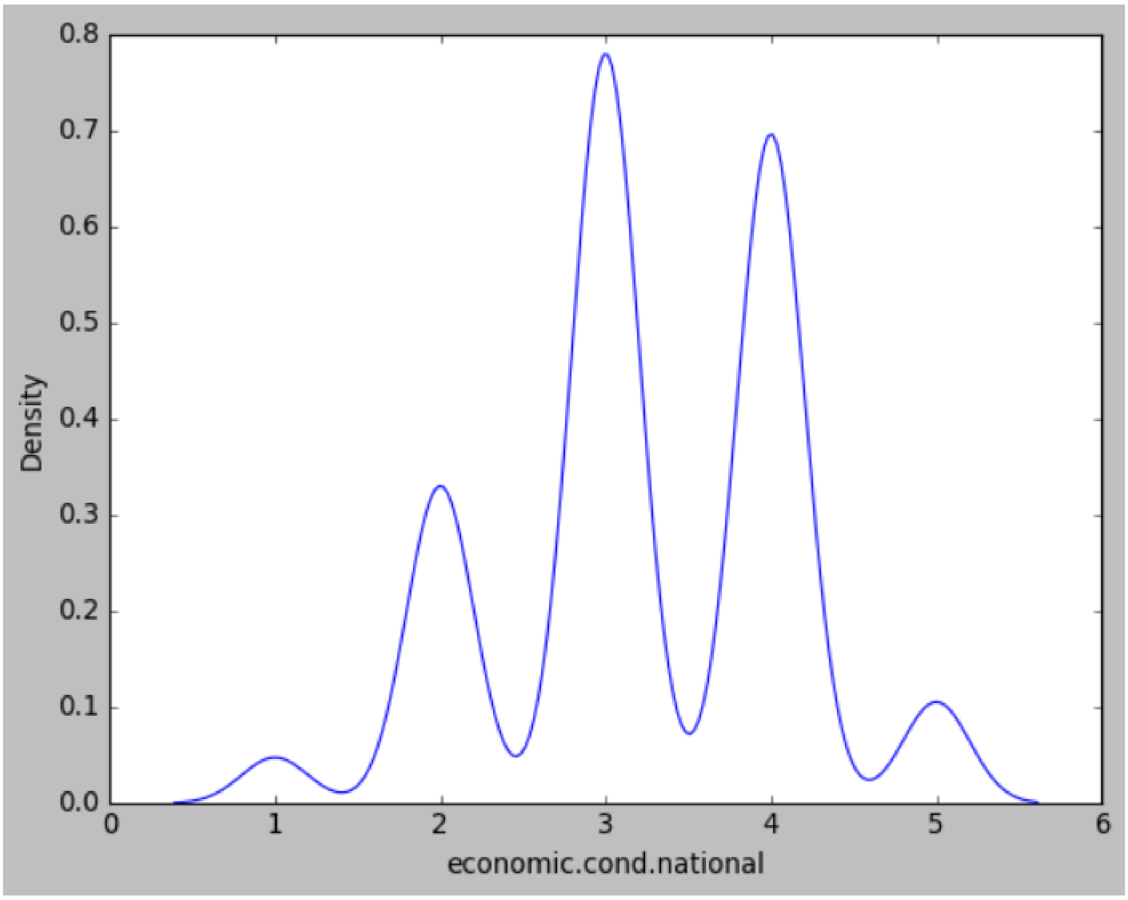
Below are the screenshots of the univariate analysis done on column name "economic.cond.national"

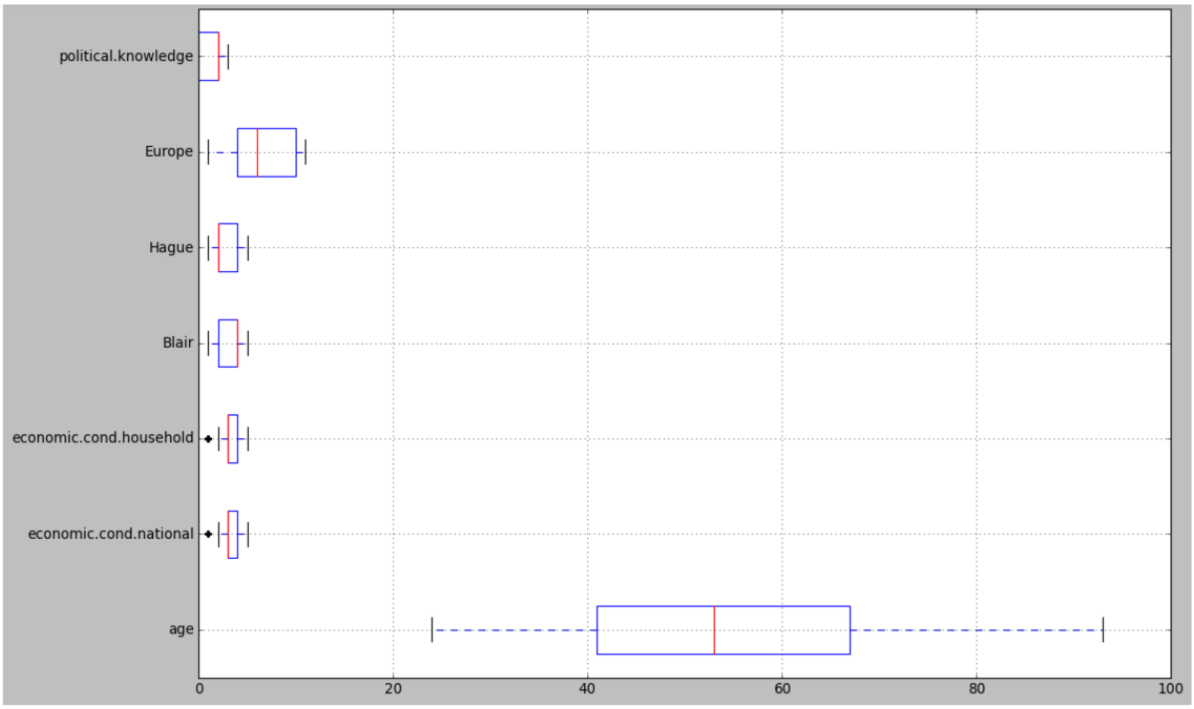
We have the histogram of the column shown in the next page. It is gauged on the scale of 1 to 5.

This column contains 1525 rows of data. There are no missing values present in this data.

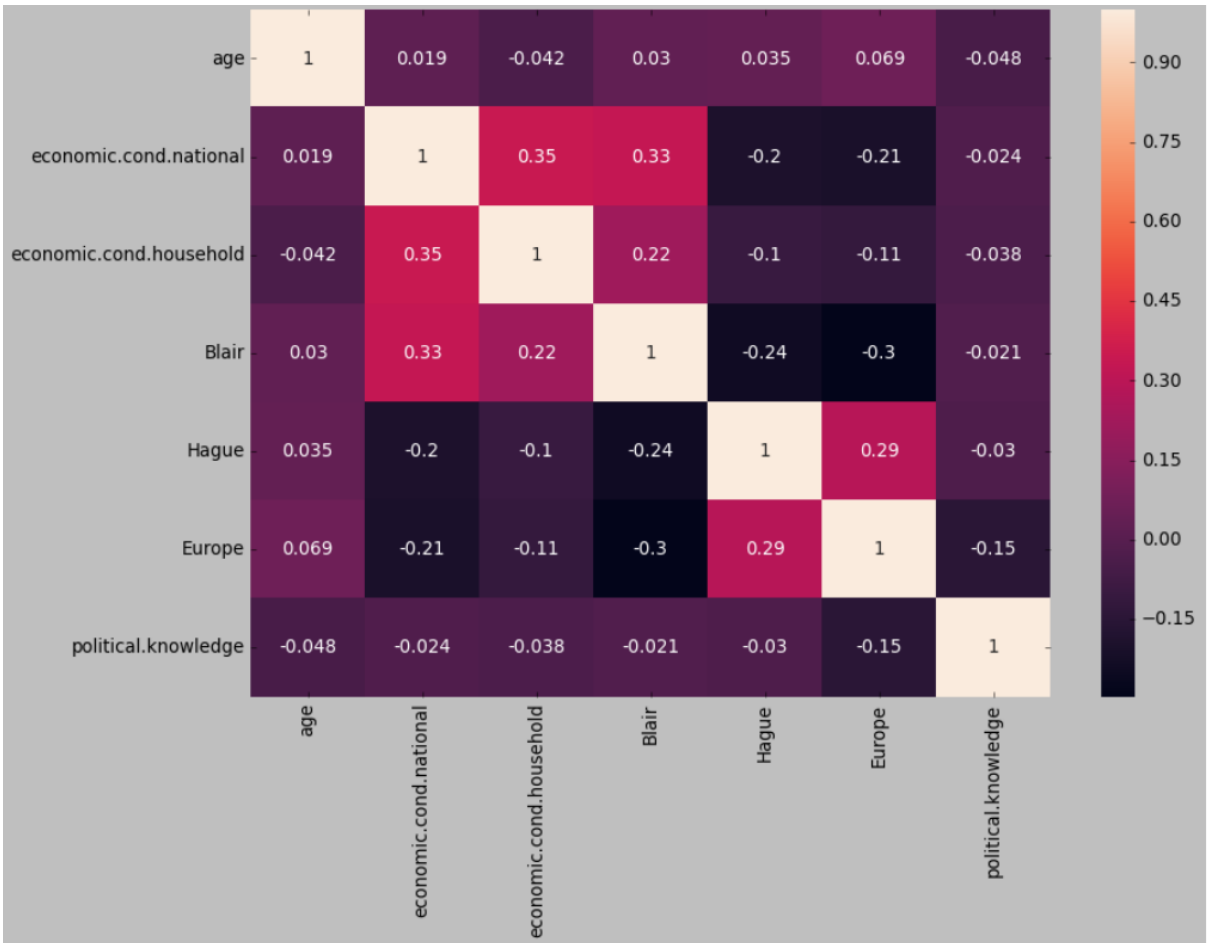


Next screenshot shows the exact distribution of the data.





The above figure is a box plot which indicates that there are no missing values present in our data.

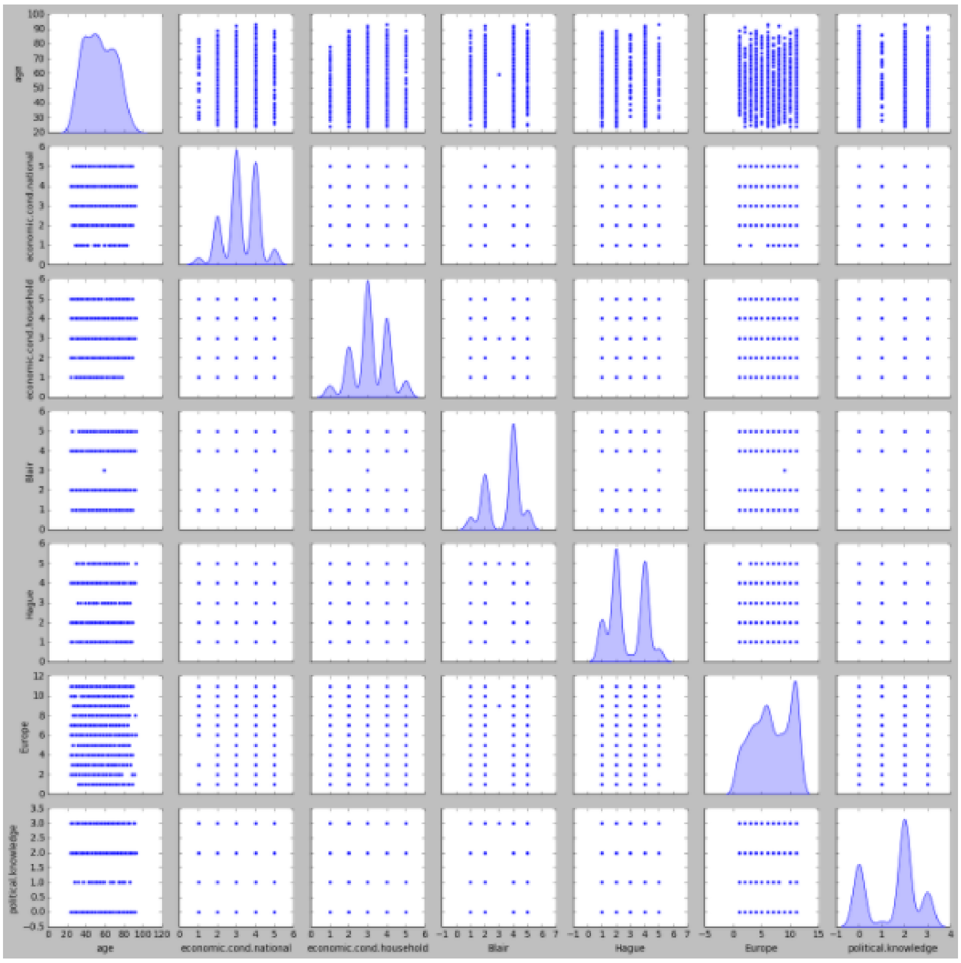


	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
age	1	0.019	-0.042	0.03	0.035	0.069	-0.048
economic.cond.national	0.019	1	0.35	0.33	-0.2	-0.21	-0.024
economic.cond.household	-0.042	0.35	1	0.22	-0.1	-0.11	-0.038
Blair	0.03	0.33	0.22	1	-0.24	-0.3	-0.021
Hague	0.035	-0.2	-0.1	-0.24	1	0.29	-0.03
Europe	0.069	-0.21	-0.11	-0.3	0.29	1	-0.15
political.knowledge	-0.048	-0.024	-0.038	-0.021	-0.03	-0.15	1

The heatmap above explains the exact correlation between different columns and how related they are.

From the figure we do not see many highly related columns, however the best we have is 0.35 between Economic cond household and Economic Cond national.

Other all columns are related with a very low percentage.



Above is the pair plot which shows the relation between columns in a scatter plot form.

You can find the detailed version of my Exploratory data analysis in my code book. I have attached only limited screenshots of the EDA here.

1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(3 pts), Data Split: Split the data into train and test (70:30) (2 pts).

I have encoded the data which has string values. Only two columns vote and gender had to be encoded and I have get used get dummies and replaced them.

Yes I scaling is required for this data because all the columns has values from 1 to 5 but whereas the age columns has high values up to 70’s. Hence we might get incorrect predictions if we hadn’t done the scaling. Hence I have done scaling to the data using Min Max scaler.

I have split the data into train and test. All these can be seen in the code I've attached.

1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (3 pts). Interpret the inferences of both models (2 pts)

I have applied the Logistic Regression and Linear Discriminant Analysis.

Training data accuracy for Logistic Regression: 84%

Testing data accuracy for Logistic Regression: 82%

Training data accuracy for LDA: 83%

Testing data accuracy for LDA: 84%

The confusion Matric, ROC curve and AUC has also been done for both the models and I shall add those results in the 1.7 Question.

1.5) Apply KNN Model and Naïve Bayes Model(5 pts). Interpret the inferences of each model (2 pts)

Both the KNN Model and Naïve Bayes Model has been applied on the training and testing data.

For Naïve Bayes Model:

We have achieved an accuracy of 82% on training data and 85% accuracy on testing data.

For KNN Model:

We have achieved 87% accuracy on training data and 85% accuracy on testing data.

All the other performance metrics has been explained in the question 1.7.

1.6) Model Tuning (2 pts) , Bagging (2.5 pts) and Boosting (2.5 pts).

I've done the Model Tuning, Bagging and Boosting. I've done ADA boosting as well as Gradient Boosting.

Accuracy of Training data when bagging is used: 100%

Accuracy of Testing data when bagging is used: 82%

Accuracy of Training data when Ada boosting is used: 84%

Accuracy of Testing data when bagging is used: 84%

Accuracy of Training data when Gradient boosting is used: 89%

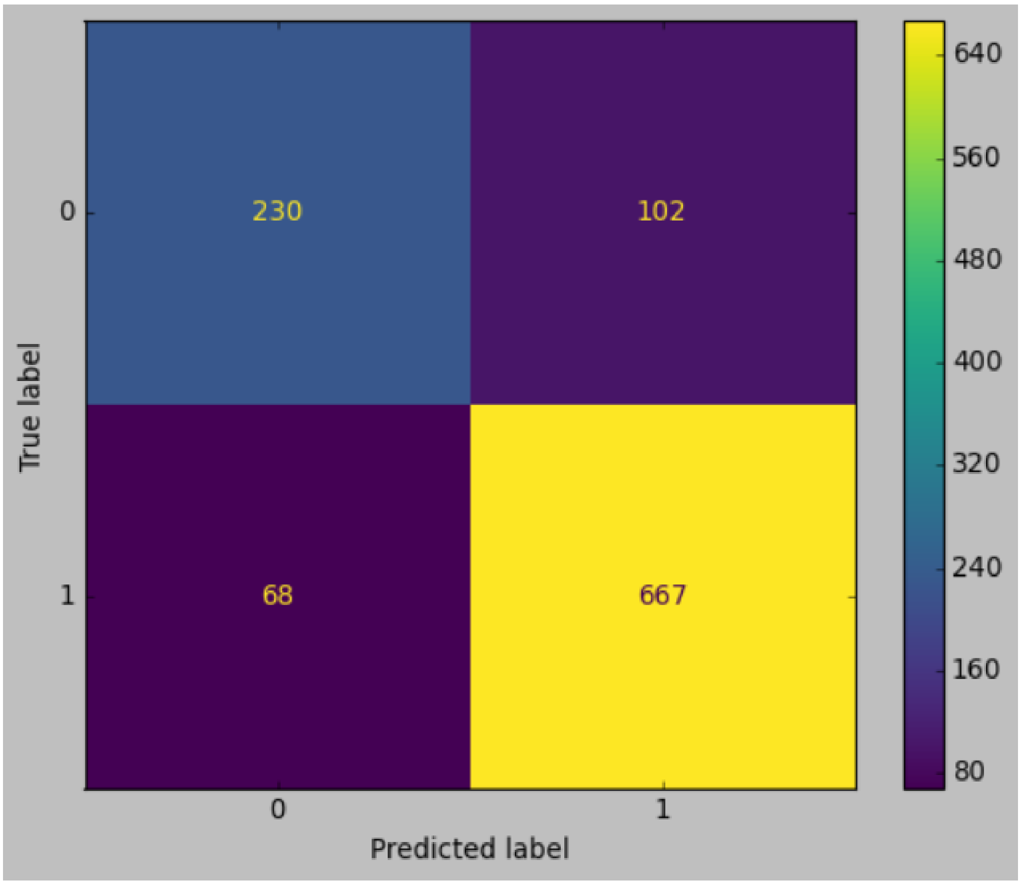
Accuracy of Testing data when bagging is used: 84%

1.7 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 (4 pts) Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized (3 pts)

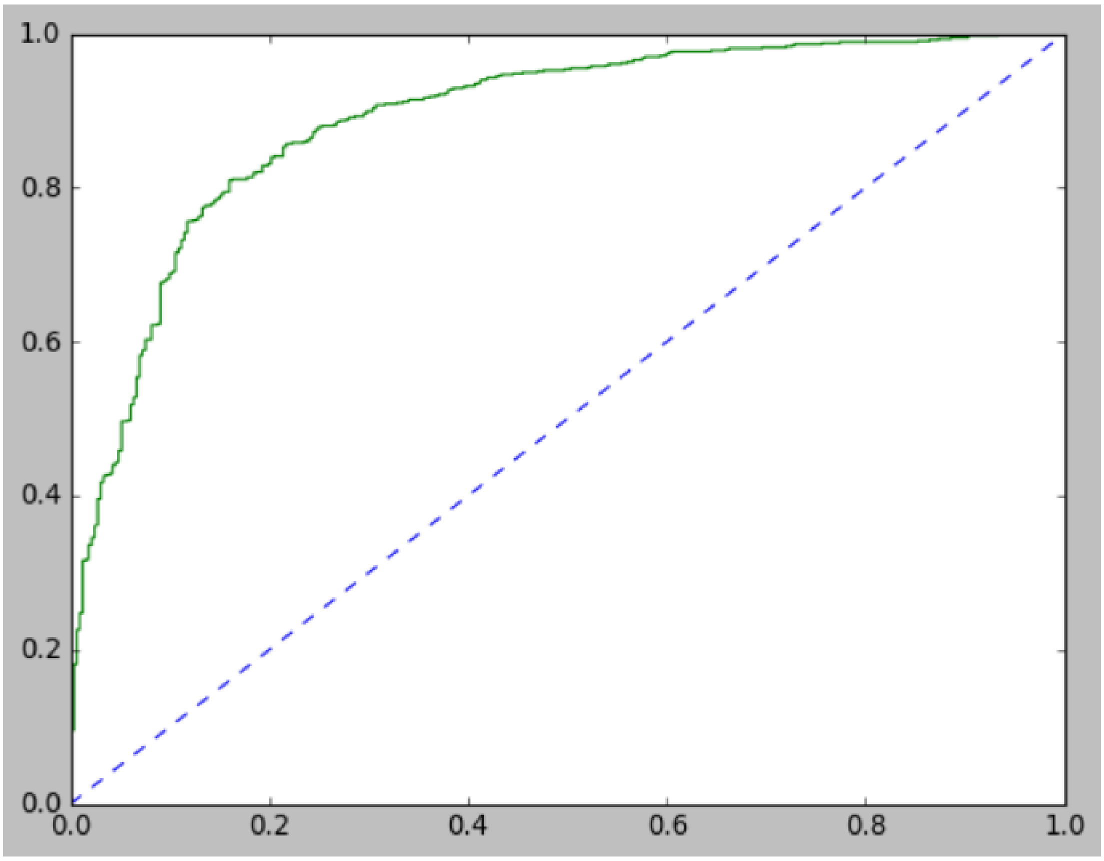
Performance Metrics of Linear Regression Training Data:

Accuracy: 84%

Confusion Matrix:



ROC Curve:

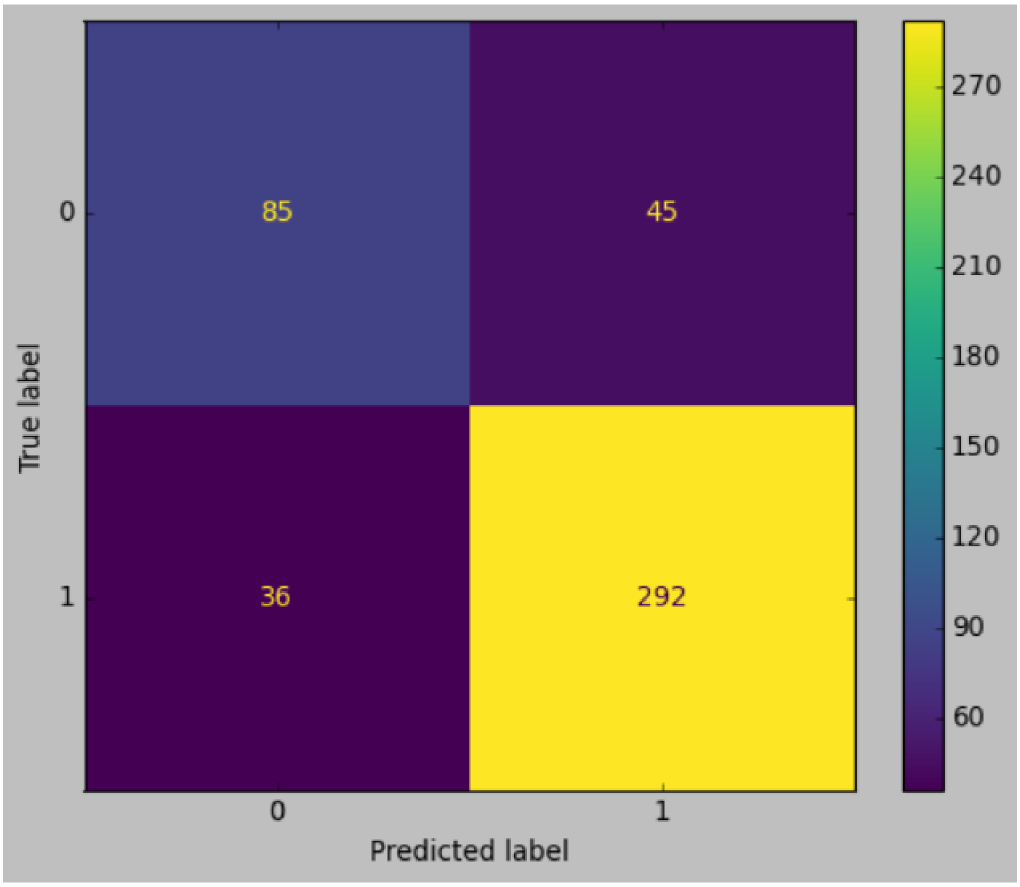


AUC Score: 89%

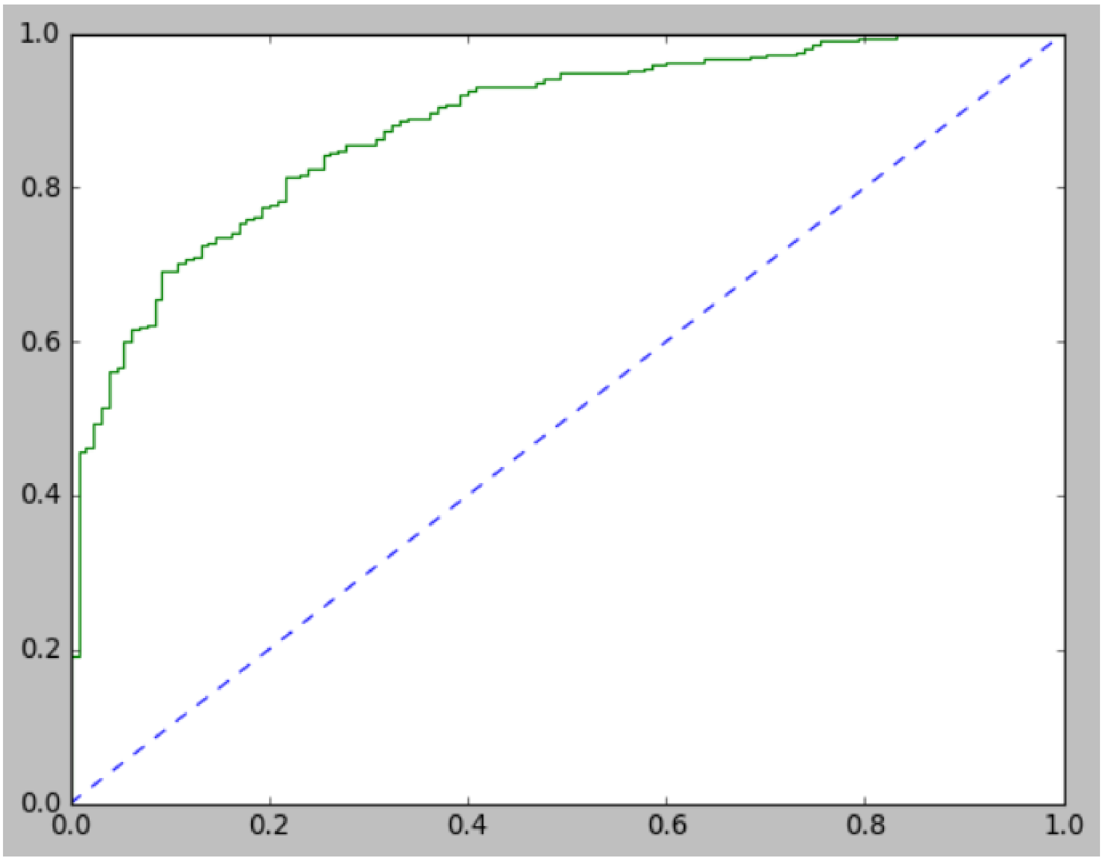
Performance Metrics of Linear Regression Testing Data

Accuracy: 82%

Confusion Matrix:


85
45
36
292
0
1
True label
0
1
Predicted label
270
240
210
180
150
120
90
60

ROC Curve:



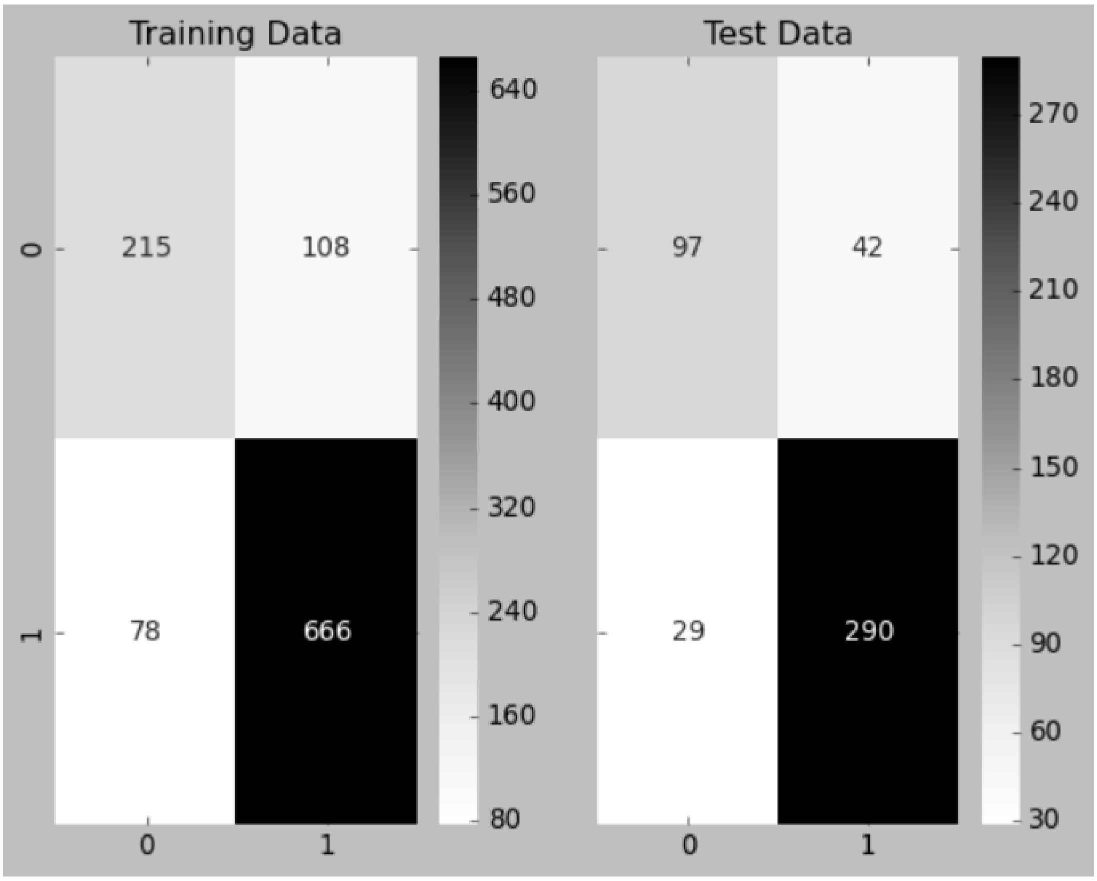
AUC Score: 89%

Performance Metrics of LDA Training and Testing Data:

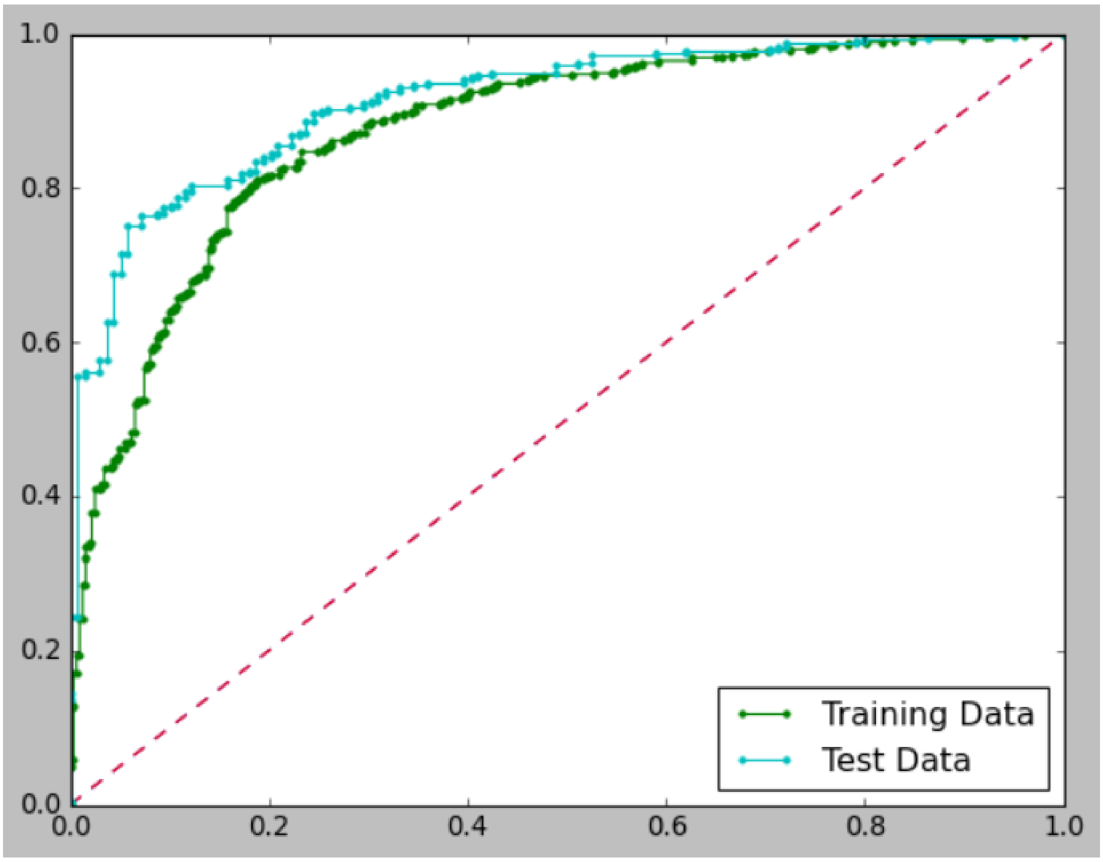
Accuracy Training Data: 83%

Accuracy Testing Data: 84%

Confusion Matrix:


Training Data
Test Data
0
1
215
108
78
666
97
42
29
290
640
560
480
400
320
240
160
80
270
240
210
180
150
120
90
60
30

ROC Curve:



AUC Score Training Data: 88%

AUC Score Testing Data: 91%

Performance Metrics of Naïve Bayes Model:

Accuracy for Training Set: 82%

Accuracy for Testing Set: 85%

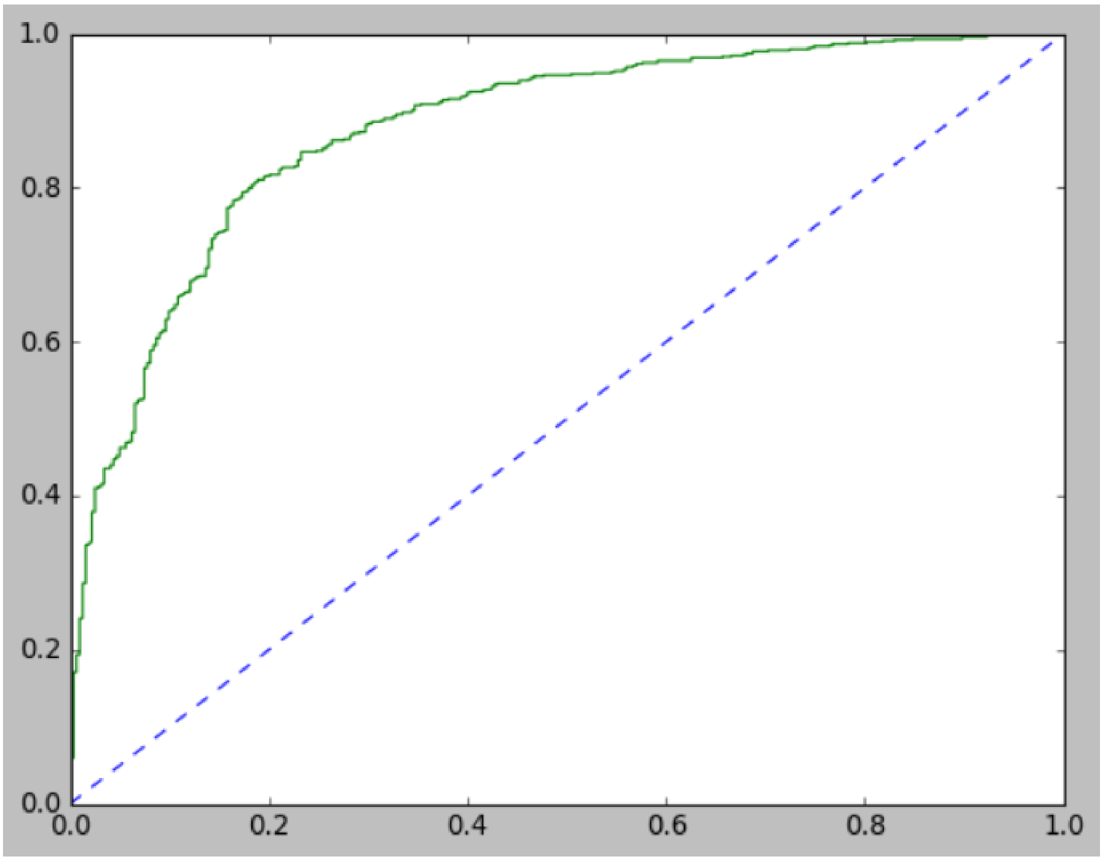
Confusion Matrix for Training Set: [[101 38]

 [32 287]]

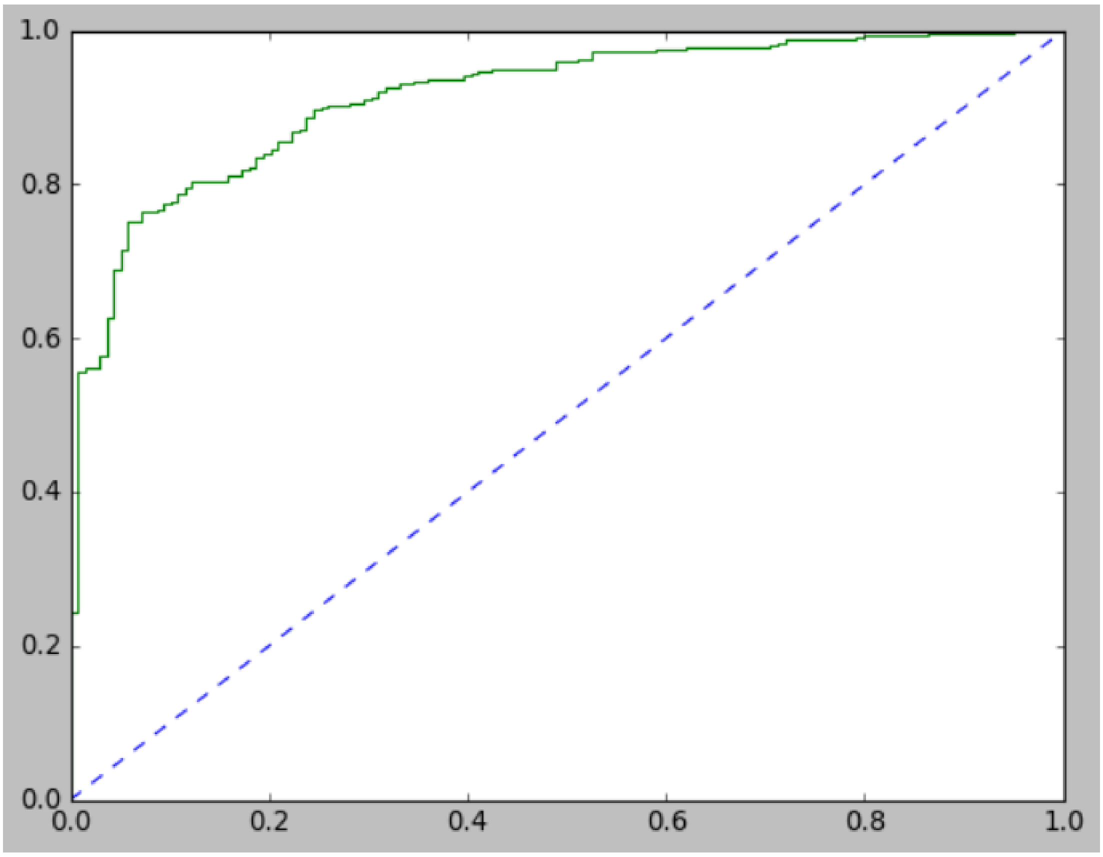
Confusion Matrix for Testing Set: [[238 85]

 [56 688]]

ROC Curve for Training Set:



ROC Curve for Testing Set:



AUC Score for Training Set: 88%

AUC Score for Testing Set: 88%

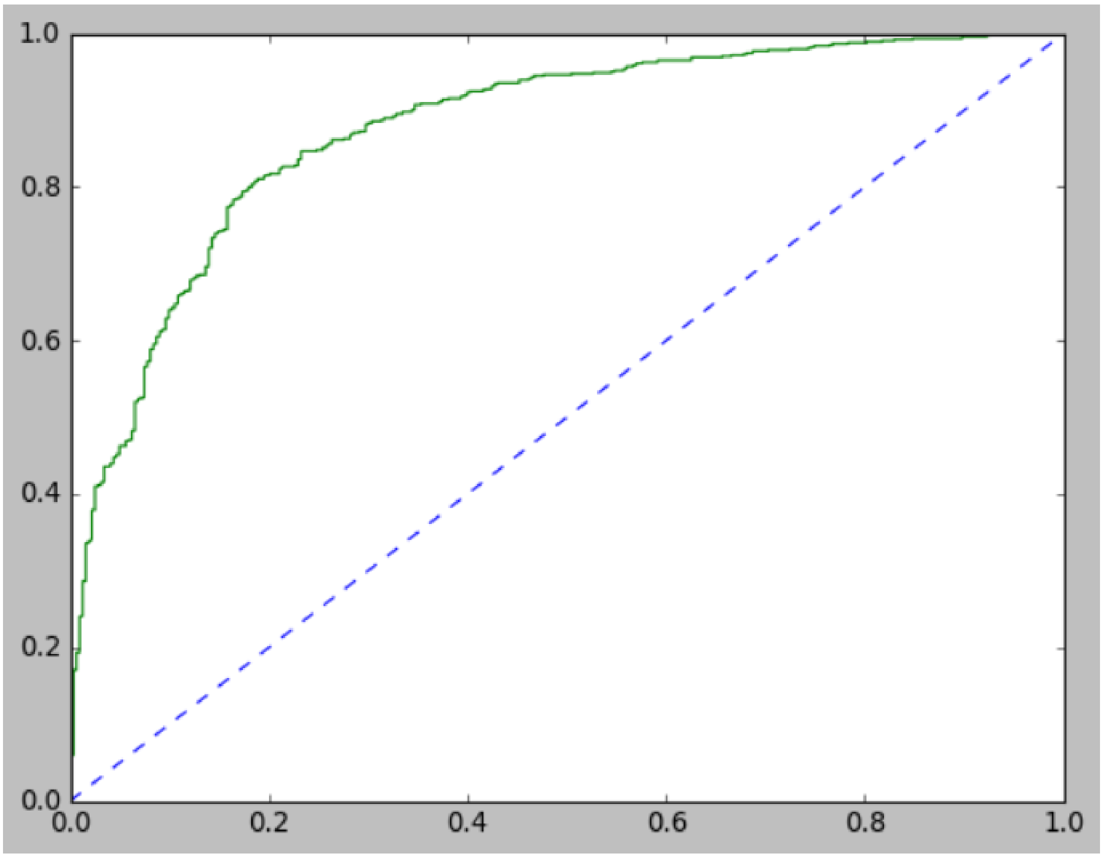
Performance Metrics of KNN Model:

Accuracy for Training Set: 87%

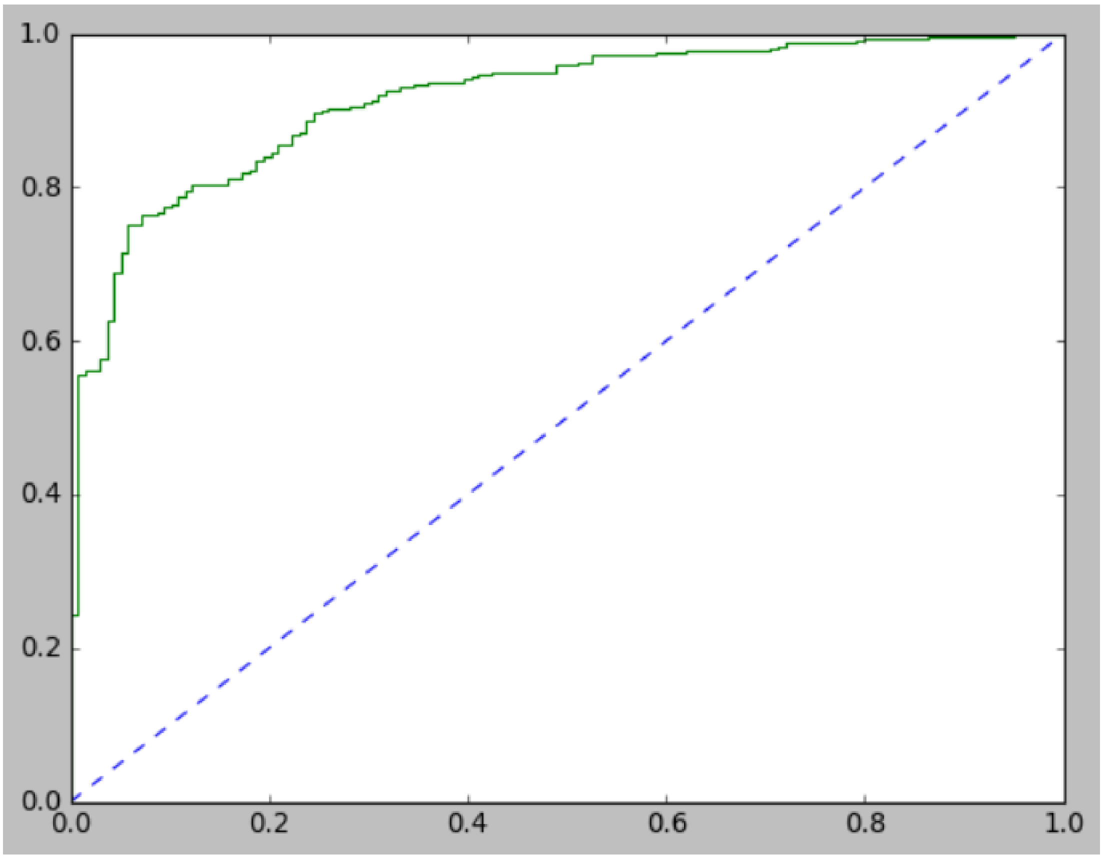
Accuracy for Testing Set: 85%

```
Confusion Matrix for Training Set: [[238  85]
 [ 56 688]]
Confusion Matrix for Testing Set: [[101  38]
 [ 32 287]]
```

ROC Curve for Training Set:

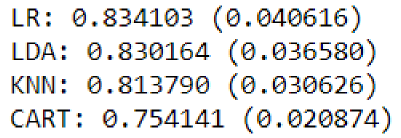


ROC Curve for Testing Set:

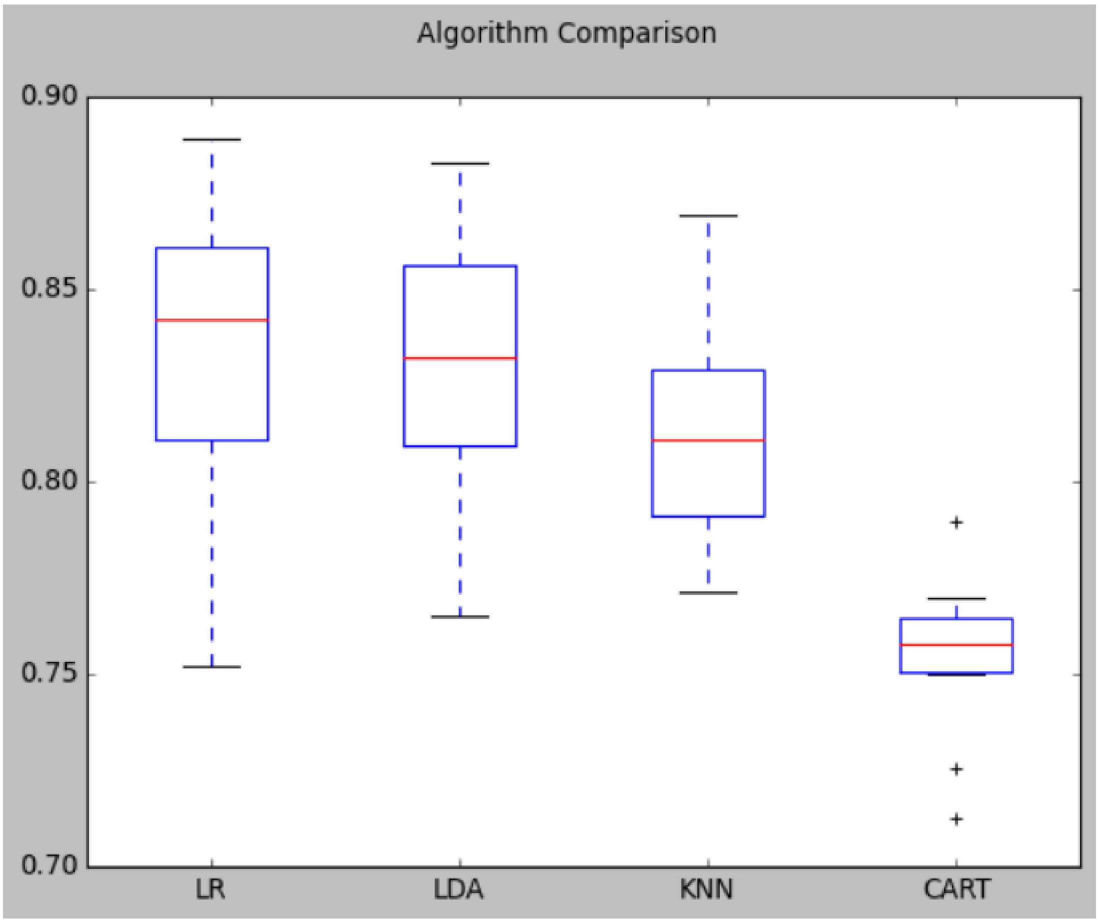


AUC Score for Training Set: 87%

AUC Score for Testing Set: 87%



```
LR: 0.834103 (0.040616)
LDA: 0.830164 (0.036580)
KNN: 0.813790 (0.030626)
CART: 0.754141 (0.020874)
```



We have not observed a huge difference in the models accuracy but we can clearly observe that the ADA Boosting model has a high amount of accuracy. We have observed 84% Accuracy with high precision.

1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.

From all the models we have predicted that the votes for Labour party are higher than the votes for Conservative part. We have observed this from all the confusion matrix that out prediction was highly accurate the Labour party has more votes compared to the Conservative party.

We can also tell that the current Economic Condition national has been one of the main factor for why Labour party has received more votes.

Also next important factor was the candidate Blair, we have assessed that blair has more popularity in people and hence we can confirm by using our prediction models that labour party might win the actual elections as well.

2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)

Number of Characters in Roosevelt File: 7571

Number of Characters in Kennedy File: 7618

Number of Characters in Nixon File: 9991

Number of words in Roosevelt File: 1360

Number of words in Kennedy File: 1390

Number of words in Nixon File: 1819

Number of sentences in Roosevelt File: 67

Number of sentences in Kennedy File: 52

Number of sentences in Nixon File: 68

2.2) Remove all the stopwords from the three speeches.

I have used the library from nltk.corpus import stopwords

from nltk.tokenize import word_tokenize.

Using this I have removed all the stop words.

2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)

Top 3 words used in Roosevelt speech: Nation, Know, Spirit

Top 3 words used in Kennedy speech: Let, Us, World

Top 3 words used in Nixon speech: Us, Let, America