PROBLEM 1

You are hired by one of the leading news channel CNBE who wants to analyse recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

1.1) Read the dataset. Do the descriptive statistics and do null value condition check.

Data is ingested and following is the glimpse of the data: -

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

Following is a brief about the variables: -

1. Vote: It is the party choice, labour or conservative.

2. Age: Represents the age of the voter

3.Economic.cond.national: What the voter assesses of the current national economic conditions, on a scale of 1 to 5

4.Economic.cond.household: What the voter assesses of the current house hold economic conditions, on a scale of 1 to 5.

5.Blair: The rating of a candidate named Blair, from the Labour Party, on a scale of 1 to 5.

6.Hague: The rating of a candidate named Hague, from the Conservative Party, on a scale of 1 to 5.

7.Europe: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.

8.Political knowledge: Knowledge of parties' positions on European integration, 0 to 3.

9.Gender: Gender of voter

Basic Info: -

```
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
```

Evident from above that, there are two object data types, rest are integers, with 1525 non null values each.

Shape of data: -

```
df.shape
```

```
(1517, 9)
```

The data consists of 1517 rows and 9 columns.

Confirming Null values: -

	Null Values
vote	0
age	0
economic.cond.national	0
economic.cond.household	0
Blair	0
Hague	0
Europe	0
political.knowledge	0
gender	0

Clearly, proved, that there are no null values present.

Checking for duplicates: -

```
print('No of duplicates in the data:',df.duplicated().sum())
```

```
No of duplicates in the data: 8
```

```
df.drop_duplicates(inplace = True)
print('No of duplicates after deleting:', df.duplicated().sum())
```

```
No of duplicates after deleting: 0
```

As evident, the relevant duplicates present in the data were cleaned effectively.

Description/Summary of Data: -

	count	mean	std	min	25%	50%	75%	max	IQR	COV
age	1525.0	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0	26.0	0.289969
economic.cond.national	1525.0	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0	1.0	0.271410
economic.cond.household	1525.0	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0	1.0	0.296132
Blair	1525.0	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0	2.0	0.352332
Hague	1525.0	2.746885	1.230703	1.0	2.0	2.0	4.0	5.0	2.0	0.448036
Europe	1525.0	6.728525	3.297538	1.0	4.0	6.0	10.0	11.0	6.0	0.490083
political.knowledge	1525.0	1.542295	1.083315	0.0	0.0	2.0	2.0	3.0	2.0	0.702404

(The columns “IQR” and “COV” are explicitly added)

As it seems from the descriptive analysis, most of the values appear to be distributed nearly normally. The COV confirms presence of uniformity in the data.

1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each

As far as the NULL values are concerned, there were none found(NULL value check done in the previous question)

```
df.isnull().sum().sum()
```

```
0
```

The above code snippet confirms presence of no outliers.

```
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
```

The Basic info shows that, there are 1525, non-null values. Vote and Gender are the only object type variables, while rest others are integers.

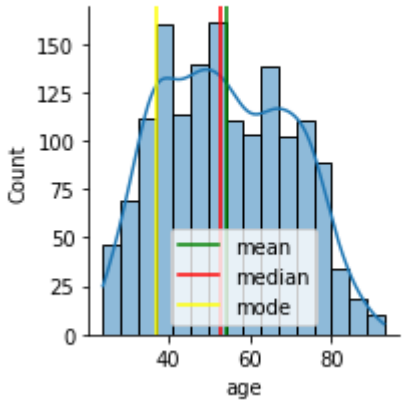
Univariate analysis: -

1. Age:

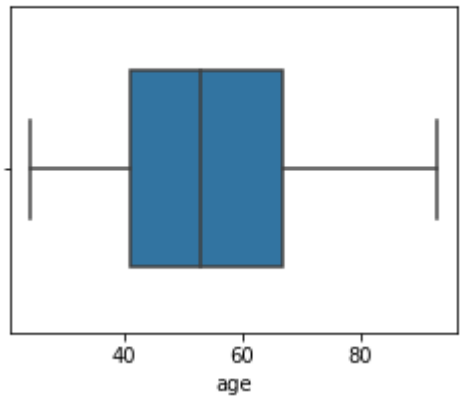
```
Description of age  is: -
count    1517.000000
mean       54.241266
std        15.701741
min        24.000000
25%        41.000000
50%        53.000000
75%        67.000000
max        93.000000
Name: age, dtype: float64

--------------------------------
Mean is:  54.2412656558998
Median is:  53.0
Mode is:  37
--------------------------------
```

Distribution of age is: -



Boxplot of age is: -



Mean and Median are quite close apart, while the mode taking a little backward seat. The std deviation is less than the mean, giving it a uniformity. The Data is close to normal distribution.

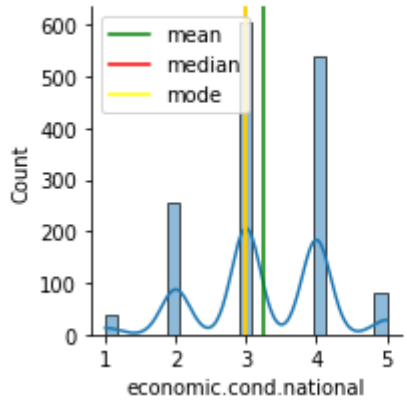
There are no outliers present.

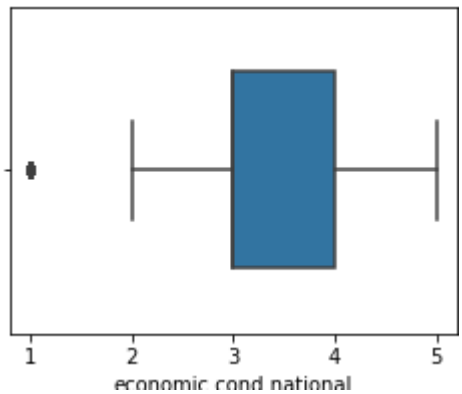
2. economic.cond.national

```
Description of economic.cond.national  is: -
count    1517.000000
mean        3.245221
std         0.881792
min         1.000000
25%         3.000000
50%         3.000000
75%         4.000000
max         5.000000
Name: economic.cond.national, dtype: float64

------------------------------------------------
Mean is:  3.245220830586684
Median is:  3.0
Mode is:  3
------------------------------------------------
```

Distribution of economic.cond.national is: -





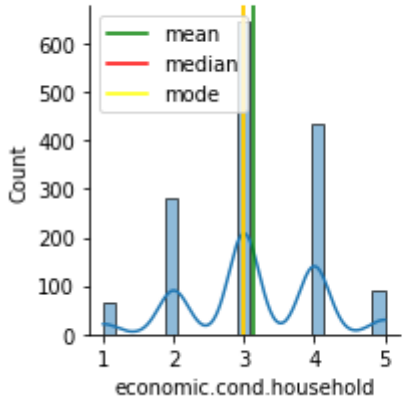
It's a discrete variable. Mean, median and Mode are close apart, with the maximum rating given being 3.

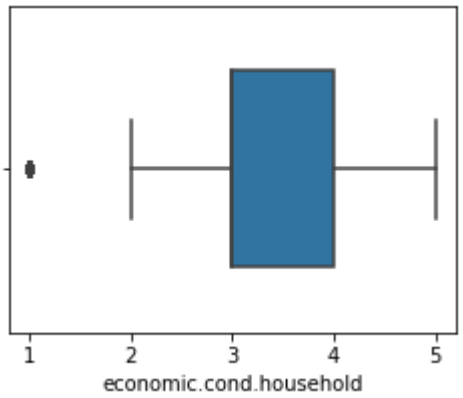
3. Econoic.cond.household

```
Description of economic.cond.household  is: -
count    1517.000000
mean        3.137772
std         0.931069
min         1.000000
25%         3.000000
50%         3.000000
75%         4.000000
max         5.000000
Name: economic.cond.household, dtype: float64

------------------------------------------------
Mean is:  3.1377719182597232
Median is:  3.0
Mode is:  3
------------------------------------------------
```

```
Distribution of  economic.cond.household  is: -
```





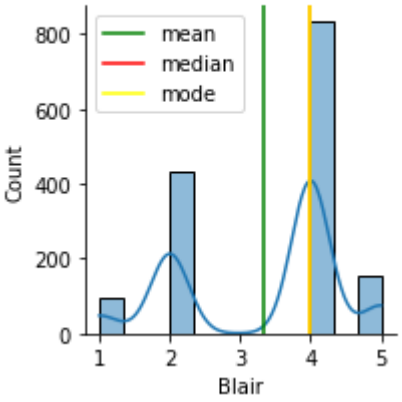
It’s a discrete variable, with mean, mode and median almost overlapping. Since the variable represents ratings, there can be outliers which would form the essential part of the data. Here a single outlier is present.

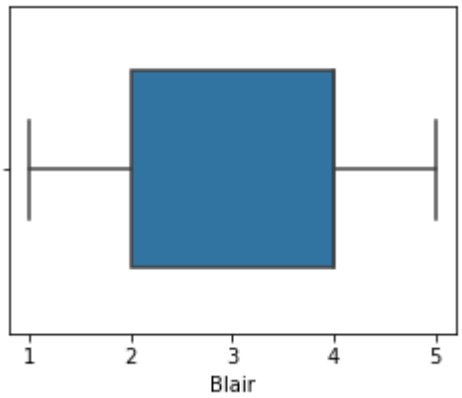
4. Blair

```
Description of Blair  is: -
count    1517.000000
mean        3.335531
std         1.174772
min         1.000000
25%         2.000000
50%         4.000000
75%         4.000000
max         5.000000
Name: Blair, dtype: float64

------------------------------
Mean is:  3.3355306526038233
Median is:  4.0
Mode is:  4
```

Distribution of Blair is: -





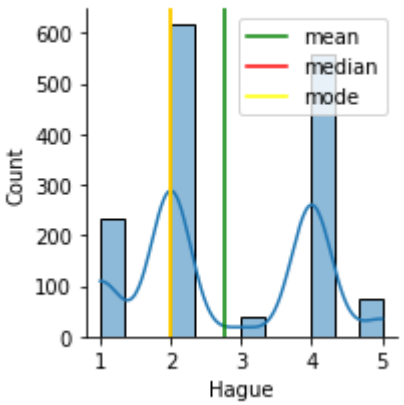
The variable represents rating of the candidate Blair as perceived by the voters. It is a discrete variable. As can be seen the max voters have rated him 4 on a scale of 5. Also, there are no outliers in the variable.

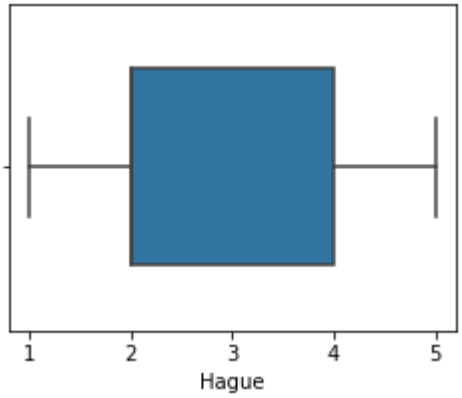
5. Hague

```
Description of Hague  is: -
count    1517.000000
mean        2.749506
std         1.232479
min         1.000000
25%         2.000000
50%         2.000000
75%         4.000000
max         5.000000
Name: Hague, dtype: float64

---------------------------
Mean is:  2.749505603164139
Median is:  2.0
Mode is:  2
```

Distribution of Hague is: -





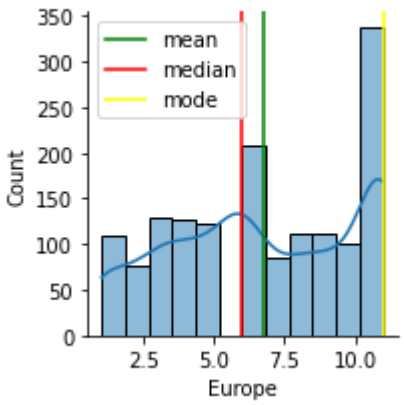
The variable represents rating of the candidate Hague as perceived by the voters. It is a discrete variable. As can be seen the max voters have rated him 2 on a scale of 5. Also, there are no outliers in the variable

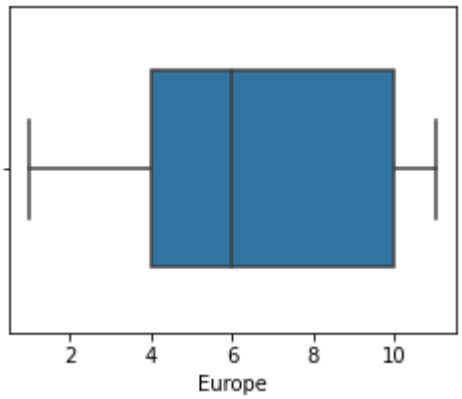
6. Europe

```
Description of Europe  is: -
count    1517.000000
mean        6.740277
std         3.299043
min         1.000000
25%         4.000000
50%         6.000000
75%        10.000000
max        11.000000
Name: Europe, dtype: float64

------------------------------
Mean is:  6.7402768622280815
Median is:  6.0
Mode is:  11
```

Distribution of Europe is: -



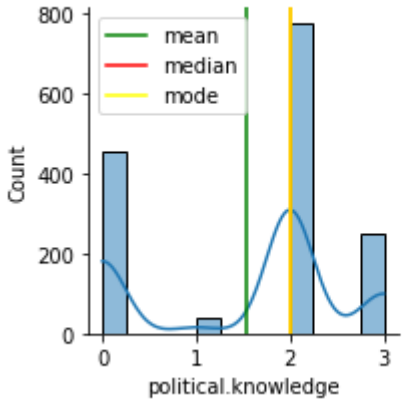


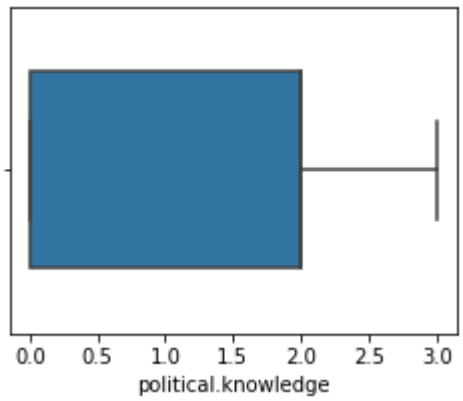
The variable represents, ratings towards European integration, on a scale 0f 11. The max rating as, can be seen is given as 11. There are no outliers in the data.

7. Political knowledge

```
Description of political.knowledge  is: -
count    1517.000000
mean        1.540541
std         1.084417
min         0.000000
25%         0.000000
50%         2.000000
75%         2.000000
max         3.000000
Name: political.knowledge, dtype: float64

-----------------------------------------
Mean is:  1.5405405405405406
Median is:  2.0
Mode is:  2
Distribution of  political.knowledge  is: -
```

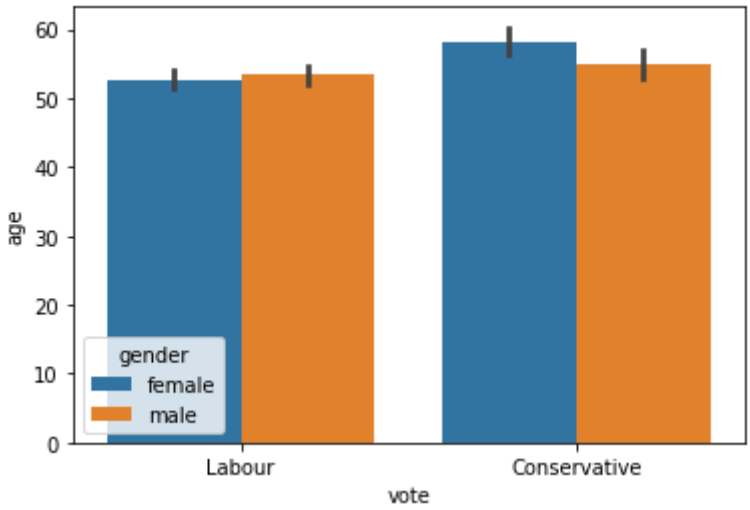




Knowledge of parties' positions on European integration, 0 to 3. It is a discrete variable, with max rating being 2. There are no outliers here.

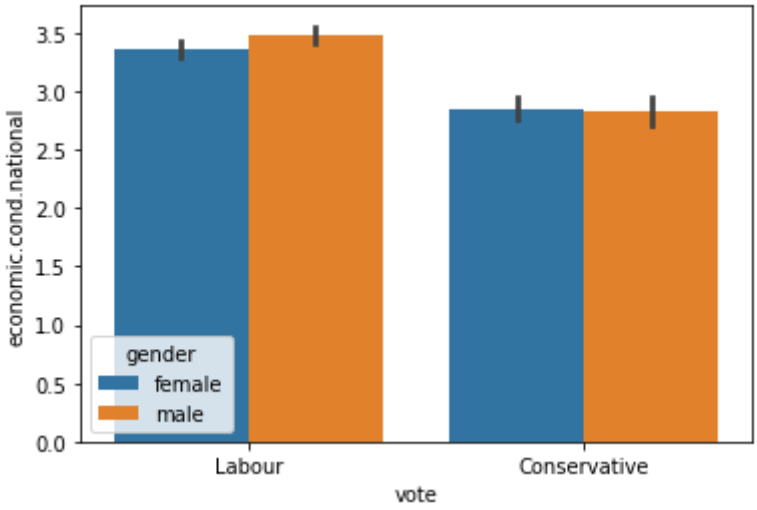
Bivariate Analysis

1. Age vs vote



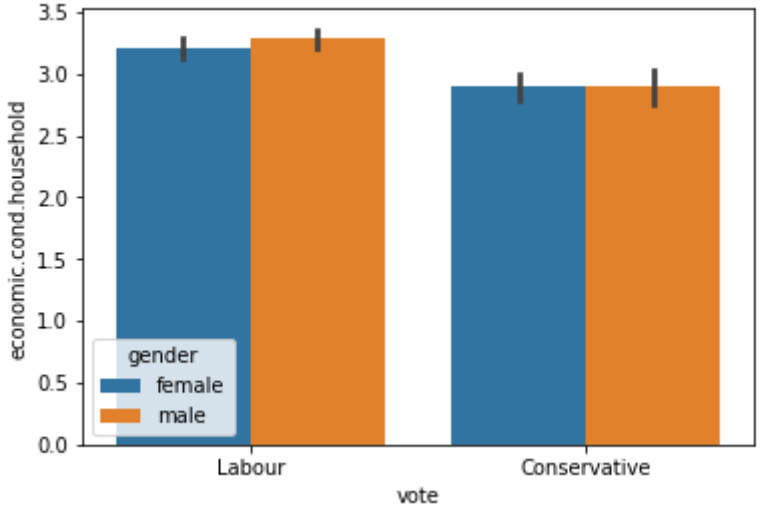
As can be seen clearly, voters with higher average age are more likely to vote in favor of the conservative party.

2. National economic conditions vs vote



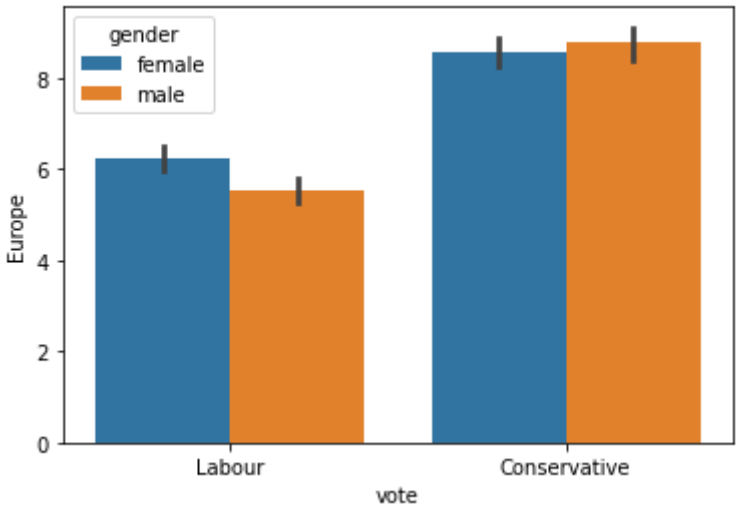
Those who rate current national economic conditions highly are more likely to vote the Labor party.

3. Household economic condition vs vote



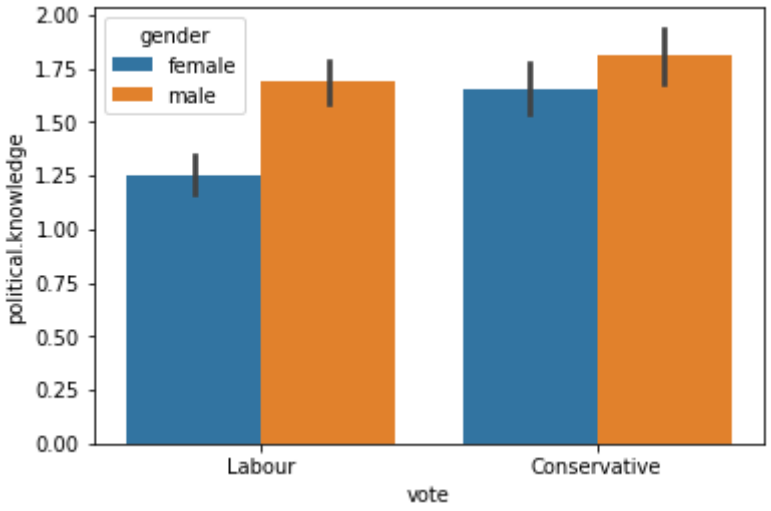
Voters rating the household conditions highly are more likely to vote for the Labor Party.

4. Europe vs vote



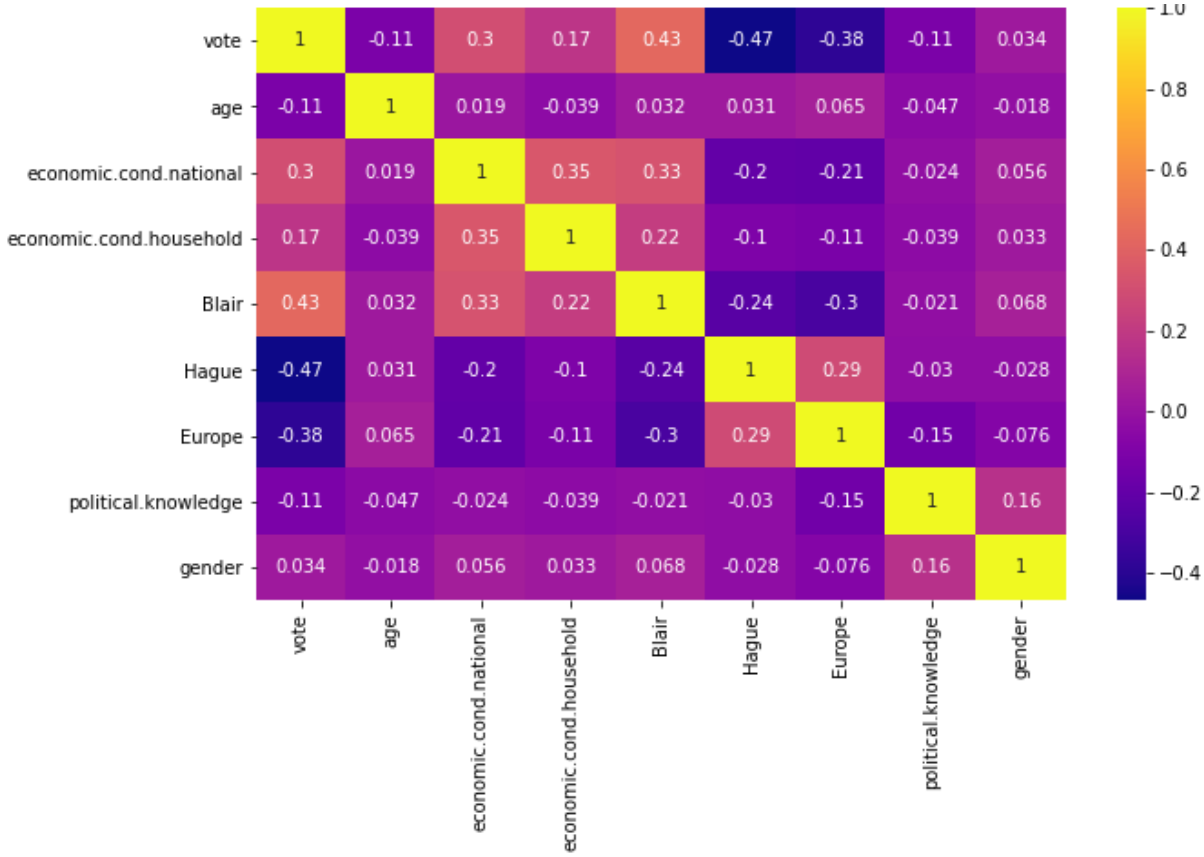
People who are more open towards the European integration, are more penchant towards voting for the Conservative Party.

5. Political Knowledge vs vote



People, who have greater political knowledge, are more likely to vote for the Conservative Party.

Correlation Check: -



There is very meagre correlation between the variables, and they appear to be independent of each other.

Outliers: The Univariate analysis has shown that there are just one or two outliers in the data, and that too in the discrete variables. Apart from that, there are none.

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

Using the pd.Categorical function, the data with string values is encoded: -

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1517 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1517 non-null   int8
 1   age                     1517 non-null   int64
 2   economic.cond.national  1517 non-null   int64
 3   economic.cond.household 1517 non-null   int64
 4   Blair                   1517 non-null   int64
 5   Hague                   1517 non-null   int64
 6   Europe                  1517 non-null   int64
 7   political.knowledge     1517 non-null   int64
 8   gender                  1517 non-null   int8
dtypes: int64(7), int8(2)
memory usage: 137.8 KB
```

Examining the class imbalance: -

```
df.vote.value_counts(normalize = True)
```

```
1    0.69677
0    0.30323
```

In this case, both the classes are equally important, but not equally distributed. Therefore, using oversampling technique like SMOTE might be useful. The model can be trained and tuned for both the Oversampled data as well, to arrive at a good and apt model.

Splitting the data into test and train: -

```
X = df.drop('vote',axis = 1)
Y = df.vote
```

```
x_train, x_test, train_labels, test_labels = train_test_split(X, Y, test_size = .3)
```

Further, SMOTE is applied: -

```
oversample = SMOTE()
x_bal, y_bal = oversample.fit_sample(X, Y)
```

```
y_bal.value_counts(normalize = True)
```

```
1    0.5
0    0.5
```

As can be seen after applying SMOTE, the classes are now balanced.

Inspecting Scaling: Scaling is essential only while using the distance based or weight-based algorithms such as ANN and KNN. It completely depends upon the model we are building. Here is there is no need for scaling.

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

Initially a basic Logistic Regression and LDA model is considered with default values, to get a basic idea about how the model performs. The model is trained with the given data without using any under sampling or over sampling technique.

Following are the classification reports obtained: -

Classification Report: Logistic Regression

Train Data: -

	precision	recall	f1-score	support
0	0.74	0.64	0.69	322
1	0.85	0.90	0.88	739
accuracy			0.82	1061
macro avg	0.80	0.77	0.78	1061
weighted avg	0.82	0.82	0.82	1061

The model performs decently with an accuracy of 82%. However, the recall score can be improved upon tuning, which will be done at a later part of this question.

Test Data: -

	precision	recall	f1-score	support
0	0.80	0.72	0.76	138
1	0.88	0.92	0.90	318
accuracy			0.86	456
macro avg	0.84	0.82	0.83	456
weighted avg	0.86	0.86	0.86	456

It is interesting to know the model is performing better over the test data. The recall scores have also increased.

Classification Report: LDA

Train Data:

	precision	recall	f1-score	support
0	0.74	0.66	0.70	322
1	0.86	0.90	0.88	739
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.82	0.83	0.82	1061

The accuracy seems fine; however, the model needs to be tuned for recall.

Test Data:

	precision	recall	f1-score	support
0	0.78	0.71	0.75	138
1	0.88	0.92	0.90	318
accuracy			0.85	456
macro avg	0.83	0.81	0.82	456
weighted avg	0.85	0.85	0.85	456

The accuracy on the model has increased, with better recall and f1 scores. Tuning the model might give even better results.

1.5) Apply KNN Model and Naïve Bayes Model. Interpret the inferences of each model

KNN Model: -

Since, KNN is a distance-based model, it is important to scale the model. Standard scaler, would bring out the variables between -1 and +1 which would be meaning less for variables such as age and ratings. Better would be to use min max scaler, as it would scale between 0 and 1.

Classification Report: Train Data

	precision	recall	f1-score	support
0	0.80	0.77	0.78	332
1	0.90	0.91	0.90	729
accuracy			0.87	1061
macro avg	0.85	0.84	0.84	1061
weighted avg	0.86	0.87	0.87	1061

The accuracy is decent, however, the recall on 0 needs to be improved.

Classification Report: Test Data

	precision	recall	f1-score	support
0	0.67	0.73	0.70	128
1	0.89	0.86	0.87	328
accuracy			0.82	456
macro avg	0.78	0.80	0.79	456
weighted avg	0.83	0.82	0.82	456

The accuracy has dropped, however, there is no case of overfitting, as the accuracy on test data is well within limits.

Naïve Bayes Model:

As it was already seen from the heatmap, there is negligible correlation between the variables, and thus Naïve Bayes is a fantastic choice for such cases. Examining the classification report would give us more insight.

Classification Report Train Data:

	precision	recall	f1-score	support
0	0.72	0.68	0.70	322
1	0.86	0.89	0.88	739
accuracy			0.82	1061
macro avg	0.79	0.78	0.79	1061
weighted avg	0.82	0.82	0.82	1061

The accuracy is fine, but recall on 0 is comparatively poor.

Classification Report Test Data:

	precision	recall	f1-score	support
0	0.78	0.74	0.76	138
1	0.89	0.91	0.90	318
accuracy			0.86	456
macro avg	0.83	0.82	0.83	456
weighted avg	0.86	0.86	0.86	456

The accuracy has improved, with better f1 scores and recall.

1.6) Model Tuning. Bagging and Boosting.

All the models are tuned with grid search CV and then trained with SMOTE as well to check if it improves the recall of 0 and 1.

1. Logistic Regression: The base model worked fine, but to ensure even smoother function a parameter grid is deployed and the best model is chosen. However, it was found that the optimized model performed even poorer than the base default model, hence it was dropped and the default model was trained with oversampled data. It gave impressive results: -

Classification Report: -

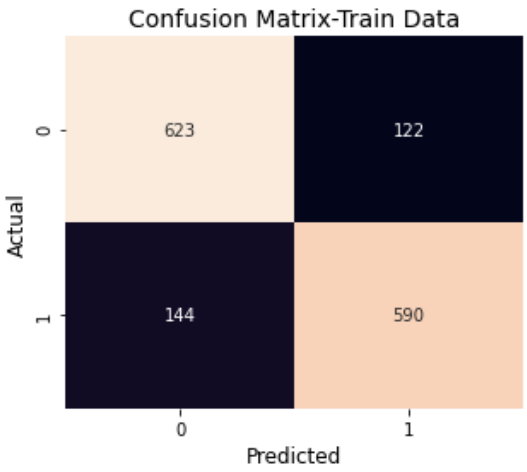
Train Data Performance: -

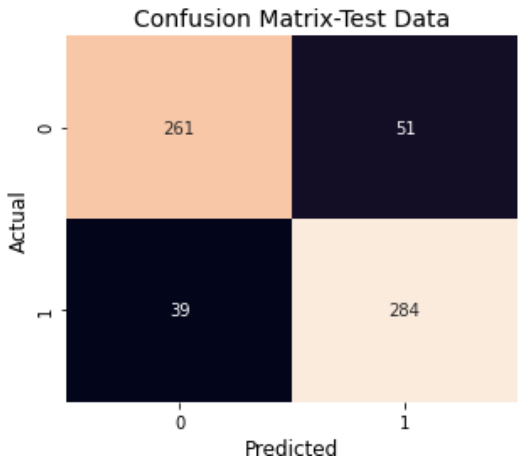
	precision	recall	f1-score	support
0	0.81	0.84	0.82	745
1	0.83	0.80	0.82	734
accuracy			0.82	1479
macro avg	0.82	0.82	0.82	1479
weighted avg	0.82	0.82	0.82	1479

Test Data Performance: -

	precision	recall	f1-score	support
0	0.87	0.84	0.85	312
1	0.85	0.88	0.86	323
accuracy			0.86	635
macro avg	0.86	0.86	0.86	635
weighted avg	0.86	0.86	0.86	635

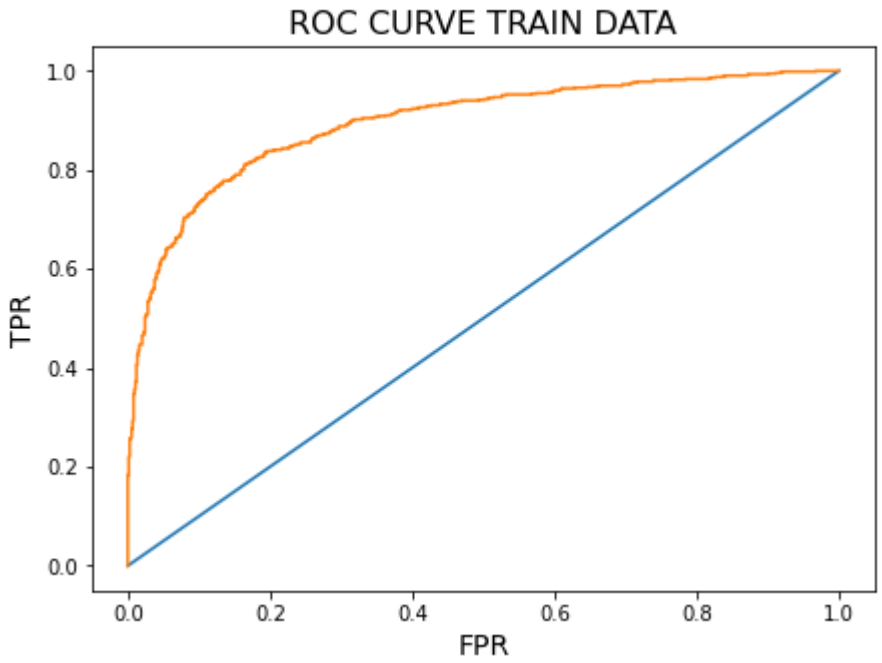
Confusion Matrix



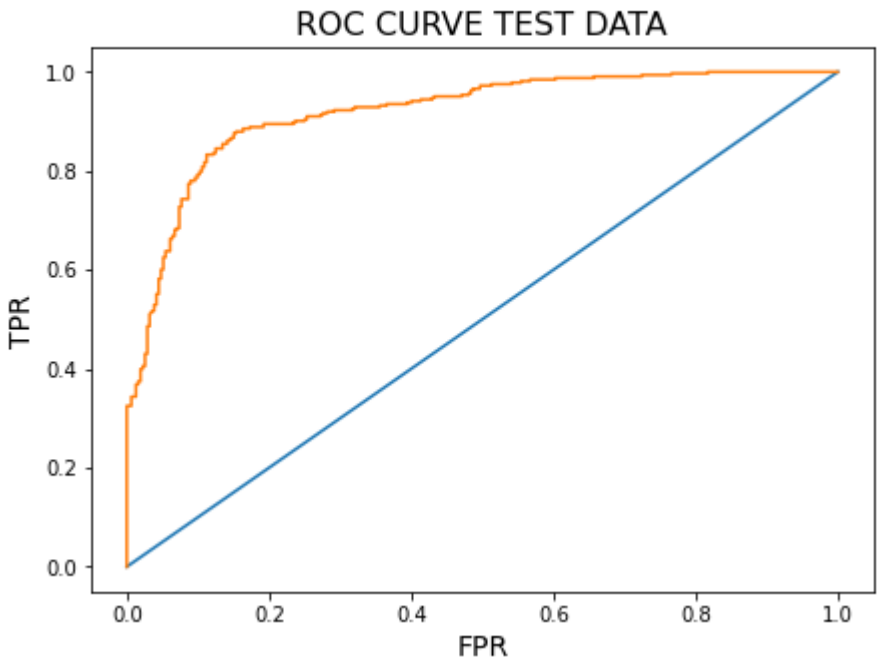


ROC Curve and AUC Score

```
ROC AUC SCORE 0.894895122798676
```




```
ROC AUC SCORE 0.9184577677224737
```



2. LDA: The LDA default model performed decently on the train data, worked fine on the test data as well. The only problem being that of class imbalance problem, which was solved using SMOTE.

Classification Report: -

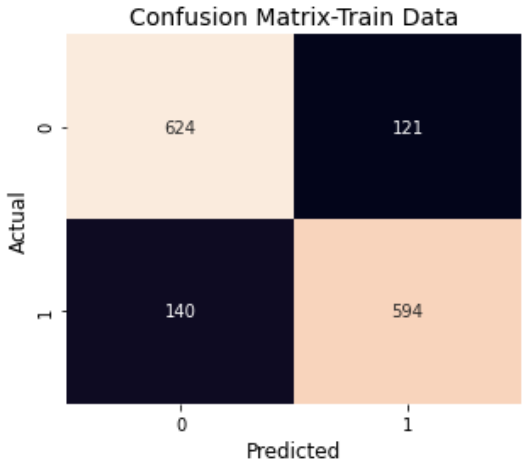
Training Data: -

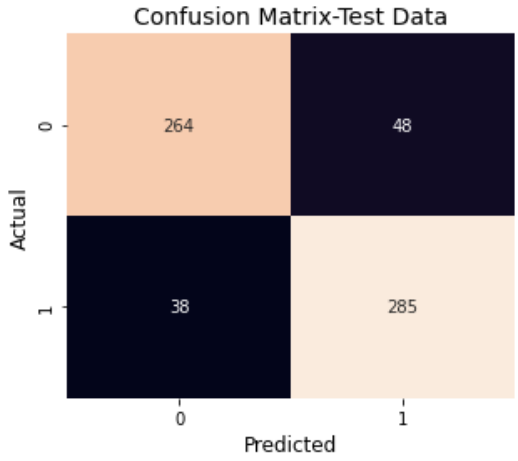
	precision	recall	f1-score	support
0	0.82	0.84	0.83	745
1	0.83	0.81	0.82	734
accuracy			0.82	1479
macro avg	0.82	0.82	0.82	1479
weighted avg	0.82	0.82	0.82	1479

Testing Data: -

	precision	recall	f1-score	support
0	0.87	0.85	0.86	312
1	0.86	0.88	0.87	323
accuracy			0.86	635
macro avg	0.87	0.86	0.86	635
weighted avg	0.86	0.86	0.86	635

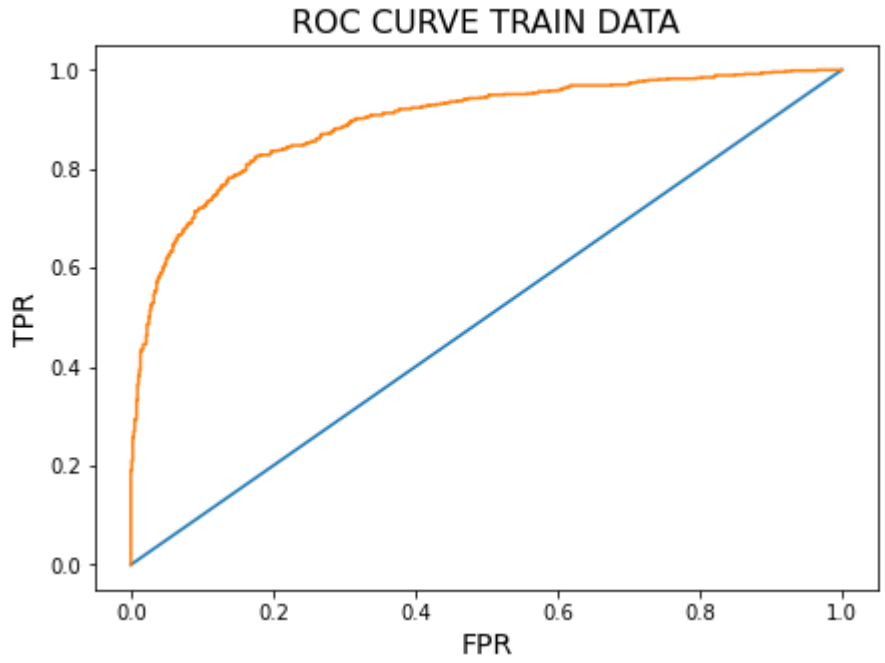
Confusion Matrix: -



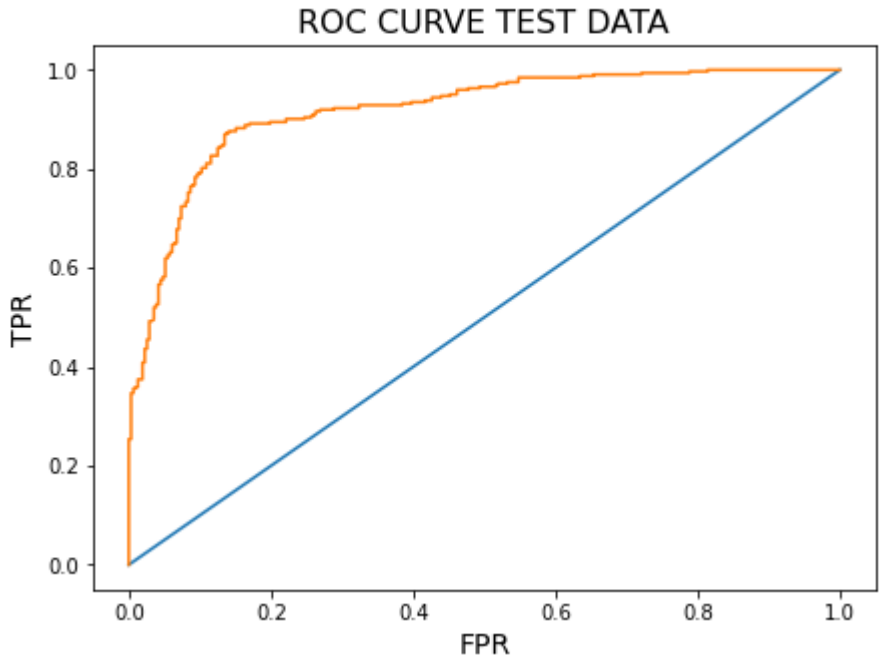


ROC Curve and AUC Score

```
ROC AUC SCORE 0.8946683612822999
```




```
ROC AUC SCORE 0.9180509248233706
```



3. KNN: Choosing K value is the point of optimization here. As a rule of thumb, the best K-values is the square root of the no of rows. Apart from the default value of 5, and keeping in mind the class imbalance problem, K was kept at 33 and 39 in the grid search. The model yielded best results for K = 39, and when the model was trained with Over sampled data.

Classification Report: -

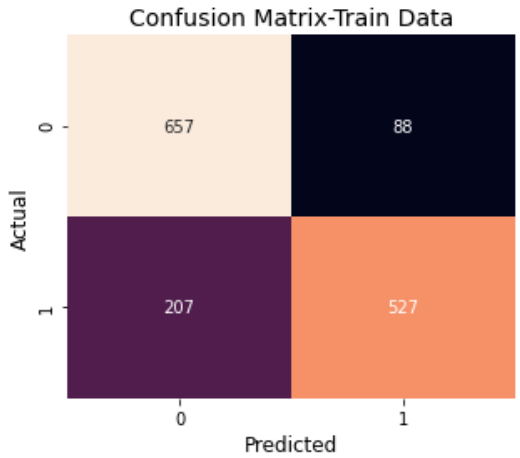
Training Data: -

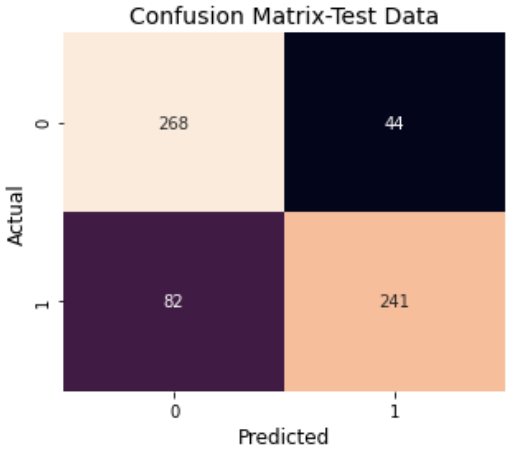
	precision	recall	f1-score	support
0	0.76	0.88	0.82	745
1	0.86	0.72	0.78	734
accuracy			0.80	1479
macro avg	0.81	0.80	0.80	1479
weighted avg	0.81	0.80	0.80	1479

Test Data: -

	precision	recall	f1-score	support
0	0.77	0.86	0.81	312
1	0.85	0.75	0.79	323
accuracy			0.80	635
macro avg	0.81	0.80	0.80	635
weighted avg	0.81	0.80	0.80	635

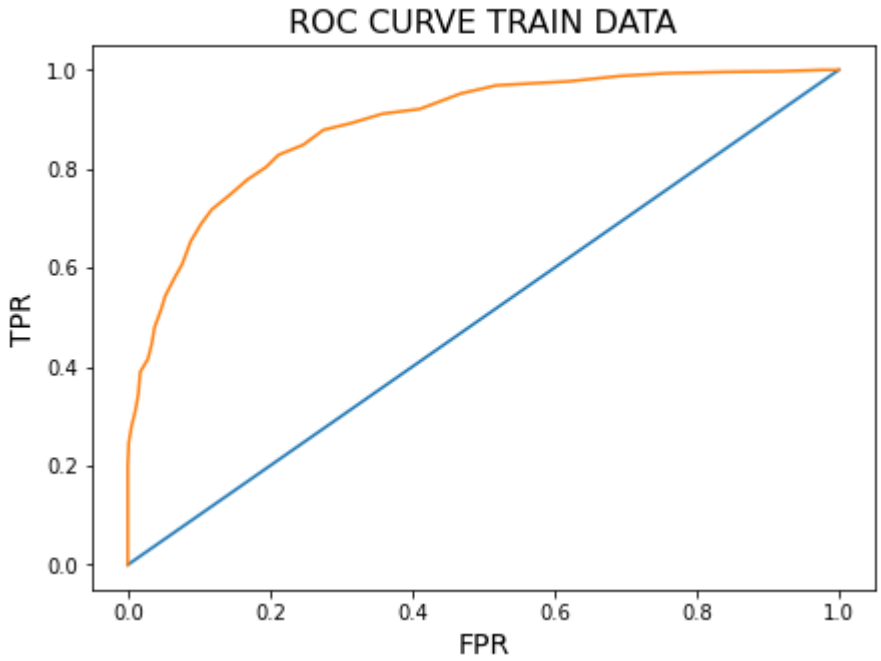
Confusion Matrix: -



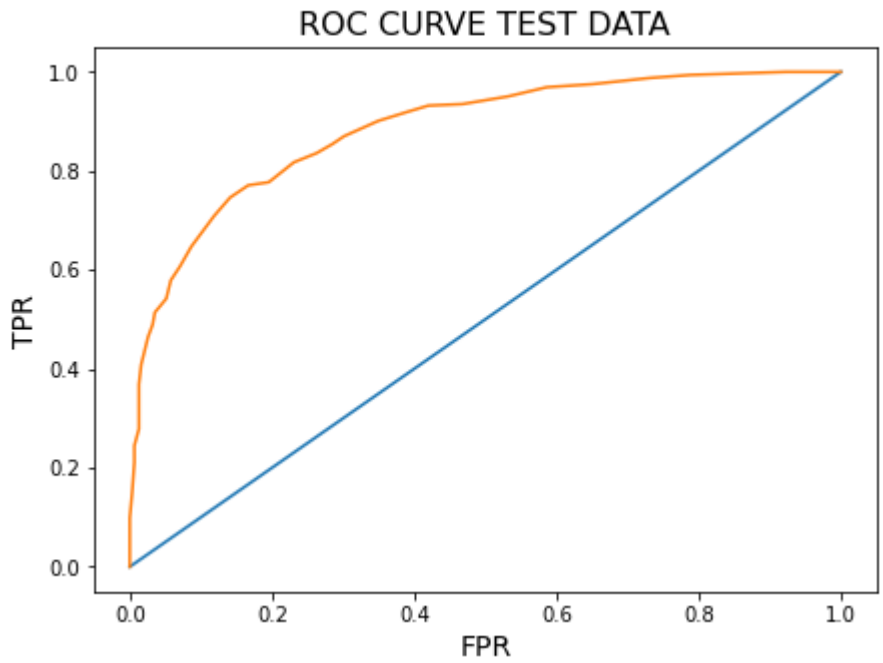


ROC Curve and AUC Score: -

```
ROC AUC SCORE 0.8898652231955085
```




```
ROC AUC SCORE 0.8837024688417877
```



4. Naïve Bayes: The Naïve Bayes base model performed better, only lacking the good recall values. It gave even better results when SMOTE was used.

Classification Report: -

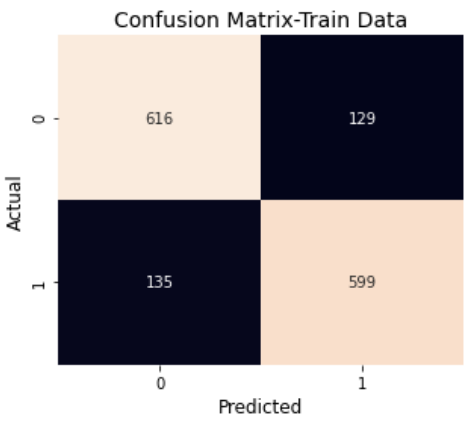
Training Data: -

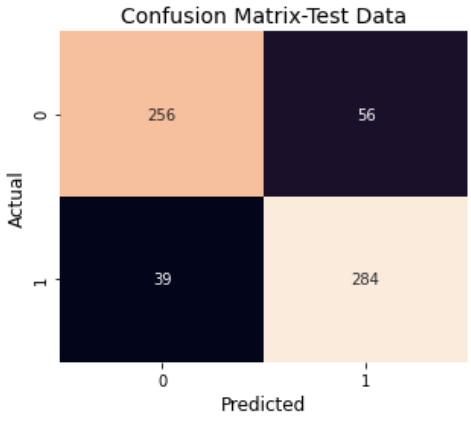
	precision	recall	f1-score	support
0	0.82	0.83	0.82	745
1	0.82	0.82	0.82	734
accuracy			0.82	1479
macro avg	0.82	0.82	0.82	1479
weighted avg	0.82	0.82	0.82	1479

Testing Data: -

	precision	recall	f1-score	support
0	0.87	0.82	0.84	312
1	0.84	0.88	0.86	323
accuracy			0.85	635
macro avg	0.85	0.85	0.85	635
weighted avg	0.85	0.85	0.85	635

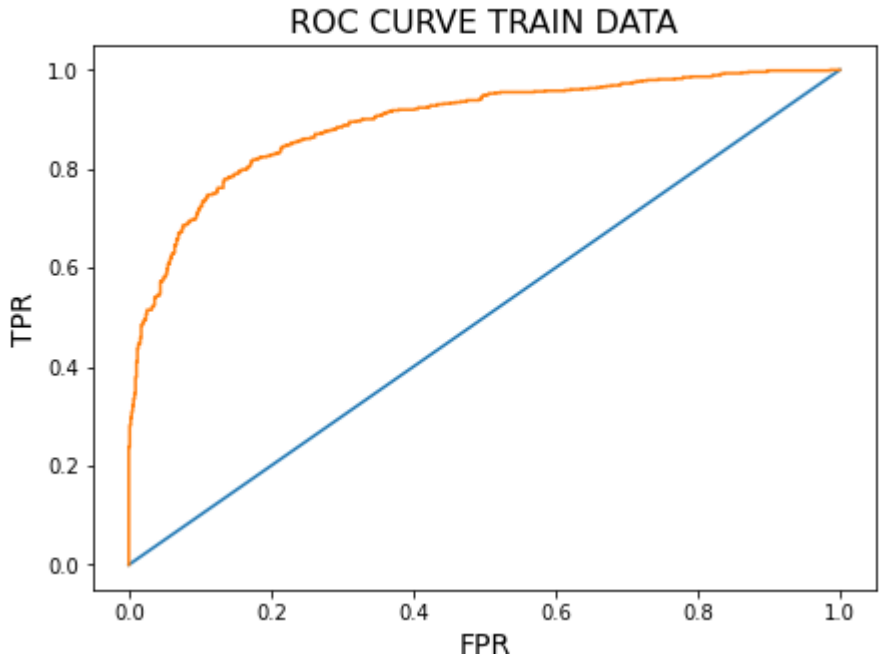
Confusion Matrix



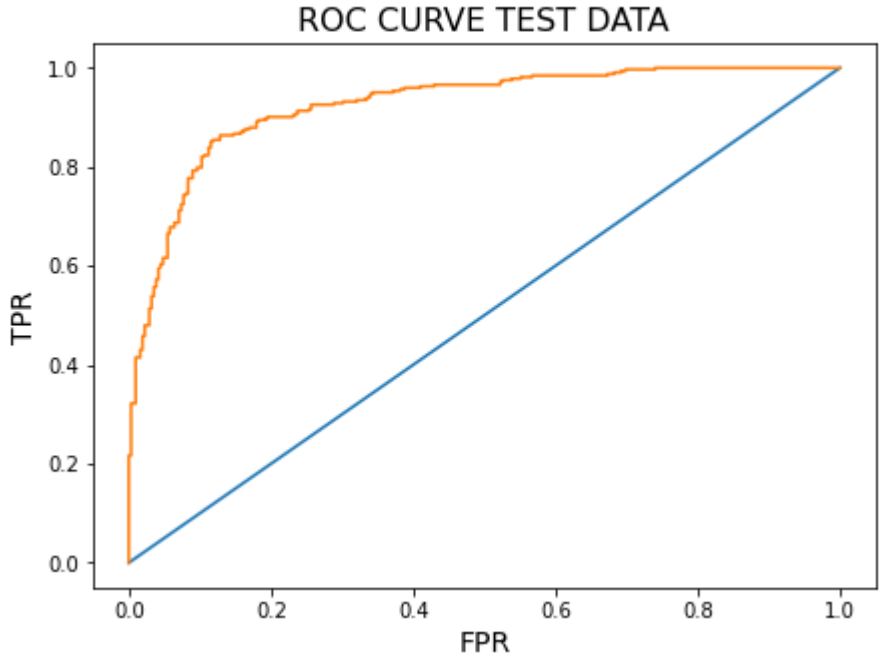


ROC Curve and AUC Score

```
ROC AUC SCORE 0.8940795128284841
```




```
ROC AUC SCORE 0.92486802413273
```



5.Ada Boost: A base of model was RFCL was fed into the grid search and then the best model was treated as the base of Ada Boost. The critical parameters were the max depth, no of estimators, min sample leaf and min sample split. Using the general rules of thumb, certain values were fed to the grid. For instance, min sample leaf is usually considered from 1% to 3% of the data. Similarly, the min sample split is taken to be 3 times that of min sample leaf. As the target class was imbalanced, Ada Boost gave best results when trained with SMOTE.

Classification Report: -

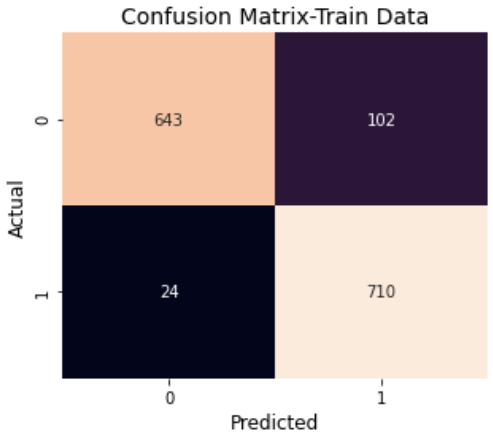
Training Data:

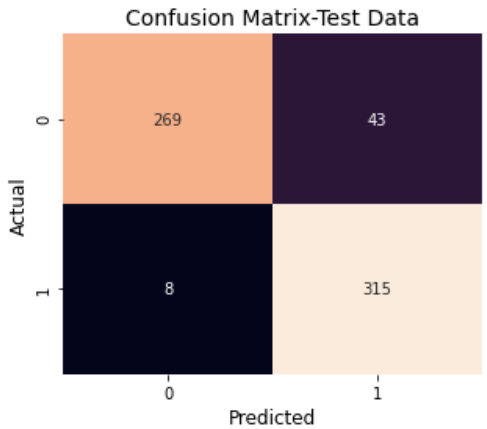
	precision	recall	f1-score	support
0	0.96	0.86	0.91	745
1	0.87	0.97	0.92	734
accuracy			0.91	1479
macro avg	0.92	0.92	0.91	1479
weighted avg	0.92	0.91	0.91	1479

Testing Data:

	precision	recall	f1-score	support
0	0.97	0.86	0.91	312
1	0.88	0.98	0.93	323
accuracy			0.92	635
macro avg	0.93	0.92	0.92	635
weighted avg	0.92	0.92	0.92	635

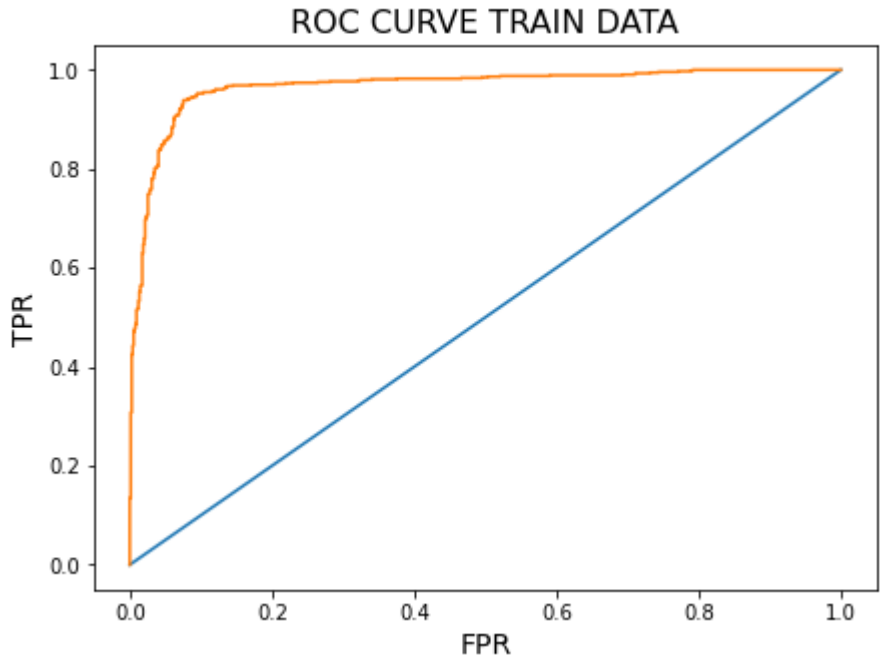
Confusion Matrix



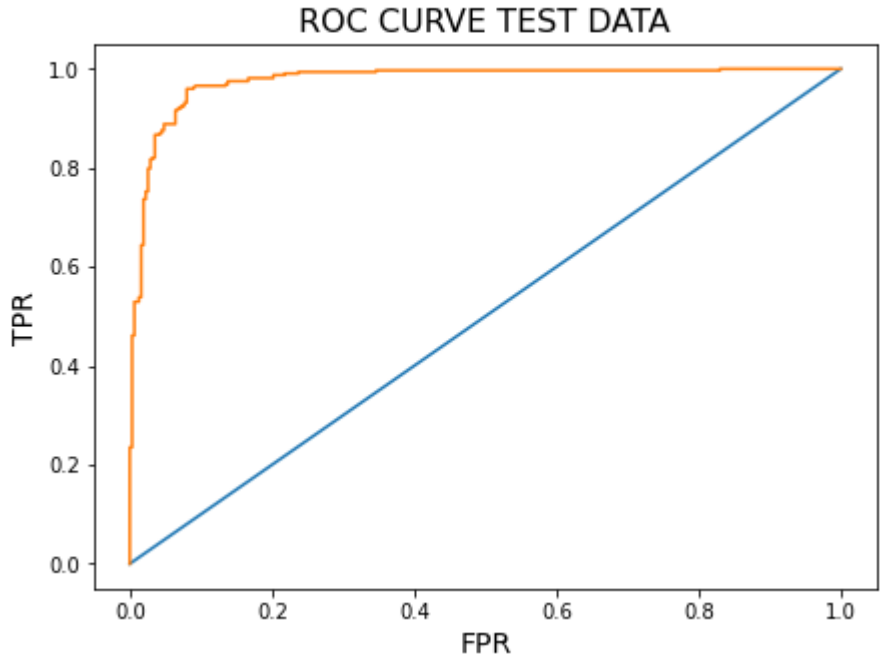


ROC Curve and AUC Score

```
ROC AUC SCORE 0.9665407896421192
```




```
ROC AUC SCORE 0.976120306422164
```



6.Gradient Boost: Gradient Boost performed decently over the base model, however, due to system constraints the model was too stubborn to optimize. Therefore, the model was trained on oversampled data.

Classification Report: -

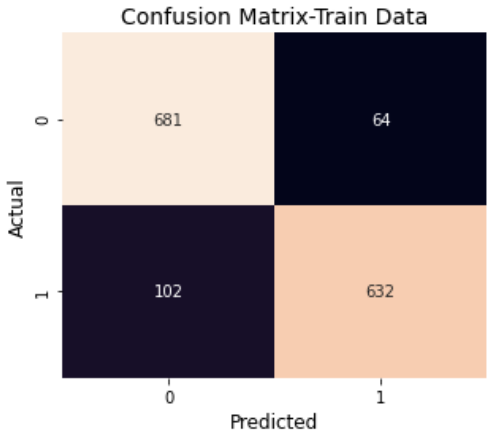
Training data: -

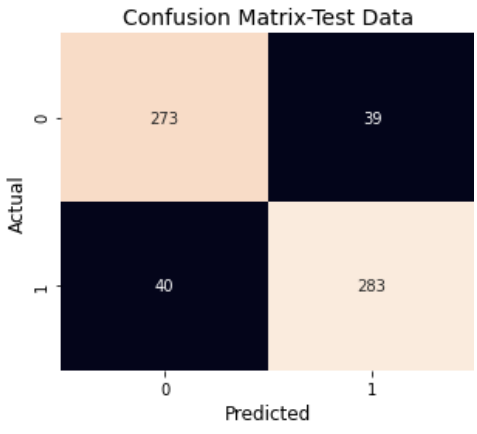
	precision	recall	f1-score	support
0	0.87	0.91	0.89	745
1	0.91	0.86	0.88	734
accuracy			0.89	1479
macro avg	0.89	0.89	0.89	1479
weighted avg	0.89	0.89	0.89	1479

Testing Data: -

	precision	recall	f1-score	support
0	0.87	0.88	0.87	312
1	0.88	0.88	0.88	323
accuracy			0.88	635
macro avg	0.88	0.88	0.88	635
weighted avg	0.88	0.88	0.88	635

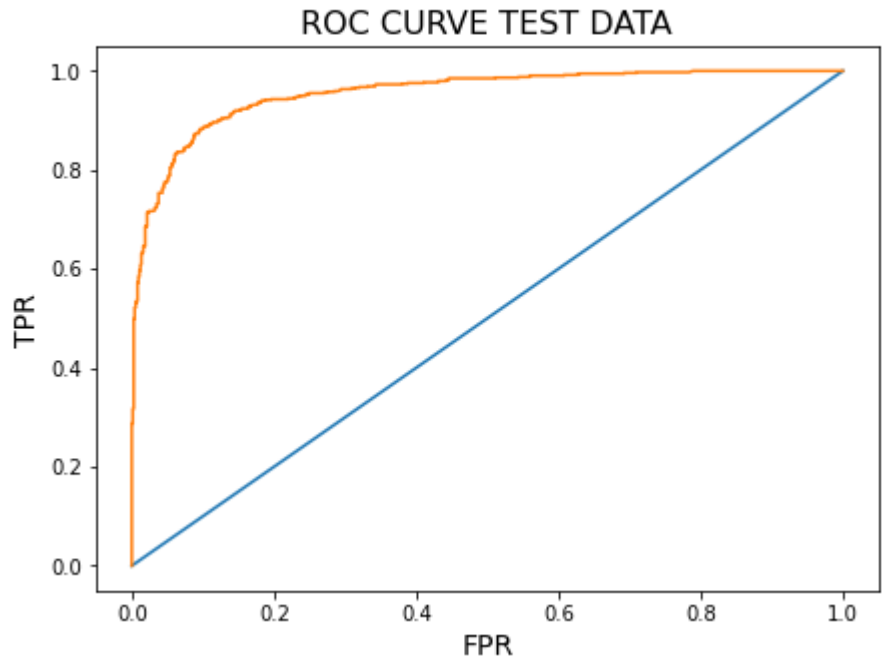
Confusion Matrix



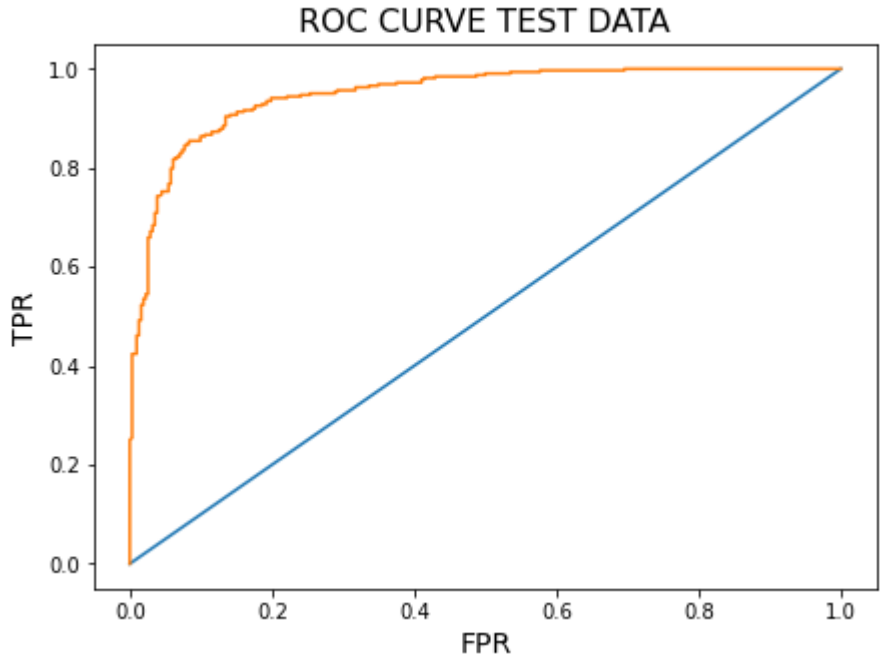


ROC Curve and AUC Score

```
ROC AUC SCORE 0.9553252381910282
```




```
ROC AUC SCORE 0.9493133285702946
```



7.Bagging Classifier: The Bagging classifier was optimized over its base model. The optimized RFCL model was used for its base version. It gave better results however, due to low recall scores, SMOTE was used to improve the Recall scores.

Classification Report:

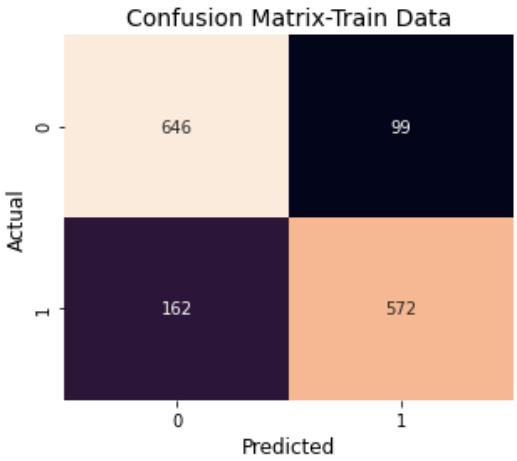
Training Data:

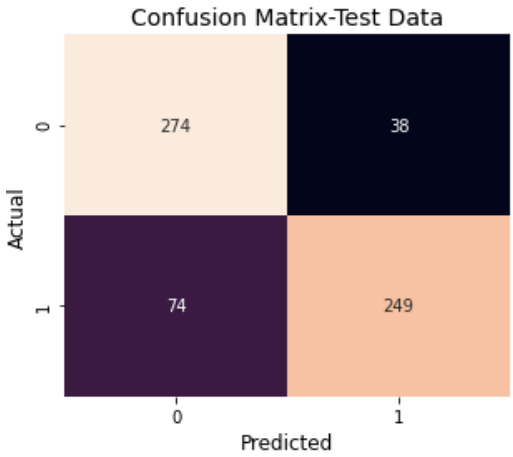
	precision	recall	f1-score	support
0	0.80	0.87	0.83	745
1	0.85	0.78	0.81	734
accuracy			0.82	1479
macro avg	0.83	0.82	0.82	1479
weighted avg	0.83	0.82	0.82	1479

Testing Data:

	precision	recall	f1-score	support
0	0.79	0.88	0.83	312
1	0.87	0.77	0.82	323
accuracy			0.82	635
macro avg	0.83	0.82	0.82	635
weighted avg	0.83	0.82	0.82	635

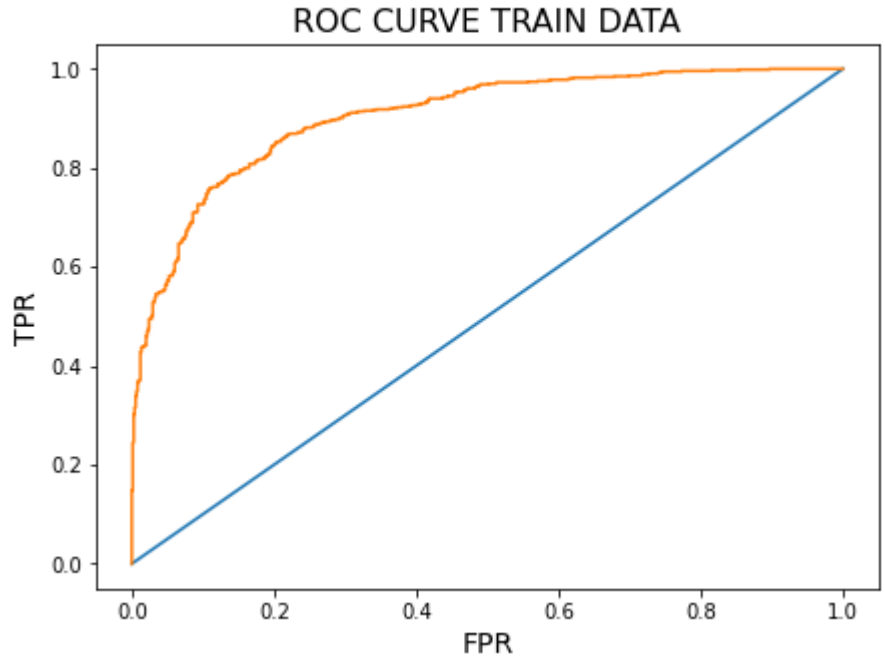
Confusion Matrix



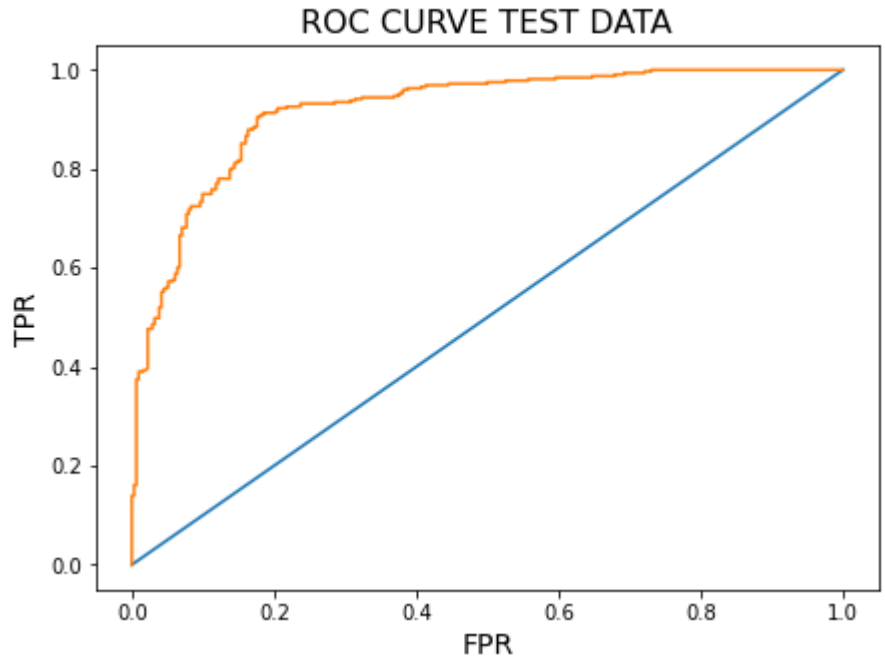


ROC Curve and AUC Score

```
ROC AUC SCORE 0.9042554358758663
```




```
ROC AUC SCORE 0.9186363816781774
```



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

All the best models, are checked in the attached Jupyter notebook file, and the best results are tabulated here for training and testing data respectively.

Model comparison for training Data: -

TRAINING DATA								
MODEL	**Accuracy (%)**	**Precision**		**Recall**		**F-1 Score**		**AUC**
		0	**1**	**0**	**1**	**0**	**1**	
Logistic Regression	82	81	83	84	80	82	82	89.4
LDA	82	82	83	84	81	83	82	89.4
KNN	80	76	86	88	72	82	78	88.9
Naïve Bayes	82	82	82	83	82	82	82	89.4
Ada Boost	92	97	88	86	98	91	93	96.65
Gradient Boost	89	87	91	91	86	89	88	95.53
Bagging	82	80	85	87	78	83	81	90.04

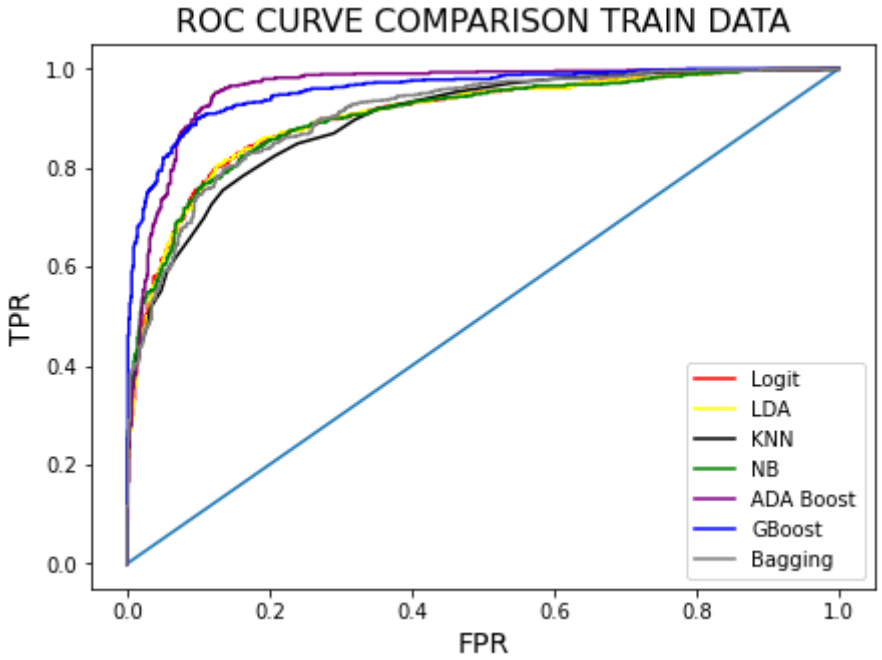
Model comparison for testing data: -

TESTING DATA								
MODEL	**Accuracy (%)**	**Precision**		**Recall**		**F-1 Score**		**AUC**
		0	**1**	**0**	**1**	**0**	**1**	
Logistic Regression	86	87	85	84	88	85	86	91.8
LDA	86	87	86	85	88	86	87	91.8
KNN	80	77	85	86	75	81	79	88.3
Naïve Bayes	85	87	84	82	88	84	86	92.4
Ada Boost	92	97	88	86	98	91	93	97.6
Gradient Boost	88	87	88	88	88	87	88	94.93
Bagging	82	79	87	88	77	83	82	91.86

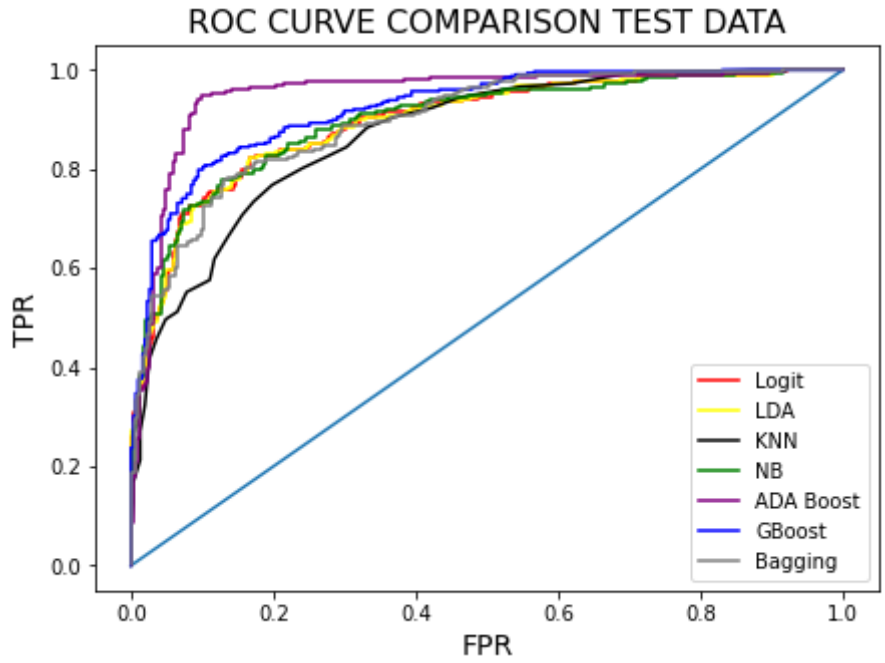
On comparing the training Data and the testing data, it can be seen clearly that, Ada Boost Model performed above all in every department. The main concern here was to predict the votes accurately with focus on both the classes in a balanced manner. Ada boost proves to be a perfect model here, with best parameters in every sense.

ROC Curve Comparison: -

Training Curves



Testing Data



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.

The business problem aims at creating an exit poll on the given data. The data is imbalanced, therefore, SMOTE is an important factor to get the recall scores right.

The EDA led us with the following trend: -

1. Voters with higher average are more likely to vote conservative Party.
2. People who think the current economic conditions for both nation and household are better, are more likely to vote for Labor party.
3. Its an obvious fact, who have rated Blair highly, are more likely to vote for Labor Party and others who have rated Hague highly, would more likely vote Hague.
4. People who are open towards European integration are considered more politically wise, and they are more likely to vote for the conservative party.

In general, people who value the economic development more are more likely to vote for the labor party, while those of the liberal mind set would vote for the conservative party.

On building the models, it was observed, that the ADA Boost Classifier gave fantastic results for both 0 and 1(Labor and Conservative). This model can definitely be deployed in production to achieve a precise exit poll.

PROBLEM 2

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

Counting no of characters: -

```
print('No of characters in Roosevelts speech:', len(rsvlt))
print('No of characters is Kennedys speech:', len(kndy))
print('No of characters in Nixons speech:', len(nixon))
```

```
No of characters in Roosevelts speech: 7571
No of characters is Kennedys speech: 7618
No of characters in Nixons speech: 9991
```

The len () command simply tells us the no of characters in each speech.

Counting no of words: -

Two approaches are shown here to count the no of words: -

```
print('No of words in Roosevelts speech is:',
      len(inaugural.words(fileids='1941-Roosevelt.txt')))
print('No of words in Kenedys speech is:',
      len(inaugural.words(fileids='1961-Kennedy.txt')))
print('No of words in Nixon speech is:',
      len(inaugural.words(fileids='1973-Nixon.txt')))
```

```
No of words in Roosevelts speech is: 1536
No of words in Kenedys speech is: 1546
No of words in Nixon speech is: 2028
```

In this approach, the words () function is used. The second approach is shown below: -

```
print('No of words in Roosevelt speech:',
      len(inaugural.raw(fileids = '1941-Roosevelt.txt').split()))
print('No of words in Kenedys speech:',
      len(inaugural.raw(fileids = '1961-Kennedy.txt').split()))
print('No of words in Nixons speech:',
      len(inaugural.raw(fileids = '1973-Nixon.txt').split()))
```

```
No of words in Roosevelt speech: 1360
No of words in Kenedys speech: 1390
No of words in Nixons speech: 1819
```

Counting no of sentences: -

```
print('No of sentences in Roosevelt speech:',
      len(inaugural.sents(fileids='1941-Roosevelt.txt')))
print('No of sentences in Kennedys speech:',
      len(inaugural.sents(fileids = '1961-Kennedy.txt')))
print('No of sentences in Nixons speech:',
      len(inaugural.sents(fileids = '1973-Nixon.txt')))
```

```
No of sentences in Roosevelt speech: 68
No of sentences in Kennedys speech: 52
No of sentences in Nixons speech: 69
```

The sents () command yields the no of sentences.

2.2) Remove all the stop words from the three speeches

Speech 1: Roosevelt

Before removing stop words: -

```
'On each national day of inauguration since 1789, the people have renewed their sense of dedication to the United States.\n\nIn
Washington\'s day the task of the people was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the people
was to preserve that Nation from disruption from within.\n\nIn this day the task of the people is to save that Nation and its i
nstitutions from disruption from without.\n\nTo us there has come a time, in the midst of swift happenings, to pause for a mome
nt and take stock -- to recall what our place in history has been, and to rediscover what we are and what we may be. If we do n
ot, we risk the real peril of inaction.\n\nLives of nations are determined not by the count of years, but by the lifetime of th
e human spirit. The life of a man is three-score years and ten: a little more, a little less. The life of a nation is the fulln
ess of the measure of its will to live.\n\nThere are men who doubt this. There are men who believe that democracy, as a form of
Government and a frame of life, is limited or measured by a kind of mystical and artificial fate that, for some unexplained rea
son, tyranny and slavery have become the surging wave of the future -- and that freedom is an ebbing tide.\n\nBut we Americans
know that this is not true.\n\nEight years ago, when the life of this Republic seemed frozen by a fatalistic terror, we proved
that this is not true. We were in the midst of shock -- but we acted. We acted quickly, boldly, decisively.\n\nThese later year
```

Th stop words, and the punctuations those come as standard with the natural language processing toolkit, are removed along with '- -'. The extract after removing looks like this: -

```
'national day inauguration since 1789 people renewed sense dedication united states washington day task people create weld toge
ther nation lincoln day task people preserve nation disruption within day task people save nation institutions disruption witho
ut us come time midst swift happenings pause moment take stock recall place history rediscover may risk real peril inaction liv
es nations determined count years lifetime human spirit life man three score years ten little little less life nation fullness
measure live men doubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplai
ned reason tyranny slavery become surging wave future freedom ebbing tide americans know true eight years ago life republic see
med frozen fatalistic terror proved true midst shock acted acted quickly boldly decisively later years living years fruitful ye
ars people democracy brought us greater security hope better understanding life ideals measured material things vital present f
uture experience democracy successfully survived crisis home put away many evil things built new structures enduring lines main
tained fact democracy action taken within three way framework constitution united states coordinate branches government continu
e freely function bill rights remains inviolate freedom elections wholly maintained prophets downfall american democracy seen d
ire predictions come naught democracy dying know seen revive grow know cannot die built unhampered initiative individual men wo
men joined together common enterprise enterprise undertaken carried free expression free majority know democracy alone forms go
vernment enlists full force men enlightened know democracy alone constructed unlimited civilization capable infinite progress i
```

Speech 2: Kennedy

Before removing stop words: -

```
Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend
:lergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well a
; a beginning -- signifying renewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our
forebears l prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal
ıands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for w
ıich our forebears fought are still at issue around the globe -- the belief that the rights of man come not from the generosity
of the state, but from the hand of God.\n\nWe dare not forget today that we are the heirs of that first revolution. Let the wor
l go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans -
```

After removing stop words: -

'vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change s worn almighty god solemn oath forebears l prescribed nearly century three quarters ago world different man holds mortal hands p ower abolish forms human poverty forms human life yet revolutionary beliefs forebears fought still issue around globe belief ri ghts man come generosity state hand god dare forget today heirs first revolution let word go forth time place friend foe alike torch passed new generation americans born century tempered war disciplined hard bitter peace proud ancient heritage unwilling witness permit slow undoing human rights nation always committed committed today home around world let every nation know whethe r wishes us well ill shall pay price bear burden meet hardship support friend oppose foe order assure survival success liberty much pledge old allies whose cultural spiritual origins share pledge loyalty faithful friends united little cannot host coopera tive ventures divided little dare meet powerful challenge odds split asunder new states welcome ranks free pledge word one form

Speech 3: Nixon

Before removing stop words: -

'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and go od country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of see mingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are a bout to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnati on at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilit ies greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis p ast year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendships, and

After removing stop words: -

'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and go od country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of see mingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are a bout to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnati on at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilit ies greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis p ast year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendships, and by our missions to Peking and to Moscow, we were able to establish the base for a new and more durable pattern of relationships among the nations of the world. Because of America\'s bold initiatives, 1972 will be long remembered as the year of the greates t progress since the end of World War II toward a lasting peace in the world.\n\nThe peace we seek in the world is not the flim

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

Speech 1: Roosevelt

Following words were maximum in Frequency: -

	Word	Frequency
0	nation	12
1	know	10
2	spirit	9

Speech 2: Kennedy

	Word	Frequency
0	let	16
1	us	12
2	world	8

Speech 3: Nixon

	Word	Frequency
0	us	26
1	let	22
2	america	21

2.4) Plot the word cloud of each of the three speeches. (after removing the stop words)

Word cloud for Roosevelt's speech: -

Word cloud for Kennedy's speech: -

Word cloud for Nixon's speech: -