[MACHINE LEARNING]

There are two case studies , one is U.S Election Prediction and another is Election Speech analysis of three American President Franklin Roosevelt , John F Kennedy and Richard Nixon

TABLE OF CONTENTS

Problem-1

You are hired by one of the leading news channels CNBE who wants to analyze 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 Data set. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks)

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

Table 1 - Dataframe Head

	Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
count	1525.000000	1525	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525
unique	NaN	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2
top	NaN	Labour	NaN	NaN	NaN	NaN	NaN	NaN	NaN	female
freq	NaN	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN	812
mean	763.000000	NaN	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295	NaN
std	440.373894	NaN	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315	NaN
min	1.000000	NaN	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	NaN
25%	382.000000	NaN	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000	NaN
50%	763.000000	NaN	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000	NaN
75%	1144.000000	NaN	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000	NaN
max	1525.000000	NaN	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000	NaN

Table 2-Dataframe Description

The following assumptions can be made from the description of the Election data

The data set gives the demographic , national and economical condition view points, and view points on respective party leaders of 1525 voters.

There are two political parties , "Labour" and "Conservative" and Blair and Hague are the respective leaders of these parties.

The age of the voters ranges between 24yrs to 93yrs. 75% voters are in the age of up to 67yrs.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Unnamed: 0               1525 non-null   int64 
 1   vote                     1525 non-null   object
 2   age                      1525 non-null   int64 
 3   economic.cond.national   1525 non-null   int64 
 4   economic.cond.household  1525 non-null   int64 
 5   Blair                    1525 non-null   int64 
 6   Hague                    1525 non-null   int64 
 7   Europe                   1525 non-null   int64 
 8   political.knowledge      1525 non-null   int64 
 9   gender                   1525 non-null   object
dtypes: int64(8), object(2)
memory usage: 119.3+ KB
```

There are two categorical variables one is "vote" and other is "gender" the rest are numerical.

```
Unnamed: 0                 False
vote                       False
age                        False
economic.cond.national     False
economic.cond.household    False
Blair                      False
Hague                      False
Europe                     False
political.knowledge        False
gender                     False
dtype: bool
```

There are no missing values in the data set.

```
age                         0.144621
economic.cond.national     -0.240453
economic.cond.household    -0.149552
Blair                      -0.535419
Hague                       0.152100
Europe                     -0.135947
political.knowledge        -0.426838
dtype: float64
```

Skew ness is a statistical term and it is a way to estimate or measure the shape of a distribution. It is an important statistical methodology that is used to estimate the asymmetrical behavior rather than computing frequency distribution. Skewness can be two types:

Symmetrical: A distribution can be called symmetric if it appears the same from the left and right from the center point.

Asymmetrical: A distribution can be called asymmetric if it doesn't appear the same from the left and right from the center point.

Distribution on the basis of skewness value:

Skewness = 0: Then normally distributed.

Skewness > 0: Then more weight in the left tail of the distribution.

Skewness < 0: Then more weight in the right tail of the distribution.

After looking at the above skewness values of each variables in Election data following observations have been made:

1. All the variables are asymmetrically distributed

1. Only "Age" is positively skewed, which means that majority of the data distribution is on the left side of the mean. So most of the population age is more than the average age, so we are dealing with a comparatively older age population here.

1. All other variables are negatively skewed which means majority of the data distribution is on the right side of the mean.

1.2 Perform Uni-variate and Bi-variate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

There are 1525 observations on 9 variables.

```
vote                       False
age                        False
economic.cond.national     False
economic.cond.household    False
Blair                      False
Hague                      False
Europe                     False
political.knowledge        False
gender                     False
dtype: bool
```


```
vote                       False
age                        False
economic.cond.national     False
economic.cond.household    False
Blair                      False
Hague                      False
Europe                     False
political.knowledge        False
gender                     False
dtype: bool
```


There are no NaN or missing values in the data set.

Uni-variate Analysis

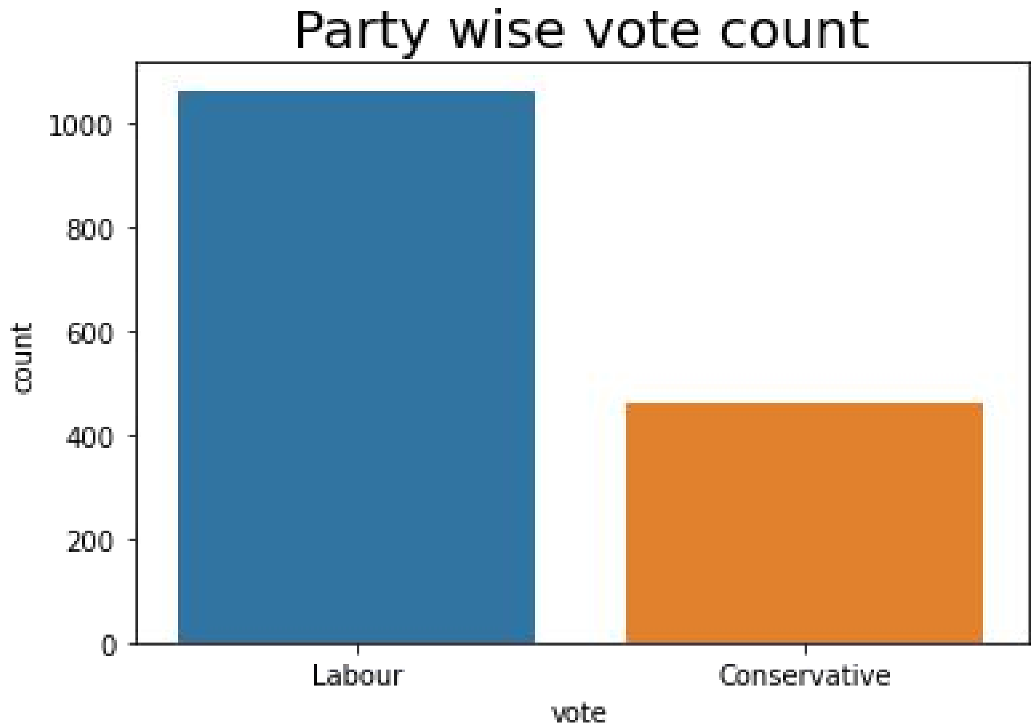


Table 3-Party wise vote count

Here as we can see there is a huge difference between the Labour party votes and conservative party votes. Our target variable is "Vote" and seeing this kind of difference between the two classes, we can say that the data set is having a class imbalance problem.

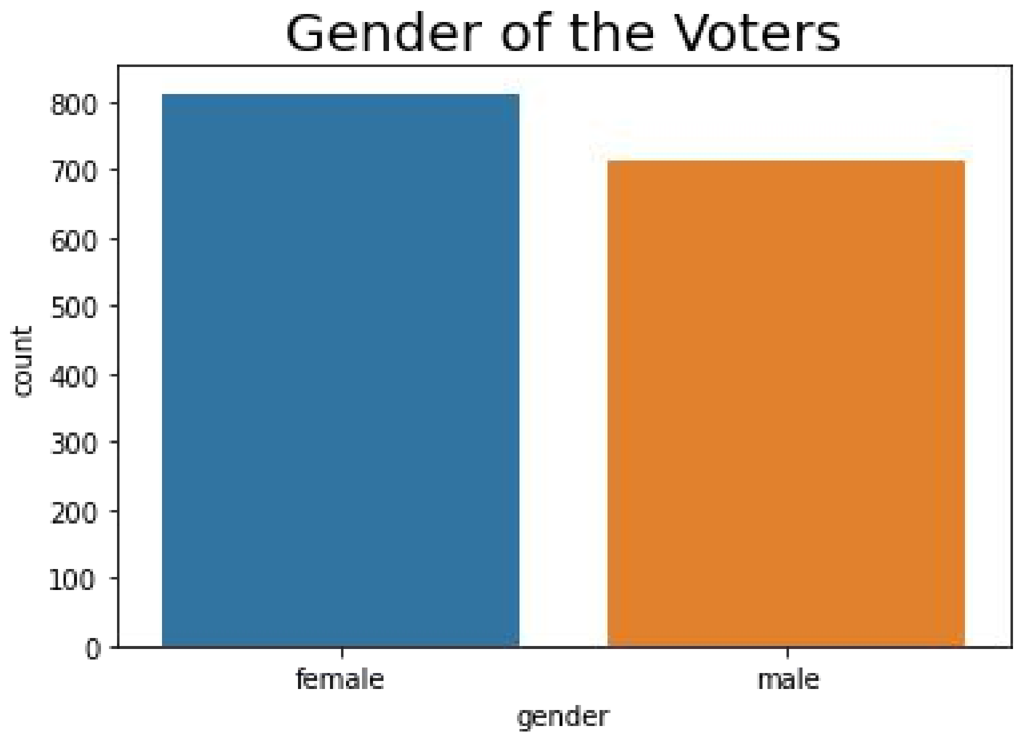


Table 4-Gender wise voters

Number of female voters is more than the male voters.

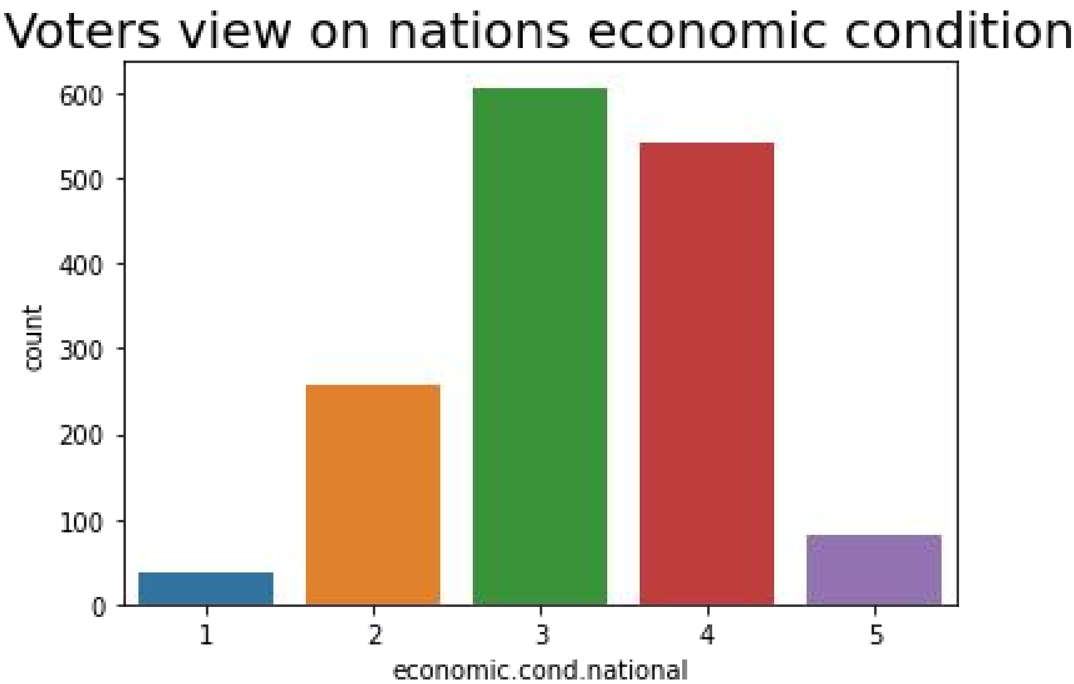


Table 5-voters view on economic condition

Most of the voters have rated 3 to nation's economic condition

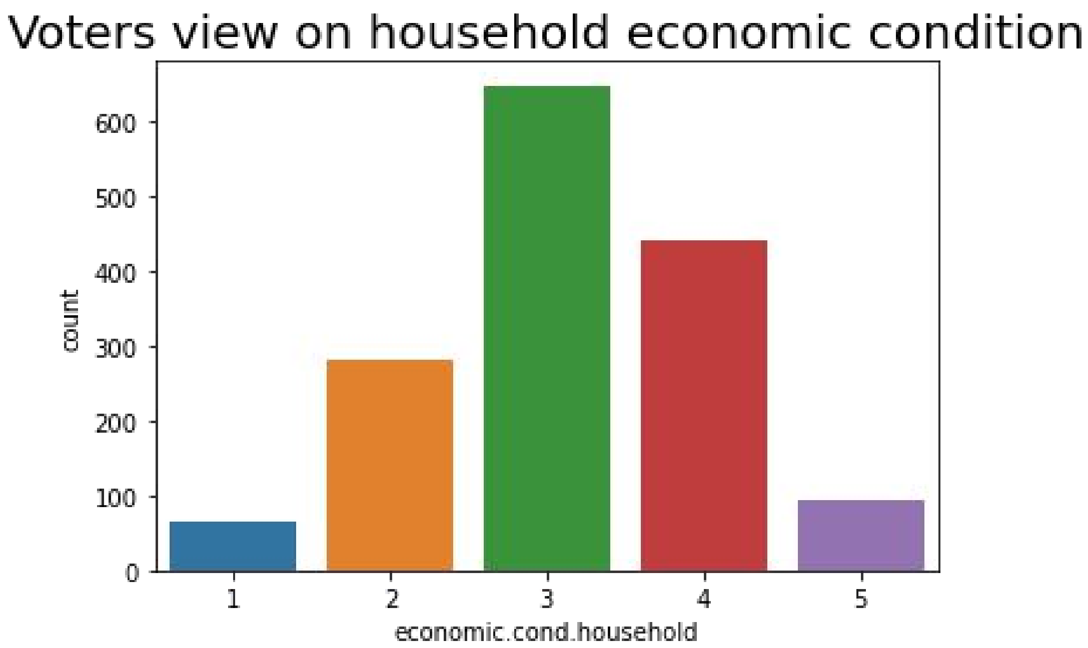


Table 6-voters view on house hold economic cond

Most of the voters have given 3 rating to household economic condition.

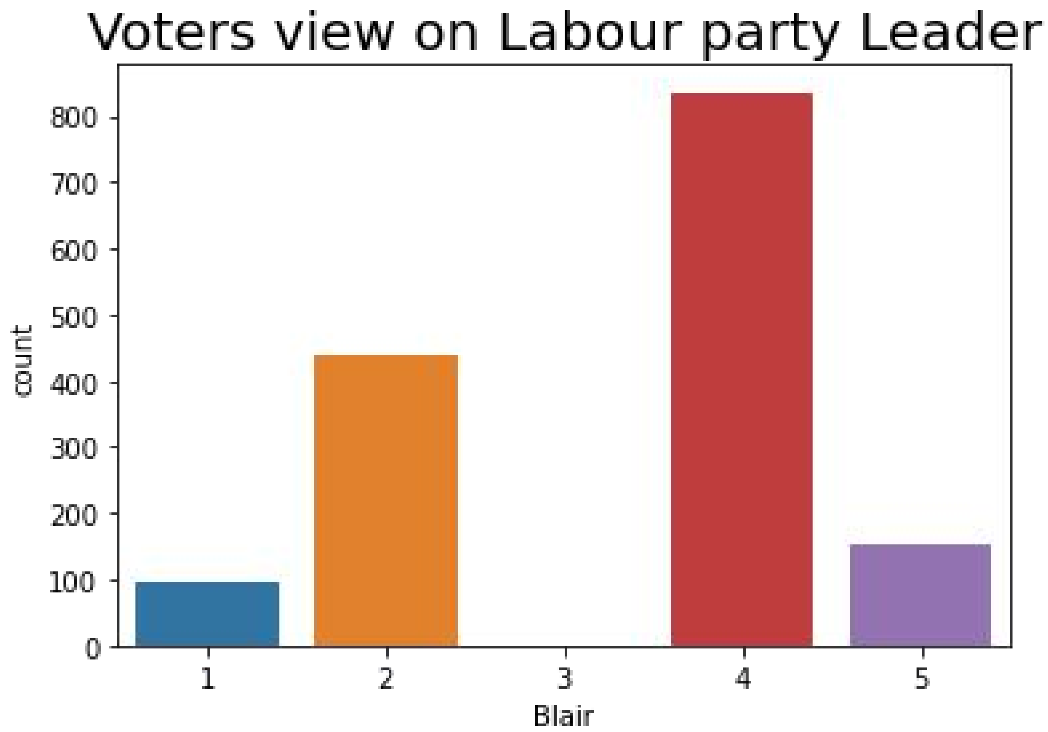


Table 7-Voters rating on Blair

Most of the people have given 4 rating to the leader of Labour party

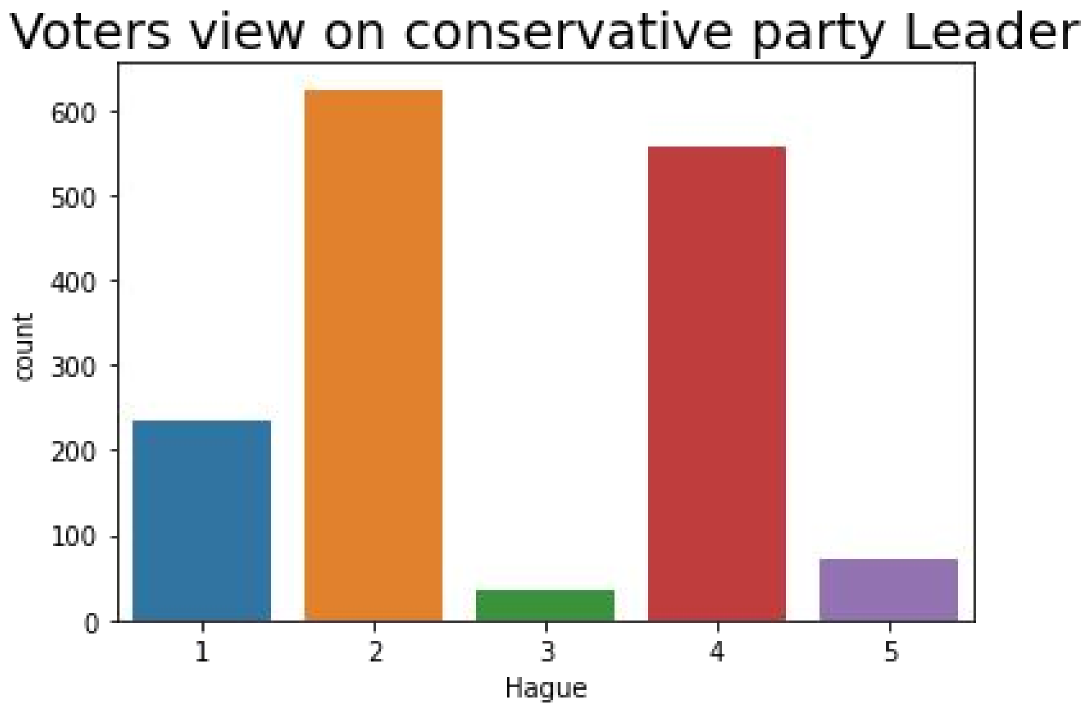


Table 8-voters view on Hague

Most of the voters have given 2 rating to the leader of conservative party leader.

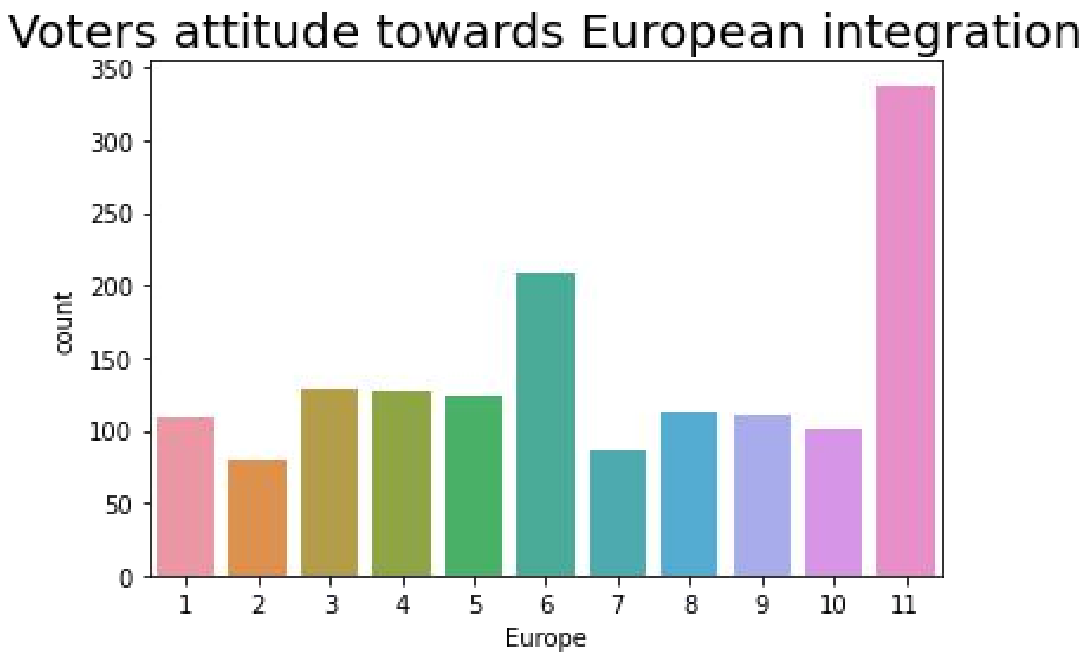


Table 9-voters attitude towards EU

Here higher the score, higher is the voters eurosceptic sentiment i.e they oppose increasing power of european union in their nation.

So, as it can be seen from the above graphs, most of the voters have given 11 rating which means European union is a boiling issue in this election and most voters have opposed the increasing power of EU in their nation.

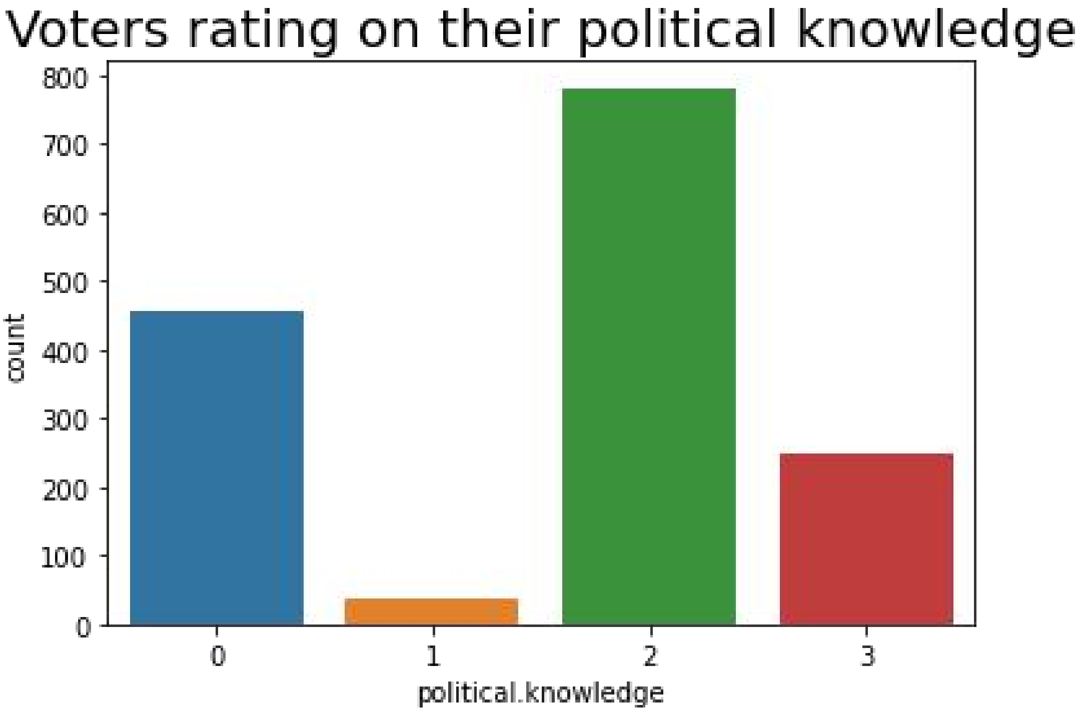


Table 10-Voters estimation of their political knowledge

Most of the voters have rated themselves 2 in political knowledge.

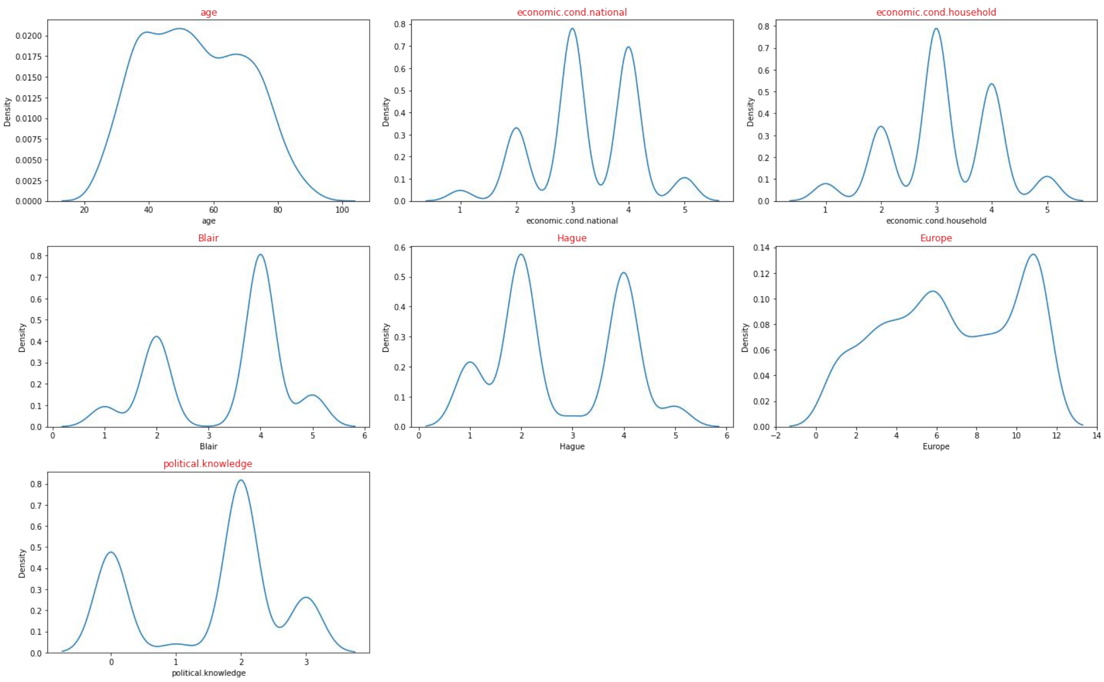


Table 11-Distribution curve

What we have interpreted from count plot, can also be seen here in distribution plots of each variable.

Here the distribution of "Age" shows that most of the voters are in the age bracket of 24 to 67 yrs.

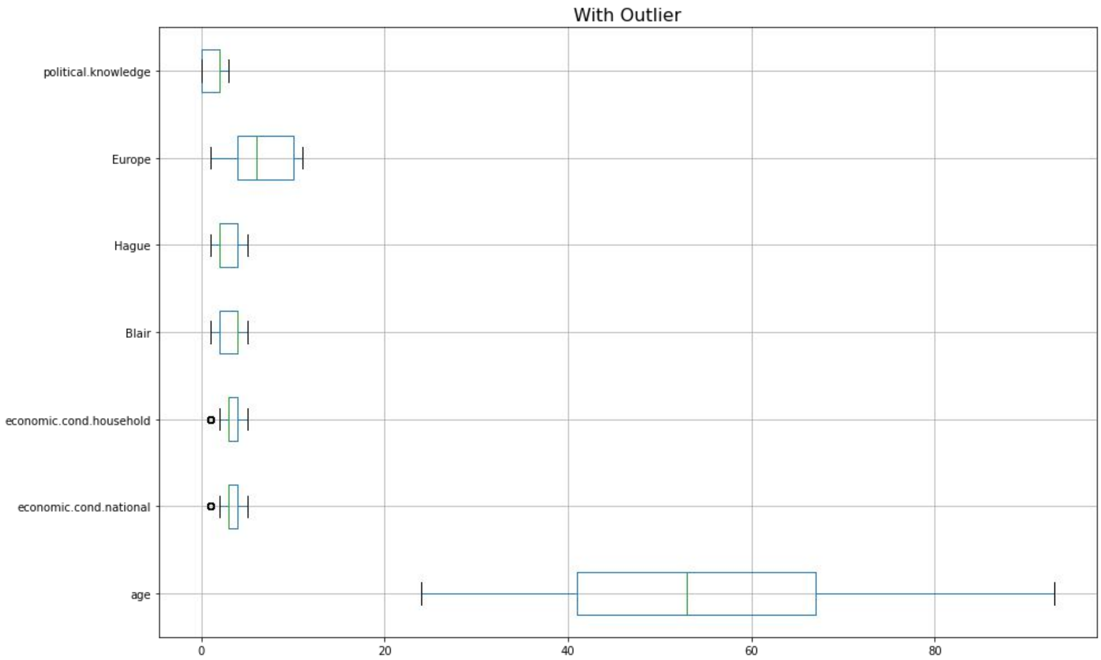


Table 12-Box plot with outliers

There are outliers observed in "economic condition household" and "economic condition national" variables.

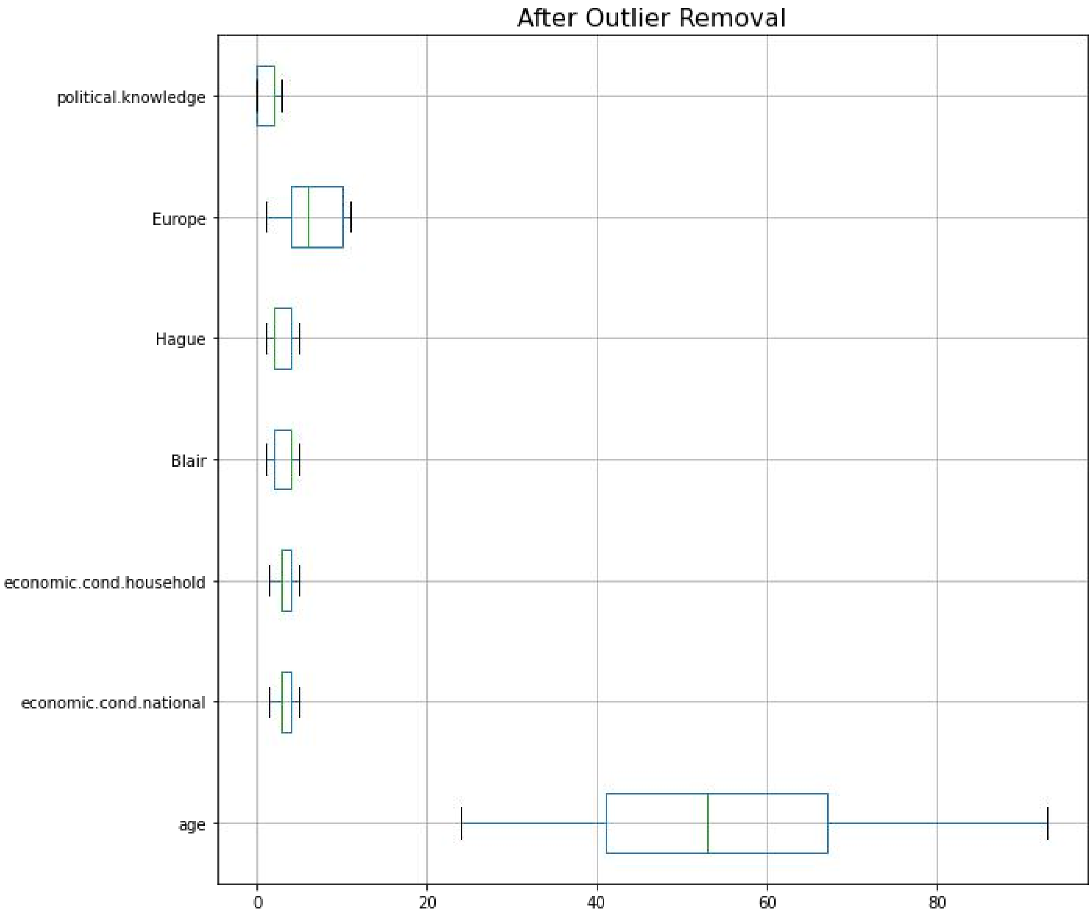


Table 13-Box plot without outlier

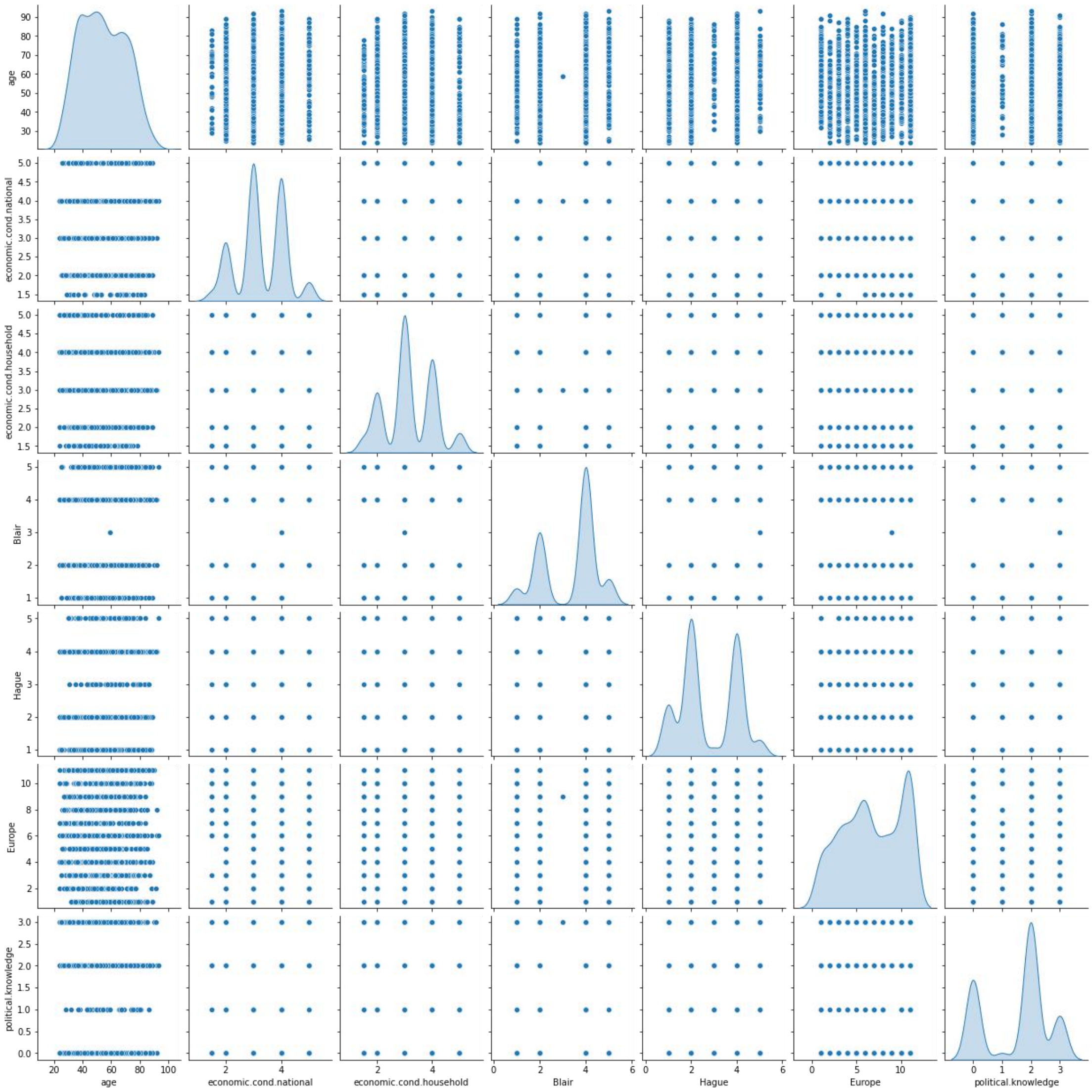


Table 14-Pairplot

from the above pair plot it can be observed that there are no correlations between the variables.

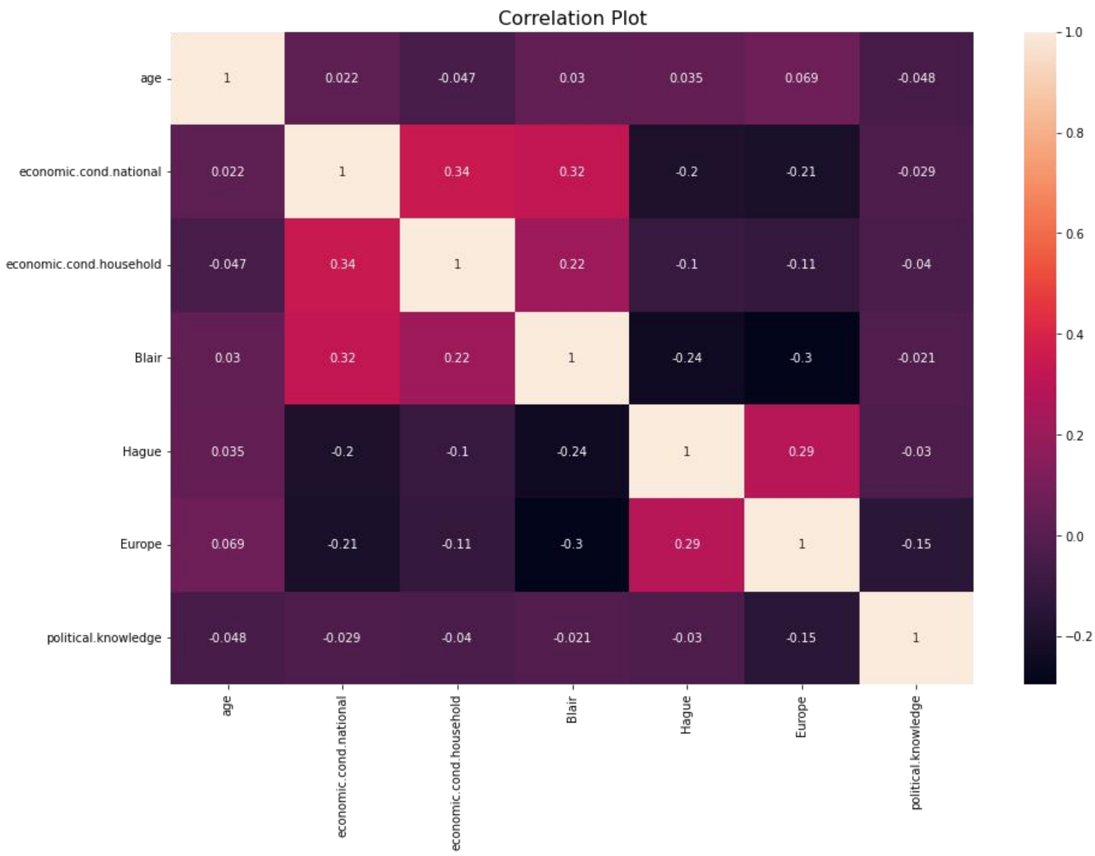


Table 15-Correlation plot

There are no major correlations observed.

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). (4 Marks)

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	1	43.0	3.0	3.0	4.0	1.0	2.0	2.0	1
1	1	36.0	4.0	4.0	4.0	4.0	5.0	2.0	0
2	1	35.0	4.0	4.0	5.0	2.0	3.0	2.0	0
3	1	24.0	4.0	2.0	2.0	1.0	4.0	0.0	1
4	1	41.0	2.0	2.0	1.0	1.0	6.0	2.0	0

Table 16-Encoded dataframe

“Gender” has been encoded male 0 and female 1 and Vote as Labour : 1 and Conservative: 0

```
age                         15.711209
economic.cond.national       0.852938
economic.cond.household      0.885286
Blair                        1.174824
Hague                        1.230703
Europe                       3.297538
political.knowledge          1.083315
gender                       0.499109
dtype: float64
```

The Standard Deviation is a statistic that indicates how much variance or dispersion there is in a group of statistics. A low Standard Deviation means that the value is close to the mean of the set (also known as the expected value), and a high Standard Deviation means that the value is spread over a wider area.

Looking at the above std values we can say that the "age" variable is having highest std and also not in the same scale so there is a need of scaling.

```
age                         246.842075
economic.cond.national        0.727503
economic.cond.household       0.783731
Blair                         1.380212
Hague                         1.514631
Europe                       10.873759
political.knowledge           1.173571
gender                        0.249110
dtype: float64
```

The variance is a numerical value that represents how broadly individuals in a group may change. The variance will be larger if the individual observations change largely from the group mean and vice versa.

Here looking at the above variances of each variables we can say that some of the variables like "Age" and "Europe" are having high variance values , so we have to scale the data to bring all the variables on the same scale.

1.4 Apply Logistic Regression and LDA (linear discriminant analysis). (4 marks)

Logistics Regression

```
              precision    recall  f1-score   support

           0       0.77      0.69      0.73       332
           1       0.87      0.91      0.89       735

    accuracy                           0.84      1067
   macro avg       0.82      0.80      0.81      1067
weighted avg       0.84      0.84      0.84      1067
```

Logistics regression applied and we got the above classification report on train data.

	precision	recall	f1-score	support
0	0.70	0.65	0.67	130
1	0.87	0.89	0.88	328
accuracy			0.82	458
macro avg	0.78	0.77	0.78	458
weighted avg	0.82	0.82	0.82	458

Test data classification report of Logistics regression is plotted above.

Linear Discriminant Analysis

	precision	recall	f1-score	support
0	0.69	0.66	0.67	130
1	0.87	0.88	0.87	328
accuracy			0.82	458
macro avg	0.78	0.77	0.77	458
weighted avg	0.82	0.82	0.82	458

LDA classification report on test data is plotted above.

	precision	recall	f1-score	support
0	0.76	0.71	0.73	332
1	0.87	0.90	0.89	735
accuracy			0.84	1067
macro avg	0.82	0.80	0.81	1067
weighted avg	0.84	0.84	0.84	1067

LDA classification report on train data is plotted above.

1.5 Apply KNN Model and Naive Bayes Model. Interpret the results. (4 marks)

KNN Model

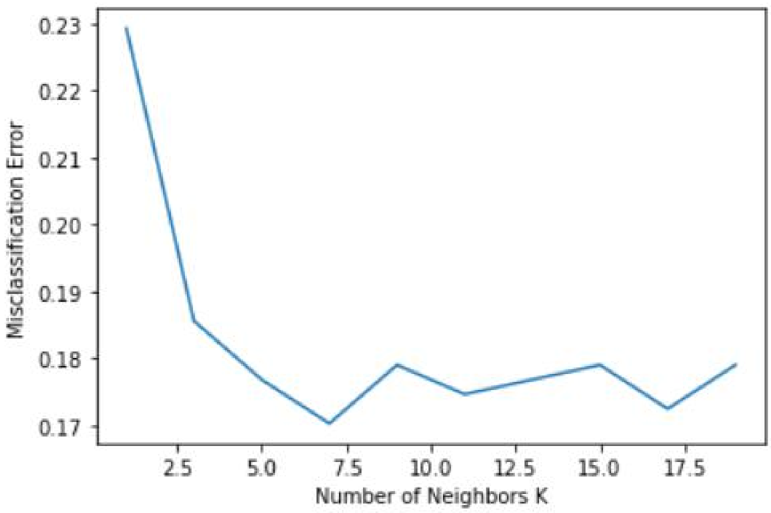
Different K values have been taken and accuracy score has been calculated.

```
Accuracy Score for K=3 is  0.8144104803493449
Accuracy Score for K=5 is  0.8231441048034934
Accuracy Score for K=9 is  0.8209606986899564
```

Miss-classification error w.r.t different K values have been plotted below.

```
[0.22925764192139741,
 0.18558951965065507,
 0.17685589519650657,
 0.17030567685589515,
 0.17903930131004364,
 0.1746724890829694,
 0.17685589519650657,
 0.17903930131004364,
 0.17248908296943233,
 0.17903930131004364]
```

We can plot a graph taking above values of MCE and K



K=3 is best.

Classification report of KNN using train data

```
[[104 228]
 [ 20 715]]
              precision    recall  f1-score   support

           0       0.84      0.31      0.46       332
           1       0.76      0.97      0.85       735

    accuracy                           0.77      1067
   macro avg       0.80      0.64      0.65      1067
weighted avg       0.78      0.77      0.73      1067
```

Classification report of KNN using test data

```
[[  8 122]
 [  1 327]]
              precision    recall  f1-score   support

           0       0.89      0.06      0.12       130
           1       0.73      1.00      0.84       328

    accuracy                           0.73       458
   macro avg       0.81      0.53      0.48       458
weighted avg       0.77      0.73      0.64       458
```

Naive Bayes

Naive Bayes applied and the classification report on train data is plotted below.

```
[[ 94  36]
 [ 43 285]]
              precision    recall  f1-score   support

           0       0.69      0.72      0.70       130
           1       0.89      0.87      0.88       328

    accuracy                           0.83       458
   macro avg       0.79      0.80      0.79       458
weighted avg       0.83      0.83      0.83       458
```

Naive bayes classification report on test data.

```
[[240  92]
 [ 87 648]]
              precision    recall  f1-score   support

           0       0.73      0.72      0.73       332
           1       0.88      0.88      0.88       735

    accuracy                           0.83      1067
   macro avg       0.80      0.80      0.80      1067
weighted avg       0.83      0.83      0.83      1067
```

1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. (7 marks)

Ada Boost Classifier

Classification report on train data

```
[[236  96]
 [ 72 663]]
              precision    recall  f1-score   support

           0       0.77      0.71      0.74       332
           1       0.87      0.90      0.89       735

    accuracy                           0.84      1067
   macro avg       0.82      0.81      0.81      1067
weighted avg       0.84      0.84      0.84      1067
```

AdaBoost Classification report on test data.

```
[[ 90  40]
 [ 41 287]]
              precision    recall  f1-score   support

           0       0.69      0.69      0.69       130
           1       0.88      0.88      0.88       328

    accuracy                           0.82       458
   macro avg       0.78      0.78      0.78       458
weighted avg       0.82      0.82      0.82       458
```

Gradient Boosting

Classification report on train data.

```
[[262  70]
 [ 51 684]]
              precision    recall  f1-score   support

           0       0.84      0.79      0.81       332
           1       0.91      0.93      0.92       735

    accuracy                           0.89      1067
   macro avg       0.87      0.86      0.87      1067
weighted avg       0.89      0.89      0.89      1067
```

Classification report on test data.

```
[[ 96  34]
 [ 43 285]]
              precision    recall  f1-score   support

           0       0.69      0.74      0.71       130
           1       0.89      0.87      0.88       328

    accuracy                           0.83       458
   macro avg       0.79      0.80      0.80       458
weighted avg       0.84      0.83      0.83       458
```

Bagging

Classification report of Bagging classifier on train data.

```
[[305  27]
 [  9 726]]
              precision    recall  f1-score   support

           0       0.97      0.92      0.94       332
           1       0.96      0.99      0.98       735

    accuracy                           0.97      1067
   macro avg       0.97      0.95      0.96      1067
weighted avg       0.97      0.97      0.97      1067
```

Classification report on test data.

```
[[ 92  38]
 [ 37 291]]
              precision    recall  f1-score   support

           0       0.71      0.71      0.71       130
           1       0.88      0.89      0.89       328

    accuracy                           0.84       458
   macro avg       0.80      0.80      0.80       458
weighted avg       0.84      0.84      0.84       458
```

Hyper-parameter Tuning

AdaBoost Classifier

Classification report of Hyper tuned model with train data.

```
[[128 204]
 [ 30 705]]
              precision    recall  f1-score   support

           0       0.81      0.39      0.52       332
           1       0.78      0.96      0.86       735

    accuracy                           0.78      1067
   macro avg       0.79      0.67      0.69      1067
weighted avg       0.79      0.78      0.75      1067
```

Classification report of Hyper tuned model with test data.

```
[[ 57  73]
 [ 13 315]]
              precision    recall  f1-score   support

           0       0.81      0.44      0.57       130
           1       0.81      0.96      0.88       328

    accuracy                           0.81       458
   macro avg       0.81      0.70      0.72       458
weighted avg       0.81      0.81      0.79       458
```

Gradient Boosting

Classification report of hyper tuned model using train data.

```
[[188 144]
 [ 40 695]]
              precision    recall  f1-score   support

           0       0.82      0.57      0.67       332
           1       0.83      0.95      0.88       735

    accuracy                           0.83      1067
   macro avg       0.83      0.76      0.78      1067
weighted avg       0.83      0.83      0.82      1067
```

Classification report of hyper tuned model using test data.

```
[[ 67  63]
 [ 24 304]]
              precision    recall  f1-score   support

           0       0.74      0.52      0.61       130
           1       0.83      0.93      0.87       328

    accuracy                           0.81       458
   macro avg       0.78      0.72      0.74       458
weighted avg       0.80      0.81      0.80       458
```

Table 17

KNN

Classification report of KNN model after hyper tuning using train data

```
[[ 12 320]
 [  4 731]]
              precision    recall  f1-score   support

           0       0.75      0.04      0.07       332
           1       0.70      0.99      0.82       735

    accuracy                           0.70      1067
   macro avg       0.72      0.52      0.44      1067
weighted avg       0.71      0.70      0.59      1067
```

Classification report of KNN model after hyper tuning using test data

```
[[  8 122]
 [  1 327]]
```

	precision	recall	f1-score	support
0	0.89	0.06	0.12	130
1	0.73	1.00	0.84	328
accuracy			0.73	458
macro avg	0.81	0.53	0.48	458
weighted avg	0.77	0.73	0.64	458

Linear Discriminant Analysis

Classification report after hyper tuned LDA model using train data

	precision	recall	f1-score	support
0	0.76	0.71	0.73	332
1	0.87	0.90	0.89	735
accuracy			0.84	1067
macro avg	0.82	0.80	0.81	1067
weighted avg	0.84	0.84	0.84	1067

Table 18

Classification report after hyper tuned LDA model using test data

	precision	recall	f1-score	support
0	0.69	0.66	0.67	130
1	0.87	0.88	0.87	328
accuracy			0.82	458
macro avg	0.78	0.77	0.77	458
weighted avg	0.82	0.82	0.82	458

Logistics Regression

Classification report of hyper tuned LR model on train data.

	precision	recall	f1-score	support
0	0.77	0.69	0.73	332
1	0.87	0.91	0.89	735
accuracy			0.84	1067
macro avg	0.82	0.80	0.81	1067
weighted avg	0.84	0.84	0.84	1067

Classification report of hyper tuned LR model on test data.

	precision	recall	f1-score	support
0	0.70	0.65	0.67	130
1	0.87	0.89	0.88	328
accuracy			0.82	458
macro avg	0.78	0.77	0.78	458
weighted avg	0.82	0.82	0.82	458

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 the models and write inference which model is best/optimized. (7 marks)

All hyper tuned model AUC score and ROC curve

AdaBoost

AUC Score - 0.865

ROC curve in train data

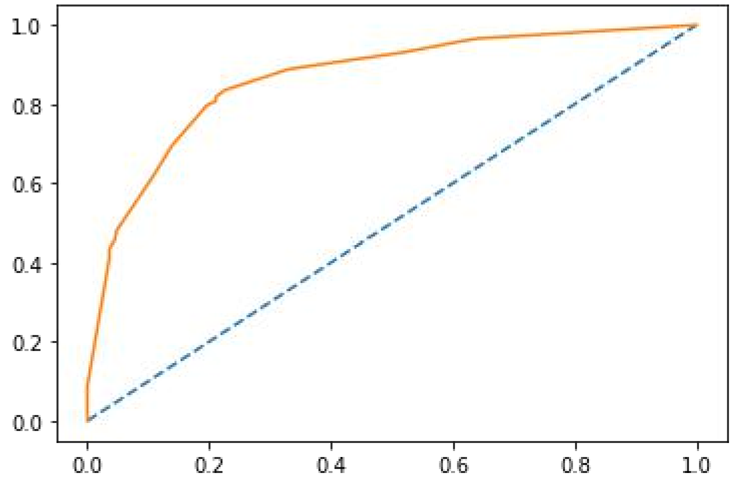


Table 19-ROC curve of Adaboost-train

ROC curve in test data

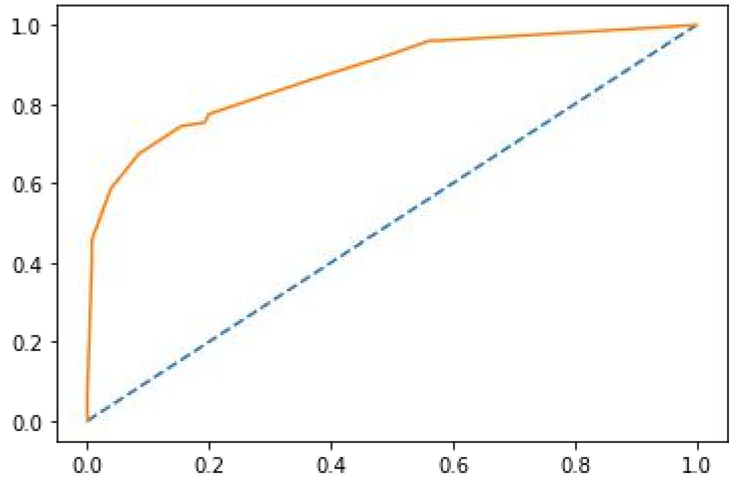


Table 20-ROC curve of Adaboost-test

Gradient Boosting

AUC score : 0.90

ROC Curve in train data

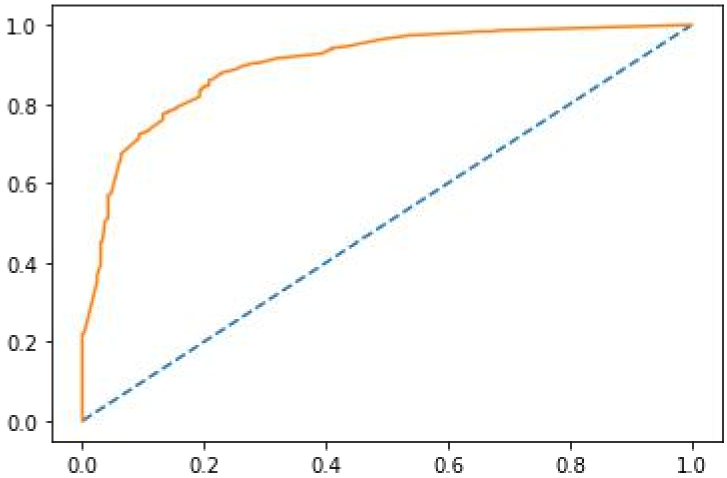


Table 21-ROC curve of Gradient Boosting-train

ROC Curve in test data

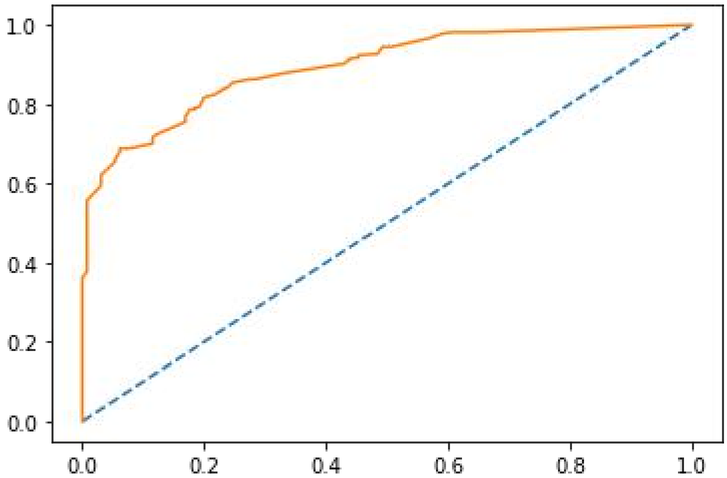


Table 22-ROC curve of Gradient boosting -test

KNN

AUC score : 0.792

ROC Curve using train data

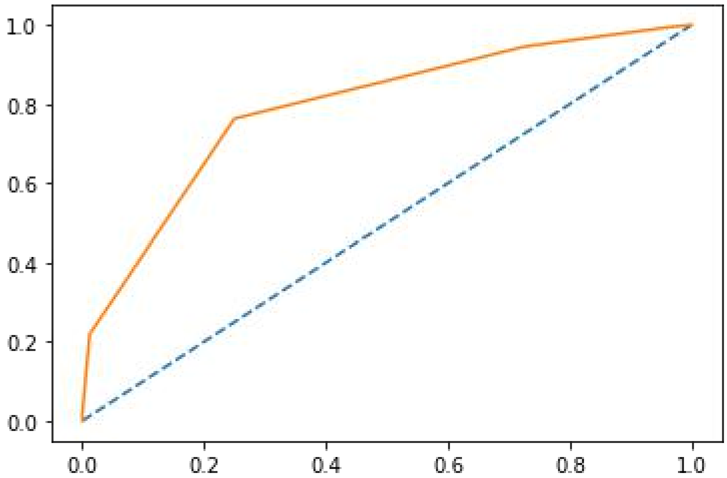


Table 23-KNN ROC curve-train

ROC Curve using test data

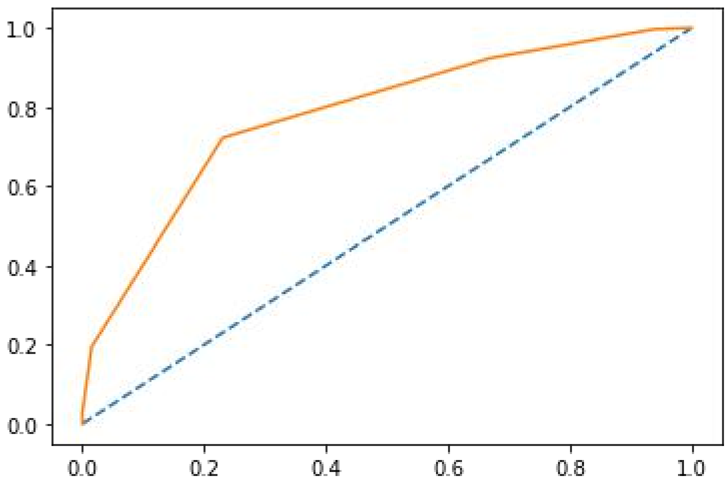


Table 24-KNN ROC Curve test

LDA

AUC score : 0.749

ROC Curve using train data

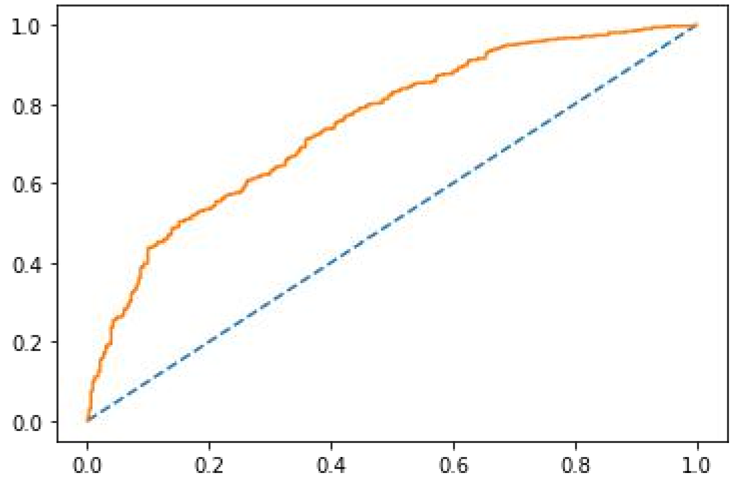


Table 25-LDA ROC curve -train

ROC Curve using test data

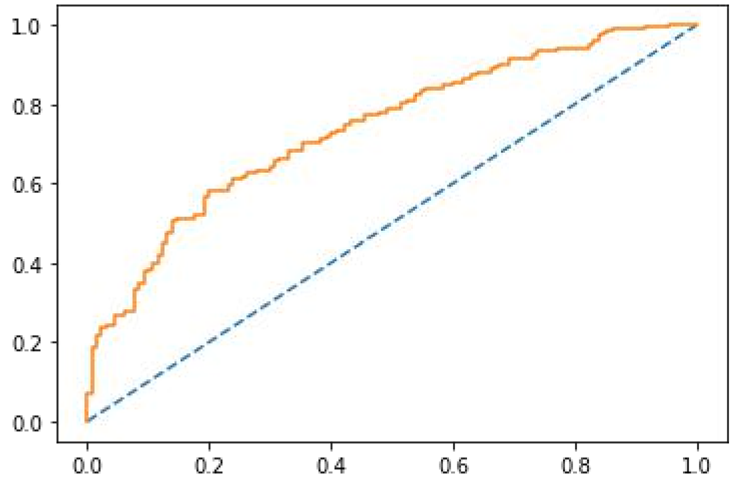


Table 26-LDA ROC curve-test

Logistics Regression

AUC Score : 0.76

ROC Curve using train data

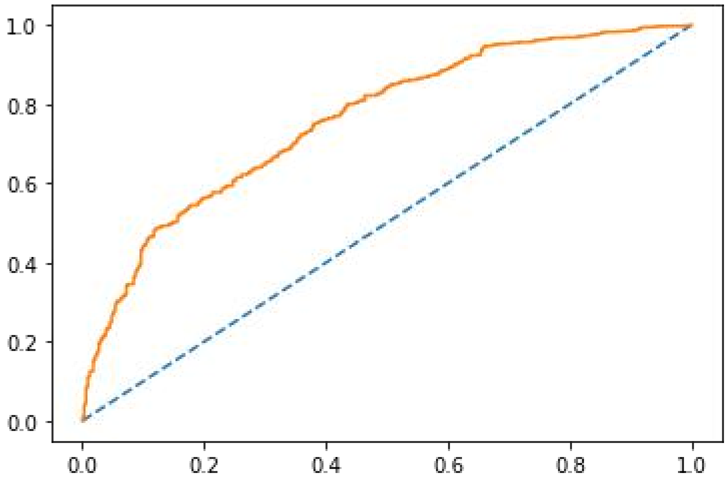


Table 27-LR ROC curve -train

ROC Curve using test data

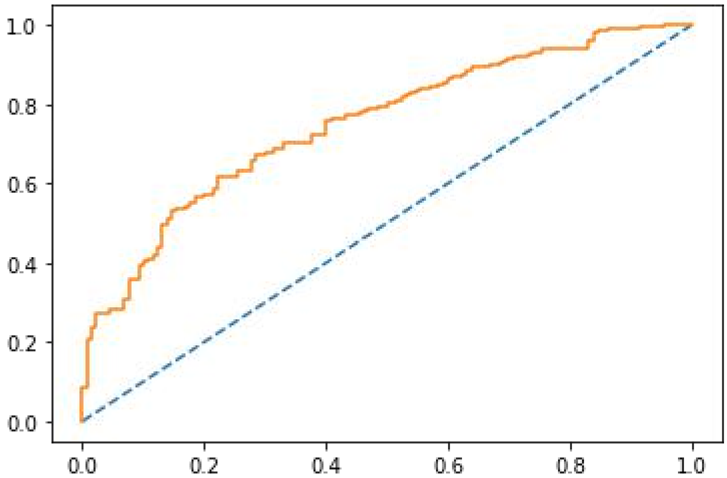


Table 28-LR ROC curve -test

Bagging

AUC Score : 0.997

ROC Curve using train data

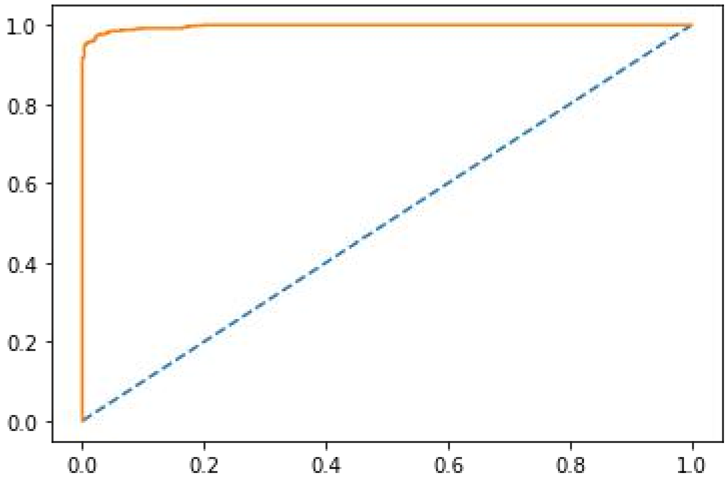


Table 29-Bagging ROC curve-train

ROC Curve using test data

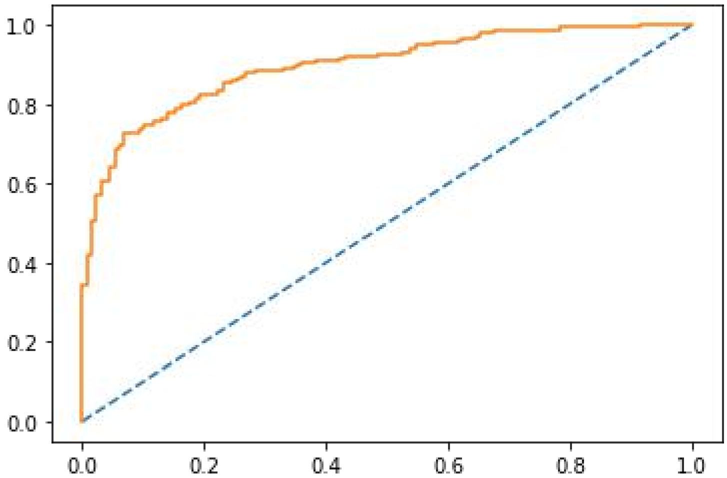


Table 30-Bagging ROC curve-test

Table 31

Feature Importance

	Imp
Hague	0.321766
Europe	0.175909
Blair	0.134201
age	0.132741
political.knowledge	0.106378
economic.cond.household	0.061174
economic.cond.national	0.055050
gender	0.012781

Comparison of All model performance

Class of interest is 1 (I.e Labour=1)

Recall refers to the percentage of total relevant results correctly classified by the algorithm and hence we will compare Recall of class "1" for all models.

F1-score metric is to find an equal balance between precision and recall, which is extremely useful in most scenarios when we are working with imbalanced data sets

Model (Regular)	Model (Hyper-tuned)	Recall		F1 Score		Remarks
		Train	Test	Train	Test	
Logistics Regression		0.91	0.89	0.89	0.88	
Linear Discriminant Analysis		0.90	0.88	0.87	0.89	
KNN		0.97	0.97	0.85	0.86	
Naive Bayes		0.88	0.87	0.88	0.88	
Bagging		0.99	0.89	0.98	0.89	
AdaBoost Classifier		0.90	0.88	0.89	0.88	
Gradient Boosting		0.93	0.87	0.92	0.88	Best model
	AdaBoost Classifier	0.96	0.96	0.86	0.88	
	Gradient Boosting	0.95	0.93	0.88	0.87	
	KNN	0.99	1	0.82	0.84	Over fitting problem
	Linear Discriminant Analysis	0.90	0.88	0.89	0.87	
	Logistics Regression	0.91	0.89	0.89	0.88	

Gradient boosting is the best model considering the recall score and F1 score values.

1.8 Based on these predictions, what are the insights? (5 marks)

Insights :

1. There are more supporters of Labour party than conservative party.
2. Voters view on Labour party leader Blair is very positive.
3. There are more female voters than male voters.
4. European union integration can be an issue in the election so the party which is supporting this issue can have better chances of winning.
5. The party which has more voters base in females are likely to win.
6. Issues related to women safety and female LFPR can be given importance.
7. Voters don’t give economic condition a priority.

Problem-2

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

President Franklin D. Roosevelt in 1941

President John F. Kennedy in 1961

President Richard Nixon in 1973

2.1 Find the number of characters, words, and sentences for the mentioned documents. – 3 Marks

	Total No of Characters	Total No of Words	Total No of Sentences
Roosevelt	7571	1536	67
Kennedy	7618	1546	52
Nixon	9991	2028	68

2.2 Remove all the stopwords from all three speeches. – 3 Marks

```
['national',
 'day',
 'inauguration',
 'since',
 '1789,',
 'people',
 'renewed',
 'sense',
 'dedication',
 'united',
 'states.',
 "washington's",
 'day',
 'task',
 'people',
 'create',
 'weld',
 'together',
 'nation.',
```

All stopwords removed

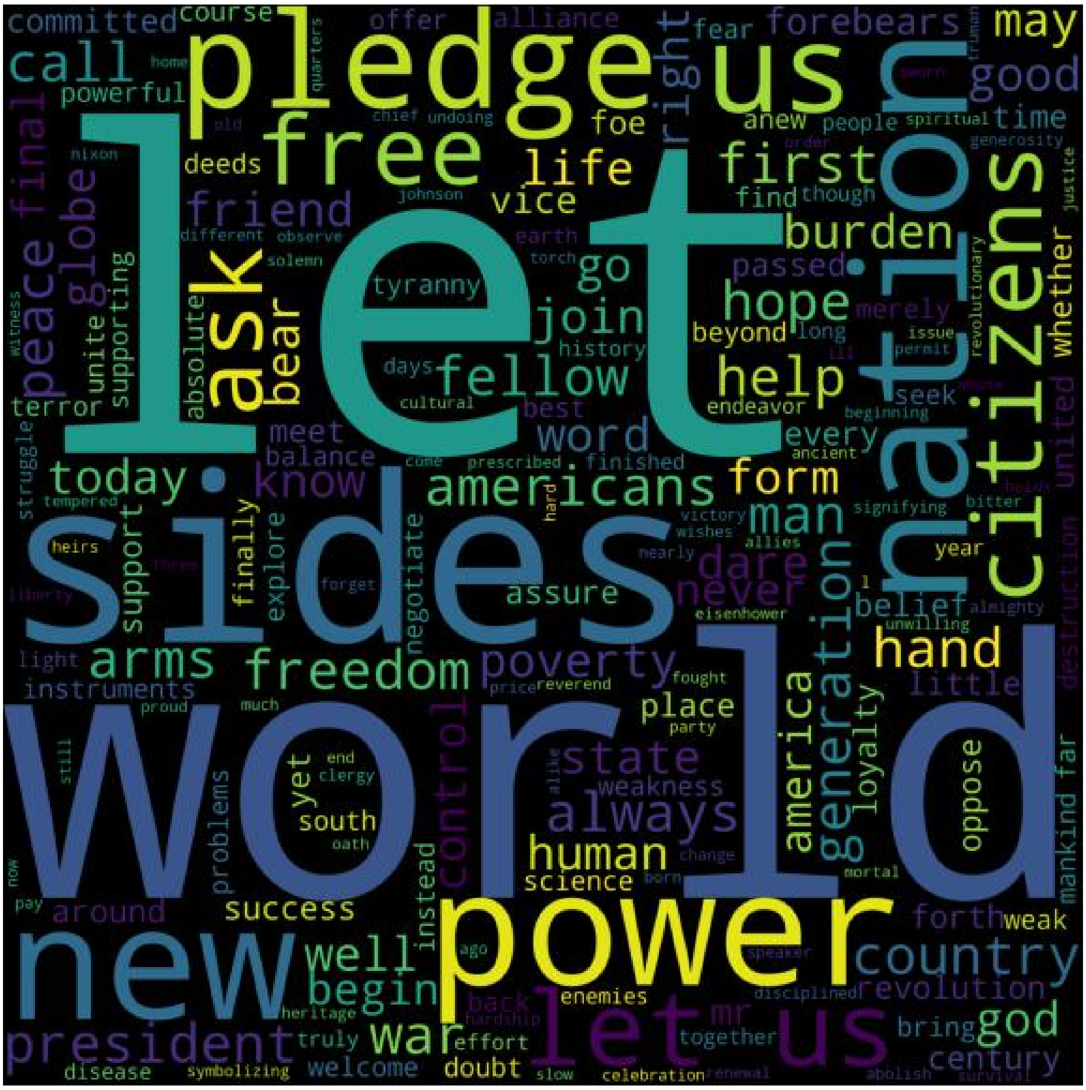
2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) – 3 Marks

	Most frequent word	Frequency
Roosevelt	Nation	12
Kennedy	Let	16
Nixon	us	26

2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks [refer to the End-to-End Case Study done in the Mentored Learning Session]

Word Cloud for Roosevelt

Word cloud of Kennedy



Word cloud of Nixon