

Machine Learning Project Report

Deepti Yadav
Aug'22
Date : 28/08/2022

Table of Contents

Problem 1.....	4
1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.....	5
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers	12
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).	30
1.4 Apply Logistic Regression and LDA (linear discriminant analysis)	33
1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results	35
1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.	38
Bagging using RandomForest.....	46
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.	52
1.8 Based on these predictions, what are the insights?	64
Problem 2:.....	66
2.1 Find the number of characters, words, and sentences for the mentioned documents.	66
2.2 Remove all the stopwords from all three speeches.	75
2.3 Which word occurs the most number of times in his inaugural address for each president?	76
2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords)	78

List of Figures

Figure 1 – Proportion of Votes	14
Figure 2 – Proportion of gender	14
Figure 3 – Votes Vs Gender	15
Figure 4 – National Economic condition	15
Figure 5 – National Economic condition of Labour voters	16
Figure 6 – National Economic condition of conservative voters	16
Figure 7 – Household Economic condition	17
Figure 8 – Household Economic condition of Labour voters	17
Figure 9 – Household Economic condition of Conservative voters	18
Figure 10 – Boxplot for vote v/s age	19
Figure 11 – Boxplot for vote v/s household economic condition	19
Figure 12 – Boxplot for vote for Blair	20
Figure 13 – Boxplot for vote for Hague	20
Figure 14 – Boxplot for vote v/s Eurosceptic sentiment	21
Figure 15 – Boxplot for vote v/s political Knowledge	22
Figure 16 – Boxplot for vote v/s gender	22
Figure 17 – Boxplot for gender v/s national economic condition	23
Figure 18 – Boxplot for gender v/s household economic condition	24
Figure 19 – Boxplot for vote to Blair for among different gender	24
Figure 20 – Boxplot for vote to Hague among different gender	25
Figure 21 – Boxplot for Eurosceptic sentiment among different gender	25
Figure 22 – Boxplot for political knowledge among different gender	26
Figure 23 - Pairplot	27
Figure 24 – Heat Map	28
Figure 25 – Misclassificationn error for different k value	42
Figure 24 – Feature importance for different variables	46

List of Tables

Table 1- Original Dataset Description	5
Table 2- Dataset Description after dropping Unnamed column	6
Table 3 - Dataset Information	6
Table 4- Dataset Description (after dropping Unnamed column)	7
Table 5 - Dataset Information (after dropping Unnamed column)	7
Table 6 – Five point summary (numerical variables)	7
Table 7 – Five point summary (categorical variables)	8
Table 8 - Missing values Check	8
Table 9 – Checking for duplicates	8
Table 10 – Univariate Analysis	12
Table 11 – Skewness Analysis	13
Table 12 – Correlation Values	29
Table 13 – Data set with encoding	30
Table 14 – Data info with categorical variables	31

Table 14 – Feature importance values for different variables	46
Table 15 – Comparison of all model	52
Table 16 – Speeches into Data frame	73
Table 17 – Number of words, characters and sentences	73
Table 18 – Number of Uppercase Words	75
Table 19 – Number of Uppercase Letters.....	75
Table 20 – Number of Numeric	75
Table 21 – Lower case conversion	75
Table 22 – Remove punctuation.....	76
Table 23 – Removing stop words and perform stemming	76
Table 24 – Word count after removing stopwords and cleaning.....	76

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.

Data Dictionary:

Variable Name	Description
vote	Party choice: Conservative or Labour
age	in years
economic.cond.national	Assessment of current national economic conditions, 1 to 5.
economic.cond.household:	Assessment of current household economic conditions, 1 to 5.
Blair	Assessment of the Labour leader, 1 to 5.
Hague	Assessment of the Conservative leader, 1 to 5.
Europe	an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
political.knowledge	Knowledge of parties' positions on European integration, 0 to 3.
gender	female or male.

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.

Table 1- Original Dataset Description

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

Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender	
1520	1521	Conservative	67	5	3	2	4	11	3	male
1521	1522	Conservative	73	2	2	4	4	8	2	male
1522	1523	Labour	37	3	3	5	4	2	2	male
1523	1524	Conservative	61	3	3	1	4	11	2	male
1524	1525	Conservative	74	2	3	2	4	11	0	female

no. of rows: 1525

no. of columns: 10

The dataset has Unnamed column which we are going to drop as it contains serial numbers.

Table 2- Dataset Description after dropping Unnamed column

	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

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1520	Conservative	67	5	3	2	4	11	3	male
1521	Conservative	73	2	2	4	4	8	2	male
1522	Labour	37	3	3	5	4	2	2	male
1523	Conservative	61	3	3	1	4	11	2	male
1524	Conservative	74	2	3	2	4	11	0	female

no. of rows: 1525

no. of columns: 9

Table 3 - Dataset Information

```
<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
```

Table 4- Dataset Description (after dropping Unnamed column)

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	Ideal	F	VS1	60.4	59.0	4.35	4.43	2.65	779

Table 5 - Dataset Information (after dropping Unnamed column)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
#   Column      Non-Null Count  Dtype
---  -
0   carat        26967 non-null  float64
1   cut          26967 non-null  object
2   color        26967 non-null  object
3   clarity      26967 non-null  object
4   depth        26270 non-null  float64
5   table        26967 non-null  float64
6   x            26967 non-null  float64
7   y            26967 non-null  float64
8   z            26967 non-null  float64
9   price        26967 non-null  int64
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

we need some summary statistics of our data frame. For this, we can use describe() method. It can be used to generate various summary statistics.

Table 6 – Five point summary (numerical variables)

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

Table 7 – Five point summary (categorical variables)

	count	unique	top	freq
vote	1525	2	Labour	1063
gender	1525	2	female	812

Table 8 - Missing values Check

```

vote          0
age           0
economic.cond.national  0
economic.cond.household  0
Blair         0
Hague        0
Europe       0
political.knowledge  0
gender        0
dtype: int64

```

The above output shows that there is no “null” value in our dataset.

Table 9 – Checking for duplicates

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
67	Labour	35	4	4	5	2	3	2	male
626	Labour	39	3	4	4	2	5	2	male
870	Labour	38	2	4	2	2	4	3	male
983	Conservative	74	4	3	2	4	8	2	female
1154	Conservative	53	3	4	2	2	6	0	female
1236	Labour	36	3	3	2	2	6	2	female
1244	Labour	29	4	4	4	2	2	2	female
1438	Labour	40	4	3	4	2	2	2	male

These duplicates need to be dropped because they do not add any value to the study, be it associated with different people.

Number of rows after dropping them = 1517

Columns name after renaming

'vote', 'age', 'economic_cond_national', 'economic_cond_household', 'Blair', 'Hague', 'Europe',
'political_knowledge', 'gender'

Value counts for categorical variables

VOTE : 2

Conservative 460

Labour 1057

Name: vote, dtype: int64

Clearly there is a major imbalance in the data

GENDER : 2

male 709

female 808

Name: gender, dtype: int64

ECONOMIC_COND_NATIONAL : 5

1 37

5 82

2 256

4 538

3 604

Name: economic_cond_national, dtype: int64

ECONOMIC_COND_HOUSEHOLD : 5

1 65

5 92

2 280

4 435

3 645

Name: economic_cond_household, dtype: int64

BLAIR : 5

3 1

1 97

5 152

2 434

4 833

Name: Blair, dtype: int64

HAGUE : 5

3 37

5 73

1 233

4 557

2 617

Name: Hague, dtype: int64

EUROPE : 11

2 77

7 86

10 101

1 109

9 111

8 111

5 123

4 126

3 128

6 207

11 338

Name: Europe, dtype: int64

POLITICAL_KNOWLEDGE : 4

1 38

3 249

0 454

2 776

Inference:

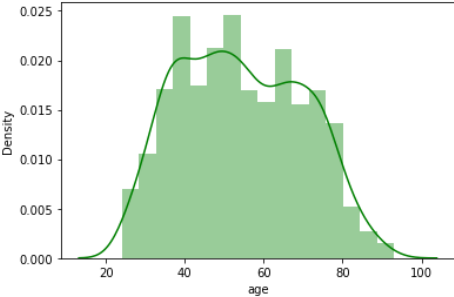
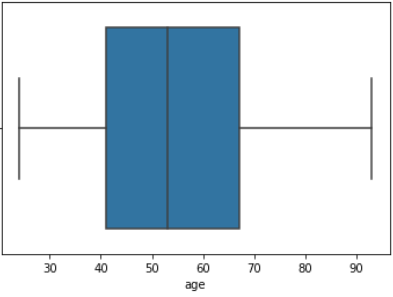
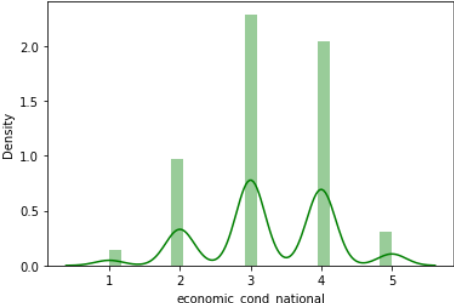
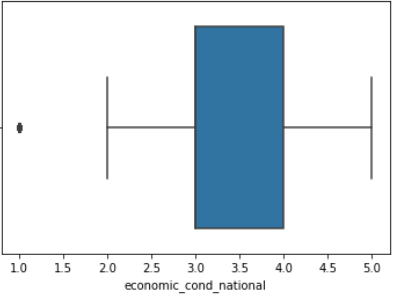
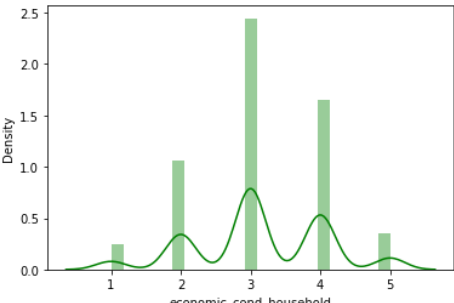
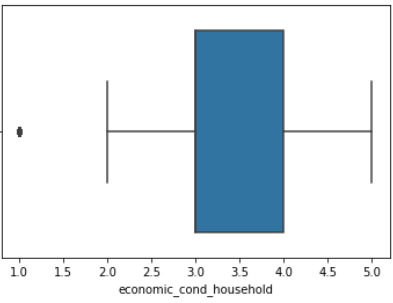
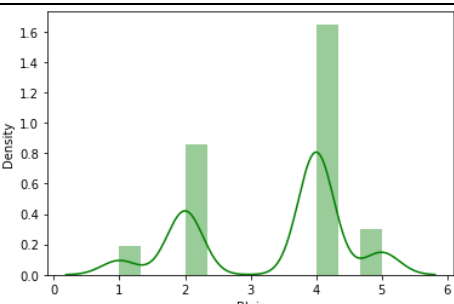
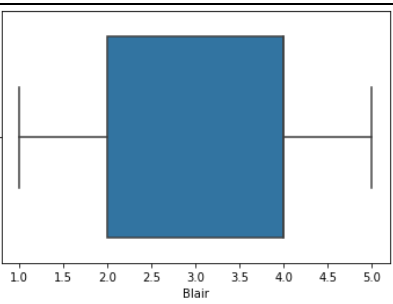
- There are total 1525 rows representing voters and 10 columns with 9 variables. Out of 10, 2 columns are of object type and 8 columns are of integer type.
- Data does not contain missing values.
- The first column is an index ("Unnamed: 0") as these are only serial numbers, we dropped it as it does not add value to our analysis.
- After dropping "Unnamed: 0", data now contains 1525 rows and 9 columns.
- There are 2 types of voting parties- Labour and Conservative. Labour party tops in number of votes.

- There are 2 types of genders voting- Male and Female with Female being the topmost voters.
- Minimum age of an individual voting is 24 years and maximum age is 93 years. Mean voting age is 54 years.
- Minimum assessment of current national economic conditions is 1 and a maximum assessment is 5 with an average assessment of 3.
- Minimum assessment of current household economic conditions 1 and a maximum assessment is 5 with an average assessment of 3.
- Minimum assessment of the Labour leader Blair is 1 and maximum assessment is 5 with an average assessment of 4.
- Minimum assessment of the Conservative leader Hague is 1 and maximum assessment is 5 with an average assessment of 2.
- There are 34 duplicate rows but looking at the row for every column entry, it is found that all columns are not duplicate. So, dropping duplicate rows is not recommended.
- 75% of the voters on a 11-point scale that measures respondent's attitudes toward European integration represent high 'Eurosceptic' sentiment with a maximum scale of 11 and a minimum scale of 1.
- On an average knowledge of parties positions on European integration is 2. Approximately 25% of parties do not hold positions on European integration with a maximum holding of 3.
- After dropping duplicate rows and unnamed columns, total number of column is 9 and total number of rows is 1517.

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers

Univariate Analysis

Table 10 – Univariate Analysis

Description	Distribution plot	Boxplot
Description of age <hr/> 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 Distribution of age <hr/>		
Description of economic_cond_national <hr/> count 1517.000000 mean 3.245221 std 0.881792 min 1.000000 25% 3.000000 50% 3.000000 75% 4.000000 max 5.000000 		
Description of economic_cond_household <hr/> count 1517.000000 mean 3.137772 std 0.931069 min 1.000000 25% 3.000000 50% 3.000000 75% 4.000000 max 5.000000 		
Description of Blair <hr/> count 1517.000000 mean 3.335531 std 1.174772 min 1.000000 25% 2.000000 50% 4.000000 75% 4.000000 max 5.000000 		

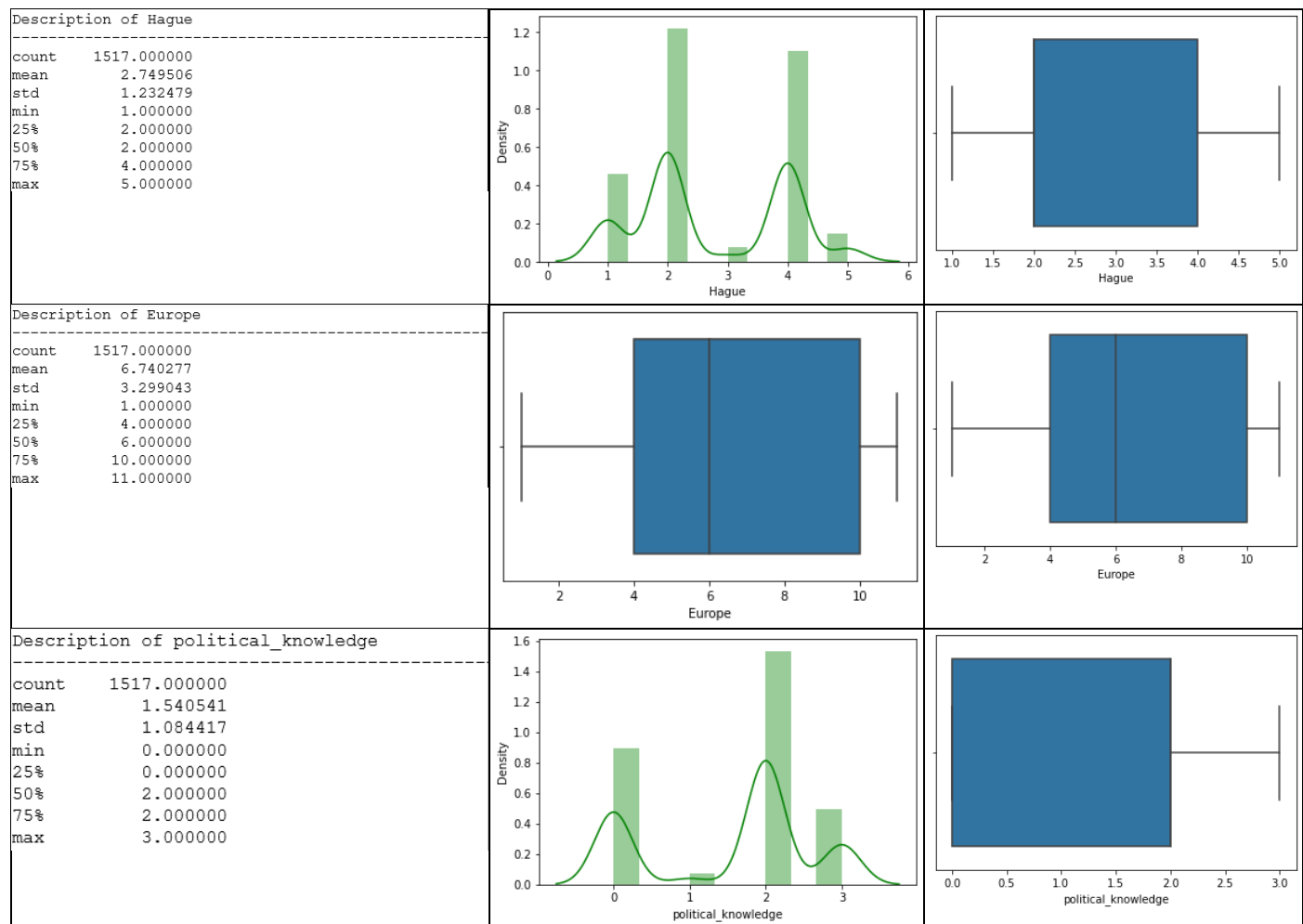


Table 11 – Skewness Analysis

Skewness values	
age	0.139800
economic.cond.national	-0.238474
economic.cond.household	-0.144148
Blair	-0.539514
Hague	0.146191
Europe	-0.141891
political.knowledge	-0.422928
dtype: float64	

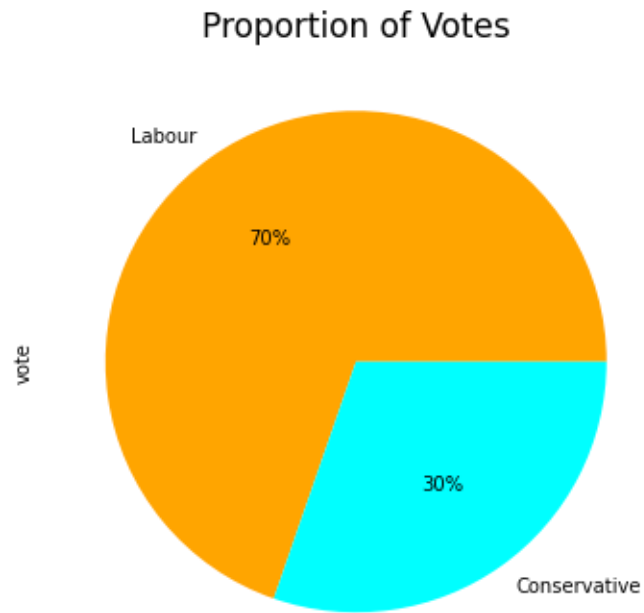


Figure 1 – Proportion of Votes

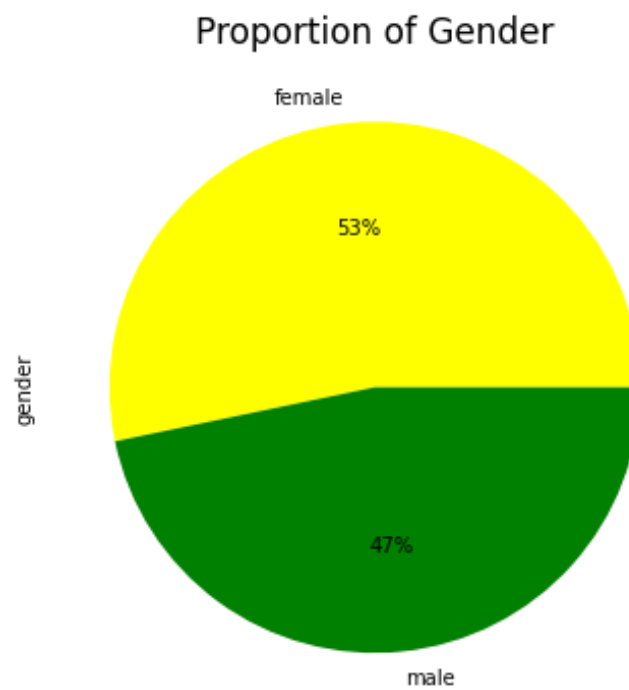


Figure 2 – Proportion of gender

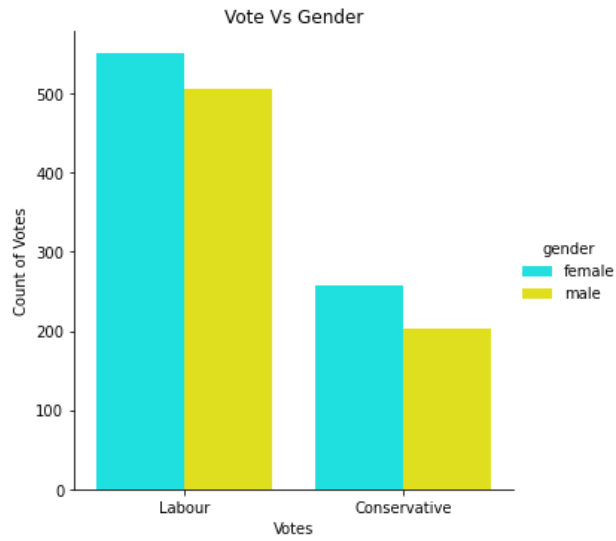


Figure 3 – Votes Vs Gender

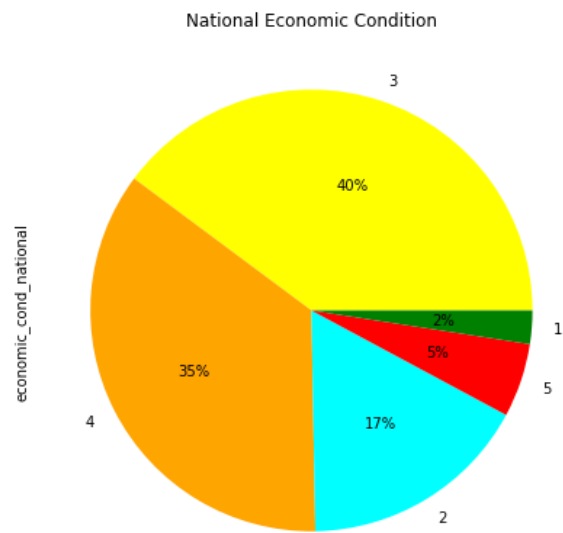


Figure 4 – National Economic condition

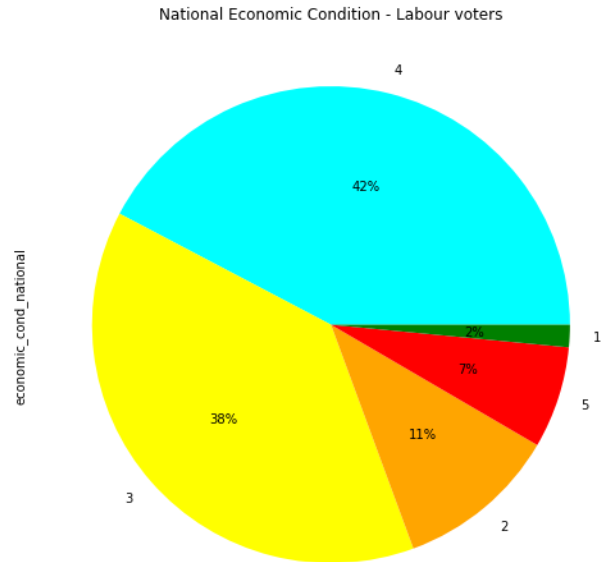


Figure 5 – National Economic condition of Labour voters

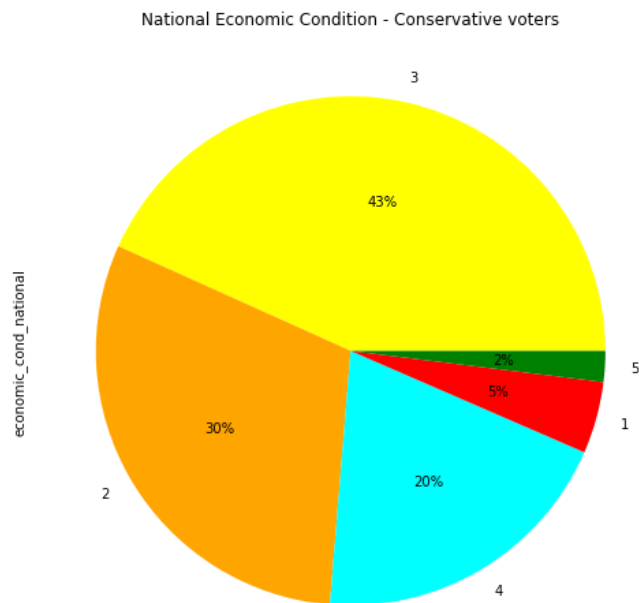


Figure 6 – National Economic condition of conservative voters

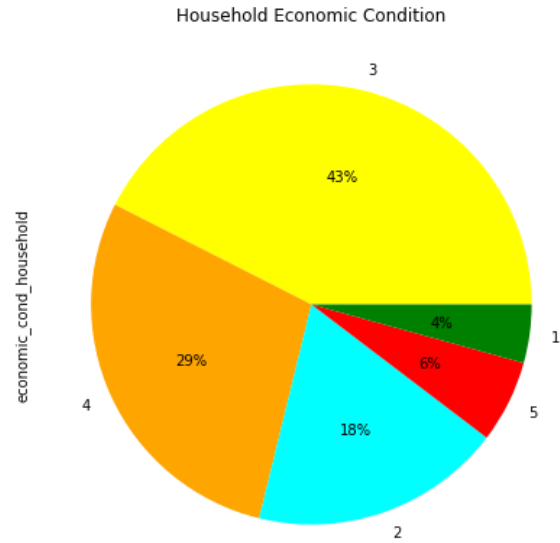


Figure 7 – Household Economic condition

Household Economic Condition - Labour Voters

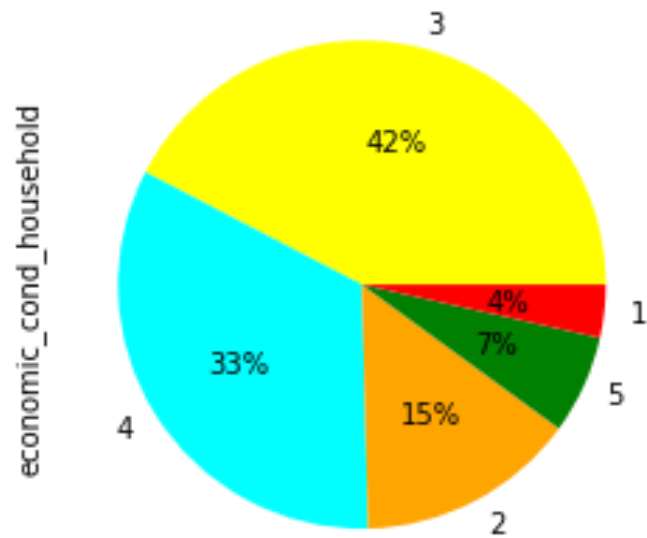


Figure 8 – Household Economic condition of Labour voters

Household Economic Condition - Conservative voters

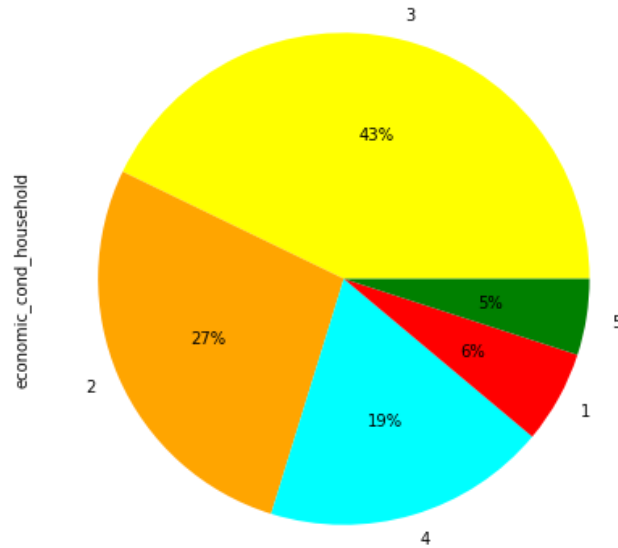


Figure 9 – Household Economic condition of Conservative voters

Insights:

- All the variables except age are not normally distributed (right skewed) as they are categorical in nature. Age is almost normally distributed with multimode which signifies presence of some clusters.
- The medians of variables "Blair", "Hague", "economic_cond_national" and "economic_cond_household" are identical to the first quartile, which is why there is an overlap in the Boxplot. This could be because data might have identical large proportion of low values.
- We can also confirm presence of outliers in variables "economic_cond_national" and "economic_cond_household". Since the lower quartile and middle quartile values are same (i.e., 0), variable "political_knowledge" does not have a lower whisker and middle whisker.

Bivariate & Multivariate Analysis with observations

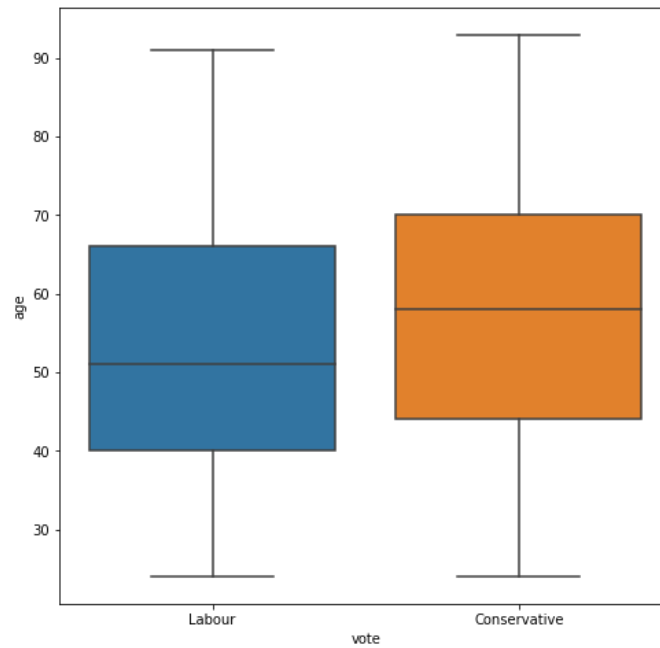


Figure 10 – Boxplot for vote v/s age

Average age of voters for Labour party is 50 while for conservative it is 55.

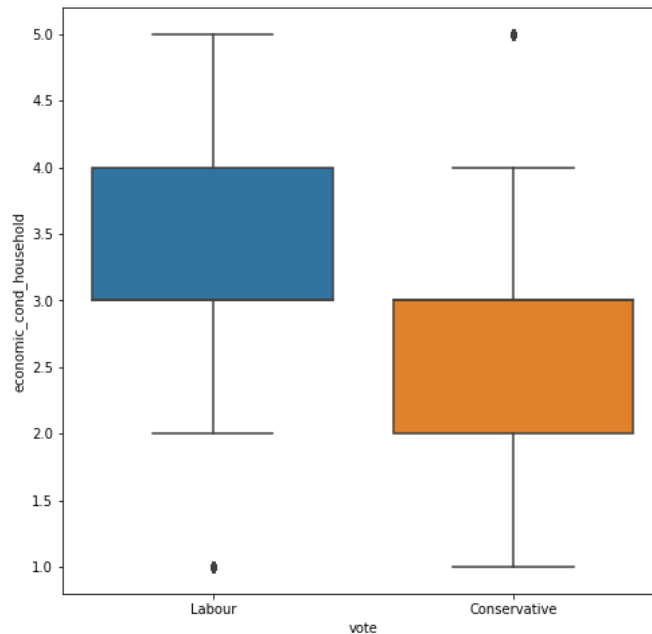


Figure 11 – Boxplot for vote v/s household economic condition

75% percent of voters for Labour party have economic_condition_house value as 4 while for Conservative party is 3.

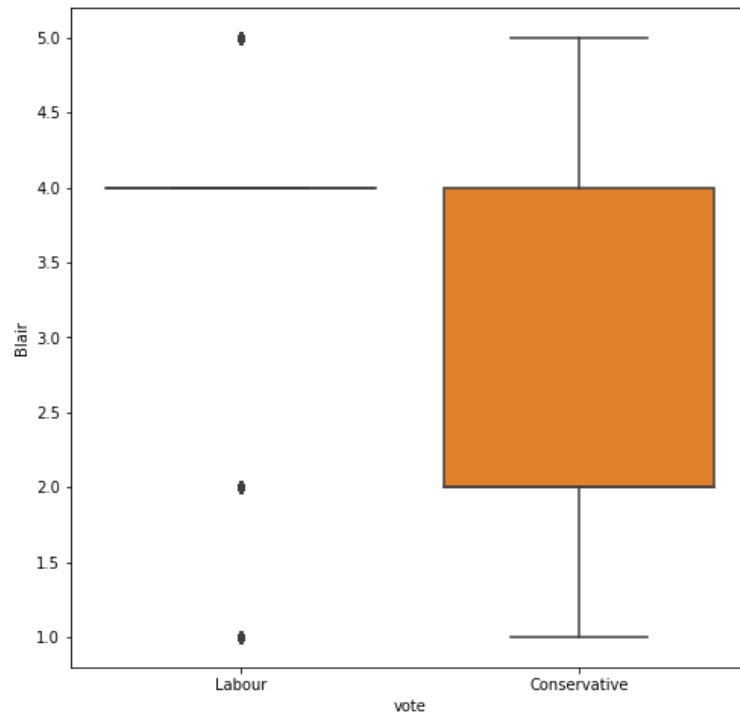


Figure 12 – Boxplot for vote for Blair

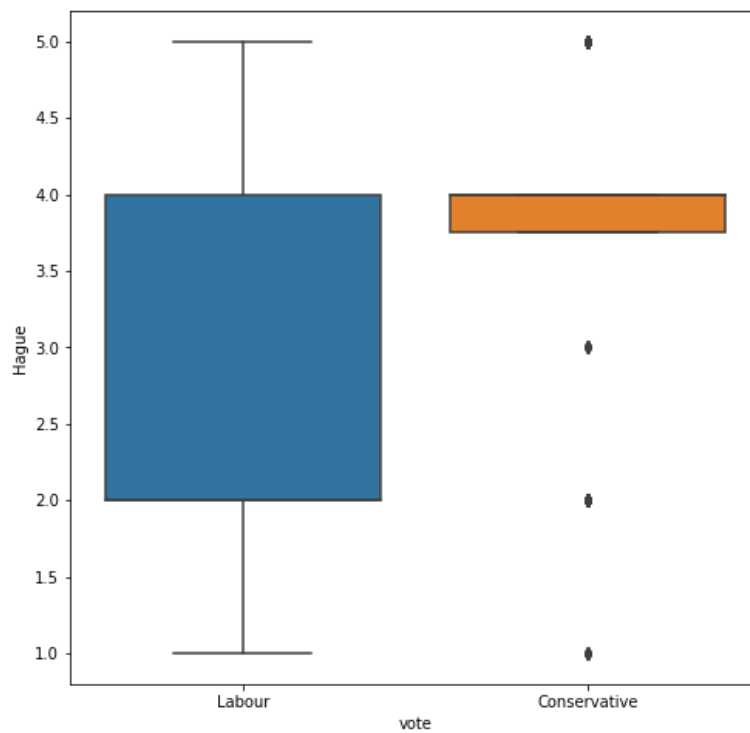


Figure 13 – Boxplot for vote for Hague

Conservative party voters have assessment of Blair ranging from 2 to 4 (25% - 75% range) while 75% Labour party voters have Blair assessment as 4 which is obvious because Blair is from Labour party

Labour party voters have assessment of Hague ranging from 2 to 4 (25% - 75% range) while 75% Conservative party voters have Hague assessment as 4 which is obvious because Hague is from Conservative party

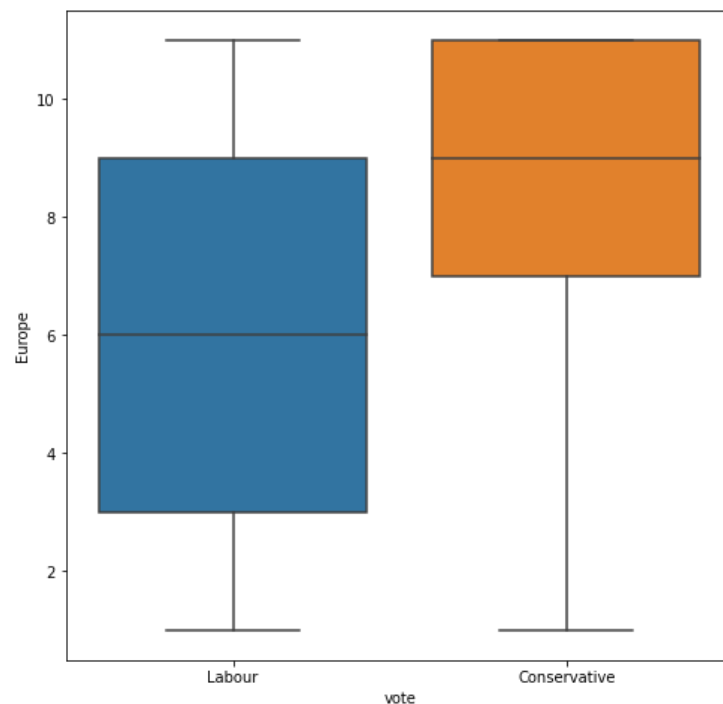


Figure 14 – Boxplot for vote v/s Eurosceptic sentiment

75% voters from Conservative party have Eurosceptic sentiment of scale 10 while for Labour party it is 9.

25% voters from Conservative party have Eurosceptic sentiment of scale 7 while for Labour party it is 3.

Hence voters from Conservative party have high Eurosceptic sentiment

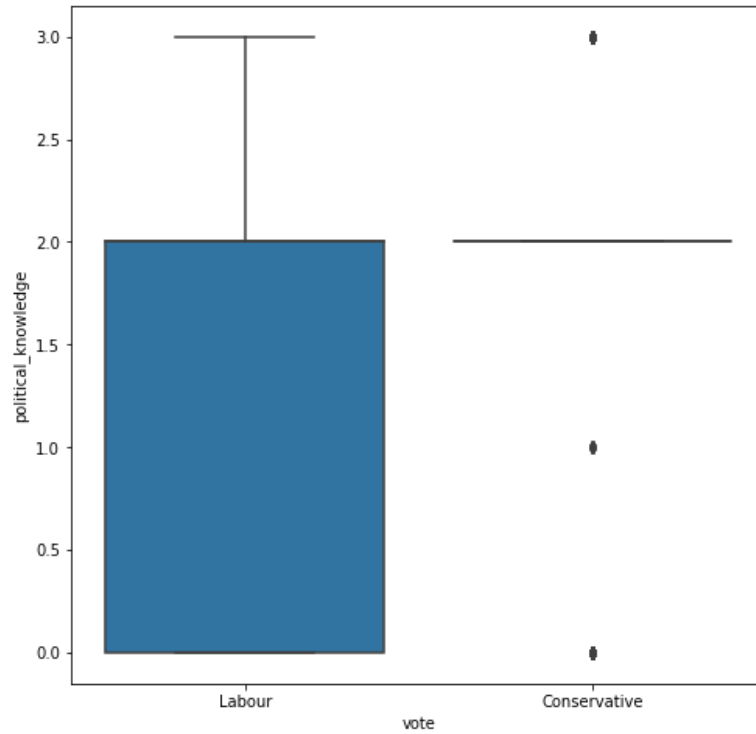


Figure 15 – Boxplot for vote v/s political Knowledge

Conservative party voters have higher political knowledge.

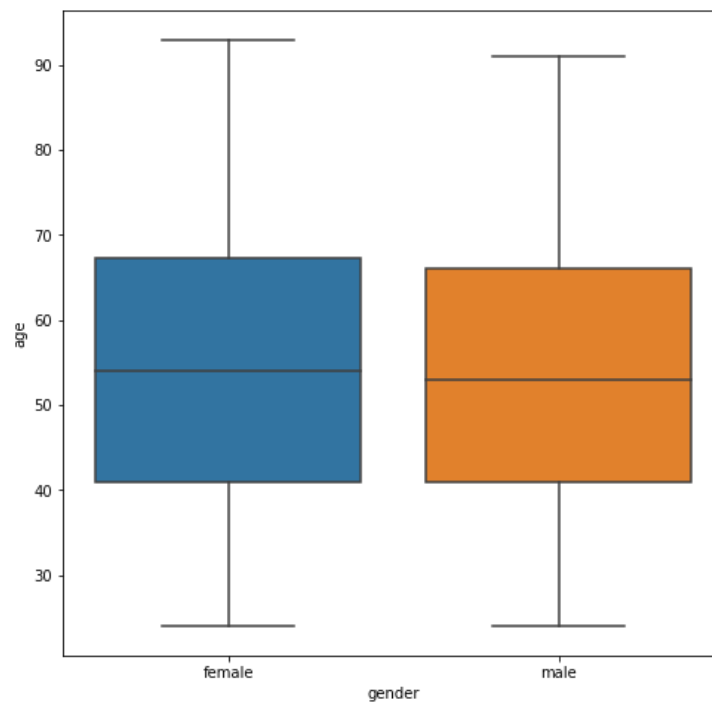


Figure 16 – Boxplot for vote v/s gender

From above we can see that data has almost similar distribution of gender's age.

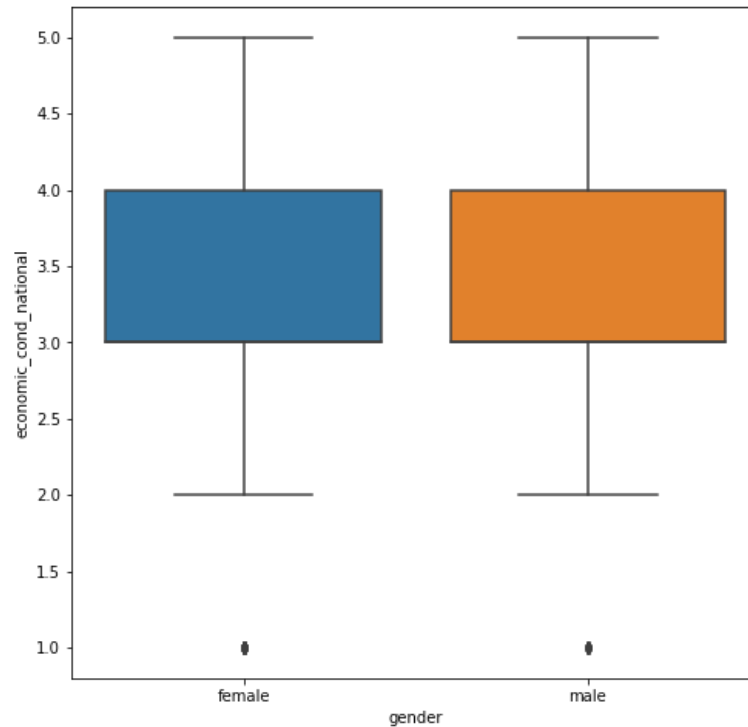


Figure 17 – Boxplot for gender v/s national economic condition

From above we can see that data has almost similar distribution of different gender's economical_cond_national

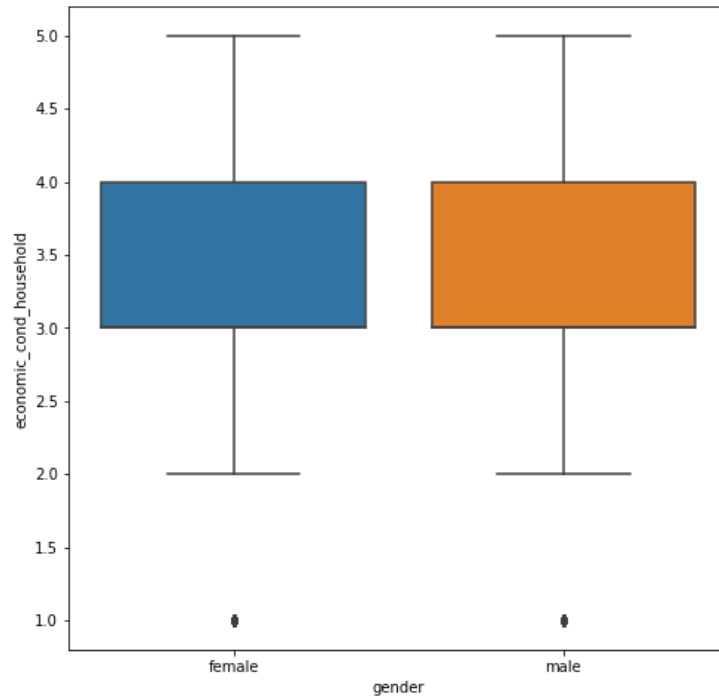


Figure 18 – Boxplot for gender v/s household economic condition

From above we can see that data has almost similar distribution of different gender's economical_cond_household

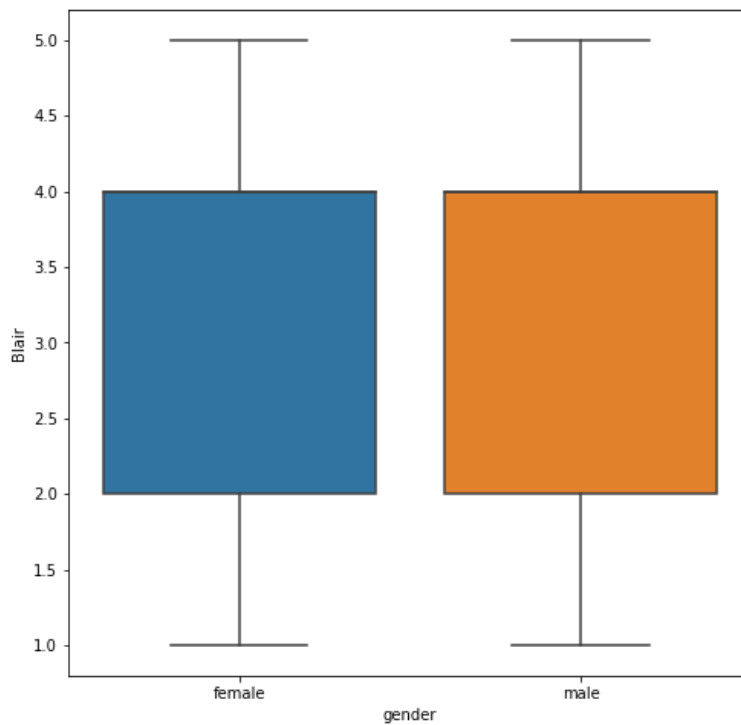


Figure 19 – Boxplot for vote to Blair for among different gender

From above we can see that data has almost similar distribution of different gender for Blair assessment.

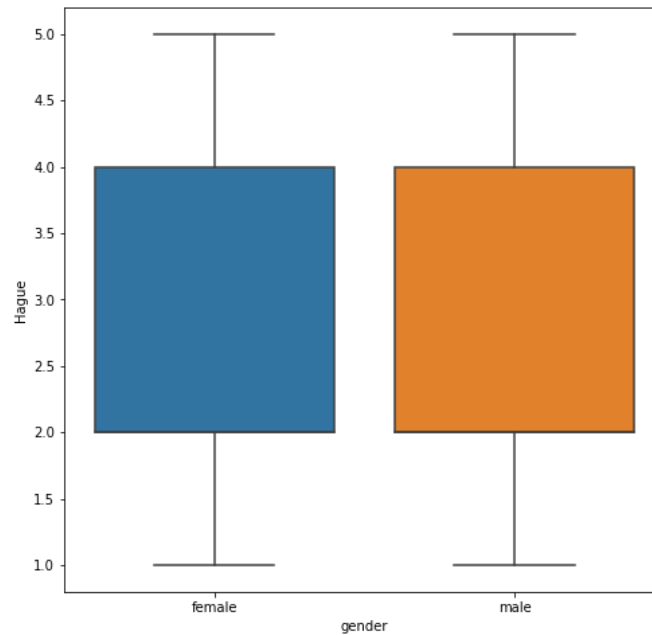


Figure 20 – Boxplot for vote to Hague among different gender

From above we can see that data has almost similar distribution of different gender for Hague assessment.

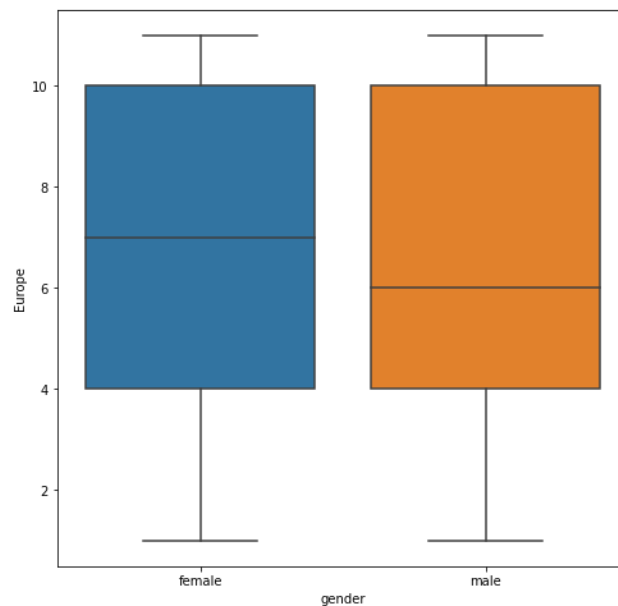


Figure 21 – Boxplot for Eurosceptic sentiment among different gender

From above we can see that data has almost similar distribution of different gender for Eurosceptic sentiment. On an average Female have little higher value of Eurosceptic sentiment

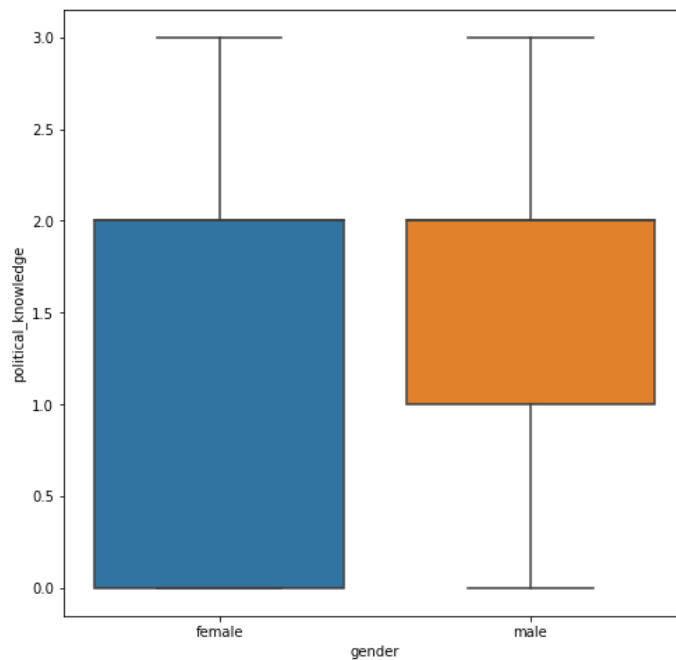


Figure 22 – Boxplot for political knowledge among different gender

From above we can see that on an average male have higher political Knowledge.

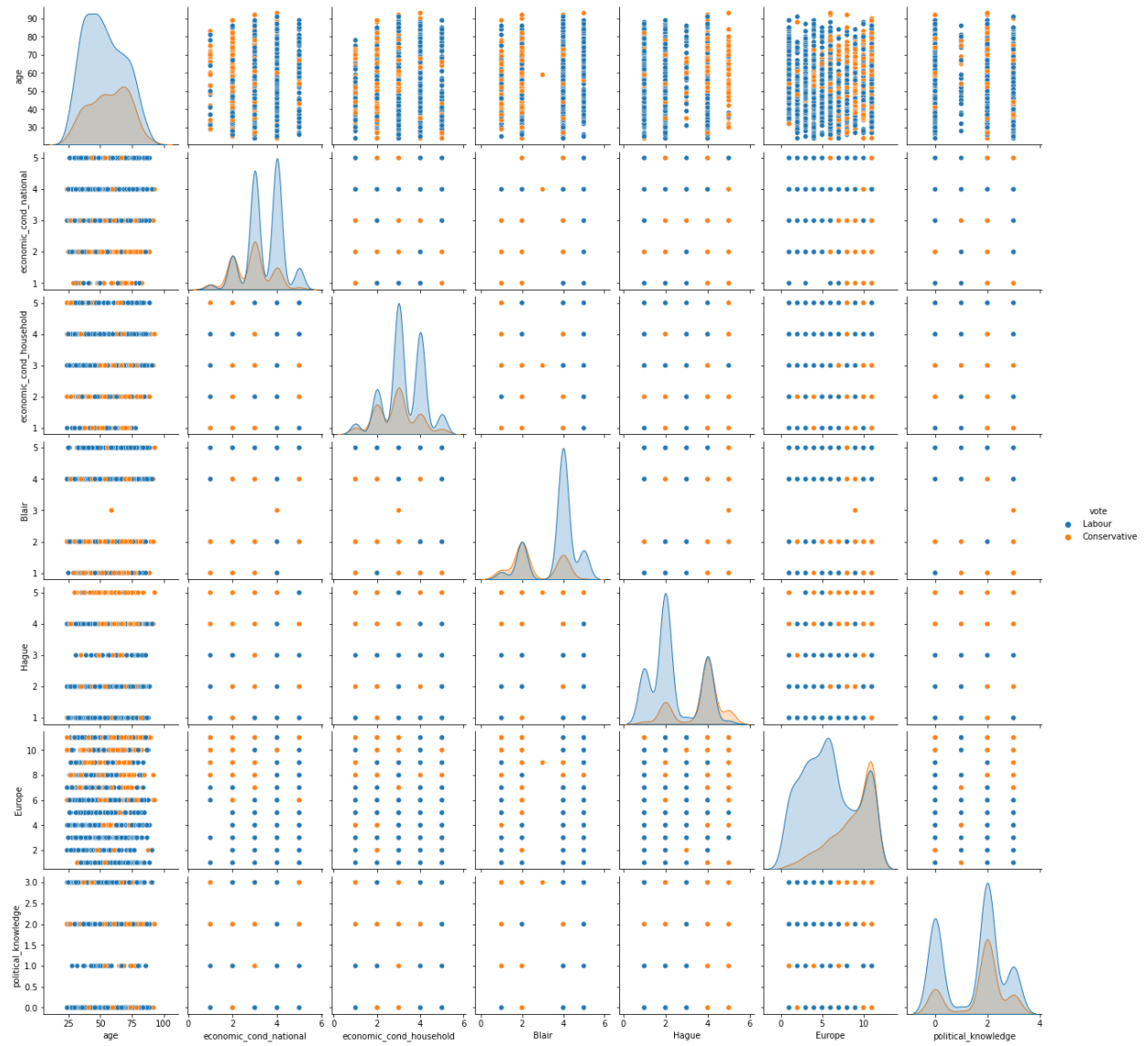


Figure 23 - Pairplot



Figure 24 – Heat Map

Table 12 – Correlation Values

			correlation
economic_cond_national	economic_cond_household		0.347687
		Blair	0.326141
	Blair	Europe	0.295944
	Hague	Europe	0.285738
	Blair	Hague	0.243508
economic_cond_household		Blair	0.215822
economic_cond_national		Europe	0.209150
		Hague	0.200790
	Europe	political_knowledge	0.151197
		economic_cond_household	0.112897
economic_cond_household		Hague	0.100392
	age	Europe	0.064562
political_knowledge		age	0.046598
	age	economic_cond_household	0.038868
economic_cond_household		political_knowledge	0.038528
	age	Blair	0.032084
	Hague	age	0.031144
		political_knowledge	0.029906
	political_knowledge	economic_cond_national	0.023510
		Blair	0.021299
economic_cond_national		age	0.018687

There is weak correlation between different variables

Removing Outliers : We are not removing outliers as they are not present in continuous variables

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).

Data Encoding:

There are three common approaches for converting ordinal and categorical variables to numerical values. They are:

- Ordinal Encoding
- One-Hot Encoding
- Dummy Variable Encoding

Here, we will use Label encoder for target variable vote & Dummy Variable Encoding for gender.

Table 13 – Data set with encoding

	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
0	1	43	3.0	3.0	4	1	2	2	0
1	1	36	4.0	4.0	4	4	5	2	1
2	1	35	4.0	4.0	5	2	3	2	1
3	1	24	4.0	2.0	2	1	4	0	0
4	1	41	2.0	2.0	1	1	6	2	1
...
1520	0	67	5.0	3.0	2	4	11	3	1
1521	0	73	2.0	2.0	4	4	8	2	1
1522	1	37	3.0	3.0	5	4	2	2	1
1523	0	61	3.0	3.0	1	4	11	2	1
1524	0	74	2.0	3.0	2	4	11	0	0

1517 rows × 9 columns

Vote to Labour : 1

Vote to conservative : 0

Gender Male : 1

Gender Female : 0

Table 14 – Data info with categorical variables

```
<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   int64
1   age                                1517 non-null   int64
2   economic_cond_national             1517 non-null   float64
3   economic_cond_household            1517 non-null   float64
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_male                        1517 non-null   uint8
dtypes: float64(2), int64(6), uint8(1)
memory usage: 140.4 KB
```

Now all variables are numeric, and we can now proceed with model building

Scaling:

In general, algorithms that exploit distances or similarities (e.g. in the form of scalar product) between data samples are sensitive to feature transformations i.e. Feature Scaling is performed when we are dealing with Gradient Descent Based algorithms (Linear and Logistic Regression, Neural Network) and Distance-based algorithms (KNN, K-means, SVM) as these are very sensitive to the range of the data points.

The Machine Learning algorithms that require the feature scaling are mostly KNN (K-Nearest Neighbors), Neural Networks, Linear Regression, and Logistic Regression.- The machine learning algorithms that do not require feature scaling is mostly non-linear ML algorithms such as Decision trees, Random Forest, AdaBoost, Naïve Bayes, etc.

Here, we are building a model, to predict which party a voter will vote for on the basis of the given information and to create an exit poll that will help in predicting overall win and seats covered by a particular party. To do our analysis, we are expected to build model using Logistic Regression, LDA, KNN Model and Naïve Bayes Model. For now, we are not scaling the data and will do the scaling based on the models we will run ahead. Hence, as mentioned scaling might be necessary for two models and might not be necessary for the other two

Splitting data into training and test set in 30% test data

X_train (1061, 8)

X_test (456, 8)

y_train (1061,)

y_test (456,)

Total Obs 1517

y_train_value_counts

1 754

0 307

y_test_value_counts

1 303

0 153

1.4 Apply Logistic Regression and LDA (linear discriminant analysis)

Logistic Regression Model without model tuning

The basic model is built with default parameters and the train and test performance results have been discussed

```
classification_report on training data
      precision    recall  f1-score   support

     0       0.75      0.65      0.69       307
     1       0.86      0.91      0.89       754

 accuracy          0.84       1061
 macro avg          0.81      0.78      0.79       1061
 weighted avg       0.83      0.84      0.83       1061

classification_report on test data
      precision    recall  f1-score   support

     0       0.75      0.72      0.73       153
     1       0.86      0.88      0.87       303

 accuracy          0.82       456
 macro avg          0.80      0.80      0.80       456
 weighted avg       0.82      0.82      0.82       456
```

Accuracy on Training data : 84%

Accuracy on Test data : 82%

AUC on Training data: 0.890

AUC on Test data: 0.890

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.89, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class. The model has good fit with no under or over fitting.

Linear Discriminant Analysis without Model Tuning

The basic model is built with default parameters and the train and test performance results have been discussed

```
classification_report on training data
      precision    recall  f1-score   support

    0       0.74      0.65      0.69       307
    1       0.86      0.91      0.89       754

 accuracy
macro avg       0.80      0.78      0.79      1061
weighted avg       0.83      0.83      0.83      1061
```

```
classification_report on test data
      precision    recall  f1-score   support

    0       0.77      0.73      0.74       153
    1       0.86      0.89      0.88       303

 accuracy
macro avg       0.82      0.81      0.81       456
weighted avg       0.83      0.83      0.83       456
```

Accuracy on Training data : 83%

Accuracy on Test data : 83%

AUC on Training data: 0.889

AUC on Test data: 0.888

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.89, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class

The model has good fit with no under or over fitting.

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results

KNN Model

Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. The KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

KNN has the following basic steps:

1. Calculate distance
2. Find closest neighbors
3. Vote for labels

As KNN is Distance-based algorithms and very sensitive to the range of the data points. We are applying z-score scaling to our dataset.

Table 15 – Train Dataset with Scaling

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
991	-1.296710	-1.518921	0.923619	-2.018037	1.029070	1.332089	0.452231	-0.936950
1274	-0.910337	0.892714	-0.196335	0.550300	1.029070	-0.202156	-1.407526	1.067292
649	0.441968	0.892714	-0.196335	0.550300	1.029070	0.104693	0.452231	-0.936950
677	-0.459569	-0.313103	-0.196335	0.550300	-0.593283	1.332089	-1.407526	1.067292
538	-0.652755	2.098531	-0.196335	0.550300	-0.593283	0.411542	-1.407526	1.067292
...
717	-0.137592	-0.313103	-0.196335	0.550300	-1.404459	-0.202156	0.452231	-0.936950
908	-0.717151	-0.313103	0.923619	-1.161925	-0.593283	0.718391	0.452231	-0.936950
1100	1.279109	0.892714	-0.196335	1.406413	1.029070	1.332089	-1.407526	-0.936950
236	-1.489897	-0.313103	-0.196335	-1.161925	0.217894	-0.202156	-1.407526	-0.936950
1065	2.245042	-0.313103	2.043572	0.550300	-0.593283	-1.736401	-1.407526	1.067292

1061 rows × 8 columns

Table 16 – Test Dataset with Scaling

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
504	1.027267	-0.255557	-0.110599	-1.082257	-0.643946	0.316674	0.360390	-0.936239
369	-0.714354	-0.255557	-1.206968	0.602075	-0.643946	0.316674	1.268333	1.068103
1075	2.146881	1.882680	2.082140	1.444241	-0.643946	-1.754684	0.360390	1.068103
1031	-0.465551	-1.324675	-0.110599	-1.082257	0.982865	0.316674	0.360390	-0.936239
1329	-1.336362	1.882680	0.985771	0.602075	0.982865	0.316674	-1.455497	1.068103
...
562	-1.087559	0.813562	-1.206968	0.602075	-0.643946	0.316674	-0.547554	1.068103
928	-0.776555	-1.324675	-1.206968	-1.924423	-0.643946	0.020765	0.360390	-0.936239
276	2.084680	-0.255557	-0.110599	0.602075	-1.457352	-0.275143	-1.455497	-0.936239
1128	-0.092346	0.813562	-0.110599	0.602075	-0.643946	0.908490	-1.455497	-0.936239
1151	-1.149759	-0.255557	-0.110599	0.602075	-1.457352	-0.275143	1.268333	1.068103

456 rows × 8 columns

KNN model without Model Tuning

Default value of neighbors is equal to 5. First we will build KNN Model with k=5

Parameters : default

```

classification_report on training data
              precision    recall  f1-score   support

         0           0.77       0.71       0.74         307
         1           0.88       0.91       0.90         754

    accuracy                  0.85         1061
   macro avg           0.83       0.81       0.82         1061
  weighted avg           0.85       0.85       0.85         1061

```

```

classification_report on test data
              precision    recall  f1-score   support

         0           0.78       0.68       0.72         153
         1           0.85       0.90       0.87         303

    accuracy                  0.83         456
   macro avg           0.81       0.79       0.80         456
  weighted avg           0.82       0.83       0.82         456

```

Accuracy on Training data : 85%

Accuracy on Test data : 83%

AUC on Training data: 0.928

AUC on Test data: 0.928

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.93, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class

The model has good fit with no under or over fitting.

Naïve Bayes model without Model Tuning

For naive bayes algorithm while calculating likelihoods of numerical features it assumes the feature to be normally distributed and then we calculate probability using mean and variance of that feature only and also it assumes that all the predictors are independent to each other. Scale doesn't matter. Performing a feature scaling in this algorithm may not have much effect. We have scaled the data while performing KNN and will be using the same.

```
classification_report on training data
              precision    recall  f1-score   support

     0           0.73       0.66       0.70         307
     1           0.87       0.90       0.88         754

 accuracy              0.83         1061
 macro avg           0.80       0.78       0.79         1061
weighted avg           0.83       0.83       0.83         1061


classification_report on test data
              precision    recall  f1-score   support

     0           0.74       0.73       0.73         153
     1           0.87       0.87       0.87         303

 accuracy              0.82         456
 macro avg           0.80       0.80       0.80         456
weighted avg           0.82       0.82       0.82         456
```

Accuracy on Training data : 83%

Accuracy on Test data : 82%

AUC on Training data: 0.885

AUC on Test data: 0.879

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.88, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class

The model has good fit with no under or over fitting.

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

Logistic Regression Model with model tuning

Applying GridSearchCV for Logistic Regression

Grid search is the process of performing hyper parameter tuning to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified.

Parameters in Grid Search CV:

```
'penalty':['l2','none'],  
'solver':['liblinear','lbfgs', 'sag', 'newton-cg'],  
  tol':[0.0001,0.00001],}  
n_jobs=-1,  
max_iter=10000  
cv = 5  
verbose = 0
```

The parameters used in GridsearchCV can be explained as :

penalty: It imposes a penalty to the logistic model for having too many variables. This results in shrinking the coefficients of the less contributive variables toward zero. l2 is default and we have used none also.

solver: lbfgs by default that decides what solver to use for fitting the model. Other options are 'newton-cg', 'liblinear', 'sag', and 'saga'. Here we are using newton-cg as it adaptively controls the accuracy of the solution without loss of the rapid convergence properties.

Tolerance for stopping criteria : Default is .00001 and we have added .000001

n_jobs: controls the number of cores on which the package will attempt to run in parallel.

max_iter: Defines the maximum number of iterations by the solver during model fitting. Here, we are using 10000.

verbose: non-negative integer (0 by default) that defines the verbosity.

cv: cross validation generator or an iterable, in this case, there is a 5-fold cross-validation.

scoring: choosing scoring F1 since it computes the Harmonic Mean between Recall and Precision, it tells us whether both Type I and Type II error is low or high on an average

```
classification report for training data
      precision    recall  f1-score   support

    0       0.74      0.64      0.69        307
    1       0.86      0.91      0.88        754

 accuracy          0.83        1061
 macro avg          0.80      0.77      0.79        1061
weighted avg          0.83      0.83      0.83        1061
```

```
classification report for test data
      precision    recall  f1-score   support

    0       0.76      0.74      0.75        153
    1       0.87      0.88      0.88        303

 accuracy          0.84        456
 macro avg          0.82      0.81      0.81        456
weighted avg          0.83      0.84      0.83        456
```

Accuracy on Training data : 83%

Accuracy on Test data : 84%

AUC on Training data: 0.890

AUC on Test data: 0.890

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.89, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class

The model has good fit with no under or over fitting

Linear Discriminant Analysis with Model Tuning

Applying GridSearchCV for Linear Discriminant Analysis

Parameters in Grid Search CV:

```
solver': ['lsqr', 'eigen']  
tol':[0.0001,0.001, 0.01],}  
n_jobs=-1,  
cv = 5
```

The parameters used in GridsearchCV can be explained as :

solver: svd by default that decides what solver to use for fitting the model. Svd does not compute the covariance matrix, therefore this solver is recommended for data with a large number of features.

Other options are 'lsqr', 'eigen' which can be combined with shrinkage or custom covariance estimator

Tolerance for stopping criteria : Default is .0001 and we have added .001 & .01.

n_jobs: controls the number of cores on which the package will attempt to run in parallel. -1`` means using all processors

cv: cross validation generator or an iterable, in this case, there is a 5-fold cross-validation.

scoring: choosing scoring F1 since it computes the Harmonic Mean between Recall and Precision, it tells us whether both Type I and Type II error is low or high on an average

classification_report on training data				
	precision	recall	f1-score	support
0	0.74	0.65	0.69	307
1	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

classification_report on test data				
	precision	recall	f1-score	support
0	0.77	0.73	0.74	153
1	0.86	0.89	0.88	303
accuracy			0.83	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456

Accuracy on Training data : 83%

Accuracy on Test data : 83%

AUC on Training data: 0.889

AUC on Test data: 0.888

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.89, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class

The model has good fit with no under or over fitting

KNN with Model Tuning

Default value `n_neighbors=5`

We will perform on KNN with no of neighbours to be 1,3,5..41 and find the optimal number of neighbours from `K=1,3,5,7.....41` using the Mis classification error

Misclassification error (MCE) = 1 - Test accuracy score.

Calculate MCE for each model with neighbours = 1,3,5,7,....41 and find the model with lowest MCE

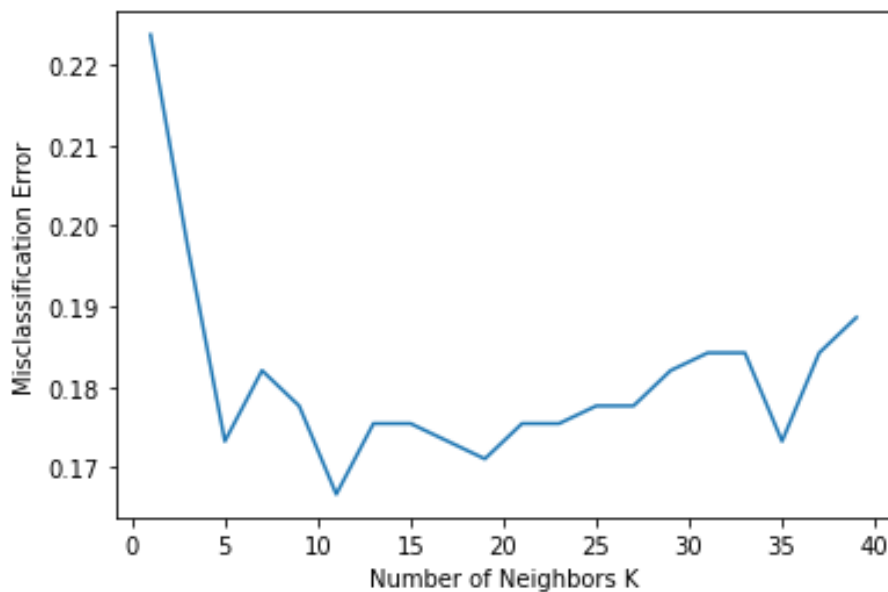


Figure 25 – Misclassification error for different k value

Looking at the above graph we `k =11` gt misclassification error. We will check with grid search CV

Applying GridSearchCV for KNN

```
GridSearchCV
GridSearchCV(cv=3, estimator=KNeighborsClassifier(),
              param_grid={'metric': ['minkowski', 'euclidean', 'canberra'],
                           'n_neighbors': range(5, 20), 'weights': ['uniform']})
  ▸ estimator: KNeighborsClassifier
    ▸ KNeighborsClassifier
```

`N_neighbors`: Number of neighbors to use by default for k-neighbors queries. 11 is used based on leasr misclassification error.

Metrics: default.

Weights: weight function used in prediction.

Possible values:

‘Uniform’: uniform weights. All points in each neighborhood are weighted equally

‘Distance’: weight points by the inverse of their distance and in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away

	precision	recall	f1-score	support
0	0.82	0.23	0.36	307
1	0.76	0.98	0.85	754
accuracy			0.76	1061
macro avg	0.79	0.61	0.61	1061
weighted avg	0.78	0.76	0.71	1061

	precision	recall	f1-score	support
0	0.83	0.23	0.36	153
1	0.71	0.98	0.83	303
accuracy			0.73	456
macro avg	0.77	0.60	0.59	456
weighted avg	0.75	0.73	0.67	456

Accuracy on Training data : 76%

Accuracy on Test data : 76%

AUC on Training data: 0.909

AUC on Test data: 0.889

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.89, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class

The model has good fit with no under or over fitting.

Naïve Bayes with Model Tuning

Applying GridSearchCV for NB

Parameters in Grid Search CV:

```
{'var_smoothing': np.logspace(0,-9, num =100)}
```

The parameters used in GridsearchCV can be explained as :

Explaining the parameters used to find the optimal combinations :

param_grid_NB: Dictionary that contains all of the parameters to try

var_smoothing : Stability calculation to widen (or smooth) the curve and therefore account for more samples that are further away from the distribution mean

.np.logspace : Returns numbers spaced evenly on a log scale, starting from 0, ending at -9, and generating 100 samples

Num: samples, equally spaced on a log scale

	precision	recall	f1-score	support
0	0.73	0.66	0.70	307
1	0.87	0.90	0.88	754
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

	precision	recall	f1-score	support
0	0.76	0.71	0.73	153
1	0.86	0.88	0.87	303
accuracy			0.82	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

Accuracy on Training data : 83%

Accuracy on Test data : 82%

AUC on Training data: 0.885

AUC on Test data: 0.879

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.89, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class

The model has good fit with no under or over fitting

Bagging:

Idea of Bagging:

To fit several independent models and “average” their predictions to obtain a model with a lower variance. However, in practice, it requires too much data to fit fully independent models. So, we rely on the good “approximate properties” of bootstrap samples (representativity and independence) to fit models that are almost independent.

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. When samples are drawn with replacement, then the method is known as Bagging.

Bootstrap Aggregation (or bagging for short) is a simple and very powerful ensemble method. It is a general procedure that can be used to reduce the variance for those algorithms that have high variance. An algorithm that has high variance are decision trees, like classification and regression trees (CART). Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees. Hence, we will build the Bagging model using Decision Tree as the base estimator and then fit the model. We know that Decision trees are sensitive to the specific data on which they are trained. If the training data is changed (e.g., a tree is trained on a subset of the training data) the resulting decision tree can be quite different and in turn the predictions can be quite different.

Bagging using RandomForest

First, we make a random forest model and use it as an estimator in bagging

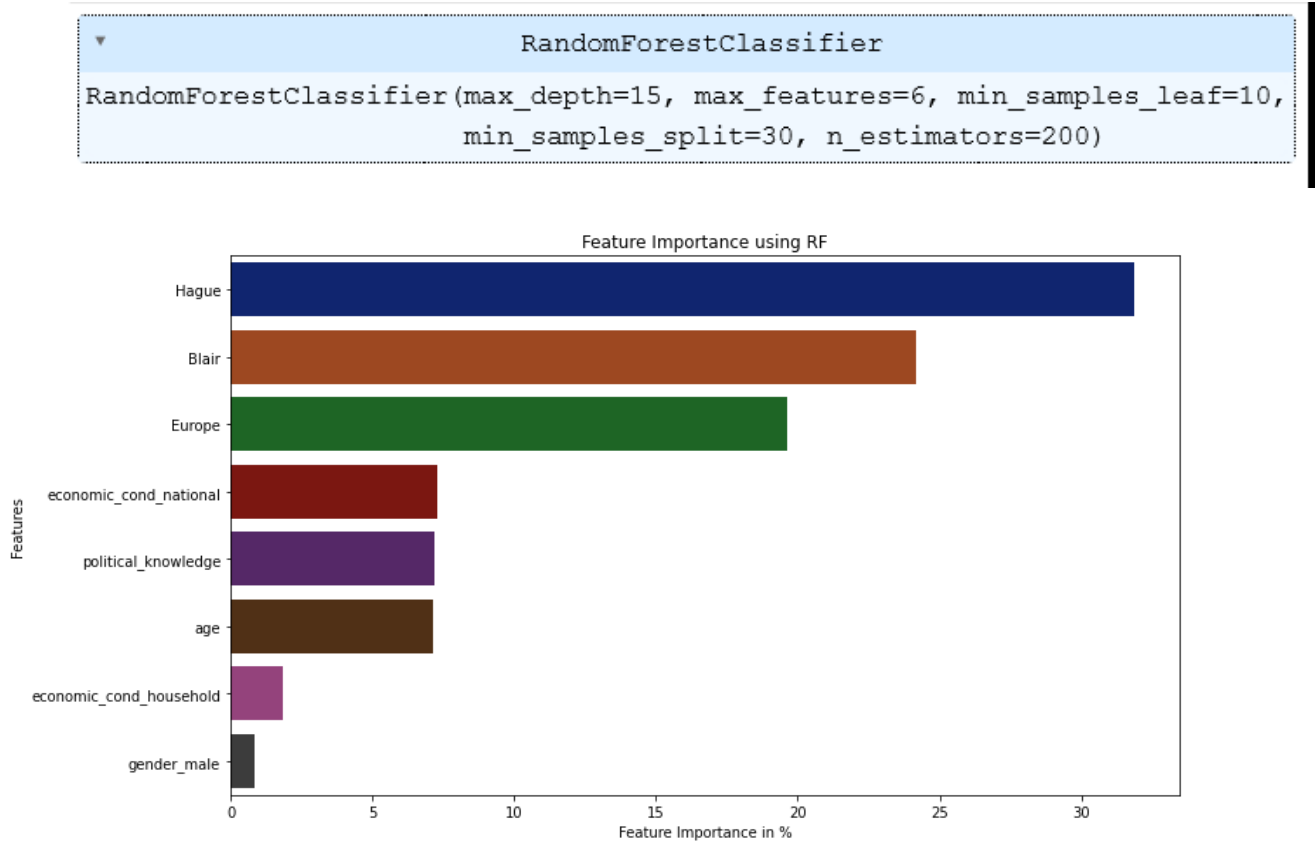


Figure 26 – Feature importance % for different variables

Table 17 – Feature importance values for different variables

	Imp
Hague	0.318666
Blair	0.241779
Europe	0.196180
economic_cond_national	0.073092
political_knowledge	0.072076
age	0.071332
economic_cond_household	0.018493
gender_male	0.008381

From above we can see that gender has least importance in predicting votes so it can be dropped while building model

```

classification report for training data
              precision    recall  f1-score   support

     0       0.73       0.66       0.70       307
     1       0.87       0.90       0.88       754

 accuracy          0.83       1061
 macro avg         0.80       0.78       0.79       1061
 weighted avg      0.83       0.83       0.83       1061

```

```

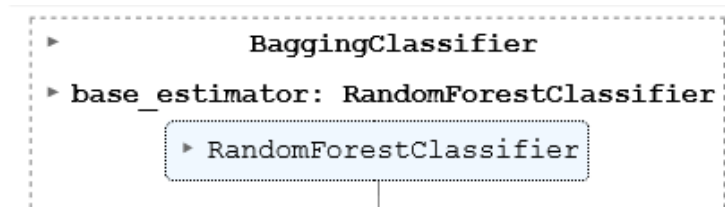
classification report for test data
              precision    recall  f1-score   support

     0       0.76       0.71       0.73       153
     1       0.86       0.88       0.87       303

 accuracy          0.82       456
 macro avg         0.81       0.80       0.80       456
 weighted avg      0.82       0.82       0.82       456

```

We will now check with bagging to see the model performance with Random forest as base estimator



```

classification report for training data
              precision    recall  f1-score   support

     0       0.79       0.69       0.74       307
     1       0.88       0.93       0.90       754

 accuracy          0.86       1061
 macro avg         0.83       0.81       0.82       1061
 weighted avg      0.85       0.86       0.85       1061

```



```

classification report for test data
              precision    recall  f1-score   support

     0       0.78        0.68        0.73        153
     1       0.85        0.90        0.88        303

 accuracy          0.83        456
  macro avg       0.82        0.79        0.80        456
 weighted avg     0.83        0.83        0.83        456

```

Accuracy on Training data : 86%

Accuracy on Test data : 83%

AUC on Training data: 0.879

AUC on Test data: 0.879

We can see from above that with bagging, accuracy, precision, recall and f1 score has increased.

Model Validness:

The model performance score seems to be pretty good in training and testing. The test and train performance is within $\pm 10\%$. We are more interested in recall as it tells us how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

Lastly, the AUC value has a value of 0.89, by observing this we can say that there is a very high chance that the classifier will be able to distinguish the positive class values from the negative class

The model has good fit with no under or over fitting

Boosting:

In Boosting, Base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The idea is to combine several weak models to produce a powerful ensemble. It makes the boosting algorithms prone to overfitting. Examples: AdaBoost, Gradient Boosting.

Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model Boosting is focused on reducing the bias. It makes the boosting algorithms prone to overfitting.

We are using three types of Boosting Algorithms which are as follows:1. AdaBoost (Adaptive Boosting) algorithm.2. Gradient Boosting algorithm.3. XG Boost algorithm.

Below are the key parameters for tuning:

n_estimators: It controls the number of weak learners.

learning_rate:Controls the contribution of weak learners in the final combination. There is a trade-offbetween learning_rate and n_estimators.

We will apply 3 boosting techniques, Ada Boost, Gradient Boosting and XG boosting and analyses the performance.

Ada Boost

```
accuracy 0.8501413760603205
confusion matrix for training data
[[214  93]
 [ 66 688]]
classification_report for training data
              precision    recall  f1-score   support

      0       0.76      0.70      0.73      307
      1       0.88      0.91      0.90      754

   accuracy          0.85      1061
  macro avg          0.82      1061
weighted avg          0.85      1061

accuracy 0.8135964912280702
confusion matrix for test data
[[103  50]
 [ 35 268]]
classification_report for test data
              precision    recall  f1-score   support

      0       0.75      0.67      0.71      153
      1       0.84      0.88      0.86      303

   accuracy          0.81      456
  macro avg          0.79      456
weighted avg          0.81      456
```

Gradient Boosting

```
accuracy 0.8925541941564562
confusion matrix for training data
[[239  68]
 [ 46 708]]
classification_report for training data
              precision    recall  f1-score   support

     0           0.84       0.78       0.81         307
     1           0.91       0.94       0.93         754

 accuracy          0.89          1061
  macro avg         0.88          1061
 weighted avg       0.89          1061


accuracy 0.8355263157894737
confusion matrix for test data
[[105  48]
 [ 27 276]]
classification_report for test data
              precision    recall  f1-score   support

     0           0.80       0.69       0.74         153
     1           0.85       0.91       0.88         303

 accuracy          0.84          456
  macro avg         0.82          456
 weighted avg       0.83          456
```

XGBoost Classifier

```
accuracy 0.9066918001885014
confusion matrix for training data
[[248  59]
 [ 40 714]]
classification_report for training data
              precision    recall  f1-score   support

     0           0.86       0.81       0.83         307
     1           0.92       0.95       0.94         754

 accuracy          0.91          1061
  macro avg         0.89          1061
 weighted avg       0.91          1061
```

```

accuracy 0.8289473684210527
confusion matrix for test data
[[106  47]
 [ 31 272]]
classification_report for test data
      precision    recall  f1-score   support

     0       0.77      0.69      0.73       153
     1       0.85      0.90      0.87       303

 accuracy
macro avg      0.81      0.80      0.80       456
weighted avg    0.83      0.83      0.83       456

```

Gradient Boosting:

Accuracy on Training data : 89%

Accuracy on Test data : 84%

AUC on Training data: 0.879

AUC on Test data: 0.879

XG Boosting:

Accuracy on Training data : 91%

Accuracy on Test data : 83%

AUC on Training data: 0.879

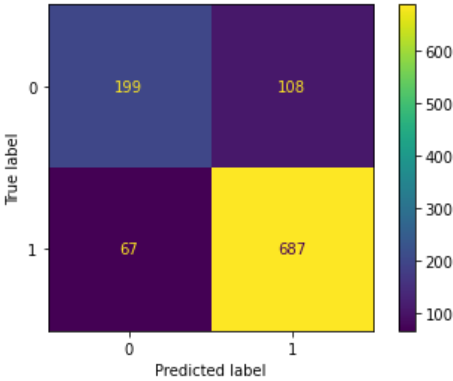
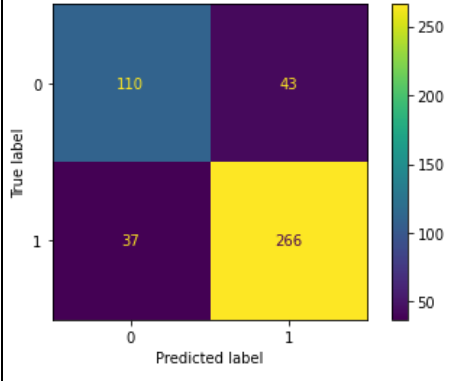
AUC on Test data: 0.879

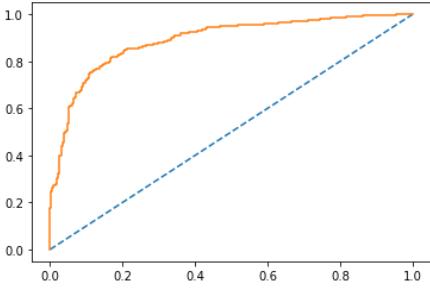
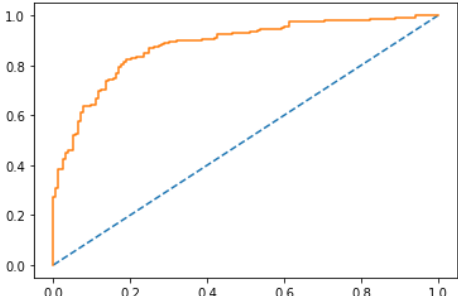
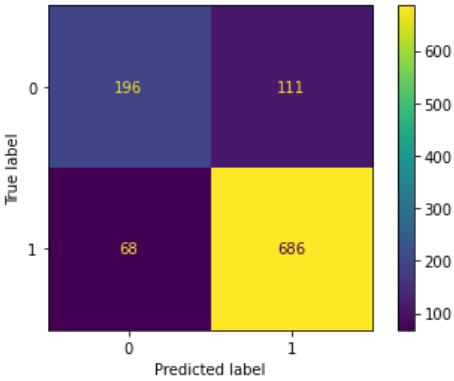
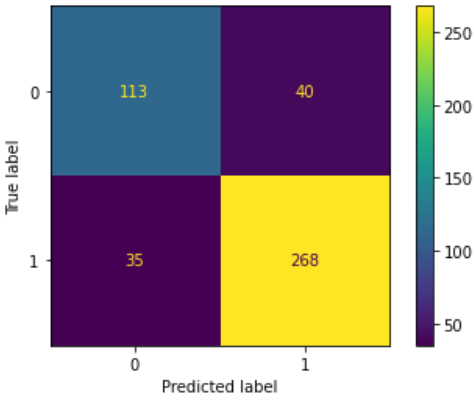
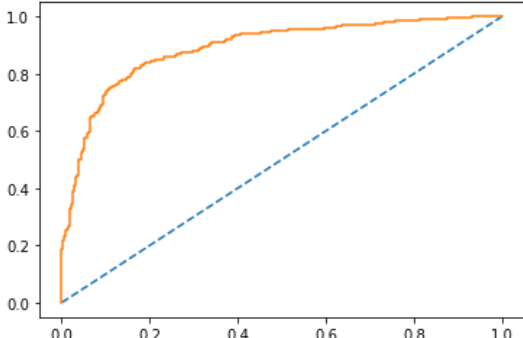
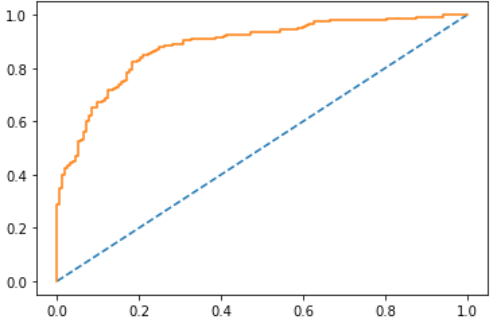
From above, we can see that even the XG boost classifier is giving higher values of recall, accuracy etc. with respect to other models but the looking at recall for class 0 for train and test data, model is overfit. Gradient boosting is best fit model for both class 0 and 1.

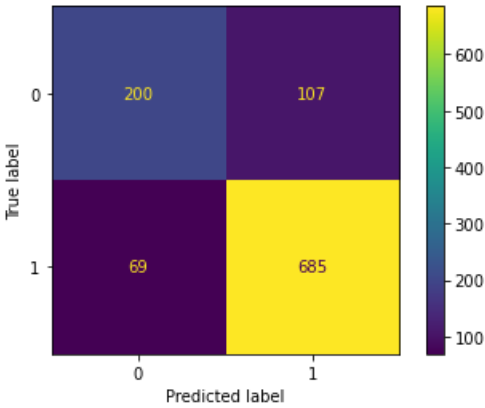
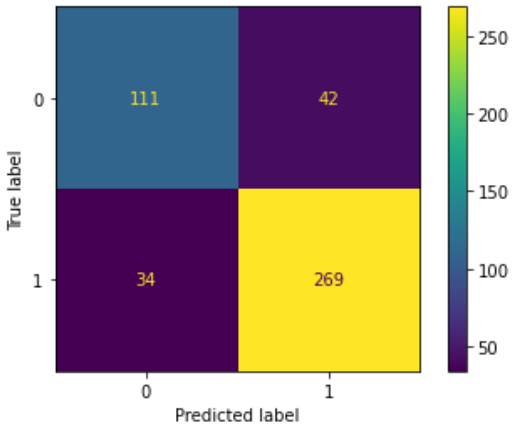
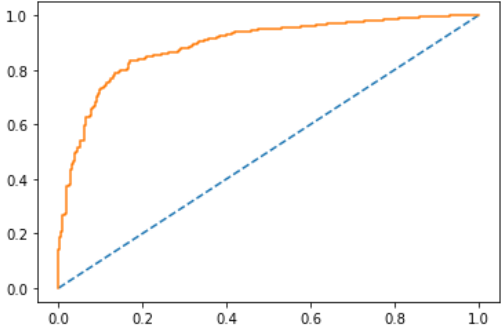
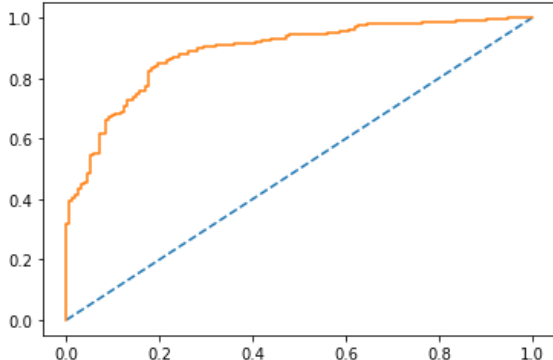
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

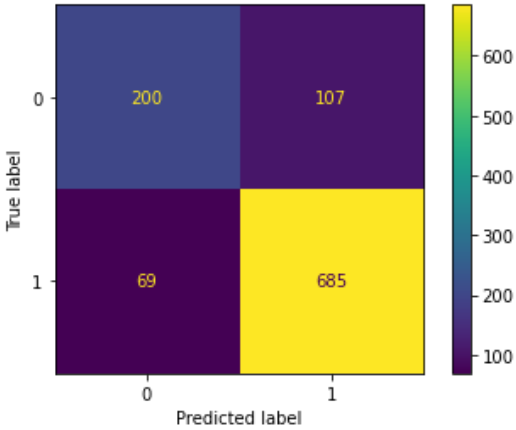
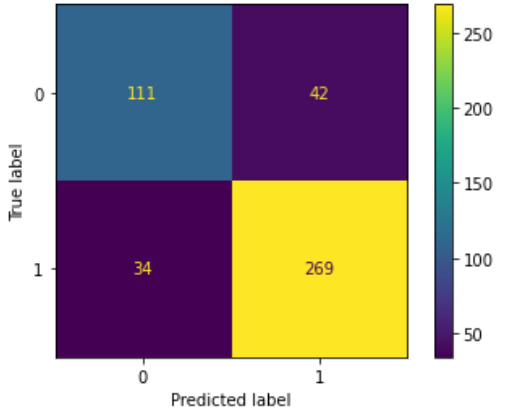
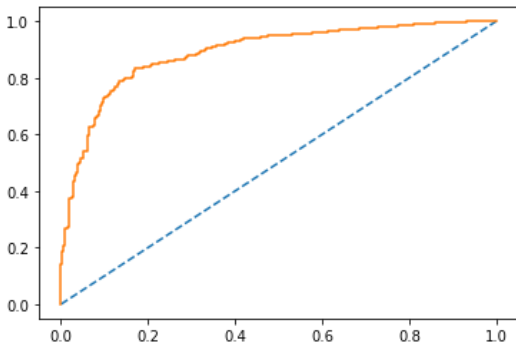
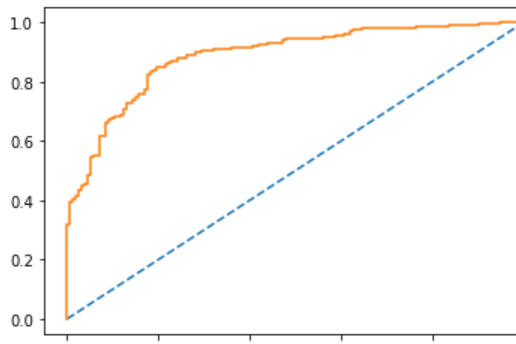
Since we are building a model to predict if whether employee will opt for tour or not, for practical purposes, we will be more interested in correctly classifying 1 (employees opting for tour) than 0(employees not opting for tour).Below is comparison of all models.

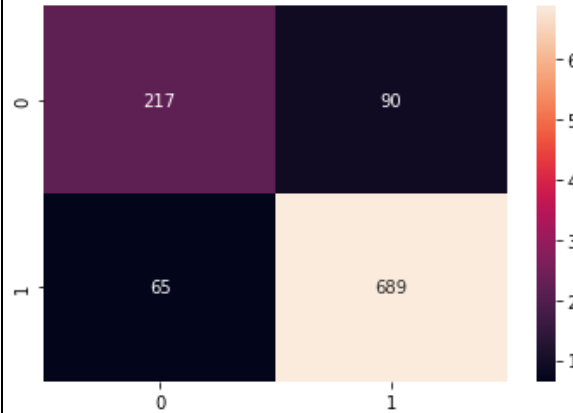
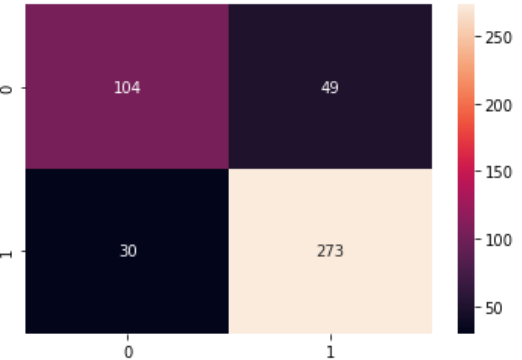
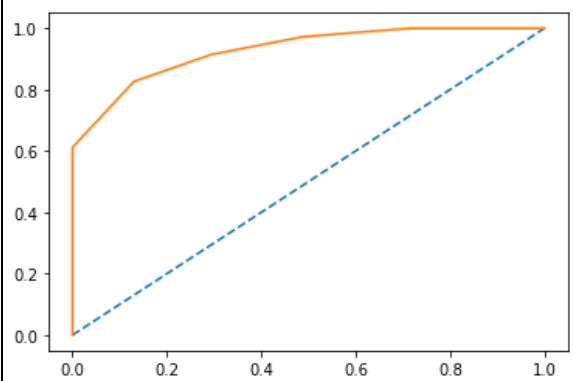
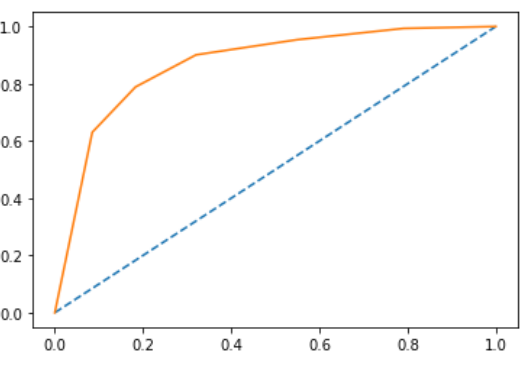
Table 18 – Comparison of all model

Logistic Regression (Basic model)					
Performance Indicator	Train data			Test data	
Accuracy	0.83			0.82	
Confusion matrix					
Classification report	<pre> classification_report on training data precision recall f1-score support 0 0.75 0.65 0.69 307 1 0.86 0.91 0.89 754 accuracy 0.83 macro avg 0.81 weighted avg 0.83 </pre>			<pre> classification_report on test data precision recall f1-score support 0 0.75 0.72 0.73 153 1 0.86 0.88 0.87 303 accuracy 0.82 macro avg 0.80 weighted avg 0.82 </pre>	

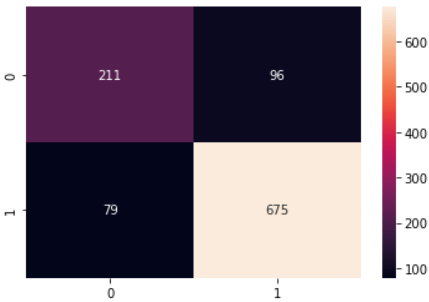
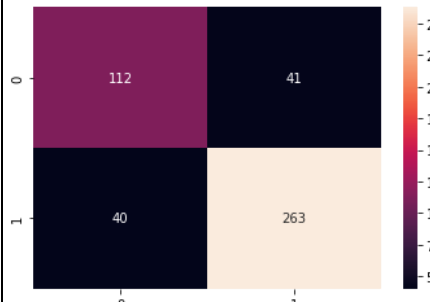
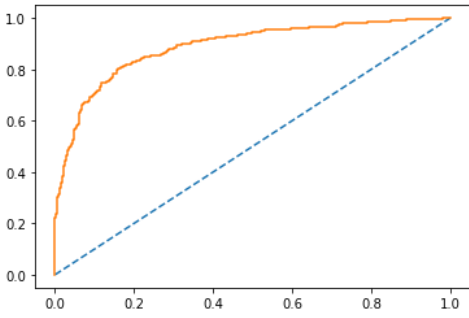
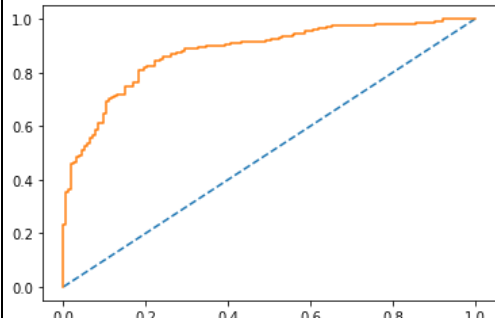
ROC curve		
ROC_AUC score	0.890	0.890
Logistic Regression (Tuned model)		
Accuracy	0.83	0.83
Confusion matrix		
Classification report	<pre> classification report for training data precision recall f1-score support 0 0.74 0.64 0.69 307 1 0.86 0.91 0.88 754 accuracy 0.83 1061 macro avg 0.80 0.77 0.79 1061 weighted avg 0.83 0.83 0.83 1061 </pre>	<pre> classification report for test data precision recall f1-score support 0 0.76 0.74 0.75 153 1 0.87 0.88 0.88 303 accuracy 0.84 456 macro avg 0.82 0.81 0.81 456 weighted avg 0.83 0.84 0.83 456 </pre>
ROC curve		

ROC_AUC score	0.890	0.890
Comparison between base and tuned model of Logistic Regression: In terms of performance of basic and tuned models, there is improvement in the performance without overfitting or underfitting. Hence, we can safely choose the tuned model for further comparison between other Machine Learning models.		
Linear Discriminant analysis (Basic model)		
Accuracy	0.83	0.83
Confusion matrix		
Classification report	<pre> classification_report on training data precision recall f1-score support 0 0.74 0.65 0.69 307 1 0.86 0.91 0.89 754 accuracy 0.83 1061 macro avg 0.80 0.78 0.79 1061 weighted avg 0.83 0.83 0.83 1061 </pre>	<pre> classification_report on test data precision recall f1-score support 0 0.77 0.73 0.74 153 1 0.86 0.89 0.88 303 accuracy 0.83 456 macro avg 0.82 0.81 0.81 456 weighted avg 0.83 0.83 0.83 456 </pre>
ROC curve		
ROC_AUC score	0.889	0.888

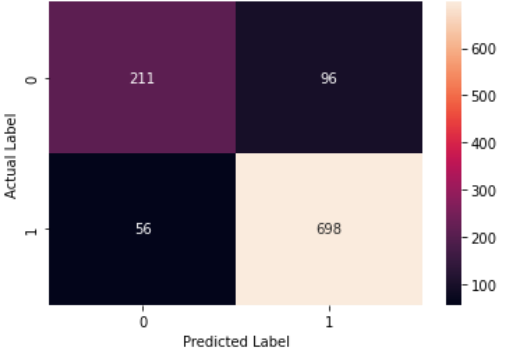
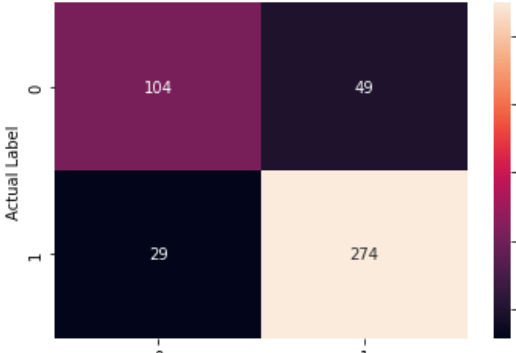
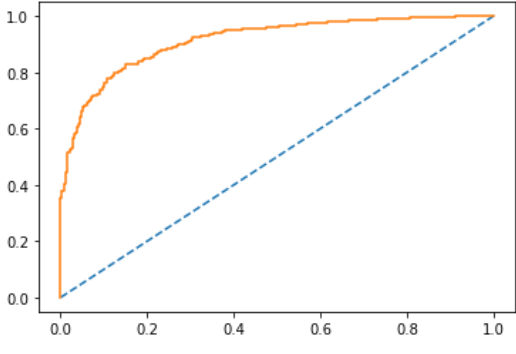
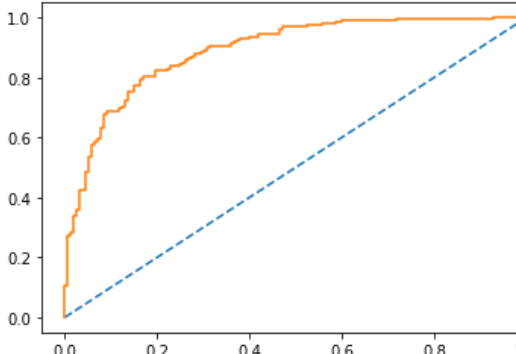
Linear Discriminant analysis (Tuned model)		
Accuracy	0.83	0.83
Confusion matrix		
Classification report	<pre> classification_report on training data precision recall f1-score support 0 0.74 0.65 0.69 307 1 0.86 0.91 0.89 754 accuracy 0.83 1061 macro avg 0.80 0.78 0.79 1061 weighted avg 0.83 0.83 0.83 1061 </pre>	<pre> classification_report on test data precision recall f1-score support 0 0.77 0.73 0.74 153 1 0.86 0.89 0.88 303 accuracy 0.83 456 macro avg 0.82 0.81 0.81 456 weighted avg 0.83 0.83 0.83 456 </pre>
ROC curve		
ROC_AUC score	0.889	0.888
Comparison between base and tuned model of Linear Discriminant analysis: In terms of performance of basic and tuned models, there is no change in the performance without overfitting or underfitting. Hence, we can safely choose the tuned model for further comparison between other Machine Learning models.		

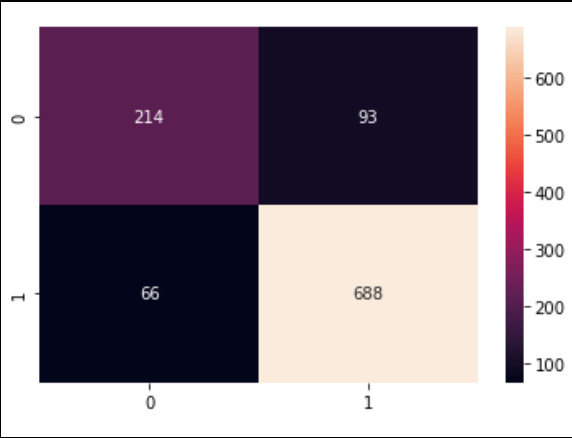
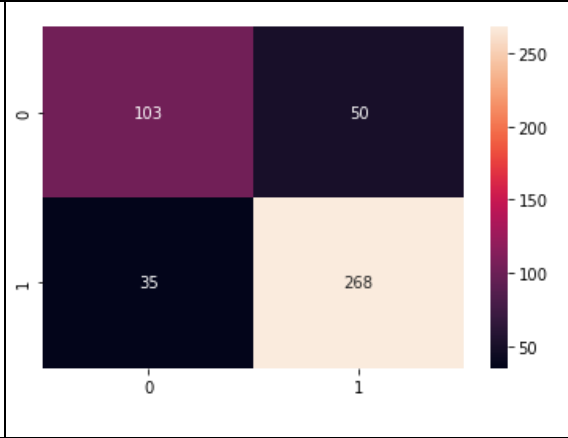
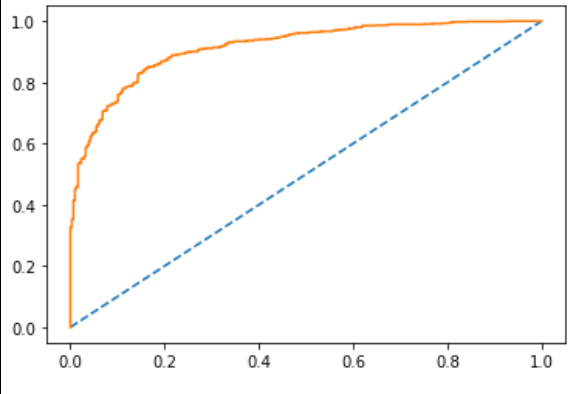
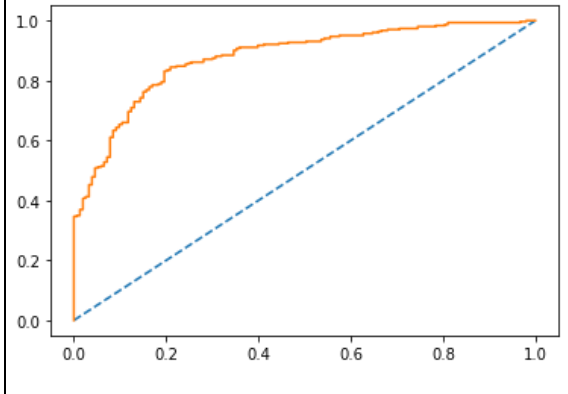
KNN (Basic model)																				
Accuracy	0.85	0.83																		
Confusion matrix	 <p>Confusion matrix for training data:</p> <table border="1"> <thead> <tr> <th></th><th>Actual 0</th><th>Actual 1</th></tr> </thead> <tbody> <tr> <th>Predicted 0</th><td>217</td><td>90</td></tr> <tr> <th>Predicted 1</th><td>65</td><td>689</td></tr> </tbody> </table>		Actual 0	Actual 1	Predicted 0	217	90	Predicted 1	65	689	 <p>Confusion matrix for test data:</p> <table border="1"> <thead> <tr> <th></th><th>Actual 0</th><th>Actual 1</th></tr> </thead> <tbody> <tr> <th>Predicted 0</th><td>104</td><td>49</td></tr> <tr> <th>Predicted 1</th><td>30</td><td>273</td></tr> </tbody> </table>		Actual 0	Actual 1	Predicted 0	104	49	Predicted 1	30	273
	Actual 0	Actual 1																		
Predicted 0	217	90																		
Predicted 1	65	689																		
	Actual 0	Actual 1																		
Predicted 0	104	49																		
Predicted 1	30	273																		
Classification report	<pre> classification_report on training data precision recall f1-score support 0 0.77 0.71 0.74 307 1 0.88 0.91 0.90 754 accuracy 0.85 1061 macro avg 0.83 1061 weighted avg 0.85 1061 </pre>	<pre> classification_report on test data precision recall f1-score support 0 0.78 0.68 0.72 153 1 0.85 0.90 0.87 303 accuracy 0.83 456 macro avg 0.81 456 weighted avg 0.82 456 </pre>																		
ROC curve																				
ROC_AUC score	0.928	0.928																		
KNN (Tuned model)																				
Accuracy	0.85	0.82																		

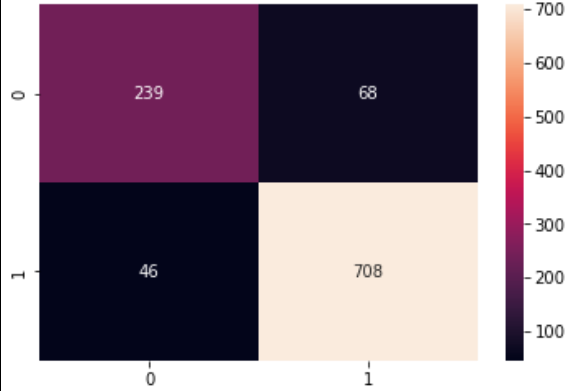
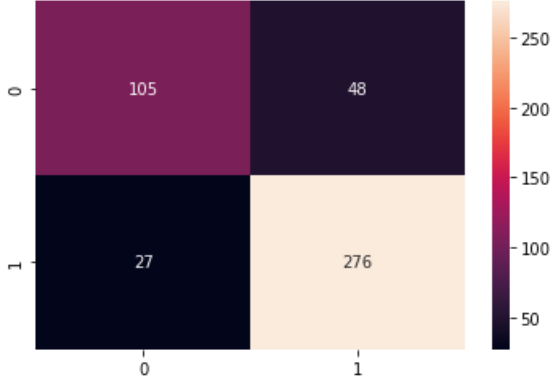
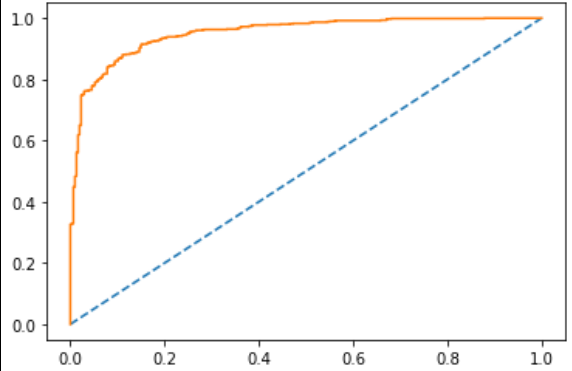
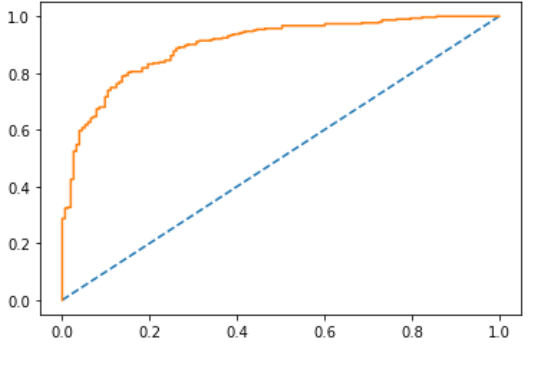
Confusion matrix		
Classification report	<pre> precision recall f1-score support 0 0.82 0.23 0.36 307 1 0.76 0.98 0.85 754 accuracy 0.76 1061 macro avg 0.79 1061 weighted avg 0.78 1061 </pre>	<pre> precision recall f1-score support 0 0.83 0.23 0.36 153 1 0.71 0.98 0.83 303 accuracy 0.73 456 macro avg 0.77 456 weighted avg 0.75 456 </pre>
ROC curve		
ROC_AUC score	0.906	0.876
<p>Comparison between base and tuned model of KNN:</p> <p>Accuracy: The basic model performs better compared to that of basic model.</p> <p>Recall: Recall of the tuned model for class 0 is very poor (23%).</p> <p>Precision: It has decreased in tuned model.</p> <p>AUC: The AUC test score is much better at 87.6% in regularized model as compared to basic test score.</p> <p>F1-score: The f1-score increased/dropped in tuned model phase as compared to basic test performance.</p> <p>Hence, choosing the basic model would be a wise decision</p>		

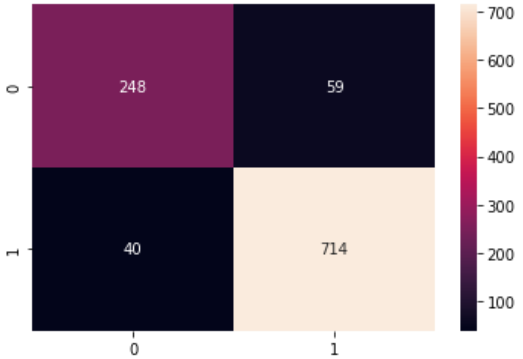
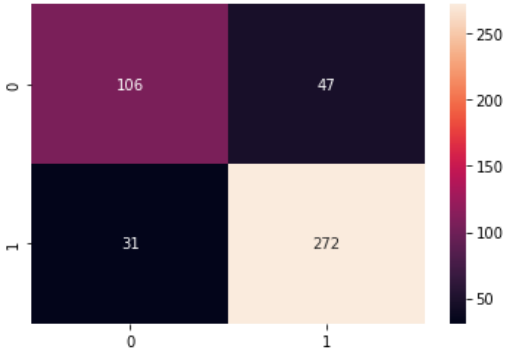
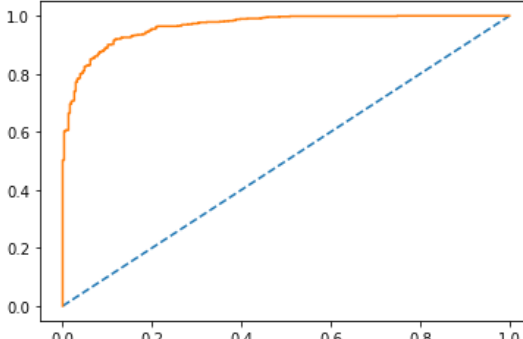
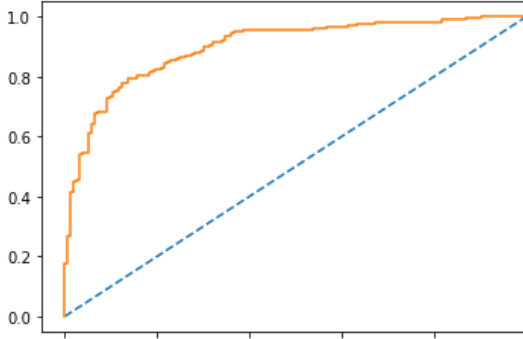
Naïve Bayes (Basic model)		
Accuracy	0.84	0.82
Confusion matrix		
Classification report	<pre> classification_report on training data precision recall f1-score support 0 0.73 0.69 0.71 307 1 0.88 0.90 0.89 754 accuracy 0.84 1061 macro avg 0.80 1061 weighted avg 0.83 1061 </pre>	<pre> classification_report on test data precision recall f1-score support 0 0.74 0.73 0.73 153 1 0.87 0.87 0.87 303 accuracy 0.82 456 macro avg 0.80 456 weighted avg 0.82 456 </pre>
ROC curve		
ROC_AUC score	0.888	0.876
Naïve Bayes (Tuned model)		
Accuracy	0.83	0.82

Confusion matrix	<div>Figure 20.1: Confusion Matrix of NB Model With GridSearch-Train Data</div>	<div>Figure 20.2: Confusion Matrix of NB Model With GridSearch-Test Data</div>																																																												
Classification report	<table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>0</td><td>0.73</td><td>0.66</td><td>0.70</td><td>307</td></tr><tr><td>1</td><td>0.87</td><td>0.90</td><td>0.88</td><td>754</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.83</td><td>1061</td></tr><tr><td>macro avg</td><td>0.80</td><td>0.78</td><td>0.79</td><td>1061</td></tr><tr><td>weighted avg</td><td>0.83</td><td>0.83</td><td>0.83</td><td>1061</td></tr></tbody></table>		precision	recall	f1-score	support	0	0.73	0.66	0.70	307	1	0.87	0.90	0.88	754	accuracy			0.83	1061	macro avg	0.80	0.78	0.79	1061	weighted avg	0.83	0.83	0.83	1061	<table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>0</td><td>0.76</td><td>0.71</td><td>0.73</td><td>153</td></tr><tr><td>1</td><td>0.86</td><td>0.88</td><td>0.87</td><td>303</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.82</td><td>456</td></tr><tr><td>macro avg</td><td>0.81</td><td>0.80</td><td>0.80</td><td>456</td></tr><tr><td>weighted avg</td><td>0.82</td><td>0.82</td><td>0.82</td><td>456</td></tr></tbody></table>		precision	recall	f1-score	support	0	0.76	0.71	0.73	153	1	0.86	0.88	0.87	303	accuracy			0.82	456	macro avg	0.81	0.80	0.80	456	weighted avg	0.82	0.82	0.82	456
	precision	recall	f1-score	support																																																										
0	0.73	0.66	0.70	307																																																										
1	0.87	0.90	0.88	754																																																										
accuracy			0.83	1061																																																										
macro avg	0.80	0.78	0.79	1061																																																										
weighted avg	0.83	0.83	0.83	1061																																																										
	precision	recall	f1-score	support																																																										
0	0.76	0.71	0.73	153																																																										
1	0.86	0.88	0.87	303																																																										
accuracy			0.82	456																																																										
macro avg	0.81	0.80	0.80	456																																																										
weighted avg	0.82	0.82	0.82	456																																																										
ROC curve	<div>Figure 21.1: AUC-ROC Train Data-NB Model with GridSearch</div>	<div>Figure 21.2: AUC-ROC Test Data-NB Model with GridSearch</div>																																																												
ROC_AUC score	0.885	0.879																																																												
<div>Comparison between base and tuned model of Naïve Bayes:</div> <div>In terms of performance of basic and tuned models, there is very little improvement in the performance without overfitting or underfitting. Hence, we can safely choose the tuned model for further comparison between other Machine Learning models.</div>																																																														

Bagging classifier model with Random Forest as base estimator		
Accuracy	0.86	0.83
Confusion matrix		
Classification report	<pre> classification report for training data precision recall f1-score support 0 0.79 0.69 0.74 307 1 0.88 0.93 0.90 754 accuracy 0.86 1061 macro avg 0.83 1061 weighted avg 0.85 1061 </pre>	<pre> classification report for test data precision recall f1-score support 0 0.78 0.68 0.73 153 1 0.85 0.90 0.88 303 accuracy 0.83 456 macro avg 0.82 456 weighted avg 0.83 456 </pre>
ROC curve		
ROC_AUC score	0.879	0.879
Ada Boost model		
Accuracy	0.85	0.81

Confusion matrix	 <p>Heatmap showing confusion matrix for training data. The x-axis represents actual classes (0, 1) and the y-axis represents predicted classes (0, 1). The color scale ranges from 100 (dark purple) to 600 (light orange). The values are: True Positive (TP) = 214, True Negative (TN) = 93, False Positive (FP) = 66, False Negative (FN) = 688.</p>	 <p>Heatmap showing confusion matrix for test data. The x-axis represents actual classes (0, 1) and the y-axis represents predicted classes (0, 1). The color scale ranges from 50 (dark purple) to 250 (light orange). The values are: True Positive (TP) = 103, True Negative (TN) = 50, False Positive (FP) = 35, False Negative (FN) = 268.</p>
Classification report	<pre>confusion matrix for training data classification_report for training data precision recall f1-score support 0 0.76 0.70 0.73 307 1 0.88 0.91 0.90 754 accuracy 0.85 macro avg 0.82 weighted avg 0.85</pre>	<pre>confusion matrix for test data classification_report for test data precision recall f1-score support 0 0.75 0.67 0.71 153 1 0.84 0.88 0.86 303 accuracy 0.81 macro avg 0.79 weighted avg 0.81</pre>
ROC curve	 <p>ROC curve for training data. The x-axis is False Positive Rate (FPR) and the y-axis is True Positive Rate (TPR), both ranging from 0.0 to 1.0. A solid orange line represents the model's performance, and a dashed blue line represents the random baseline. The area under the curve (AUC) is 0.879.</p>	 <p>ROC curve for test data. The x-axis is False Positive Rate (FPR) and the y-axis is True Positive Rate (TPR), both ranging from 0.0 to 1.0. A solid orange line represents the model's performance, and a dashed blue line represents the random baseline. The area under the curve (AUC) is 0.879.</p>
ROC_AUC score	0.879	0.879
Gradient Boosting model		
Accuracy	0.89	0.84

Confusion matrix		
Classification report	<pre>classification_report for traning data precision recall f1-score support 0 0.84 0.78 0.81 307 1 0.91 0.94 0.93 754 accuracy 0.89 0.89 0.89 1061 macro avg 0.88 0.86 0.87 1061 weighted avg 0.89 0.89 0.89 1061</pre>	<pre>classification_report for test data precision recall f1-score support 0 0.80 0.69 0.74 153 1 0.85 0.91 0.88 303 accuracy 0.84 0.84 0.84 456 macro avg 0.82 0.80 0.81 456 weighted avg 0.83 0.84 0.83 456</pre>
ROC curve		
ROC_AUC score	0.879	0.879
XG Boosting model		
Accuracy	0.91	0.83

Confusion matrix		
Classification report	<pre> classification_report for training data precision recall f1-score support 0 0.86 0.81 0.83 307 1 0.92 0.95 0.94 754 accuracy 0.91 0.91 0.91 1061 macro avg 0.89 0.88 0.88 1061 weighted avg 0.91 0.91 0.91 1061 </pre>	<pre> classification_report for test data precision recall f1-score support 0 0.77 0.69 0.73 153 1 0.85 0.90 0.87 303 accuracy 0.83 0.83 0.83 456 macro avg 0.81 0.80 0.80 456 weighted avg 0.83 0.83 0.83 456 </pre>
ROC curve		
ROC_AUC score	0.879	0.879

Among all the models, gradient boosting has shown best performance. XG boosting is not selected here although train and test accuracies are good because if we see recall for class 1, the difference between train and test is well in 10% and hence there is overfitting involved.

1.8 Based on these predictions, what are the insights?

The business issue essentially spun around fostering a model to anticipate which party a citizen would vote in favor of depending on the data about the citizens. The model will in this way be utilized to make an exit poll that will help in predicting the overall win and seats covered by a specific party.

For this to achieve, the analyses assumed CNBE wish to focus more on accurately predicting the Labor's win and hence that has been the class of choice for prediction. The analysis and building of Machine Learning models based on a restricted dataset of 1525 citizens with specific details of the electors. This notwithstanding, regardless of limitations, has assisted us with finding not many key bits of knowledge and patterns alongside exhibiting the ideal model which could be used by CNBE to anticipate the previously mentioned.

Insights:

- 1) Majority of the voters are between the ages 33 – 75 and there are no voters' data capture between the age 18 to 24.
- 2) Majority of people think that household and national economic condition is satisfactory as most have ranked them in 3 or 4 out of 5.
- 3) Conservatives consists of slightly higher proportion of aged voters (50 years and above).
- 4) 50% of the voters are of age above 53 years and only the bottom 25% voters are aged less than 41 years.
- 5) There are more female voters than male voters.
- 6) Conservative voters have better political knowledge of political parties' position on European integration than their Labour counterparts.
- 7) Labour voters appears to have a pro-European integration opinion as opposed to Conservative voters.
- 8) 43% Conservative voters have rated the national economic condition average with score of 3, further indicating, the overall assessment to be between poor and average.
- 9) Candidates can focus to improve the image of economic conditions to gather more crowd favor.

Business Recommendations:

CNBE must gather data of voters aged between 18 and 24 so as to make the predictions more accurate.

It needs to be addressed that, the larger the number of voters, better the Machine Learning models can be optimized.

The dataset must also include additional assessment ratings about migration policy including refugee settlement, employment generation, income tax regime, etc.

Based on the existing data, the most optimized model is found to be the regularized Gradient Boosting Classifier model

This however would need re-tuning of hyperparameters with larger dataset to accurately predict the win for the Labour party. Irrespective of the size of the dataset, the regularized Gradient Boosting Classifier model could be deployed to build an exit poll which will still perform with great degree of accuracy

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.

1941-Roosevelt.txt

'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 institutions from disruption from without.\n\nTo us there has come a time, in the midst of swift happenings, to pause for a moment 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 not, we risk the real peril of inaction.\n\nLives of nations are determined not by the count of years, but by the lifetime of the 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 fullness 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 reason, 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 years have been living years -- fruitful years for the people of this democracy. For they have brought to us greater security and, I hope, a better understanding that life\'s ideals are to be measured in other than material things.\n\nMost vital to our present and our future is this experience of a democracy which successfully survived crisis at home; put away many evil things; built new structures on enduring lines; and, through it all, maintained the fact of its democracy.\n\nFor action has been taken within the three-way framework of the Constitution of the United States. The coordinate branches of the Government continue freely to function. The Bill of Rights remains inviolate. The freedom of elections is wholly maintained. Prophets of the downfall of American democracy have seen their dire predictions come to naught.\n\nDemocracy is not dying.\n\nWe know it because we have seen it revive--and grow.\n\nWe know it cannot die -- because it is built on the unhampered initiative of individual men and women joined together in a common enterprise -- an enterprise undertaken and carried through by the free expression of a free majority.\n\nWe know it because democracy

alone, of all forms of government, enlists the full force of men's enlightened will. We know it because democracy alone has constructed an unlimited civilization capable of infinite progress in the improvement of human life. We know it because, if we look below the surface, we sense it still spreading on every continent -- for it is the most humane, the most advanced, and in the end the most unconquerable of all forms of human society. A nation, like a person, has a body--a body that must be fed and clothed and housed, invigorated and rested, in a manner that measures up to the objectives of our time. A nation, like a person, has a mind -- a mind that must be kept informed and alert, that must know itself, that understands the hopes and the needs of its neighbors -- all the other nations that live within the narrowing circle of the world. And a nation, like a person, has something deeper, something more permanent, something larger than the sum of all its parts. It is that something which matters most to its future -- which calls forth the most sacred guarding of its present. It is a thing for which we find it difficult -- even impossible -- to hit upon a single, simple word. And yet we all understand what it is -- the spirit -- the faith of America. It is the product of centuries. It was born in the multitudes of those who came from many lands -- some of high degree, but mostly plain people, who sought here, early and late, to find freedom more freely. The democratic aspiration is no mere recent phase in human history. It is human history. It permeated the ancient life of early peoples. It blazed anew in the middle ages. It was written in Magna Charta. In the Americas its impact has been irresistible. America has been the New World in all tongues, to all peoples, not because this continent was a new-found land, but because all those who came here believed they could create upon this continent a new life -- a life that should be new in freedom. Its vitality was written into our own Mayflower Compact, into the Declaration of Independence, into the Constitution of the United States, into the Gettysburg Address. Those who first came here to carry out the longings of their spirit, and the millions who followed, and the stock that sprang from them -- all have moved forward constantly and consistently toward an ideal which in itself has gained stature and clarity with each generation. The hopes of the Republic cannot forever tolerate either undeserved poverty or self-serving wealth. We know that we still have far to go; that we must more greatly build the security and the opportunity and the knowledge of every citizen, in the measure justified by the resources and the capacity of the land. But it is not enough to achieve these purposes alone. It is not enough to clothe and feed the body of this Nation, and instruct and inform its mind. For there is also the spirit. And of the three, the greatest is the spirit. Without the body and the mind, as all men know, the Nation could not live. But if the spirit of America were killed, even though the Nation's body and mind, constricted in an alien world, lived on, the America we know would have perished. That spirit -- that faith -- speaks to us in our daily lives in ways often unnoticed, because they seem so obvious. It speaks to us here in the Capital of the Nation. It speaks to us through the processes of governing in the sovereignties of 48 States. It speaks to us in our counties, in our cities, in our towns, and in our villages. It speaks to us from the other nations of the hemisphere, and from those across the seas -- the enslaved, as well as the free. Sometimes we fail to hear or heed these voices of freedom because to us the privilege of our freedom is such an old, old story. The destiny of America was proclaimed in words of prophecy spoken by our first President in his first inaugural in 1789 -- words almost directed, it would seem, to this year of 1941: "The preservation of the sacred fire of liberty and the destiny of the republican

model of government are justly considered deeply, finally, staked on the experiment intrusted to the hands of the American people." \n\nIf we lose that sacred fire--if we let it be smothered with doubt and fear -- then we shall reject the destiny which Washington strove so valiantly and so triumphantly to establish. The preservation of the spirit and faith of the Nation does, and will, furnish the highest justification for every sacrifice that we may make in the cause of national defense.\n\nIn the face of great perils never before encountered, our strong purpose is to protect and to perpetuate the integrity of democracy.\n\nFor this we muster the spirit of America, and the faith of America.\n\nWe do not retreat. We are not content to stand still. As Americans, we go forward, in the service of our country, by the will of God.\n'

1961-Kennedy.txt

'Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well as a beginning -- signifying renewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our forebears I prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for which 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 word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans -- born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage -- and unwilling to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are committed today at home and around the world.\n\nLet every nation know, whether it wishes us well or ill, that we shall pay any price, bear any burden, meet any hardship, support any friend, oppose any foe, in order to assure the survival and the success of liberty.\n\nThis much we pledge -- and more.\n\nTo those old allies whose cultural and spiritual origins we share, we pledge the loyalty of faithful friends. United, there is little we cannot do in a host of cooperative ventures. Divided, there is little we can do -- for we dare not meet a powerful challenge at odds and split asunder.\n\nTo those new States whom we welcome to the ranks of the free, we pledge our word that one form of colonial control shall not have passed away merely to be replaced by a far more iron tyranny. We shall not always expect to find them supporting our view. But we shall always hope to find them strongly supporting their own freedom -- and to remember that, in the past, those who foolishly sought power by riding the back of the tiger ended up inside.\n\nTo those peoples in the huts and villages across the globe struggling to break the bonds of mass misery, we pledge our best efforts to help them help themselves, for whatever period is required -- not because the Communists may be doing it, not because we seek their votes, but because it is right. If a free society cannot help the many who are poor, it cannot save the few who are rich.\n\nTo our sister republics south of our border, we offer a special pledge -- to convert our good words into good deeds -- in a new alliance for progress -- to assist free men and free governments in casting off the chains of poverty. But this peaceful revolution of hope cannot

become the prey of hostile powers. Let all our neighbors know that we shall join with them to oppose aggression or subversion anywhere in the Americas. And let every other power know that this Hemisphere intends to remain the master of its own house.

To that world assembly of sovereign states, the United Nations, our last best hope in an age where the instruments of war have far outpaced the instruments of peace, we renew our pledge of support--to prevent it from becoming merely a forum for invective -- to strengthen its shield of the new and the weak -- and to enlarge the area in which its writ may run.

Finally, to those nations who would make themselves our adversary, we offer not a pledge but a request: that both sides begin anew the quest for peace, before the dark powers of destruction unleashed by science engulf all humanity in planned or accidental self-destruction.

We dare not tempt them with weakness. For only when our arms are sufficient beyond doubt can we be certain beyond doubt that they will never be employed.

But neither can two great and powerful groups of nations take comfort from our present course -- both sides overburdened by the cost of modern weapons, both rightly alarmed by the steady spread of the deadly atom, yet both racing to alter that uncertain balance of terror that stays the hand of mankind's final war.

So let us begin anew -- remembering on both sides that civility is not a sign of weakness, and sincerity is always subject to proof. Let us never negotiate out of fear. But let us never fear to negotiate.

Let both sides explore what problems unite us instead of belaboring those problems which divide us.

Let both sides, for the first time, formulate serious and precise proposals for the inspection and control of arms -- and bring the absolute power to destroy other nations under the absolute control of all nations.

Let both sides seek to invoke the wonders of science instead of its terrors. Together let us explore the stars, conquer the deserts, eradicate disease, tap the ocean depths, and encourage the arts and commerce.

Let both sides unite to heed in all corners of the earth the command of Isaiah -- to "undo the heavy burdens ... and to let the oppressed go free."

And if a beachhead of cooperation may push back the jungle of suspicion, let both sides join in creating a new endeavor, not a new balance of power, but a new world of law, where the strong are just and the weak secure and the peace preserved.

All this will not be finished in the first 100 days. Nor will it be finished in the first 1,000 days, nor in the life of this Administration, nor even perhaps in our lifetime on this planet. But let us begin.

In your hands, my fellow citizens, more than in mine, will rest the final success or failure of our course. Since this country was founded, each generation of Americans has been summoned to give testimony to its national loyalty. The graves of young Americans who answered the call to service surround the globe.

Now the trumpet summons us again -- not as a call to bear arms, though arms we need; not as a call to battle, though embattled we are -- but a call to bear the burden of a long twilight struggle, year in and year out, "rejoicing in hope, patient in tribulation" -- a struggle against the common enemies of man: tyranny, poverty, disease, and war itself.

Can we forge against these enemies a grand and global alliance, North and South, East and West, that can assure a more fruitful life for all mankind? Will you join in that historic effort?

In the long history of the world, only a few generations have been granted the role of defending freedom in its hour of maximum danger. I do not shrink from this responsibility -- I welcome it. I do not believe that any of us would exchange places with any other people or any other generation. The energy, the faith, the devotion which we bring to this endeavor will light our country and all who serve it -- and the glow from that fire can truly light the world.

And so,

my fellow Americans: ask not what your country can do for you -- ask what you can do for your country.\n\nMy fellow citizens of the world: ask not what America will do for you, but what together we can do for the freedom of man.\n\nFinally, whether you are citizens of America or citizens of the world, ask of us the same high standards of strength and sacrifice which we ask of you. With a good conscience our only sure reward, with history the final judge of our deeds, let us go forth to lead the land we love, asking His blessing and His help, but knowing that here on earth God's work must truly be our own.\n'

'1973-Nixon.txt'

'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and good country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly 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 about to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilities greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis past 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 greatest 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 flimsy peace which is merely an interlude between wars, but a peace which can endure for generations to come.\n\nIt is important that we understand both the necessity and the limitations of America's role in maintaining that peace.\n\nUnless we in America work to preserve the peace, there will be no peace.\n\nUnless we in America work to preserve freedom, there will be no freedom.\n\nBut let us clearly understand the new nature of America's role, as a result of the new policies we have adopted over these past four years.\n\nWe shall respect our treaty commitments.\n\nWe shall support vigorously the principle that no country has the right to impose its will or rule on another by force.\n\nWe shall continue, in this era of negotiation, to work for the limitation of nuclear arms, and to reduce the danger of confrontation between the great powers.\n\nWe shall do our share in defending peace and freedom in the world. But we shall expect others to do their share.\n\nThe time has passed when America will make every other nation's conflict our own, or make every other nation's future our responsibility, or presume to tell the people of other nations how to manage their own affairs.\n\nJust as we respect the right of each nation to determine its own future, we also recognize the responsibility of each nation to secure its own future.\n\nJust as America's role is indispensable in preserving the world's peace, so is each nation's role indispensable in preserving its own peace.\n\nTogether with the rest of the world, let us resolve to move forward from the beginnings we have made. Let us continue to bring down the walls of hostility which have divided the world for too long, and to build in their

place bridges of understanding -- so that despite profound differences between systems of government, the people of the world can be friends.\n\nLet us build a structure of peace in the world in which the weak are as safe as the strong -- in which each respects the right of the other to live by a different system -- in which those who would influence others will do so by the strength of their ideas, and not by the force of their arms.\n\nLet us accept that high responsibility not as a burden, but gladly -- gladly because the chance to build such a peace is the noblest endeavor in which a nation can engage; gladly, also, because only if we act greatly in meeting our responsibilities abroad will we remain a great Nation, and only if we remain a great Nation will we act greatly in meeting our challenges at home.\n\nWe have the chance today to do more than ever before in our history to make life better in America -- to ensure better education, better health, better housing, better transportation, a cleaner environment -- to restore respect for law, to make our communities more livable -- and to insure the God-given right of every American to full and equal opportunity.\n\nBecause the range of our needs is so great -- because the reach of our opportunities is so great -- let us be bold in our determination to meet those needs in new ways.\n\nJust as building a structure of peace abroad has required turning away from old policies that failed, so building a new era of progress at home requires turning away from old policies that have failed.\n\nAbroad, the shift from old policies to new has not been a retreat from our responsibilities, but a better way to peace.\n\nAnd at home, the shift from old policies to new will not be a retreat from our responsibilities, but a better way to progress.\n\nAbroad and at home, the key to those new responsibilities lies in the placing and the division of responsibility. We have lived too long with the consequences of attempting to gather all power and responsibility in Washington.\n\nAbroad and at home, the time has come to turn away from the condescending policies of paternalism -- of "Washington knows best." A person can be expected to act responsibly only if he has responsibility. This is human nature. So let us encourage individuals at home and nations abroad to do more for themselves, to decide more for themselves. Let us locate responsibility in more places. Let us measure what we will do for others by what they will do for themselves.\n\nThat is why today I offer no promise of a purely governmental solution for every problem. We have lived too long with that false promise. In trusting too much in government, we have asked of it more than it can deliver. This leads only to inflated expectations, to reduced individual effort, and to a disappointment and frustration that erode confidence both in what government can do and in what people can do.\n\nGovernment must learn to take less from people so that people can do more for themselves.\n\nLet us remember that America was built not by government, but by people -- not by welfare, but by work -- not by shirking responsibility, but by seeking responsibility.\n\nIn our own lives, let each of us ask -- not just what will government do for me, but what can I do for myself?\n\nIn the challenges we face together, let each of us ask -- not just how can government help, but how can I help?\n\nYour National Government has a great and vital role to play. And I pledge to you that where this Government should act, we will act boldly and we will lead boldly. But just as important is the role that each and every one of us must play, as an individual and as a member of his own community.\n\nFrom this day forward, let each of us make a solemn commitment in his own heart: to bear his responsibility, to do his part, to live his ideals -- so that together, we can see the dawn of a new age of progress for America, and together, as we celebrate our 200th anniversary as a nation, we can do so proud in the fulfillment of our promise to ourselves and to the world.\n\nAs America's longest and

most difficult war comes to an end, let us again learn to debate our differences with civility and decency. And let each of us reach out for that one precious quality government cannot provide -- a new level of respect for the rights and feelings of one another, a new level of respect for the individual human dignity which is the cherished birthright of every American.

Above all else, the time has come for us to renew our faith in ourselves and in America.

In recent years, that faith has been challenged.

Our children have been taught to be ashamed of their country, ashamed of their parents, ashamed of America's record at home and of its role in the world.

At every turn, we have been beset by those who find everything wrong with America and little that is right. But I am confident that this will not be the judgment of history on these remarkable times in which we are privileged to live.

America's record in this century has been unparalleled in the world's history for its responsibility, for its generosity, for its creativity and for its progress.

Let us be proud that our system has produced and provided more freedom and more abundance, more widely shared, than any other system in the history of the world.

Let us be proud that in each of the four wars in which we have been engaged in this century, including the one we are now bringing to an end, we have fought not for our selfish advantage, but to help others resist aggression.

Let us be proud that by our bold, new initiatives, and by our steadfastness for peace with honor, we have made a break-through toward creating in the world what the world has not known before -- a structure of peace that can last, not merely for our time, but for generations to come.

We are embarking here today on an era that presents challenges great as those any nation, or any generation, has ever faced.

We shall answer to God, to history, and to our conscience for the way in which we use these years.

As I stand in this place, so hallowed by history, I think of others who have stood here before me. I think of the dreams they had for America, and I think of how each recognized that he needed help far beyond himself in order to make those dreams come true.

Today, I ask your prayers that in the years ahead I may have God's help in making decisions that are right for America, and I pray for your help so that together we may be worthy of our challenge.

Let us pledge together to make these next four years the best four years in America's history, so that on its 200th birthday America will be as young and as vital as when it began, and as bright a beacon of hope for all the world.

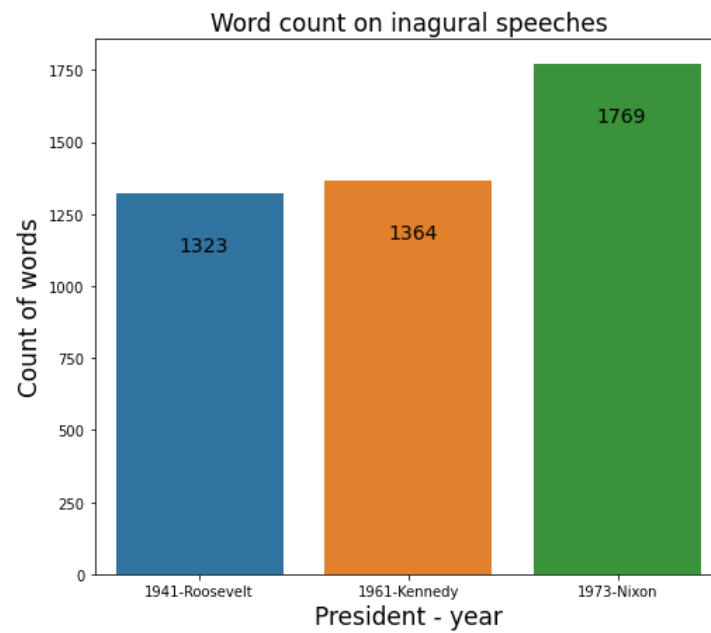
Let us go forward from here confident in hope, strong in our faith in one another, sustained by our faith in God who created us, and striving always to serve His purpose.

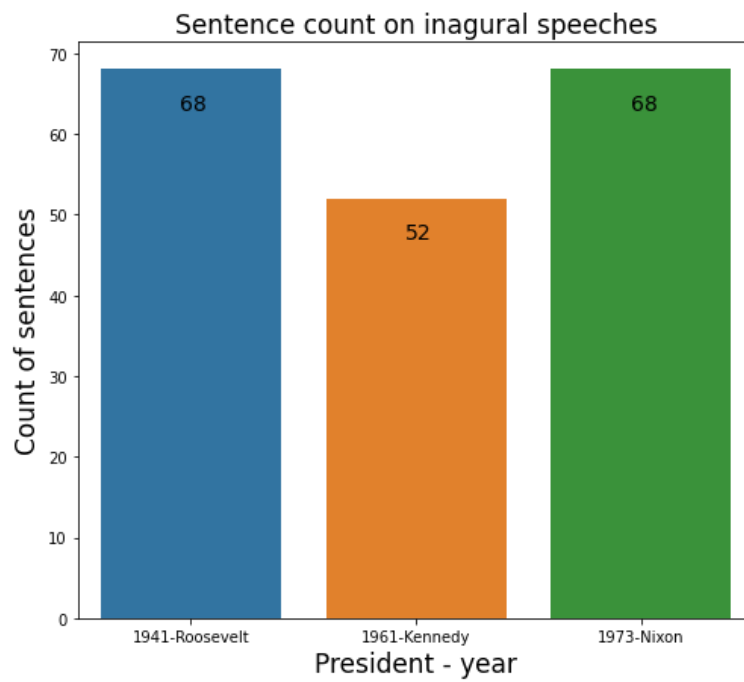
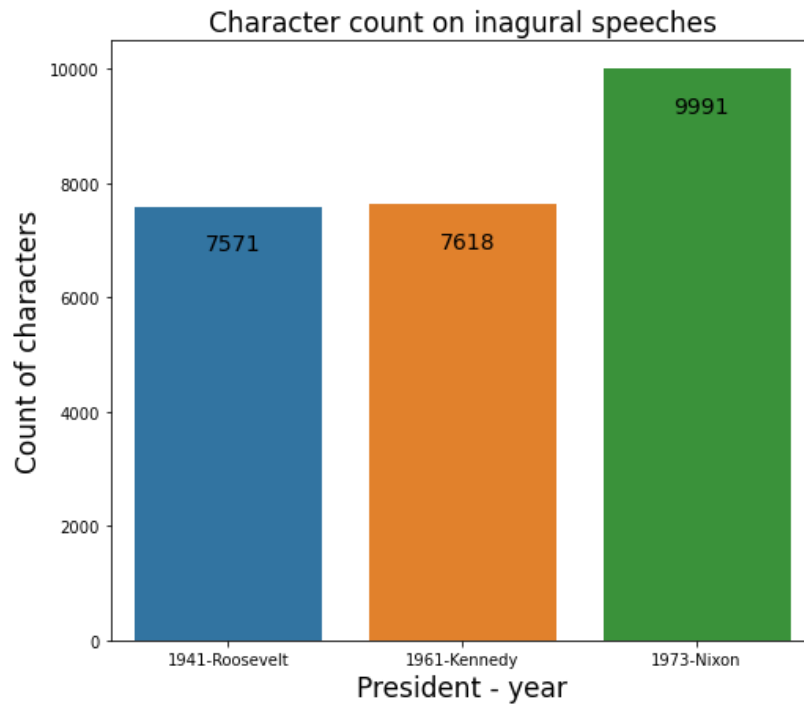
Table 19 – Speeches into Data frame

	text
1941-Roosevelt	On each national day of inauguration since 178...
1961-Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...
1973-Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...

Table 20 – Number of words, characters and sentences

	text	word_count	char_count	sents_count
1941-Roosevelt	On each national day of inauguration since 178...	1323	7571	68
1961-Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364	7618	52
1973-Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769	9991	68





2.2 Remove all the stopwords from all three speeches.

Basic Pre-Processing:

Table 21 – Number of Uppercase Words

	text	upper
1941-Roosevelt	On each national day of inauguration since 178...	3
1961-Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	5
1973-Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	14

Table 22 – Number of Uppercase Letters

	text	upper_letter
1941-Roosevelt	On each national day of inauguration since 178...	119
1961-Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	94
1973-Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	132

Table 23 – Number of Numeric

	text	numerics
1941-Roosevelt	On each national day of inauguration since 178...	14
1961-Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	7
1973-Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10

Table 24 – Lower case conversion

1941-Roosevelt	on each national day of inauguration since 178...
1961-Kennedy	vice president johnson, mr. speaker, mr. chief...
1973-Nixon	mr. vice president, mr. speaker, mr. chief jus...

Table 25 – Remove punctuation

1941-Roosevelt on each national day of inauguration since 178...
 1961-Kennedy vice president johnson mr speaker mr chief jus...
 1973-Nixon mr vice president mr speaker mr chief justice ...

Removal of StopWords & Stemming

List to stop words is extended with (['mr','on','it','the','in','let','to','us','shall','since'])

Stemming -refers to the removal of suffices, like “ing”, “ly”, “s”, etc. by a simple rule-based approach

Table 26 – Removing stop words and perform stemming

1941-Roosevelt nation day inaugur 1789 peopl renew sen dedic ...
 1961-Kennedy vice presid johnson speaker chief justic presi...
 1973-Nixon vice presid speaker chief justic senat cook mr...

Number of words after stop words removal = word_count 1 in below table

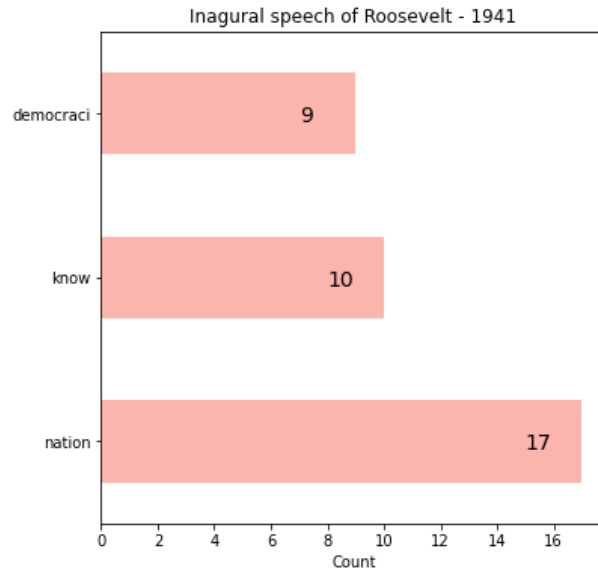
Table 27 – Word count after removing stopwords and cleaning

	text	word_count	char_count	sents_count	upper	upper_letter	numerics	word_count1
1941-Roosevelt	nation day inaugur 1789 peopl renew sens dedic...	1323	7571	68	3	119	14	616
1961-Kennedy	vice presid johnson speaker chief justic presi...	1364	7618	52	5	94	7	657
1973-Nixon	vice presid speaker chief justic senat cook mr...	1769	9991	68	14	132	10	774

2.3 Which word occurs the most number of times in his inaugural address for each president?

Top three words occurs the most number of times in inaugural address by Roosevelt

nation 17
 know 10
 democraci 9

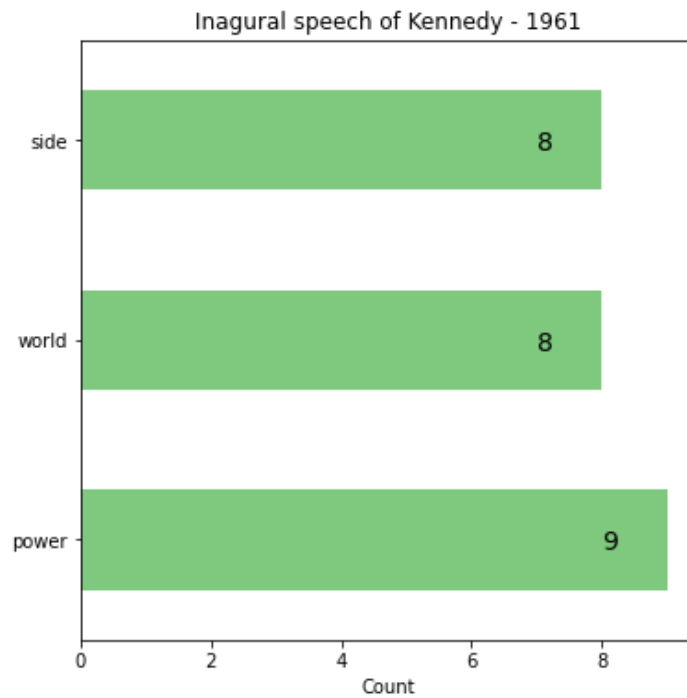


Top three words occurs the most number of times in inaugural address by Kennedy

power 9

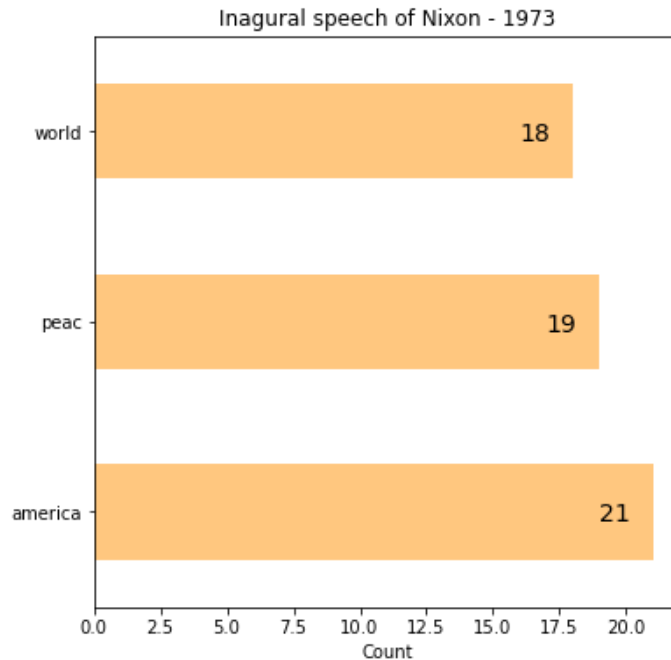
world 8

side 8



Top three words occurs the most number of times in inaugural address by Nixon

america	21
peac	19
world	18



2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords)

