DSBA ML Project

Vikash Kumar

05.02.2023

Table of Contents

Problem statement 3
Perform data upload and EDA 3
Checking Skewness of data 6
Data information 8
Scaling of data 17
Model Building 18
Model 1: Logistic Regression 19
LogisticRegression 20
Model 2 Linear Discriminant Analysis 21
Model 3: Performing KNN Model 24
Model 4: Performing Naïve Bayes Model 27
Tuning of Models :Hyper-Parameters Tuning 29
Bagging 37
Boosting 38

Table 1 Top 5 of dataset 3
Figure 2 Skewness of variables 7

Problem statement

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

1.1) Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.

Perform data upload and EDA

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

Table 1 Top 5 of dataset

```
Range Index: 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)
```

Table 2 Data info dataset

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

Table 3 Data type of dataset

Insights:

1.There is total 1525 entries in dataset with 9 variables.

2.The data has only 2 object category like Vote and Gender.

3.Seems there is NO missing values in data set.

4.One variable 'Unnamed ‘is found which may be indication of Serial Number.

Figure 3 Datatype of dataset

```
Number of duplicate rows = 8
```

	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

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

```
Dataset Before the removal of Duplicates (1525, 9)
Dataset after the removal of Duplicates (1517, 9)
```

Table 3 Duplicate values in dataset

```
Missing Values in data is:
 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
```

NO missing value is detected in data set

Table 4 Null value in dataset

	count	mean	std	min	25%	50%	75%	max
age	1517.00	54.24	15.70	24.00	41.00	53.00	67.00	93.00
economic.cond.national	1517.00	3.25	0.88	1.00	3.00	3.00	4.00	5.00

	count	mean	std	min	25%	50%	75%	max
economic.cond.household	1517.00	3.14	0.93	1.00	3.00	3.00	4.00	5.00
Blair	1517.00	3.34	1.17	1.00	2.00	4.00	4.00	5.00
Hague	1517.00	2.75	1.23	1.00	2.00	2.00	4.00	5.00
Europe	1517.00	6.74	3.30	1.00	4.00	6.00	10.00	11.00
political.knowledge	1517.00	1.54	1.08	0.00	0.00	2.00	2.00	3.00

Table 5 Fivefold statistical summary of data

Insights:

1.The Mean and median are close to each other so it can be presumed that data has symmetry in distribution to large extent.
2.Possibility of Outliers in Age can be observed because the maximum value is 93.Which cannot be ruled out for age purpose.
3.Details of outlier can be detected during its dedicated operation.
4.Standard deviation is almost close to 1 in most of the cases like Economic conditions national and household, Blair, Hague, Political knowledge.
5.While its differing in case of age and Europe.

Checking Skewness of data

Skewness is degree of asymmetry observed in a in a statistical distribution, in which the curve appears distorted
or skewed either to the left or to the right.
Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.
In this statistical distribution. Distributions can exhibit right(positive) skewness or
left (negative) skewness to varying degrees.
While Normal distribution (bell curve) exhibits zero skewness.
Skewness is computed for each row or each column of the data present in the DataFrame object

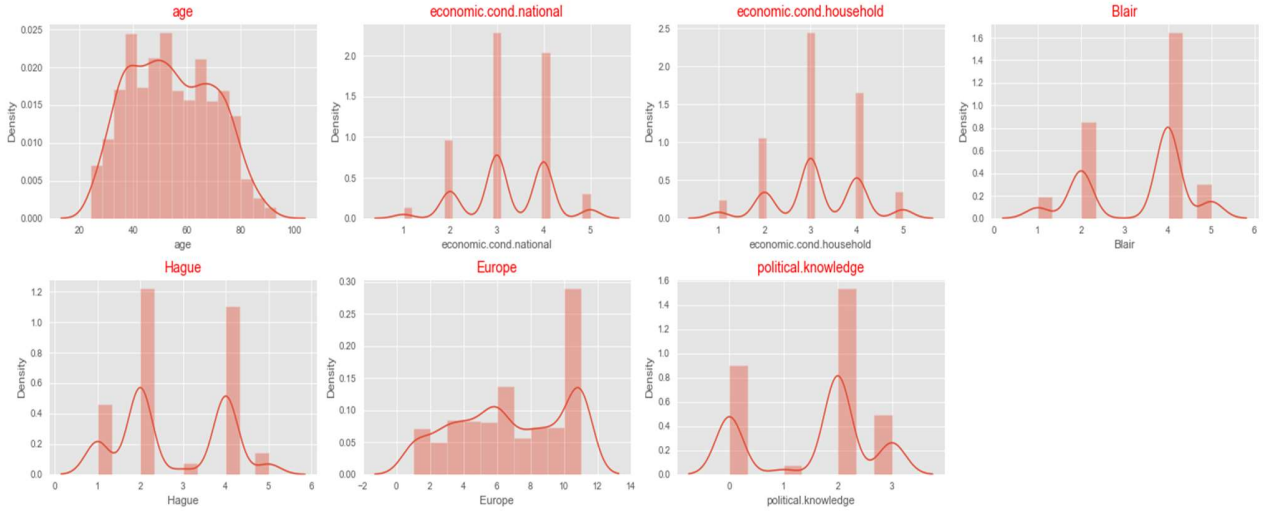


Figure 2 Skewness of variables

Variable	Skew
age	0.14
economic.cond.national	-0.24
economic.cond.household	-0.14
Blair	-0.54
Hague	0.15
Europe	-0.14
political.knowledge	-0.42

dtype: float64

Table 6 Skewness of of data

Insights:

1. In most of cases the skewness is between -0.5 and 0.5, We can say that the data are fairly sy mmetrical.

2.Age is almost normally distributed. 3.Only in case of 'Vote' skewness is between -1 and — 0.5 in this case the data can be treated as moderately skewed

4.A negative skewness value in the output indicates an asymmetry in the distribution correspon ding to row and the tail is larger towards the left-hand side of the distribution. Determination b y Box Plot:

We can determine whether or not a distribution is skewed based on the location of the median value in the box plot/ In our case the median is in the middle (with sliglt movement towards ne gative side) of the box and the whiskers are roughly equal on each side, the distribution is sym metrical (or “no” skew)..(Fig showing in Outlier detection)

1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

Data information

```
Missing Values in data is:
 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
```

Table 7 Null Value in dataset

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

Table 7 Datatype in dataset

```
Shape of data (1517, 9)
```

Table 8 shape of dataset

Univariate Analysis

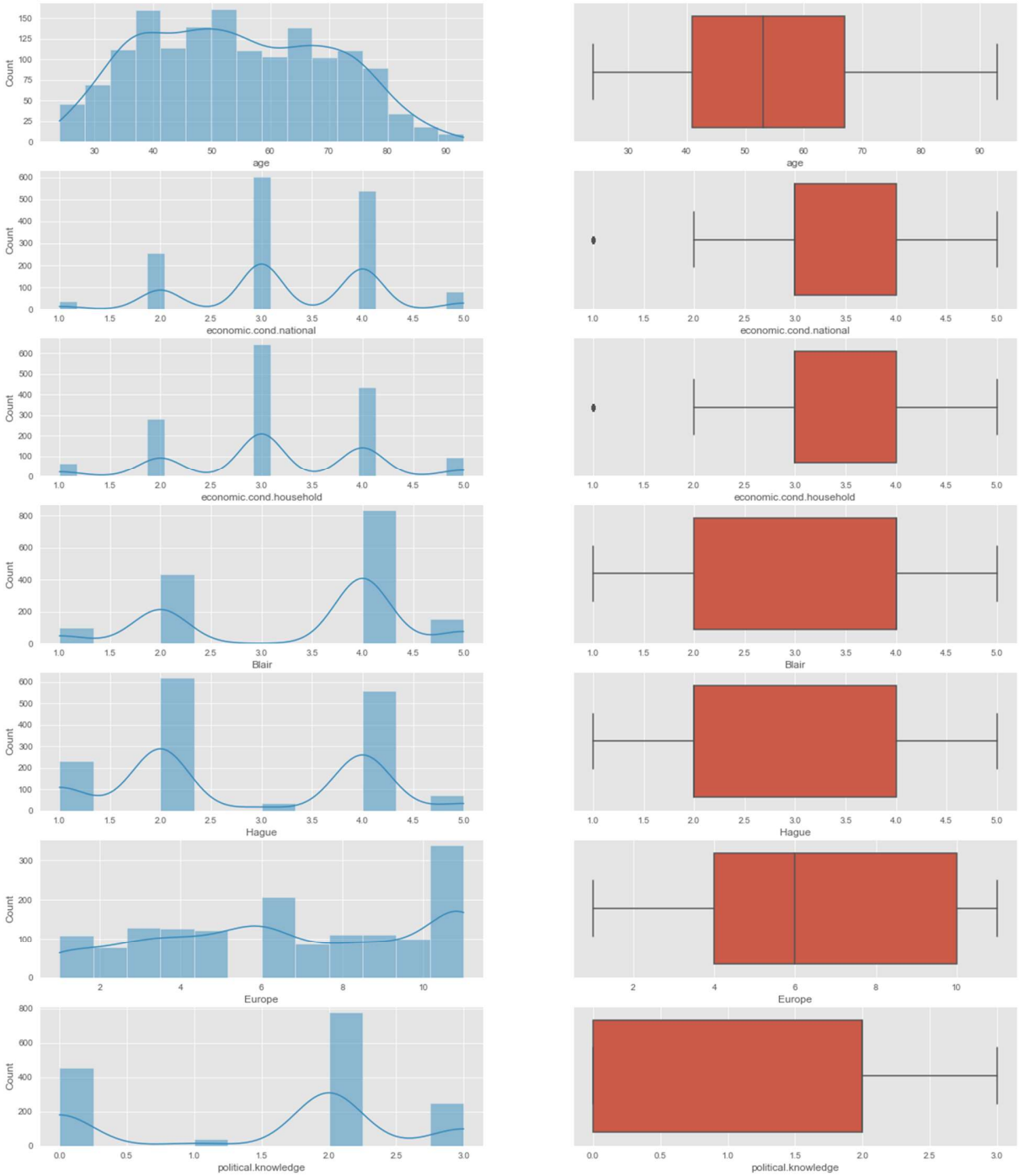


Figure 2 Data distribution of Numerical variables

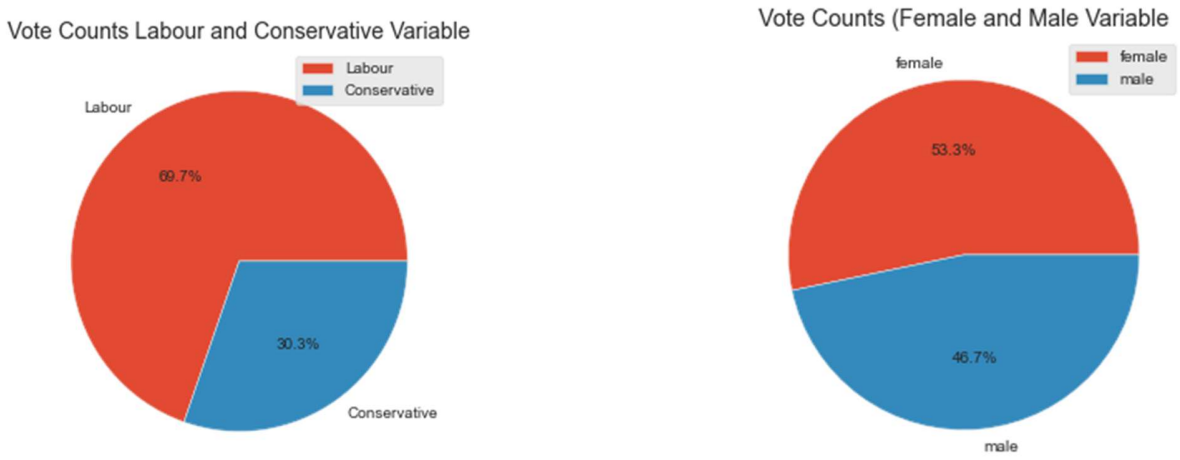


Figure 2 Data distribution Categorical variables

Insight of Data Distribution:

1.In two variables like "economic.cond.national" and "economic.cond.household" both has showing that the mid-level economic conditions population are more than others.
2.Wide range of age groups
3.High assessment of for labor leader because bar is highest for 4 on scale of 800.
4.Assestment for Conservative leaders are comparative low in score because of bar no 2 has highest elevation and less for no 4 in scale of 600.
5.Highest number of people are are Eurosceptic because its percentage is more than double of remaining population.
6.Political knowledge of parties position in Europe integration are fairly high as bar 2 has highest elevation on scale of 800.of scale of rating 0-3.
7.Females are more in number than of males.
8.As per visualization the labor party has more than 50% better choice than of conservative.
9.The Labor party is choice of almost 70% voters while 30% for conservative.
10.The female population are 7% more over the male population.

Bivariate Analysis

Bivariate analysis is stated to be an analysis of any concurrent relation between two variables or attributes. This study explores the relationship of two variables as well as the depth of this relationship to figure out if there are any discrepancies between two variables and any causes of this difference

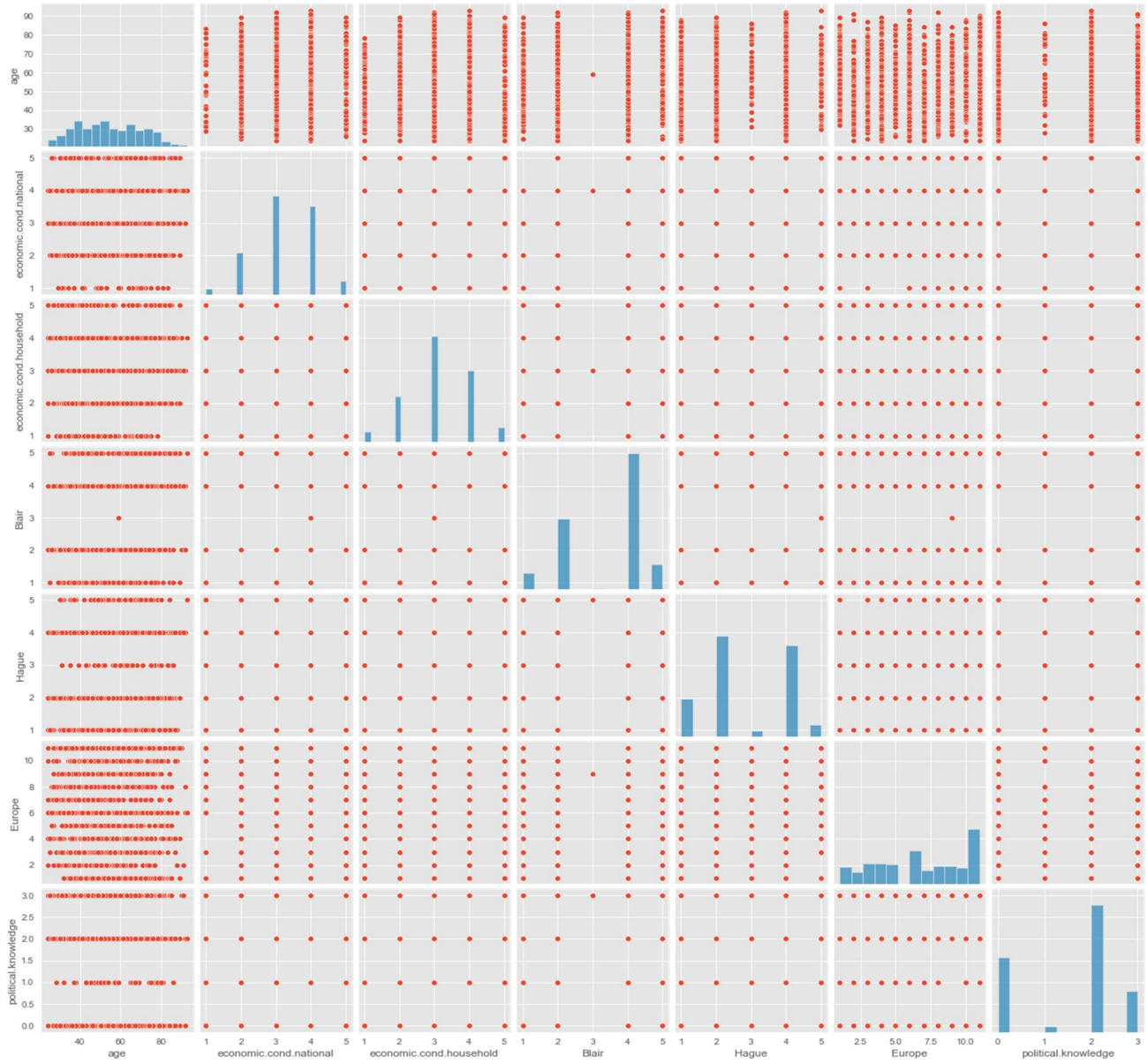
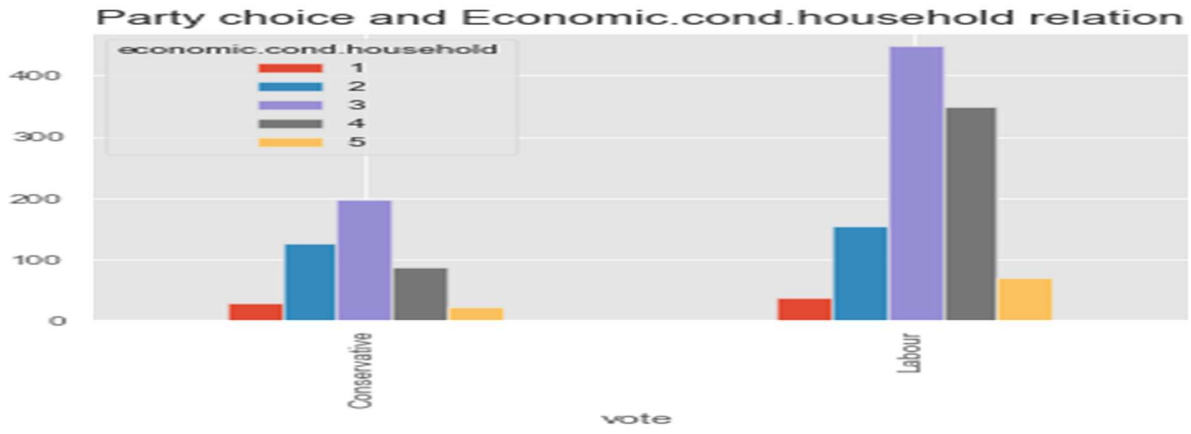
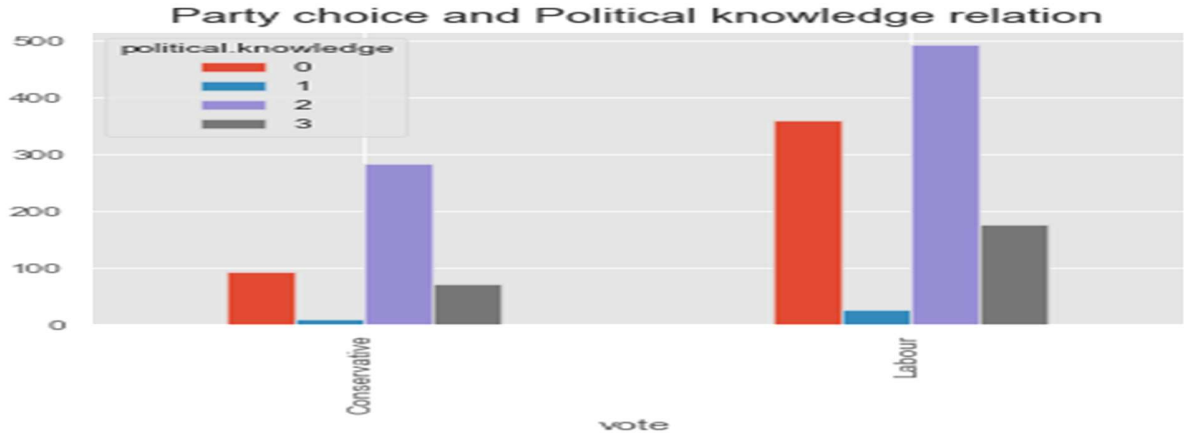
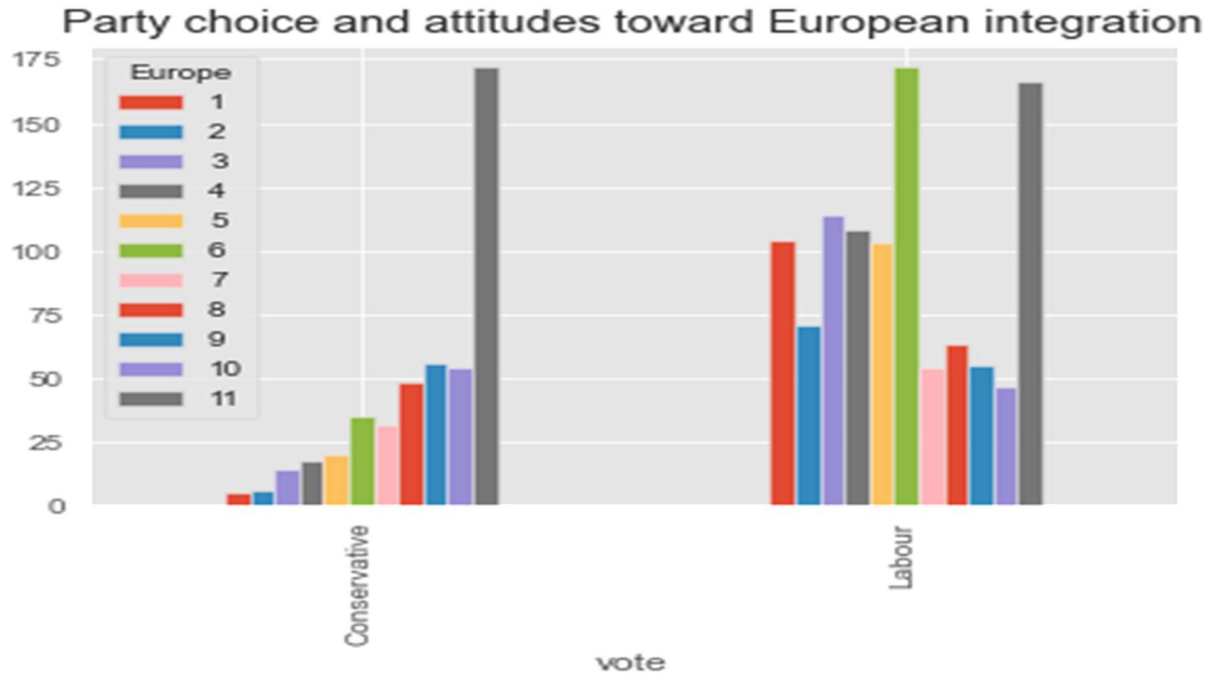


Figure 3 Bivariate Analysis


Party choice and Economic.cond.household relation
economic.cond.household
1
2
3
4
5
400
300
200
100
0
Conservative
Labour
vote


Party choice and Political knowledge relation
political.knowledge
0
1
2
3
500
400
300
200
100
0
Conservative
Labour
vote


Party choice and attitudes toward European integration
Europe
1
2
3
4
5
6
7
8
9
10
11
175
150
125
100
75
50
25
0
Conservative
Labour
vote

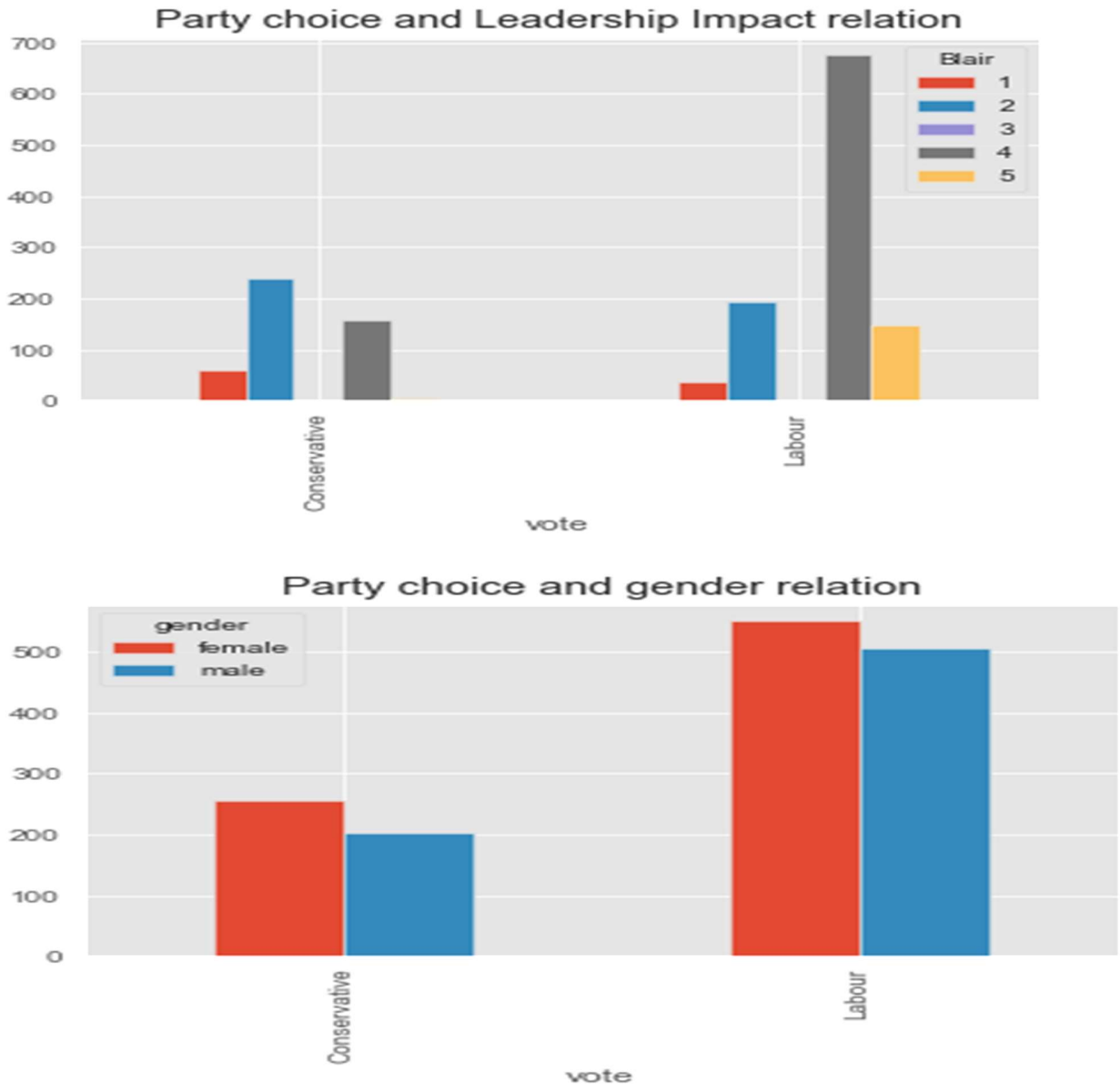


Figure 4 Bivariate Analysis of variables with Party choice(vote)

Inferences of Bivariate variables with Vote:

All economic condition of Household has orientation towards Labour party.

Labour party has edge with class 2 voters nearly 500 followed by class 0 aprox 380 and class 3 almost 180.While class 1 has low voter both the parties.

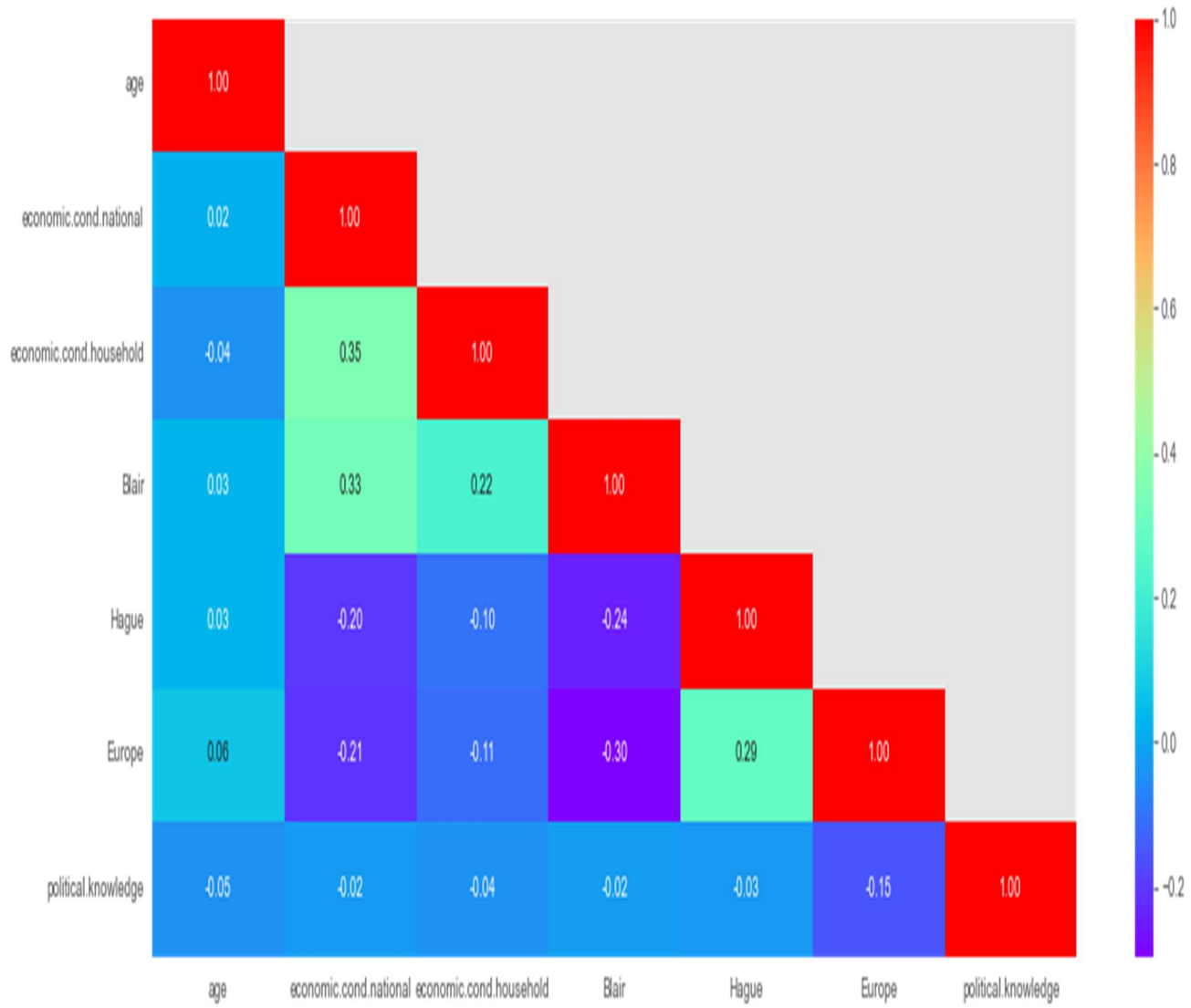
Labour party has edge with class 2 voters nearly 500 followed by class 0 aprox 380 and class 3 almost 180.While class 1 has low voter both the parties.

Labour party leader has high impact among almost 700 with rating of 4

Both male and female has oriented towards the Labour party with large margin.

Multivariate Analysis

Multivariate analysis encompasses all statistical techniques that are used to analyze more than two variables at once. The aim is to find patterns and correlations between several variables simultaneously—allowing for a much deeper, more complex understanding of a given scenario than you'll get with bivariate analysis.



There is no Such high correlation are being observe among variables.

Figure 5 Multivariate Analysis of all variables

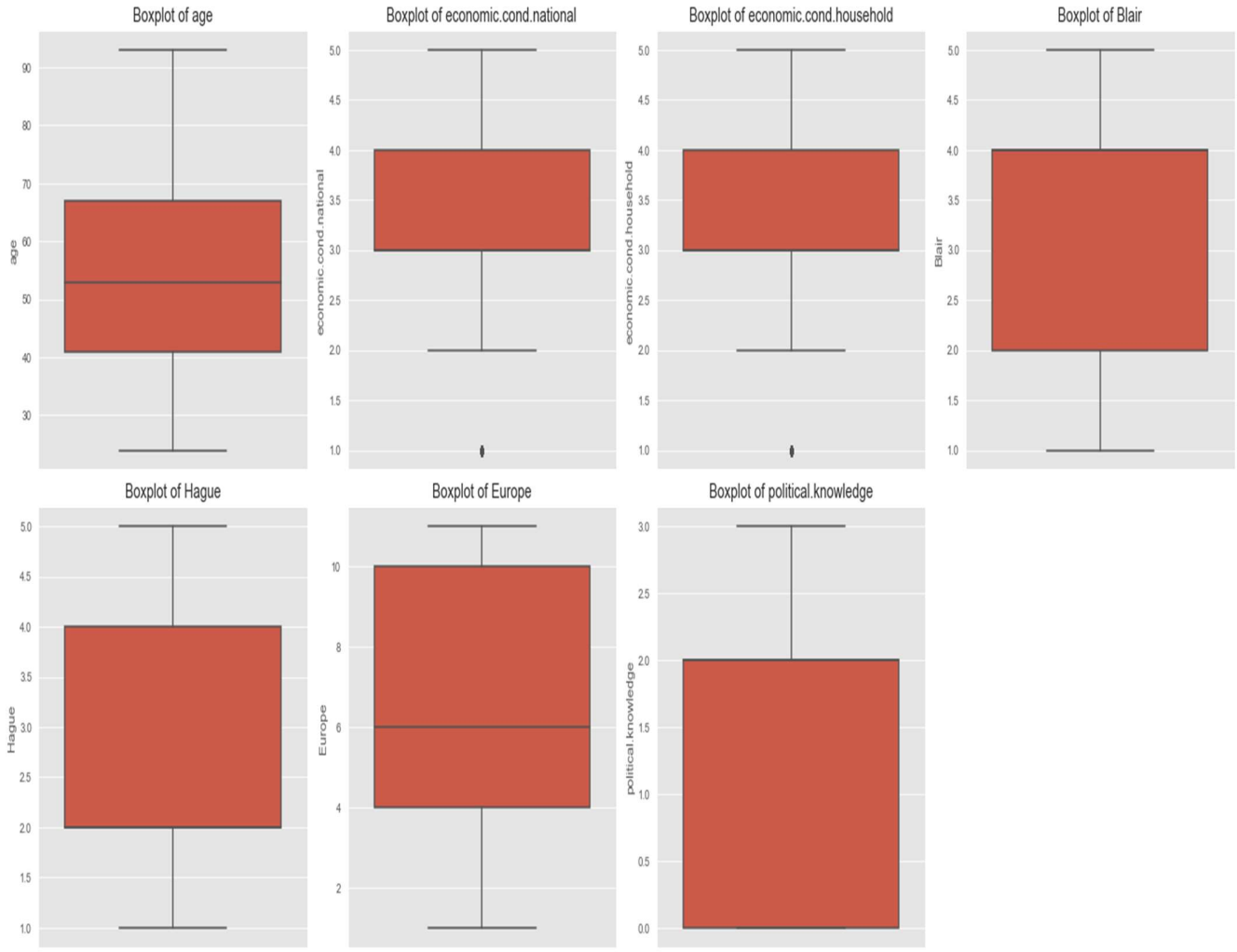


Figure 6 Numerical variable plotting for Outlier Detection

Insights:

1. Only Two has one point each as outlier can be detected in Boxplot but it seems some error because the both the variables are on 1-5 scale. and No point beyond this is dataset.
2. All the variables are are on the close mean position.

Economic condition nation and household the variable has Ordinal variable, hence there is NO NEED of TREATMENT of Outliers.

3. Economic condition nation and household has same mean.
4. In Europe the 3rd Quartile has large number of datapoints than others quartile with maens 0f 6.
5. .Ploitical knowledge has all data point fall in Quartile 1 and 2.
6. Both the variable has Ordinal variable so the outliers not to remove for further processing.

1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed

Encoding of Categorical Columns

One Hot encoding for Categorical data

The Two variables are present as "Object"(Gender) as original form but for further processing these has to encoding. So, we chose for One Hot encoding. Both has two categories so by this method the columns of dataset will remain same after drop first. While for target variable will go for manual encoding.

One-Hot-Encoding is used to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1 or 0 based on the presence or absence of the categorical value in the record.

This is required to do since the machine learning algorithms only works on the numerical data. That is why there is a need to convert the categorical column into numerical one.

get dummies is the method which creates dummy variable for each categorical variable.

It is considered a good practice to set parameter drop first as True whenever get dummies is used. It reduces the chances of multicollinearity which will be covered in coming courses and the number of features are also less as compared to drop first=False

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

Table 9 One Hot encoding for Gender which qualify for it drop first=True will drop one of the two

Encoding of Target Variable "Vote"(Party choice) as per instruction

```
1    1057
0     460
Name: vote, dtype: int64
```

Table 10 Coding of Vote (Target Variable)

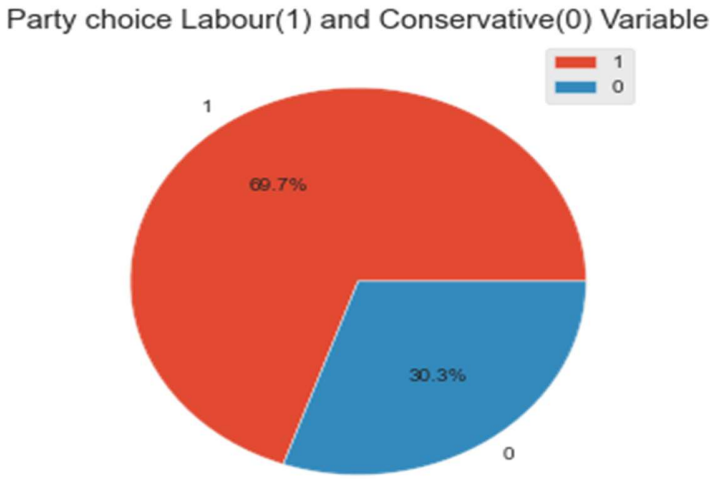


Figure 6 Value Counts of Categorical variable-Vote

Scaling of data

NO SCALING is required because(except KNN)

Often the variables of the data set are of different scales i.e., one variable is in millions and other in only 100. For e.g., in our data set Age is having values with in and 100 and other variables too are within range of 100. Since the data in these variables are of same scales, it is easy to compare these variables without scaling.
Some machine learning algorithms are sensitive to feature scaling while others are virtually invariant to it. Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. Also, distance algorithms like KNN, K-means, and SVM are most affected by the range of features. This is because behind the scenes they are using distances between data points to determine their similarity.

Scaling Types:

- Normalization: Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
- Standardization: Standardization is another scaling technique where the values are cantered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ. It is also referred to as root mean square deviation.
Standard deviation is almost close to 1 in most of the cases like Economic conditions national and housheols,Blair,Hague, Political knowledge, While its differing in case of age and Europe

The square of the standard deviation is the variance. It is also a measure of dispersion.

Model Building

- Split Data as Train and Test set
- Logistic Regression Classifier, Evaluate Model Performance and Predict Test Data
- Linear Discriminant Analysis Classifier and Evaluate Model Performance and Predict Test Data
- Naïve Bayes Model Classifier Analysis Evaluate Model Performance and Predict Test Data
- KNN Model Evaluate Model Performance and Predict Test Data

Extracting the target column into separate vectors for training set and test set

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender_male
0	43	3	3	4	1	2	2	0
1	36	4	4	4	4	5	2	1
2	35	4	4	5	2	3	2	1
3	24	4	2	2	1	4	0	0
4	41	2	2	1	1	6	2	1

Table 10 Top 5 of target column into separate vectors for training set and test set

Data Split: Spliting the data into training and test in ratio of 70:30

Discussion:
1.The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. 2.The train_test_split() method is used to split our data into train and test sets. Samples from the original training dataset are split into the two subsets using random selection. This is to ensure that the train and test datasets are representative of the original dataset.
Procedure: First, we need to divide our data into features (X) and labels (y). The dataFrame gets divided into X_train, X_test , y_train and y_test. X_train and y_train sets are used for training and fitting the model. The X_test and y_test sets are used for testing the model if it's predicting the right outputs/labels. we can explicitly test the size of the train and test sets.
3. The procedure has one main configuration parameter, which is the size of the train and test sets. This is most commonly expressed as a percentage between 0 and 1 for either the train or test datasets, we keep our train sets larger than the test sets.
4.There is no optimal split percentage on train and test but 75:25 and 70:30 are most common in analysis.
5.The objective is to estimate the performance of the machine learning model on new data: data not used to train the model 6.This is done to ensure that datasets are a representative sample (e.g. random sample) of the original dataset, which in turn, should be a representative sample of observations from the problem domain
7.Stratified Train-Test Splits: Some classification problems do not have a balanced number of examples

for each class label. As such, it is desirable to split the dataset into train and test sets in a way that preserves the same proportions of examples in each class as observed in the original dataset.

```
X_train (1061, 8)
X_test (456, 8)
y_train (1061,)
y_test (456,)
Total Obs 1517
```

Table 11 Dimension of data after train test split

1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

Model 1: Logistic Regression

Logistic regression is statistical method to make predictions on binary classes, the target/outcome variable of these models can only have two possible classes. In examples where the tarhet variable is of categorical nature, the model uses a log of odds as the dependent variable. Logistic regression there after compute the probability of an event occurrences.

For this project we can add in the OneVsRestClassifer described in the beginning of this chapter to make our multiclass problem binary in order to fit to a logistics regression model.

Logical reason behind the selection of different values for the parameters involved in each model.

That we use parameter C as our regularization parameter. Parameter C = 1/λ. Lambda (λ) controls the trade-off between allowing the model to increase it's complexity as much as it wants with trying to keep it simple. For example, if λ is very low or 0, the model will have enough power to increase it's complexity (overfit) by assigning big values to the weights for each parameter. If, in the other hand, we increase the value of λ, the model will tend to underfit, as the model will become too simple. Parameter C will work the other way around. For small values of C, we increase the regularization strength which will create simple models which underfit the data. For big values of C, we low the power of regularization which implies the model is allowed to increase it's complexity, and therefore, overfit the data.

LogisticRegression

#Formulation of LOGISTIC REGRESSION model on the train data.

```
TRAINING RESULTS:
Accuracy Score: 0.84
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.65      0.70       307
           1       0.87      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

Confusion Matrix:
[[200 107]
 [ 68 686]]

Average Accuracy:     0.8313

Standard Deviation:   0.0264
```

Table 12 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Training data

```
TEST RESULTS:
Accuracy Score: 0.8246

Classification Report:
              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


Confusion Matrix:
[[110  43]
 [ 37 266]]
```

Table 12 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
Accuracy of logistic regression classifier on train set: 0.84
Accuracy of logistic regression classifier on test set: 0.82
```

Validness of Model - Logistic Regression:

1.The Train and test results are very close to each other with difference of less tha 2% ,
this can be considered as good case of classification.
2.The accuracy score is also pretty good at 84% and 82% on test data.
3.The variation in precision are minimum while there is slight variation in Recall and F1 score but seems consistent for both train and test set. 4.In AUC score for both train(89%) and test data(87%) are close to each other also.
5.In confusion matrix the True positives are 54& and True Negatives are 24% whereas 8% are predict as positive and 9% are predicted as Negative which is not in actual.
6.Hence there is No case of over fitting or Under fitting of data.
7.The model would be stable and will perform good.
8.As per this model can predict with 82% of accuracy that Labour party will get the edge over conservatives.

Model 2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) algorithm for classification predictive modelling problems LDA makes predictions by estimating the probability that a new set of inputs belongs to each class. The class that gets the highest probability is the output class and a prediction is made.

The model uses Bayes Theorem to estimate the probabilities. Briefly Bayes' Theorem can be used to estimate the probability of the output class (k) given the input (x) using the probability of each class and the probability of the data belonging to each class:

P(Y=x|X=x) = (PIk * fk(x)) / sum(PIl * fl(x))

Where PIk refers to the base probability of each class (k) observed in your training data (e.g. 0.5 for a 50-50 split in a two class problem). In Bayes' Theorem this is called the prior probability.

PIk = nk/n

The f(x) above is the estimated probability of x belonging to the class. A Gaussian distribution function is used for f(x). Plugging the Gaussian into the above equation and simplifying we end up with the equation below. This is called a discriminate function and the class is calculated as having the largest value will be the output classification (y):

Dk(x) = x * (muk/siga^2) – (muk^2/(2*sigma^2)) + ln(PIk)

Dk(x) is the discriminate function for class k given input x, the muk, sigma^2 and PIk are all estimated from data.

When we have meaningfully labelled data, Applying LDA to a dataset will tell us how linearly separable our dataset is (which, depending on the context, may also be a commentary on how meaningful our labels are). That, in turn, is a good marker of how hard we will have to work to generate a model with reasonable classification performance. Furthermore, by looking at what variables load highly, we can determine what elements of the dataset are the strongest signals for each of our classes

```
TRAINING RESULTS:

Accuracy Score: 0.8341

Classification Report:
              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


Confusion Matrix:
[[200 107]
 [ 69 685]]

Average Accuracy:     0.8257

Standard Deviation:   0.0282
```

Table 14 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Training data

```
TEST RESULTS:

Accuracy Score: 0.8333

Classification Report:
              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


Confusion Matrix:
[[111  42]
 [ 34 269]]
Accuracy of Linear Discriminant Analysis Classifier on train set:0.83
Accuracy of Linear Discriminant Analysis Classifier on test  set:0.83
```

Table 15 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

classification Report of the training data for cutoff of 0.35 :

	precision	recall	f1-score	support
0	0.80	0.53	0.64	307
1	0.83	0.95	0.88	754
accuracy			0.82	1061
macro avg	0.81	0.74	0.76	1061
weighted avg	0.82	0.82	0.81	1061

Classification Report of the training data for cutoff of 0.4 :

	precision	recall	f1-score	support
0	0.80	0.57	0.67	307
1	0.84	0.94	0.89	754
accuracy			0.83	1061
macro avg	0.82	0.76	0.78	1061
weighted avg	0.83	0.83	0.82	1061

Classification Report of the training data for cutoff of 0.45 :

	precision	recall	f1-score	support
0	0.76	0.62	0.68	307
1	0.85	0.92	0.89	754
accuracy			0.83	1061
macro avg	0.81	0.77	0.78	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of the training data for cutoff of 0.5 :

	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 of the training data for cutoff of 0.55 :

	precision	recall	f1-score	support
0	0.71	0.68	0.70	307
1	0.87	0.89	0.88	754

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

```
Classification Report of the training data for cutoff of  0.6 :

              precision    recall  f1-score   support

           0       0.70      0.71      0.71       307
           1       0.88      0.87      0.88       754

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

Table 16 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Training data on different cuts for better accuracy

Validness of Model - LDA:

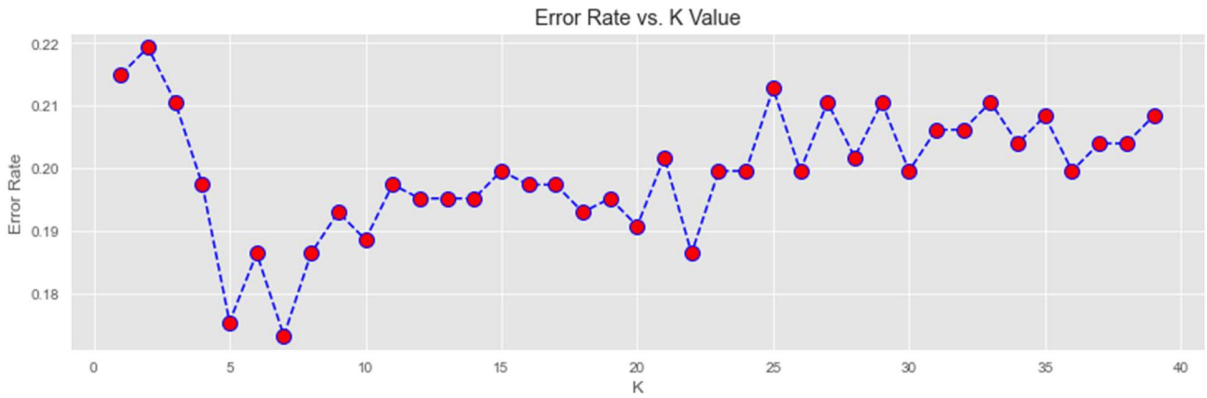
1.The accuracy score is also pretty good at 83% for train and test both.
2.The variation in precision and slight variation in Recall and F1 score but seems consistent for both train and test set. 4.In AUC score is almost 89% for both train and test data.
3.In confusion matrix classification ,True positives are 60 % and True Negatives are 23% whereas 8% are predict as positive and 4.14% are predicted as Negative which is not in actual.
5.Hence there is No case of over fitting or Under fitting of data.
6.The model would be stable and will perform good.
7.As per this model the Labour party will get the edge over conservatives.
8.This model would be also stable and good for prediction.

1.5)Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

Model 3: Performing KNN Model

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. To tackle the high variance of the Hold-out method, the k-fold method is used. The idea is simple, divide the whole dataset into 'k' sets preferably of equal sizes. Then the first set is selected as the test set and the rest 'k-1' sets are used to train the data. Error is calculated for

this particular dataset. Then the steps are repeated, i.e. the second set is selected as the test data, and the remaining 'k-1' sets are used as the training data. Again, the error is calculated. Similarly, the process continues for 'k' times. In the end, the CV error is given as the mean of the total errors calculated individually.



The Lowest Error rate at 7 so the Neighbor would be is 7.

Figure 7 Error rate Vs K value

knn = KNeighborsClassifier(n_neighbors=7,weights='distance',algorithm='brute',metric = 'minkowski')

```
TRAINING RESULTS:

Accuracy Score: 1.0000

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       307
           1       1.00      1.00      1.00       754

    accuracy                           1.00      1061
   macro avg       1.00      1.00      1.00      1061
weighted avg       1.00      1.00      1.00      1061


Confusion Matrix:
[[307   0]
 [  0 754]]

Average Accuracy:     0.8059

Standard Deviation:   0.0348
```

Table 17 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Training data

```
TEST RESULTS:

Accuracy Score: 0.82

Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.65      0.70       153
           1       0.84      0.90      0.87       303

    accuracy                           0.82       456
   macro avg       0.80      0.78      0.79       456
weighted avg       0.81      0.82      0.81       456


Confusion Matrix:
[[ 99  54]
 [ 29 274]]
```

Table 18 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 1.000
AUC for the Test Data: 0.868
AUC for the Training Data: 1.000
AUC for the Test Data: 0.868
```

Table 17 Accuracy Score on train and test data and AUC score

Validness of Model - KNN:

1.The Train and test results are very far to each other with difference of more than 17%, this cannot be considered as good for classification
2.Large variation in all parameters of scores. 3.Although The model accuracy score is also pretty good at 82% on test and 100% on train.
3. Hence the case can be considered as case of OVER fitting.
4.It would be advisable to DROP this model although the Scaling is not done .

Model 4: Performing Naïve Bayes Model

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
Naïve Bayes Classifier is a probabilistic classifier and is based on Bayes Theorem.
It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
Principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Bayes theorem gives us the probability of Event A to happen given that event B has occurred.

Improvement of Model Accuracy Logic

Total number of fits is 1000 since the cv is defined as 10 and there are 100 candidates (Var smoothing has 100 defined parameters). Therefore, the calculation for a total number of fits → 10 x [100] = 1000.

estimator is the machine learning model of interest, provided the model has a scoring function; in this case, the model assigned is GaussianNB(). param_grid is a dictionary with parameters names (string) as keys and lists of parameter settings to try as values; this enables searching over any sequence of parameter settings. verbose is the verbosity: the higher, the more messages; in this case, it is set to 1. cv is the cross-validation generator or an iterable, in this case, there is a 10-fold cross-validation. n_jobs is the maximum number of concurrently running workers; in this case, it is set to -1 which implies that all CPUs are used.

```
NB = GaussianNB()

TRAINING RESULTS:

Accuracy Score: 0.84

Classification Report:
              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      0.79      0.80      1061
weighted avg       0.83      0.84      0.83      1061


Confusion Matrix:
[[211  96]
 [ 79 675]]

Average Accuracy:     0.83

Standard Deviation:   0.0349
```

Table 19 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Training data

```
TEST RESULTS:

Accuracy Score: 0.82

Classification Report:
              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


Confusion Matrix:
[[112  41]
 [ 40 263]]
```

Table 19 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 0.888
AUC for the Test Data: 0.876
AUC for the Training Data: 0.888
AUC for the Test Data: 0.876
```

Table 20 Accuracy Score and AUC score on train and test data

Validness of Model - Naïve Bayes

1.The Train and test results are very close to each other with difference of less tha 2.5% , this can be considered as good case of classification.
2.The accuracy score is also pretty good at 84% and 82% on test data.
3.The variation in precision are minimum while there is slight variation in Recall and F1 score on train and test but seems consistent for both train and test set. 4.In AUC score for both train(89%) and test data(88%) are close to each other also.
5.In confusion matrix the the True positives are 54& and True Negatives are 24% whereas 8% are predict as positive and 9% are predicted as Negative which is not in actual.
6.Hence there is No case of over fitting or Under fitting of data.
7.The model would be stable and will perform good.
8.As per this model can predict with 82% of accuracy that Labour party will get the edge over conservatives.

1.6) Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.

Tuning of Models :Hyper-Parameters Tuning

The machine learning models are like a vote. There are some default parameter values for this vote, which we can tune or change the learning rate of the algorithm and get a better model. This is known as Hyper-Parameter Tuning

So based on the above given acuracy result will performance (Grid search as instructed)

Logistic Regression

LDA

KNN

Gradient boost Classifier

adaaBoost Classifier

Parameters

Just a quick summary of the parameters that we will be listing here for completeness,

n_jobs : Number of cores used for the training process. If set to -1, all cores are used.

n_estimators : Number of classification trees in learning model (set to 10 per default)

max_depth : Maximum depth of tree, or how much a node should be expanded. if set to too high a number would run the risk of overfitting as one would be growing the tree too deep.

verbose : Controls whether want to output any text during the process. A value of 0 suppresses all text while a value of 3 outputs the tree learning process at every iteration.

1.Tuning of Logistic Regression Models by Grid search

Grid search is best described as exhaustive guess and check.
We have a problem: In our case which the data is not too much (1526 rows) we use different range of "C" to reduce the "Overfitting". A high C means "Trust this training data a lot", while a low value says "This data may not be fully representative of the real-world data.

find the hyperparameters that result in the best cross validation score, and a set of values to try in the hyperparameter grid - the domain. The grid search method for finding the answer is to try all combinations of values in the domain and hope that the best combination is in the grid (in reality, we

will never know if we found the best settings unless we have an infinite hyperparameter grid which would then require an infinite amount of time to run).

Grid search suffers from one limiting problem: it is extremely computationally expensive because we have to perform cross validation with every single combination of hyperparameters in the grid! Let's see how many total hyperparameter settings there are in our simple little grid we developed.

Application:
If we run this on the entire dataset take these functions and put them in a script. However, I would advise against using grid search unless we have a very small hyperparameter grid because this is such as exhaustive method.

LR param grid choosing criteria:

1.Logistic Regression requires two parameters 'C' and 'penalty' to be optimised by GridSearchCV. So we have set these two parameters as a list of values form which GridSearchCV will select the best value of parameter.
2. Principal Component Analysis requires a parameter 'n_components' to be optimised. 'n_components' signifies the number of components to keep after reducing the dimension.

Best Param

{'C': 0.08858667904100823, 'max_iter': 1000, 'penalty': 'l2', 'solver': 'sag'}

```
TRAINING RESULTS:

Accuracy Score: 0.83

Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.63      0.69       307
           1       0.86      0.92      0.89       754

    accuracy                           0.83      1061
   macro avg       0.81      0.77      0.79      1061
weighted avg       0.83      0.83      0.83      1061


Confusion Matrix:
[[194 113]
 [ 63 691]]

Average Accuracy:     0.83

Standard Deviation:   0.0349
```

Table 21 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Train data

```
TEST RESULTS:

Accuracy Score: 0.8421

Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.72      0.75       153
           1       0.86      0.90      0.88       303

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


Confusion Matrix:
[[110  43]
 [ 29 274]]
```

Table 22 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 0.890
AUC for the Test Data: 0.880
Accuracy of logistic regression classifier on train set: 0.83
Accuracy of logistic regression classifier on test set: 0.83
```

Table 23 Accuracy Score and AUC score on train and test data

Validness of Model: Logistic Regression GridSearchCV

1.After the applying of gridsearch model the test accuracy has been gone up from 83% to 84% on test data.
2.The variation in precision and slight variation in Recall and F1 score but seems consistent for both train and test set.
3.In AUC score are 89% on train data while 88% on test data.
4.In confusion matrix classification ,True positives are 60 % and True Negatives are 23% whereas 8% are predict as positive and 14% are predicted as Negative which is not in actual.
5.Hence there is No case of over fitting or Under fitting of data.
6.The model would be stable and will perform good.
7.As per this model the Labour party will get the edge over conservatives by accuracy of 83%.
8.This model would be also stable and good for prediction.

2.Tuning of Model LDA by GridSearchCV

```
TRAINING RESULTS:

Accuracy Score: 0.83

Classification Report:
              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


Confusion Matrix:
[[200 107]
 [ 69 685]]

Average Accuracy:     0.83

Standard Deviation:   0.0349
```

Table 24 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Train data

```
TEST RESULTS:

Accuracy Score: 0.83

Classification Report:
              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


Confusion Matrix:
[[111  42]
 [ 34 269]]
```

Table 25 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 0.889
AUC for the Test Data: 0.888
Accuracy of Linear Discriminant (GS) classifier on train set: 0.83
Accuracy of Linear Discriminant (GS) classifier on test set: 0.83
```

Table 26 Accuracy Score and AUC score on train and test data

Validness of Model : LDA GridSearchCV

1.The accuracy score is also pretty good at 83% for train and test both.
2.The variation in precision and slight variation in Recall and F1 score but seems consistant for both train and test set.
3.In AUC score is almost 89% for both train and test data.
4.In confusion matrix classification ,True positives are 60 % and True Negatives are 23% whereas 8% are predict as positive and 14% are predict as Negative which is not in actual.
5.Hence there is No case of over fitting or Under fitting of data.
6.The model would be stable and will perform good.
7.As per this model the Labour party will get the edge over conservatives by accuracy of 83%.
8.This model would be also stable and good for prediction.

3.Tuning of Model KNN by GridSearchCV

Logic to Choose the Hyperparameter: KNN

The most important hyperparameter for KNN is the number of neighbours (n_neighbors).

Test values between at least 1 and 15, perhaps just the odd numbers. n_neighbors in [1 to 15] It may also be interesting to test different distance metrics (metric) for choosing the composition of the neighbourhood.

metric in ['euclidean', 'manhattan', 'minkowski']

It may also be interesting to test the contribution of members of the neighbourhoods via different weightings (weights).

weights in ['uniform', 'distance']

```
GridSearch best score 0.9440000000000002
GridSearch best params {'metric': 'euclidean', 'n_neighbors': 7, 'weights'
: 'distance'}
GridSearch best estimator KNeighborsClassifier(algorithm='brute', metric='
euclidean', n_neighbors=7,
                     weights='distance')


TRAINING RESULTS:
Accuracy Score: 1.00
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       307
           1       1.00      1.00      1.00       754
```

```
    accuracy                           1.00      1061
   macro avg       1.00      1.00      1.00      1061
weighted avg       1.00      1.00      1.00      1061

Confusion Matrix:
[[307   0]
 [  0 754]]

Average Accuracy:     0.83
Standard Deviation:   0.0349
```

Table 27 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Train data

```
TEST RESULTS:

Accuracy Score: 0.82

Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.65      0.70       153
           1       0.84      0.90      0.87       303

    accuracy                           0.82       456
   macro avg       0.80      0.78      0.79       456
weighted avg       0.81      0.82      0.81       456


Confusion Matrix:
[[ 99  54]
 [ 29 274]]
```

Table 28 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 1.000
AUC for the Test Data: 0.868
Accuracy of KNeighbors(GS) classifier on train set: 1.00
Accuracy of KNeighbors(GS) classifier on test set: 0.82
```

Table 29 Accuracy Score and AUC score on train and test data

Validness of Model - KNN GridSearchCV

1.The Train and test results are very far to each other with difference of more than 17%, this cannot be considered as good for classification.
2.Large variation in all parameters of scores.
3.Although The model accuracy score is also pretty good at 82% on test and 100% on train.
4.Hence the case can be considered as case of OVER fitting .
5.It would be advisable to DROP this model although the Scaling is not done .

4.Tuning of Model Naive Bayes by GridSearchCV

Best fit at:

```
Fitting 30 folds for each of 100 candidates, totalling 3000 fits
{'var_smoothing': 0.657933224657568}
```

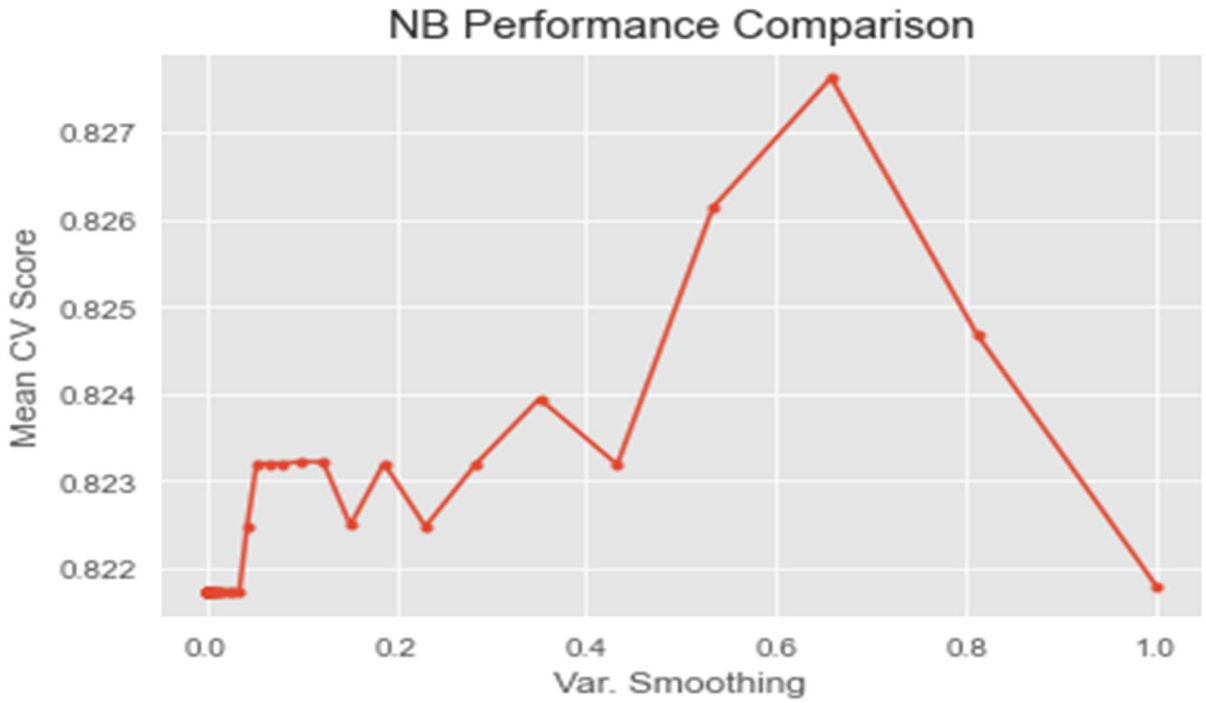


Figure 8 NB performance Comparison

```
TRAINING RESULTS:

Accuracy Score: 0.71

Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       307
           1       0.71      1.00      0.83       754

    accuracy                           0.71      1061
   macro avg       0.36      0.50      0.42      1061
weighted avg       0.51      0.71      0.59      1061


Confusion Matrix:
```

```
[[  0 307]
 [  0 754]]

Average Accuracy:     0.83

Standard Deviation:   0.0349
```

Table 30 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Train data

```
TEST RESULTS:

Accuracy Score: 0.66

Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       153
           1       0.66      1.00      0.80       303

    accuracy                           0.66       456
   macro avg       0.33      0.50      0.40       456
weighted avg       0.44      0.66      0.53       456


Confusion Matrix:
[[  0 153]
 [  0 303]]
```

Table 31 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 0.428
AUC for the Test Data: 0.414
Accuracy of Naïve Bayes Model (GS) classifier on train set: 0.71
Accuracy of Naïve Bayes Model (GS) classifier on test set: 0.66
```

Table 32 Accuracy Score and AUC score on train and test data

Validness of Model - Naïve Bayes

1.The Train and test results are very close to each other with difference of almost 10% ,
2.The accuracy score is pretty low to 71% and 66% on test data.
3.The variation in precision are minimum while there is slight variation in Recall and F1 score on train and test but seems consistent for both train and test set. 4.In AUC score for train(42%) and test data(41%) which is less than of 50%.
5.In confusion matrix the the all are classified as True positives and predicted positive No data are predict either Negative .
6.The model would be NOT stable and better to drop.

Bagging

Bagging performs best with algorithms that have high variance. A popular example are decision trees, often constructed without pruning.

Bagging Classifier:

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. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

Bagging is a special case of the model averaging approach, in case of regression problem we take mean of the output and in case of classification we take the majority vote.

This algorithm encompasses several works from the literature. When random subsets of the dataset are drawn as random subsets of the samples, then this algorithm is known as Pasting. If samples are drawn with replacement, then the method is known as Bagging. When random subsets of the dataset are drawn as random subsets of the features, then the method is known as Random Subspaces. Finally, when base estimators are built on subsets of both samples and features, then the method is known as Random Patches.

```
TRAINIG RESULTS:
==============================
CONFUSION MATRIX:
[[307   0]
 [  0 754]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    1.00   1.00      1.00       1.00          1.00
recall       1.00   1.00      1.00       1.00          1.00
f1-score     1.00   1.00      1.00       1.00          1.00
support    307.00 754.00      1.00    1061.00       1061.00
TESTING RESULTS:
==============================
CONFUSION MATRIX:
[[104  49]
 [ 26 277]]
ACCURACY SCORE:
0.8355
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.80   0.85      0.84       0.82          0.83
recall       0.68   0.91      0.84       0.80          0.84
f1-score     0.73   0.88      0.84       0.81          0.83
support    153.00 303.00      0.84     456.00        456.00
```

Table 33 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 1.000
AUC for the Test Data: 0.896
Accuracy of Bagging(Random forest) Classifier on train set:1.00
Accuracy of Bagging(Random forest on test  set:0.84
```

Table 34 Accuracy Score and AUC score on train and test data

Validness of Model - Random Forest(Bagging)

1.The Train and test results are very far to each other with difference of more than 17%,
this cannot be considered as good for classification.
2.Large variation in all parameters of scores. 3.Although The model accuracy score is also pretty good at 82% on test and 100% on train.
3.Hence the case can be considered as case of OVER fitting .
4.It would be advisable to DROP this model although the Scaling is not done

Boosting

1.Boosting ensemble algorithms creates a sequence of models that attempt to correct the mistakes of the models before them in the sequence.
2.In case of boosting, machine learning models are used one after the other and the predictions made by first layer models are used as input to next layer models. The last layer of models will use the predictions from all previous layers to get the final predictions.
3.So boosting enables each subsequent model to boost the performance of the previous one by overcoming or reducing the error of the previous model.
4.Boosting is more helpful if we have biased base models.
5.Boosting can be used to solve regression and classification problems. Here we have to solve the classification problem.

We can construct an AdaBoost model and Gradient Boost model for classification using the AdaBoostClassifier and Gradient Boost Classifier class.

AdaBoostClassifier: An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. AdaBoost was perhaps the first successful boosting ensemble algorithm. It generally works by weighting instances in the dataset by how easy or difficult they are to classify, allowing the algorithm to pay or or less attention to them in the construction of subsequent models.

Gradient Boost Classifier: Gradient Boosting Regression is an analytical technique that is designed to explore the relationship between two or more variables (X, and y). Its analytical output identifies important factors (Xi) impacting the dependent variable (y) and the nature of the relationship between each of these factors and the dependent variable.

The main differences, the Gradient Boosting is a generic algorithm to find approximate solutions to the additive modelling problem, while AdaBoost can be seen as a special case with a particular loss function. Hence, Gradient Boosting is much more flexible.

Model AdaBoost Classifier

```
TRAINIG RESULTS:
================================
CONFUSION MATRIX:
[[209  98]
 [ 62 692]]
ACCURACY SCORE:
0.8492
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.77   0.88      0.85       0.82          0.85
recall       0.68   0.92      0.85       0.80          0.85
f1-score     0.72   0.90      0.85       0.81          0.85
support    307.00 754.00      0.85    1061.00       1061.00
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[104  49]
 [ 37 266]]
ACCURACY SCORE:
0.8114
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.74   0.84      0.81       0.79          0.81
recall       0.68   0.88      0.81       0.78          0.81
f1-score     0.71   0.86      0.81       0.78          0.81
support    153.00 303.00      0.81     456.00        456.00
```

Table 35 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 0.909
AUC for the Test Data: 0.881
Accuracy of Addaboost Classifier on train set: 0.85
Accuracy of Addaboost Classifier on test set: 0.81
```

Table 34 Accuracy Score and AUC score on train and test data

Validness of Model - Adaboost classifier

1.The Train(85%) and test(81%) results are very close to each other with difference of 5% ,
2.The accuracy score is also pretty good at 85% and 81% on train and test data respectively.
3.The variation in precision ,Recall and F1 score are aprox 5% but seems consistent for both train and test set.
4.In AUC score for both train(90%) and test data(88%) are close to each other also.

5.In confusion matrix the True positives are 58% and True Negatives are 23% whereas 10% are predict as positive and 8% are predict as Negative which is not in actual.
6.Hence there is No case of over fitting or Under fitting of data.
7.The model would be stable and will perform good.
8.As per this model can predict with 81% of accuracy that Labour party will get the edge over conservatives
Gradient Boost Classifier

Model Gradient Boost Classifier

```
TRAINIG RESULTS:
==============================
CONFUSION MATRIX:
[[214  93]
 [ 49 705]]
ACCURACY SCORE:
0.8662
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.81   0.88      0.87       0.85          0.86
recall       0.70   0.94      0.87       0.82          0.87
f1-score     0.75   0.91      0.87       0.83          0.86
support    307.00 754.00      0.87    1061.00       1061.00
TESTING RESULTS:
==============================
CONFUSION MATRIX:
[[ 99  54]
 [ 20 283]]
ACCURACY SCORE:
0.8377
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   0.83   0.84      0.84       0.84          0.84
recall      0.65   0.93      0.84       0.79          0.84
f1-score    0.73   0.88      0.84       0.81          0.83
support   153.00 303.00      0.84     456.00        456.00
```

Table 36 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

```
AUC for the Training Data: 0.925
AUC for the Test Data: 0.900
Accuracy of Gradient boost Classifier on train set: 0.87
Accuracy of Gradient boost Classifier on test set: 0.84
```

Table 37 Accuracy Score and AUC score on train and test data

Validness of Model -Gradient boost Classifier

1.The Train and test results are very close to each other with difference of less than 4%
2.The accuracy score is also pretty good at 87% and 84% on train and test data respectively.
3.The variation in precision are minimum while there is slight variation in Recall and F1 score but seems consistent for both train and test set.
4.In AUC score for both train(89%) and test data(87%) are close to each other also.
5.In confusion matrix the True positives are 62% and True Negatives are 22% whereas 11% are predict as positive and 4% are predicted as Negative which is not in actual.
6.Hence there is No case of over fitting or Under fitting of data.
7.The model would be stable and will perform good.
8.As per this model can predict with 84% of accuracy that Labour party will get the edge over conservatives.

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, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)

Accuracy, Confusion Matrix, Plot ROC curve

Accuracy is a measure of the difference between the mean value or experimental value of a set of measurements and the true value. Therefore, Accuracy = Mean Value – True Value The smaller the difference between the mean value and true value, the larger is the accuracy

f1-score:\ f1-score is the weighted harmonic mean of precision and recall. The best possible f1-score would be 1.0 and the worst would be 0.0. f1-score is the harmonic mean of precision and recall. So, f1-score is always lower than accuracy measures as they embed precision and recall into their computation. The weighted average of f1-score should be used to compare classifier models, not global accuracy.

Precision:
Precision can be defined as the percentage of correctly predicted positive outcomes out of all the predicted positive outcomes. It can be given as the ratio of true positives (TP) to the sum of true and false positives (TP + FP).

So, Precision identifies the proportion of correctly predicted positive outcome. It is more concerned with the positive class than the NEGATIVE class.

Mathematically, precision can be defined as the ratio of TP to (TP + FP).

Recall:
Recall can be defined as the percentage of correctly predicted positive outcomes out of all the actual positive outcomes. It can be given as the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN). Recall is also called Sensitivity.

Recall identifies the proportion of correctly predicted actual POSITIVES.

Mathematically, recall can be given as the ratio of TP to (TP + FN)

Confusion matrix :
A confusion matrix is a tool for summarizing the performance of a classification algorithm. A confusion matrix will give us a clear picture of classification model performance and the types of errors produced by the model. It gives us a summary of correct and incorrect predictions broken down by each category. The summary is represented in a tabular form.

Four types of outcomes are possible while evaluating a classification model performance. These four outcomes are described below:-

True Positives (TP) – True Positives occur when we predict an observation belongs to a certain class and the observation actually belongs to that class.

True Negatives (TN) – True Negatives occur when we predict an observation does not belong to a certain class and the observation actually does not belong to that class.

False Positives (FP) – False Positives occur when we predict an observation belongs to a certain class but the observation actually does not belong to that class. This type of error is called Type I error.

False Negatives (FN) – False Negatives occur when we predict an observation does not belong to a certain class but the observation actually belongs to that class. This is a very serious error and it is called Type II error

ROC - AUC ROC Curve Another tool to measure the classification model performance visually is ROC Curve. ROC Curve stands for Receiver Operating Characteristic Curve. An ROC Curve is a plot which shows the performance of a classification model at various classification threshold levels.

The ROC Curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels.

True Positive Rate (TPR) is also called Recall. It is defined as the ratio of TP to (TP + FN).

False Positive Rate (FPR) is defined as the ratio of FP to (FP + TN).

In the ROC Curve, we will focus on the TPR (True Positive Rate) and FPR (False Positive Rate) of a single point. This will give us the general performance of the ROC curve which consists of the

TPR and FPR at various threshold levels. So, an ROC Curve plots TPR vs FPR at different classification threshold levels. If we lower the threshold levels, it may result in more items being classified as positve. It will increase both True Positives (TP) and False Positives (FP).

Comparing ROC AUC curve and Scores

Average ROC_AUC curve for each modeling

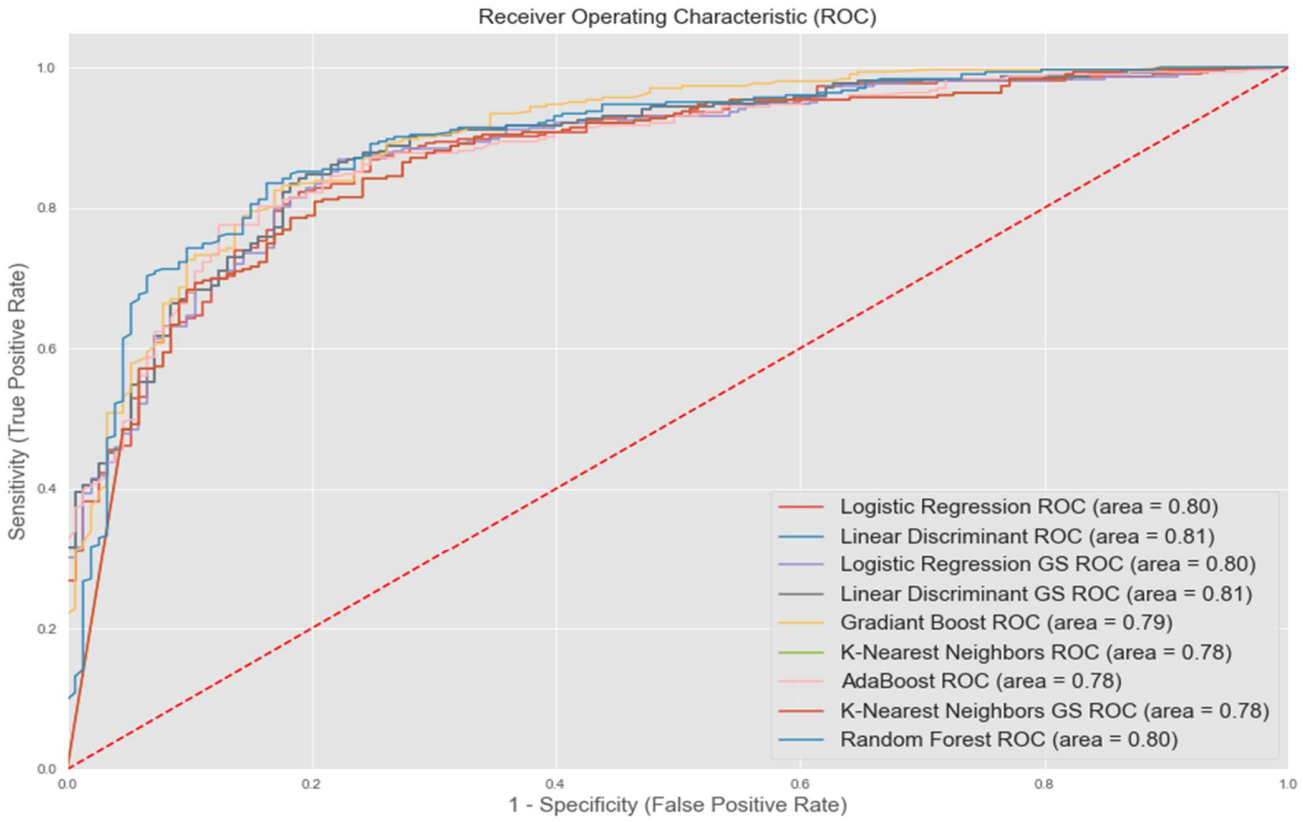


Figure 9 Comparing ROC AUC curve and Scores

Here the comparison of Average score of all models has been depicted while the accurate score and curve has been depicted in concerned models.

Discussions score

The scale of AUC over model stability
0.5 = No discrimination
0.5-0.7 = Poor discrimination
0.7-0.8 = Acceptable discrimination
0.8-0.9= Excellent discrimination
0.9-1 = Outstanding discrimination

By these standards, a model with an AUC score below 0.7 would be considered poor and anything higher would be considered acceptable or better.
In our all models the AUC scores are above 0.7 .

Model comparison Discussion and choosing best Model

All models have accuracy more than 80% on test data and AUC score also above the .70.In few models there are large variation among the train and test accuracy. One Model only predicting the positive results.
The Model those having more than 10% of variation in accuracy also showing the variation as other parameters too.
Among all the Three Model can be identify for classification of this case are:\

1.Logistic Regression
2.Naïve Bayes Model
3.Gradient Boosting Classifier\

Among these three the Best Model for the prediction in this problem statement is Gradient Boosting Classifier
because:\

1.The Train and test results are very close to each other with difference of less than 4%
2.The accuracy score is also pretty good at 87% and 84% on train and test data respectively.
3.The variation in precision are minimum while there is slight variation in Recall and F1 score but seems consistent for both train and test set.
4.In AUC score for both train(89%) and test data(87%) are close to each other also and support this model.
5.In confusion matrix the True positives are 62% and True Negatives are 22% whereas 11% are predict as positive and 5% are predicted as Negative which is not in actual.
Total Positive : 84%
Total Negative: 16%
AUC score and ROC_AUC curve also advocated to this model.
Precision : 84 on test data while 88 on train data
F1 Score: 88 on test and 91 on train data set.
The Total Positive will help the news channels CNBE who wants to analyse recent elections on exit poll.
6.As per Gradient Boosting Classifier model can predict with 84% of accuracy that Labour party will get the edge over conservatives.

Best Model: Gradient Boosting Classifier

```
TRAINIG RESULTS:
================================
CONFUSION MATRIX:
[[214  93]
 [ 49 705]]
ACCURACY SCORE:
0.8662
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.81   0.88      0.87       0.85          0.86
recall       0.70   0.94      0.87       0.82          0.87
f1-score     0.75   0.91      0.87       0.83          0.86
support    307.00 754.00      0.87    1061.00       1061.00
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[ 99  54]
 [ 20 283]]
ACCURACY SCORE:
0.8377
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   0.83   0.84      0.84       0.84          0.84
recall      0.65   0.93      0.84       0.79          0.84
f1-score    0.73   0.88      0.84       0.81          0.83
support   153.00 303.00      0.84     456.00        456.00
```

Table 38 Accuracy Score, Classification report confusion Matrix & Standard Deviation of Test data

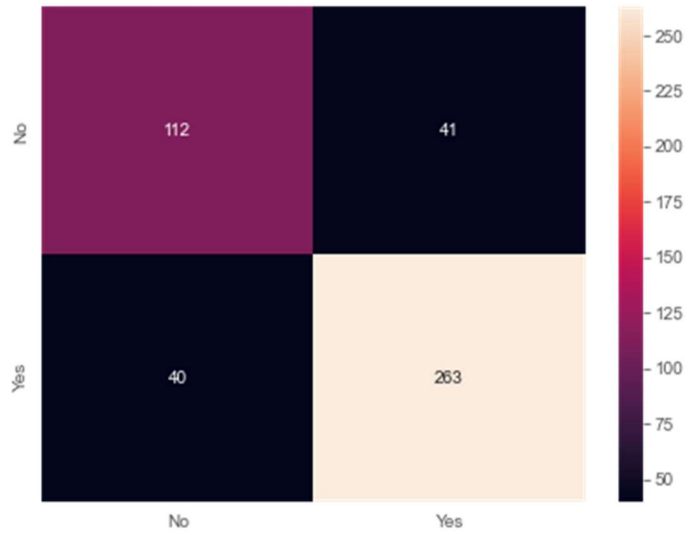


Figure 9 Classification Matrix of Gradient Boost Classifier

```
AUC for the Training Data: 0.925
AUC for the Test Data: 0.900
Accuracy of Gradient boost Classifier on train set: 0.87
Accuracy of Gradient boost Classifier on test set: 0.84
```

Table 39 Accuracy Score and AUC score on train and test data

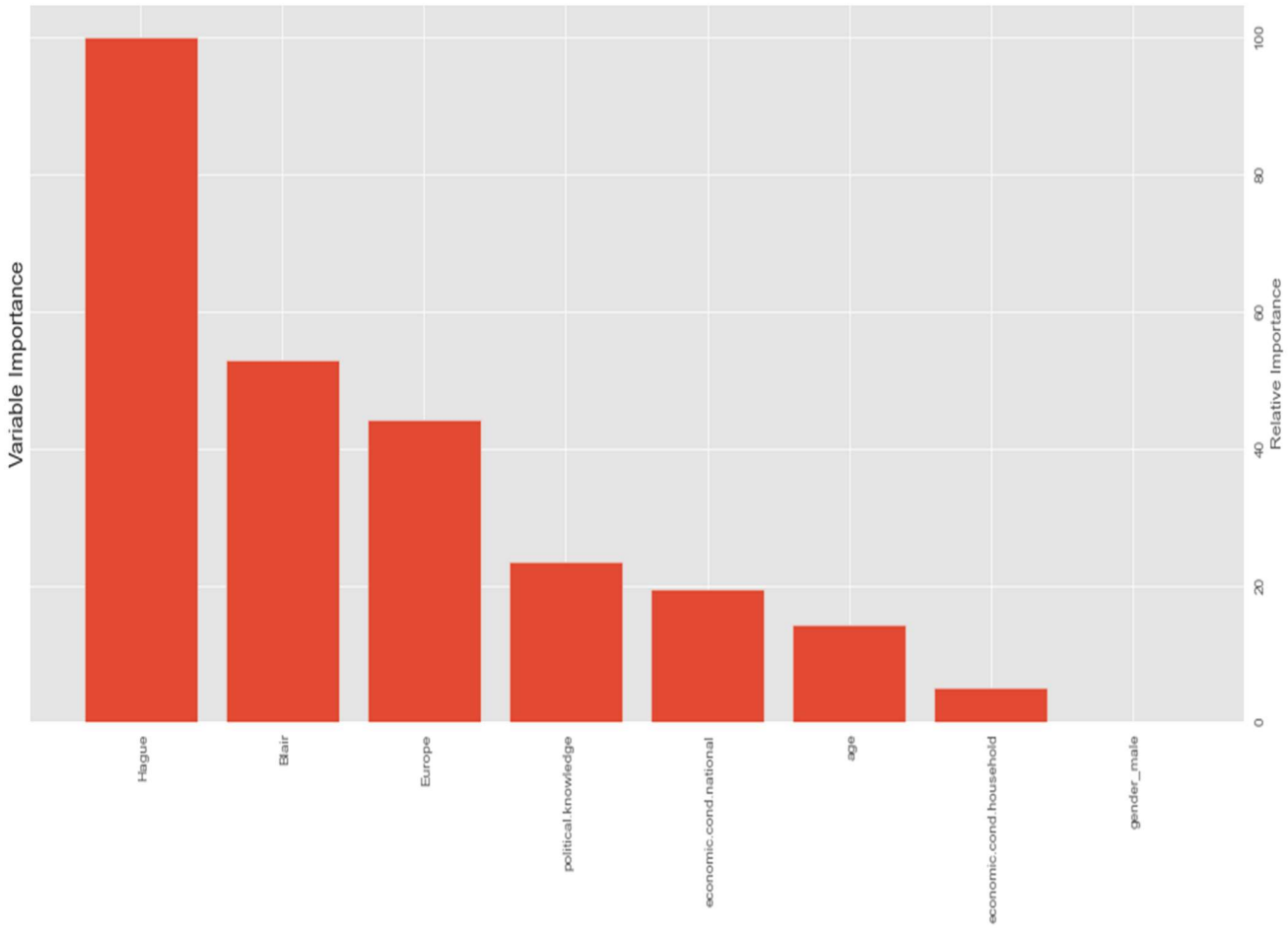


Figure 9 Feature importance of Gradient Boost Classifier

The Hague are most important feature followed by Blair and Europe. These Three factors may play the deciding role in win of Labour party. Whereas Gender not seems playing a large role in priority list. But person having Political knowledge has better importance than Economic national and age factor.

1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

Recommendations:

After 1526 exit poll data analysis, its coming out that the Labour party is going to taste the win as per data and Modelling prediction with accuracy of almost 84%. Because:

1.The population having political knowledge (especially class2) has high orientation followed by those population having NO political Knowledge towards Labour party.

2.Population with attitudes toward European integration has high orientation for labour party rather than Conservatives where has the conservative party has also very high positive impact in this class.

3.Leadership Impact of labour party leadership has great impression over population which will propagate to get the election win.

4.Both Male and population are supporting the Labour party with open heart, while female population having more in population over man.

5.All type economic condition of Household has large orientation towards Labour party to win.

6.Labour party has to work upon the population who are having of Attitudes toward European integration. This factor seems strong for conservative party as well.

7. Lower Ranking in Assessment of the Conservative leader with give top most impact on win of Labour party. The anti-incumbency can be seen or damage of Conservative leader my other means will giving edge to Labour party.

8.Labour party will win by huge margin over conservative party the election as per given data set of exit poll

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:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

(Hint: use .words(), .raw(), .sent() for extracting counts

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

Finding the Number of Words in each speech and combined speech

```
Total number of characters in speech of Roosevelt are: 7571
Total number of characters in speech of Kennedy are: 7618
Total number of characters in speech of Nixon are: 9991
Grand Total number of characters in all speeches are: 25180
```

Finding the Number of Words in each speech and combined speech

```
Total number of words in speech of Roosevelt are: 1536
Total number of words in speech of Kennedy are: 1546
Total number of words in speech of Nixon are: 2028
Grand Total number of words in all speeches are: 5110
```

Finding the Number of Sentences in each speech and combined speech

```
Total number of sentences in speech of Roosevelt are: 68
Total number of sentences in speech of Kennedy are: 52
Total number of sentences in speech of Nixon are: 69
Grand Total number of words in all speeches are: 189
```

2.2) Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.¶

Finding the Number of Words in each speech and combined speech

1941-Roosevelt

```
national day inauguration since 1789, people renewed sense dedication Unit
ed States. Washington's day task people create weld together nation. Linco
ln's day task people preserve Nation disruption within. day task people sa
ve Nation institutions disruption without. us come time, midst swift happe
nings, pause moment take stock -- recall place history been, rediscover ma
y be. not, risk real peril inaction. Lives nations determined count years,
lifetime human spirit. life man three-score years ten: little more, little
less. life nation fullness measure live. men doubt this. men believe democ
racy, form Government frame life, limited measured kind mystical artificia
l fate that, unexplained reason, tyranny slavery become surging wave futur
e -- freedom ebbing tide. Americans know true. Eight years ago, life Repub
lic seemed frozen fatalistic terror, proved true. midst shock -- acted. ac
```

ted quickly, boldly, decisively. later years living years -- fruitful year
s people democracy. brought us greater security and, hope, better understa
nding life's ideals measured material things. vital present future experie
nce democracy successfully survived crisis home; put away many evil things
; built new structures enduring lines; and, all, maintained fact democracy
. action taken within three-way framework Constitution United States. coor
dinate branches Government continue freely function. Bill Rights remains i
nviolate. freedom elections wholly maintained. Prophets downfall American
democracy seen dire predictions come naught. Democracy dying. know seen re
vive--and grow. know cannot die -- built unhampered initiative individual
men women joined together common enterprise -- enterprise undertaken carri
ed free expression free majority. know democracy alone, forms government,
enlists full force men's enlightened will. know democracy alone constructe
d unlimited civilization capable infinite progress improvement human life.
know because, look surface, sense still spreading every continent -- human
e, advanced, end unconquerable forms human society. nation, like person, b
ody--a body must fed clothed housed, invigorated rested, manner measures o
bjectives time. nation, like person, mind -- mind must kept informed alert
, must know itself, understands hopes needs neighbors -- nations live with
in narrowing circle world. nation, like person, something deeper, somethin
g permanent, something larger sum parts. something matters future -- calls
forth sacred guarding present. thing find difficult -- even impossible --
hit upon single, simple word. yet understand -- spirit -- faith America. p
roduct centuries. born multitudes came many lands -- high degree, mostly p
lain people, sought here, early late, find freedom freely. democratic aspi
ration mere recent phase human history. human history. permeated ancient l
ife early peoples. blazed anew middle ages. written Magna Charta. Americas
impact irresistible. America New World tongues, peoples, continent new-fou
nd land, came believed could create upon continent new life -- life new fr
eedom. vitality written Mayflower Compact, Declaration Independence, Const
itution United States, Gettysburg Address. first came carry longings spiri
t, millions followed, stock sprang -- moved forward constantly consistentl
y toward ideal gained stature clarity generation. hopes Republic cannot fo
rever tolerate either undeserved poverty self-serving wealth. know still f
ar go; must greatly build security opportunity knowledge every citizen, me
asure justified resources capacity land. enough achieve purposes alone. en
ough clothe feed body Nation, instruct inform mind. also spirit. three, gr
eatest spirit. Without body mind, men know, Nation could live. spirit Amer
ica killed, even though Nation's body mind, constricted alien world, lived
on, America know would perished. spirit -- faith -- speaks us daily lives
ways often unnoticed, seem obvious. speaks us Capital Nation. speaks us pr
ocesses governing sovereignties 48 States. speaks us counties, cities, tow
ns, villages. speaks us nations hemisphere, across seas -- enslaved, well
free. Sometimes fail hear heed voices freedom us privilege freedom old, ol
d story. destiny America proclaimed words prophecy spoken first President
first inaugural 1789 -- words almost directed, would seem, year 1941: "The
preservation sacred fire liberty destiny republican model government justl
y considered deeply, finally, staked experiment intrusted hands American p
eople." lose sacred fire--if let smothered doubt fear -- shall reject dest
iny Washington strove valiantly triumphantly establish. preservation spiri
t faith Nation does, will, furnish highest justification every sacrifice m
ay make cause national defense. face great perils never encountered, stron
g purpose protect perpetuate integrity democracy. muster spirit America, f

aith America. retreat. content stand still. Americans, go forward, service country, God.
Old length: 7571
New length: 4905

1961-Kennedy
Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, observe today victory party, celebration freedom -- symbolizing end, well beginning -- signifying renewal, well change. sworn Almighty God solemn oath forebears l prescribed nearly century three quarters ago. world different now. man holds mortal hands power abolish forms human poverty forms human life. yet revolutionary beliefs forebears fought still issue around globe -- belief rights man come generosity state, hand God. dare forget today heirs first revolution. Let word go forth time place, friend foe alike, torch passed new generation Americans -- born century, tempered war, disciplined hard bitter peace, proud ancient heritage -- unwilling witness permit slow undoing human rights Nation always committed, committed today home around world. Let every nation know, whether wishes us well ill, shall pay price, bear burden, meet hardship, support friend, oppose foe, order assure survival success liberty. much pledge -- more. old allies whose cultural spiritual origins share, pledge loyalty faithful friends. United, little cannot host cooperative ventures. Divided, little -- dare meet powerful challenge odds split asunder. new States welcome ranks free, pledge word one form colonial control shall passed away merely replaced far iron tyranny. shall always expect find supporting view. shall always hope find strongly supporting freedom -- remember that, past, foolishly sought power riding back tiger ended inside. peoples huts villages across globe struggling break bonds mass misery, pledge best efforts help help themselves, whatever period required -- Communists may it, seek votes, right. free society cannot help many poor, cannot save rich. sister republics south border, offer special pledge -- convert good words good deeds -- new alliance progress -- assist free men free governments casting chains poverty. peaceful revolution hope cannot become prey hostile powers. Let neighbors know shall join oppose aggression subversion anywhere Americas. let every power know Hemisphere intends remain master house. world assembly sovereign states, United Nations, last best hope age instruments war far outpaced instruments peace, renew pledge support--to prevent becoming merely forum invective -- strengthen shield new weak -- enlarge area writ may run. Finally, nations would make adversary, offer pledge request: sides begin anew quest peace, dark powers destruction unleashed science engulf humanity planned accidental self-destruction. dare tempt weakness. arms sufficient beyond doubt certain beyond doubt never employed. neither two great powerful groups nations take comfort present course -- sides overburdened cost modern weapons, rightly alarmed steady spread deadly atom, yet racing alter uncertain balance terror stays hand mankind's final war. let us begin anew -- remembering sides civility sign weakness, sincerity always subject proof. Let us never negotiate fear. let us never fear negotiate. Let sides explore problems unite us instead belaboring problems divide us. Let sides, first time, formulate serious precise proposals inspection control arms -- bring absolute power destroy nations absolute control nations. Let sides seek invoke wonders science instead terrors. Together let us explore stars, conquer deserts, eradicate disease, tap ocean depths, encourage arts commerce. Let sides un

ite heed corners earth command Isaiah -- "undo heavy burdens ... let oppre ssed go free." beachhead cooperation may push back jungle suspicion, let s ides join creating new endeavor, new balance power, new world law, strong weak secure peace preserved. finished first 100 days. finished first 1,000 days, life Administration, even perhaps lifetime planet. let us begin. han ds, fellow citizens, mine, rest final success failure course. Since countr y founded, generation Americans summoned give testimony national loyalty. graves young Americans answered call service surround globe. trumpet summo ns us -- call bear arms, though arms need; call battle, though embattled - - call bear burden long twilight struggle, year year out, "rejoicing hope, patient tribulation" -- struggle common enemies man: tyranny, poverty, dis ease, war itself. forge enemies grand global alliance, North South, East W est, assure fruitful life mankind? join historic effort? long history worl d, generations granted role defending freedom hour maximum danger. shrink responsibility -- welcome it. believe us would exchange places people gene ration. energy, faith, devotion bring endeavor light country serve -- glow fire truly light world. so, fellow Americans: ask country -- ask country. fellow citizens world: ask America you, together freedom man. Finally, whe ther citizens America citizens world, ask us high standards strength sacri fice ask you. good conscience sure reward, history final judge deeds, let us go forth lead land love, asking blessing help, knowing earth God's work must truly own.
Old length: 7618
New length: 5057

1973-Nixon

Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eis enhower, fellow citizens great good country share together: met four years ago, America bleak spirit, depressed prospect seemingly endless war abroad destructive conflict home. meet today, stand threshold new era peace world . central question us is: shall use peace? Let us resolve era enter postwa r periods often been: time retreat isolation leads stagnation home invites new danger abroad. Let us resolve become: time great responsibilities grea tly borne, renew spirit promise America enter third century nation. past y ear saw far-reaching results new policies peace. continuing revitalize tra ditional friendships, missions Peking Moscow, able establish base new dura ble pattern relationships among nations world. America's bold initiatives, 1972 long remembered year greatest progress since end World War II toward lasting peace world. peace seek world flimsy peace merely interlude wars, peace endure generations come. important understand necessity limitations America's role maintaining peace. Unless America work preserve peace, peac e. Unless America work preserve freedom, freedom. let us clearly understan d new nature America's role, result new policies adopted past four years. shall respect treaty commitments. shall support vigorously principle count ry right impose rule another force. shall continue, era negotiation, work limitation nuclear arms, reduce danger confrontation great powers. shall s hare defending peace freedom world. shall expect others share. time passed America make every nation's conflict own, make every nation's future respo nsibility, presume tell people nations manage affairs. respect right natio n determine future, also recognize responsibility nation secure future. Am erica's role indispensable preserving world's peace, nation's role indispe nsable preserving peace. Together rest world, let us resolve move forward beginnings made. Let us continue bring walls hostility divided world long,

build place bridges understanding -- despite profound differences systems government, people world friends. Let us build structure peace world weak safe strong -- respects right live different system -- would influence oth ers strength ideas, force arms. Let us accept high responsibility burden, gladly -- gladly chance build peace noblest endeavor nation engage; gladly , also, act greatly meeting responsibilities abroad remain great Nation, r emain great Nation act greatly meeting challenges home. chance today ever history make life better America -- ensure better education, better health , better housing, better transportation, cleaner environment -- restore re spect law, make communities livable -- insure God-given right every Americ an full equal opportunity. range needs great -- reach opportunities great -- let us bold determination meet needs new ways. building structure peace abroad required turning away old policies failed, building new era progres s home requires turning away old policies failed. Abroad, shift old polici es new retreat responsibilities, better way peace. home, shift old policie s new retreat responsibilities, better way progress. Abroad home, key new responsibilities lies placing division responsibility. lived long conseque nces attempting gather power responsibility Washington. Abroad home, time come turn away condescending policies paternalism -- "Washington knows bes t." person expected act responsibly responsibility. human nature. let us e ncourage individuals home nations abroad themselves, decide themselves. Le t us locate responsibility places. Let us measure others themselves. today offer promise purely governmental solution every problem. lived long false promise. trusting much government, asked deliver. leads inflated expectati ons, reduced individual effort, disappointment frustration erode confidenc e government people do. Government must learn take less people people them selves. Let us remember America built government, people -- welfare, work -- shirking responsibility, seeking responsibility. lives, let us ask -- g overnment me, myself? challenges face together, let us ask -- government h elp, help? National Government great vital role play. pledge Government ac t, act boldly lead boldly. important role every one us must play, individu al member community. day forward, let us make solemn commitment heart: bea r responsibility, part, live ideals -- together, see dawn new age progress America, together, celebrate 200th anniversary nation, proud fulfillment p romise world. America's longest difficult war comes end, let us learn deba te differences civility decency. let us reach one precious quality governm ent cannot provide -- new level respect rights feelings one another, new l evel respect individual human dignity cherished birthright every American. else, time come us renew faith America. recent years, faith challenged. ch ildren taught ashamed country, ashamed parents, ashamed America's record h ome role world. every turn, beset find everything wrong America little rig ht. confident judgment history remarkable times privileged live. America's record century unparalleled world's history responsibility, generosity, cr eativity progress. Let us proud system produced provided freedom abundance , widely shared, system history world. Let us proud four wars engaged cent ury, including one bringing end, fought selfish advantage, help others res ist aggression. Let us proud bold, new initiatives, steadfastness peace ho nor, made break-through toward creating world world known -- structure pea ce last, merely time, generations come. embarking today era presents chall enges great nation, generation, ever faced. shall answer God, history, con science way use years. stand place, hallowed history, think others stood m e. think dreams America, think recognized needed help far beyond order mak e dreams come true. Today, ask prayers years ahead may God's help making d ecisions right America, pray help together may worthy challenge. Let us pl

```
edge together make next four years best four years America's history, 200t
h birthday America young vital began, bright beacon hope world. Let us go
forward confident hope, strong faith one another, sustained faith God crea
ted us, striving always serve purpose.
Old length:  9991
New length:  6266
```

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

The Most frequent word in Roosevelt speech after the stopwords

```
 ('--', 22),
 ('know', 9),
 ('us', 8),
 ('life', 6)
```

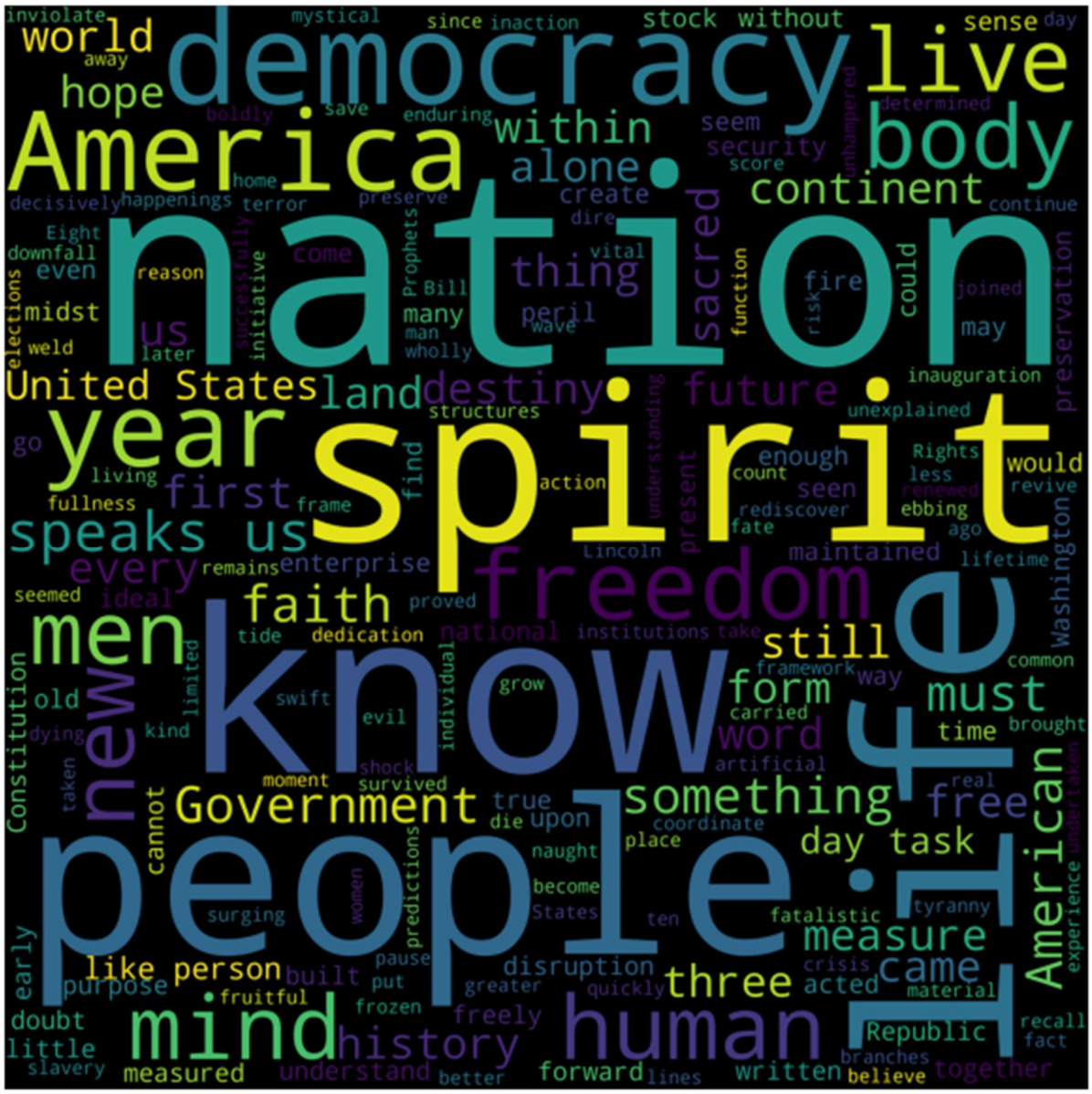
The Most frequent word in Kennedy speech after the stopwords

```
 ('--', 24),
 ('us', 11),
 ('Let', 8)
```

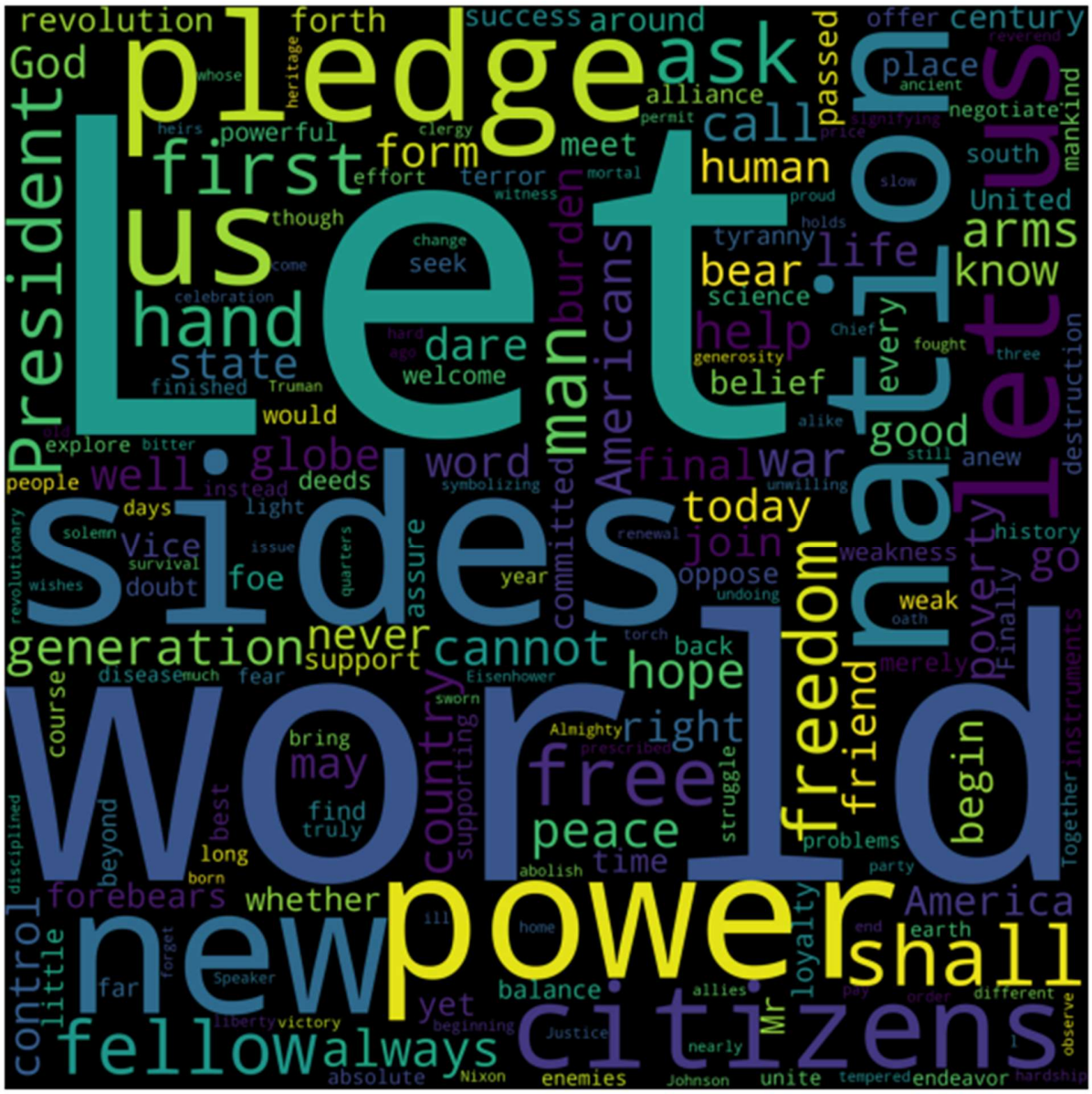
The Most frequent word in Nixon speech after the stopwords

```
 ('us', 25),
 ('--', 17),
 ('new', 15)
```

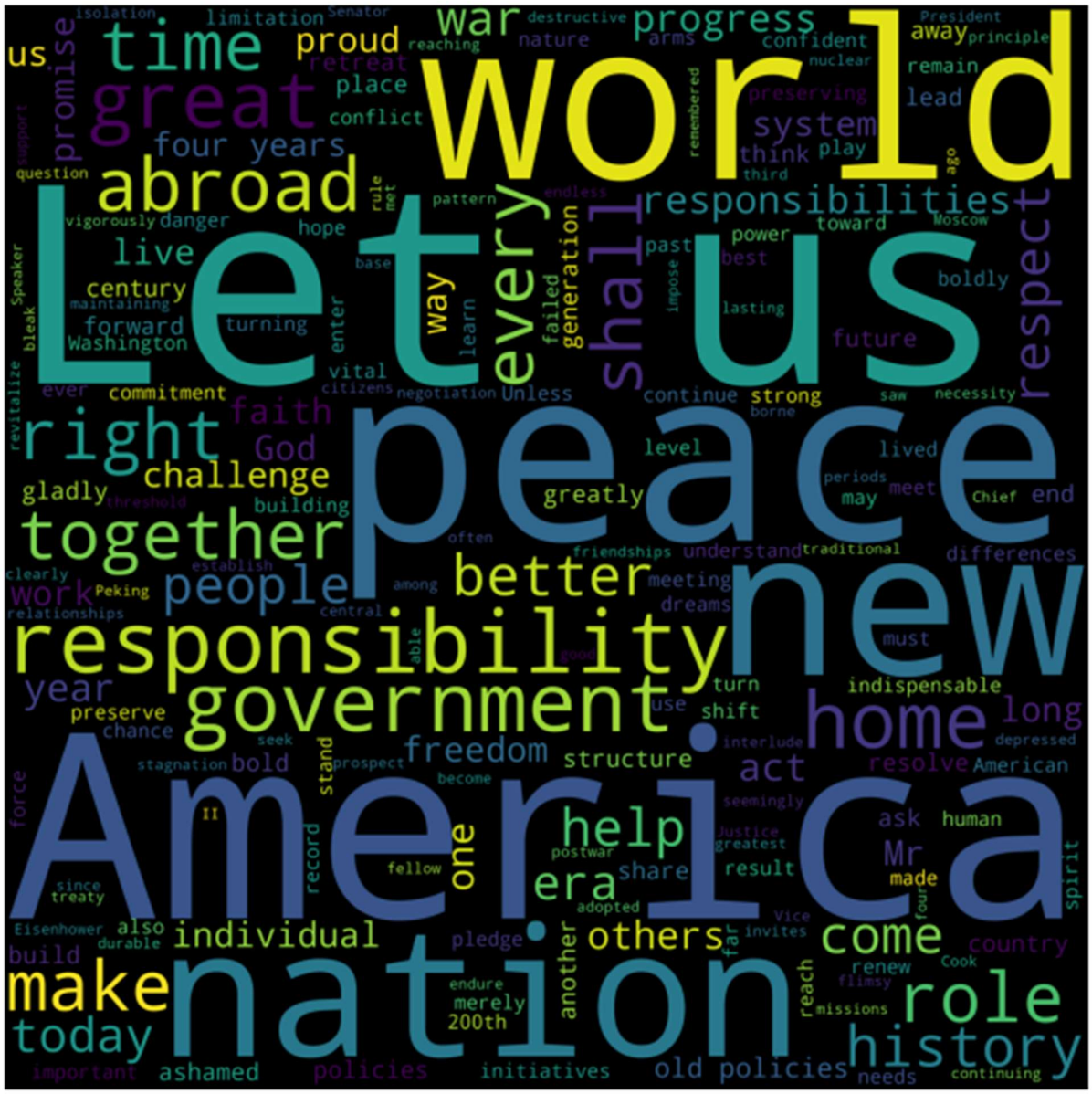
2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)



Word Cloud of Inaugural speech of Roosevelt (after cleaning)!!



Word Cloud Inaugural speech of Nixon (after cleaning)!!



Word Cloud Inaugural speech of Nixon (after cleaning)!!
