MACHINE LEARNING PROJECT REPORT

By Karthik Sreeram R

JANUARY 16, 2020

UNIVERSITY OF TEXAS AT AUSTIN AND GREAT LAKES

Purpose
This document is the business report for my final project in the subject “Machine Learning “ This document gives us a detailed explanation of various approaches used, their insight and inferences. Tools used analysis: Python and Jupiter notebook. Packages used: NumPy, pandas, seaborn, os, matplotlib, SciPy, stats model, sklearn and sweetviz

Contents

Problem 1: Logistic Regression , LDA KNN Model , Naive Bayes ,Bagging and Boosting 1

Business scenario 1

1.1) Read the dataset. Do the descriptive statistics and do null value condition check. 1

a.) Dataset Head 1

Inference: 1

b.) Summary of the dataset: 1

Inference: 2

c.) Type of the variables in dataset 2

Inference: 2

d.) Remove the Unnamed column. 3

e.) Dataset has any null values. 3

Inference: 3

f.) Data has any duplicities? 3

Inference: 3

g.) Information of the data after removing Outliers: 3

Inference: 4

h.) Data has any zero values. 4

Inference: 4

i.) Unique count of values in object variables. 4

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

a) Check for outlier 5

Inference: 5

b.) Any null values? 5

Inference: 5

c) Univariate Analysis 5

Inference: 6

d) Bi variate analysis (Between Target variable (Vote) and other variables) 7

Inference: 7

e.) Multivariate Analysis 8

1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not? , Data Split: Split the data into train and test (70:30) 10

a) Encode data 10

Head of dataset after encoding 10

Inference: 10

b) Data requires scaling? 10

Inference: 10
c) Data Split: 11
Inference: 11
1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis). Interpret the inferences of both models 11
A.) Logistic Regression 11
Logistic model score 11
Important attributes for Logistic regression: 11
B.) LDA 12
LDA model score 12
Important attributes using LDA: 12
Inference: 12
1.5) Apply KNN Model and Naïve Bayes Model and Interpret the inferences of both models 13
A.) KNN Model 13
KNN model score 13
B.) Naïve Bayes 13
Naïve Bayes model score 13
Inference: 13
1.6) Model Tuning, Bagging and Boosting . 13
A.) Random Forest is built for Bagging 14
B) Bagging 14
Bagging model score 14
Important attributes for Bagging: 14
c.) Ada Boosting 15
Ada boosting score 15
Important attributes for Ada Boosting: 15
d.) Gradient Boosting 15
Important attributes for Gradient Boosting: 16
Inference: 16
1.7) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized 17
a) Performance Metrics for Logistic Regression 17
Accuracy score for Training data 17
Accuracy score for Testing data 17
Confusion Matrix for Logistic Regression 17

Classification Report for Logistic Regression: 18
ROC curve and ROC_AUC score for Logistic Regression 18
b) Performance Metrics for LDA 19
Accuracy score for Training data 19
Accuracy score for Testing data 19
Confusion Matrix For LDA 19
Classification Report For LDA 19
ROC curve and ROC_AUC score for LDA 20
c) Performance Metrics for KNN 20
Accuracy score for Training data 20
Accuracy score for Testing data 20
Confusion Matrix For KNN 20
Classification Report For KNN 21
ROC curve and ROC_AUC score for KNN 21
d) Performance Metrics for Naïve Bayes 21
Accuracy score for Training data 21
Accuracy score for Testing data 21
Confusion Matrix For Naïve Bayes 22
Classification Report For Naïve Bayes 22
ROC curve and ROC_AUC score for Naïve Bayes 22
e) Performance Metrics for Bagging 23
Accuracy score for Training data 23
Accuracy score for Testing data 23
Confusion Matrix For Bagging 23
Classification Report For Bagging 23
ROC curve and ROC_AUC score for Bagging 24
f) Performance Metrics for ADA boosting 24
Accuracy score for Training data 24
Accuracy score for Testing data 24
Confusion Matrix For Ada Boosting 24
Classification Report For Ada boosting 25
ROC curve and ROC_AUC score for Ada boosting 25
g) Performance Metrics for Gradient boosting 25
Accuracy score for Training data 25
Accuracy score for Testing data 25
Confusion Matrix For gradient Boosting 26

Classification Report For gradient boosting 26
ROC curve and ROC_AUC score for gradient boosting 26
Compare all models on the basis of the performance metrics in a structured tabular manner. 27
Comparison in Table form: 27
Inference: 27
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. 28
Business Insights: 28
Recommendation: 28
Problem 2: Text and Sentimental Analytics 29
Business scenario 29
2.1) Find the number of characters, words and sentences for the mentioned documents. 29
2.2) Remove all the stop words from the three speeches. 29
a.) Sample of data after removing stop words 29
b.) President Roosevelt speech as text after removing stop words and punctuation and changing to lower case 29
c.) President Kennedy speech as text after removing stop words and punctuation and changing to lower case 30
d.) President Nixon speech as text after removing stop words and punctuation and changing to lower case 30
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 stop words) 31
2.4) Plot the word cloud of each of the three speeches. (after removing the stop words) 31
a.) President Roosevelt speech word cloud 31
b.) President Kennedy speech word cloud 32
c.) President Nixon speech word cloud 32

Problem 1: Logistic Regression , LDA KNN Model , Naive Bayes ,Bagging and Boosting

Business scenario

You are hired by one of the leading news channel 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. Do the descriptive statistics and do null value condition check.

a.) Dataset Head

	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

Inference:

Dataset has 10 columns.
The first column (Unnamed column :0) is of no use for analysis and can be removed.

b.) Summary of the dataset:

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

Inference:

Unnamed column can be ignored. Most of variables are having same scale. While looking at range of the values between minimum, 50 percentile and maximum, data seems to have no outliers

c.) Type of the variables in dataset

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

Inference:

Dataset has no null values.

Gender column and target column (vote) are of object data type i.e., contains strings value. Remining columns are of numerical datatype (integer).

Dataset has 1525 observations (Rows of data)

d.) Remove the Unnamed column.

Head of dataset after removing unwanted column:

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

e.) Dataset has any null values.

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

Inference:

Dataset has no null values.

f.) Data has any duplicities?

```
Number of Duplicates  8
```

Inference:

Dataset had 8 duplicated rows.

g.) Information of the data after removing Outliers:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1517 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1517 non-null   object
 1   age                     1517 non-null   int64
 2   economic.cond.national  1517 non-null   int64
 3   economic.cond.household 1517 non-null   int64
 4   Blair                   1517 non-null   int64
 5   Hague                   1517 non-null   int64
 6   Europe                  1517 non-null   int64
 7   political.knowledge     1517 non-null   int64
 8   gender                  1517 non-null   object
dtypes: int64(7), object(2)
memory usage: 158.5+ KB
```

Inference:

Removed duplicates and now data has 1517 rows.

h.) Data has any zero values.

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

Inference:

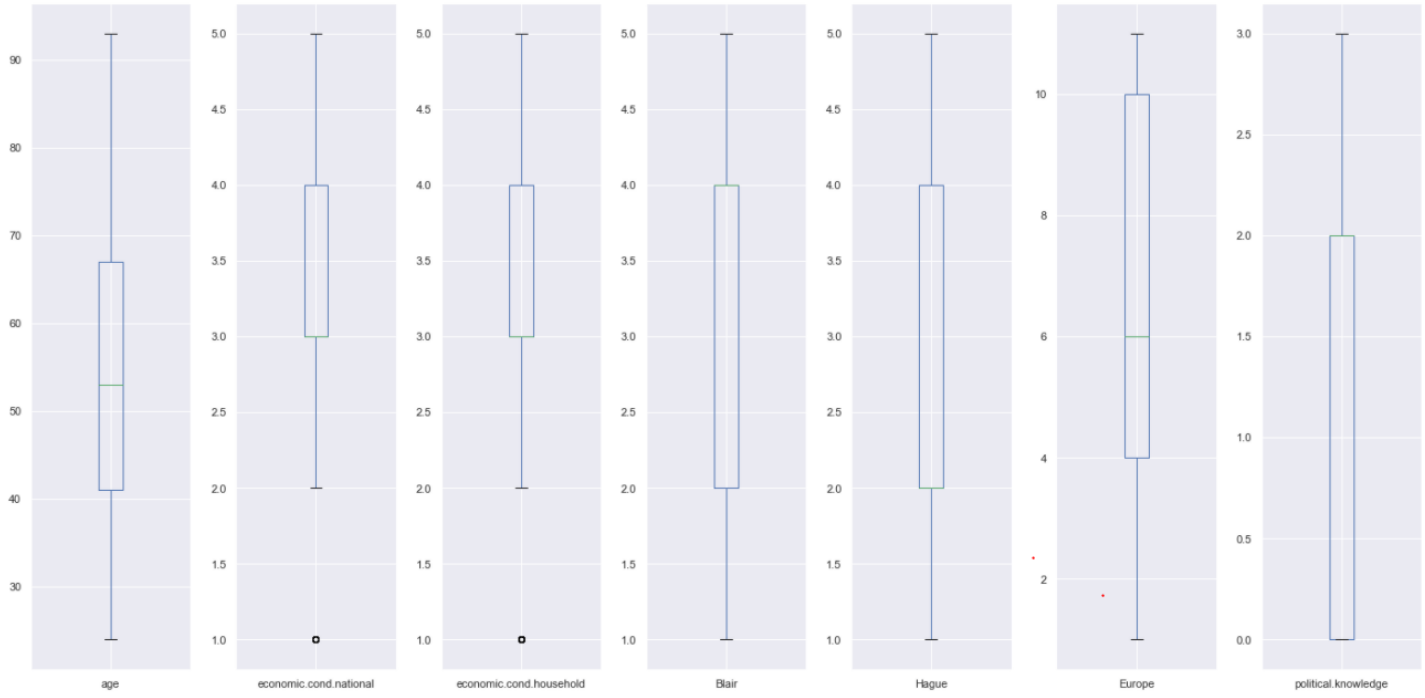
Data has no zero values

i.) Unique count of values in object variables.

```
VOTE :  2
Conservative     460
Labour          1057
Name: vote, dtype: int64


GENDER :  2
male      709
female    808
Name: gender, dtype: int64
```

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

a) Check for outlier




```
the number of outliers are  0
```

Inference:

Data has No outliers

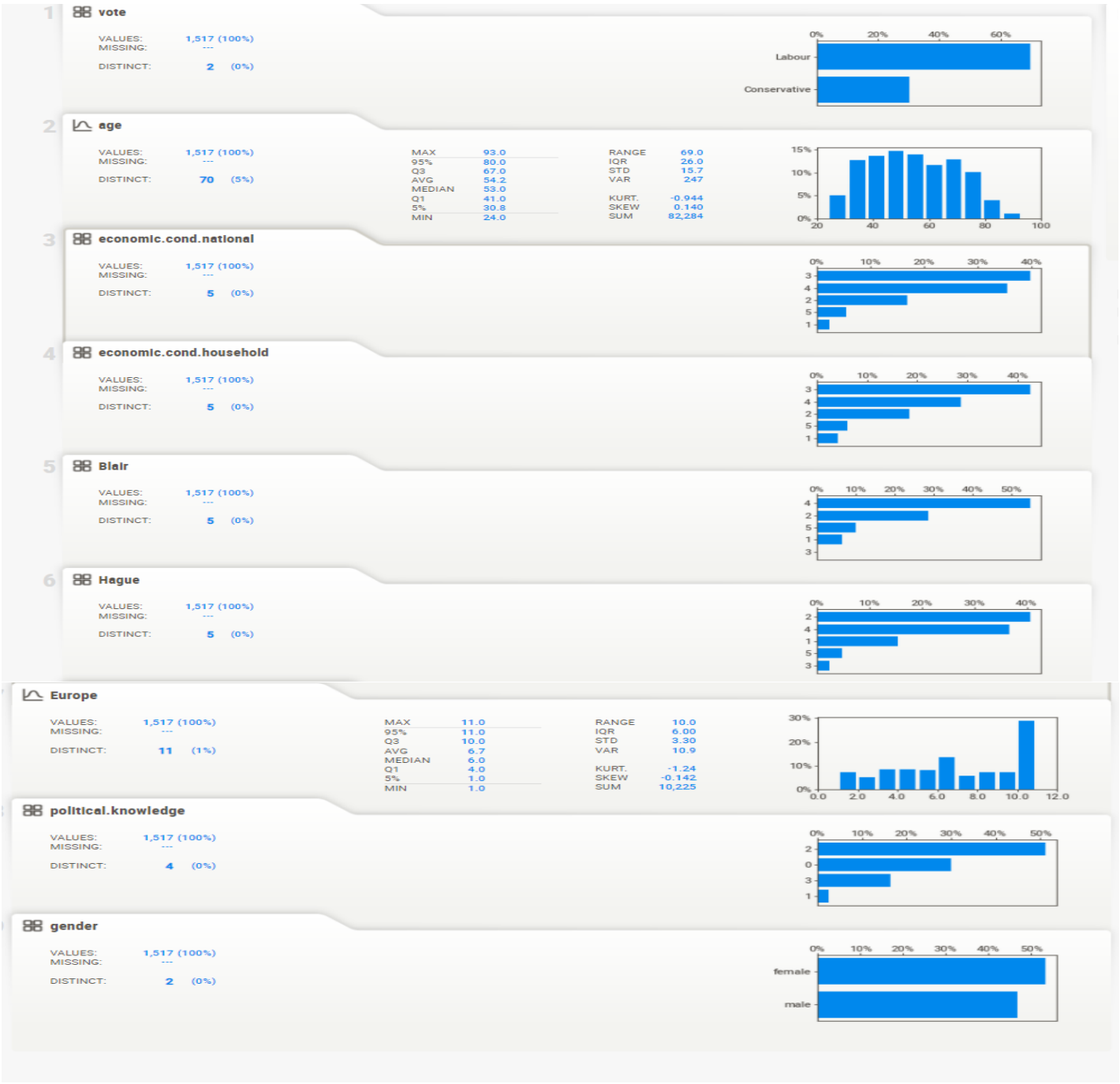
b.) Any null values?

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

Inference:

Data has no null values

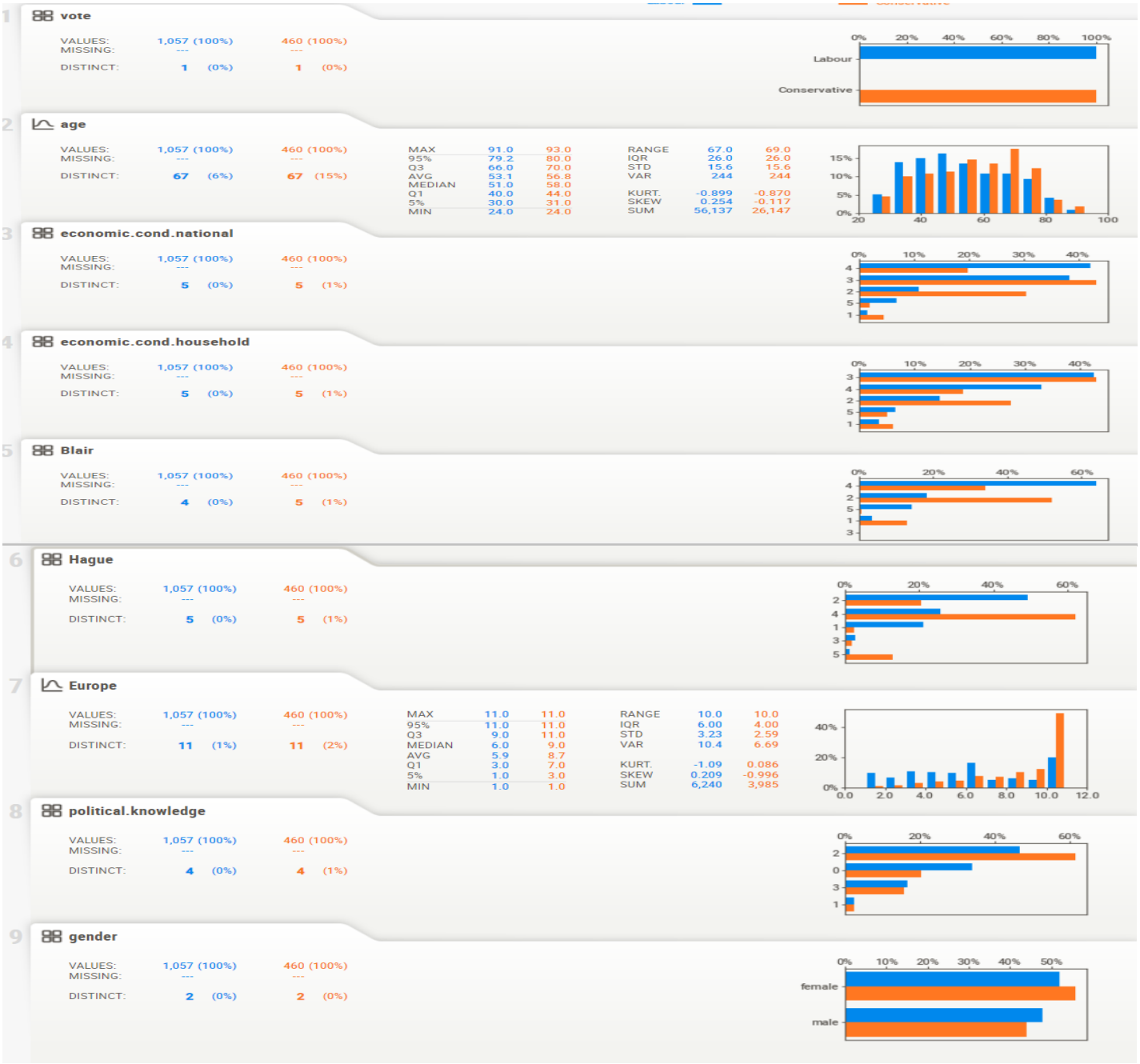
c) Univariate Analysis



Inference:

The class proportion of target variable “vote “is more than 10 percent. so its balanced for modeling.
More than Two third of the voters are from economic condition of 3 and 4.
More than 50 percentage of people have assessed 4 and above for the labor leader (Blair)
More than 50 percentage of people have assessed 2 and below for the conservative leader (Hague)
More than 50 percentage of people have political knowledge of below 2.
Most of the people have preferred the Eurosceptic sentiments
There are slightly more female voters than male voters.

d) Bi variate analysis (Between Target variable (Vote) and other variables)



Inference:

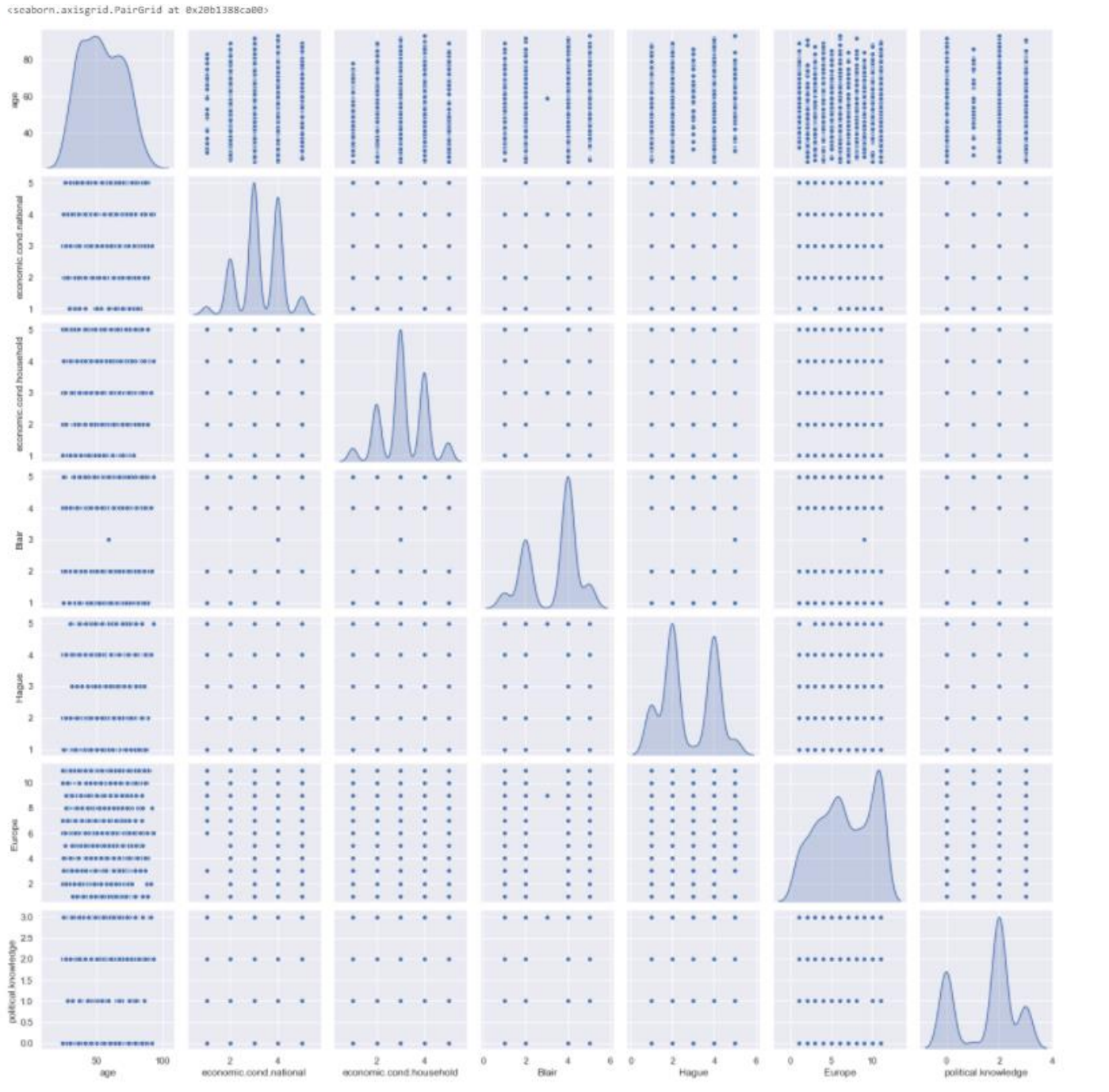
There is more voter's observation for Labor than conservative in the data set.
Majority of Voters belonging to economic condition of 4 and 5 prefer Labor (In number of observation). Voters belonging to 2 and below prefer conservative party.
More voters from Conservative party have preference towards Europe sceptic sentiments than labor party
Majority of females prefer conservative party and majority of male prefer labor

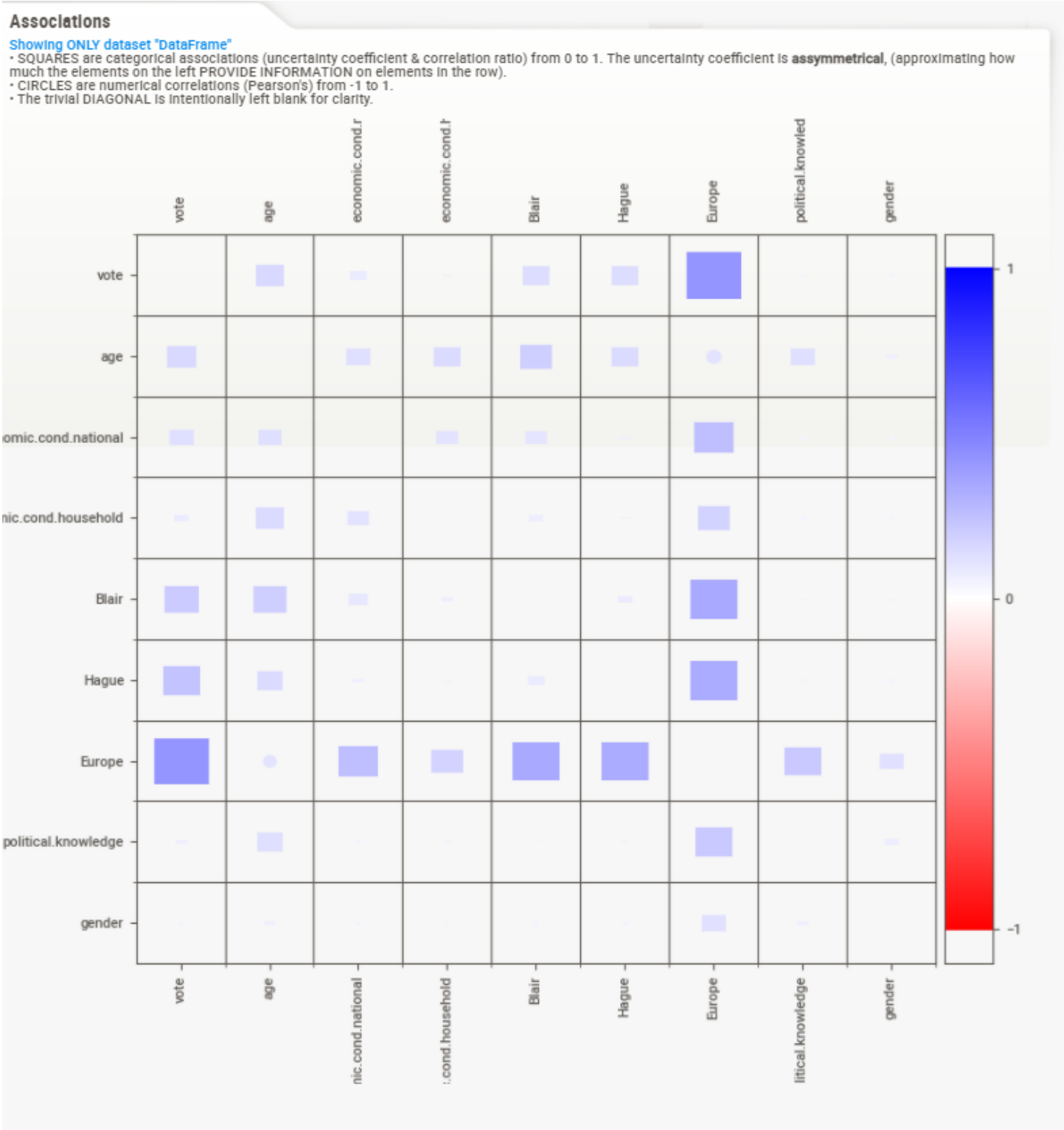
Majority Voters with political knowledge of 2 prefer conservative. Majority of Voters with political knowledge of 0 and 3 prefer labor

Most number of voters between age 20 – 50 they prefer Labor. Majority of Voters above the age of 50 prefer conservative

e.) Multivariate Analysis

[22]: <seaborn.axisgrid.PairGrid at 0x28b1388ca80>





Inference:

Target variable Vote has positive Categorical association with Europe, Hague, Blair and age.
Most of the variable shows multiple peaks specifying data has multiple classes

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

a) Encode data

Head of dataset after encoding

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	Labour	43	3	3	4	1	2	2	2
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	2
4	Labour	41	2	2	1	1	6	2	1

Inference:

Gender column is encoded to 1 (if male) and 2 (if female)

b) Data requires scaling?

data summary

	count	mean	std	min	25%	50%	75%	max
age	1517.0	54.241266	15.701741	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1517.0	3.245221	0.881792	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1517.0	3.137772	0.931069	1.0	3.0	3.0	4.0	5.0
Blair	1517.0	3.335531	1.174772	1.0	2.0	4.0	4.0	5.0
Hague	1517.0	2.749506	1.232479	1.0	2.0	2.0	4.0	5.0
Europe	1517.0	6.740277	3.299043	1.0	4.0	6.0	10.0	11.0
political.knowledge	1517.0	1.540541	1.084417	0.0	0.0	2.0	2.0	3.0
gender	1517.0	1.532630	0.499099	1.0	1.0	2.0	2.0	2.0

Inference:

Data scaling not required for this dataset. The scale of all features is relevant.

c) Data Split:

```
numpy.matrix
```

Inference:

Data is successfully split into train and test (70:30) and random state is 1

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

A.) Logistic Regression

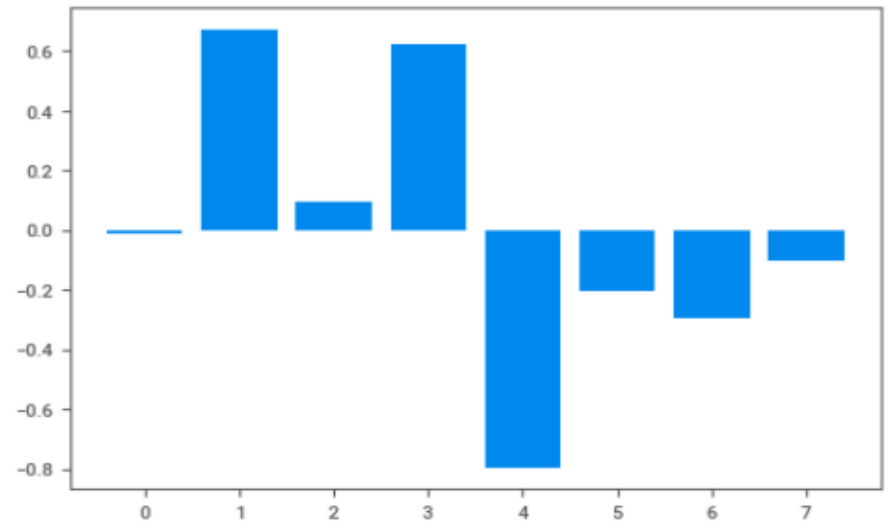
```
: LogisticRegression(solver='liblinear')
```

Logistic model score

```
0.8289473684210527
```

Important attributes for Logistic regression:

```
Feature: 0,Score :-0.01230
Feature: 1,Score :0.67157
Feature: 2,Score :0.09575
Feature: 3,Score :0.62326
Feature: 4,Score :-0.79487
Feature: 5,Score :-0.20179
Feature: 6,Score :-0.29544
Feature: 7,Score :-0.10248
```



B.) LDA

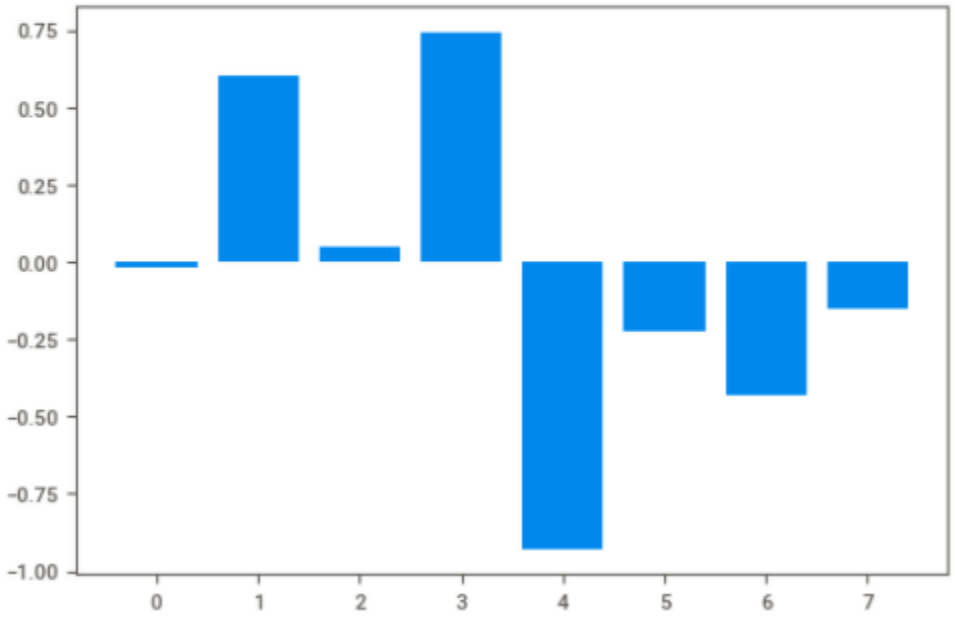
```
LinearDiscriminantAnalysis()
```

LDA model score

```
0.8333333333333334
```

Important attributes using LDA:

```
Feature: 0,Score :-0.02004
Feature: 1,Score :0.60492
Feature: 2,Score :0.05007
Feature: 3,Score :0.74240
Feature: 4,Score :-0.92663
Feature: 5,Score :-0.22361
Feature: 6,Score :-0.43033
Feature: 7,Score :-0.14908
```



Inference:

Logistic and LDA models are built for the given data set

Based on the score parameter the LDA model shows slightly better prediction than Logistic model

The LDA model will be able to predict the right party name which a voter will vote with 83 percentage accuracy based on the test data

The Logistic model will be able to predict the right party name which a voter will vote with 82 percentage accuracy based on the test data.

The best 3 features through Logistic regression and LDA are Economic.cond.national ,Blair and Hague

1.5) Apply KNN Model and Naïve Bayes Model and Interpret the inferences of both models

A.) KNN Model

```
KNeighborsClassifier(weights='distance')
```

KNN model score

```
0.8157894736842105
```

B.) Naïve Bayes

```
GaussianNB()
```

Naïve Bayes model score

```
0.8223684210526315
```

Inference:

The KNN and Naïve Bayes model are built for the given data

Based on the score parameter the Naïve Bayes shows slightly better prediction than KNN model

The KNN model will be able to predict the right party name which a voter will vote with 81 percentage accuracy based on the test data

The Naïve Bayes model will be able to predict the right party name which a voter will vote with 82 percentage accuracy based on the test data.

1.6) Model Tuning, Bagging and Boosting .

A.) Random Forest is built for Bagging

```
RandomForestClassifier(n_estimators=50, random_state=1)
```

B) Bagging

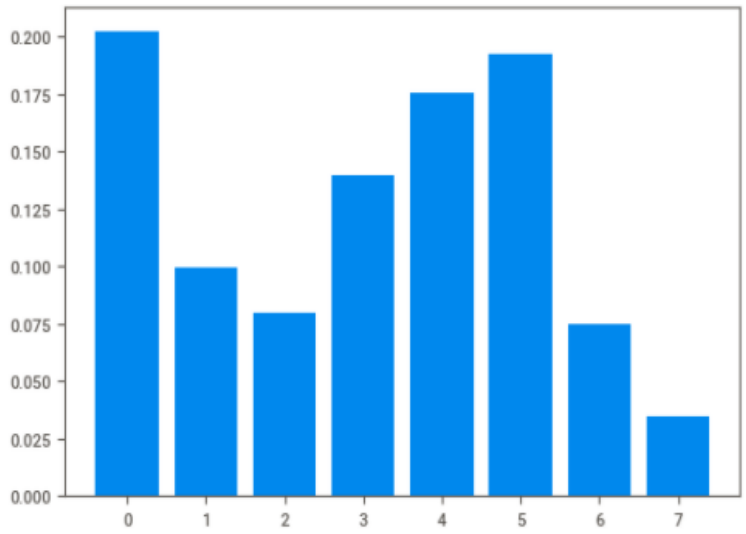
```
BaggingClassifier(base_estimator=RandomForestClassifier(n_estimators=50,
                                                        random_state=1),
                  n_estimators=50, random_state=1)
```

Bagging model score

```
0.8289473684210527
```

Important attributes for Bagging:

```
Feature: 0,Score :0.20270
Feature: 1,Score :0.09975
Feature: 2,Score :0.08011
Feature: 3,Score :0.13939
Feature: 4,Score :0.17587
Feature: 5,Score :0.19250
Feature: 6,Score :0.07489
Feature: 7,Score :0.03479
```



c.) Ada Boosting

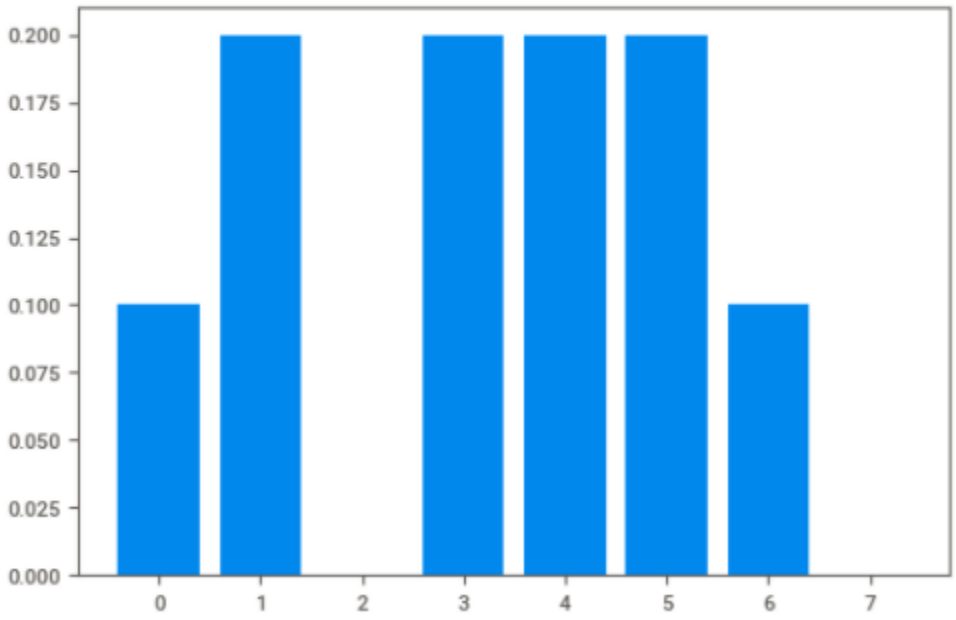
```
AdaBoostClassifier(n_estimators=10, random_state=1)
```

Ada boosting score

```
0.8201754385964912
```

Important attributes for Ada Boosting:

```
Feature: 0,Score :0.10000
Feature: 1,Score :0.20000
Feature: 2,Score :0.00000
Feature: 3,Score :0.20000
Feature: 4,Score :0.20000
Feature: 5,Score :0.20000
Feature: 6,Score :0.10000
Feature: 7,Score :0.00000
```



d.) Gradient Boosting

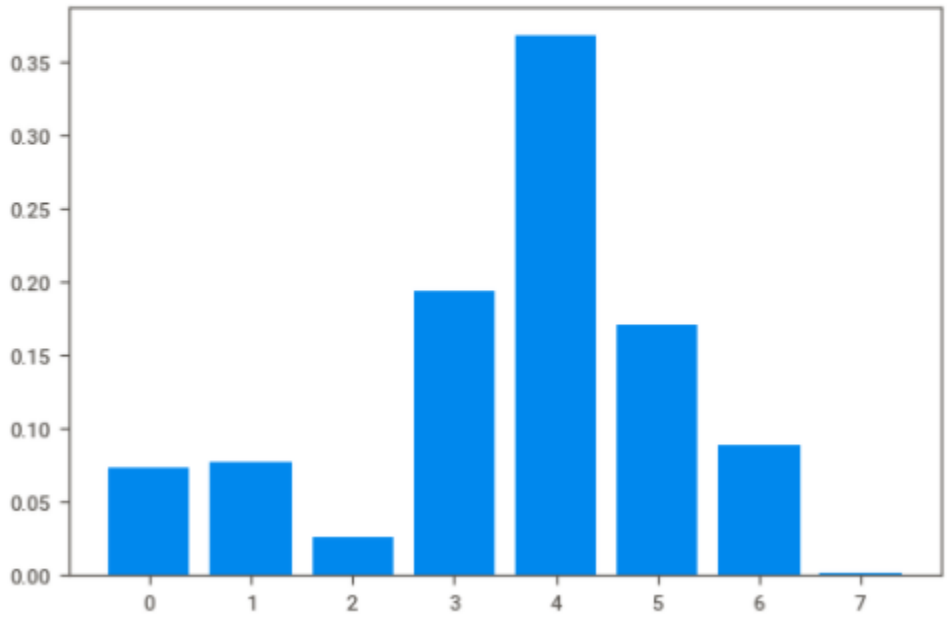
```
GradientBoostingClassifier(n_estimators=50, random_state=1)
```

Gradient boosting score

```
0.8289473684210527
```

Important attributes for Gradient Boosting:

```
Feature: 0,Score :0.07292
Feature: 1,Score :0.07714
Feature: 2,Score :0.02613
Feature: 3,Score :0.19418
Feature: 4,Score :0.36841
Feature: 5,Score :0.17084
Feature: 6,Score :0.08899
Feature: 7,Score :0.00139
```



Inference:

The Bagging (Random forest classifier) and Boosting (Ada boosting and Gradient boosting) are built for the given data

Based on the score parameter the Bagging and Gradient boosting are showing slightly better prediction than Ada boosting.

The Bagging model and Gradient boosting will be able to predict the right party name which a voter will vote with 82.8 percentage accuracy based on the test data

The Ada boosting will be able to predict the right party name which a voter will vote with 82 percentage accuracy based on the test data.

The Top features of Bagging model are: Age, Blair, Hague, and Europe
The Top Features of Ada boosting are: Economic.cond.national , Blair , Hague and Europe
The Top Feature of Gradient boosting are: Blair, Hague and Europe

1.7) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized

a) Performance Metrics for Logistic Regression

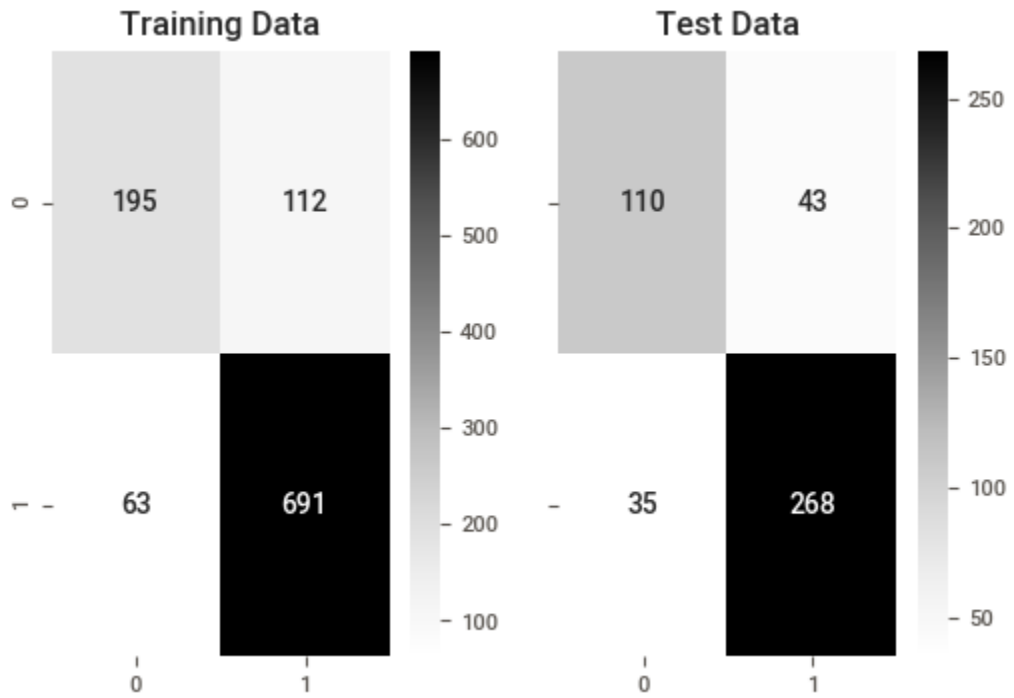
Accuracy score for Training data

```
Accuracy score  for Logistic regression  train variables
0.8350612629594723
```

Accuracy score for Testing data

```
Accuracy score for Logistic regression test variables
0.8289473684210527
```

Confusion Matrix for Logistic Regression

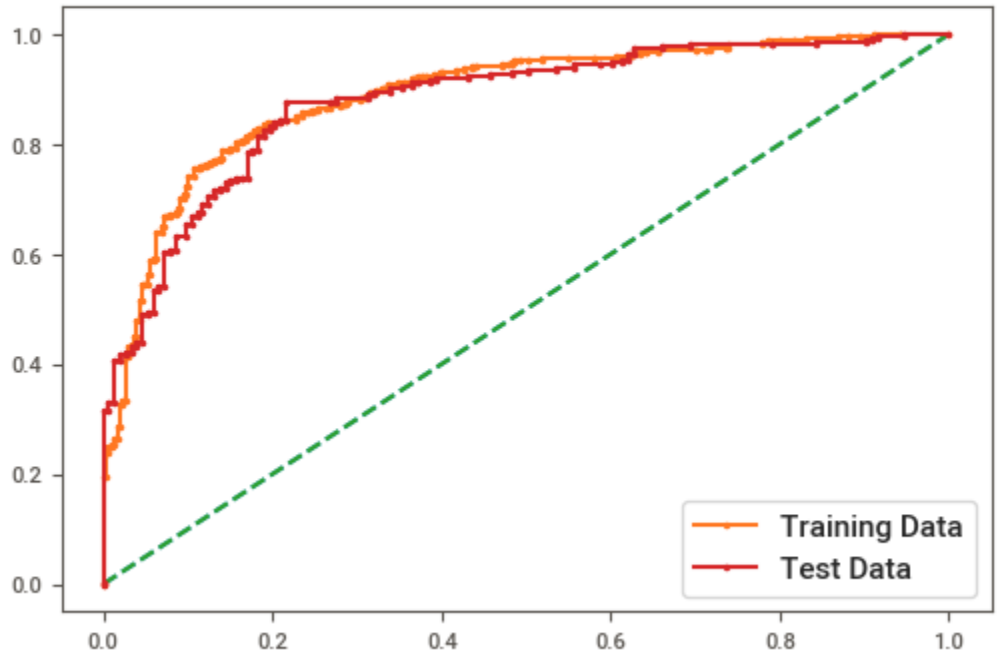


Classification Report for Logistic Regression:

```
Logistic regression Classfication report
Classification Report of the training data:

                precision    recall  f1-score   support

Conservative       0.76      0.64      0.69       307
      Labour       0.86      0.92      0.89       754

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


Classification Report of the test data:

                precision    recall  f1-score   support

Conservative       0.76      0.72      0.74       153
      Labour       0.86      0.88      0.87       303

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

ROC curve and ROC_AUC score for Logistic Regression

```
AUC and ROC FOR Logistic regression
AUC for the Training Data: 0.890
AUC for the Test Data: 0.880
```



b) Performance Metrics for LDA

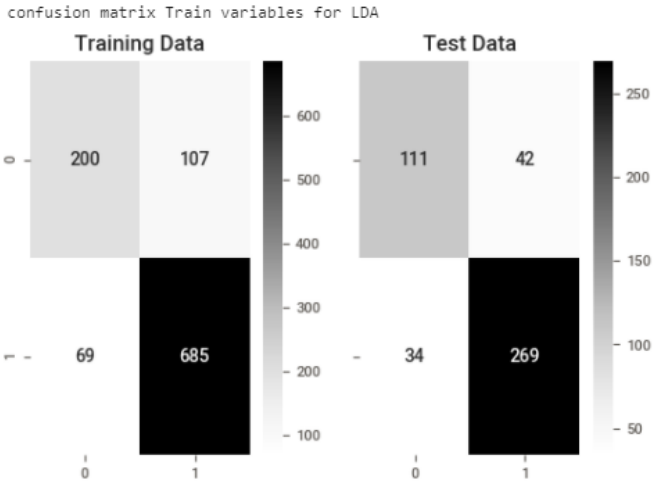
Accuracy score for Training data

```
Accuracy score  for LDA  train variables
0.8341187558906692
```

Accuracy score for Testing data

```
Accuracy score for LDA test variables
0.8333333333333334
```

Confusion Matrix For LDA

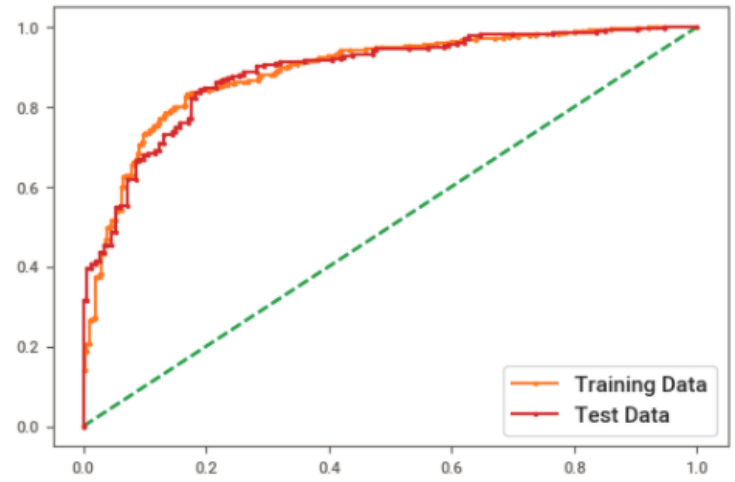


Classification Report For LDA

```
LDA Classfication report
Classification Report of the training data:
              precision    recall  f1-score   support

Conservative       0.74      0.65      0.69       307
      Labour       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 test data:
              precision    recall  f1-score   support

Conservative       0.77      0.73      0.74       153
      Labour       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
```

ROC curve and ROC_AUC score for LDA

```
AUC and ROC FOR LDA
AUC for the Training Data: 0.889
AUC for the Test Data: 0.888
```



c) Performance Metrics for KNN

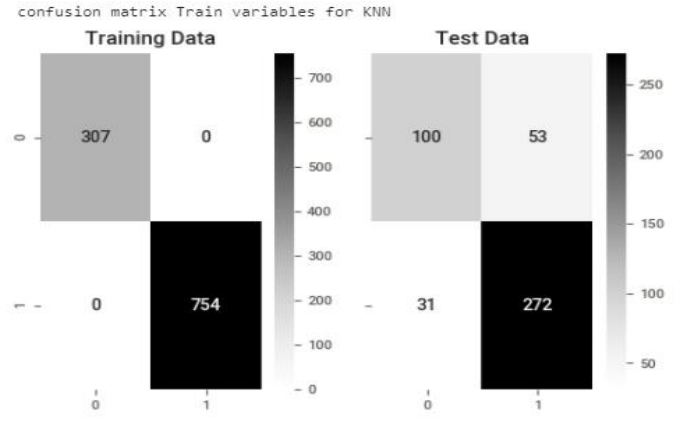
Accuracy score for Training data

```
Accuracy score  for KNN  train variables
1.0
```

Accuracy score for Testing data

```
Accuracy score for KNN test variables
]: 0.8157894736842105
```

Confusion Matrix For KNN

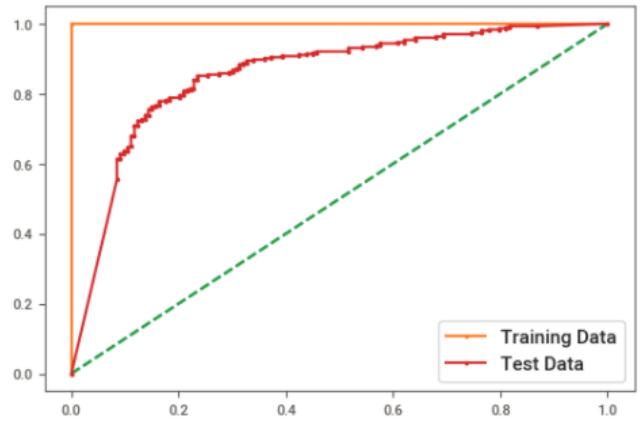


Classification Report For KNN

```
KNN Classfication report
Classification Report of the training data:
              precision    recall  f1-score   support

Conservative       1.00      1.00      1.00       307
      Labour       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


Classification Report of the test data:
              precision    recall  f1-score   support

Conservative       0.76      0.65      0.70       153
      Labour       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
```

ROC curve and ROC_AUC score for KNN

```
AUC and ROC FOR knn
AUC for the Training Data: 1.000
AUC for the Test Data: 0.858
```



d) Performance Metrics for Naïve Bayes

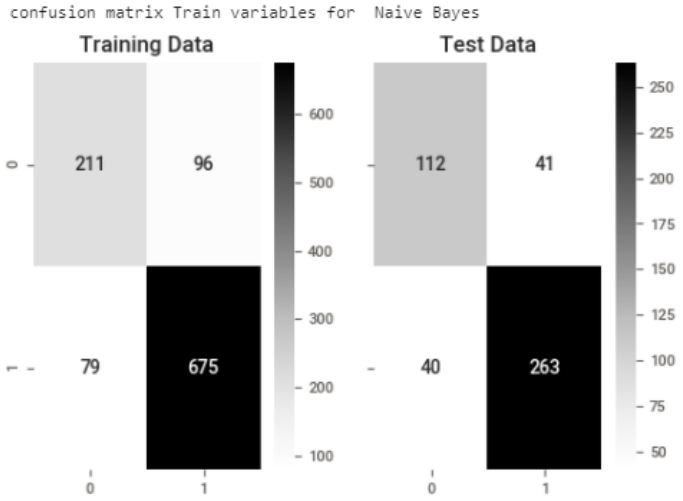
Accuracy score for Training data

```
Accuracy score  for Naive Bayes train variables
0.8350612629594723
```

Accuracy score for Testing data

```
Accuracy score for  Naive Bayes test variables
0.8223684210526315
```

Confusion Matrix For Naïve Bayes

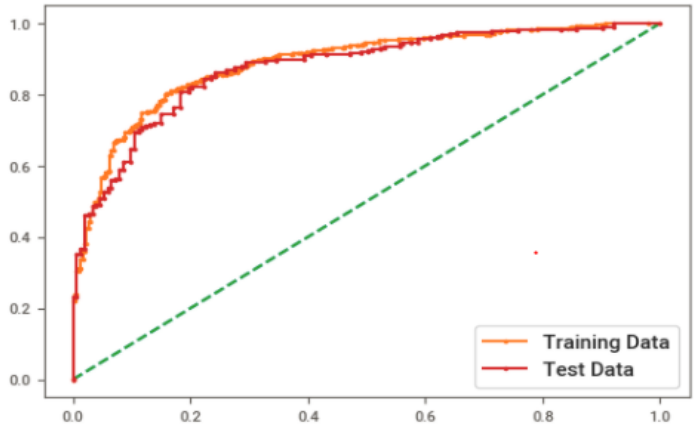


Classification Report For Naïve Bayes

```
 Naive Bayes Classfication report
Classification Report of the training data:

                precision    recall  f1-score   support

Conservative       0.73      0.69      0.71       307
      Labour       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


Classification Report of the test data:

                precision    recall  f1-score   support

Conservative       0.74      0.73      0.73       153
      Labour       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
```

ROC curve and ROC_AUC score for Naïve Bayes

```
AUC and ROC FOR  Naive Bayes
AUC for the Training Data: 0.888
AUC for the Test Data: 0.876
```



e) Performance Metrics for Bagging

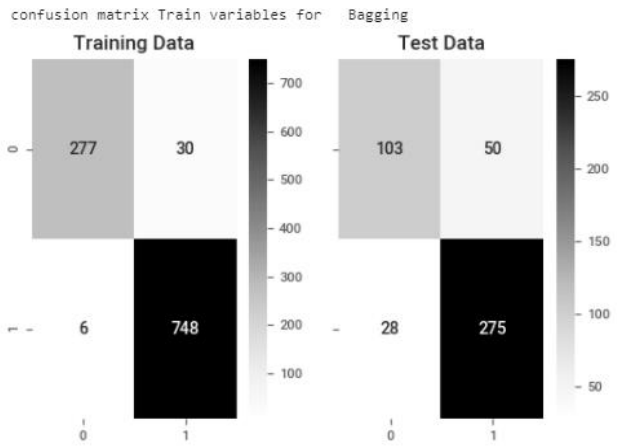
Accuracy score for Training data

```
Accuracy score  for Bagging train variables
0.9660697455230914
```

Accuracy score for Testing data

```
Accuracy score for Bagging test variables
0.8289473684210527
```

Confusion Matrix For Bagging

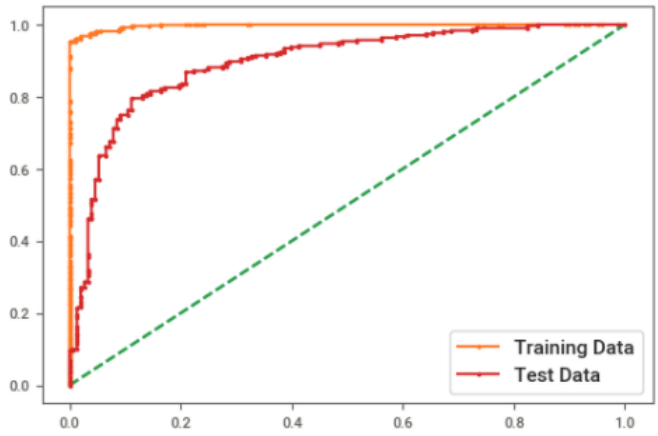


Classification Report For Bagging

```
  Classfication report  for   Bagging
Classification Report of the training data:

              precision    recall  f1-score   support

Conservative       0.98      0.90      0.94       307
      Labour       0.96      0.99      0.98       754

    accuracy                           0.97      1061
   macro avg       0.97      0.95      0.96      1061
weighted avg       0.97      0.97      0.97      1061


Classification Report of the test data:

              precision    recall  f1-score   support

Conservative       0.79      0.67      0.73       153
      Labour       0.85      0.91      0.88       303

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

ROC curve and ROC_AUC score for Bagging

```
AUC and ROC FOR   Bagging
AUC for the Training Data: 0.997
AUC for the Test Data: 0.896
```



f) Performance Metrics for ADA boosting

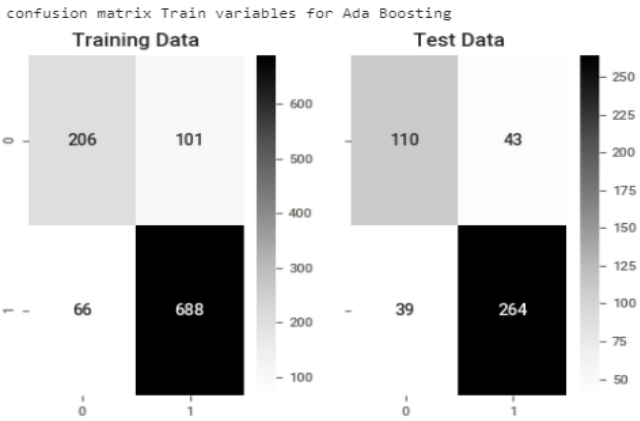
Accuracy score for Training data

```
Accuracy score  for Ada Boosting train variables
0.8426013195098964
```

Accuracy score for Testing data

```
   Accuracy score forAda Boosting test variables
]: 0.8201754385964912
```

Confusion Matrix For Ada Boosting

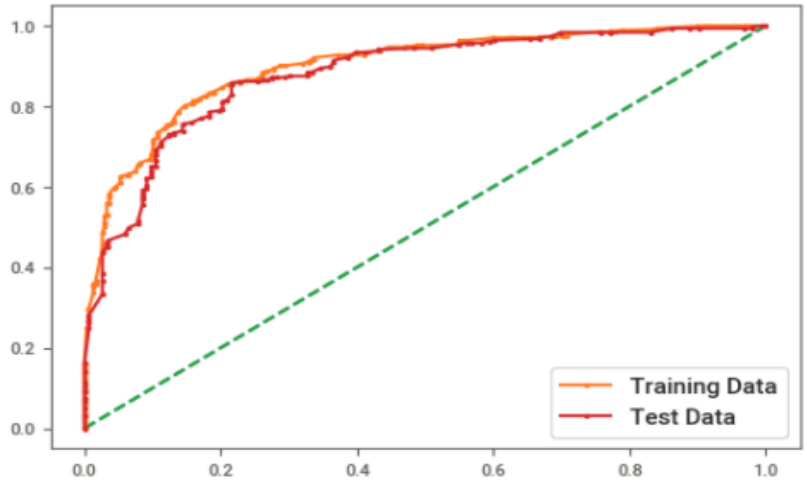


Classification Report For Ada boosting

```
  Classfication report  for   Ada Boosting
Classification Report of the training data:

                precision    recall  f1-score   support

Conservative       0.76      0.67      0.71       307
      Labour       0.87      0.91      0.89       754

    accuracy                           0.84      1061
   macro avg       0.81      0.79      0.80      1061
weighted avg       0.84      0.84      0.84      1061


Classification Report of the test data:

                precision    recall  f1-score   support

Conservative       0.74      0.72      0.73       153
      Labour       0.86      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
```

ROC curve and ROC_AUC score for Ada boosting

```
AUC and ROC FOR   Ada Boosting
AUC for the Training Data: 0.898
AUC for the Test Data: 0.878
```



g) Performance Metrics for Gradient boosting

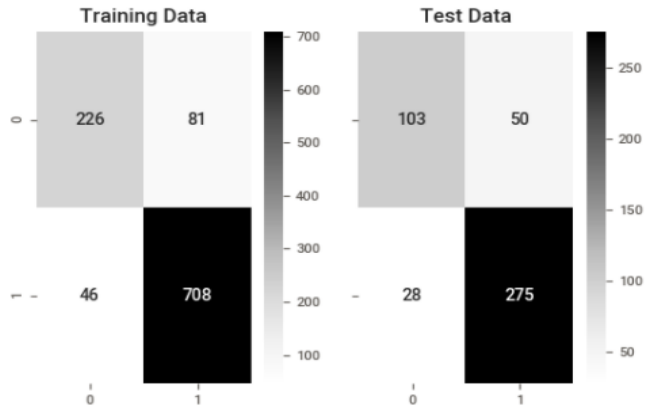
Accuracy score for Training data

```
Accuracy score  for Gradient Boosting train variables
0.8803016022620169
```

Accuracy score for Testing data

```
Accuracy score gradient Boosting test variables
0.8289473684210527
```

Confusion Matrix For gradient Boosting

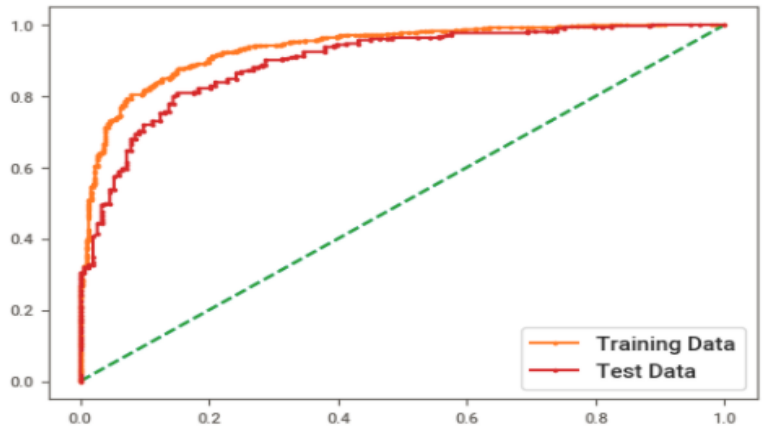


Classification Report For gradient boosting

```
  Classfication report  for   gradient Boosting
Classification Report of the training data:

               precision    recall  f1-score   support

Conservative       0.83      0.74      0.78       307
      Labour       0.90      0.94      0.92       754

    accuracy                           0.88      1061
   macro avg       0.86      0.84      0.85      1061
weighted avg       0.88      0.88      0.88      1061


Classification Report of the test data:

               precision    recall  f1-score   support

Conservative       0.79      0.67      0.73       153
      Labour       0.85      0.91      0.88       303

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

ROC curve and ROC_AUC score for gradient boosting

```
AUC and ROC FOR   Gradient Boosting
AUC for the Training Data: 0.935
AUC for the Test Data: 0.897
```



Compare all models on the basis of the performance metrics in a structured tabular manner.

Comparison in Table form:

	Logistic reg Train	Logistic reg Test	LDA Train	LDA Test	KNN Train	KNN Test	Naive Bayes Train	Naive Bayes Test	Bagging Train	Bagging Test	Ada Boosting Train	Ada Boosting Test	Gradient Boosting Train	Gradient Boosting Test
Accuracy	0.84	0.83	0.83	0.83	1.0	0.82	0.84	0.82	0.97	0.83	0.84	0.82	0.88	0.83
AUC	0.89	0.88	0.89	0.89	1.0	0.86	0.89	0.88	1.00	0.90	0.90	0.88	0.94	0.90
Recall	0.92	0.88	0.91	0.89	1.0	0.90	0.90	0.87	0.99	0.91	0.91	0.87	0.94	0.91
Precision	0.86	0.86	0.86	0.86	1.0	0.84	0.88	0.87	0.96	0.85	0.87	0.86	0.90	0.85
F1 Score	0.89	0.87	0.89	0.88	1.0	0.87	0.89	0.87	0.98	0.88	0.89	0.87	0.92	0.88

Inference:

Based on comparing the performance metrics, Gradient Boosting performs better than other models .Gradient boosting accuracy between training and testing is minimal .(88 percent for train and 83 percent for testing). Its recall rate is also 94 percent for Training and 90 percent for testing.

Even though Bagging has better score for many metrics, there is a huge difference between the training and testing.

So, I conclude that Gradient Boosting is the best model for this data set .

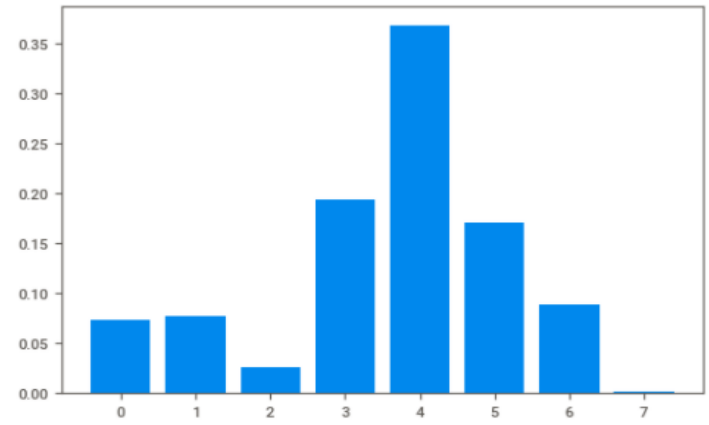
1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.

The gradient boosting model will be able to predict which party a voter will vote for on the basis of the given information with 83 percentage accuracy .

Business Insights:

The important factors which determine whether a voter will vote for a party is
Europe, Hague, Blair and political.knowledge

```
Feature: 0,Score :0.07292
Feature: 1,Score :0.07714
Feature: 2,Score :0.02613
Feature: 3,Score :0.19418
Feature: 4,Score :0.36841
Feature: 5,Score :0.17084
Feature: 6,Score :0.08899
Feature: 7,Score :0.00139
```



Recommendation:

If people prefer Europe sentiments, then there is a high chance for them to vote for Conservative party.

Around 55 percentage of voters have given assessment score of 4 for the Blair
Only 37 percentage of voters have given assessment score of 4 for the Hague.
So, Based on Assessment labor party leader Blair has more assessment majority score of 4 than Conservative party Hague (Majority score for Hague is 2).

If the voter has political knowledge of 2 then there is more chance for them to vote for conservative.
If the voter has political knowledge of 0 and 3 then there is more chance for them to vote for Labor.

More than 51 percentage of people belonging to age of 50 or less. The majority of voters of this age group less than 50 supports Labor party.

Problem 2: Text and Sentimental Analytics

Business scenario

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

President Franklin D. Roosevelt in 1941

President John F. Kennedy in 1961

President Richard Nixon in 1973

2.1) Find the number of characters, words and sentences for the mentioned documents.

```
1941-Roosevelt.txt Number of characters: 7571 , Number of words: 1536 , Number of sentence 68
1961-Kennedy.txt Number of characters: 7618 , Number of words: 1546 , Number of sentence 52
1973-Nixon.txt Number of characters: 9991 , Number of words: 2028 , Number of sentence 69
```

2.2) Remove all the stop words from the three speeches.

a.) Sample of data after removing stop words

```
stop words removed
['mr.',
 'vice',
 'presid',
 'mr.',
 'speaker',
 'mr.',
 'chief',
 'justic',
 'senat',
 'cook',
 'mrs.',
 'eisenhow',
 'fellow',
 'citizen',
 'great',
 'good',
 'countri',
 'share',
 'togeth',
 'met',
 'four',
 'year',
 'ago',
```

b.) President Roosevelt speech as text after removing stop words and punctuation and changing to lower case

"nation day inaugur sinc peopl renew sens dedic unit state washington 's day task peopl creat weld togeth nation lincoln 's day task peopl preserv nation disrupt within day task p eopl save nation institut disrupt without us come time midst swift happen paus moment take stock recal place histori rediscov may risk real peril inact live nation determin count year lifetim human spirit life man three-scor year ten littl littl less life nation full measur live men doubt men believ democraci form govern frame life limit measur kind mystic artifici fate unexplain reason tyranni slaveri becom surg wave futur freedom eb tide american know true eight year ago life republ seem frozen fatalist terror prove true midst sho ck act act quick bold decis later year live year fruit year peopl democraci brought us greater secur hope better understand life 's ideal measur materi thing vital present futur e xperi democraci success surviv crisi home put away mani evil thing built new structur endur line maintain fact democraci action taken within three-way framework constitut unit sta te coordin branch govern continu freeli function bill right remain inviol freedom elect wholli maintain prophet downfal american democraci seen dire predict come naught democraci die know becaus seen reviv grow know die becaus built unhamp initi individu men women join togeth common enterpris enterpris undertaken carri free express free major know becaus d emocraci alon form govern enlist full forc men 's enlighten know becaus democraci alon construct unlimit civil capabl infinit progress improv human life know becaus look surfac se ns still spread everi contin human advanc end unconquer form human societi nation like person bodi bodi must fed cloth hous invigor rest manner measur object time nation like pers on mind mind must kept inform alert must know understand hope need neighbor nation live within narrow circl world nation like person someth deeper someth perman someth larger sum part someth matter futur call forth sacr guard present thing find difficult even imposs hit upon singl simpl word yet understand spirit faith america product centuri born multitud came mani land high degre plain peopl sought earli late find freedom freeli democrat aspir mere recent phase human histori human histori permeat ancient life earli peopl blaze ane w middl age written magna charta america impact irresist america new world tongu peopl becaus contin new-found land becaus came believ could creat upon contin new life life new fr eedom vital written mayflow compact declar independ constitut unit state gettysburg address first came carri long spirit million follow stock sprang move forward constant consist toward ideal gain statur clariti generat hope republ forev toler either undeserv poverti self-serv wealth know still far go must great build secur opportun knowledg everi citizen measur justifi resourc capac land enough achiev purpos alon enough cloth feed bodi nation instruct inform mind also spirit three greatest spirit without bodi mind men know nation could live spirit america kill even though nation 's bodi mind constrict alien world live america know would perish spirit faith speak us daili live way often unnot becaus seem ob vious speak us capit nation speak us process govern sovereignti state speak us counti citi town villag speak us nation hemispher across sea enslav well free sometim fail hear heed voic freedom becaus us privileg freedom old old stori destini america proclaim word propheci spoken first presid first inaugur word almost direct would seem year preserv sacr fire liberti destini republican model govern consid deepli final stake experi intrust hand american peopl lose sacr fire let smother doubt fear shall reject destini washington strove v aliant triumphant establish preserv spirit faith nation doe furnish highest justif everi sacrific may make caus nation defens face great peril never befor encount strong purpos pr otect perpetu integr democraci muster spirit america faith america retreat content stand still american go forward servic countri god "

c.) President Kennedy speech as text after removing stop words and punctuation and changing to lower case

"vice presid johnson mr. speaker mr. chief justic presid eisenhow vice presid nixon presid truman reverend clergi fellow citizen observ today victori parti celebr freedom symbol e nd well begin signifi renew well chang sworn befor almighti god solemn oath forebear l prescrib near centuri three quarter ago world veri differ man hold mortal hand power abolish form human poverti form human life yet revolutionari belief forebear fought still issu around globe belief right man come generos state hand god dare forget today heir first revol ut let word go forth time place friend foe alik torch pass new generat american born centuri temper war disciplin hard bitter peac proud ancient heritag unwil wit permit slow undo human right nation alway commit commit today home around world let everi nation know whether wish us well ill shall pay ani price bear ani burden meet ani hardship support ani fri end oppos ani foe order assur surviv success liberti much pledg old alli whose cultur spiritu origin share pledg loyalti faith friend unit littl host cooper ventur divid littl dar e meet power challeng odd split asund new state welcom rank free pledg word one form coloni control shall pass away mere replac far iron tyranni shall alway expect find support vi ew shall alway hope find strong support freedom rememb past foolish sought power ride back tiger end insid peopl hut villag across globe struggl break bond mass miseri pledg best effort help help themselv whatev period requir becaus communist may becaus seek vote becaus right free societi help mani poor save rich sister republ south border offer special pl edg convert good word good deed new allianc progress assist free men free govern cast chain poverti peac revolut hope becom prey hostil power let neighbor know shall join oppos ag gress subvers anywher america let everi power know hemispher intend remain master hous world assembl sovereign state unit nation last best hope age instrument war far outpac instr ument peac renew pledg support prevent becom mere forum invect strengthen shield new weak enlarg area writ may run final nation would make themselv adversari offer pledg request s ide begin anew quest peac befor dark power destruct unleash scienc engulf human plan accident self-destruct dare tempt weak onli arm suffici beyond doubt certain beyond doubt neve r employ neither two great power group nation take comfort present cours side overburden cost modern weapon right alarm steadi spread dead atom yet race alter uncertain balanc ter ror stay hand mankind 's final war let us begin anew rememb side civil sign weak sincer alway subject proof let us never negoti fear let us never fear negoti let side explor probl em unit us instead belabor problem divid us let side first time formul serious precis propos inspect control arm bring absolut power destroy nation absolut control nation let side seek invok wonder scienc instead terror togeth let us explor star conquer desert erad diseas tap ocean depth encourag art commerc let side unit heed corner earth command isaiah un do heavi burden let oppress go free beachhead cooper may push back jungl suspicion let side join creat new endeavor new balanc power new world law strong weak secur peac preserv f inish first day finish first day life administr even perhap lifetim planet let us begin hand fellow citizen mine rest final success failur cours sinc countri found generat america n summon give testimoni nation loyalti grave young american answer call servic surround globe trumpet summon us call bear arm though arm need call battl though embattl call bear b urden long twilight struggl year year rejoic hope patient tribul struggl common enemi man tyranni poverti diseas war forg enemi grand global allianc north south east west assur fr uit life mankind join histor effort long histori world onli generat grant role defend freedom hour maximum danger shrink respons welcom believ ani us would exchang place ani peopl ani generat energi faith devot bring endeavor light countri serv glow fire truli light world fellow american ask countri ask countri fellow citizen world ask america togeth freedo m man final whether citizen america citizen world ask us high standard strength sacrific ask good conscienc onli sure reward histori final judg deed let us go forth lead land love ask bless help know earth god 's work must truli "

d.) President Nixon speech as text after removing stop words and punctuation and changing to lower case

"mr. vice presid mr. speaker mr. chief justic senat cook mrs. eisenhow fellow citizen great good countri share togeth met four year ago america bleak spirit depress prospect seem endless war abroad destruct conflict home meet today stand threshold new era peac world central question befor us shall use peac let us resolv era enter postwar period often time retreat isol lead stagnat home invit new danger abroad let us resolv becom time great respons great born renew spirit promis america enter third centuri nation past year saw far-r each result new polici peac continu revit tradit friendship mission peke moscow abl establish base new durabl pattern relationship among nation world becaus america 's bold initi long rememb year greatest progress sinc end world war ii toward last peac world peac seek world flimsi peac mere interlud war peac endur generat come import understand necess limi t america 's role maintain peac unless america work preserv peac peac unless america work preserv freedom freedom let us clear understand new natur america 's role result new poli ci adopt past four year shall respect treati commit shall support vigor principl countri right impos rule anoth forc shall continu era negoti work limit nuclear arm reduc danger c onfront great power shall share defend peac freedom world shall expect share time pass america make everi nation 's conflict make everi nation 's futur respons presum tell peopl n ation manag affair respect right nation determin futur also recogn respons nation secur futur america 's role indispens preserv world 's peac nation 's role indispens preserv peac togeth rest world let us resolv move forward begin made let us continu bring wall hostil divid world long build place bridg understand despit profound differ system govern peopl w orld friend let us build structur peac world weak safe strong respect right live differ system would influenc strength idea forc arm let us accept high respons burden glad glad be caus chanc build peac noblest endeavor nation engag glad also becaus onli act great meet respons abroad remain great nation onli remain great nation act great meet challeng home c hanc today ever befor histori make life better america ensur better educ better health better hous better transport cleaner environ restor respect law make communiti livabl insur god-given right everi american full equal opportun becaus rang need great becaus reach opportun great let us bold determin meet need new way build structur peac abroad requir turn away old polici fail build new era progress home requir turn away old polici fail abroad shift old polici new retreat respons better way peac home shift old polici new retreat res pons better way progress abroad home key new respons lie place divis respons live long consequ attempt gather power respons washington abroad home time come turn away condescend p olici patern washington know best person expect act respons onli respons human natur let us encourag individu home nation abroad themselv decid themselv let us locat respons place let us measur themselv whi today offer promis pure government solut everi problem live long fals promis trust much govern ask deliv lead onli inflat expect reduc individu effort d isappoint frustrat erod confid govern peopl govern must learn take less peopl peopl themselv let us rememb america built govern peopl welfar work shirk respons seek respons live l et us ask govern challeng face togeth let us ask govern help help nation govern great vital role play pledg govern act act bold lead bold import role everi one us must play indivi du member communiti day forward let us make solemn commit heart bear respons part live ideal togeth see dawn new age progress america togeth celebr 200th anniversari nation proud fulfil promis ourselv world america 's longest difficult war come end let us learn debat differ civil decenc let us reach one precious qualiti govern provid new level respect righ t feel one anoth new level respect individu human digniti cherish birthright everi american abov els time come us renew faith ourselv america recent year faith challeng children t aught asham countri asham parent asham america 's record home role world everi turn beset find everyth wrong america littl right confid judgment histori remark time privileg live america 's record centuri unparallel world 's histori respons generos creativ progress let us proud system produc provid freedom abund wide share ani system histori world let us p roud four war engag centuri includ one bring end fought selfish advantag help resist aggress let us proud bold new initi steadfast peac honor made break-through toward creat world world known befor structur peac last mere time generat come embark today era present challeng great ani nation ani generat ever face shall answer god histori conscienc way use yea r stand place hallow histori think stood befor think dream america think recogn need help far beyond order make dream come true today ask prayer year ahead may god 's help make de cis right america pray help togeth may worthi challeng let us pledg togeth make next four year best four year america 's histori 200th birthday america young vital began bright be acon hope world let us go forward confid hope strong faith one anoth sustain faith god creat us strive alway serv purpos "

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 stop words)

```
The top 3 words by President Roosevelt : ['nation', 'know', 'peopl']
The top 3 words by President kennedy: ['let', 'us', 'power']
The top 3 words by President Nixon : ['us', 'let', 'america']
```

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

a.) President Roosevelt speech word cloud

b.) President Kennedy speech word cloud

c.) President Nixon speech word cloud