10/1/2023

Election Exit Poll Prediction - Report

PGP-DSBA

Karthick Raj S

Table of Contents

List of Tables

List of Figures

Problem:

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.

Dataset for Problem: Election_Data.xlsx

Data Ingestion: 11 marks

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks)

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

Data Preparation: 4 marks

1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). (4 Marks)

Modeling: 22 marks

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

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

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

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized. (7 marks)

Inference: 5 marks

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

Data Description:

S.no	Variable Name	Description
1	vote	Party choice: Conservative or Labour
2	age	Respondents' age in years
3	economic.cond.national	Assessment of current national economic conditions, 1 to 5.
4	economic.cond.household	Assessment of current household economic conditions, 1 to 5.
5	Blair	Assessment of the Labour leader, 1 to 5.
6	Hague	Assessment of the Conservative leader, 1 to 5.
7	Europe	Respondents' attitudes toward European integration, 1-11
8	political.knowledge	Knowledge of parties' positions on European integration, 0 to 3.
9	gender	Respondents' gender,female or male.

Table 1 Data Description

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

Descriptive Statistics:

The First Five Rows of Dataset

	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

Table 2 First Five Rows

- There are 1525 rows and 9 columns.
- Vote is the target variable with Labour and Conservative categories.
- Other than age in the remaining 8, all are categorical variable.

```
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   vote                     1525 non-null   object
 1   age                      1525 non-null   int64
 2   economic.cond.national   1525 non-null   int64
 3   economic.cond.household  1525 non-null   int64
 4   Blair                    1525 non-null   int64
 5   Hague                    1525 non-null   int64
 6   Europe                   1525 non-null   int64
 7   political.knowledge      1525 non-null   int64
 8   gender                   1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

Table 3 Data Info

- Vote and Gender are object.
- All 7 are integer datatype but other than age, all should be changed to category datatype.

- There are no null values. Also checked with isnull() function.

```
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 4 Null Values count

Duplicates:

There are 8 duplicates in the dataset.

Duplicates are not treated as the id column or unique voter id is not present. They might be genuine duplicates. It won't have any affect in the analysis.

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

Table 5 Duplicates List

Except vote, age and gender, other variables are converted to category datatype.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   vote                     1525 non-null   object
 1   age                      1525 non-null   int64
 2   economic.cond.national   1525 non-null   category
 3   economic.cond.household  1525 non-null   category
 4   Blair                    1525 non-null   category
 5   Hague                    1525 non-null   category
 6   Europe                   1525 non-null   category
 7   political.knowledge      1525 non-null   category
 8   gender                   1525 non-null   object
dtypes: category(6), int64(1), object(2)
memory usage: 46.2+ KB
```

Table 6 Data Info After conversion

The Descriptive statistics for the dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
vote	1525	2	Labour	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	1525.0	NaN	NaN	NaN	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	5.0	3.0	607.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
economic.cond.household	1525.0	5.0	3.0	648.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Blair	1525.0	5.0	4.0	836.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Hague	1525.0	5.0	2.0	624.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Europe	1525.0	11.0	11.0	338.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
political.knowledge	1525.0	4.0	2.0	782.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gender	1525	2	female	812	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 7 Descriptive Statistics

Observations:

- **vote:** Vote has two categories with labour having 2/3rd of the data.
- **age:** age is a continuous variable with mean of 54 and range of 24 to 93. The age is slightly rightly skewed (Skewness: 0.144). we can also see the range of age is 24 to 93 (older respondents are higher than younger respondents).
- **gender:** 53% of the voters are females.
- **economic.cond.national:** National Economic condition is neutral mostly
- **economic.cond.household:** Household Economic condition is neutral mostly
- **Blair:** More than 50% of the voters feel better towards Blair.
- **Hague:** Around 40% of the voters are not too sure about Hague and 45% of the voters feels good about Hague.
- **Europe:** The attitude towards the European integration is not popular among the voters as only 29% feels strongly about that and other remaining voters feels neutral or not strongly about the European integration.
- **political.knowledge:** Most of the voters have political knowledge as the survey have 67% have 2 and above.

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

EDA

Univariate:

CountPlot:

Observation:

- 70% of the respondents in the survey supports Blair's Labour Party.
- Labour party will get votes more than twice of conservative party based on the survey.

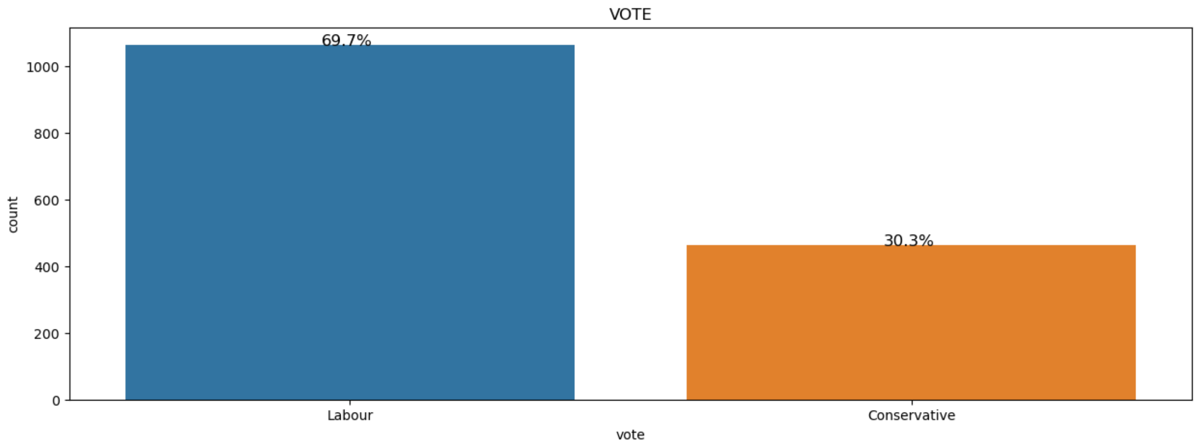

Figure A Countplot of Vote

```
VOTE
Labour          1063
Conservative     462
Name: vote, dtype: int64
vote  in Percentage
Labour          69.704918
Conservative    30.295082
Name: vote, dtype: float64
***********************************************************************
```

Observation:

- 80% of the respondents have an good National economic conditions.

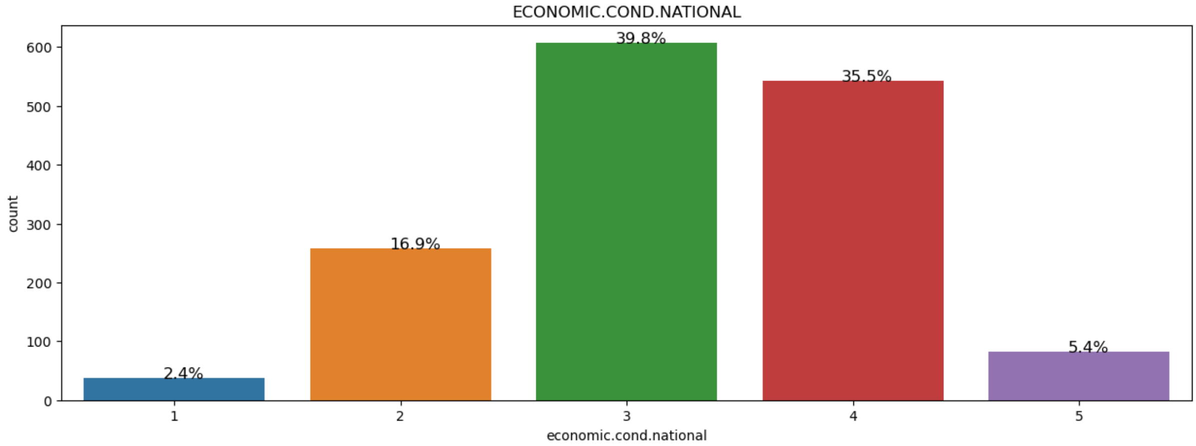

Figure B Countplot of National Economic conditions

```
ECONOMIC.COND.NATIONAL
3    607
4    542
2    257
5     82
1     37
Name: economic.cond.national, dtype: int64
economic.cond.national  in Percentage
3    39.803279
4    35.540984
2    16.852459
5     5.377049
1     2.426230
Name: economic.cond.national, dtype: float64
**********************************************************************
```

Observation:

- The Household economic condition is similar to the national economic conditions.

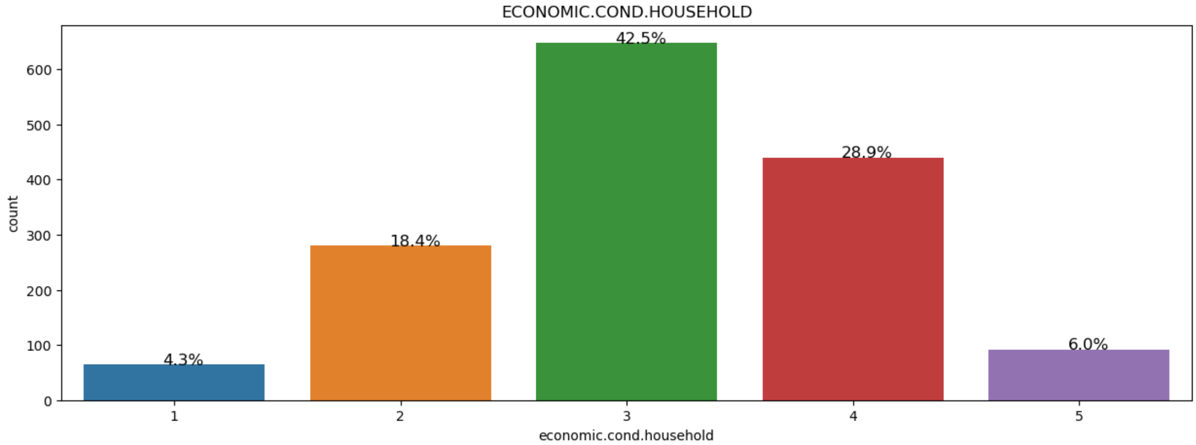

Figure C Countplot of Household Economic conditions

```
ECONOMIC.COND.HOUSEHOLD
3    648
4    440
2    280
5     92
1     65
Name: economic.cond.household, dtype: int64
economic.cond.household  in Percentage
3    42.491803
4    28.852459
2    18.360656
5     6.032787
1     4.262295
Name: economic.cond.household, dtype: float64
**********************************************************************
```

Observation:

- Among Hague and Blair, Blair has the highest supporters.

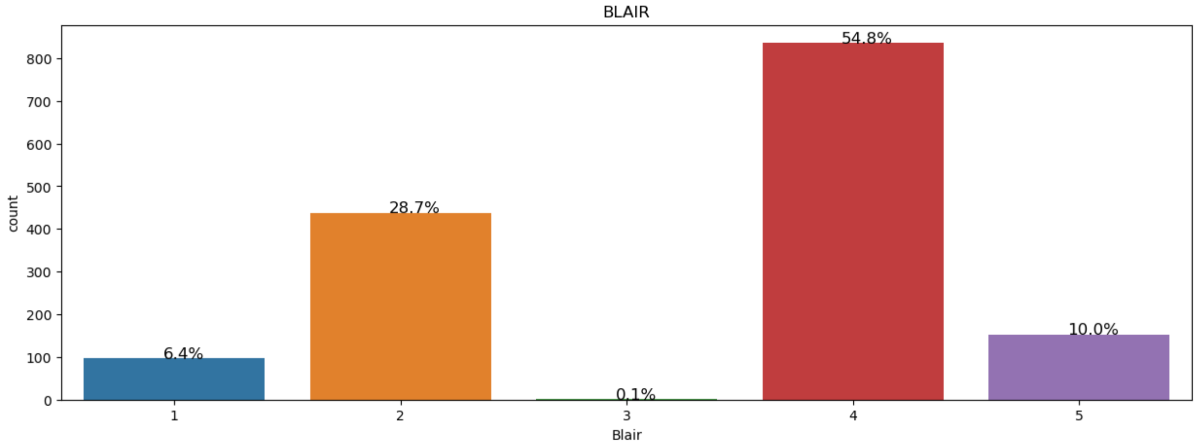

Figure D Countplot of Blair Assessment

```
BLAIR
4    836
2    438
5    153
1     97
3      1
Name: Blair, dtype: int64
Blair  in Percentage
4    54.819672
2    28.721311
5    10.032787
1     6.360656
3     0.065574
Name: Blair, dtype: float64
***********************************************************************
```

Observation:

- Hague has more respondents who feel not strongly about him than the respondents who supports him.
- This is an expected behaviour of respondents as they are supporting Blair.

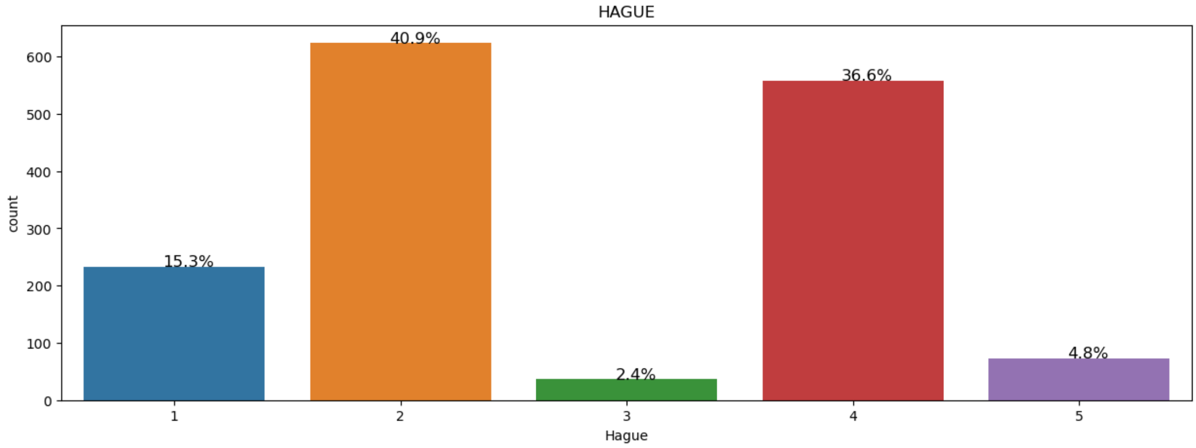

Figure E Countplot of Hague Assessment

```
HAGUE
2    624
4    558
1    233
5     73
3     37
Name: Hague, dtype: int64
Hague  in Percentage
2    40.918033
4    36.590164
1    15.278689
5     4.786885
3     2.426230
Name: Hague, dtype: float64
**********************************************************************
```

Observation:

- The respondents have mixed opinion towards the European Integration i.e. they are not strong supporting or opposing in terms of attitude.

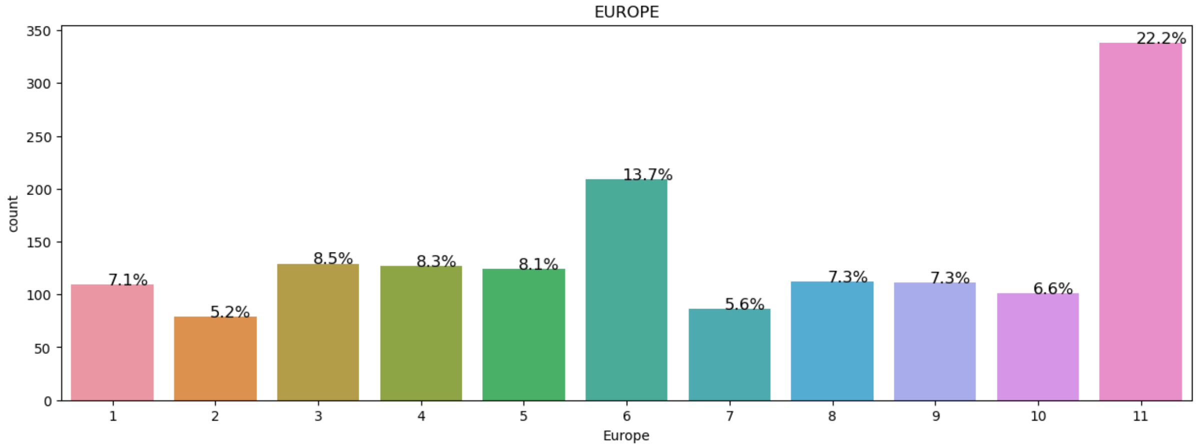

Figure F Countplot of Attitude towards European Integration

```
EUROPE
11    338
6     209
3     129
4     127
5     124
8     112
9     111
1     109
10    101
7      86
2      79
Name: Europe, dtype: int64
Europe  in Percentage
11    22.163934
6     13.704918
3      8.459016
4      8.327869
5      8.131148
8      7.344262
9      7.278689
1      7.147541
10     6.622951
7      5.639344
2      5.180328
Name: Europe, dtype: float64
**********************************************************************
```

Observation:

- Most of the respondents have an basic political Knowledge.

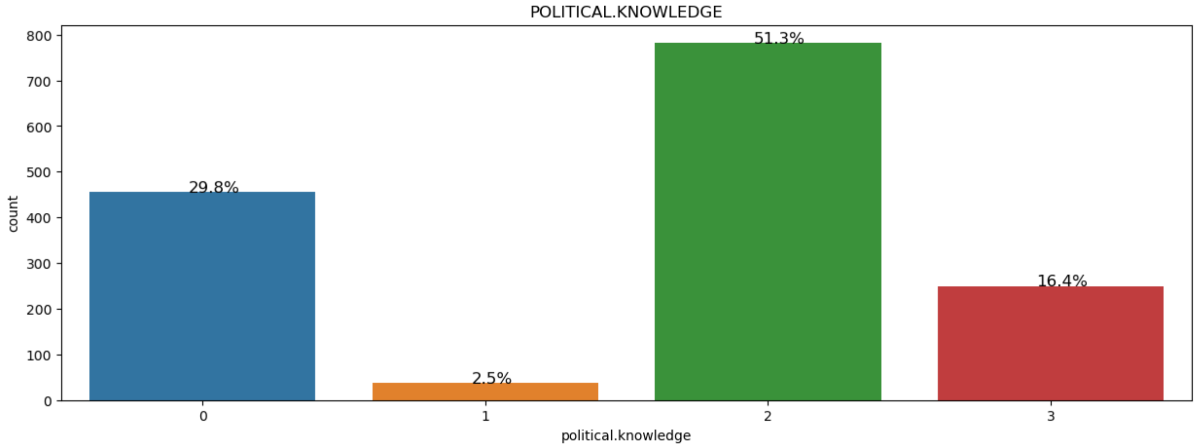

Figure G Countplot of Political Knowledge

```
POLITICAL.KNOWLEDGE
2    782
0    455
3    250
1     38
Name: political.knowledge, dtype: int64
political.knowledge  in Percentage
2    51.278689
0    29.836066
3    16.393443
1     2.491803
Name: political.knowledge, dtype: float64
**********************************************************************
```

Observation:

- Female respondents are higher in the survey.
- Gender Ratio is 1.13 in the survey.

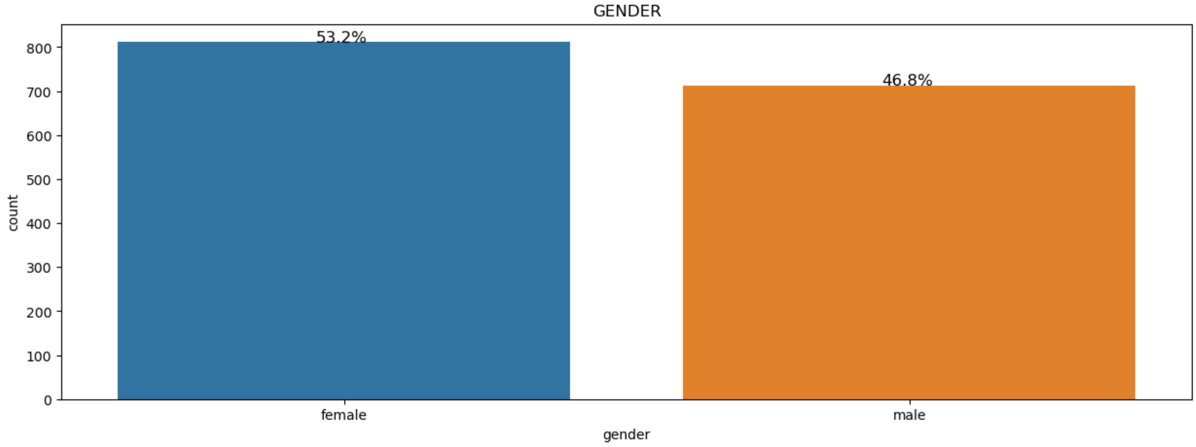

Figure H Countplot of gender

```
GENDER
female    812
male      713
Name: gender, dtype: int64
gender  in Percentage
female    53.245902
male      46.754098
Name: gender, dtype: float64
***********************************************************************
```

Histogram:

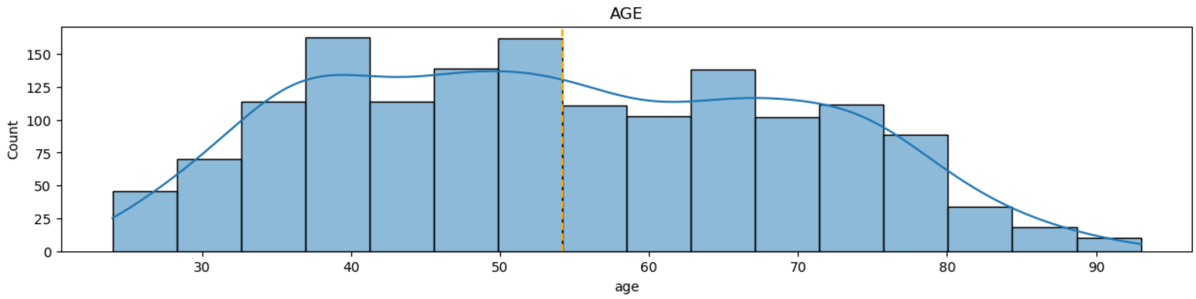

Figure I Histogram of age

The Histogram supports the skewness. It is clearly visible that the older age respondents are more in the survey than the young respondents.

Boxplot:

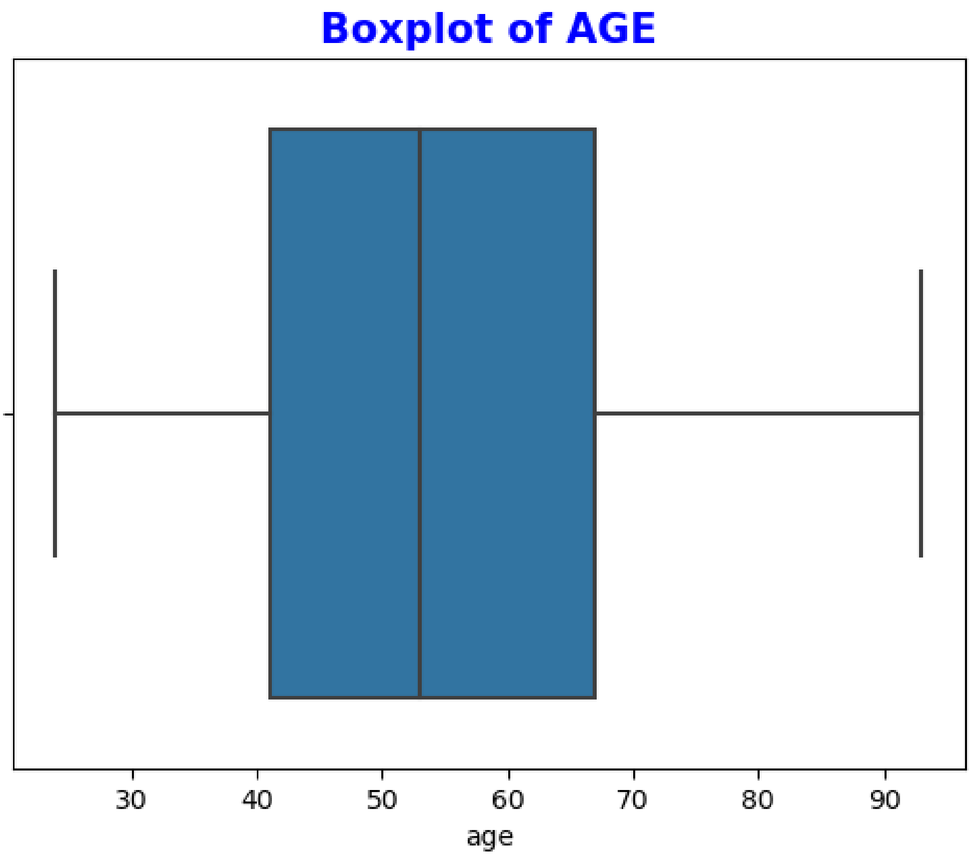

Figure J Boxplot of Age

There is no outlier in the dataset. Outlier treatment won't be necessary. This also supports the histogram and skewness.

Bivariate:

Histogram:

- The age of the respondents is not having much of a difference across two party supporters.

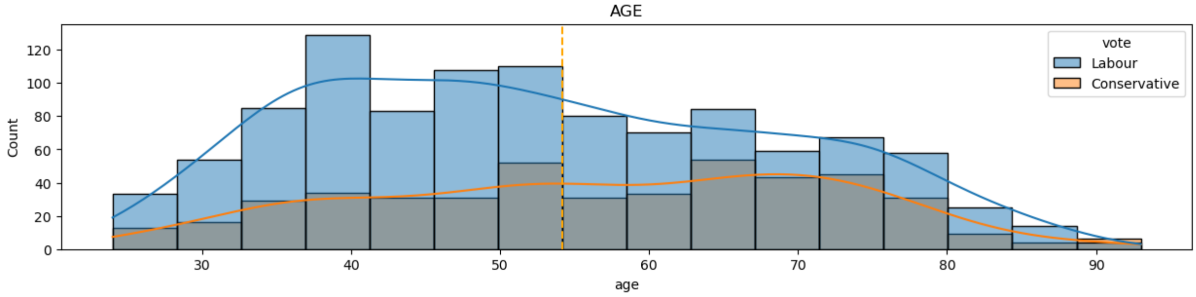

Figure K Histogram of Age with Vote

Countplot:

- The respondents support towards the two party is clearly visible in the below charts.
- Blair's supporters are voting to the labour party and Hague's are voting to the conservative party.

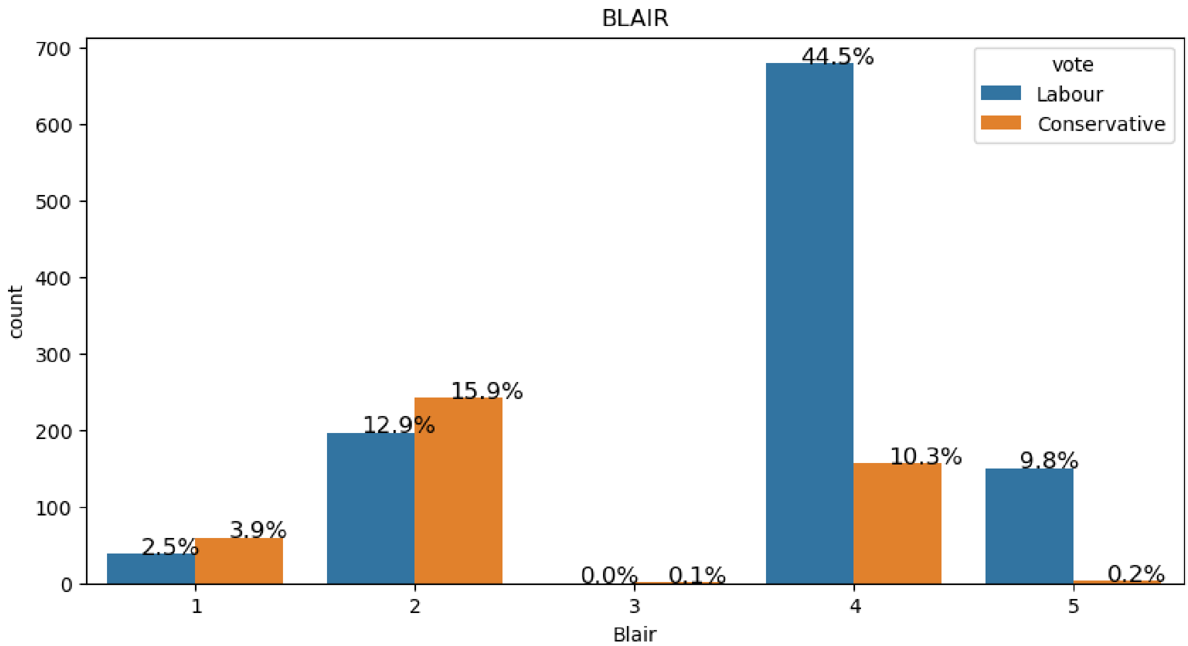

Figure L Countplot of Blair Assessment with Vote

- Nearly Half of the respondents who feel good about hague are voting for the labour party too.

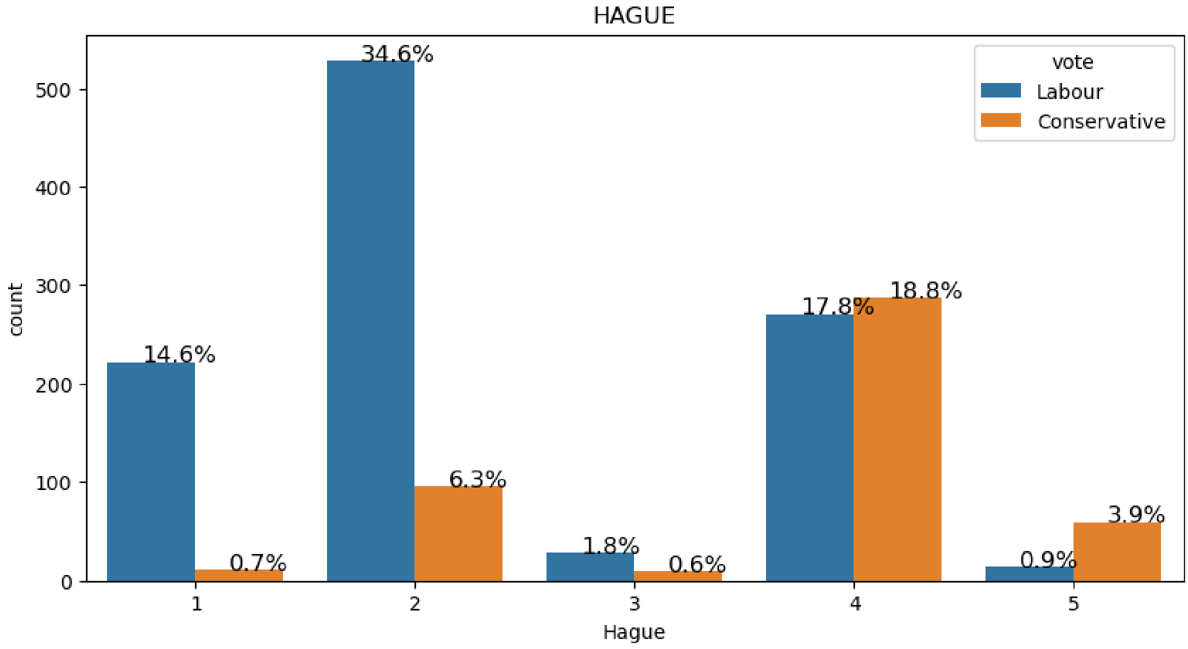

Figure M Countplot of Hague Assessment with Vote

- The male and female supporters of both party are nearly same.

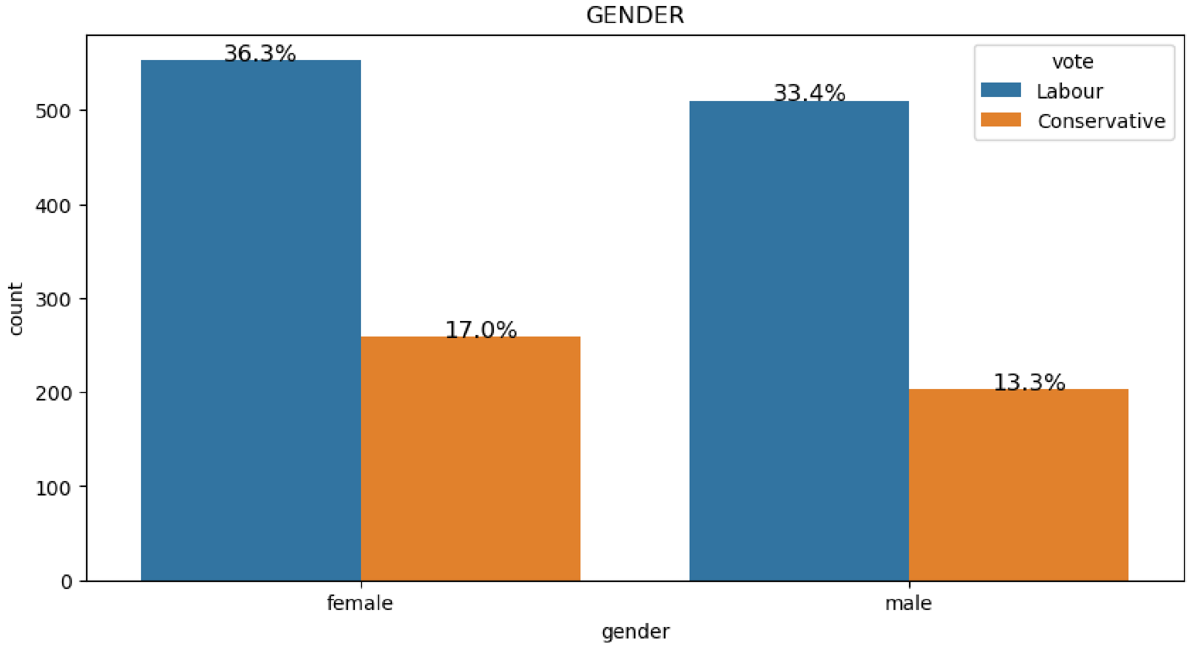

Figure N Countplot of Gender with vote

Boxplot:

- The mean age of categories is more or less the same.
- There is not much of a difference except a few.

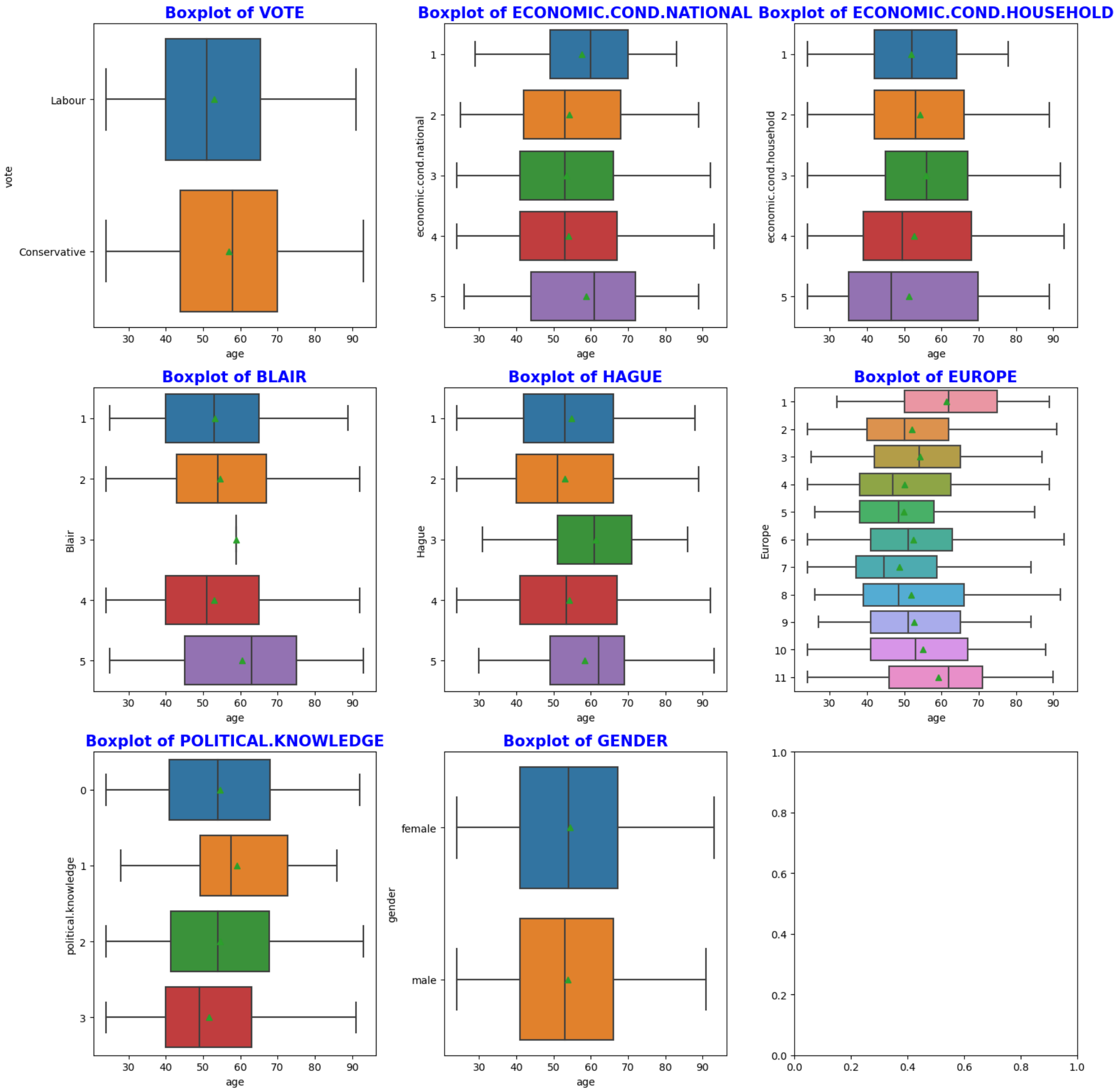

Figure O Boxplot of Age with Categorical variables

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

Encoding:

Label encoding is used for encoding.

The dataset has been encoded with 0,1 for Vote and gender.

Male is mapped as 1 and Female as 0.

Labour is Mapped as 0 and Conservative as 1.

Data Split:

The Dataset is split into train and test with 70:30 ratio.

The X train and X test will have 8 columns. Y Train and Y test will have only one column (Target variable).

The Train data has 1067 Rows and test will have 458 rows.

Scaling:

The Scaling is not necessary. There is no need to scale the data except for the KNN algorithms which measures the distance for algorithm calculation.

Except age, all of the numerical variables are categorical.

For this dataset, scaling is not that much important with these categorical variables.

During optimisation, before and after scaling models are tested to see if the model is improving due to scaling.

Before and After Scaling - Log

	Model	Accuracy	Precision-Labour	Recall-Labour	F1_Score-Labour	Precision-Conservative	Recall-Conservative	F1_Score-Conservative
0	Log	82.31	86.65	89.02	87.82	70.25	65.38	67.73
1	Log_scaled	82.10	86.61	88.72	87.65	69.67	65.38	67.46

Table 8 Model Results - Before and After Scaling for Log Reg

Before and After Scaling - Log

	Model	Accuracy	Precision-Labour	Recall-Labour	F1_Score-Labour	Precision-Conservative	Recall-Conservative	F1_Score-Conservative
2	LDA	81.88	86.79	88.11	87.44	68.80	66.15	67.45
3	LDA_scaled	81.66	86.53	88.11	87.31	68.55	65.38	66.93

Table 9 Model Result - Before and After Scaling for LDA

There is no much difference between the two models. So scaling is not necessary except the distance calculation algorithms like KNN.

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

Logistic Regression:

A simple Logistic Regression is applied without any tuning or parameters.

As we are interested in predicting both the party vote, F1-Score is taken as the model performance parameter.

The Classification Report of Simple Logistic regression is:

```
Train Accuracy: 0.8397375820056232
Test Accuracy: 0.8231441048034934

Classification Report Train
              precision    recall  f1-score   support

           0       0.87      0.91      0.89       735
           1       0.77      0.69      0.73       332

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


Classification Report Test
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       328
           1       0.70      0.65      0.68       130

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.78       458
weighted avg       0.82      0.82      0.82       458
```

Table 10 Classification Report of Log Reg

The Accuracy of Log Reg is around 80%.

The Labour Party Vote is predicted well in both train and test. The Conservative Party Vote is not performing well in the test (compared with train) and also in training when compare with Labour category.

LDA:

A Simple Linear Discriminant Analysis is applied without tuning for building the model.

The Classification Report of LDA is:

```
Train Accuracy: 0.8369259606373008
Test Accuracy: 0.8187772925764192

Classification Report Train
              precision    recall  f1-score   support

           0       0.87      0.90      0.88       735
           1       0.76      0.70      0.73       332

    accuracy                           0.84      1067
   macro avg       0.81      0.80      0.81      1067
weighted avg       0.83      0.84      0.84      1067


Classification Report Test
              precision    recall  f1-score   support

           0       0.87      0.88      0.87       328
           1       0.69      0.66      0.67       130

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.77       458
weighted avg       0.82      0.82      0.82       458
```

Table 11 Classification Report of LDA

The results are similar to Log Regression.

The Accuracy of LDA is also around 80%. Log Reg is performing a bit well when compared with LDA in unknown dataset.

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

KNN:

The X train and X test are scaled as the algorithm has distance calculation.

Scaled Data Descriptive Stats

	count	mean	std	min	25%	50%	75%	max
age	1067.0	1.664814e-17	1.000469	-1.934589	-0.858459	-0.035537	0.850687	2.433231
economic.cond.national	1067.0	4.827961e-17	1.000469	-2.531749	-0.295510	-0.295510	0.822609	1.940728
economic.cond.household	1067.0	2.297444e-16	1.000469	-2.317265	-0.156914	-0.156914	0.923261	2.003437
Blair	1067.0	-8.657034e-17	1.000469	-2.005100	-1.153072	0.550983	0.550983	1.403011
Hague	1067.0	1.465037e-16	1.000469	-1.439264	-0.623273	-0.623273	1.008709	1.824700
Europe	1067.0	-6.492776e-17	1.000469	-1.719829	-0.811534	-0.206005	1.005054	1.307819
political.knowledge	1067.0	-5.327406e-17	1.000469	-1.465782	-1.465782	0.411756	0.411756	1.350525
gender	1067.0	-9.406201e-17	1.000469	-0.953292	-0.953292	-0.953292	1.048997	1.048997

Table 12 Descriptive Stats after scaling

The StandardScaler is used for scaling the data. The data are scaled to mean 0 and std dev of 1.

KNNClassifier is used for the model with minkowski distance as the default metric.

```
Train Accuracy: 0.8631677600749765
Test Accuracy: 0.8275109170305677

Classification Report Train
              precision    recall  f1-score   support

           0       0.89      0.92      0.90       735
           1       0.80      0.75      0.77       332

    accuracy                           0.86      1067
   macro avg       0.84      0.83      0.84      1067
weighted avg       0.86      0.86      0.86      1067


Classification Report Test
              precision    recall  f1-score   support

           0       0.89      0.87      0.88       328
           1       0.69      0.72      0.70       130

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

Table 13 Classification Report of KNN

The Classification Report of simple KNN model outperforms both LDA and Log Regression models. Accuracy and F1-Score are better than the other two models. There might be a bit of overfitting incase of training model for KNN. This can be resolved in hyper tuning the models.

Naïve Bayes:

Gaussian Naïve Bayes is used for the Naïve Bayes model.

The Naïve Basyes model can be the baseline model for the other models.

```
Train Accuracy: 0.8331771321462043
Test Accuracy: 0.8253275109170306

Classification Report Train
              precision    recall  f1-score   support

           0       0.88      0.88      0.88       735
           1       0.74      0.72      0.73       332

    accuracy                           0.83      1067
   macro avg       0.81      0.80      0.80      1067
weighted avg       0.83      0.83      0.83      1067


Classification Report Test
              precision    recall  f1-score   support

           0       0.89      0.87      0.88       328
           1       0.68      0.72      0.70       130

    accuracy                           0.83       458
   macro avg       0.78      0.79      0.79       458
weighted avg       0.83      0.83      0.83       458
```

Table 14 Classification Report of NB

This can be used as a baseline parameter for other model performances.

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

Model Tuning:

The model has been hyper tuned for Logistic Regression, LDA and KNN.

Log Reg:

The Tuning parameter used for Log reg are

solver:['newton-cg','liblinear','lbfgs','sag','newton-cholesky', 'saga']

tol:[0.0001,0.00001,0.000001,0.000001]

C:[100, 10, 1.0, 0.1, 0.01,0.001]

After Tuning, the optimised model is having parameter of

```
LogisticRegression
LogisticRegression(C=100, max_iter=10000, n_jobs=-1, solver='saga', tol=1e-06)
```

The Classification Report for Optimised model is

```
Train Accuracy: 0.8406747891283973
Test Accuracy: 0.8231441048034934

Classification Report Train
              precision    recall  f1-score   support

           0       0.87      0.91      0.89       735
           1       0.77      0.69      0.73       332

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


Classification Report Test
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       328
           1       0.70      0.65      0.68       130

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.78       458
weighted avg       0.82      0.82      0.82       458
```

Table 15 Classification Report of Tuned Log Reg

There is not any improvement with the optimised model. Both simple and optimised model of Log Reg has same Results.

LDA:

The Tuning parameter used for LDA are

solver: ['svd','lsqr','eigen']

The Classification Report for Optimised model is

```
Train Accuracy: 0.8369259606373008
Test Accuracy: 0.8187772925764192

Classification Report Train
              precision    recall  f1-score   support

           0       0.87      0.90      0.88       735
           1       0.76      0.70      0.73       332

    accuracy                           0.84      1067
   macro avg       0.81      0.80      0.81      1067
weighted avg       0.83      0.84      0.84      1067


Classification Report Test
              precision    recall  f1-score   support

           0       0.87      0.88      0.87       328
           1       0.69      0.66      0.67       130

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.77       458
weighted avg       0.82      0.82      0.82       458
```

Table 16 Classification Report of Tuned LDA

Same as Log Reg, there is not any improvement in the optimised model.

KNN:

The KNN Tuning parameters are

n_neighbors : list(range(1,50,1))

weights: ['uniform','distance']

metric: ['euclidean','chebyshev','manhattan']

After Tuning, the optimised model parameters are

```
{'model_KNN__metric': 'euclidean',
 'model_KNN__n_neighbors': 15,
 'model_KNN__weights': 'uniform'}
```

The Classification Report for Optimised model is

```
Train Accuracy: 0.8416119962511716
Test Accuracy: 0.8209606986899564

Classification Report Train
              precision    recall  f1-score   support

           0       0.87      0.91      0.89       735
           1       0.77      0.70      0.73       332

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


Classification Report Test
              precision    recall  f1-score   support

           0       0.87      0.88      0.88       328
           1       0.69      0.68      0.68       130

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

Table 17 Classification Report of Tuned KNN

The KNN optimised model solves the over fitting issue in basic KNN model. The optimised model is not improved.

Bagging:

Random Forest:

The random Forest Model is also a type of Bagging which uses Decision Tree.

The model parameter after tuning is

```
{'n_estimators': 680,
 'min_samples_leaf': 3,
 'max_samples': 0.1,
 'max_features': 6,
 'criterion': 'gini'}
```

The Classification Report for the model is

```
Train Accuracy: 0.8481724461105904
Test Accuracy: 0.8253275109170306

Classification Report Train
              precision    recall  f1-score   support

           0       0.86      0.93      0.89       735
           1       0.81      0.67      0.73       332

    accuracy                           0.85      1067
   macro avg       0.83      0.80      0.81      1067
weighted avg       0.85      0.85      0.84      1067


Classification Report Test
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       328
           1       0.71      0.65      0.68       130

    accuracy                           0.83       458
   macro avg       0.79      0.77      0.78       458
weighted avg       0.82      0.83      0.82       458
```

Table 18 Classification Report of RF

The Feature importance for Random Forest is

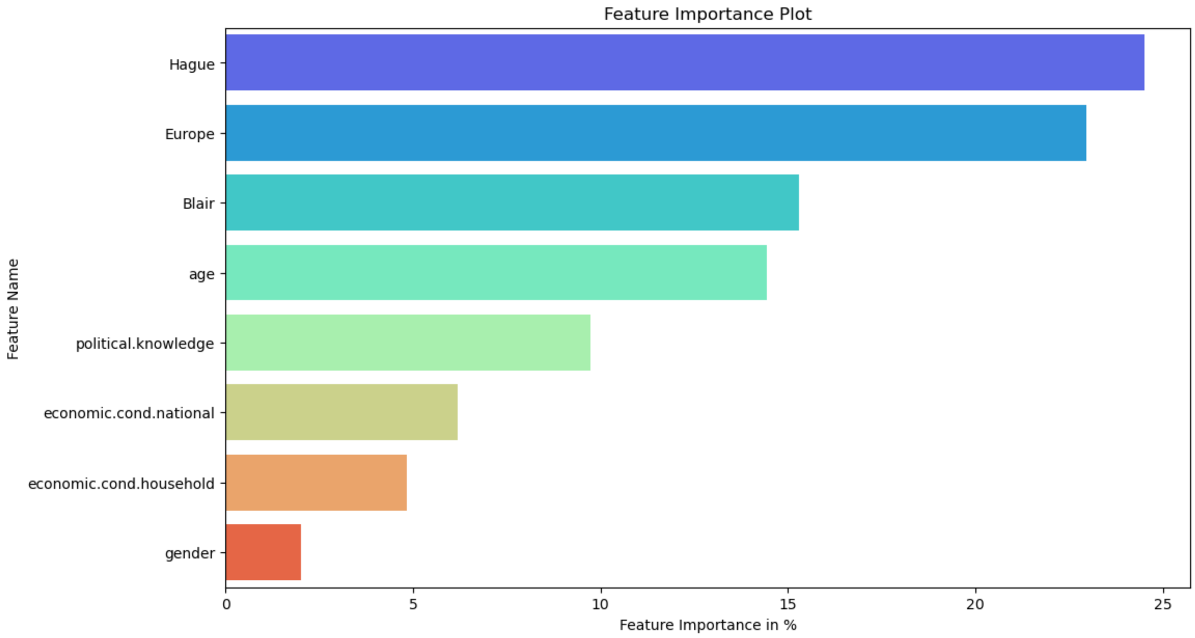

Figure P Feature Importance Plot

	Imp
Hague	0.245114
Europe	0.229691
Blair	0.153047
age	0.144254
political.knowledge	0.097370
economic.cond.national	0.061942
economic.cond.household	0.048343
gender	0.020239

Table 19 Feature Importance Value

LDA - Bagging:

The Bootstrap Aggregation is using LDA as the model.

The Optimised parameter for LDA Bagging is

```
{'n_estimators': 900, 'max_samples': 0.9, 'max_features': 0.91}
```

The Classification Report for LDA Bagging is

```
Train Accuracy: 0.8388003748828491
Test Accuracy: 0.8253275109170306

Classification Report Train
              precision    recall  f1-score   support

           0       0.87      0.91      0.89       735
           1       0.77      0.69      0.73       332

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


Classification Report Test
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       328
           1       0.71      0.65      0.68       130

    accuracy                           0.83       458
   macro avg       0.79      0.77      0.78       458
weighted avg       0.82      0.83      0.82       458
```

Table 20 Classification Report of LDAB

LDA Bagging is also in the same performance range of accuracy as others. It might have some slight changes based on the preference we needed in terms of Precision, Recall or F1-Score.

Boosting:

XG Boosting is used for this model.

The parameter after tuning is

```
{'tol': 1e-05, 'n_estimators': 237, 'learning_rate': 0.013972000000003973}
```

The Classification report for XG Boosting model is

```
Train Accuracy: 0.915651358950328
Test Accuracy: 0.8144104803493449

Classification Report Train
              precision    recall  f1-score   support

           0       0.93      0.95      0.94       735
           1       0.89      0.83      0.86       332

    accuracy                           0.92      1067
   macro avg       0.91      0.89      0.90      1067
weighted avg       0.91      0.92      0.91      1067


Classification Report Test
              precision    recall  f1-score   support

           0       0.88      0.86      0.87       328
           1       0.67      0.69      0.68       130

    accuracy                           0.81       458
   macro avg       0.77      0.78      0.77       458
weighted avg       0.82      0.81      0.82       458
```

Table 21 Classification Report of XG Boosting

The XG Boosting performs well in the test data than other models. Comparing all the models will help to select a final model.

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

Train Vs Test:

Confusion Matrix:

The Confusion matrix has Predicted in X-axis (Columns) and Actual in Y-axis (Rows).

The confusion matrix for Train and Test for each model are below

Log Reg:

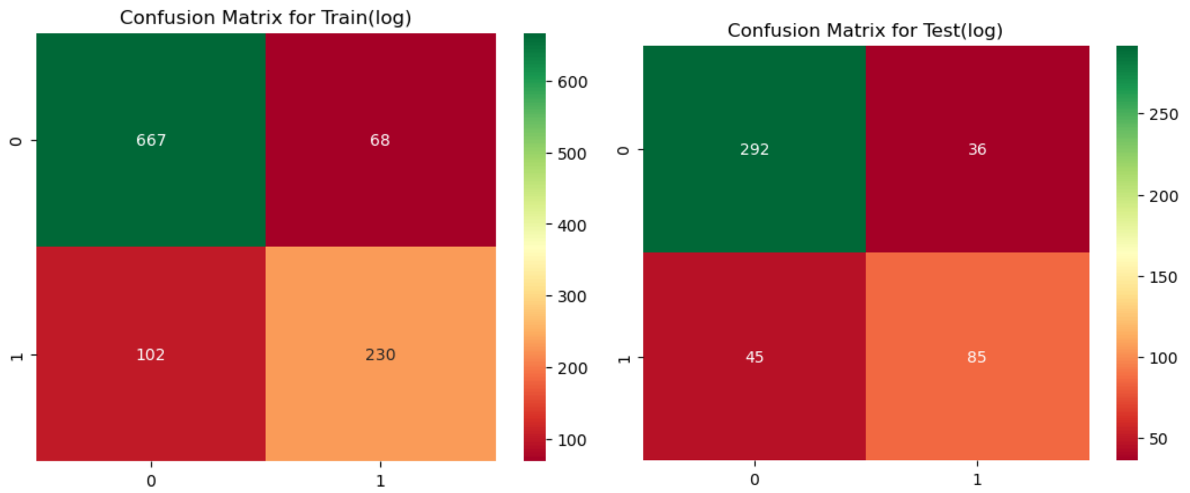

Figure Q Confusion matrix of Log Reg (Train vs Test)

LDA:

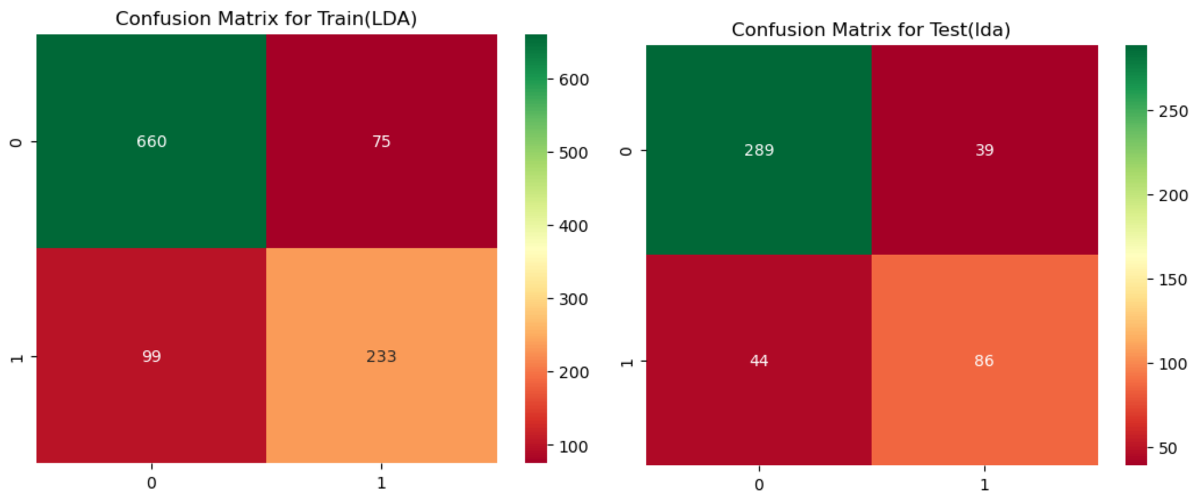

Figure R Confusion matrix of LDA (Train vs Test)

KNN:

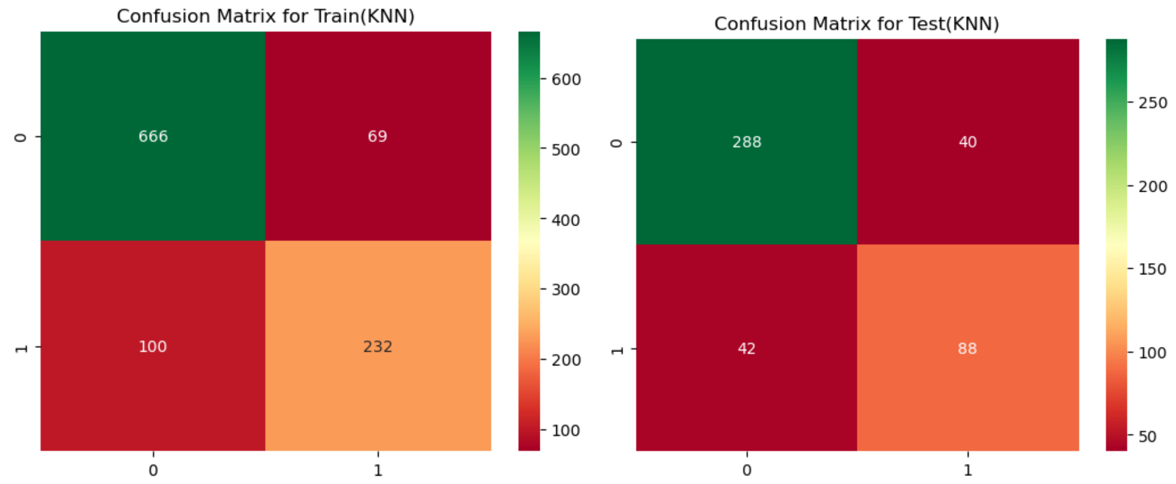

Figure S Confusion matrix of KNN (Train vs Test)

Naïve Bayes:

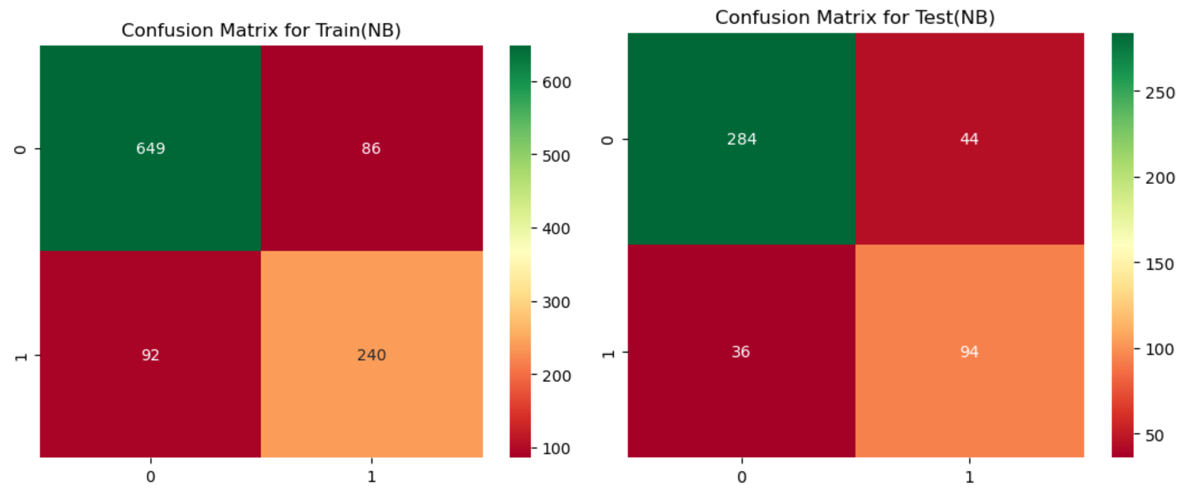

Figure T Confusion matrix of NB (Train vs Test)

Random Forest:

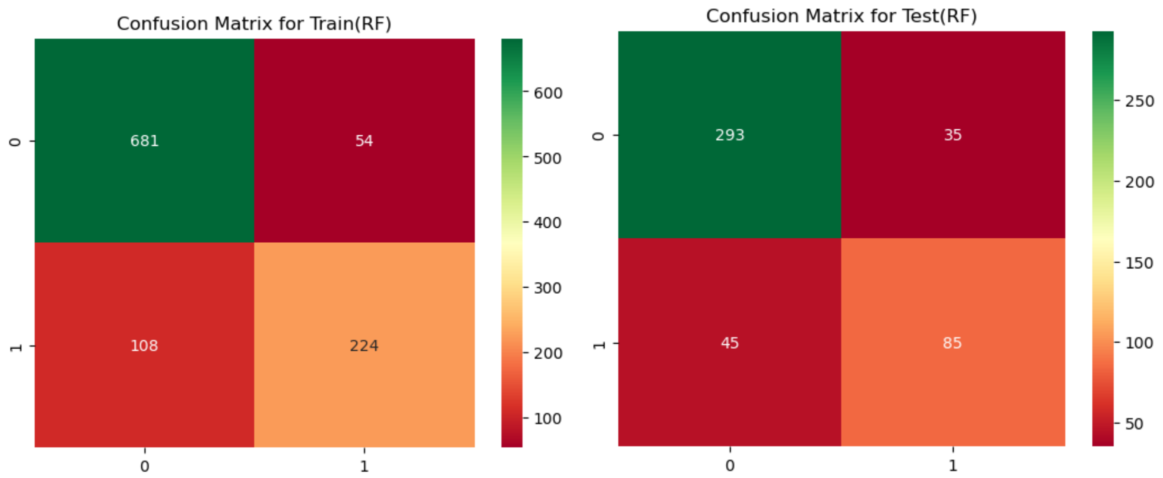

Figure U Confusion matrix of RF (Train vs Test)

LDA – Bagging:

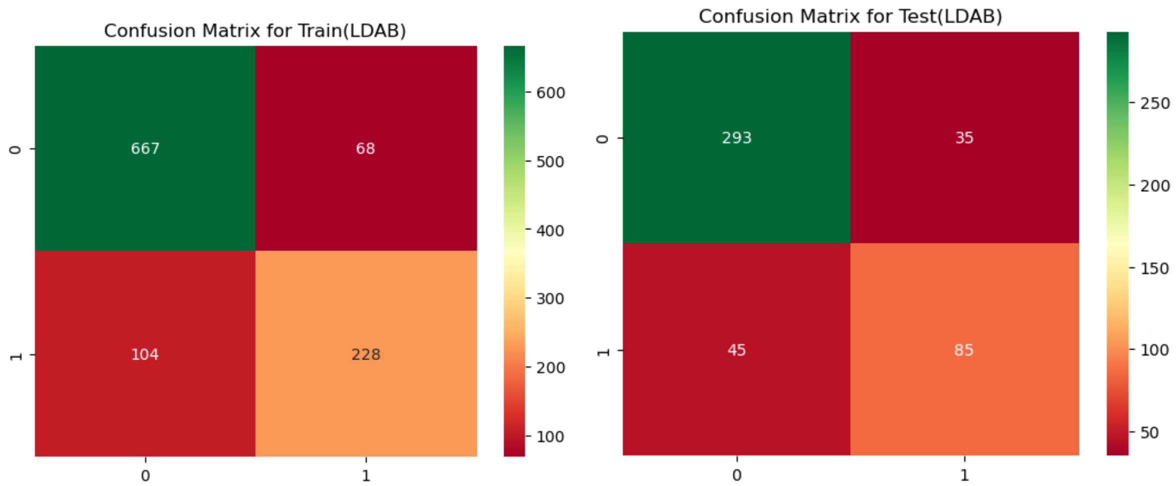

Figure V Confusion matrix of LDAB (Train vs Test)

Boosting:

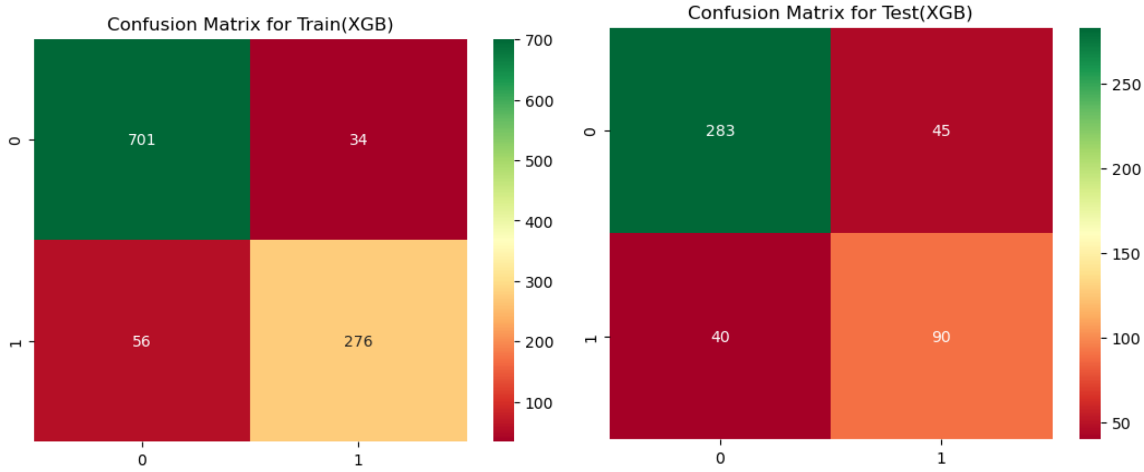

Figure W Confusion matrix of XGB (Train vs Test)

From the overall Confusion matrices, XG Boost is performing well in correctly predicting the vote for which label.

Roc:

Log Reg:

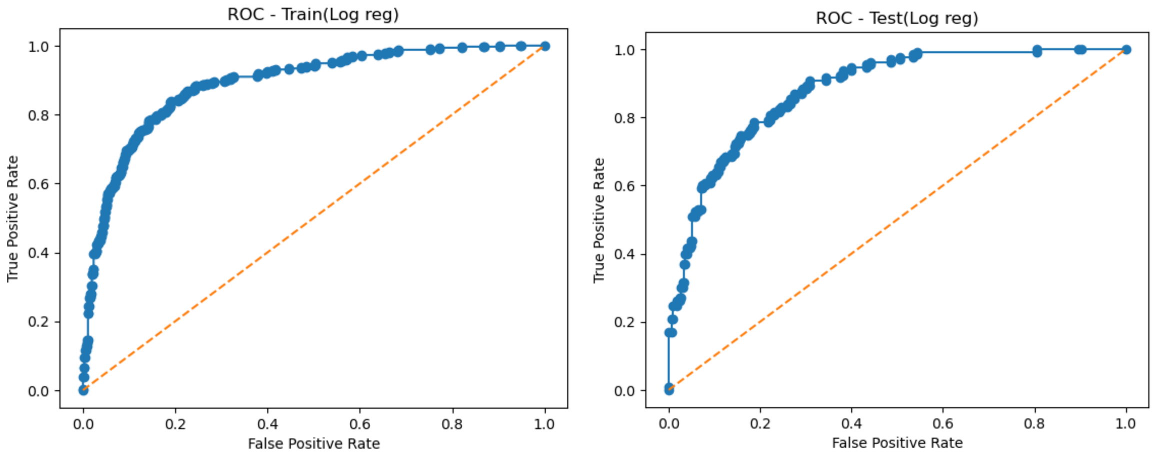

Figure X ROC of Log Reg (Train vs Test)

LDA:

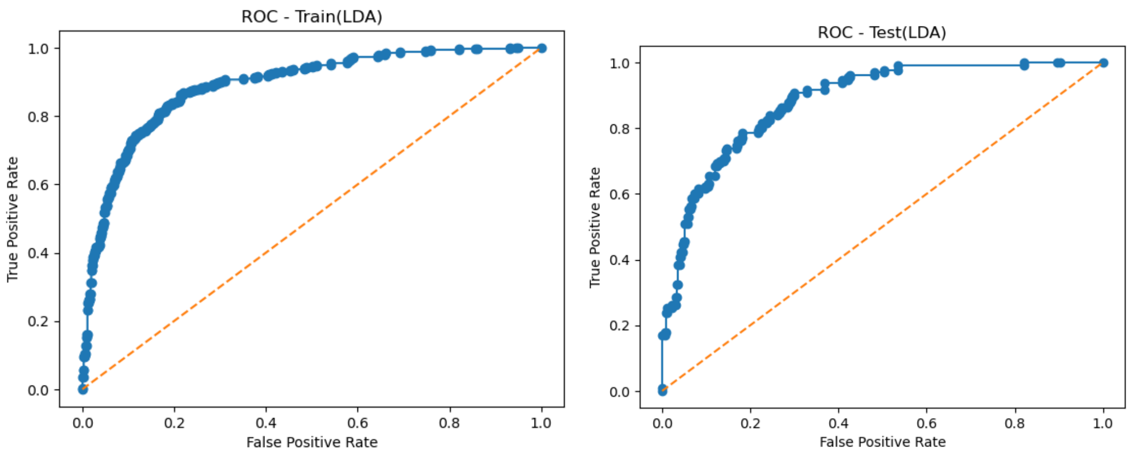

Figure Y ROC of LDA (Train vs Test)

KNN:

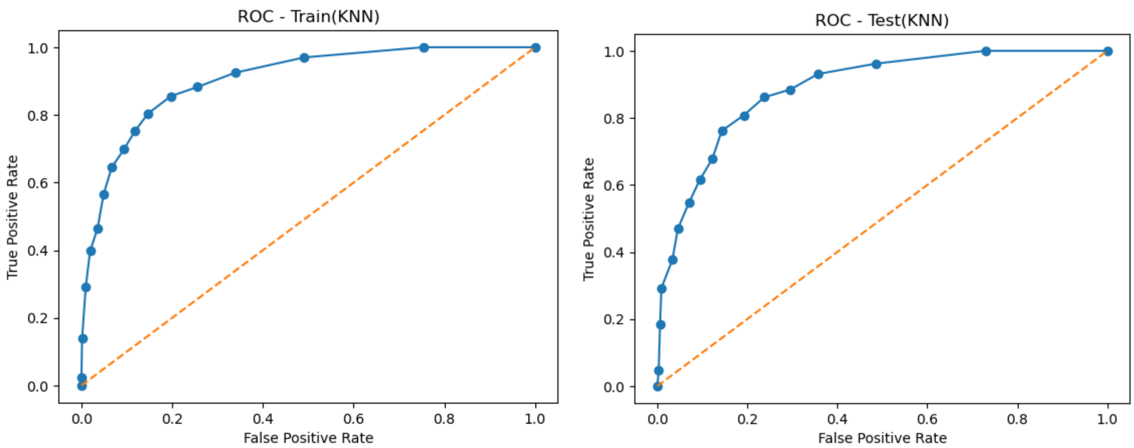

Figure Z ROC of KNN (Train vs Test)

Naïve Bayes:

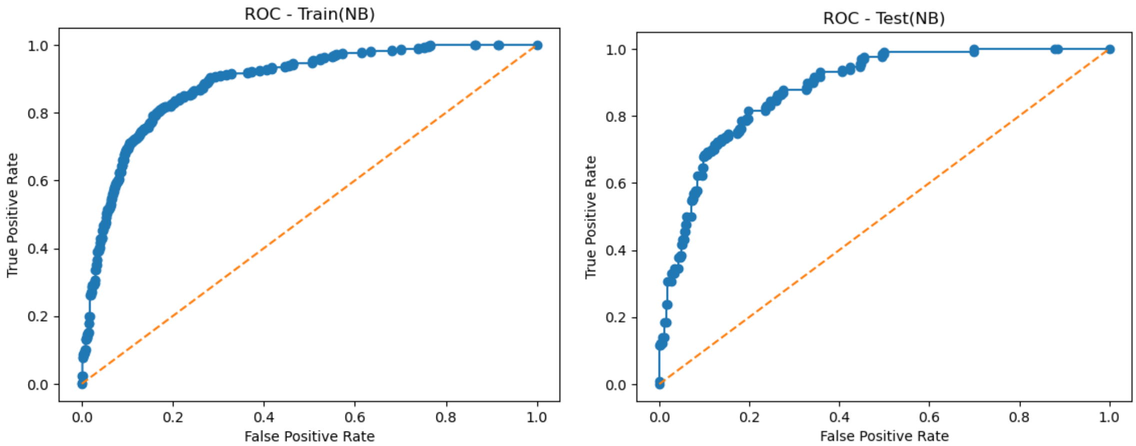

Figure AA ROC of NB (Train vs Test)

Random Forest:

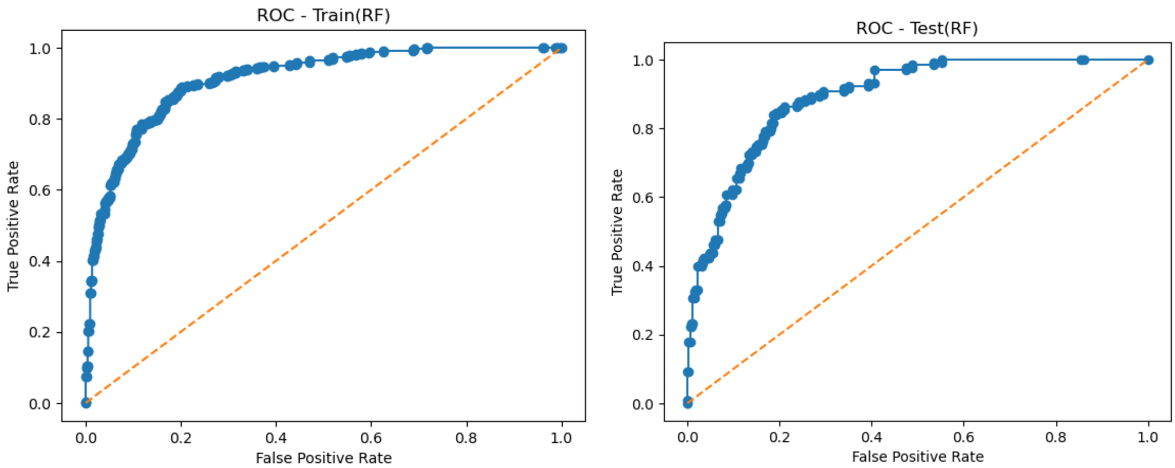

Figure BB ROC of RF (Train vs Test)

LDA – Bagging:

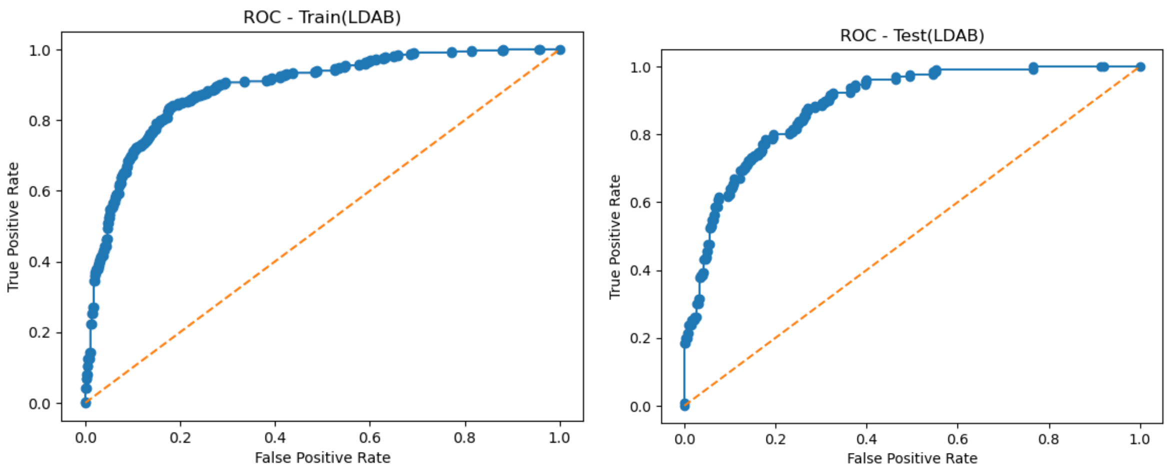

Figure CC ROC of LDAB (Train vs Test)

Boosting:

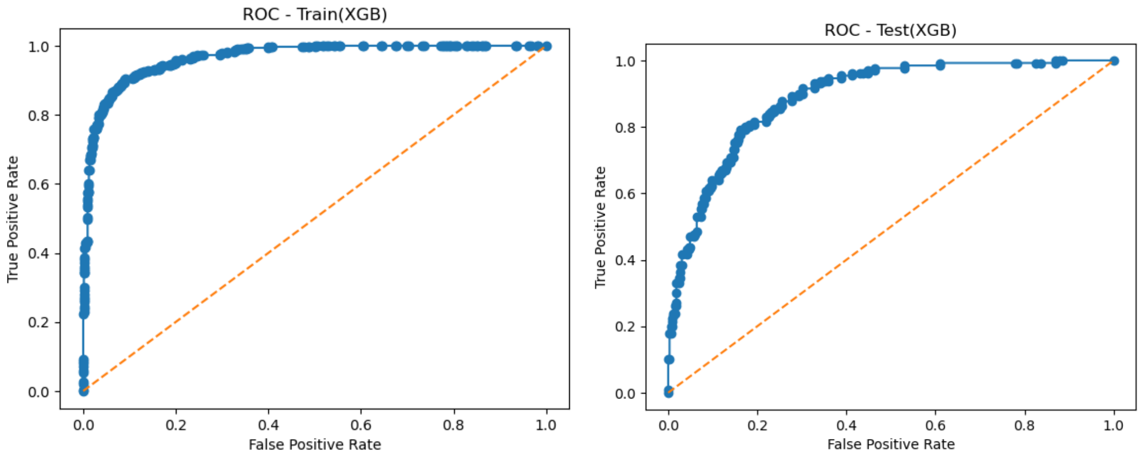

Figure DD ROC of XGB (Train vs Test)

ROC is also supporting the fact that Boosting performs well than the other models. But it is not very clear as all the models are having only a slight difference.

The Model performance scores can be used to decide the final model.

Train Scores:

Train Scores

	Model	Accuracy	Precision-Labour	Recall-Labour	F1_Score-Labour	Precision-Conservative	Recall-Conservative	F1_Score-Conservative	Auc
0	NB	83.32	87.58	88.30	87.94	73.62	72.29	72.95	88.65
1	Log	84.07	86.74	90.75	88.70	77.18	69.28	73.02	88.94
2	LDA	83.69	86.96	89.80	88.35	75.65	70.18	72.81	88.92
3	KNN	84.16	86.95	90.61	88.74	77.08	69.88	73.30	90.50
4	LDAB	83.88	86.51	90.75	88.58	77.03	68.67	72.61	88.93
5	XGB	91.57	92.60	95.37	93.97	89.03	83.13	85.98	96.63
6	RF	84.82	86.31	92.65	89.37	80.58	67.47	73.44	91.23

Table 22 Training Model Scores

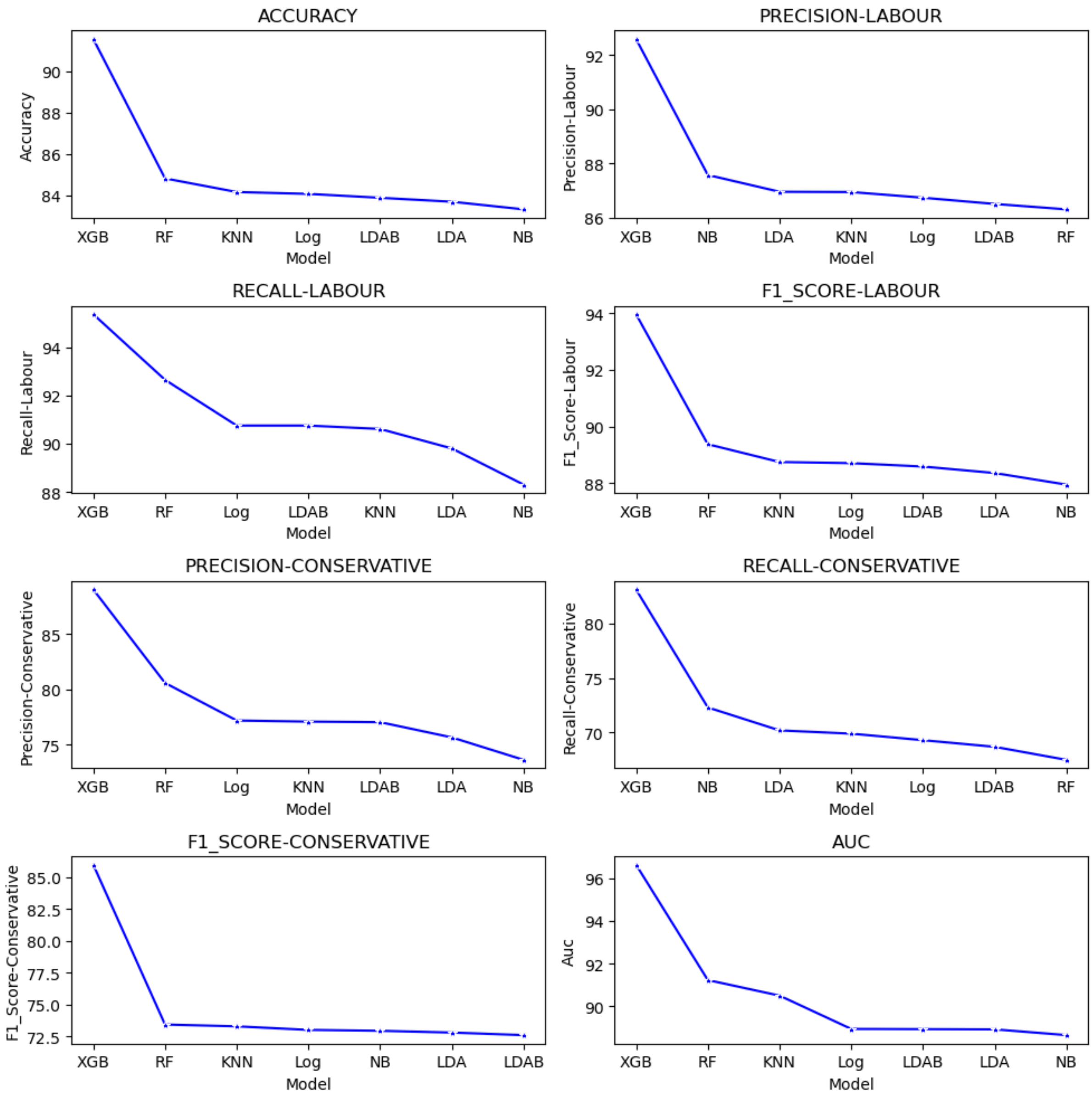

Figure EE Lineplot for Training model scores

From the Scores and the Graph, it can be visually easily understood that the XGB model performs well in training and it outperforms the baseline model i.e., NB

Test Scores:

Test Scores

	Model	Accuracy	Precision-Labour	Recall-Labour	F1_Score-Labour	Precision-Conservative	Recall-Conservative	F1_Score-Conservative	Auc
0	NB	82.53	88.75	86.59	87.65	68.12	72.31	70.15	88.45
1	Log	82.31	86.65	89.02	87.82	70.25	65.38	67.73	88.25
2	LDA	81.88	86.79	88.11	87.44	68.80	66.15	67.45	88.38
3	KNN	82.10	87.27	87.80	87.54	68.75	67.69	68.22	88.82
4	LDAB	82.53	86.69	89.33	87.99	70.83	65.38	68.00	88.66
5	XGB	81.44	87.62	86.28	86.94	66.67	69.23	67.92	88.92
6	RF	82.53	86.69	89.33	87.99	70.83	65.38	68.00	77.36

Table 23 Testing model scores

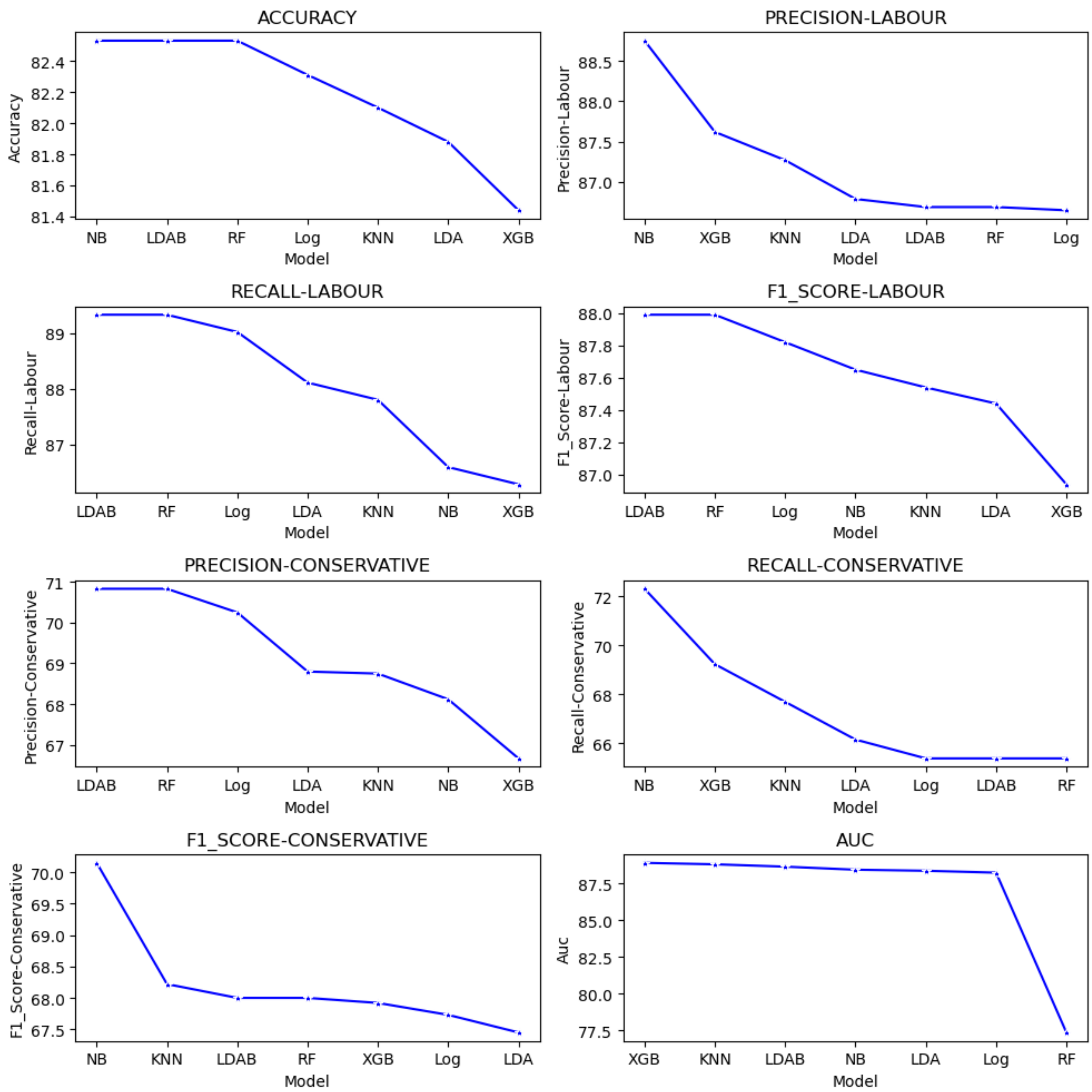

Figure FF Lineplot of testing model scores

Even though the XGB performs well in the training data, for the unknown dataset, both bagging methods - Random Forest and LDAB is the best model for prediction.

The F1-score and precision, recall of the RF are better in comparison with other models, even though the model has less AUC. So LDAB can be the final model as the AUC score is also high.

1.8 Based on these predictions, what are the insights?

Based on the predictions from the RF model, the insights for the Election are

- The vote of a respondent is based on the importance of his attitude towards these features with top 5 being the most important.
    - ✓ ***Hague***
    - ✓ ***Europe***
    - ✓ ***Blair***
    - ✓ ***age***
    - ✓ ***political.knowledge***
    - ✓ economic.cond.national
    - ✓ economic.cond.household
    - ✓ gender
- The Attitude of the people towards, Hague, Blair and EU influence the voters most.
- Hague needs to increase his popularity among the voters to win the election.
- Blair already has huge popularity.
- Blair has huge support among the voters and he has twice the vote as Hague.
- Hague supporters are split into two categories
    1. Supporters who have good opinion and vote for him
    2. Supporters who have good opinion and didn't vote for him.

    The second case has to be analysed further by the party to find the reason for those behaviours.

- Based on the predictions, Blair has the upper hand in the election. It is predicted that the Labour party might win with huge margins.
- Based on the Business Problem, whether to predict which party votes accurately, the model can be used for that problem statement with the respective Precision/Recall Measures.
- As a Final Model, RF and LDAB will the best ones to make predictions. With AUC also into consideration, LDAB can be used for better predictions.