PROJECT

MACHINE LEARNING

Table of contents

List of Figures

List of Tables

PROBLEM 1

Problem Statement

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

Data Dictionary

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

Q1.1. 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.

Sample of the 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 1. Sample of Election Dataset.

Basic Information of the Dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   vote                     1525 non-null   object
 1   age                      1525 non-null   int64 
 2   economic_cond_national   1525 non-null   int64 
 3   economic_cond_household  1525 non-null   int64 
 4   Blair                    1525 non-null   int64 
 5   Hague                    1525 non-null   int64 
 6   Europe                   1525 non-null   int64 
 7   political_knowledge      1525 non-null   int64 
 8   gender                   1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

Data Types of Variables:

Feature	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender
Data Type	object	int64	int64	int64	int64	int64	int64	int64	object

Table 2. Data Types of All Features in the Election Dataset.

Insights:

1. There are 9 features (columns) with 1525 observations (rows) in the dataset.
2. The dataset has one numerical variable i.e., age and all remaining variables are of categorical type.
3. Out of 8 categorical variables, vote and gender are nominal and their data type is object and remaining six variables are ordinal and their data type is int64.
4. The **target variable** in this dataset is **vote**.

Description of the Dataset

Continuous Numerical Features:

	count	mean	std	min	25%	50%	75%	max
age	1525.0	54.2	15.7	24.0	41.0	53.0	67.0	93.0

Table 3. Description of Numerical Features in Election Dataset.

Insights:

1. Minimum age of the voter is 24 years.
2. Maximum age of the voter is 93 years.

Categorical Features:

	count	unique	top	freq
economic_cond_national	1525	5	3	607
economic_cond_household	1525	5	3	648
Blair	1525	5	4	836
Hague	1525	5	2	624
Europe	1525	11	11	338
political_knowledge	1525	4	2	782
gender	1525	2	female	812

Table 4. Description of Categorical Features in Election Dataset.

Checking Null values in the Dataset:

	Number_of_Null_Values
Feature	
vote	0
age	0
economic_cond_national	0
economic_cond_household	0
Blair	0
Hague	0
Europe	0
political_knowledge	0
gender	0

Table 5. Null Values in the Election Dataset.

- There are **no null values** in the dataset.

Skewness & Kurtosis:

- Skewness is a measure of lack of symmetry in a distribution.
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

Feature	Skewness	Kurtosis
Age	0.14	-0.94

Table 6. Skewness and Kurtosis of Numeric Features in the Election Dataset.

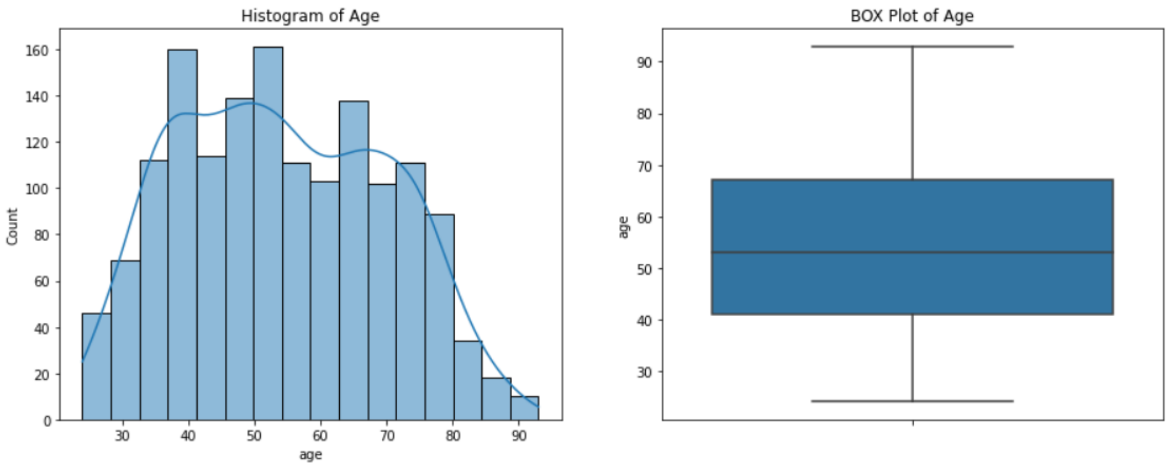


Figure 1. Histogram and Box Plot of Continuous Numerical Features in Election Dataset.

Insights:

From above plots and tables, we can conclude below points,

1. Age feature is slightly right skewed distribution (Positively skewness of 0.14).
2. Age feature has negative kurtosis (-0.94). It means that the data is heavy tailed relative normal distribution.

Q1.2. 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.

EXPLORATORY DATA ANALYSIS

- Shape of the dataset, data types of features and null values have been already discussed in Question 1.1.

Checking Duplicated Observations:

- There are 8 duplicate observations.
- Duplicate records are listed in below table along with their index.
- Duplicate records are dropped from the dataset by keeping the first original record.

	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 7. Duplicate Observations in the Election Dataset.

CHECKING FOR ANOMALIES

Checking Normalized Value Counts in Discrete Numerical and Categorical Variables

```
Feature:  economic_cond_national
3     39.82
4     35.46
2     16.88
5      5.41
1      2.44
Name: economic_cond_national, dtype: float64
---------------------------------------------
Feature:  economic_cond_household
3     42.52
4     28.68
2     18.46
5      6.06
1      4.28
Name: economic_cond_household, dtype: float64
---------------------------------------------
Feature:  Blair
4     54.91
2     28.61
5     10.02
1      6.39
3      0.07
Name: Blair, dtype: float64
---------------------------------------------
Feature:  Hague
2     40.67
4     36.72
1     15.36
5      4.81
3      2.44
Name: Hague, dtype: float64
```

```
Feature:  Europe
11     22.28
6      13.65
3       8.44
4       8.31
5       8.11
9       7.32
8       7.32
1       7.19
10      6.66
7       5.67
2       5.08
Name: Europe, dtype: float64
---------------------------------------------
Feature:  political_knowledge
2     51.15
0     29.93
3     16.41
1      2.50
Name: political_knowledge, dtype: float64
---------------------------------------------
Feature:  gender
female    53.26
male      46.74
Name: gender, dtype: float64
```

Checking Unique Entries in Discrete Numerical and Categorical Variables

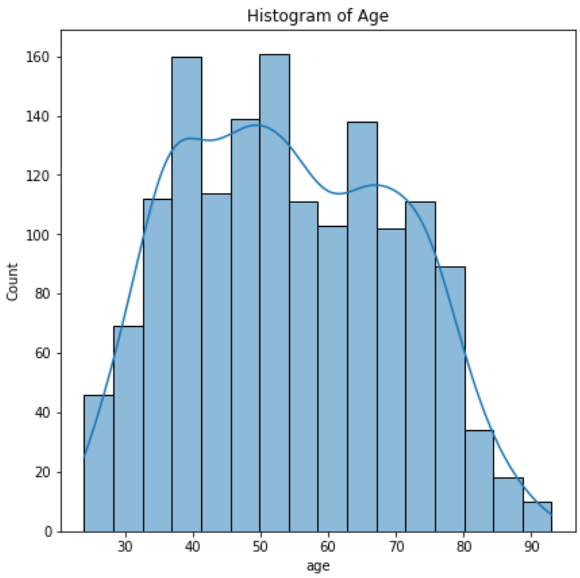
```
Feature:  economic_cond_national
[3 4 2 1 5]
----------------------------
Feature:  economic_cond_household
[3 4 2 1 5]
----------------------------
Feature:  Blair
[4 5 2 1 3]
----------------------------
Feature:  Hague
[1 4 2 5 3]
----------------------------
Feature:  Europe
[ 2  5  3  4  6 11  1  7  9 10  8]
----------------------------
Feature:  political_knowledge
[2 0 3 1]
----------------------------
Feature:  gender
['female' 'male']
```

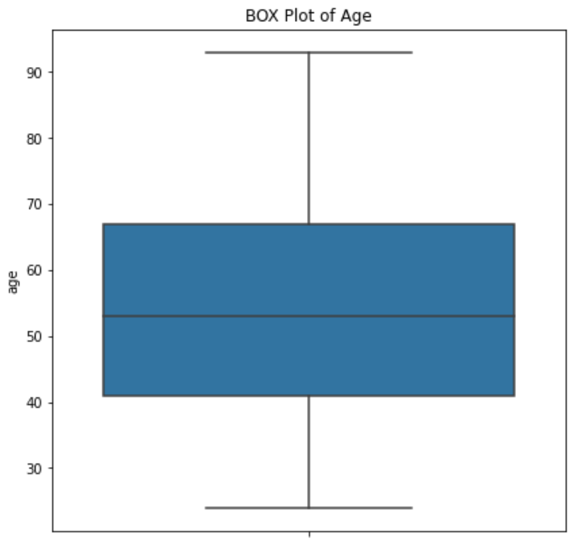
Insights

1. There are **no anomalies** in the sublevels of discrete numerical and categorical features.
2. There are **few sublevels with negligible count** in all the features except Europe and Gender features.

UNIVARIATE ANALYSIS

Histogram, Boxplot and Swarmplot of Continuous Numerical Features (Age)





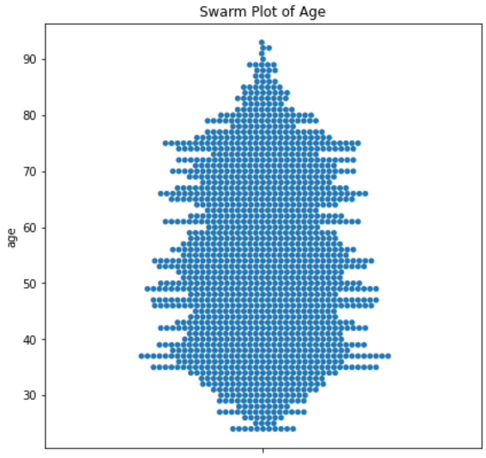


Figure 2. Histogram, Box Plot and Swarmplot of Continuous Numerical Features.

Histogram, Boxplot, Swarmplot and Bar Plot of Continuous Numerical Features with Vote as Hue

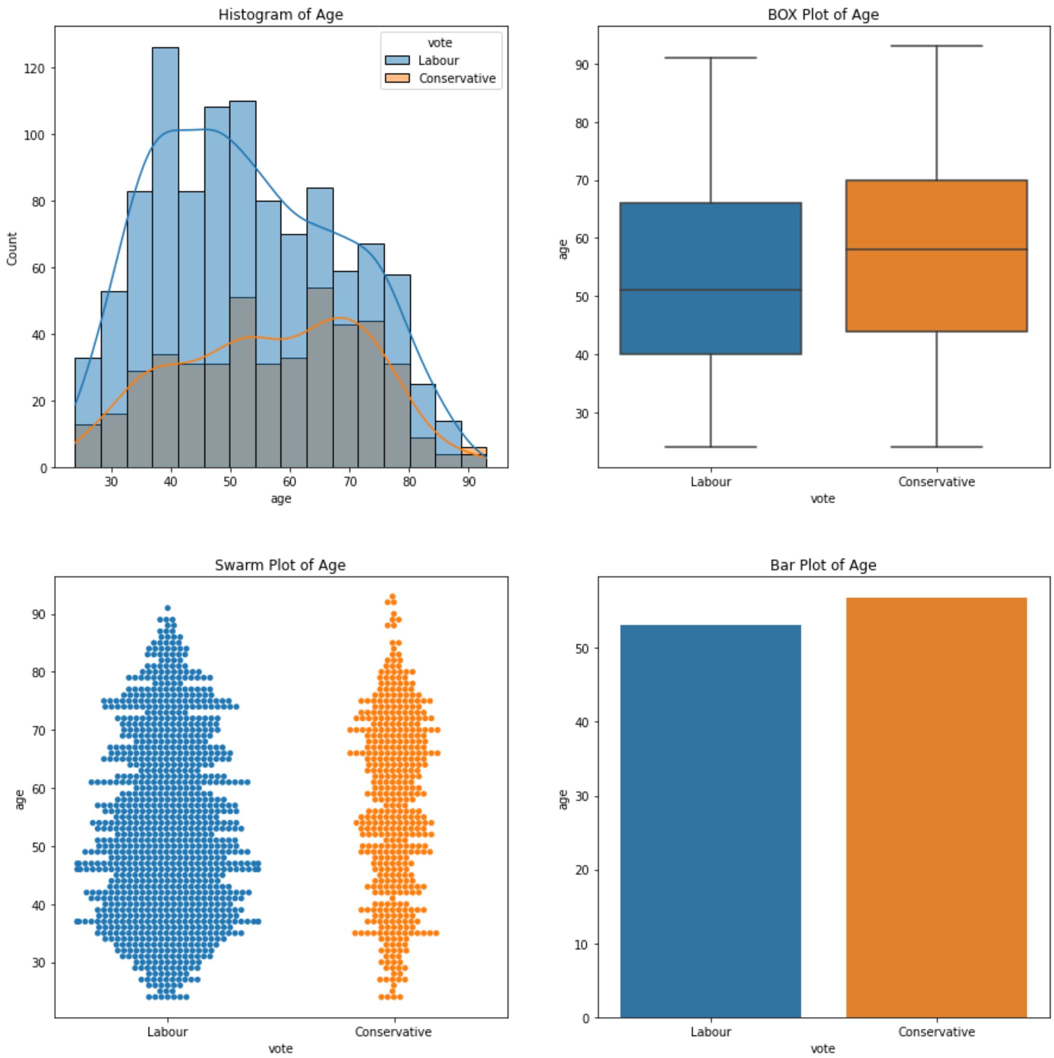
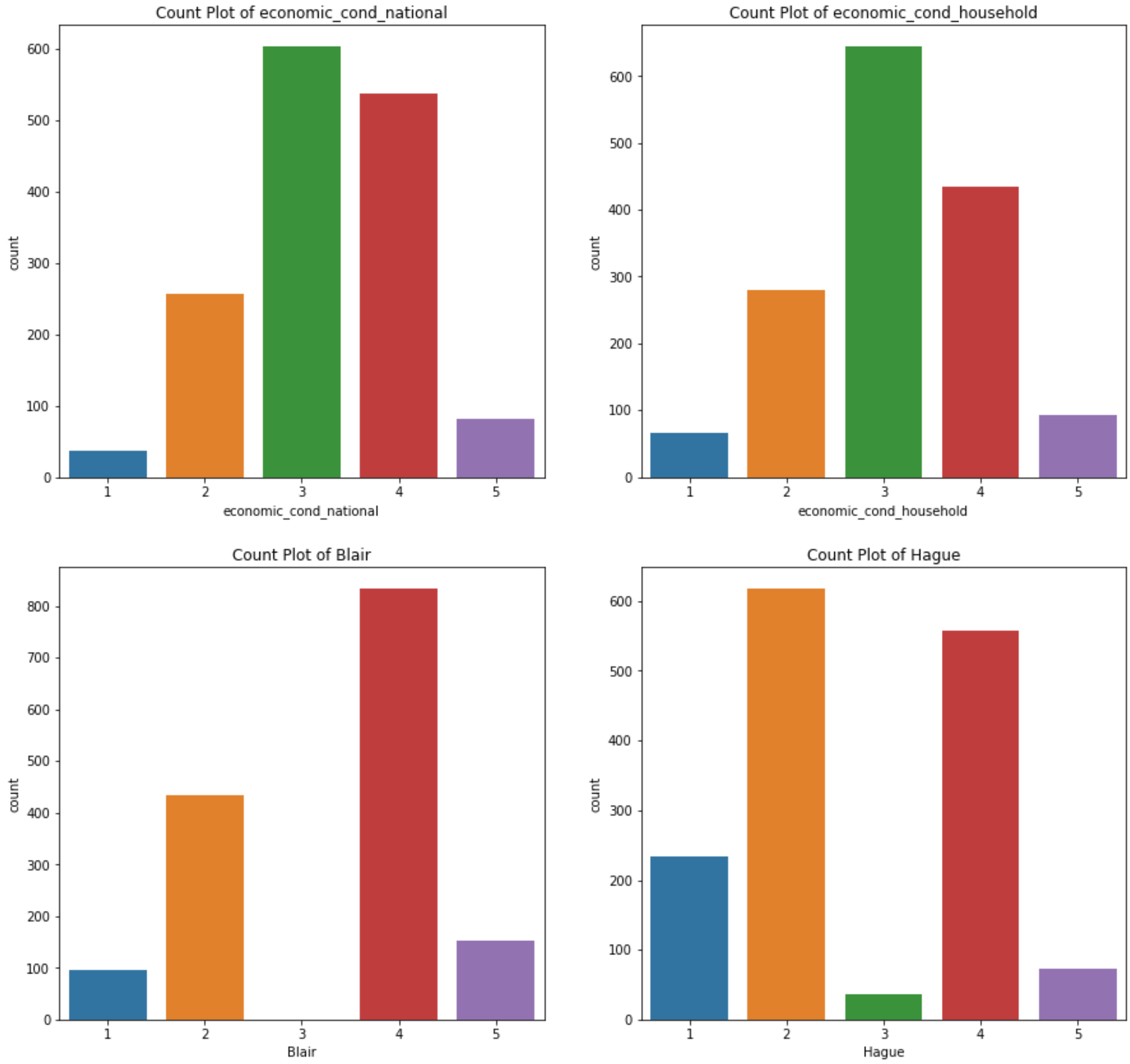


Figure 3. Histogram, Box Plot, Swarmplot and Bar Plot for Continuous Numerical Features with Vote as Hue.

Inferences:

1. Age feature is normally distributed with slightly right skewed.
2. Age has similar distributions in both the classes of target feature.
3. Age do not have outliers in both the classes of target feature.
4. Median age of voters voting for Conservative party is more than that of Labour party.
5. Mean age of voters voting for Conservative party is more than that of Labour party.

Count Plots of Discrete Numerical and Categorical Features



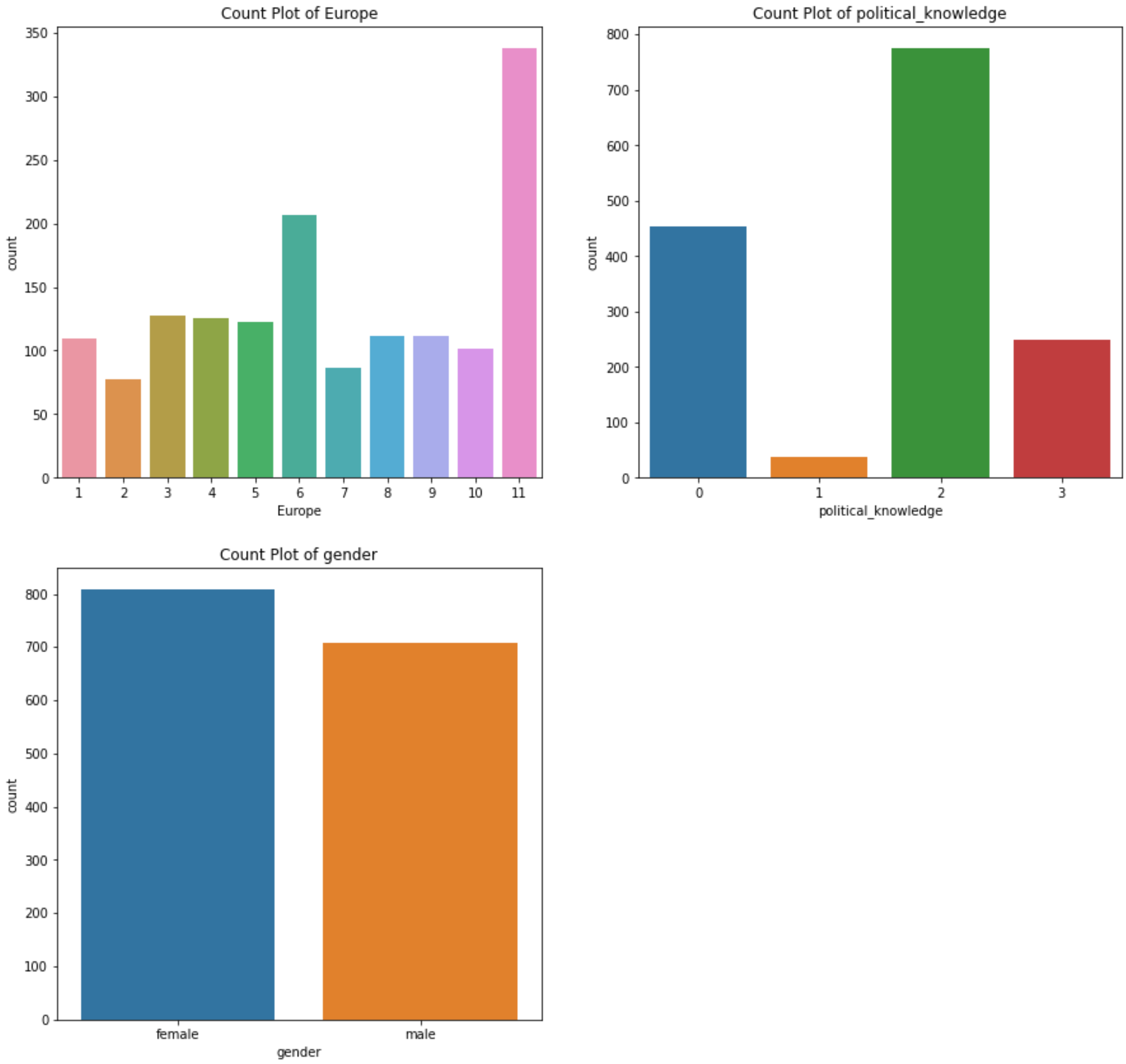


Figure 4. Count Plots of Discrete Numerical and Categorical Features.

Insights

From above count plots, we can conclude below points.

1. There are a greater number of voters who assessed economic_condtion_national as 3. There are a smaller number of voters who assessed economic_condition_national as 1. Decreasing order of voters as below.

   3 > **4** > **2** > **5** > **1**

2. There are a greater number of voters who assessed economic_condtion_household as 3. There are a smaller number of voters who assessed economic_condtion_household as 1. Decreasing order of voters as below.

   3 > **4** > **2** > **5** > **1**

3. There are a greater number of voters who assessed Blair as 4. There are no voters who assessed Blair as 3. Decreasing order of voters as below.

4 > **2** > **5** > **1** > **3**

4. There are a greater number of voters who assessed Hague as 2. There are a smaller number of voters who assessed Blair as 3. Decreasing order of voters as below.

2 > **4** > **1** > **5** > **3**

5. There are a greater number of voters who have Eurosceptic sentiment as 11. There are a smaller number of voters who have Eurosceptic sentiment as 2. Decreasing order of voters as below.

11> **6** > **3** > **4** > **5** > **9** > **8** > **1** > **10** > **7** > **2**

6. There are a greater number of voters who have political knowledge as 2. There are a smaller number of voters who have political knowledge as 1. Decreasing order of voters as below.

2 > **0** > **3** > **1**

7. There are a greater number of female voters compared to male voters.

Distribution of Classes in Target Column

The percentage of voters voted for Labour party is 69.68%

The percentage of voters voted for Conservative party is 30.32%

The data is well balanced with the classes in target feature. We can proceed with model building process.

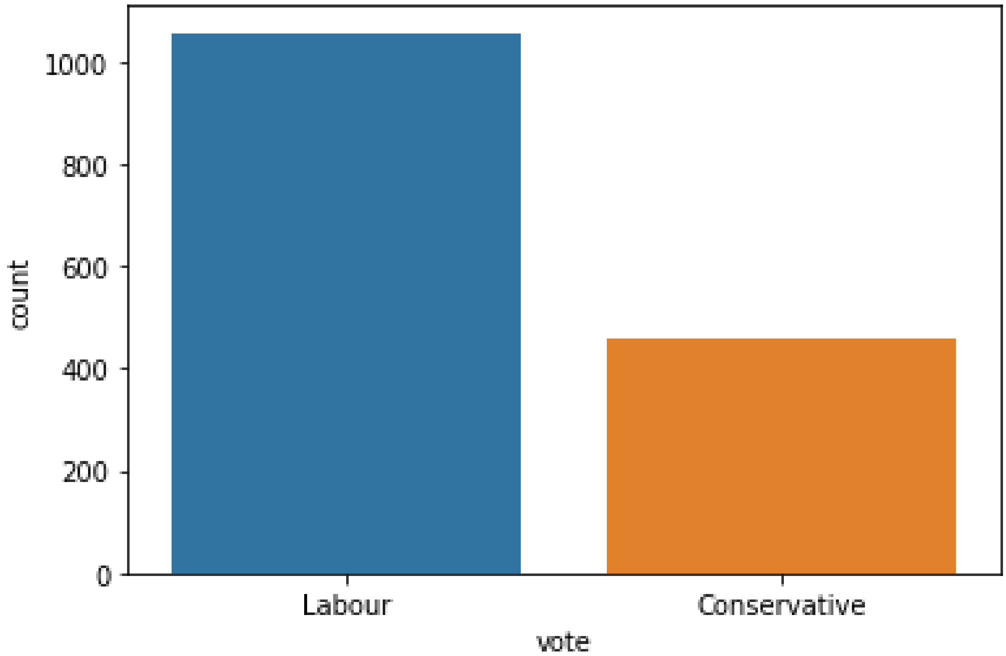


Figure 5. Count Plot of Target Feature (Vote).

BIVARIATE ANALYSIS

Pair Plot of Numerical Features with Vote as Hue

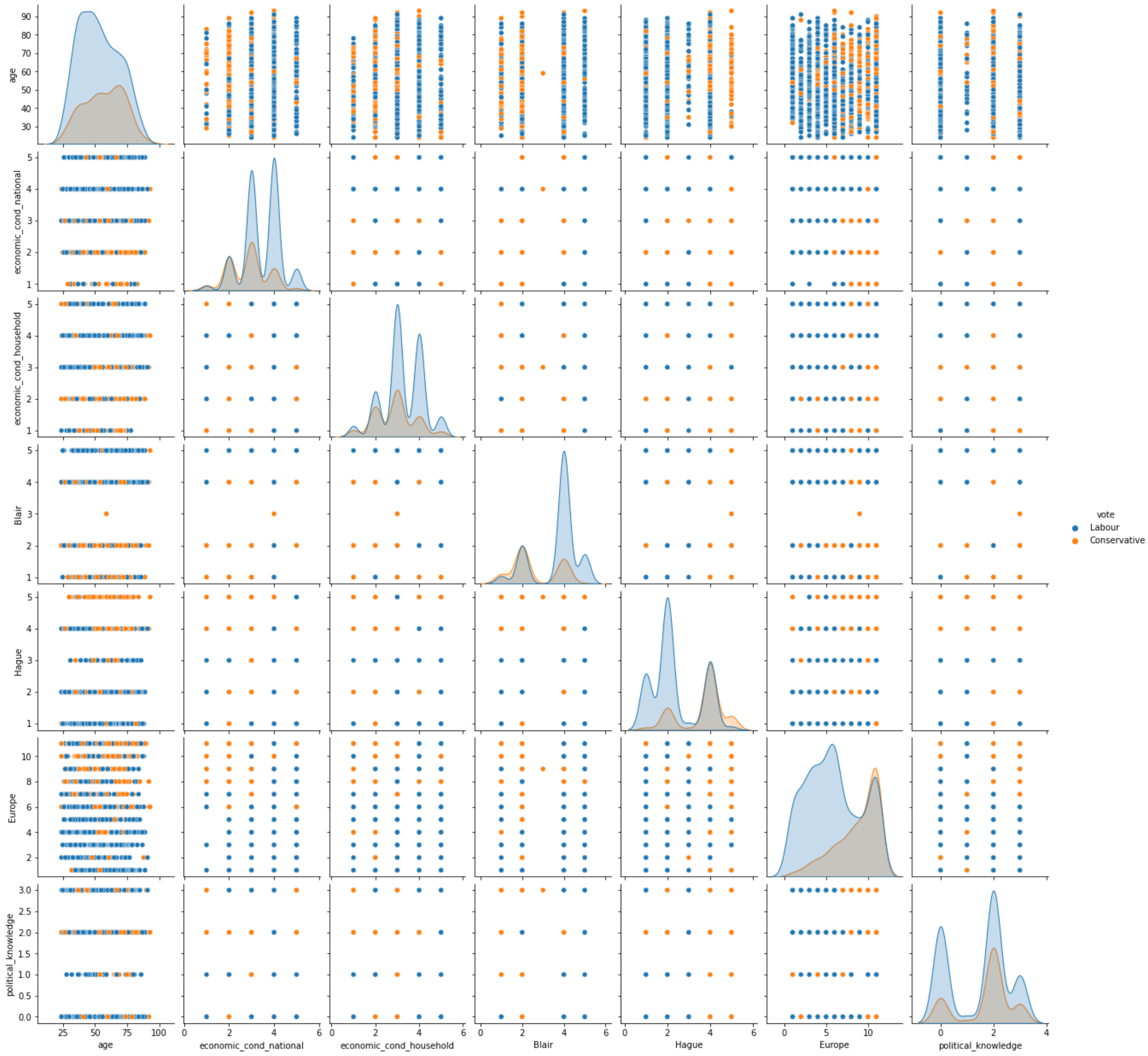


Figure 6. Pair Plot for Numeric Features in Election Dataset.

Heat Map

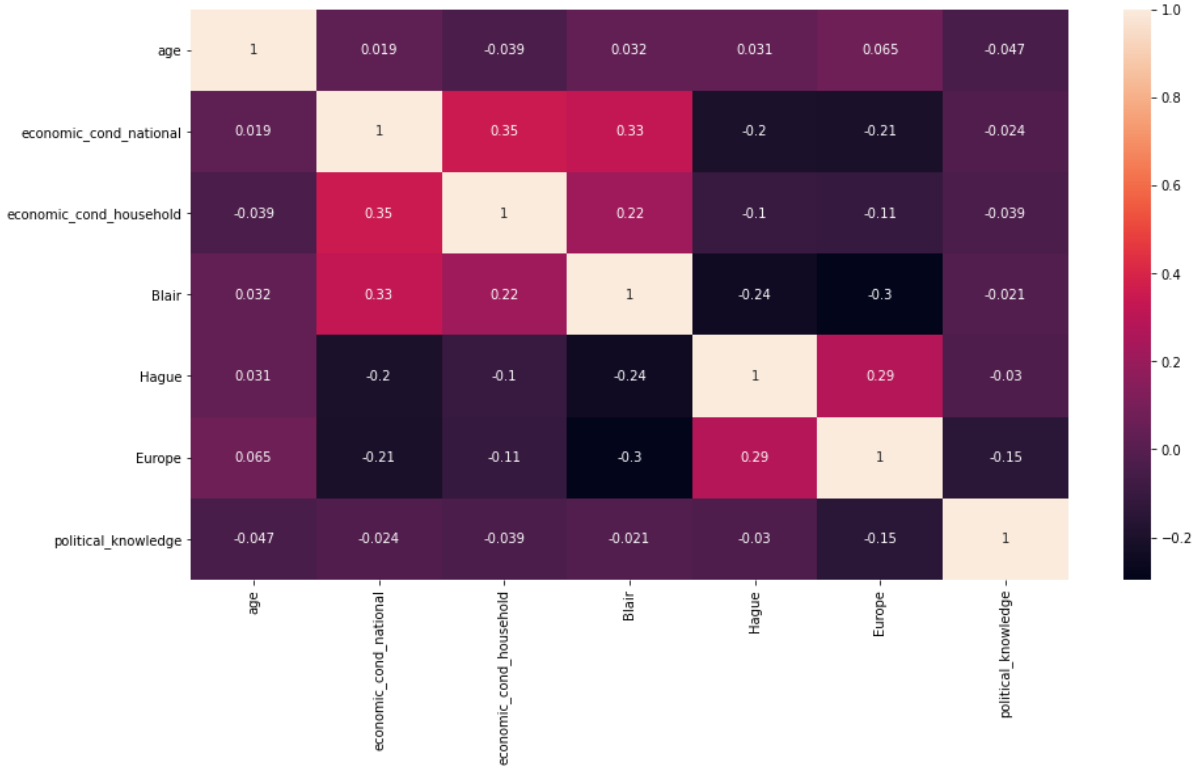


Figure 7. Heatmap for Numeric Features in Election Dataset.

Note:

From above Pair-Plot and Heatmap, it can be noticed that there is **no significant correlation** between the predictor variables.

Box Plot of Age Vs Vote

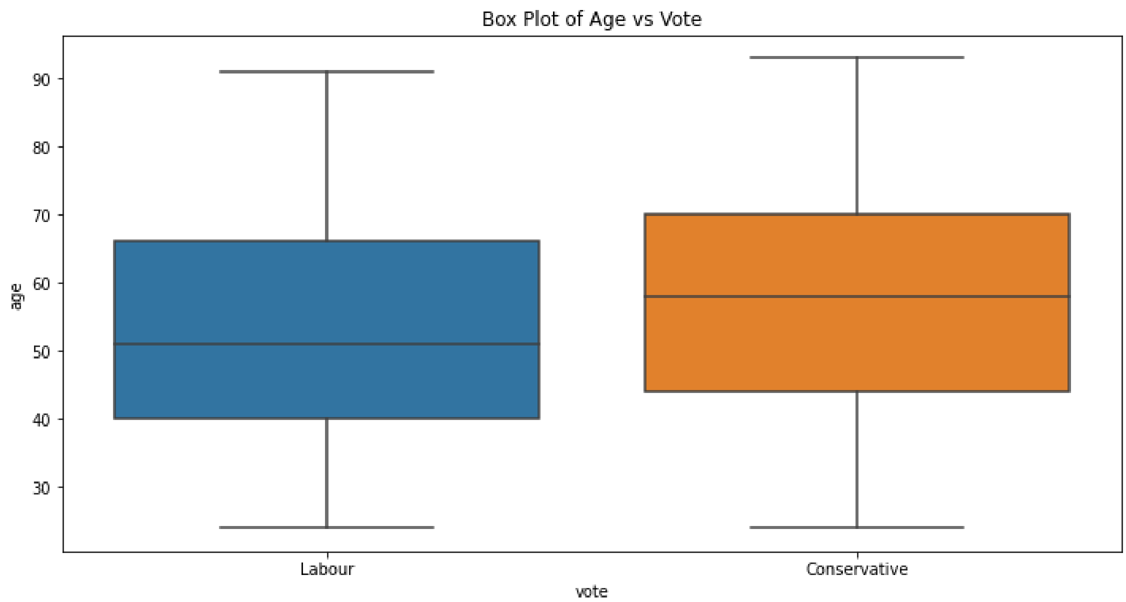


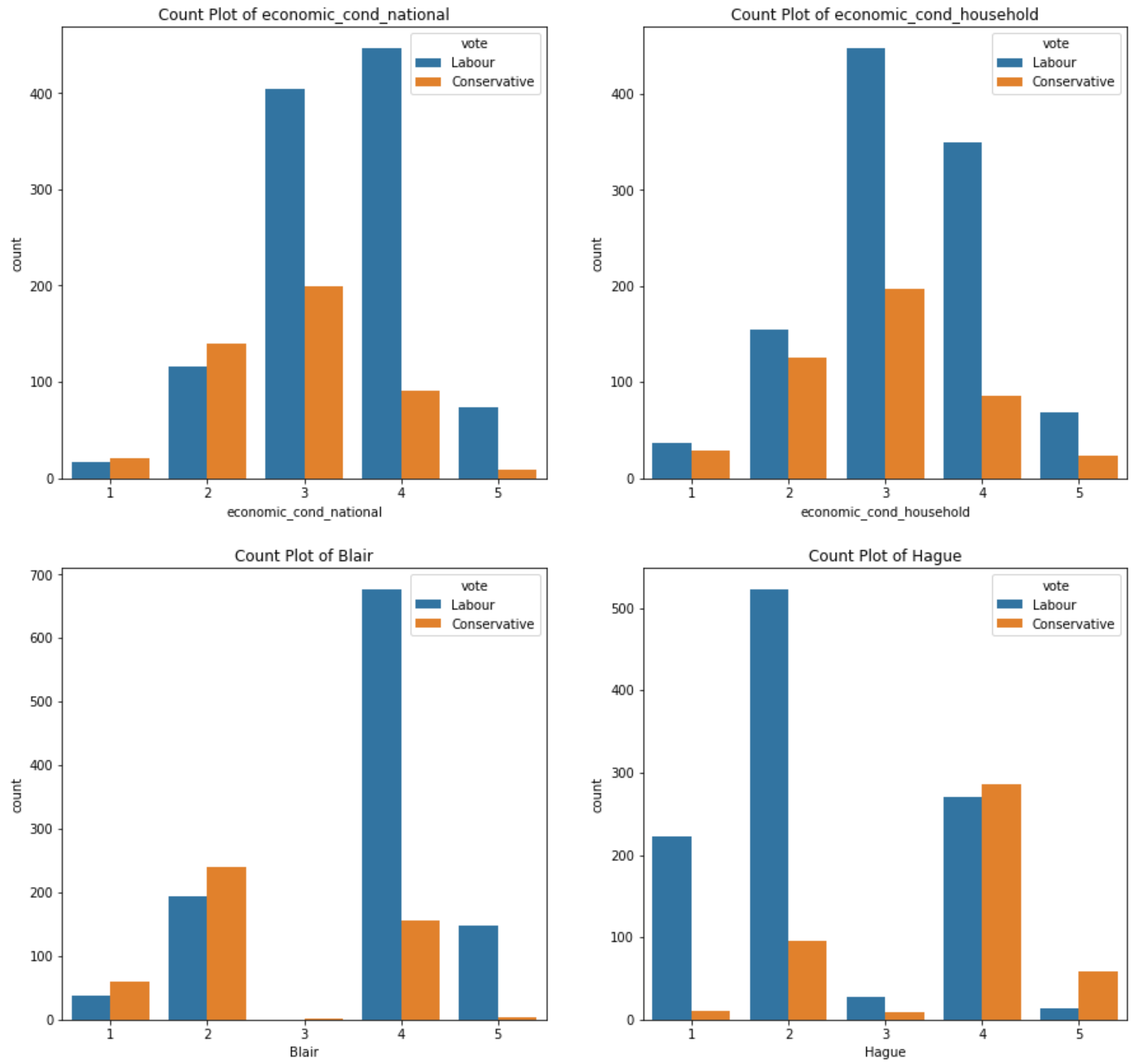
Figure 8. Box Plot of Age Vs Vote.

Insights:

From above bar plots, we can write below inferences,

1. Median age of voters those who have voted for Conservative party is slightly more than that of voters voted for Labour party.
2. There are no outliers in age feature of both classes.

Count Plots of Discrete Numerical and Categorical Features with Holliday Package as Hue



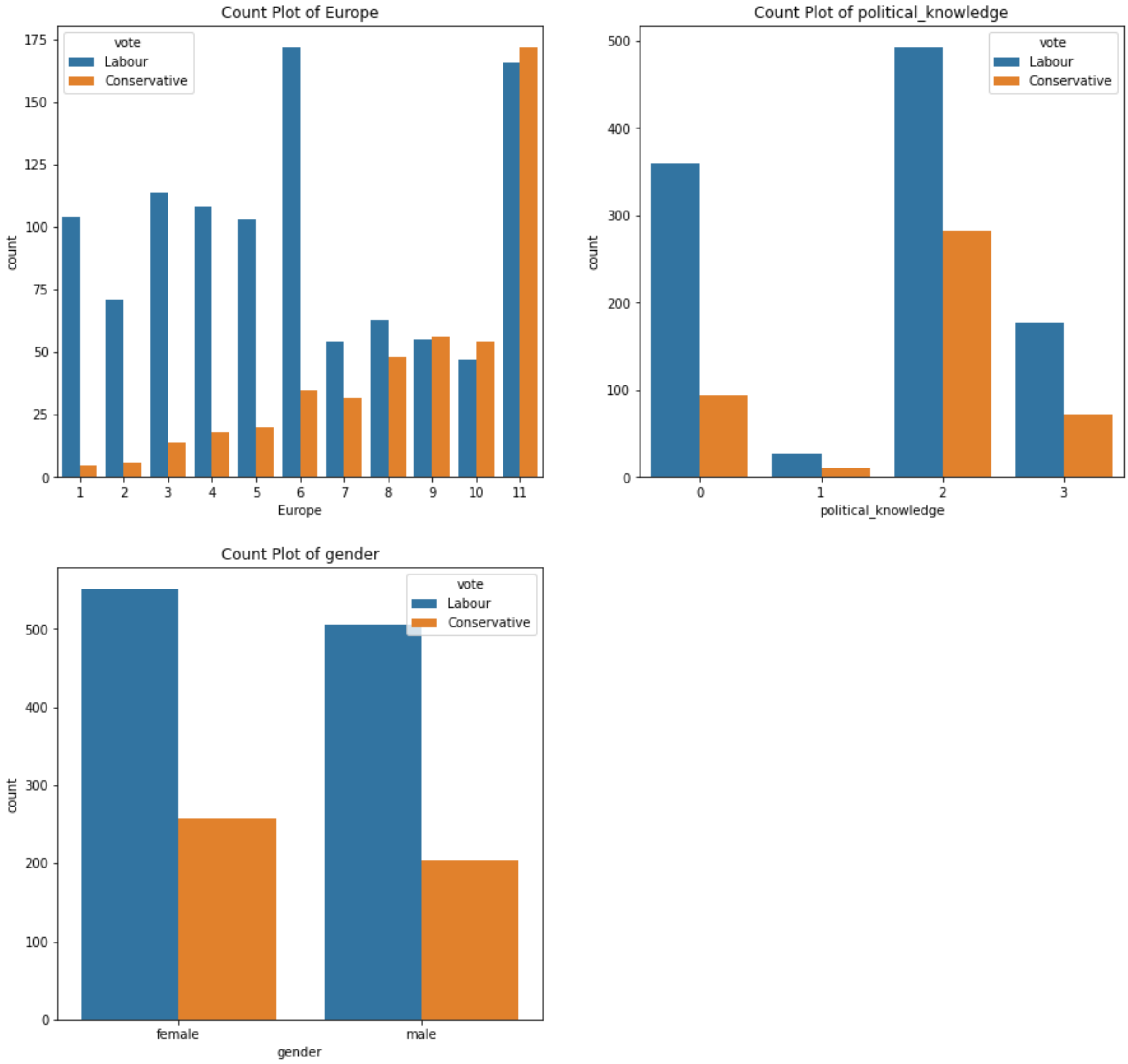


Figure 9. Count Plots of Discrete Numerical and Categorical Features with Vote as Hue

Insights:

From above count plots, we can conclude below points.

1. The voters who assessed **high ratings (3, 4 & 5) to economic_condtion_national** are preferring to **vote Labour party** whereas voters who assessed **low ratings (1 & 2) to economic_condtion_national** are preferring to **vote Conservative party**.
2. The voters who assessed **high ratings (3, 4 & 5) to economic_condtion_household** are preferring to **vote Labour party** whereas voters who **assessed low ratings (1 & 2) to economic_condtion_household** are preferring to **vote Conservative party**.
3. The voters who assessed **high ratings (4 & 5) to Blair** (Labour party leader) are preferring to **vote Labour party** whereas voters who assessed **low ratings (1 & 2) to Blair** (Labour party leader) are preferring to **vote Conservative party**.

4. The voters who assessed **high ratings (4 & 5) to Hague** (Conservative party leader) are preferring to **vote Conservative party** whereas voters who assessed **low ratings (1 & 2) to Hague** (Conservative party leader) are preferring to **vote Labour party**.
5. The voters who have **high Eurosceptic sentiment (9,10 & 11)** are preferring to **vote Conservative party** whereas voters who **have low Eurosceptic sentiment (less than 9)** are preferring to **vote Labour party.**
6. All the voters with **different levels of political knowledge** are preferring to **vote Labour party** than conservative party.
7. Both **male and female voters** are preferring **to vote Labour party** than conservative party.

Checking Outliers in Continuous Numerical Features (Age):

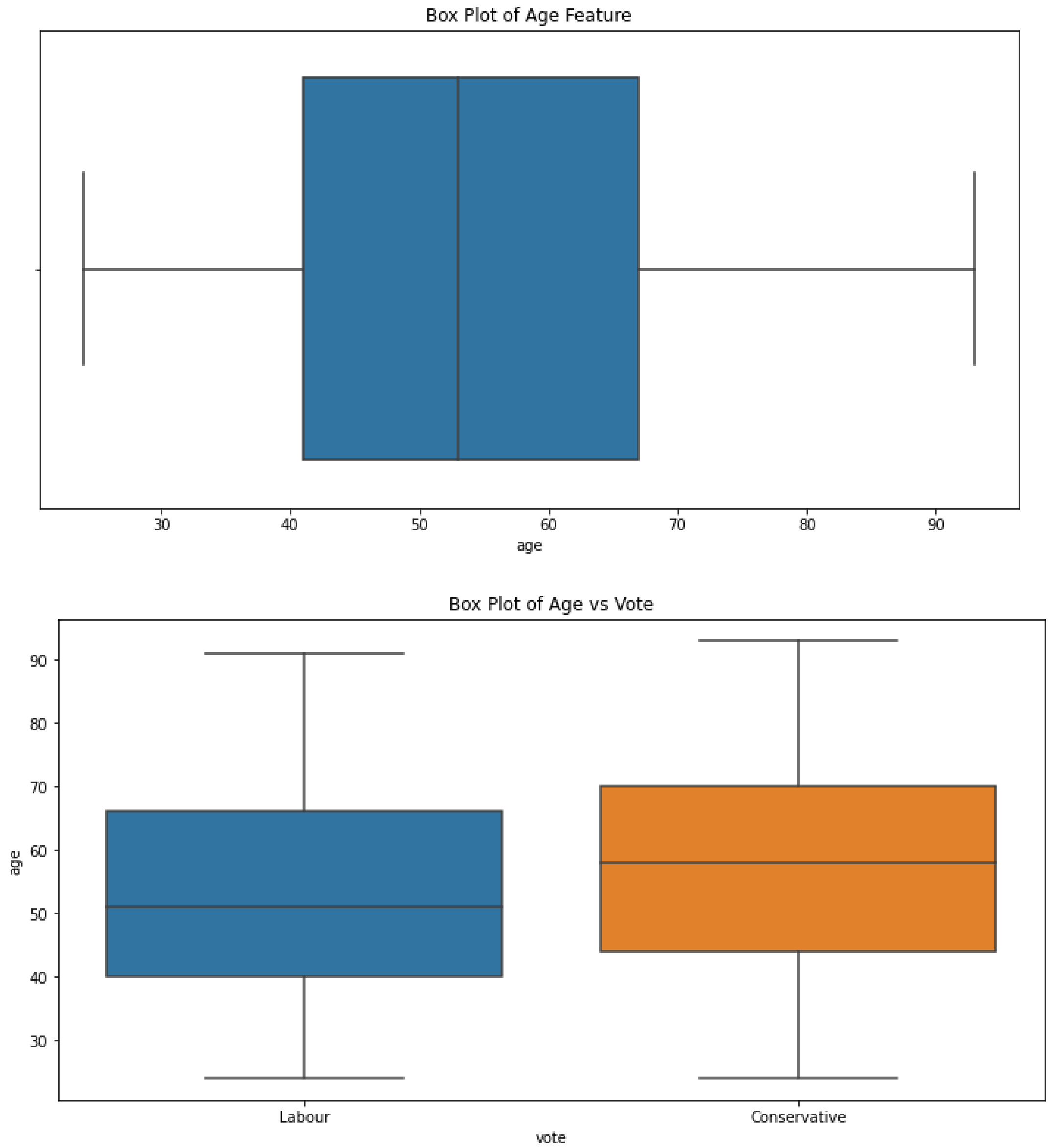


Figure 10. Box Plots of Continuous Numerical Features (Age) in the Election Dataset.

Insights:

1. There are no outliers in the age of voters.
2. There are no outliers in the age feature even in both classes of voters.

Q1.3. 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.

Necessity of Scaling

1. Generally, Scaling improves the performance of **all distance-based models like Linear Discriminant Analysis and KNN**. Even Scaling influences the coefficients obtained for different features in logistic regression model. By scaling, units can be avoided in coefficients and standardized coefficients are obtained. Also scaling improves the speed of convergence of the models.
2. If we don't scale the data, it gives higher weightage to features which have higher magnitude. Hence, it is always advisable to **bring all the features to the same scale** before proceeding to model building.
3. In this dataset, the magnitudes of the statistical parameters like Mean, Standard Deviation, Variance, Minimum and Maximum are significantly different for all features (Refer below table). **Hence, scaling is required to bring all the features into a common scale before proceeding to model building.**
4. Z-Score method is used to scale the data i.e., finding z-score value for each and every observation in the dataset by using following formula.

$$Z\ Score = \frac{(\mathrm{x} - \mu)}{Sigma}$$

Where, x = Value of the observation

μ = Mean

Sigma = Standard Deviation

5. **Scaling is required for Logistic Regression, Linear Discriminant Analysis and KNN models.** Scaled dataset is used for these models.
6. For other models like **Naive Bayes, Bagging, Random Forest, Ada Boosting and Gradient Boosting** models, **scaling is not required**. Hence, non-scaled dataset is used for these models.

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge
count	1517.0	1517.0	1517.0	1517.0	1517.0	1517.0	1517.0
mean	54.2	3.2	3.1	3.3	2.7	6.7	1.5
std	15.7	0.9	0.9	1.2	1.2	3.3	1.1
variance	246.5	0.8	0.9	1.4	1.5	10.9	1.2
min	24.0	1.0	1.0	1.0	1.0	1.0	0.0
25%	41.0	3.0	3.0	2.0	2.0	4.0	0.0
50%	53.0	3.0	3.0	4.0	2.0	6.0	2.0
75%	67.0	4.0	4.0	4.0	4.0	10.0	2.0
max	93.0	5.0	5.0	5.0	5.0	11.0	3.0

Table 8. Mean, Standard Deviation and Variance of All Numeric Features.

Encoding the Data

1. As Target variable vote is categorical, minority class – **Conservative party** is assigned as **label 1** and majority class – **Labour party** is assigned as **label 0**.
2. As gender feature is categorical, dummies are created for this feature by dropping the first dummy variable.

Sample of the Encoded Dataset

	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
0	0	43	3	3	4	1	2	2	0
1	0	36	4	4	4	4	5	2	1
2	0	35	4	4	5	2	3	2	1
3	0	24	4	2	2	1	4	0	0
4	0	41	2	2	1	1	6	2	1

Table 9. Sample of the Encoded Election Dataset.

Splitting the Data into Train and Test Sets

1. Both independent and target features have been divided into train and test sets.
2. No. of observations in test set is selected as 0.3 times of total data points.
3. Then no. of observations in train set will be 0.7 times of total data points.

Checking the Training and Test Data

```
size of xtrain:  (1061, 8)
size of xtest:  (456, 8)
size of ytrain:  (1061,)
size of ytest:  (456,)
```

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
991	34	2	4	1	4	11	2	0
1274	40	4	3	4	4	6	0	1
649	61	4	3	4	4	7	2	0
677	47	3	3	4	2	11	0	1
538	44	5	3	4	2	8	0	1

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
504	71	3	3	2	2	8	2	0
369	43	3	2	4	2	8	3	1
1075	89	5	5	5	2	1	2	1
1031	47	2	3	2	4	8	2	0
1329	33	5	4	4	4	8	0	1

Table 10. Samples of Predictors Train and Predictors Test Datasets.

```
991       1
1274      0
649       1
677       0
538       0
Name: vote, dtype: int32
```

```
504       0
369       0
1075      0
1031      1
1329      0
Name: vote, dtype: int32
```

Table 11. Samples of Target Train and Target Test Data.

Distribution of Target Class in Train and Test sets

```
0     71.065033
1     28.934967
Name: vote, dtype: float64
```

```
0     66.447368
1     33.552632
Name: vote, dtype: float64
```

Table 12. Distribution of Target Class in Train and Test sets.

From above table, we can notice that target class (0s and 1s) is almost uniformly distributed between train and test datasets.

Sample of Scaled Datasets

Predictor variables have been **scaled by using z-score method**. Initially train dataset has been scaled by using its mean and standard deviation. Then test dataset has been **scaled by using train dataset parameters** (mean and standard deviation) **to avoid data leakage**.

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
991	-1.296710	-1.455581	0.902100	-2.018037	1.029070	1.332089	0.452231	0
1274	-0.910337	0.877307	-0.163744	0.550300	1.029070	-0.202156	-1.407526	1
649	0.441968	0.877307	-0.163744	0.550300	1.029070	0.104693	0.452231	0
677	-0.459569	-0.289137	-0.163744	0.550300	-0.593283	1.332089	-1.407526	1
538	-0.652755	2.043751	-0.163744	0.550300	-0.593283	0.411542	-1.407526	1

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
504	1.085923	-0.289137	-0.163744	-1.161925	-0.593283	0.411542	0.452231	0
369	-0.717151	-0.289137	-1.229589	0.550300	-0.593283	0.411542	1.382110	1
1075	2.245042	2.043751	1.967945	1.406413	-0.593283	-1.736401	0.452231	1
1031	-0.459569	-1.455581	-0.163744	-1.161925	1.029070	0.411542	0.452231	0
1329	-1.361106	2.043751	0.902100	0.550300	1.029070	0.411542	-1.407526	1

Table 13. Samples of Predictors Train and Predictors Test Datasets after Scaling.

Q1.4. Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both models (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)

LOGISTIC REGRESSION MODEL

Initially Logistic Regression model has been built **with default hyperparameters** as shown below.

penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=1, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	686	68
Actual 1	111	196

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	268	35
Actual 1	41	112

Table 14. Confusion Matrix for Train and Test Datasets in Logistic Regression Model with Default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.86	0.91	0.88	754
1	0.74	0.64	0.69	307
accuracy			0.83	1061
macro avg	0.80	0.77	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.87	0.88	0.88	303
1	0.76	0.73	0.75	153
accuracy			0.83	456
macro avg	0.81	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456

Table 15. Classification Reports for Train and Test Datasets in Logistic Regression Model with Default Hyperparameters.

ROC Curves for Train and Test Datasets

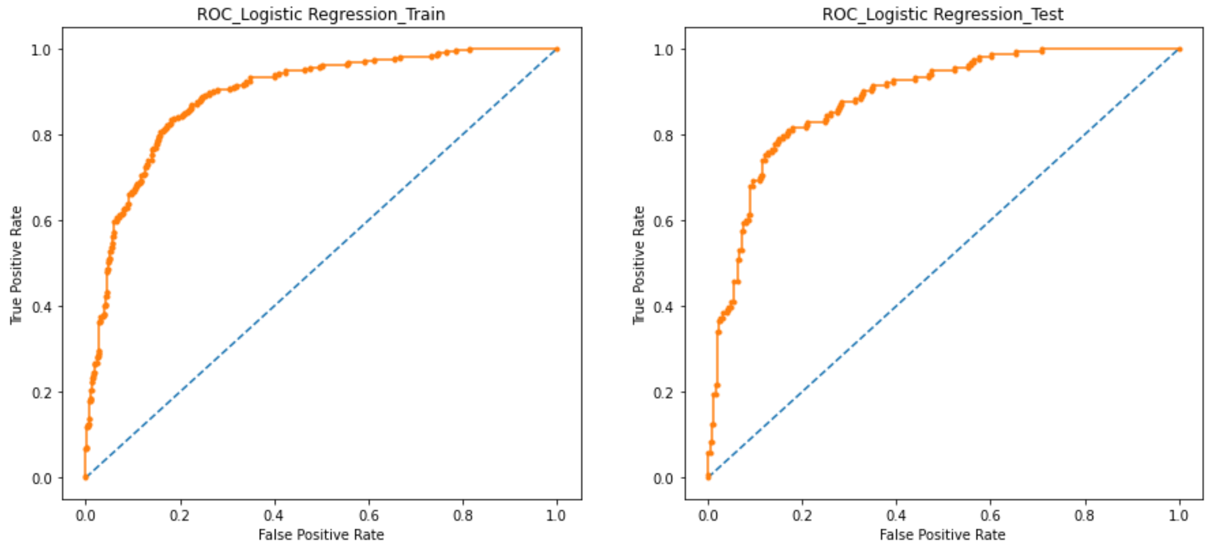


Figure 11. ROC Curves for Train and Test Datasets in Logistic Regression Model with Default Hyperparameters.

Feature Importance (Coefficients)

Predictor	Hague	Blair	Europe	economic_cond_national	political_knowledge	age	gender_male	economic_cond_household
Coefficients	1.012	-0.7	0.684	-0.54	0.345	0.231	-0.192	-0.059

Table 16. Coefficients of Features in Logistic Regression Model with Default Parameters.

Inferences

1. Accuracy for test dataset (0.83) is equal to accuracy for train dataset (0.83). Hence, there is **no overfitting in logistic regression model** and the model is valid.
2. ROC_AUC score for train and test datasets are 0.89 and is 0.88 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.87, 0.88 and 0.88 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.76, 0.73 and 0.75 respectively. Recall and F1 scores are less than 0.75. These scores may be improved by tuning hyperparameters.
5. **Hague, Blair and Europe** are three most important features for predicting the target variable.
6. The **decreasing order of features according their importance** as given below
   Hague > Blair > Europe > economic_cond_national > political_knowledge > age > gender_male > economic_cond_household

LINEAR DISCRIMINANT ANALYSIS (LDA)

Initially Linear Discriminant Analysis model has been built **with default hyperparameters** as shown below.

solver='svd', shrinkage=None, priors=None, n_components=None, store_covariance=False, tol=0.0001.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	685	69
Actual 1	107	200

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	269	34
Actual 1	42	111

Table 17. Confusion Matrix for Train and Test Datasets in LDA Model with Default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.86	0.91	0.89	754
1	0.74	0.65	0.69	307
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 Test Dataset

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

Table 18. Classification Reports for Train and Test Datasets in LDA Model with Default Hyperparameters.

ROC Curves for Train and Test Datasets

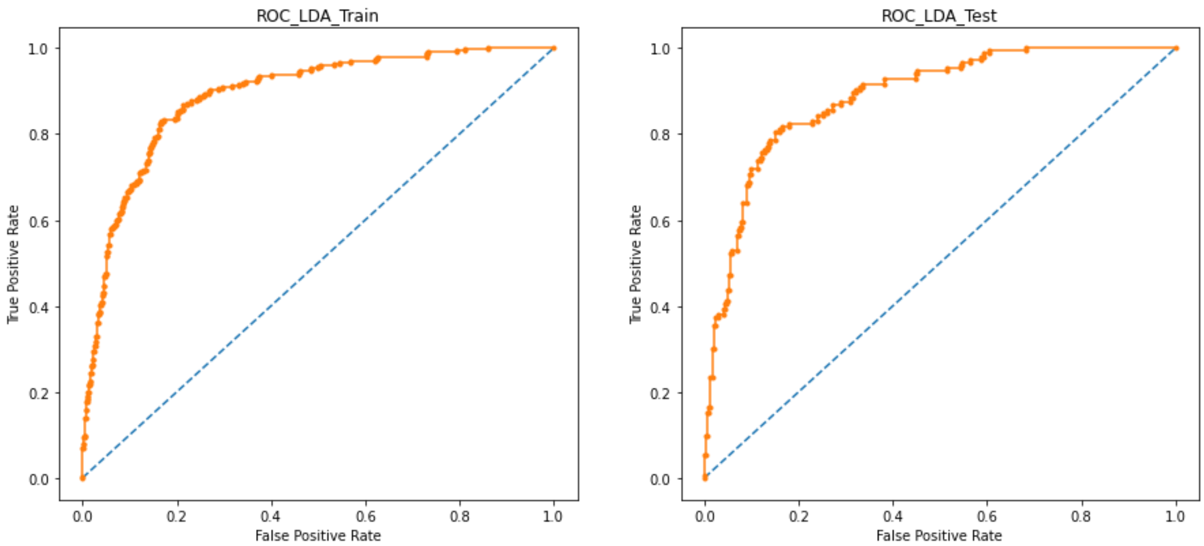


Figure 12. ROC Curves for Train and Test Datasets in LDA Model with Default Hyperparameters.

Feature Importance (Coefficients)

Predictor	Hague	Blair	Europe	economic_cond_national	political_knowledge	age	gender_male	economic_cond_household
Coefficients	1.142	-0.867	0.729	-0.519	0.463	0.311	-0.149	-0.047

Table 19. Coefficients of Features in LDA Model with Default Parameters.

Inferences

1. Accuracy for test dataset (0.83) is equal to accuracy for train dataset (0.83). Hence, there is **no overfitting in Linear Discriminant Analysis model** and the model is valid.
2. ROC_AUC score for train and test datasets are 0.89 and is 0.89 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.86, 0.89 and 0.88 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.77, 0.73 and 0.74 respectively. Recall and F1 scores are less than 0.75. These scores may be improved by tuning hyperparameters.
5. **Hague and Blair** are two most important features for predicting the target variable
6. The **decreasing order of features according their importance** as given below Hague > Blair > Europe > economic_cond_national > political_knowledge > age > gender_male > economic_cond_household

Q1.5. 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)

NAÏVE BAYES MODEL

Initially Naïve Bayes model has been built **with default hyperparameters** as shown below.

priors=None, var_smoothing=1e-09

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	675	79
Actual 1	96	211

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	263	40
Actual 1	41	112

Table 20. Confusion Matrix for Train and Test Datasets in Naïve Bayes Model with Default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.88	0.90	0.89	754
1	0.73	0.69	0.71	307
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 Test Dataset

	precision	recall	f1-score	support
0	0.87	0.87	0.87	303
1	0.74	0.73	0.73	153
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

Table 21. Classification Reports for Train and Test Datasets in Naïve Bayes Model with Default Hyperparameters.

ROC Curves for Train and Test Datasets

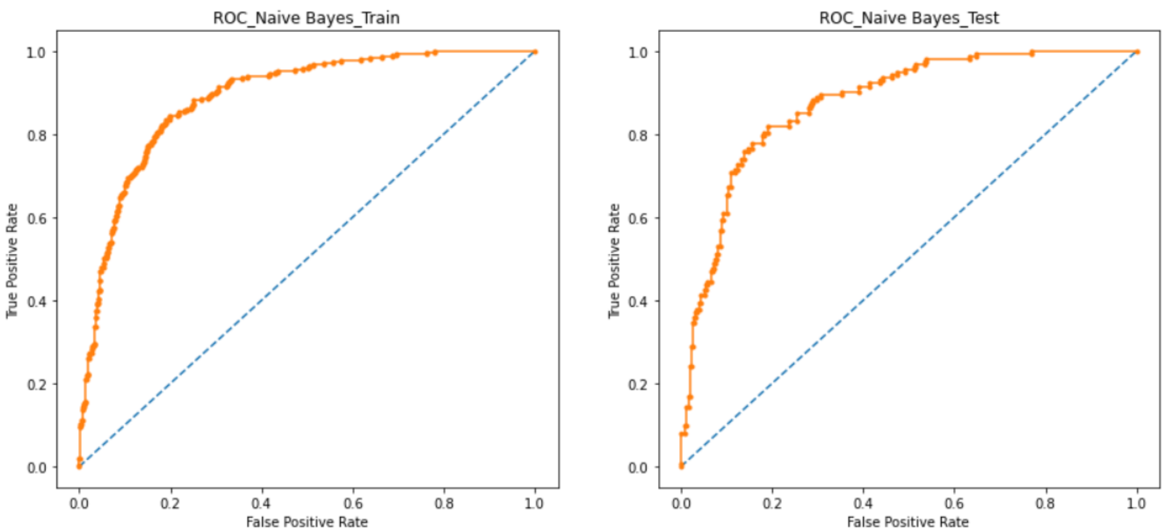


Figure 13. ROC Curves for Train and Test Datasets in Naïve Bayes Model with Default Hyperparameters.

Inferences

1. Accuracy for test dataset (0.82) is slightly less than that of train dataset (0.84). Hence, there is **no overfitting in Naïve Bayes model** and the model is valid.

2. ROC_AUC score for train and test datasets are 0.89 and is 0.88 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.87, 0.87 and 0.87 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.74, 0.73 and 0.73 respectively. Precision, recall and F1 scores are less than 0.75. These scores may be improved by tuning hyperparameters.

KNN MODEL

Initially k-nearest neighbors model has been built **with default hyperparameters** as shown below.

n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	689	65
Actual 1	91	216

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	269	34
Actual 1	45	108

Table 22. Confusion Matrix for Train and Test Datasets in KNN Model with Default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.88	0.91	0.90	754
1	0.77	0.70	0.73	307
accuracy			0.85	1061
macro avg	0.83	0.81	0.82	1061
weighted avg	0.85	0.85	0.85	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.86	0.89	0.87	303
1	0.76	0.71	0.73	153
accuracy			0.83	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.82	0.83	0.83	456

Table 23. Classification Reports for Train and Test Datasets in KNN Model with Default Hyperparameters.

ROC Curves for Train and Test Datasets

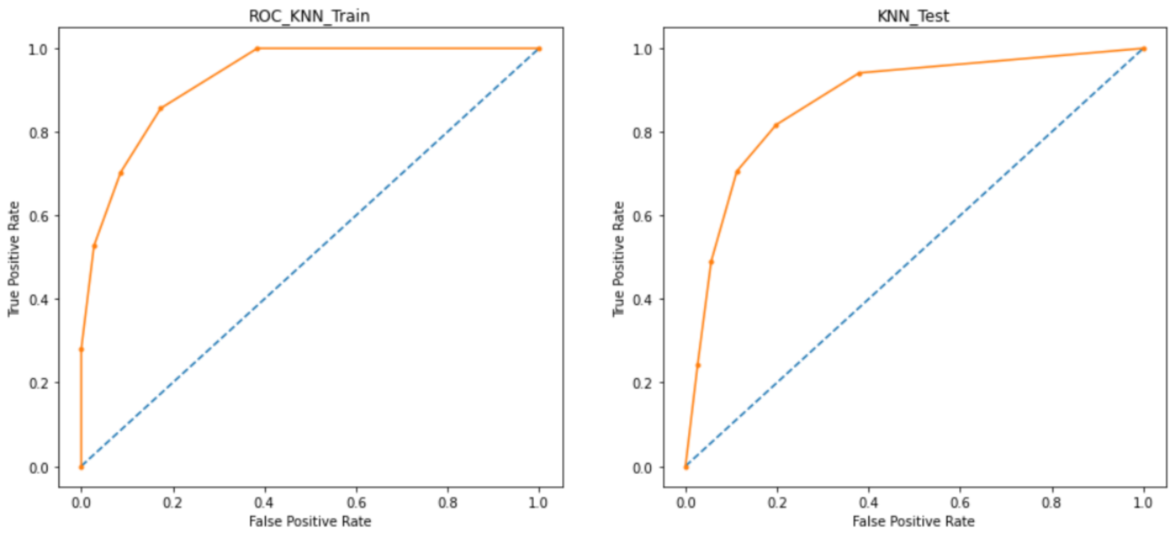


Figure 14. ROC Curves for Train and Test Datasets in KNN Model with Default Hyperparameters.

Inferences

1. Accuracy for test dataset (0.83) is slightly less than that of train dataset (0.85). Hence, there is **no overfitting in KNN model** and the model is valid.
2. ROC_AUC score for train and test datasets are 0.93 and is 0.87 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.86, 0.89 and 0.87 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.76, 0.71 and 0.73 respectively. These scores are less than 0.75. We can try to improve these scores by tuning hyperparameters.

Q1.6. 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 parameters. 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.

BAGGING

Initially Bagging Classifier model has been built **with default hyperparameters** as shown below.

base_estimator=None, n_estimators=10, max_samples=1.0, max_features=1.0, bootstrap=True, bootstrap_features=False, oob_score=False, warm_start=False, n_jobs=None, random_state=1, verbose=0

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	752	2
Actual 1	17	290

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	270	33
Actual 1	58	95

Table 24. Confusion Matrix for Train and Test Datasets in Bagging Classifier Model with Default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.98	1.00	0.99	754
1	0.99	0.94	0.97	307
accuracy			0.98	1061
macro avg	0.99	0.97	0.98	1061
weighted avg	0.98	0.98	0.98	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.82	0.89	0.86	303
1	0.74	0.62	0.68	153
accuracy			0.80	456
macro avg	0.78	0.76	0.77	456
weighted avg	0.80	0.80	0.80	456

Table 25. Classification Reports for Train and Test Datasets in Bagging Classifier Model with Default Hyperparameters.

ROC Curves for Train and Test Datasets

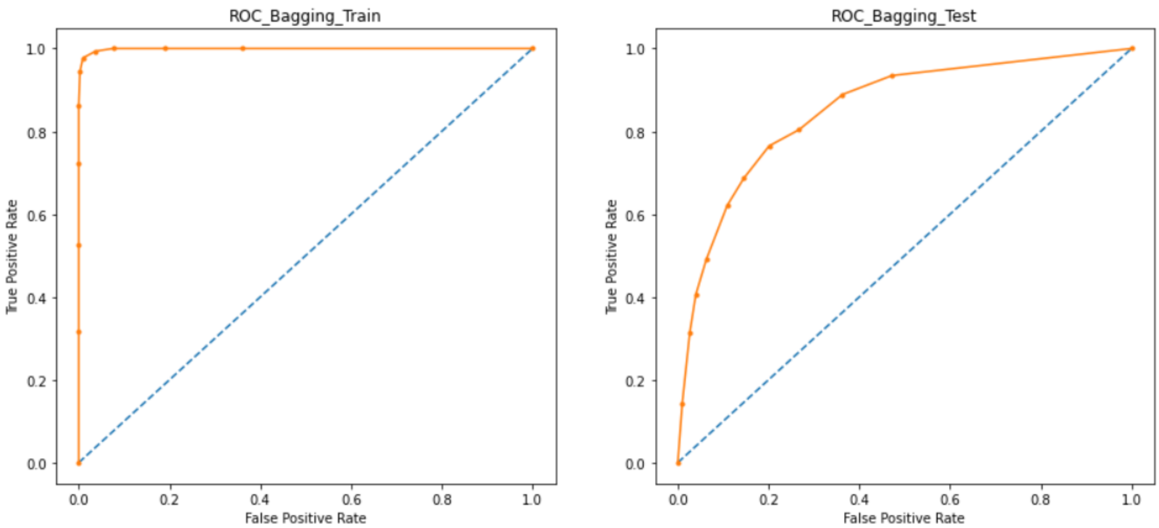


Figure 15. ROC Curves for Train and Test Datasets in Bagging Classifier Model with Default Hyperparameters.

Inferences

1. Accuracy for test dataset (0.80) is much lesser than accuracy for train dataset (0.98). Hence, there is **overfitting in Bagging Classifier model** and the **model is not valid at present**. Model should be verified to avoid overfitting by tuning hyperparameters.
2. ROC_AUC score for train and test datasets are 1 and is 0.85 respectively. These scores are pretty good but **the model is overfitted**.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.82, 0.89 and 0.86 respectively. These scores are good enough to use the model for predictions but before proceeding for predictions, overfitting should be removed.

4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.74, 0.62 and 0.68 respectively. These scores are less than 0.75. We can try to improve these scores by tuning hyperparameters.

RANDOM FOREST MODEL

Initially Random Forest model has been built **with default hyperparameters** as shown below. n_estimators=100, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=1, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	754	0
Actual 1	0	307

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	276	27
Actual 1	51	102

Table 26. Confusion Matrix for Train and Test Datasets in Random Forest Model with Default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	1.00	1.00	1.00	754
1	1.00	1.00	1.00	307
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 Test Dataset

	precision	recall	f1-score	support
0	0.84	0.91	0.88	303
1	0.79	0.67	0.72	153
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.83	0.83	0.82	456

Table 27. Classification Reports for Train and Test Datasets in Random Forest Model with Default Hyperparameters.

ROC Curves for Train and Test Datasets

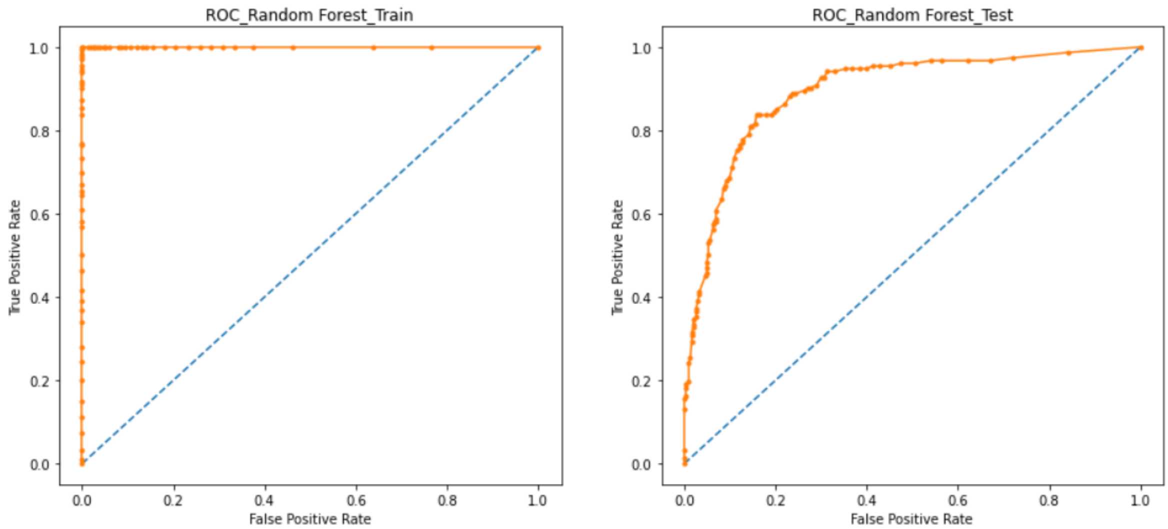


Figure 16. ROC Curves for Train and Test Datasets in Random Forest Model with Default Hyperparameters.

Feature Importance

	age	Europe	Hague	Blair	economic_cond_national	economic_cond_household	political_knowledge	gender_male
Feature_Importances	0.213	0.188	0.179	0.133	0.092	0.081	0.078	0.036

Table 28. Features Importance in Random Forest Model with Default Parameters.

Inferences

1. Accuracy for test dataset (0.88) is much lesser than accuracy for train dataset (1). Hence, there is **overfitting in Random Forest model** and the **model is not valid at present**. Model should be verified to avoid overfitting by tuning hyperparameters.
2. ROC_AUC score for train and test datasets are 1 and is 0.9 respectively. These scores are pretty good but **the model is overfitted**.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.84, 0.91 and 0.88 respectively. These scores are good enough to use the model for predictions but before proceeding for predictions, overfitting should be removed.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.79, 0.67 and 0.72 respectively. Recall and F1 score are less than 0.75. We can try to improve these scores by tuning hyperparameters.
5. **Age, Europe and Hague** are three most important features for predicting the target variable.
6. The **decreasing order of features according their importance** as given below

Age > Europe > Hague > Blair > economic_cond_national > economic_cond_household > political_knowledge > gender_male

AdaBoosting Model

Initially AdaBoosting model has been built **with default hyperparameters** as shown below.

base_estimator=None, n_estimators=50, learning_rate=1.0,

algorithm='SAMME.R', random_state=1

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	688	66
Actual 1	97	210

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	266	37
Actual 1	48	105

Table 29. Confusion Matrix for Train and Test Datasets in AdaBoosting Model with Default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.88	0.91	0.89	754
1	0.76	0.68	0.72	307
accuracy			0.85	1061
macro avg	0.82	0.80	0.81	1061
weighted avg	0.84	0.85	0.84	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.85	0.88	0.86	303
1	0.74	0.69	0.71	153
accuracy			0.81	456
macro avg	0.79	0.78	0.79	456
weighted avg	0.81	0.81	0.81	456

Table 30. Classification Reports for Train and Test Datasets in AdaBoosting Model with Default Hyperparameters.

ROC Curves for Train and Test Datasets

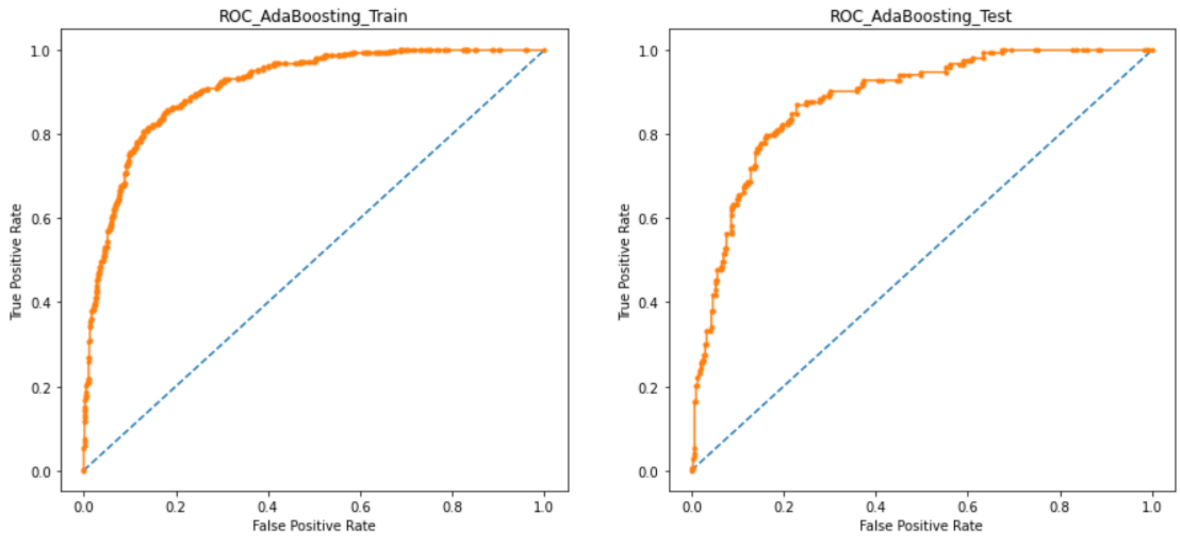


Figure 17. ROC Curves for Train and Test Datasets in AdaBoosting Model with Default Hyperparameters.

Feature Importance

	age	Europe	Blair	Hague	economic_cond_household	economic_cond_national	political_knowledge	gender_male
Feature_Importances	0.4	0.2	0.14	0.12	0.08	0.04	0.02	0.0

Table 31. Features Importance in AdaBoosting Model with Default Parameters.

Inferences

1. Accuracy for test dataset (0.81) is slightly less than accuracy for train dataset (0.85). Hence, there is **no overfitting in AdaBoosting model** and the **model is valid.**
2. ROC_AUC score for train and test datasets are 0.91 and is 0.88 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.85, 0.88 and 0.86 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.74, 0.69 and 0.71 respectively. These scores are less than 0.75. We can try to improve these scores by tuning hyperparameters.
5. **Age and Europe** are two most important features for predicting the target variable
6. **Gender** is not at all useful for prediction of target variable.
7. The **decreasing order of features according their importance** as given below
   Age > Europe > Blair > Hague > economic_cond_household > economic_cond_national > political_knowledge > gender_male

GRADIENT BOOSTING MODEL

Initially Gradient Boosting model has been built **with default hyperparameters** as shown below.

loss='deviance', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=3, min_impurity_decrease=0.0, min_impurity_split=None, init=None, random_state=1, max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, presort='deprecated', validation_fraction=0.1, n_iter_no_change=None, tol=0.0001, ccp_alpha=0.0.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	708	46
Actual 1	68	239

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	276	27
Actual 1	48	105

Table 32. Confusion Matrix for Train and Test Datasets in Gradient Boosting Model with Default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.91	0.94	0.93	754
1	0.84	0.78	0.81	307
accuracy			0.89	1061
macro avg	0.88	0.86	0.87	1061
weighted avg	0.89	0.89	0.89	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.85	0.91	0.88	303
1	0.80	0.69	0.74	153
accuracy			0.84	456
macro avg	0.82	0.80	0.81	456
weighted avg	0.83	0.84	0.83	456

Table 33. Classification Reports for Train and Test Datasets in Gradient Boosting Model with Default Hyperparameters.

ROC Curves for Train and Test Datasets

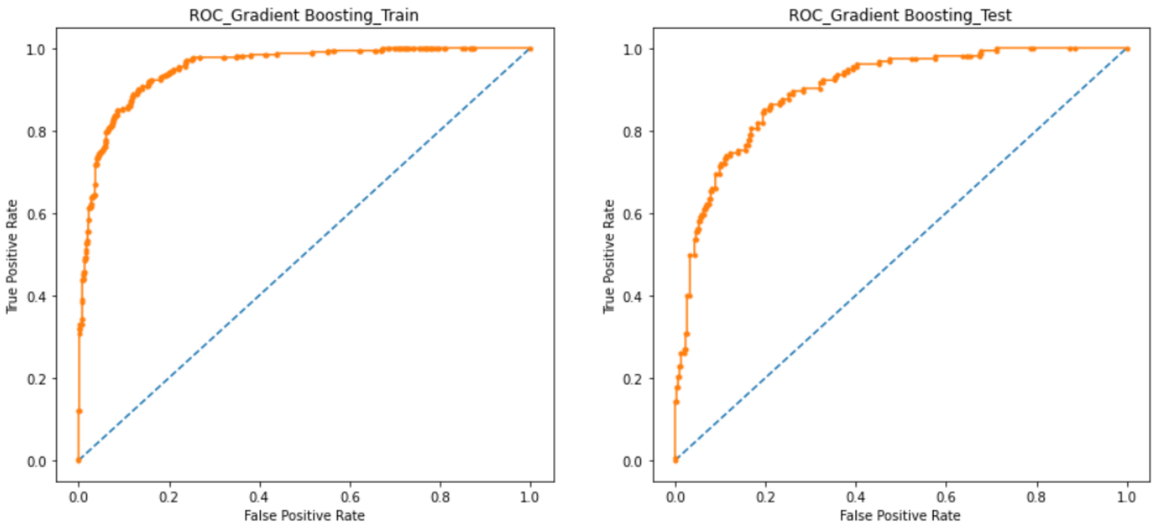


Figure 18. ROC Curves for Train and Test Datasets in Gradient Boosting Model with Default Hyperparameters.

Feature Importance

	Hague	Blair	Europe	age	political_knowledge	economic_cond_national	economic_cond_household	gender_male
Feature_Importances	0.344	0.187	0.175	0.097	0.087	0.077	0.031	0.002

Table 34. Features Importance in Gradient Boosting Model with Default Parameters.

Inferences

1. Accuracy for test dataset (0.84) is slightly less than accuracy for train dataset (0.89). Hence, there is **no overfitting in Gradient Boosting model** and the **model is valid.**
2. ROC_AUC score for train and test datasets are 0.95 and is 0.90 respectively. These scores are pretty good.

3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.85, 0.91 and 0.88 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.80, 0.69 and 0.74 respectively. Recall is less than 0.75. We can try to improve it by tuning hyperparameters.
5. **Hague, Blair** are two most important features for predicting the target variable
6. **Gender** is not at all useful for prediction of target variable.
7. The **decreasing order of features according their importance** as given below
   Hague > Blair > Europe > Age > political_knowledge > economic_cond_national > economic_cond_household > gender_male

Models Tuning by Using GridsearchCV

Tuning Hyperparameters for Logistic Regression Model

Actually, there is **no overfitting in Logistic Regression Model**. Anyway, below hyperparameters have been selected in Logistic Regression model to optimize by using GridSearchCV.

Solver: [newton-cg, lbfgs, liglinear, sag, saga]

- Solver is an algorithm to use in the optimization problem

Best Parameters

Below are the best parameters obtained in Logistic Regression model by using GridSearchCV

Solver: newton-cg

Feature Importance (Coefficients)

Predictor	Hague	Blair	Europe	economic_cond_national	political_knowledge	age	gender_male	economic_cond_household
Coefficients	1.012	-0.7	0.684	-0.54	0.345	0.231	-0.192	-0.059

Table 35. Coefficients of Features in Logistic Regression Model with Tuned Parameters.

Inferences

1. **Hague, Blair and Europe** are three most important features for predicting the target variable.
2. The **decreasing order of features according their importance** as given below
   Hague > Blair > Europe > economic_cond_national > political_knowledge > age > gender_male > economic_cond_household

Tuning Hyperparameters for Linear Discriminant Analysis Model

Actually, there is **no overfitting in Linear Discriminant Analysis Model**. Anyway, below hyperparameters have been selected to optimize by using GridSearchCV.

Solver: [svd, lsqr, eigen]

- Solver is an algorithm to use in the optimization problem

Best Parameters

Below are the best parameters obtained in Linear Discriminant Analysis model by using GridSearchCV

Solver: svd

Feature Importance (Coefficients)

Predictor	Hague	Blair	Europe	economic_cond_national	political_knowledge	age	gender_male	economic_cond_household
Coefficients	1.142	-0.867	0.729	-0.519	0.463	0.311	-0.149	-0.047

Table 36. Coefficients of Features in LDA Model with Tuned Parameters.

Inferences

1. **Hague and Blair** are two most important features for predicting the target variable
2. The **decreasing order of features according their importance** as given below

   Hague > Blair > Europe > economic_cond_national > political_knowledge > age > gender_male > economic_cond_household

Tuning Hyperparameters for KNN Model

Actually, there is **no overfitting in KNN Model**. Anyway, below hyperparameters have been selected to optimize by using GridSearchCV.

n_neighbors: [41,45,51,55,61,65,71] – Number of neighbors

- n_ neighbors are selected as approximately square root of number of observations and tuned by using GridsearchCV

Best Parameters

Below are the best parameters obtained in KNN model by using GridSearchCV

n_neighbors: 61

Tuning Hyperparameters for Bagging Model

Below hyperparameters have been selected in Bagging Model to optimize by using GridSearchCV.

n_estimators: [51,75,101]

- n_estimators are the number of base estimators (Decision Trees) used in Bagging Classifier.

Best Parameters

Below are the best parameters obtained in Bagging model by using GridSearchCV

n_estimators: 75

Tuning Hyperparameters for Random Forest Model

Below hyperparameters have been selected in Random Forest Model to optimize by using GridSearchCV.

- max_depth: [7,8,9] - The maximum depth of the tree. It is selected based the depth upto which tree has grown uniformly.
- max_feature: [3,4,5] - The number of features to consider when looking for the best split. It is selected approximately square root of number of features.
- min_samples_leaf: [1,5,10] - The minimum number of samples required to be at a leaf node. It is selected approximately 1-2% of observations in train dataset.
- min_samples_split: [2,15,30] - The minimum number of samples required to split an internal node. It is selected approximately three times of minimum samples leaf.
- n_estimators: [25,51,101] - The number of trees in the forest.

Best Parameters

Below are the best parameters obtained in Random Forest model by using GridSearchCV

Number of Estimators: 51

Maximum Features: 4

Maximum Depth: 8

Minimum samples leaf: 1

Minimum samples split: 15

Feature Importance

	Hague	Blair	Europe	age	economic_cond_national	political_knowledge	economic_cond_household	gender_male
Feature_Importances	0.272	0.206	0.179	0.121	0.092	0.072	0.047	0.011

Table 37. Features Importance in Random Forest Model with Tuned Parameters.

Inferences

1. **Hague and Blair** are two most important features for predicting the target variable.
2. The **decreasing order of features according their importance** as given below

Hauge > Blair > Europe > Age > economic_cond_national > political_knowledge > economic_cond_household > gender_male

Tuning Hyperparameters for AdaBoosting Model

Below hyperparameters have been selected in AdaBoosting Model to optimize by using GridSearchCV.

n_estimators: [15,25,50]

- n_estimators are the maximum number of estimators (trees) at which boosting is terminated

Best Parameters

Below are the best parameters obtained in AdaBoosting model by using GridSearchCV

n_estimators: 25

Feature Importance

	age	Blair	Hague	Europe	economic_cond_household	economic_cond_national	political_knowledge	gender_male
Feature_Importances	0.28	0.16	0.16	0.16	0.12	0.08	0.04	0.0

Table 38. Features Importance in AdaBoosting Model with Tuned Parameters.

Inferences

1. **Age** is the most important feature for predicting the target variable
2. **Gender** is not at all useful for prediction of target variable.
3. The **decreasing order of features according their importance** as given below

   Age > Blair > Hague > Europe > economic_cond_household > economic_cond_national > political_knowledge > gender_male

Tuning Hyperparameters for Gradient Boosting Model

Below hyperparameters have been selected in Gradient Boosting Model to optimize by using GridSearchCV.

- max_depth: [1,2,3] – It is the maximum depth of the individual regression estimators
- min_samples_leaf: [5,10,15] - The minimum number of samples required to be at a leaf node. It is selected approximately 1-2% of observations in train dataset.
- min_samples_split: [2,15,30] - The minimum number of samples required to split an internal node. It is selected approximately three times of minimum samples leaf.
- n_estimators: [25,51,101] - The number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number usually results in better performance

Best Parameters

Below are the best parameters obtained in Gradient Boosting model by using GridSearchCV

Number of Estimators: 51

Maximum Depth: 2

Minimum samples leaf: 10

Minimum samples split: 2

Feature Importance

	Hague	Blair	Europe	economic_cond_national	political_knowledge	age	economic_cond_household	gender_male
Feature_Importances	0.368	0.22	0.205	0.084	0.082	0.037	0.004	0.0

Table 39. Features Importance in Gradient Boosting Model with Tuned Parameters.

Inferences

1. **Hague, Blair and Europe** are three most important features for predicting the target variable
2. **Gender** is not at all useful for prediction of target variable.
3. The **decreasing order of features according their importance** as given below Hague > Blair > Europe > economic_cond_national > political_knowledge > Age > economic_cond_household > gender_male

Q1.7. 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).

PERFORMANCE METRICS – AFTER TUNING THE MODELS

LOGISTIC REGRESSION MODEL

Logistic Regression model has been built **with the best hyperparameters** obtained by using GridsearchCV

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	686	68
Actual 1	111	196

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	268	35
Actual 1	41	112

Table 40. Confusion Matrix for Train and Test Datasets in Logistic Regression Model with Tuned Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.86	0.91	0.88	754
1	0.74	0.64	0.69	307
accuracy			0.83	1061
macro avg	0.80	0.77	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.87	0.88	0.88	303
1	0.76	0.73	0.75	153
accuracy			0.83	456
macro avg	0.81	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456

Table 41. Classification Reports for Train and Test Datasets in Logistic Regression Model with Tuned Hyperparameters.

Inferences

1. Accuracy for test dataset (0.83) is equal to accuracy for train dataset (0.83). Hence, there is **no overfitting in logistic regression model** and the model is valid.
2. ROC_AUC score for train and test datasets are 0.89 and is 0.88 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.87, 0.88 and 0.88 respectively. These scores are good enough to use the model for predictions.

4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.76, 0.73 and 0.75 respectively. Recall is less than 0.75. Business should be consulted to reduce the class imbalance in target feature and to improve recall.

ROC Curves for Train and Test Datasets

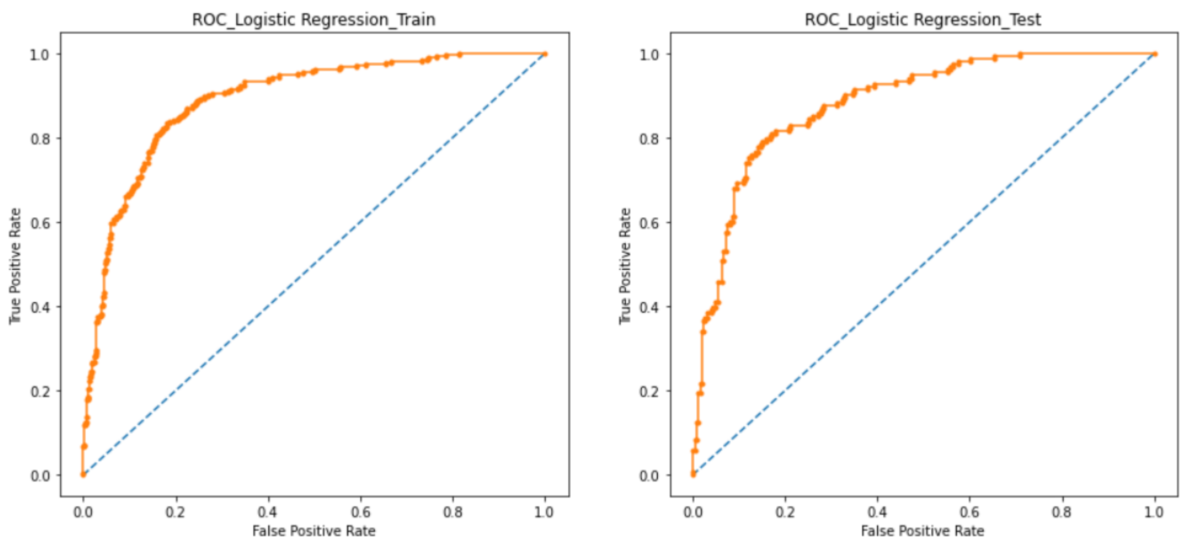


Figure 19. ROC Curves for Train and Test Datasets in Logistic Regression Model with Tuned Hyperparameters.

LINEAR DISCRIMINANT ANALYSIS (LDA)

Linear Discriminant Analysis model has been built **with best hyperparameters** obtained by using GridsearchCV.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	685	69
Actual 1	107	200

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	269	34
Actual 1	42	111

Table 42. Confusion Matrix for Train and Test Datasets in LDA Model with Tuned Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.86	0.91	0.89	754
1	0.74	0.65	0.69	307
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 Test Dataset

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

Table 43. Classification Reports for Train and Test Datasets in LDA Model with Tuned Hyperparameters.

ROC Curves for Train and Test Datasets

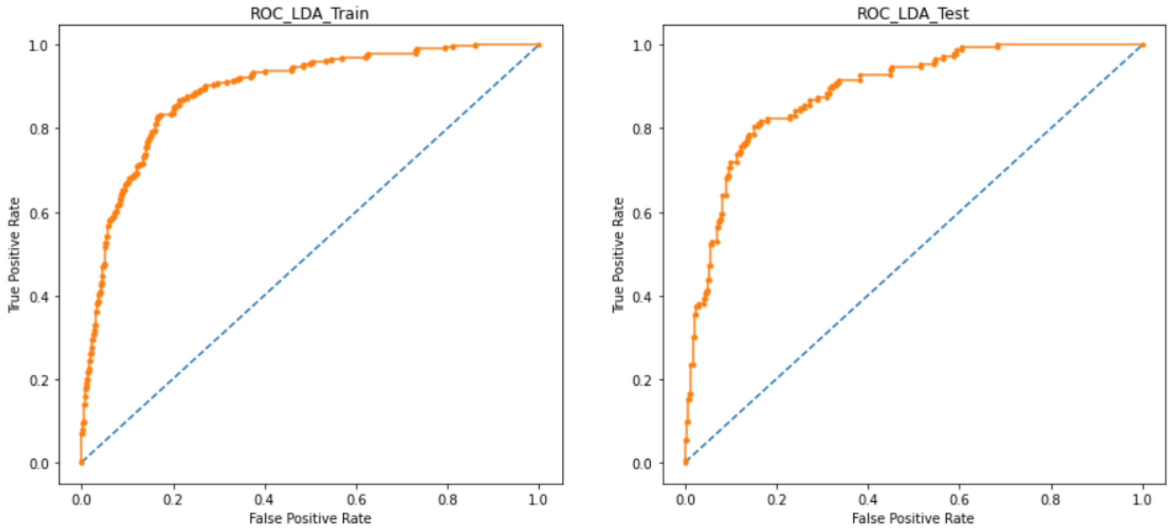


Figure 20. ROC Curves for Train and Test Datasets in LDA Model with Tuned Hyperparameters.

Inferences

1. Accuracy for test dataset (0.83) is equal to accuracy for train dataset (0.83). Hence, there is **no overfitting in Linear Discriminant Analysis model** and the model is valid.
2. ROC_AUC score for train and test datasets are 0.89 and is 0.89 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.86, 0.89 and 0.88 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.77, 0.73 and 0.74 respectively. Recall and F1scores are less than 0.75. Business should be consulted to reduce the class imbalance in target feature and to improve recall.

NAÏVE BAYES MODEL

Naïve Bayes model has been built **with default hyperparameters** only. Because there is **no overfitting in the model** with default hyperparameters and also there are no such important hyperparameters to optimize them.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	675	79
Actual 1	96	211

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	263	40
Actual 1	41	112

Table 44. Confusion Matrix for Train and Test Datasets in Naïve Bayes Model with Tuned Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.88	0.90	0.89	754
1	0.73	0.69	0.71	307
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 Test Dataset

	precision	recall	f1-score	support
0	0.87	0.87	0.87	303
1	0.74	0.73	0.73	153
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

Table 45. Classification Reports for Train and Test Datasets in Naïve Bayes Model with Tuned Hyperparameters.

ROC Curves for Train and Test Datasets

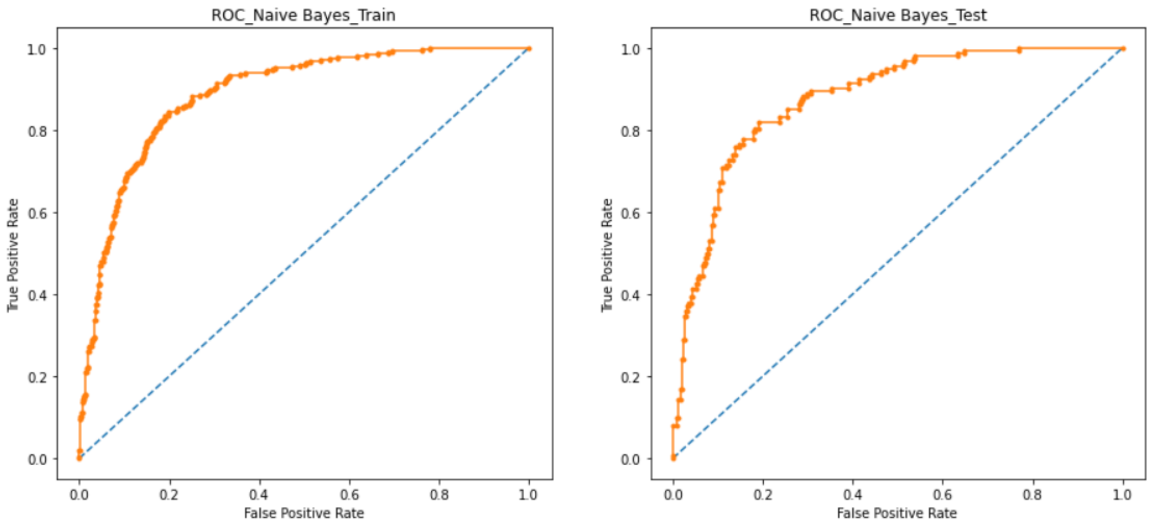


Figure 21. ROC Curves for Train and Test Datasets in Naïve Bayes Model with Tuned Hyperparameters.

Inferences

1. Accuracy for test dataset (0.82) is slightly less than that of train dataset (0.84). Hence, there is **no overfitting in Naïve Bayes model** and the model is valid.
2. ROC_AUC score for train and test datasets are 0.89 and is 0.88 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.87, 0.87 and 0.87 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.74, 0.73 and 0.73 respectively. As these scores are less than 0.75, Business

should be consulted to reduce the class imbalance in target feature and to improve recall and F1 score.

KNN MODEL

K-nearest neighbors model has been built with the best hyperparameters obtained by using GridsearchCV.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	699	55
Actual 1	122	185

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	276	27
Actual 1	52	101

Table 46. Confusion Matrix for Train and Test Datasets in KNN Model with Tuned Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.85	0.93	0.89	754
1	0.77	0.60	0.68	307
accuracy			0.83	1061
macro avg	0.81	0.76	0.78	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.84	0.91	0.87	303
1	0.79	0.66	0.72	153
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.82	0.83	0.82	456

Table 47. Classification Reports for Train and Test Datasets in KNN Model with Tuned Hyperparameters.

ROC Curves for Train and Test Datasets

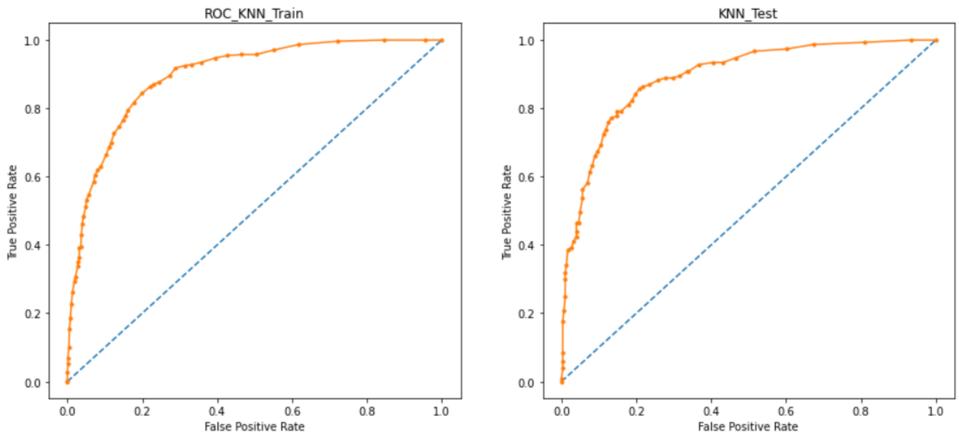


Figure 22. ROC Curves for Train and Test Datasets in KNN Model with Tuned Hyperparameters.

Inferences

1. Accuracy for test dataset (0.83) is equal to that of train dataset (0.83). Hence, there is **no overfitting in KNN model** and the model is valid.
2. ROC_AUC score for train and test datasets are 0.90 and is 0.89 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.84, 0.91 and 0.87 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.79, 0.66 and 0.72 respectively. Recall and F1 score are less than 0.75. Business should be consulted to reduce the class imbalance in target feature and to improve recall and F1 score.

BAGGING MODEL

Bagging Classifier model has been built **with the best hyperparameters** obtained by using GridsearchCV.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	754	0
Actual 1	0	307

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	265	38
Actual 1	44	109

Table 48. Confusion Matrix for Train and Test Datasets in Bagging Classifier Model with Tuned Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	1.00	1.00	1.00	754
1	1.00	1.00	1.00	307
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 Test Dataset

	precision	recall	f1-score	support
0	0.86	0.87	0.87	303
1	0.74	0.71	0.73	153
accuracy			0.82	456
macro avg	0.80	0.79	0.80	456
weighted avg	0.82	0.82	0.82	456

Table 49. Classification Reports for Train and Test Datasets in Bagging Classifier Model with Tuned Hyperparameters.

ROC Curves for Train and Test Datasets

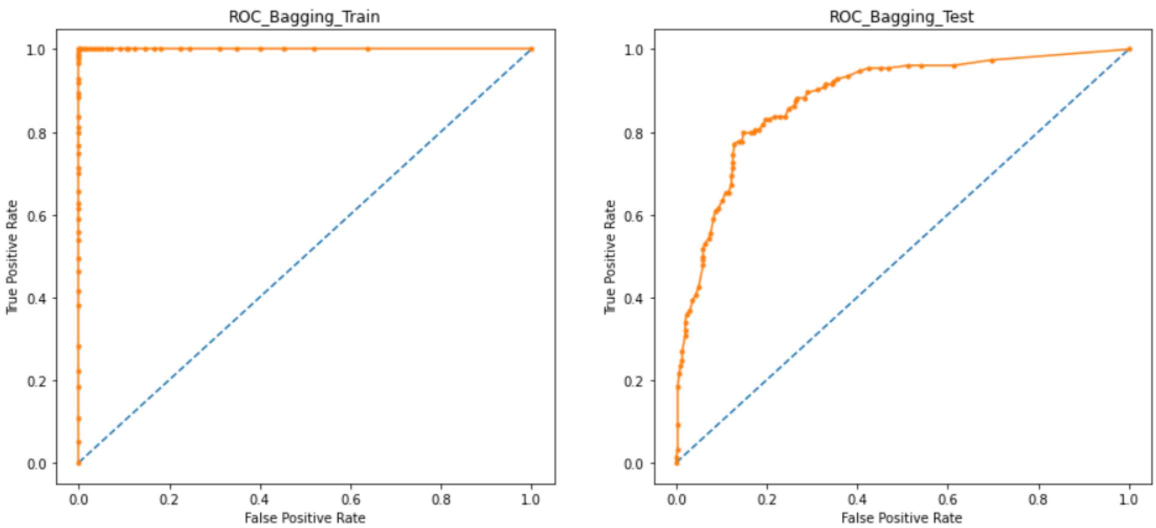


Figure 23. ROC Curves for Train and Test Datasets in Bagging Classifier Model with Tuned Hyperparameters.

Inferences

1. Accuracy for test dataset (0.82) is much lesser than accuracy for train dataset (1). Hence, there is **OVERFITTING in Bagging Classifier model even after tuning hyperparameters** and so the **model is not valid.**
2. ROC_AUC score for train and test datasets are 1 and is 0.88 respectively. These scores are pretty good but **the model is overfitted**.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.86, 0.87 and 0.87 respectively. These scores are good enough to use the model for predictions but before proceeding for predictions, overfitting should be removed.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.74, 0.71 and 0.73 respectively. These scores are less than 0.75. Business should be consulted to reduce the class imbalance in target feature and to improve these scores. Overfitting also should be addressed before implementing this model into production.

RANDOM FOREST MODEL

Random Forest model has been built **with the best hyperparameters** obtained by using GridsearchCV.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	712	42
Actual 1	77	230

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	274	29
Actual 1	49	104

Table 50. Confusion Matrix for Train and Test Datasets in Random Forest Model with default Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.90	0.94	0.92	754
1	0.85	0.75	0.79	307
accuracy			0.89	1061
macro avg	0.87	0.85	0.86	1061
weighted avg	0.89	0.89	0.89	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.85	0.90	0.88	303
1	0.78	0.68	0.73	153
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.83	0.83	0.83	456

Table 51. Classification Reports for Train and Test Datasets in Random Forest Model with Tuned Hyperparameters.

ROC Curves for Train and Test Datasets

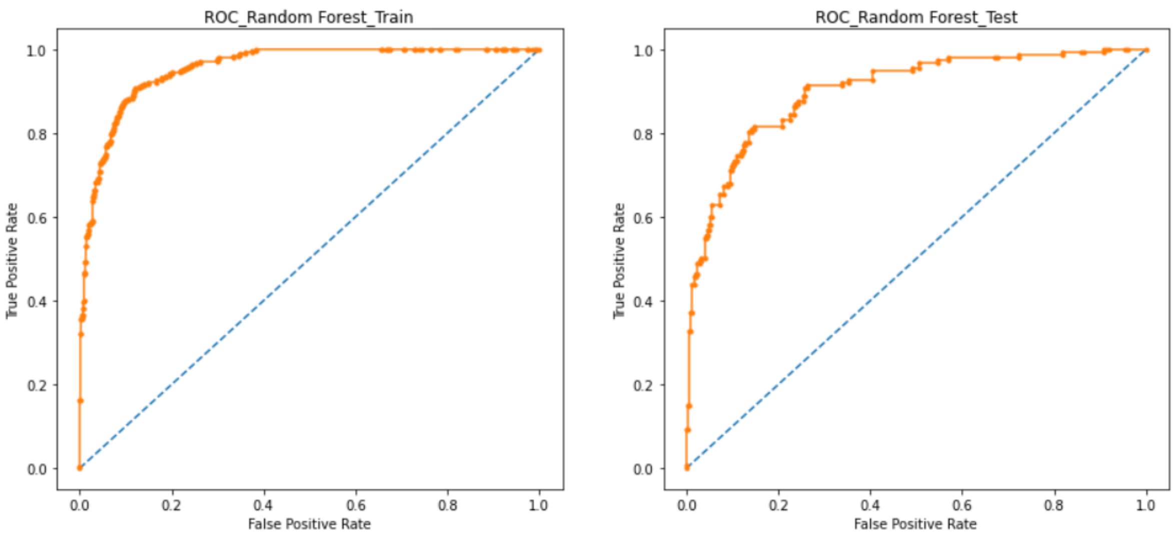


Figure 24. ROC Curves for Train and Test Datasets in Random Forest Model with Tuned Hyperparameters.

Inferences

1. Accuracy for test dataset (0.83) is slightly less than accuracy for train dataset (0.89). Hence, there is **no overfitting in Random Forest model** and the **model is valid.**
2. ROC_AUC score for train and test datasets are 0.96 and is 0.9 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.85, 0.90 and 0.88 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.78, 0.68 and 0.73 respectively. Recall and F1 score are less than 0.75. Business

should be consulted to reduce the class imbalance in target feature and to improve recall and F1 Score.

AdaBoosting Model

AdaBoosting model has been built **with the best hyperparameters** obtained by using GridsearchCV.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	689	65
Actual 1	97	210

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	266	37
Actual 1	46	107

Table 52. Confusion Matrix for Train and Test Datasets in AdaBoosting Model with Tuned Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.88	0.91	0.89	754
1	0.76	0.68	0.72	307
accuracy			0.85	1061
macro avg	0.82	0.80	0.81	1061
weighted avg	0.84	0.85	0.84	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.85	0.88	0.87	303
1	0.74	0.70	0.72	153
accuracy			0.82	456
macro avg	0.80	0.79	0.79	456
weighted avg	0.82	0.82	0.82	456

Table 53. Classification Reports for Train and Test Datasets in AdaBoosting Model with Tuned Hyperparameters.

ROC Curves for Train and Test Datasets

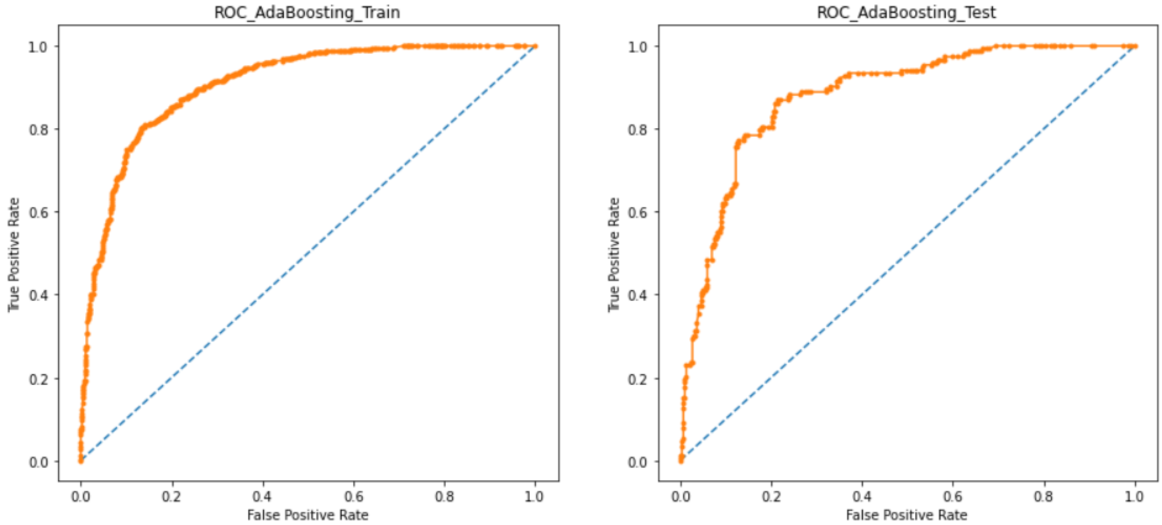


Figure 25. ROC Curves for Train and Test Datasets in AdaBoosting Model with Tuned Hyperparameters.

Inferences

1. Accuracy for test dataset (0.82) is slightly less than accuracy for train dataset (0.85). Hence, there is **no overfitting in AdaBoosting model** and the **model is valid.**
2. ROC_AUC score for train and test datasets are 0.91 and is 0.88 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.85, 0.88 and 0.87 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.74, 0.70 and 0.72 respectively. These scores are less than 0.75. Business should be consulted to reduce the class imbalance in target feature and to improve these scores.

GRADIENT BOOSTING MODEL

Gradient Boosting model has been built **with the best hyperparameters** obtained by using GridsearchCV.

Confusion Matrix for Train and Test Datasets

Confusion matrix of Train Dataset

	Predicted 0	Predicted 1
Actual 0	697	57
Actual 1	95	212

Confusion matrix of Test Dataset

	Predicted 0	Predicted 1
Actual 0	275	28
Actual 1	48	105

Table 54. Confusion Matrix for Train and Test Datasets in Gradient Boosting Model with Tuned Hyperparameters.

Classification Reports for Train and Test Datasets

Classification Report of Train Dataset

	precision	recall	f1-score	support
0	0.88	0.92	0.90	754
1	0.79	0.69	0.74	307
accuracy			0.86	1061
macro avg	0.83	0.81	0.82	1061
weighted avg	0.85	0.86	0.85	1061

Classification Report of Test Dataset

	precision	recall	f1-score	support
0	0.85	0.91	0.88	303
1	0.79	0.69	0.73	153
accuracy			0.83	456
macro avg	0.82	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

Table 55. Classification Reports for Train and Test Datasets in Gradient Boosting Model with Tuned Hyperparameters.

ROC Curves for Train and Test Datasets

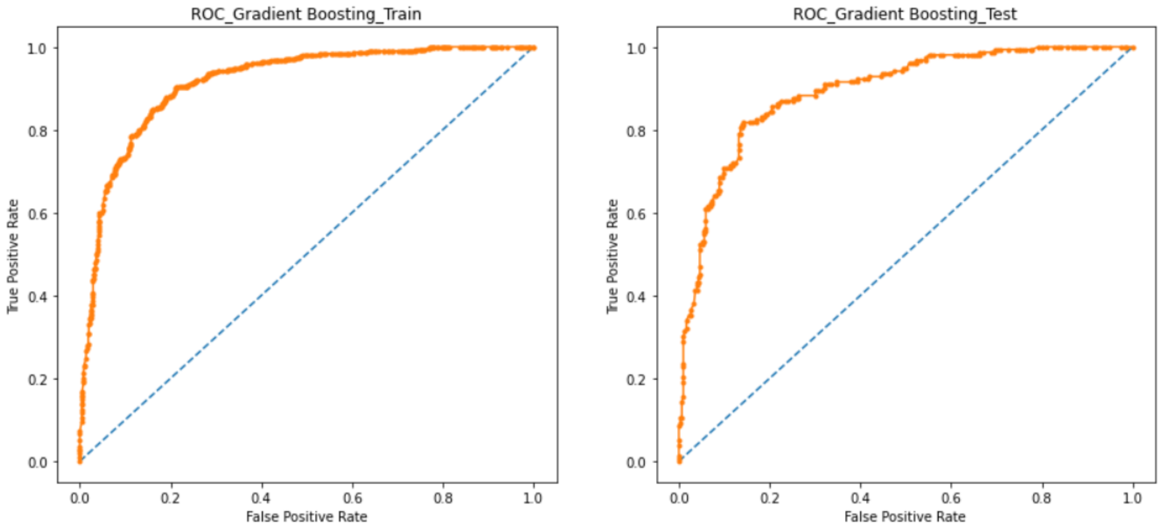


Figure 26. ROC Curves for Train and Test Datasets in Gradient Boosting Model with Tuned Hyperparameters.

Inferences

1. Accuracy for test dataset (0.83) is slightly less than accuracy for train dataset (0.86). Hence, there is **no overfitting in Gradient Boosting model** and the **model is valid.**
2. ROC_AUC score for train and test datasets are 0.92 and is 0.89 respectively. These scores are pretty good.
3. Precision, recall and F1 score **for majority class (Labour party)** in test dataset are 0.85, 0.91 and 0.88 respectively. These scores are good enough to use the model for predictions.
4. Precision, recall and F1 score **for minority class (Conservative party)** in test dataset are 0.79, 0.69 and 0.73 respectively. Recall and F1 scores are less than 0.75. Business should be consulted to reduce the class imbalance in target feature and to improve Recall and F1 scores.

COMPARISON OF MODELS

Performance Metrics of all Models with Default Hyperparameters

	Logistic Regression	LDA	Naive Bayes	KNN	Bagging	Random Forest	AdaBoosting	Gradient Boosting
Accuracy_Train	0.83	0.83	0.84	0.85	0.98	1.00	0.85	0.89
Accuracy_Test	0.83	0.83	0.82	0.83	0.80	0.83	0.81	0.84
AUC_Train	0.89	0.89	0.89	0.93	1.00	1.00	0.91	0.95
AUC_Test	0.88	0.89	0.88	0.87	0.85	0.90	0.88	0.90
Precision_Conservative_Train	0.74	0.74	0.73	0.77	0.99	1.00	0.76	0.84
Precision_Conservative_Test	0.76	0.77	0.74	0.76	0.74	0.79	0.74	0.80
Recall_Conservative_Train	0.64	0.65	0.69	0.70	0.94	1.00	0.68	0.78
Recall_Conservative_Test	0.73	0.73	0.73	0.71	0.62	0.67	0.69	0.69
F1score_Conservative_Train	0.69	0.69	0.71	0.73	0.97	1.00	0.72	0.81
F1score_Conservative_Test	0.75	0.74	0.73	0.73	0.68	0.72	0.71	0.74
Precision_Labour_Train	0.86	0.86	0.88	0.88	0.98	1.00	0.88	0.91
Precision_Labour_Test	0.87	0.86	0.87	0.86	0.82	0.84	0.85	0.85
Recall_Labour_Train	0.91	0.91	0.90	0.91	1.00	1.00	0.91	0.94
Recall_Labour_Test	0.88	0.89	0.87	0.89	0.89	0.91	0.88	0.91
F1score_Labour_Train	0.88	0.89	0.89	0.90	0.99	1.00	0.89	0.93
F1score_Labour_Test	0.88	0.88	0.87	0.87	0.86	0.88	0.86	0.88

Table 56. Performance Metrics of all Models with Default Hyperparameters.

Performance Metrics of all Models with Tuned Hyperparameters

	Logistic Regression	LDA	Naive Bayes	KNN	Bagging	Random Forest	AdaBoosting	Gradient Boosting
Accuracy_Train	0.83	0.83	0.84	0.83	1.00	0.89	0.85	0.86
Accuracy_Test	0.83	0.83	0.82	0.83	0.82	0.83	0.82	0.83
AUC_Train	0.89	0.89	0.89	0.90	1.00	0.96	0.91	0.92
AUC_Test	0.88	0.89	0.88	0.89	0.88	0.90	0.88	0.89
Precision_Conservative_Train	0.74	0.74	0.73	0.77	1.00	0.85	0.76	0.79
Precision_Conservative_Test	0.76	0.77	0.74	0.79	0.74	0.78	0.74	0.79
Recall_Conservative_Train	0.64	0.65	0.69	0.60	1.00	0.75	0.68	0.69
Recall_Conservative_Test	0.73	0.73	0.73	0.66	0.71	0.68	0.70	0.69
F1score_Conservative_Train	0.69	0.69	0.71	0.68	1.00	0.79	0.72	0.74
F1score_Conservative_Test	0.75	0.74	0.73	0.72	0.73	0.73	0.72	0.73
Precision_Labour_Train	0.86	0.86	0.88	0.85	1.00	0.90	0.88	0.88
Precision_Labour_Test	0.87	0.86	0.87	0.84	0.86	0.85	0.85	0.85
Recall_Labour_Train	0.91	0.91	0.90	0.93	1.00	0.94	0.91	0.92
Recall_Labour_Test	0.88	0.89	0.87	0.91	0.87	0.90	0.88	0.91
F1score_Labour_Train	0.88	0.89	0.89	0.89	1.00	0.92	0.89	0.90
F1score_Labour_Test	0.88	0.88	0.87	0.87	0.87	0.88	0.87	0.88

Table 57. Performance Metrics of all Models with Tuned Hyperparameters.

Inferences:

From above tables, we can derive below inferences.

1. **Bagging and Random Forest models are overfitted** when modelled with default hyperparameters. Except these two models**, no other model is overfitted**.
2. **Overfitting of Random Forest** model has been **reduced to allowable limit** by tuning hyperparameters by using GridSearchCV but **Bagging model is overfitted** even after tuning hyperparameters. **Hence, bagging model should not be used for predictions in this problem until overfitting is addressed.**
3. Gradient Boosting model has better accuracy, precision, recall and F1 score with default hyperparameters compared to tuned hyperparameters. Hence, **Model tuning is not improving the performance of Gradient Boosting model**.
4. Accuracy, ROC-AUC for different models on test dataset are ranging from (0.82 to 0.83), (0.88 to 0.9) respectively. Precision, Recall, F1 score for minority class (Conservative Party) for different models on test dataset are ranging from (0.74 to 0.79), (0.66 to 0.73) and (0.72 to 0.75) respectively. Precision, Recall, F1 score for majority class (Labour Party) for different models on test dataset are ranging from (0.84 to 0.87), (0.87 to 0.91) and (0.87 to 0.88) respectively. **As there is no much difference in performance metrics for different models on test dataset, in general any model can be used for predictions except bagging for this problem.**
5. Specifically, if **overall accuracy** of the model and **recall of minority class** (Conservative Party) **are important** for business, **Logistic Regression and LDA** models are best optimized.
6. Specifically, if **overall accuracy** of the model and **recall of majority class** (Labour Party) **are important** for business, **KNN and Gradient Boosting** models are best optimized.
7. Specifically, if **overall accuracy** of the model and **F1 score of both classes are important** for business, then **Logistic Regression and LDA** models are best optimized.
8. **Logistic regression** can be considered as **the best optimized model** by considering all performance metrics because most of the **performance metrics got highest values** in Logistic Regression Model. Apart from that Logistic Regression model can be easily explained with feature coefficients in the form of a linear equation.

Q1.8. 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 to the Management

1. The voters who assessed **high ratings (4 & 5) to Hague** (Conservative party leader) are preferring to **vote Conservative party** whereas voters who assessed **low ratings (1 & 2) to Hague** (Conservative party leader) are preferring to **vote Labour party**. **Hague is the most important feature in predicting target class.** Hague feature got the **highest coefficient in most of the models**. Business should focus on this feature to get accurate information about this feature and making use of this feature as strong predictor.
2. The voters who assessed **high ratings (4 & 5) to Blair** (Labour party leader) are preferring to **vote Labour party** whereas voters who assessed **low ratings (1 & 2) to Blair** (Labour party leader) are preferring to **vote Conservative party**. **Blair** is the **second most important feature** in predicting target class. Blair feature got the **second highest coefficient** in most of the models. Business should focus on this feature to get accurate information about this feature and making use of this feature as one of the strong predictors.
3. The voters who have **high Eurosceptic sentiment (9,10 & 11)** are preferring to **vote Conservative party** whereas voters who **have low Eurosceptic sentiment (less than 9)** are preferring to **vote Labour party. Europe** is the **third most important feature** in predicting target class. Europe feature got the **third highest coefficient** in most of the models. Business should focus on this feature to get accurate information about this feature and making use of this feature as one of the strong predictors.
4. Median age of voters voting for Conservative party is more than that of Labour party.
5. All the voters with **different levels of political knowledge** are preferring to **vote Labour party** than conservative party.

6. The voters who assessed **high ratings (3, 4 & 5) to economic_condtion_national** are preferring to **vote Labour party** whereas voters who assessed **low ratings (1 & 2) to economic_condtion_national** are preferring to **vote Conservative party**.
7. The voters who assessed **high ratings (3, 4 & 5) to economic_condtion_household** are preferring to **vote Labour party** whereas voters who **assessed low ratings (1 & 2) to economic_condtion_household** are preferring to **vote Conservative party**.
8. Age, political knowledge economic_condtion_national and economic_condtion_household are moderate predictors of target class because they got moderate coefficient values in most of the models.
9. Both **male and female voters** are preferring **to vote Labour party** than conservative party. Gender got **least coefficient** value in most of the models. Hence, it is **a weak predictor of target class. Business can ignore this variable.**
10. As there is no much difference in performance metrics for different models on test dataset, in general any model can be used for predictions except bagging for this problem.
11. Specifically, if **overall accuracy** of the model and **F1 score of both classes are important** for business, then **Logistic Regression and LDA** models are best optimized.
12. **Logistic regression** can be considered as **the best optimized model** by considering all performance metrics because most of the **performance metrics got highest values** in Logistic Regression Model. Apart from that Logistic Regression model can be easily explained with feature coefficients in the form of a linear equation.

PROBLEM 2

Problem Statement:

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

Q2.1. Find the number of characters, words and sentences for the mentioned documents.

Count of Characters, Words and Sentences Excluding Spaces (by using .split() function)

	number_of_characters	number_of_words	number_of_sentences
1941-Roosevelt.txt	6174	1360	68
1961-Kennedy.txt	6202	1390	52
1973-Nixon.txt	8122	1819	68

Table 58. Count of Characters, Words and Sentences Excluding Spaces.

Count of Characters, Words and Sentences Including Spaces (by using .words(), .raw(), .sent())

	number_of_characters	number_of_words	number_of_sentences
speech_titles			
1941-Roosevelt.txt	7571	1536	68
1961-Kennedy.txt	7618	1546	52
1973-Nixon.txt	9991	2028	69

Table 59. Count of Characters, Words and Sentences Including Spaces.

Q2.2. 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.

Cleaning Steps Performed

1. Three speeches are converted into lower case.
2. Special characters except alphabets, numbers and underscore are removed.
3. Stopwords are removed.
4. Stemming has been applied.

Word Count Before and After the Removal of Stopwords

	number_of_characters	number_of_characters_after_cleaning	number_of_words	number_of_words_after_cleaning
1941-Roosevelt.txt	6174	3418	1360	627
1961-Kennedy.txt	6202	3584	1390	693
1973-Nixon.txt	8122	4332	1819	833

Table 60. Word Count Before and After the Removal of Stopwords.

Sample Sentence After the Removal of Stopwords

The below the speech of President Franklin D. Roosevelt in 1941 after removing the stopwords.

[nation day inaugur sinc 1789 peopl renew sens dedic unit state washington day task peopl cr eat weld togeth nation lincoln day task peopl preserv nation disrupt within day task peopl sav e nation institut disrupt without us come time midst swift happen paus moment take stock rec al place histori rediscov may risk real peril inact live nation determin count year lifetim huma n spirit life man threescor year ten littl littl less life nation full measur live men doubt men bel iev democraci form govern frame life limit measur kind mystic artifici fate unexplain reason t yranni slaveri becom surg wave futur freedom eb tide american know true eight year ago life republ seem frozen fatalist terror prove true midst shock act act quickli boldli decis later year live year fruit year peopl democraci brought us greater secur hope better understand life ideal measur materi thing vital present futur experi democraci success surviv crisi home put away mani evil thing built new structur endur line maintain fact democraci action taken within thre eway framework constitut unit state coordin branch govern continu freeli function bill right re main inviol freedom elect wholli maintain prophet downfal american democraci seen dire pre dict come naught democraci die know seen reviveand grow know cannot die built unhamp ini ti individu men women join togeth common enterpris enterpris undertaken carri free express f ree major know democraci alon form govern enlist full forc men enlighten know democraci al on construct unlimit civil capabl infinit progress improv human life know look surfac sens stil l spread everi contin human advanc end unconquer form human societi nation like person bod ya bodi must fed cloth hous invigor rest manner measur object time nation like person mind mind must kept inform alert must know understand hope need neighbor nation live within nar row circl world nation like person someth deeper someth perman someth larger sum part som eth matter futur call forth sacr guard present thing find difficult even imposs hit upon singl si mpl word yet understand spirit faith america product centuri born multitud came mani land hi gh degre mostli plain peopl sought earli late find freedom freeli democrat aspir mere recent p hase human histori human histori permeat ancient life earli peopl blaze anew middl age writte n magna charta america impact irresist america new world tongu peopl contin newfound land came believ could creat upon contin new life life new freedom vital written mayflow compact declar independ constitut unit state gettysburg address first came carri long spirit million follo w stock sprang move forward constantli consist toward ideal gain statur clariti gener hope rep ubl cannot forev toler either undeserv poverti selfserv wealth know still far go must greatli bu ild secur opportun knowledg everi citizen measur justifi resourc capac land enough achiev pu rpos alon enough cloth feed bodi nation instruct inform mind also spirit three greatest spirit w ithout bodi mind men know nation could live spirit america kill even though nation bodi min d constrict alien world live america know would perish spirit faith speak us daili live way ofte n unnot seem obviou speak us capit nation speak us process govern sovereignti 48 state speak us counti citi town villag speak us nation hemispher across sea enslav well free sometim fail hear heed voic freedom us privileg freedom old old stori destini america proclaim word proph eci spoken first presid first inaugur 1789 word almost direct would seem year 1941 preserv sa cr fire liberti destini republican model govern justli consid deepli final stake experi intrust ha nd american peopl lose sacr fireif let smother doubt fear shall reject destini washington strove valiantli triumphantli establish preserv spirit faith nation furnish highest justif everi sacrific m ay make caus nation defens face great peril never encount strong purpos protect perpetu integ r democraci muster spirit america faith america retreat content stand still american go forwar d servic countri god']

Q 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)

Speech	Most Frequent Word	Top Three Words
President Franklin D. Roosevelt in 1941	nation (17 times)	nation (17 times) know (10 times) people (9 times)
President John F. Kennedy in 1961	let (16 times)	let (16 times) us (12 times) power (9 times)
President Richard Nixon in 1973	us (26 times)	us (26 times) let (22 times) america (21 times)

Table 61. Top Three Words in Three Speeches

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

1. President Franklin D. Roosevelt in 1941

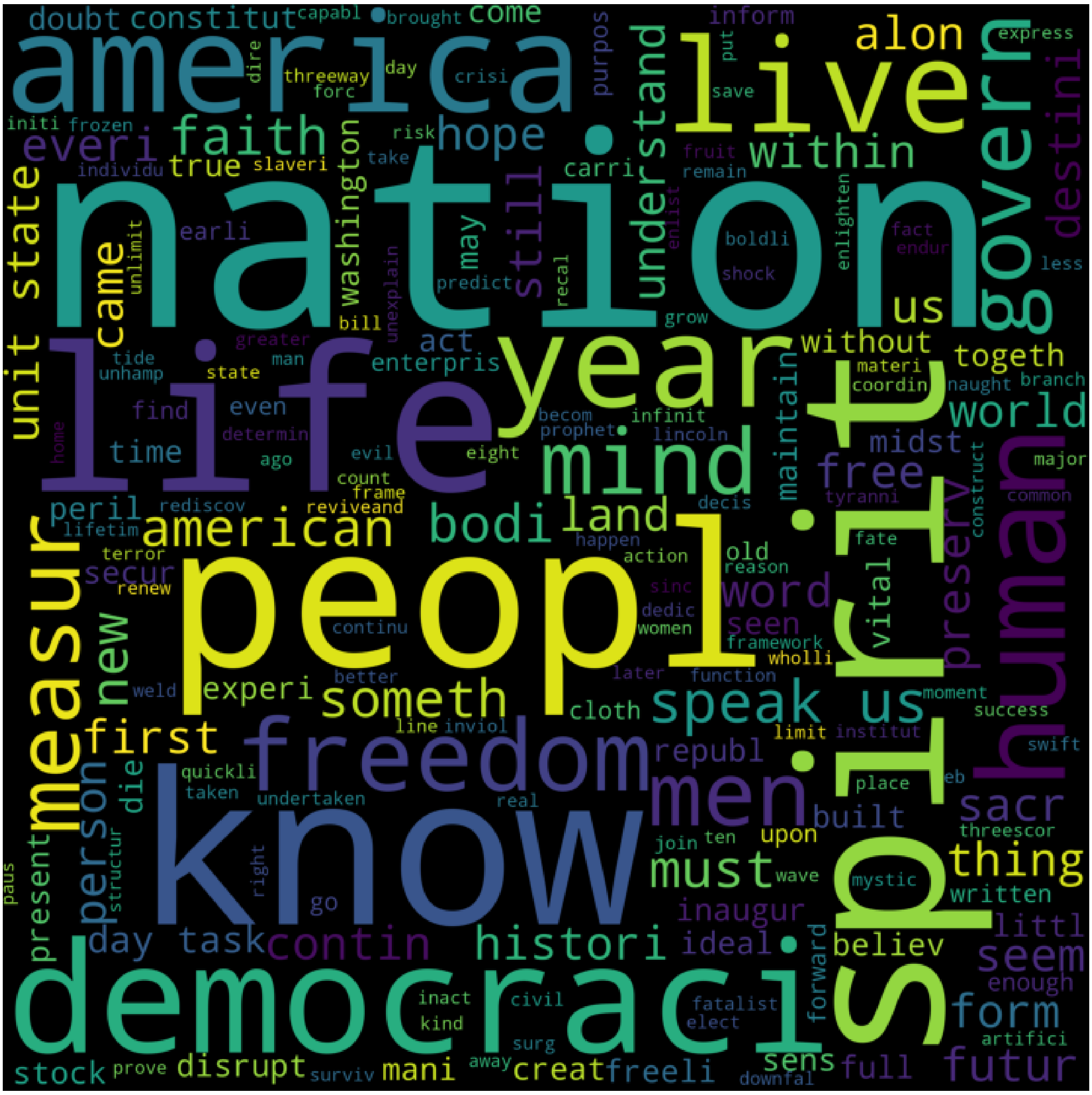


Figure 27. Word Cloud of President Franklin D. Roosevelt Speech in 1941.

2. President John F. Kennedy in 1961

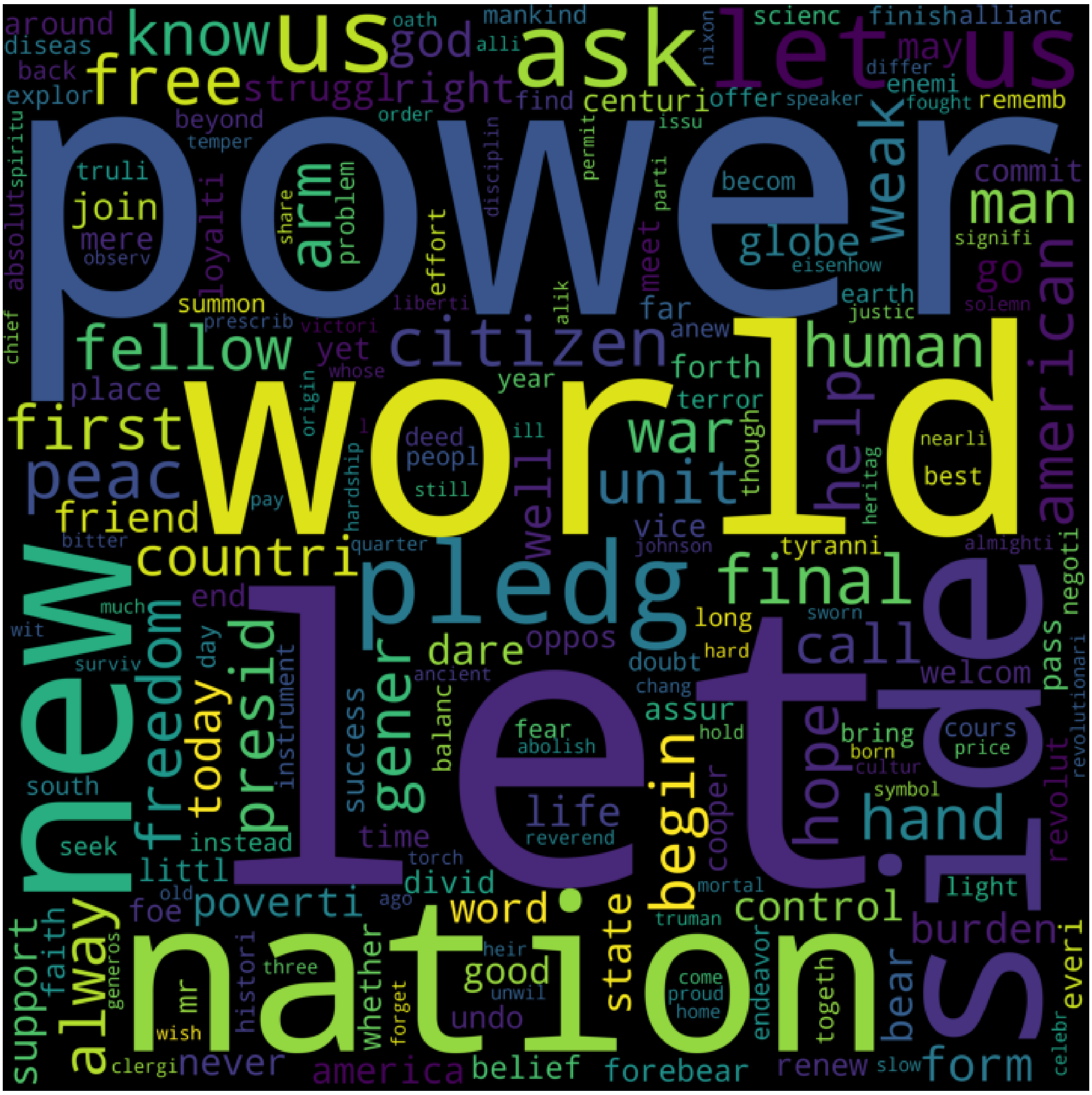


Figure 28. Word Cloud of President John F. Kennedy Speech in 1961.

3. President Richard Nixon in 1973

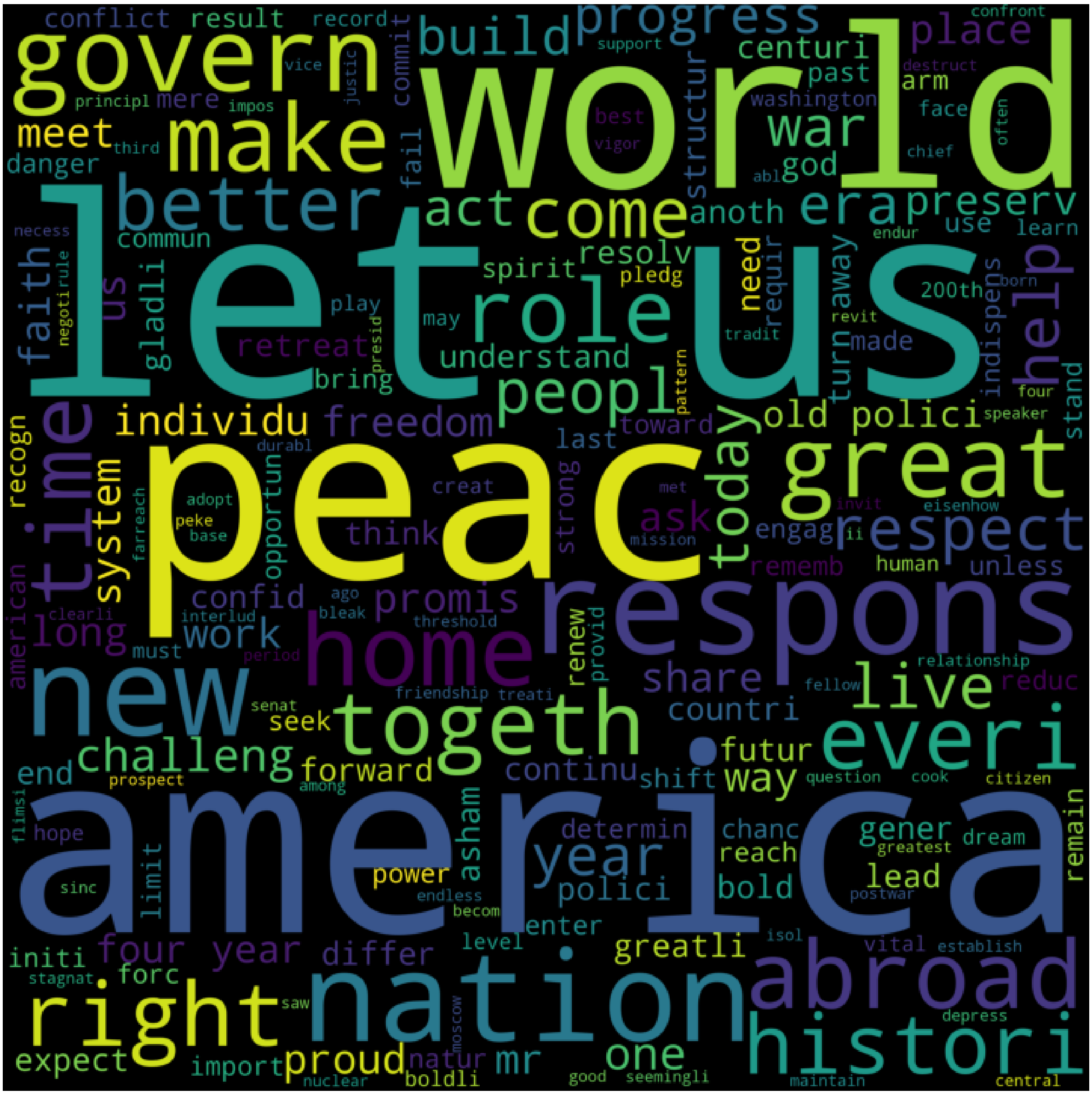


Figure 29. Word Cloud of President Richard Nixon Speech in 1973.