MACHINE LEARNING PROJECT REPORT

-APOORVA P

CONTENT

PROBLEM 1

CONTEXT

CNBE, a prominent news channel, is gearing up to provide insightful coverage of recent elections, recognizing the importance of data-driven analysis. A comprehensive survey has been conducted, capturing the perspectives of 1525 voters across various demographic and socio-economic factors. This dataset encompasses 9 variables, offering a rich source of information regarding voters' characteristics and preferences.

OBJECTIVE

The primary objective is to leverage machine learning to build a predictive model capable of forecasting which political party a voter is likely to support. This predictive model, developed based on the provided information, will serve as the foundation for creating an exit poll. The exit poll aims to contribute to the accurate prediction of the overall election outcomes, including determining which party is likely to secure the majority of seats.

DATA DESCRIPTION

1. **vote**: Party choice: Conservative or Labour
2. **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**: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment.
8. **political.knowledge**: Knowledge of parties' positions on European integration, 0 to 3.
9. **gender:** female or male.

DEFINE THE PROBLEM AND PERFORM EXPLORATORY DATA ANALYSIS

The data is as below

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

FIG 1.1 : DATA.HEAD()

```
data.shape
```

```
(1525, 9)
```

```
data.dtypes
```

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

FIG 1.2 : DATA SHAPE AND DATATYPES

In this data we have, 9 features with 1525 data points.
We can clearly check all features have correct datatypes.

```
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
```

FIG 1.3 : SUM OF NULL VALUES IN DATA

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1525 entries, 1 to 1525
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: 119.1+ KB
```

FIG 1.4 : DATA INFORMATION

9 Features - 7 numerical type & 2 object type, 1525 records, no null values

```
data.duplicated().sum()
```

```
8
```

```
data.drop_duplicates(inplace = True)
data.duplicated().sum()
```

```
0
```

```
data.reset_index(drop = True, inplace = True)
```

```
data.shape
```

```
(1517, 9)
```

FIG 1.5 : DUPLICATES IN DATA

There are 8 duplicates in the dataset that have been removed.

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

FIG 1.6 : DATA DESCRIPTION

There seems to be no presence of bad data.

```
age                      0.14
economic.cond.national  -0.24
economic.cond.household -0.14
Blair                   -0.54
Hague                    0.15
Europe                  -0.14
political.knowledge     -0.42
gender                   0.13
dtype: float64
```

FIG 1.7 : DATA SKEW

The magnitude of the Skewness of all the variables is approximately less than or around |0.5|.

ISIGHTS:

- The dataset had 8 duplicated values. So, we are dropped them.
- The data set had 1525 rows and 9 columns. After dropping the duplicate values, there are 1517 rows and 9 columns.
- It has 7 numerical data types and 2 categorical data types.
- There is no null value in any column.
- Here, we can see that there isn't much skewness in the data. All the values seems to be between -0.5 and 0.5.
- The value of 'Blair' is a little bit higher than -0.5.
- The data overall, is fairly symmetrical.

UNIVARIATE ANALYSIS:

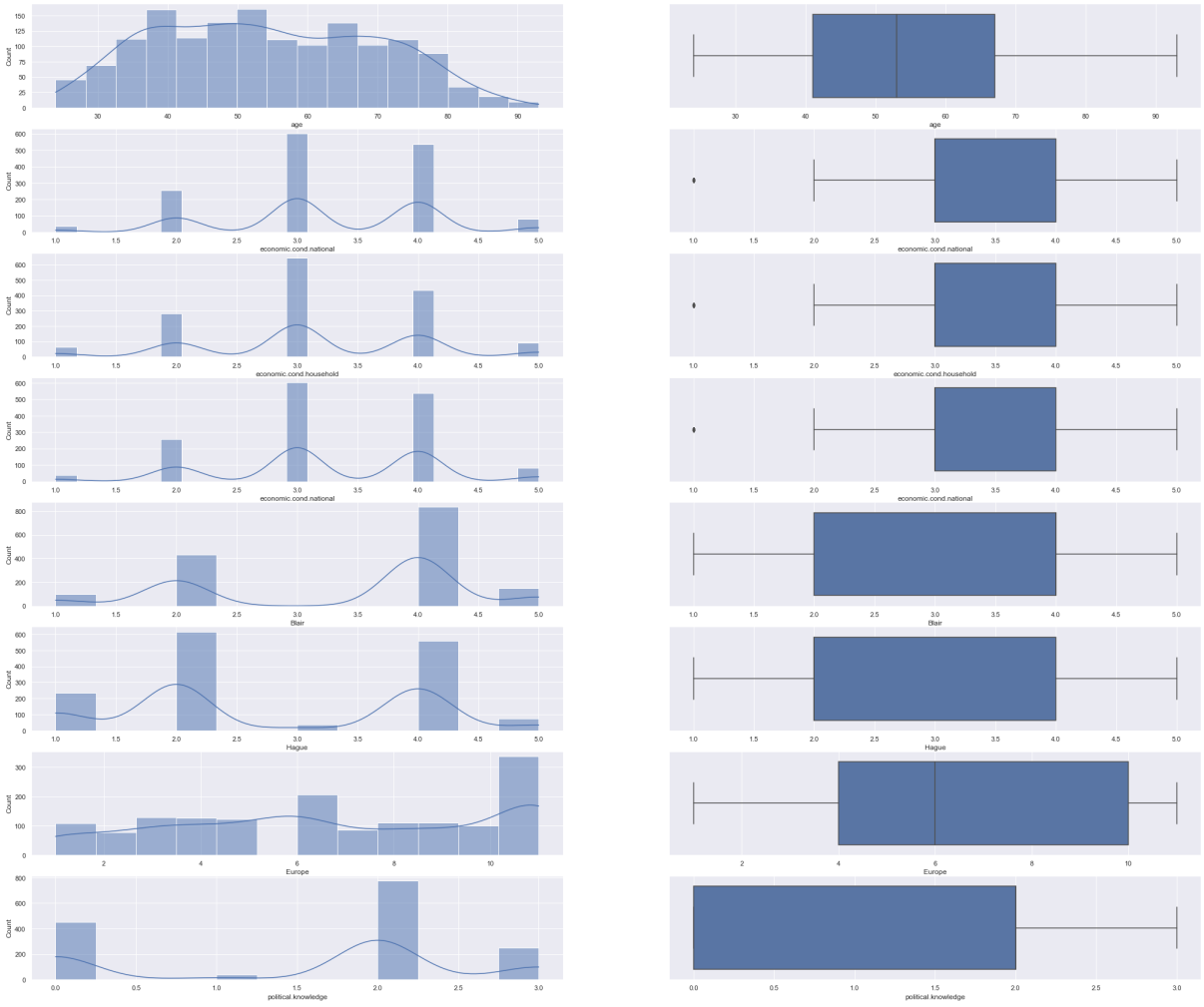


FIG 1.8 : UNIVERIATE ANALYSIS

1. Age variable is very slightly right skewed with the presence of no outliers. The variable ranges from 24 to 93.
2. Maximum number of people are aged between 40 and 70

```
economic.cond.national %age Count
                     3    39.78
                     4    35.49
                     2    16.89
                     5     5.41
                     1     2.44
Name: economic.cond.national, dtype: float64
                Average Score = 3.25
```

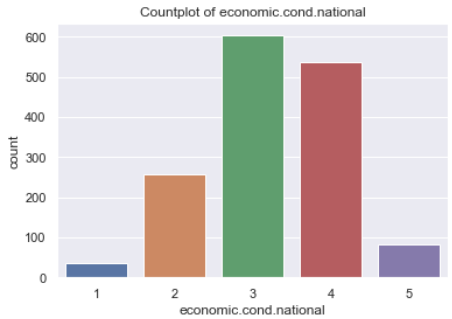


FIG 1.9 : COUNTPLOT OF ECONOMIC.COND.NATIONAL

- The top 2 variables are 3 and 4.
- 1 has the least value
- 3 has the highest value
- 3 is slightly higher than the 2nd highest variable 4
- The average score of 'economic.cond.national' is 3.25

```
economic.cond.household %age Count
                      3    42.48
                      4    28.69
                      2    18.47
                      5     6.07
                      1     4.29
Name: economic.cond.household, dtype: float64
                Average Score = 3.14
```

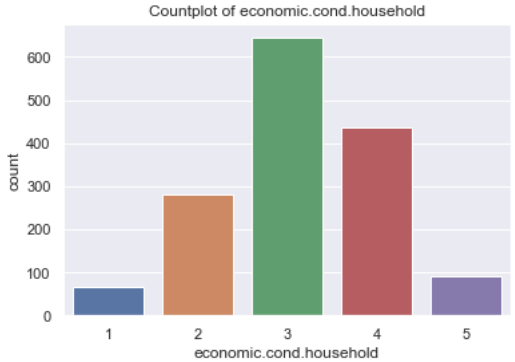


FIG 1.10 : COUNTPLOT OF ECONOMIC.COND.HOUSEHOLD

- The top 2 variables are 3 and 4.
- 1 has the least value
- 3 has the highest value
- 3 is moderately higher than the 2nd highest variable 4

- The average score of 'economic.cond.household' is 3.14

```
Blair %age Count
   4     54.88
   2     28.63
   5     10.03
   1      6.40
   3      0.07
Name: Blair, dtype: float64
   Average Score = 3.34
```

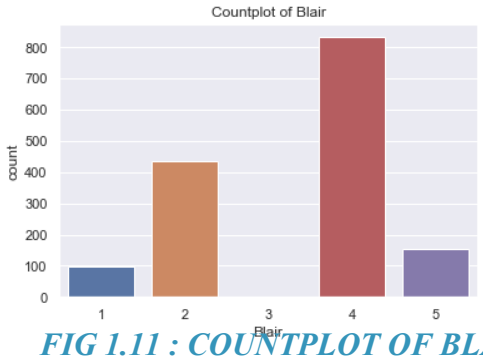


FIG 1.11 : COUNTPLOT OF BLAIR

- The top 2 variables are 2 and 4.
- 3 has the least value.
- 4 has the highest value.
- 4 is much higher than the 2nd highest variable 2.
- The average score of 'Blair' is 3.34

```
Hague %age Count
   2     40.63
   4     36.74
   1     15.37
   5      4.82
   3      2.44
Name: Hague, dtype: float64
   Average Score = 2.75
```

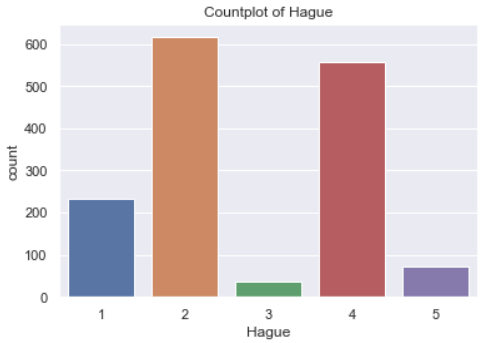


FIG 1.12 : COUNTPLOT OF HAGUE

- The top 2 variables are 2 and 4.
- 3 has the least value.
- 2 has the highest value.
- 2 is slightly higher than the 2nd highest variable 4.
- The average score of 'Blair' is 2.75

```
Europe %age Count
   11     22.30
```

```
6     13.59
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
Average Score = 6.74
```

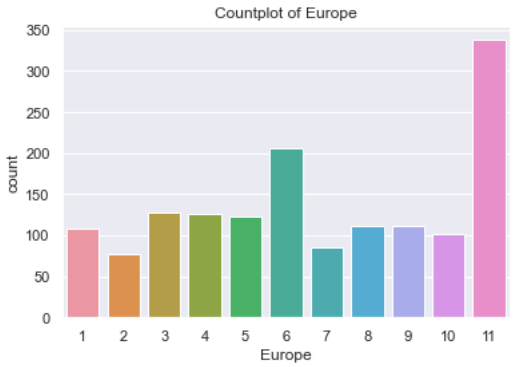


FIG 1.13 : COUNTPLOT OF EUROPE

- The top 2 variables are 11 and 6.
- 2 has the least value.
- 11 has the highest value.
- 11 is moderately higher than the 2nd highest variable 6.
- The average score of 'Europe' is 6.74

```
political.knowledge %age Count
                 2    51.19
                 0    29.88
                 3    16.42
                 1     2.51
Name: political.knowledge, dtype: float64
Average Score = 1.54
```

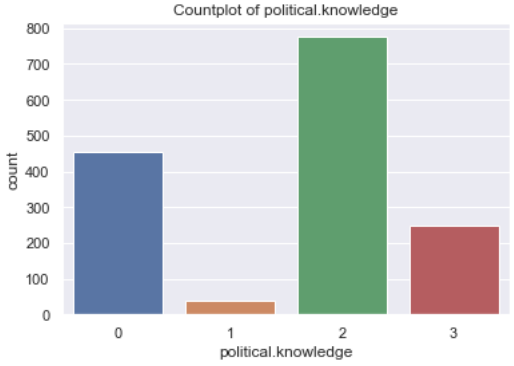


FIG 1.14 : COUNTPLOT OF POLITICAL KNOWLEDGE

- The top 2 variables are 2 and 0
- 1 has the least value.
- 2 has the highest value.

- 2 is much higher than the 2nd highest variable 0.
- We can see that, 454 out of 1517 people do not have any knowledge of parties' positions on European integration which is 29.93% of the total population.
- The average score of 'political.knowledge' is 1.54

```
gender %age Count
     0    53.23
     1    46.77
Name: gender, dtype: float64
Average Score = 0.47
```

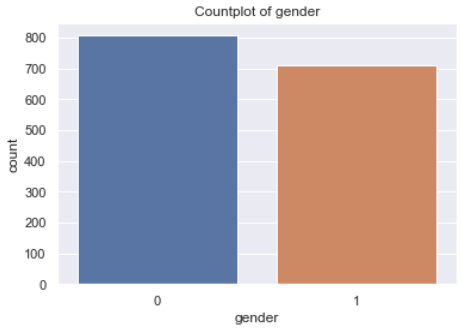


FIG 1.15 : COUNTPLOT OF GENDER

Here 1 is male and 0 is female

```
vote %age Count

     0    69.68
     1    30.32
Name: vote, dtype: float64
```

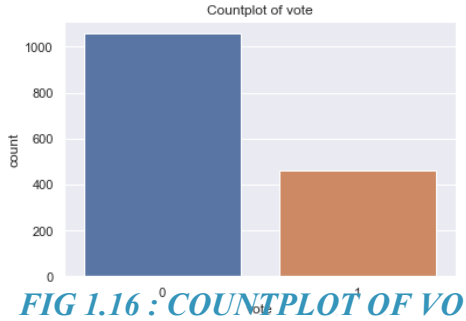


FIG 1.16 : COUNTPLOT OF VOTE

Here 1 is conservatory and 0 is labour

INSIGHTS

1. As per data, Roughly 75% of the people have rated 3 or 4 for the national economic condition with a average score of 3.25
2. Nearly 70% of the people have rated 3 or 4 for the household economic condition with a average score of 3.14
3. Roughly 65% of the people have rated Blair good with a average score of 3.34. Implying that people are happy with labor party.
4. As we can check nearly 55% of the people have rated Hague poorly with a average score of 2.75.
5. Among 1500 people surveyed at random, roughly 70% have voted for the Labour Party & 30% for the conservative party.

6. There is slight class imbalance among the classes of the target variable. But the conservative class is not massively under represented hence we don't necessarily need over/under sampling techniques to fix this issue.

BIVARIATE ANALYSIS:

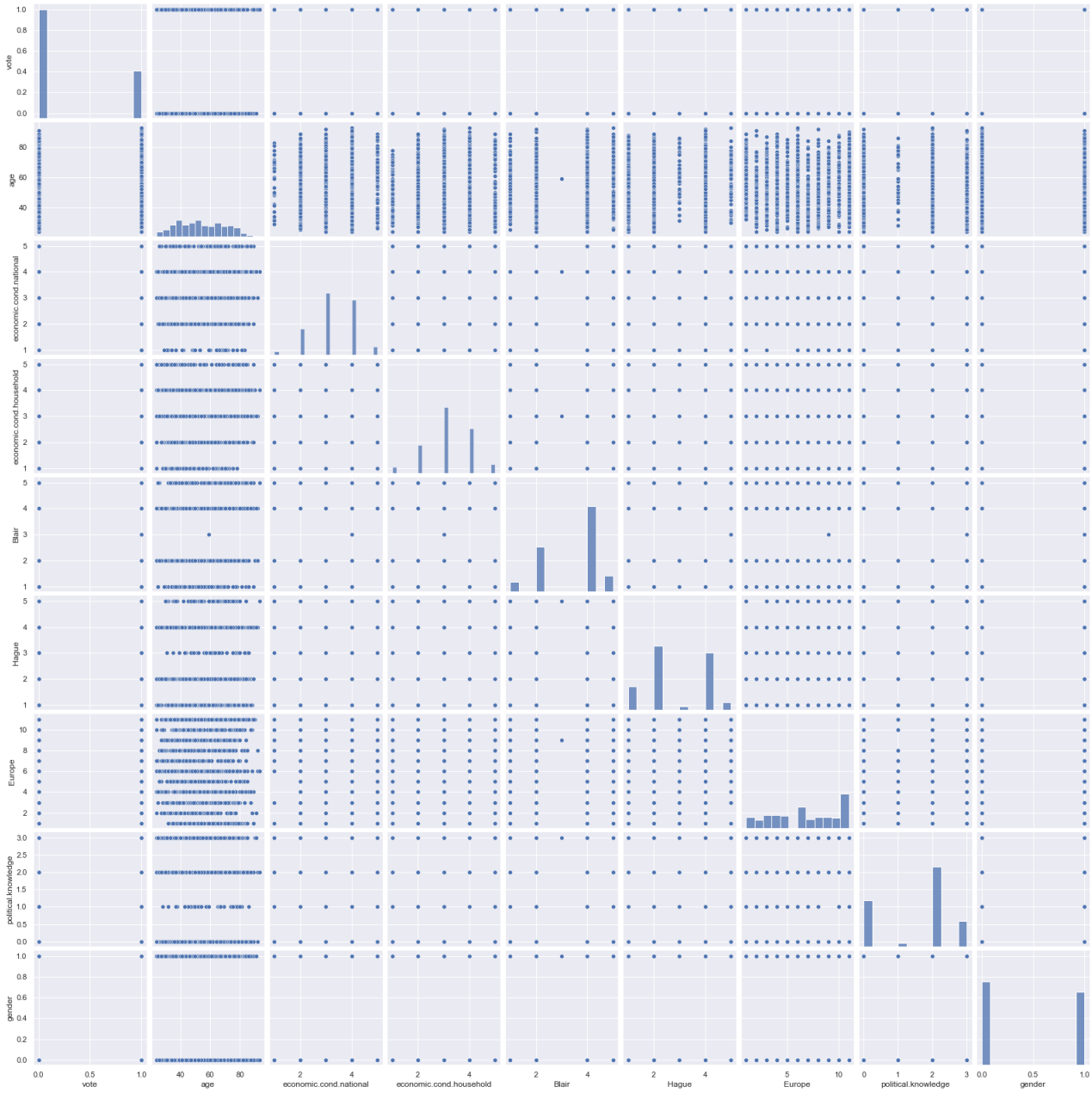


FIG 1.17 : PAIRPLOT

1. As per the visulisation we can see here is a slight positive correlation between the ratings of economic conditions of the nation & the households. This slight positive correlation exists with the ratings of labour party leader, Tony Blair as well and there is a slight negative correlation with the conservative party leader, William Hague. Implying people are generally happy with the current economic conditions and would like Labour party to continue.
2. Blair & Hague have a weak negative correlation between them as is obvious as they are standing against each other in the general election.

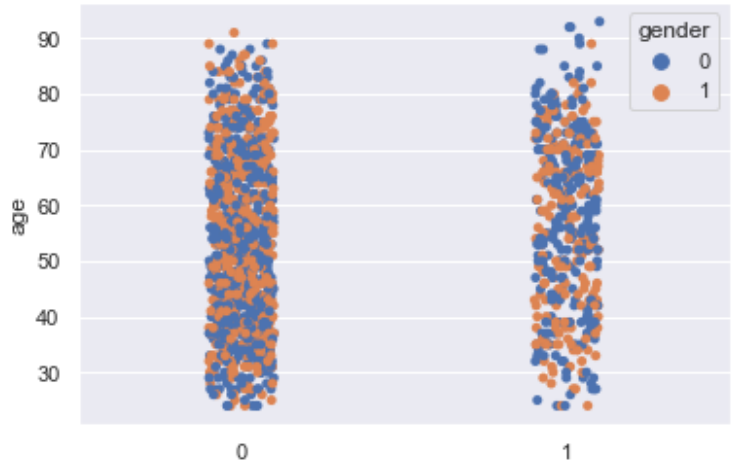


FIG 1.18 : STRIPPLOT WITH VOTE,AGE AND GENDER

- We can clearly see that, the labour party has got more votes than the conservative party.
- In every age group, the labour party has got more votes than the conservative party.
- Female votes are considerably higher than the male votes in both parties.
- In both genders, the labour party has got more votes than the conservative party.

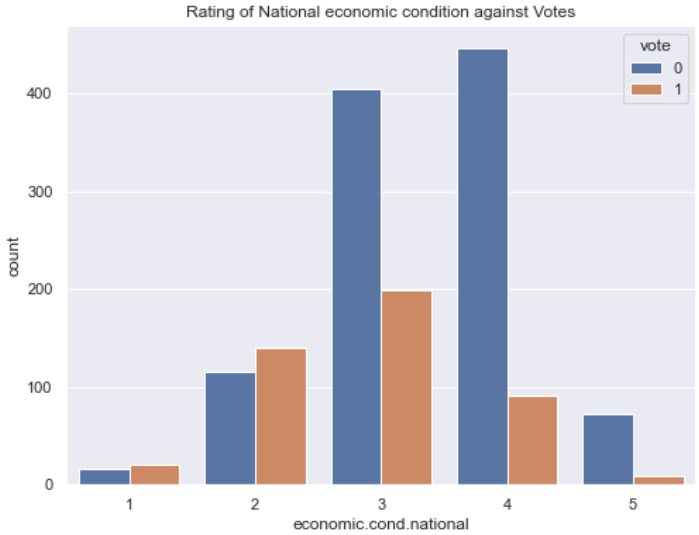


FIG 1.19 : RATING OF NATIONAL ECONOMIC CONDITION AND VOTING

- Labour party has higher votes overall.
- Out of 82 people who gave a score of 5, 73 people have voted for
- the labour party.

- Out of 538 people who gave a score of 4, 447 people have voted for the labour party. This is the highest set of people in the labour party.
- Out of 604 people who gave a score of 3, 405 people have voted for the labour party. This is the 2nd highest set of people in the labour party. The remaining 199 people who have voted for the conservative party is the highest set of people in that party.
- Out of 256 people who gave a score of 2, 116 people have voted for the labour party. 140 people have voted for the conservative party. This is the instance where the conservative party has got more votes than the labour party.
- Out of 37 people who gave a score of 1, 16 people have voted for the labour party. 21 people have voted for the conservative party.
- The score of 3, 4 and 5 have more votes in the labour party.
- The score of 1 and 2 have more votes in the conservative party.

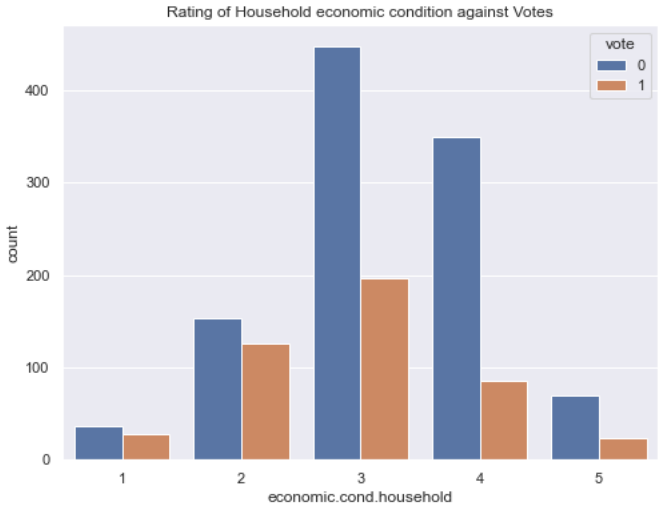


FIG 1.20 : RATING OF HOUSEHOLD ECONOMIC CONDITION AND VOTING

- Labour party has higher votes overall.
- Out of 92 people who gave a score of 5, 69 people have voted for
- the labour party.
- Out of 435 people who gave a score of 4, 349 people have voted for the labour party. This is the 2nd highest set of people in the labour party.
- Out of 645 people who gave a score of 3, 448 people have voted for the labour party. This is the highest set of people in the labour party. The remaining 197 people who have voted for the conservative party is the highest set of people in that party.
- Out of 280 people who gave a score of 2, 154 people have voted for the labour party. 126 people have voted for the conservative party.
- Out of 65 people who gave a score of 1, 37 people have voted for the labour party. 28 people have voted for the conservative party.
- In all the instances, the labour party have more votes than the conservative party.

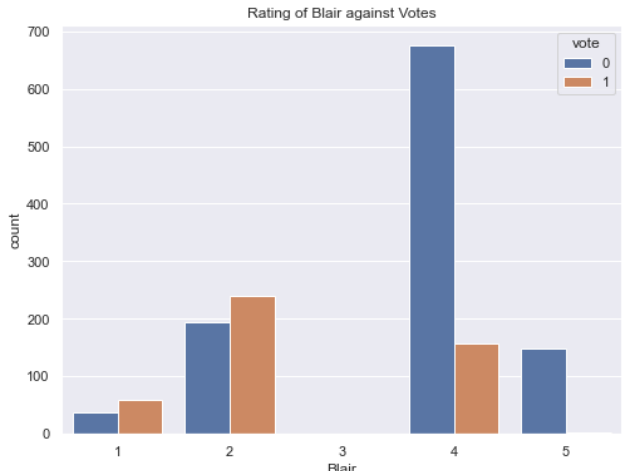


FIG 1.21 : RATING OF BLAIR AND VOTING

- Labour party has higher votes overall.
- Out of 152 people who gave a score of 5, 149 people have voted for the labour party. The remaining 3 people, despite giving a score of 5 to the labour leader, have chosen to vote for the conservative party.
- Out of 833 people who gave a score of 4, 676 people have voted for the labour party. The remaining 157 people, despite giving a score of 4 to the labour leader, have chosen to vote for the conservative party.
- Only 1 person has given a score of 3 and that person hasvoted for the conservative party.
- Out of 434 people who gave a score of 2, 240 people have voted for the conservative party. The remaining 194 people, despite giving an unsatisfactory score of 2 to the labour leader, have chosen to vote for the labour party.
- Out of 97 people who gave a score of 1, 59 people have voted for the conservative party. The remaining 38 people, despite giving the lowest score of 1 to the labourleader, have chosen to vote for the labour party.
- The score of 4 and 5 have more votes in the labour party.
- The score of 1, 2 and 3 have more votes in the conservative party.

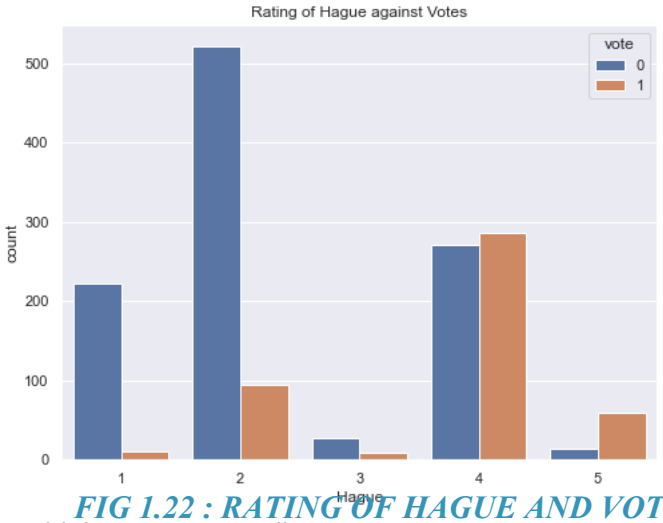


FIG 1.22 : RATING OF HAGUE AND VOTING

- Labour party has higher votes overall.
- Out of 73 people who gave a score of 5, 59 people have voted for the conservative party. The remaining 14people, despite giving a score of 5 to the conservative leader, have chosen to vote for the labour party.

- Out of 557 people who gave a score of 4, 286 people have voted for the conservative party. The remaining 271 people, despite giving a score of 4 to the conservative leader, have chosen to vote for the labour party.
- Out of 37 people who gave a score of 3, 28 have voted for the labour party. The remaining 9, despite giving an average score of 3 to the conservative party, have chosen to vote for the conservative party.
- Out of 617 people who gave a score of 2, 522 people have voted for the labour party. The remaining 95 people, despite giving an unsatisfactory score of 2 to the conservative leader, have chosen to vote for the conservative party.
- Out of 233 people who gave a score of 1, 222 people have voted for the labour party. The remaining 11 people, despite giving the lowest score of 1 to the conservative leader, have chosen to vote for the conservative party.
- The score of 4 and 5 have more votes in the conservative party, although in 4, the votes are almost equal in both the parties. Conservative party gets slightly higher.
- The score of 1, 2 and 3 have more votes in the labour party. Still, a significant percentage of people who gave a bad score to the conservative leader still chose to vote for 'Hague'.

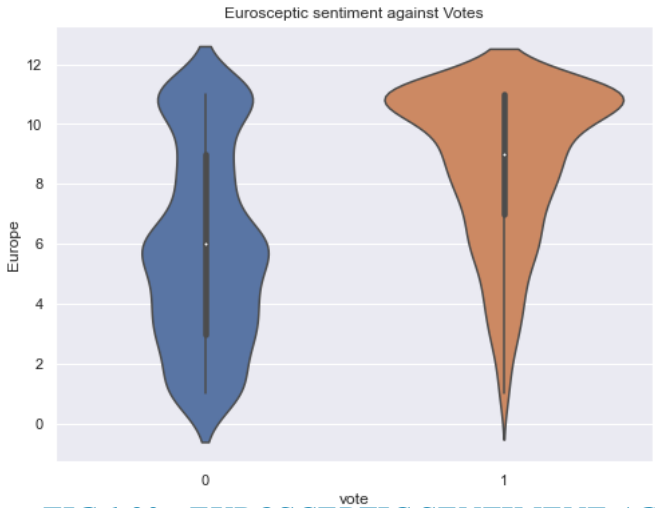


FIG 1.23 : EUROSCEPTIC SENTIMENT AGAINST VOTING

- Out of 338 people who gave a score of 11, 166 peoplehave voted for the labour party and 172 people have voted for the conservative party.
- People who gave score of 7 to 10 have voted for labourand conservative almost equally. Conservative party seem to be slightly higher in these instances.
- Out of 207 people who gave a score of 6, 172 people have voted for the labour party and 35 people have voted forthe conservative party.
- People who gave a score of 1 to 6 have predominantly voted for the labour party. As we can see, there are a totalof 770 people who have given scores from 1 to 6. Out of 770 people, 672 people have voted for the labour party. So, 87.28% of the people have chosen labour party.
- So, we can infer that lower the 'Eurosceptic' sentiment, higher the votes for labour party.

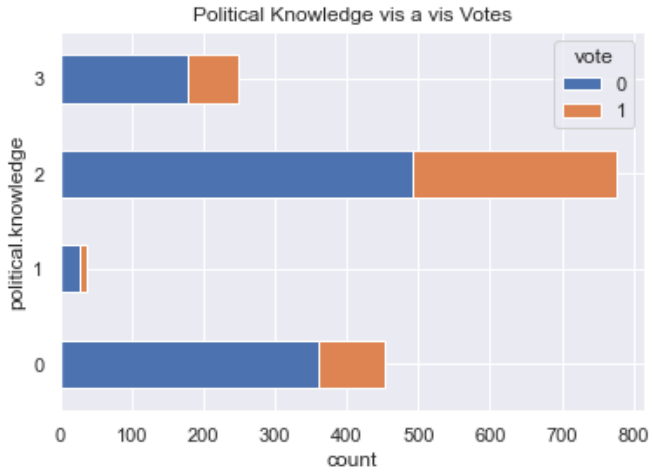


FIG 1.23 : POLITICAL KNOWLEDGE VS VOTING

- Out of 249 people who gave a score of 3, 177 people have voted for the labour party and 72 people have voted forthe conservative party.
- Out of 776 people who gave a score of 2, 493 people have voted for the labour party and 283 people have voted for the conservative party.
- Out of 38 people who gave a score of 1, 27 people have voted for the labour party and 11 people have voted for the conservative party.
- Out of 454 people who gave a score of 0, 360 people have voted for the labour party and 94 people have voted forthe conservative party.
- We can see that, in all instances, labour party gets the higher number of votes.
- Out of 1517 people, 454 people gave a score of 0. So, this means that, 29.93% of the people are casting their votes without any political knowledge.

FIG 1.24 : HEATMAP

- We can see that, mostly there is no correlation in thedataset through this matrix. There are some variables that are moderately positively correlated and some thatare slightly negatively correlated.
- ‘economic.cond.national’ with ‘economic.cond.household’ have moderate positive correlation.
- ‘Blair' with 'economic.cond.national' and ‘economic.cond.household’ have moderate positive correlation.
- ‘Europe’ with ‘Hague’ have moderate positive correlation.
- 'Hague' with 'economic.cond.national' and 'Blair' have
- moderate negative correlation.
- 'Europe' with 'economic.cond.national' and 'Blair' have moderate negative correlation.

DATA PRE-PROCESSING

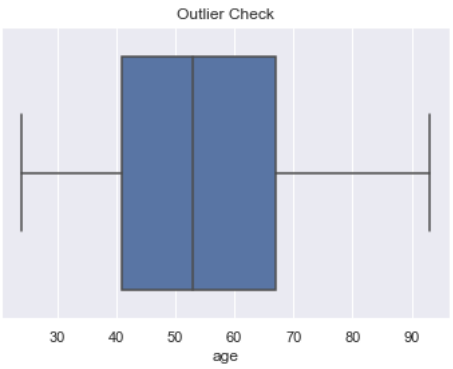


FIG 1.24 : OUTLINER CHECK

We were only interested in the outliers for the age variable since other numerical features are ordinal in nature.

There are no outliers in the age variable.

Here, only gender variable needs to be encoded since all the other independent features are numerical in nature, so we opt for One hot encoding with dropping the dummy variable.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender_1
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

FIG 1.25 : ENCODED TABLE

We have to scale the numerical variables as distance based algorithms like KNN will give highly inaccurate results.

	range	std
age	69.0	15.70
economic.cond.national	4.0	0.88
economic.cond.household	4.0	0.93
Blair	4.0	1.17
Hague	4.0	1.23
Europe	10.0	3.30
political.knowledge	3.0	1.08

FIG 1.26 : NUMERICAL DESCRIPTION

- The dataset contains features highly varying in magnitudes, units and range between the 'age' column and other columns.

- But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem.
- If left alone, these algorithms only take in the magnitude of features neglecting the units.
- The results would vary greatly between different units, 1km and 1000 metres.
- The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.
- To supress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.
- in this case, we have a lot of encoded, ordinal, categorical and continuous variables. So, we use the minmaxscaler technique to scale the data.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender_1
0	0	0.275362	0.50	0.50	0.75	0.00	0.1	0.666667	0
1	0	0.173913	0.75	0.75	0.75	0.75	0.4	0.666667	1
2	0	0.159420	0.75	0.75	1.00	0.25	0.2	0.666667	1
3	0	0.000000	0.75	0.25	0.25	0.00	0.3	0.000000	0
4	0	0.246377	0.25	0.25	0.00	0.00	0.5	0.666667	1

FIG 1.27 : SCALED TABLE

Our model will use all the variables and 'vote_Labour' is the target variable. The train-test split is a technique for evaluating the performance of a machine learning algorithm. The procedure involves taking a dataset and dividing it into two subsets.

• Train Dataset: Used to fit the machine learning model.
• Test Dataset: Used to evaluate the fit machine learning model.

The data is divided into 2 subsets, training and testing set. Earlier, we have extracted the target variable ‘vote_Labour’ in a separate vector for subsets. Random state chosen as 1.

- **Training Set:** 70percent of data.
- **Testing Set:** 30 percent of the data.

```
x_train.shape
```
```
(1061, 8)
```
```
y_train.shape
```
```
(1061,)
```
```
x_test.shape
```
```
(456, 8)
```
```
y_test.shape
```
```
(456,)
```

FIG 1.28 : TRAIN AND TEST SETS

MODEL BUILDING AND PERFORMANCE EVALUATION

APPLY LOGISTIC REGRESSION:

	precision	recall	f1-score	support
0	0.86	0.92	0.89	754
1	0.76	0.63	0.69	307
accuracy			0.83	1061
macro avg	0.81	0.77	0.79	1061
weighted avg	0.83	0.83	0.83	1061

FIG 1.29 : CLASSIFICATION REPORT FOR LOGISTIC

	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

FIG 1.30 : CLASSIFICATION REPORT FOR LOGISTIC REGRESSION TEST SET

There are no outliers present in the continuous variable 'age'. The remaining variables are categorical in nature. Our model will use all the variables and 'vote_Labour' is the target variable.

Since the data set is skewed and both type I & type II errors are costly - we look at the f1 score for both the classes along with the accuracy to adjudge the model performance.

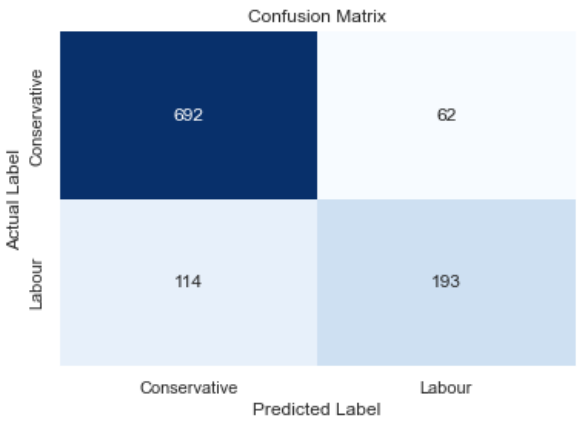


FIG 1.31 : CONFUSION MATRIX OF TRAIN DATA

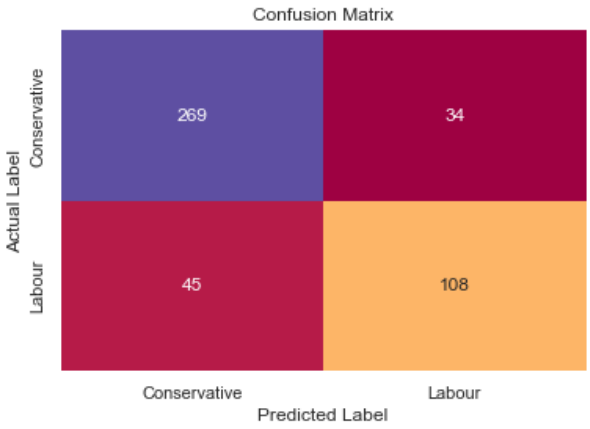


FIG 1.32 : CONFUSION MATRIX OF TEST DATA

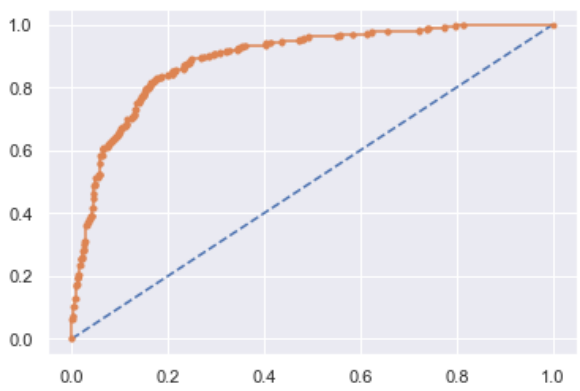


FIG 1.33 : ROC FOR TRAIN DATA

AUC : 0.890
Training Accuracy: 0.8341188
Mean square error: 0.16588124410933083

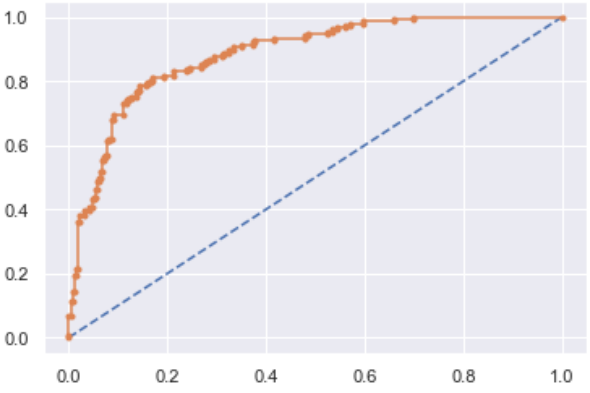


FIG 1.34 : ROC FOR TEST DATA

AUC: 0.884
Testing Accuracy: 0.8267544
Mean square error: 0.17324561403508773

data:

- Accuracy: 83.41 %
- Precision: 76%
- Recall: 63%
- F1-Score: 69%

Test data:

- Accuracy: 82.68%
- Precision: 76%
- Recall: 71%
- F1-Score: 73%

Validity of the model:

The model is not over-fitted or under-fitted.
The error in the test data is slightly higher than the train data, which is absolutely fine because t he error margin is low and the error in both train and test data is not too high. Thus, the model is not over-fitted or under-fitted.

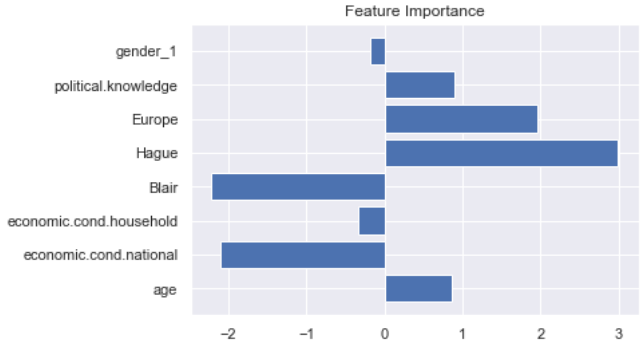


FIG 1.35 : FEATURE IMPORTANCE

Since the data is scaled, the coefficients of the Logit Function will give the feature importance.

1. We can see the Precision, Recall, Accuracy & AUC of training data for the model is inline with the testing data and is fairly high. Hence, no overfitting or underfitting has occured & the model can be used for making predictions.
2. The model is predicting better for the majority class and has a pretty inferior performance for the minority class.

APPLY LDA (LINEAR DISCRIMINANT ANALYSIS):

```
[[685  69]
 [107 200]]
              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
```

FIG 1.36 : CLASSIFICATION REPORT FOR LDA TRAIN SET

```
[[269  34]
 [ 42 111]]
              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
```

FIG 1.37 : CLASSIFICATION REPORT FOR LDA TEST SET

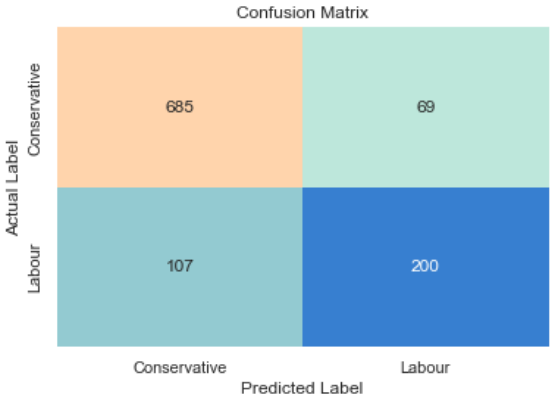


FIG 1.38 : CONFUSION MATRIX OF TRAIN DATA

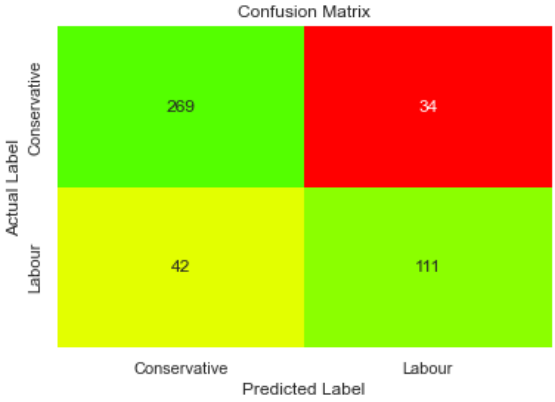


FIG 1.39 : CONFUSION MATRIX OF TEST DATA

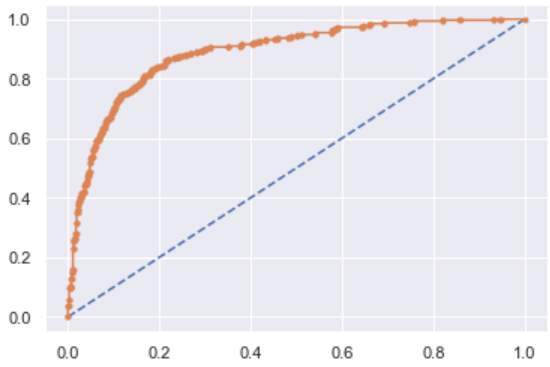


FIG 1.40 : ROC FOR TRAIN DATA

AUC : 0.889
Training Accuracy: 0.8341188

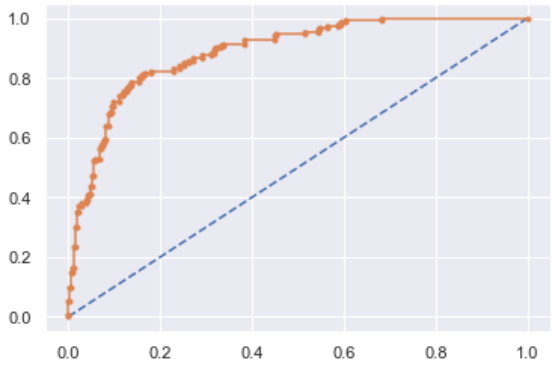


FIG 1.41 : ROC FOR TEST DATA

AUC: 0.888
Testing Accuracy: 0.8333333

data:

- Accuracy: 83.41 %
- Precision: 74%
- Recall: 65%
- F1-Score: 69%

Test data:

- Accuracy: 82.68%
- Precision: 77%
- Recall: 73%
- F1-Score: 74%

Validity of the model:

- There are no outliers present in the continuous variable 'age'.The remaining variables ar e categorical in nature. Our modelwill use all the variables and 'vote_Labour' is the targe t variable.
- The model is not over-fitted or under-fitted.
- The error in the test data is slightly higher than the train data, which is absolutely fine be cause the error margin is low and the error in both train and test data is not too high. Th us, the model is not over-fitted or under-fitted.

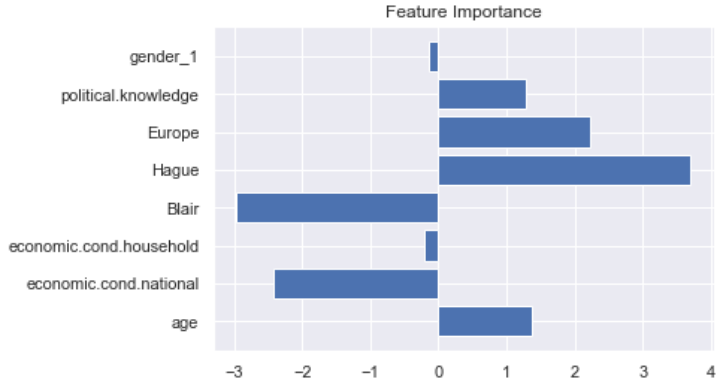


FIG 1.44 : FEATURE IMPORTANCE(LDA)

The model is similar to Logistic Regression model & is predicting better for the majority class and has a pretty inferior performance for the minority class.

APPLY KNN MODEL:

	precision	recall	f1-score	support
0	0.89	0.91	0.90	754
1	0.77	0.72	0.74	307
accuracy			0.86	1061
macro avg	0.83	0.82	0.82	1061
weighted avg	0.85	0.86	0.85	1061

FIG 1.45 : CLASSIFICATION REPORT FOR KNN

	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

FIG 1.46 : CLASSIFICATION REPORT FOR KNN TEST SET

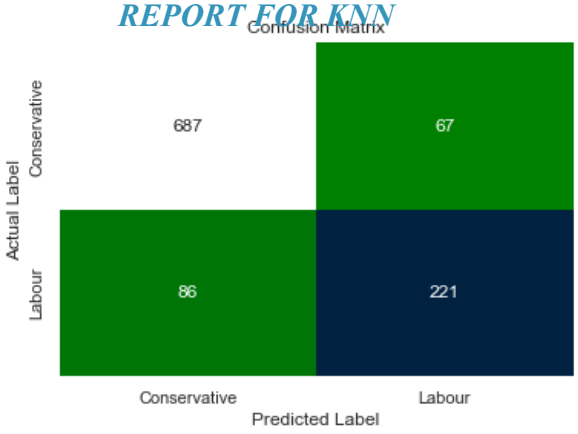


FIG 1.47 : CONFUSION MATRIX OF TRAIN DATA

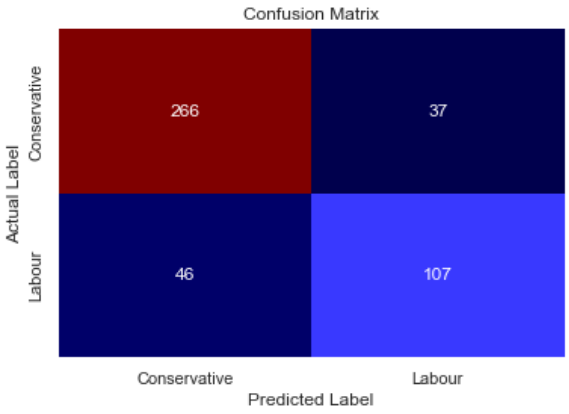


FIG 1.48 : CONFUSION MATRIX OF TEST DATA

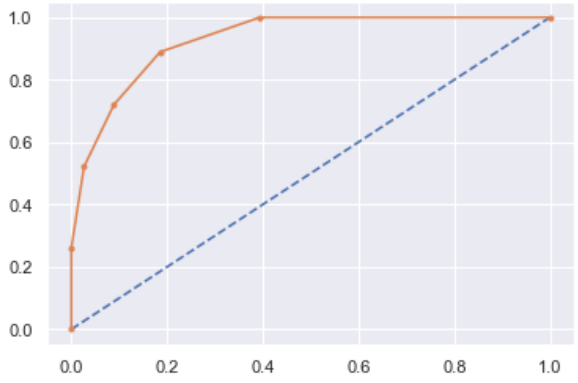


FIG 1.49 : ROC FOR TRAIN DATA

AUC : 0.930
Training Accuracy: 0.8557964

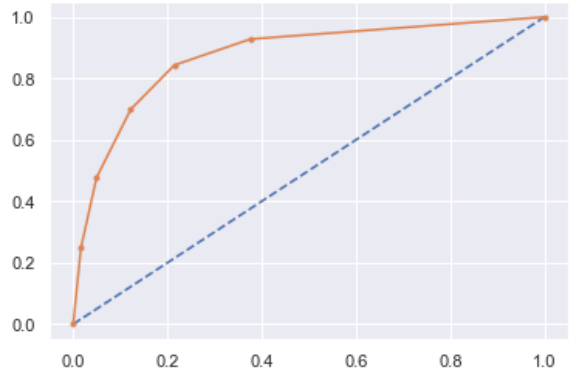


FIG 1.50 : ROC FOR TEST DATA

AUC: 0.873
Testing Accuracy: 0.8179825

data:

- Accuracy: 83.41 %
- Precision: 77%
- Recall: 72%
- F1-Score: 74%

Test data:

- Accuracy: 82.68%
- Precision: 74%
- Recall: 70%
- F1-Score: 72%

Validity of the model:

- There are no outliers present in the continuous variable 'age'. The remaining variables ar e categorical in nature. Our model will use all the variables and 'vote_Labour' is the targ et variable. We take K value as 5
- The model is not over-fitted.
- As we can see, the train data has a 83% accuracy and test data has 82% accuracy. The difference is less than 2%. Hence the model is not overfitting.

APPLY NAIVE BAYES:

	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

FIG 1.51 : CLASSIFICATION REPORT FOR NAÏVE BAYES TRAIN SET

	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

FIG 1.52 : CLASSIFICATION REPORT FOR NAÏVE BAYES TEST SET

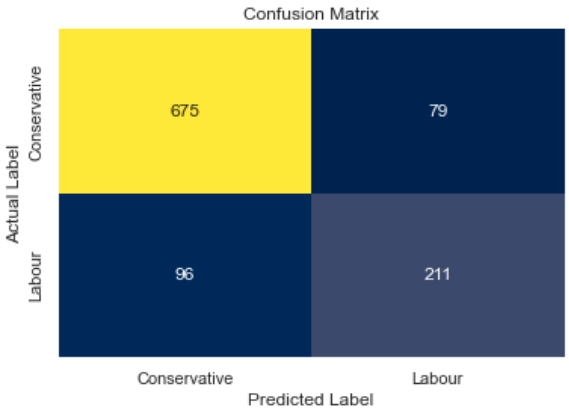


FIG 1.53 : CONFUSION MATRIX OF TRAIN DATA

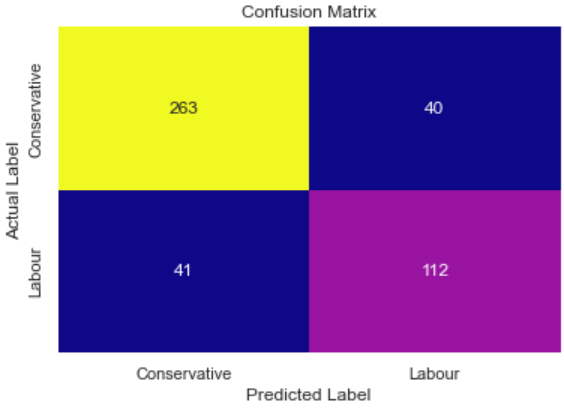


FIG 1.54 : CONFUSION MATRIX OF TEST DATA

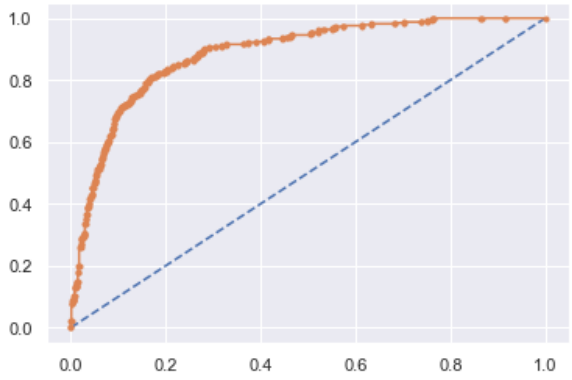


FIG 1.55 : ROC FOR TRAIN DATA

AUC : 0.886
Training Accuracy: 0.8350613

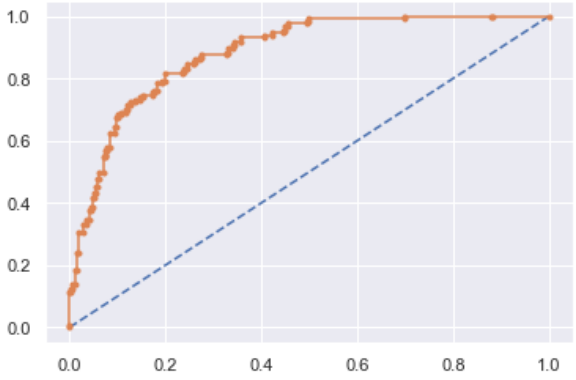


FIG 1.56 : ROC FOR TEST DATA

AUC: 0.885
Testing Accuracy: 0.8223684

data:

- Accuracy: 83.5 %
- Precision: 73%
- Recall: 69%
- F1-Score: 71%

Test data:

- Accuracy: 82.24%
- Precision: 74%
- Recall: 73%
- F1-Score: 73%

Validity of the model:

- There are no outliers present in the continuous variable 'age'. The remaining variables are categorical in nature. Our model will use all the variables and 'vote_Labour' is the target variable.
- The model is not over-fitted or under-fitted.
- The error in the test data is slightly higher than the train data, which is absolutely fine because the error margin is low and the error in both train and test data is not too high. Thus, the model is not over-fitted or under-fitted.

APPLY BOOSTING:

	precision	recall	f1-score	support
0	0.85	0.93	0.89	754
1	0.78	0.61	0.68	307
accuracy			0.84	1061
macro avg	0.82	0.77	0.79	1061
weighted avg	0.83	0.84	0.83	1061

FIG 1.57 : CLASSIFICATION REPORT FOR BOOSTING TRAIN SET

	precision	recall	f1-score	support
0	0.83	0.89	0.86	303
1	0.75	0.64	0.69	153
accuracy			0.81	456
macro avg	0.79	0.77	0.78	456
weighted avg	0.81	0.81	0.80	456

FIG 1.58 : CLASSIFICATION REPORT FOR BOOSTING TEST SET

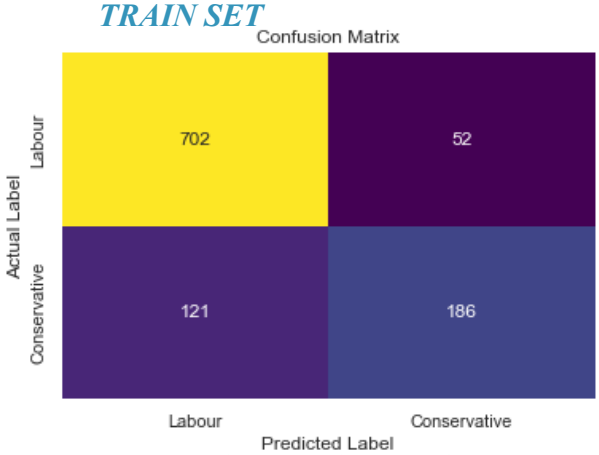


FIG 1.59 : CONFUSION MATRIX OF TRAIN DATA

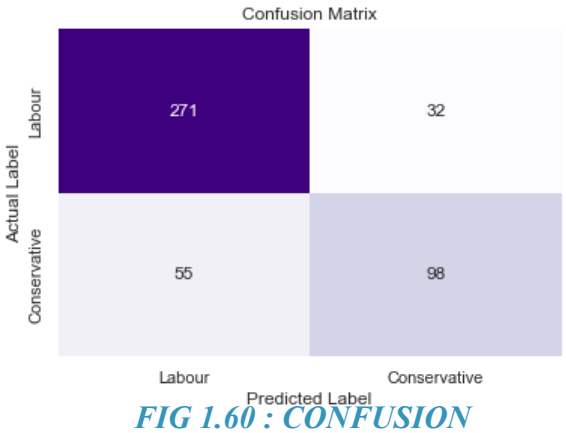


FIG 1.60 : CONFUSION MATRIX OF TEST DATA

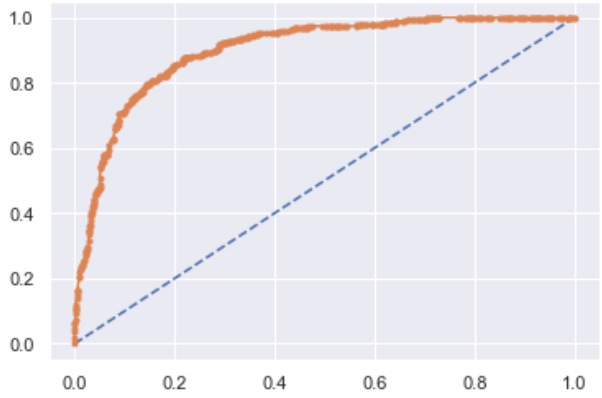


FIG 1.61 : ROC FOR TRAIN DATA

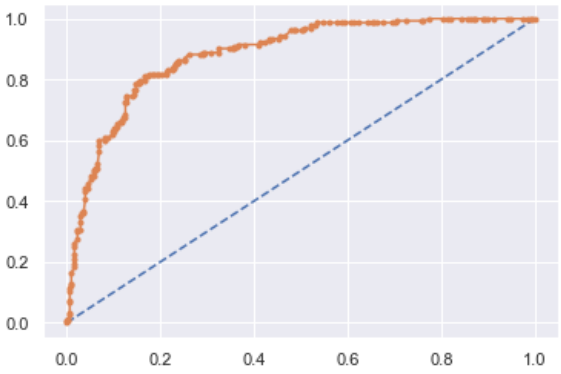


FIG 1.62 : ROC FOR TEST DATA

AUC : 0.886
Training Accuracy: 0.8369463
Model score:0.8369462770970783

AUC: 0.885
Testing Accuracy: 0.8092105
Model score: 0.8092105263157895

data:

- Accuracy: 83.69 %
- Precision: 78%
- Recall: 61%
- F1-Score: 68%

Test data:

- Accuracy: 80.92%
- Precision: 75%
- Recall: 64%
- F1-Score: 69%

Validity of the model:

We can see, The Precision, Recall, Accuracy & AUC of training data for the model is inline with the testing data and is fairly high. Hence, no overfitting or underfitting has occured & the model can be used for making predictions.

APPLY MODEL TUNING TO LOGISTIC REGRESSION:

We have observed that the performance for minority class is less. Model Tuning has a scoring parameter that may help us in forming a model that shows a better f1 - score for the minority class if we make it our class of interest. Hence, we first need to custom encode the classes to make the minority class as our class of interest (Conservative - 1, Labour - 0)

	precision	recall	f1-score	support
0	0.86	0.92	0.89	754
1	0.76	0.63	0.69	307
accuracy			0.84	1061
macro avg	0.81	0.77	0.79	1061
weighted avg	0.83	0.84	0.83	1061

FIG 1.63 : CLASSIFICATION REPORT FOR TRAIN SET

	precision	recall	f1-score	support
0	0.86	0.89	0.87	303
1	0.76	0.71	0.74	153
accuracy			0.83	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.83	0.83	0.83	456

FIG 1.64 : CLASSIFICATION REPORT FOR TEST SET

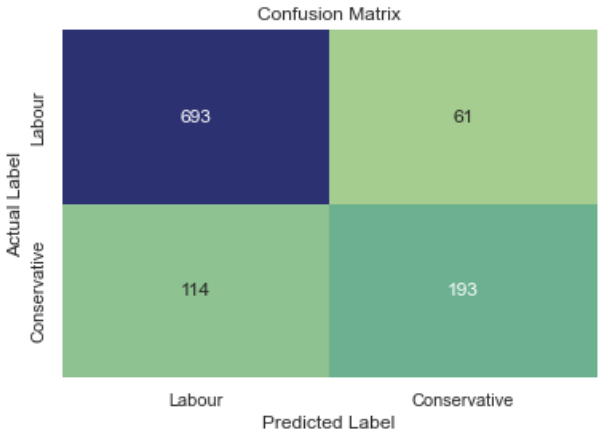


FIG 1.65 : CONFUSION MATRIX OF TRAIN DATA

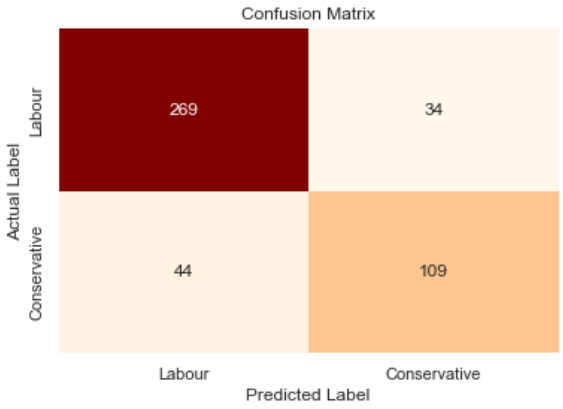


FIG 1.66 : CONFUSION MATRIX OF TEST DATA

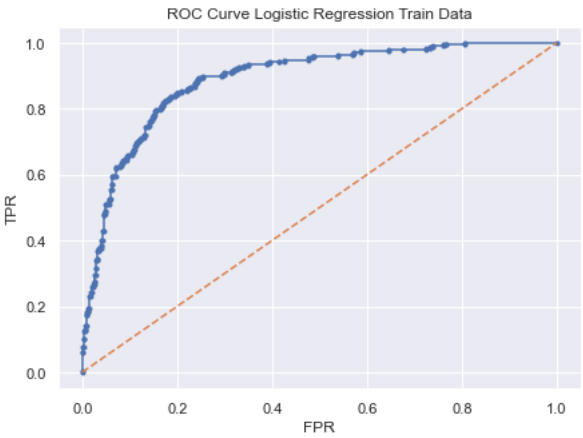


FIG 1.67 : ROC FOR TRAIN DATA

AUC : 0.89.
Training Accuracy: 0.8350613

FIG 1.68 : ROC FOR TEST DATA

AUC: 0.88
Testing Accuracy: 0.8289474

data:

- Accuracy: 83.5 %
- Precision: 76%
- Recall: 63%
- F1-Score: 69%

Test data:

- Accuracy: 82.89%
- Precision: 76%
- Recall: 71%
- F1-Score: 74%

Validity of the model:

1. No significant impact on the model even after tuning it based on the learning rate & the solver.
2. The Precision, Recall, Accuracy & AUC of training data for the model is inline with the testing data and is fairly high. Hence, no overfitting or underfitting has occured & the model can be used for making predictions.
3. The model as before is predicting better for the majority class and has a pretty inferior performance for the minority class.

APPLY MODEL TUNING TO LDA:

	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

FIG 1.69 : CLASSIFICATION REPORT FOR TRAIN SET

	precision	recall	f1-score	support
0	0.86	0.89	0.87	303
1	0.76	0.71	0.74	153
accuracy			0.83	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.83	0.83	0.83	456

FIG 1.70 : CLASSIFICATION REPORT FOR TEST SET

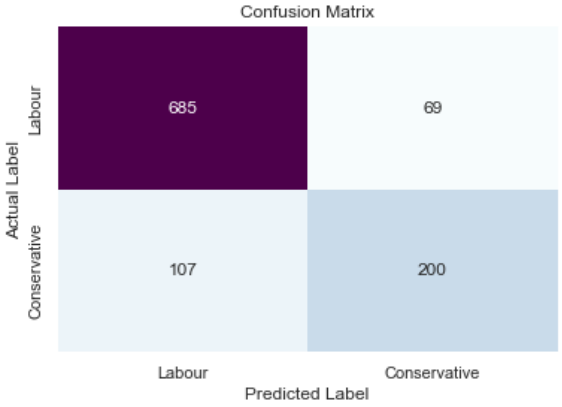


FIG 1.71 : CONFUSION MATRIX OF TRAIN DATA

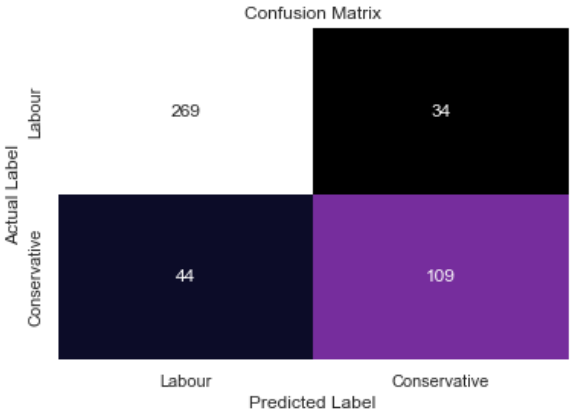


FIG 1.72 : CONFUSION MATRIX OF TEST DATA

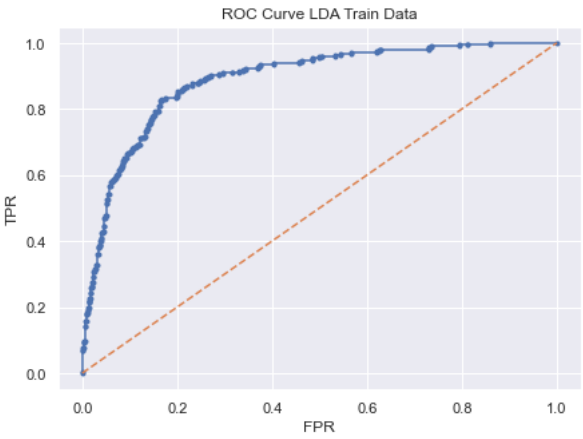


FIG 1.73 : ROC FOR TRAIN DATA

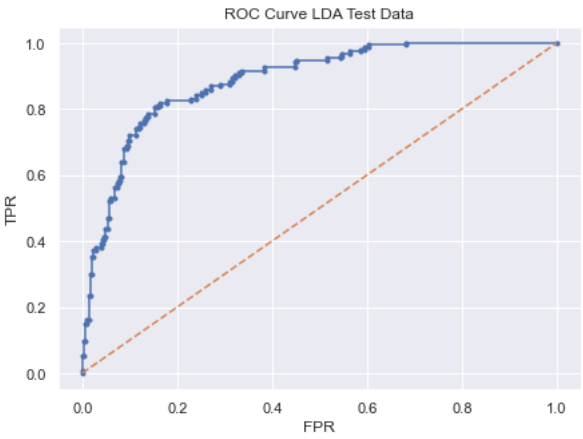


FIG 1.74 : ROC FOR TEST DATA

AUC : 0.89.
Training Accuracy: 0.8341188

AUC: 0.88
Testing Accuracy: 0.8289474

data:

- Accuracy: 83.5 %
- Precision: 74%
- Recall: 65%

- F1-Score: 69%

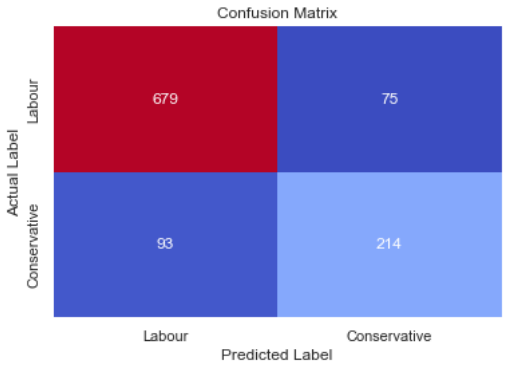
Test data:

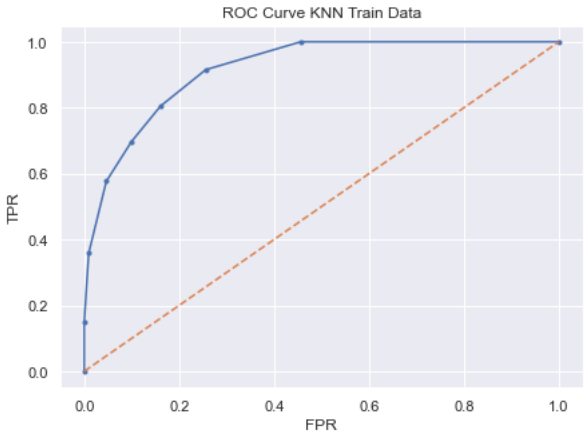
- Accuracy: 82.89%
- Precision: 76%
- Recall: 71%
- F1-Score: 74%

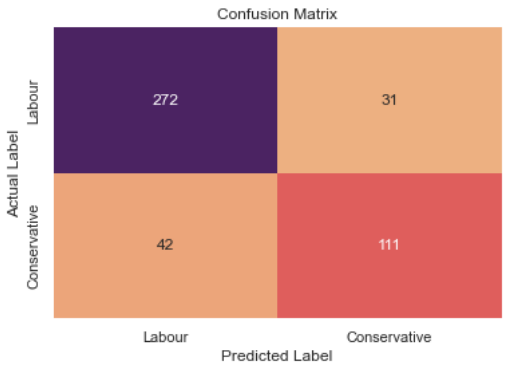
Validity of the model:

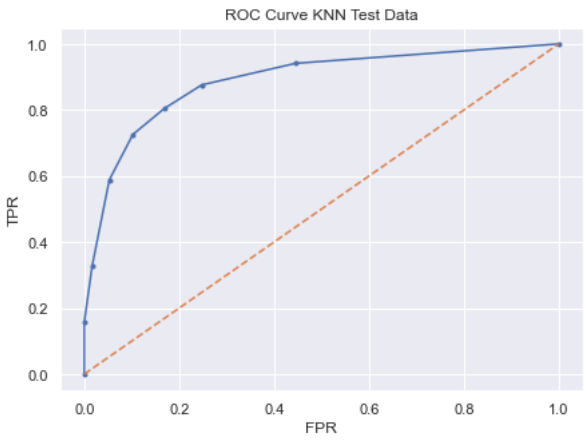
- The model is not over-fitted or under-fitted.
- The error in the test data is slightly higher than the train data, which is absolutely fine because the error margin is low and the error in both train and test data is not too high. Thus, the model is not over-fitted or under-fitted

APPLY MODEL TUNING TO KNN:









```
Training Accuracy: 0.8416588
Testing Accuracy: 0.8399123
```

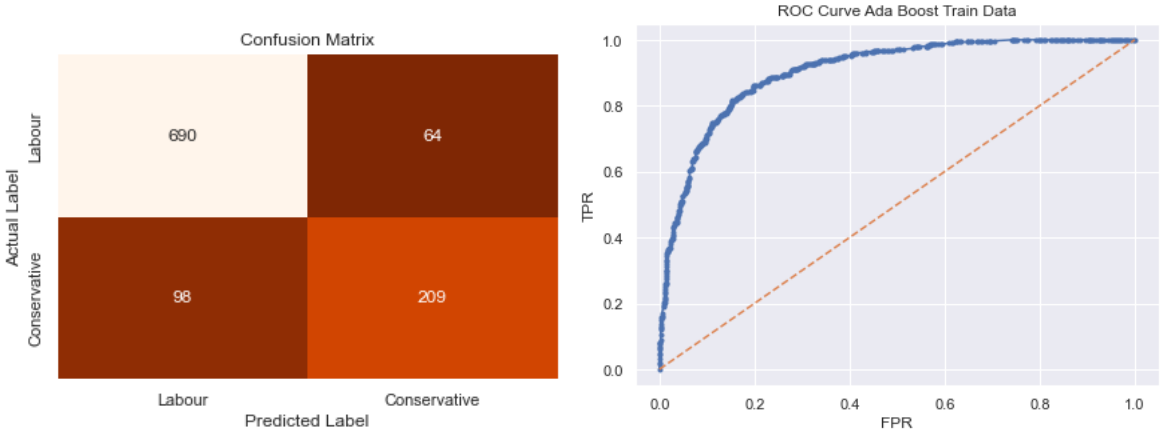
data:

- Accuracy: 84.16 %
- Precision: 74%
- Recall: 70%
- F1-Score: 72%

Test data:

- Accuracy: 82.89%
- Precision: 78%
- Recall: 73%
- F1-Score: 75%

APPLY BOOSTING (ADABOOST):



data:

- Accuracy: 85%
- Precision: 77%
- Recall: 68%
- F1-Score: 72%
- Auc:0.91

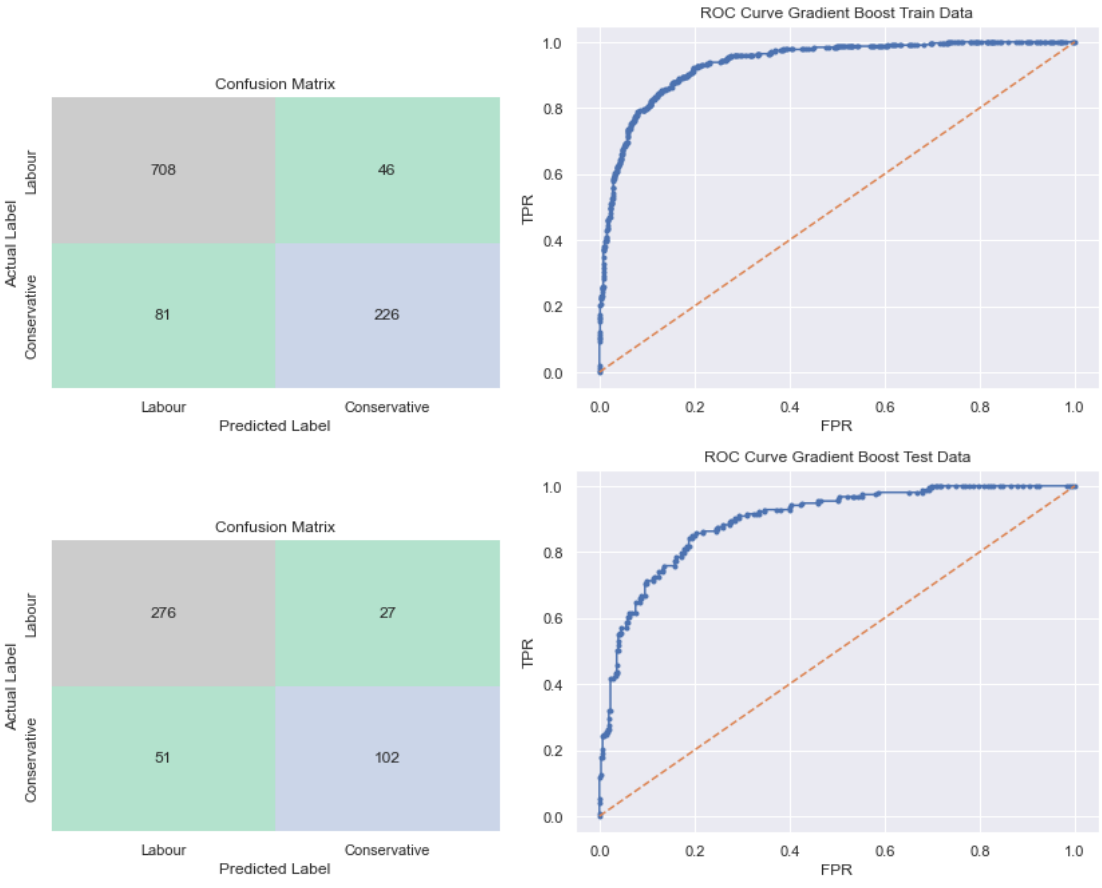
Test data:

- Accuracy: 82%
- Precision: 75%
- Recall: 71%
- F1-Score: 73%
- Auc:0.88

Validity of the model:

We can see, The Precision, Recall, Accuracy & AUC of training data for the model is inline with the testing data and is fairly high. Hence, no overfitting or underfitting has occured & the model can be used for making predictions.

APPLY BOOSTING (GRADIENT BOOST):



data:

- Accuracy: 88%
- Precision: 83%
- Recall: 74%
- F1-Score: 78%
- Auc: 0.94

Test data:

- Accuracy: 83%
- Precision: 79%
- Recall: 67%
- F1-Score: 72%
- Auc: 0.9

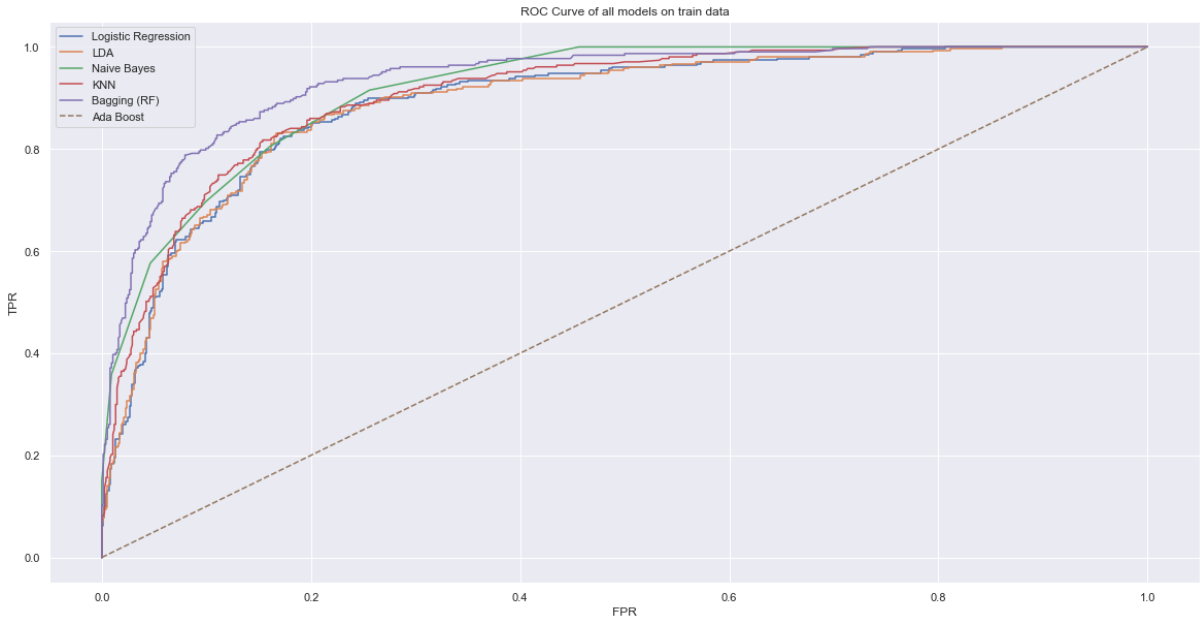
Validity of the model:

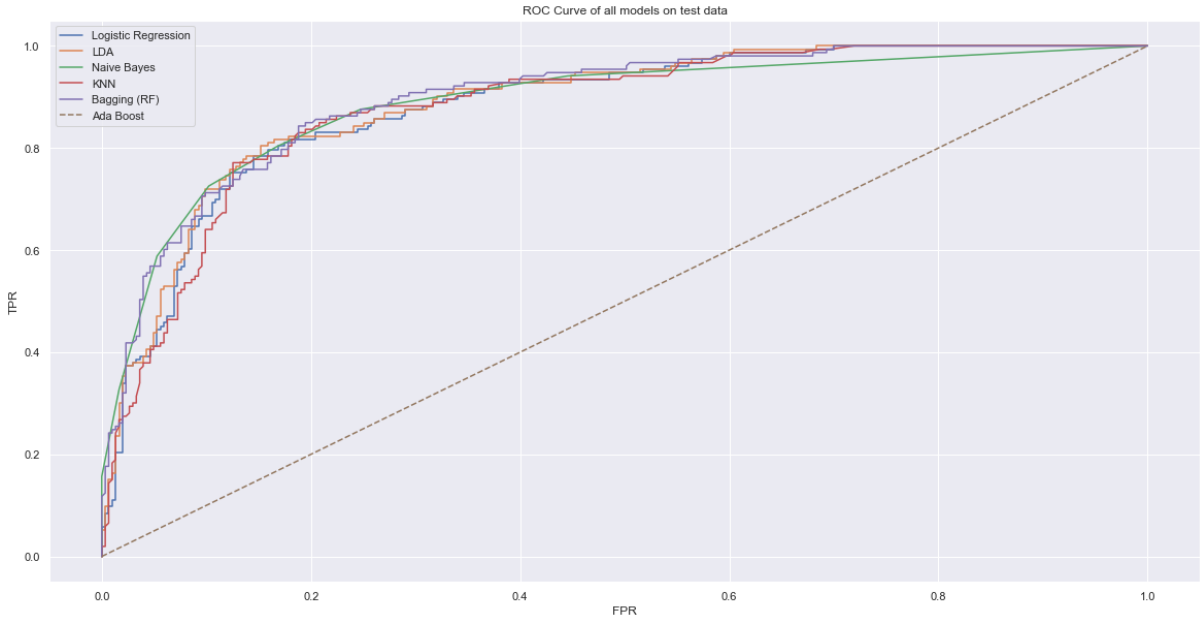
Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

Model performance of each model has already been observed.

Since Logistic Regression & LDA model showed similar performace even after tuning with the solver & learning rate, we can use any of the tuned or untuned model for comparison with other models. Here, we use the tuned model for comparison with other models.

MODEL PERFORMANCE IMPROVEMENT





COMPARISON BETWEEN THE REGULAR MODEL AND TUNED MODEL(LR)

• As we can see, there is not much difference between the performance of regular LR model and tuned LR model.

• The values are high overall and there is no over-fitting orunder- fitting. Therefore both models are equally good models.

COMPARISON BETWEEN THE REGULAR MODEL AND TUNED MODEL (LDA)

- As we can see, there is not much difference between the performance of regular LDA model and tuned LDA model.
- The values are high overall and there is no over-fitting orunder- fitting.
- Therefore both models are equally good models.

COMPARISON BETWEEN THE REGULAR MODEL AND TUNED MODEL (KNN)

- There is no over-fitting or under-fitting in the tuned KNNmodel. Overall, it is a good model.
- Comparison between the regular KNN model and tunedKNN model:
- As we can see, the regular KNN model was over-fitted.But model tuning has helped the model to recover from over-fitting.
- The values are better in the tuned KNN model.
- Therefore, the tuned KNN model is a better model.

- In all the models, tuned ones are better than the regular models. So, we compare only the tuned models and describe which model is the best/optimized.
- All the tuned models have high values and every model is good. But as we can see, the most consistent tuned model in both train and test data is the Gradient Boost model.
- The tuned gradient boost model performs the best with 88.31% accuracy score in train and 87.28% accuracy score in test. Also it has the best AUC score of 94% inboth train and test data which is the highest of all the models.
- It also has a precision score of 88% and recall of 94% which is also the highest of all the models. So, we conclude that Gradient Boost Tuned model is the best/optimized model.

INSIGHTS:

- Labour party has more than double the votes of conservative party.
- Most number of people have given a score of 3 and 4 for the national economic condition and the average score is 3.245221
- Most number of people have given a score of 3 and 4 for the household economic condition and the average scoreis 3.137772
- Blair has higher number of votes than Hague and the scores are much better for Blair than for Hague.
- The average score of Blair is 3.335531 and the average score of Hague is 2.749506. So, here we can see that,Blair has a better score.
- On a scale of 0 to 3, about 30% of the total population has zero knowledge about politics/parties.
- People who gave a low score of 1 to a certain party, still decided to vote for the same party instead of voting for the other party. This can be because of lack of political knowledge among the people.
- People who have higher Eurosceptic sentiment, has voted for the conservative party and lower the Eurosceptic sentiment, higher the votes for Labour party.
- Out of 454 people who gave a score of 0 for political knowledge, 360 people have voted for the labour partyand 94 people have voted for the conservative party.
- All models performed well on training data set as well as test dat set. The tuned models have performed better than the regular models.
- There is no over-fitting in any model except Random Forest and Bagging regular models.
- Gradient Boosting model tuned is the best/optimized model.

BUSINESS RECOMMENDATIONS:

- Hyper-parameters tuning is an import aspect of modelbuilding. There are limitations to this as to process these combinations, huge amount of processing power is required. But if tuning can be done with many sets ofparameters, we might get even better results.
- Gathering more data will also help in training the models and thus improving the predictive powers.

- We can also create a function in which all the modelspredict the outcome in sequence. This will helps in better understanding and the probability of what the outcome will be.
- Using Gradient Boosting model without scaling for predicting the outcome as it has the best optimized performance.

PROBLEM 2

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

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

Code Snippet to extract the three speeches:

"

```
import nltk
nltk.download('inaugural')
from nltk.corpus import inaugural
inaugural.fileids()
inaugural.raw('1941-Roosevelt.txt')
inaugural.raw('1961-Kennedy.txt')
inaugural.raw('1973-Nixon.txt')
```

"

DEFINE THE PROBLEM AND PERFORM EXPLORATORY DATA ANALYSIS

- President Franklin D. Roosevelt's speech have 7571 characters (including spaces).
- President John F. Kennedy's speech have 7618 characters (including spaces).
- President Richard Nixon's speech have 9991 characters (including spaces).
- There are 1526 words in President Franklin D. Roosevelt's speech.
- There are 1543 words in President John F. Kennedy's speech.
- There are 2006 words in President Richard Nixon's

  speech.

- There are 68 sentences in President Franklin D. Roosevelt's speech.
- There are 52 sentences in President John F. Kennedy's speech.
- There are 68 sentences in President Richard Nixon's

  speech.

1. Before, removing the stop-words, we have changed all the letters to lowercase and we have removed special characters.
2. Before the removal of stop-words,
    1. President Franklin D. Roosevelt's speech have 1334 words.
    2. President John F. Kennedy's speech have 1362 words.
    3. President Richard Nixon's speech have 1800 words.
3. After the removal of stop-words,
4. President Franklin D. Roosevelt's speech have 623 words.
5. President John F. Kennedy's speech have 693 words.
6. President Richard Nixon's speech have 831 words.

TEXT CLEANING

TOP 3 WORDS IN ROOSEVELT'S SPEECH:

- nation - 11
- know - 10
- spirit - 9

TOP 3 WORDS IN KENNEDY'S SPEECH:

- let-11
- us-10
- sides - 9

TOP 3 WORDS IN NIXON'S SPEECH:

- us-26
- let-22
- peace - 19

PLOT WORD CLOUD OF ALL THREE SPEECHES

Word cloud of Roosevelt's speech:

Word cloud of Kennedy's speech:

Word cloud of Nixon's speech: