Machine Learning Business Report

By: Ambrish Verma

Date: 9-Dec-2023

Contents

Problem 1: Election Result Prediction

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

Dataset for Problem: Election_Data.xlsx

Data Ingestion: 11 marks
1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks)
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)
Data Preparation: 4 marks
1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). (4 Marks)
Modeling: 22 marks
1.4 Apply Logistic Regression and LDA (linear discriminant analysis). (4 marks)
1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results. (4 marks)
1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging) and boosting. (7 marks)
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized. (7 marks)
1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.

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

- The data file provided consists of 1525 entries and has a total of 10 attributes.
- Following are the details of the attributes present in the dataset:

```
#   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Unnamed: 0               1525 non-null   int64
 1   vote                     1525 non-null   object
 2   age                      1525 non-null   int64
 3   economic.cond.national   1525 non-null   int64
 4   economic.cond.household  1525 non-null   int64
 5   Blair                    1525 non-null   int64
 6   Hague                    1525 non-null   int64
 7   Europe                   1525 non-null   int64
 8   political.knowledge      1525 non-null   int64
 9   gender                   1525 non-null   object
```

- The data provided is in random order, there was no sequence observed in the data.
- Minimum age of the persons surveyed was 24 years, Maximum age was 93 years, the mean age was 54.18 years, and the median age was 53 years. Standard deviation was 15.71 years.
- None of the attributes had any null value present.
- Attributes: ‘Vote’ and ‘Gender’ are of object datatype but store categorical variables. All other attributes except ‘age’ and the ‘unnamed:0’ also have categorical variables.

- The survey has a fairly equitable distribution of males (713) and females (812).

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

Univariate Analysis:

- Before performing the univariate analysis, the following changes were performed which make the report more readable and makes it easier for further analyses.
    - The attribute: 'Europe' was renamed to 'Eurosceptic' to denote the sentiment against the European Union.
    - The values of atribute 'gender' was changed. The value 'male' has been converted to 0 and 'female' has been converted to 1.
    - The values of attribute 'vote' was changed. The value 'Labour' has been converted to 0 and 'Conservative' has been converted to 1.
    - The attribute: 'unnamed: 0' was dropped from the dataset as it was not making any contribution to the analysis.
- For univariate analysis, the attributes were charted on countplots. The plots are displayed below in figure: Univariate Analyses - Countplots

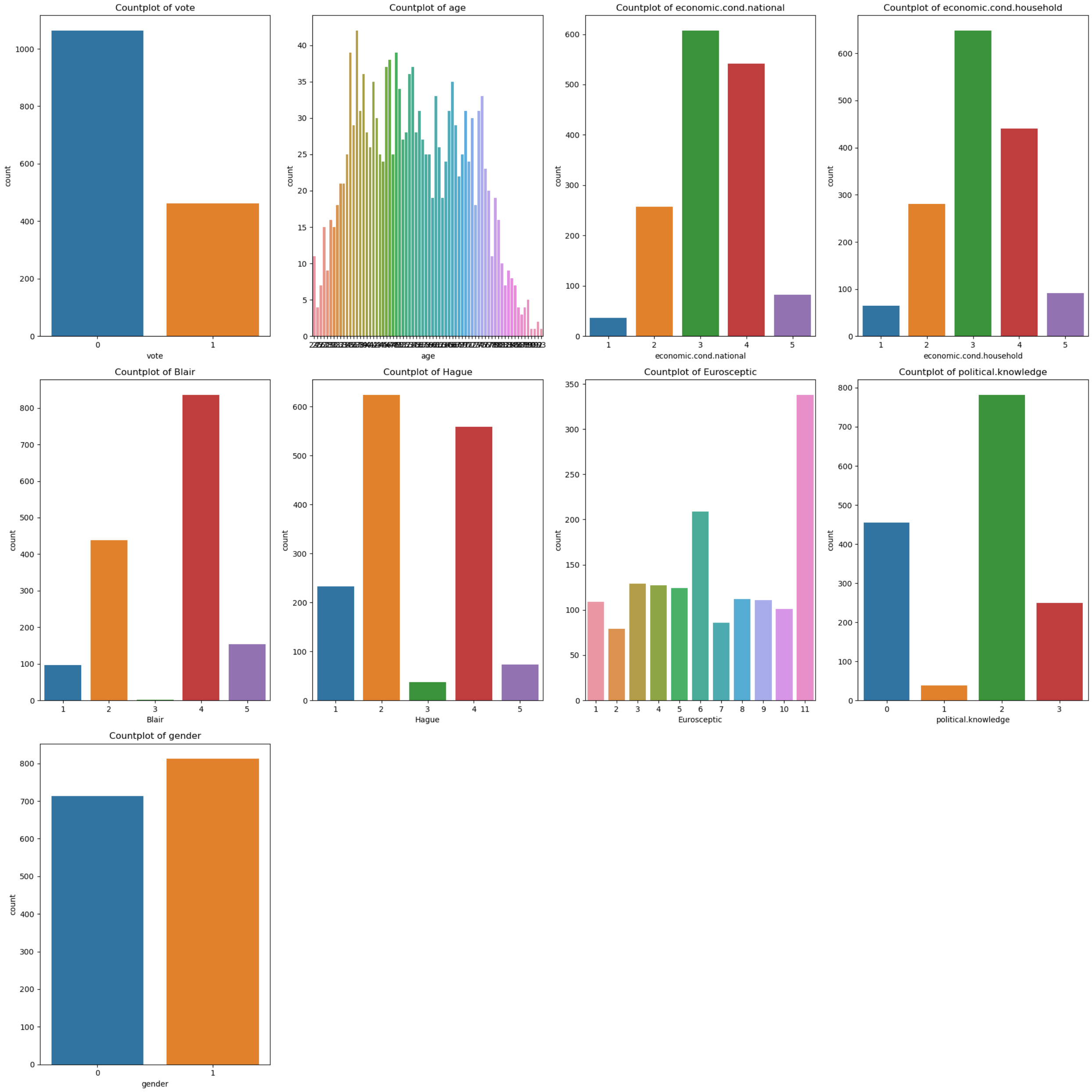

Figure 1: Univariate Analyses - Countplots

- Analysis based on the above figure:
    - ‘Labour’ party has polled votes than ‘Conservatives’
    - The survey is skewed towards the younger voters
    - Hague has a better approval rating than Blair
    - Measure of extreme Eurosceptic sentiments is much higher than the other extreme.
- To check the presence of outliers, boxplots for the attributes are charted below in the figure2: Univariate analyses – Boxplot

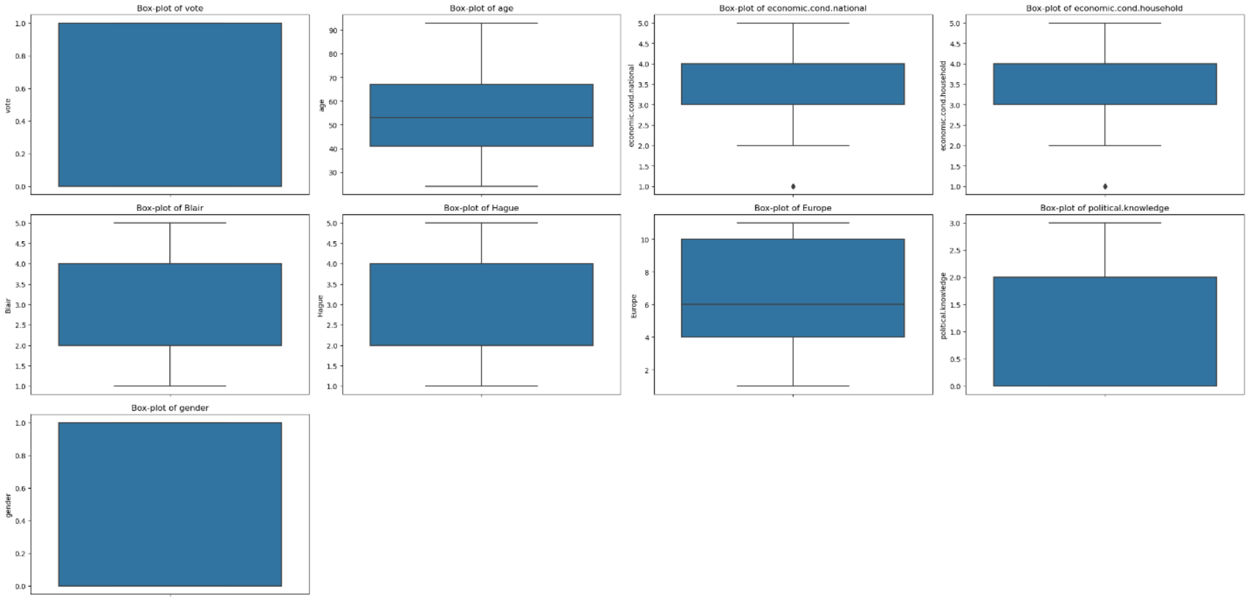

Figure 2: Univariate Analyses - Boxplots

- While in the boxplot, the number of people surveyed belong to lower category of Economic and household conditions has been shown as outliers, but they cannot be treated as such. This is because they are categorical variables.
- The age variable, which is continuous, does not have any outlier.

Bivariate Analysis:

- Heatmap is displayed below between the various attributes in the figure 3: Bivariate Analysis – Heatmap
- There is a positive correlation between National Economic condition and Blair, whereas the same attribute has a negative correlation with Hague.
- The same holds true with Household economic condition.
- There is a negative correlation between Blair and Vote. When Blair's approval ratings increase, the votes are polled for 'Labour'. Likewise, when Hague's approval ratings increase, votes are polled for 'Conservatives'.
- There is a positive correlation between 'Eurosceptic' and Hague and there is a negative correlation between 'Eurosceptic' and Blair. This indicates that the more Eurosceptic people tend to approve Hague while the people on the other extreme tend to approve Blair.

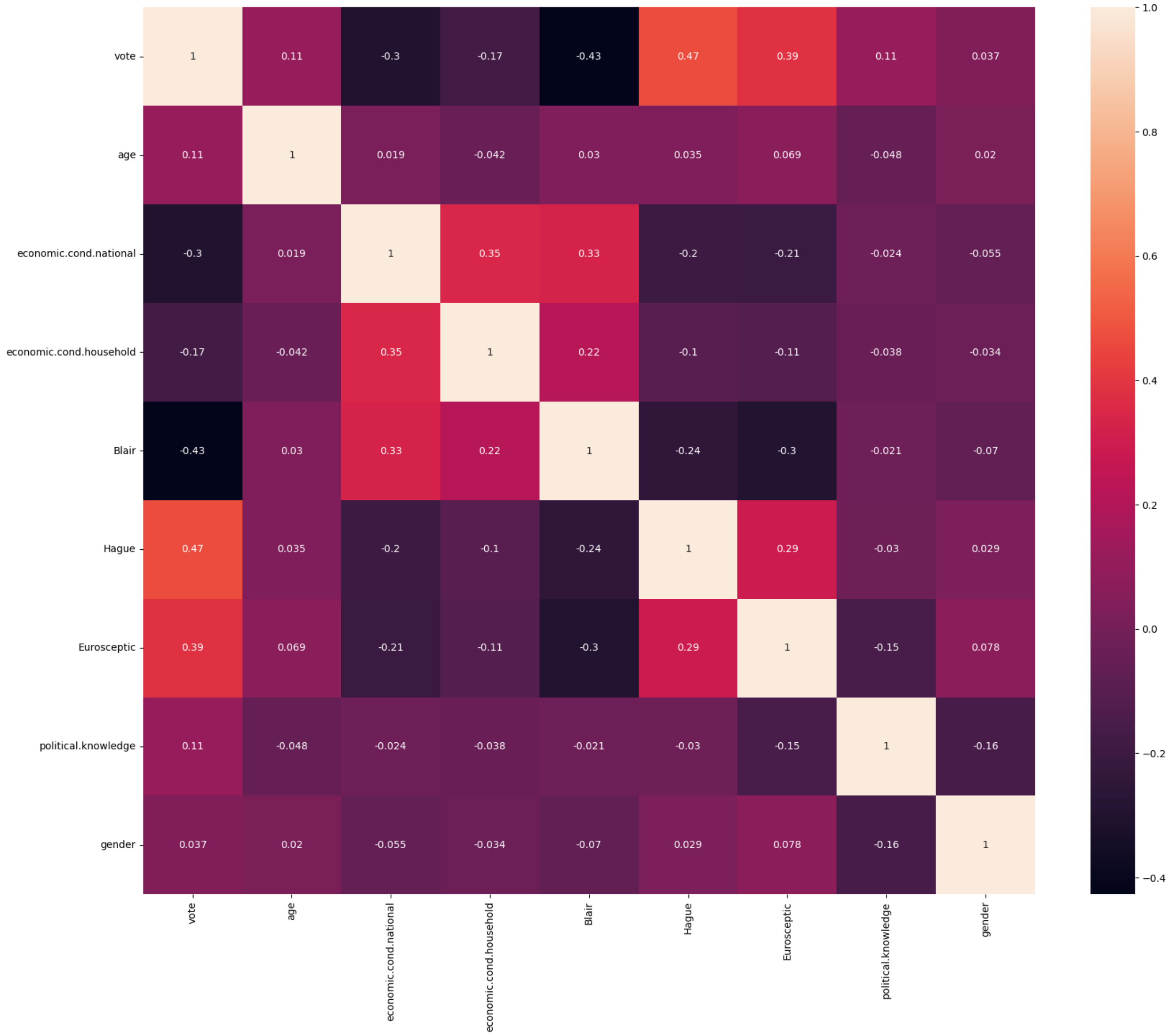

Figure 3: Bivariate Analysis - Heatmap

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

- One hot encoding for columns: Vote and gender has been applied.
- The attribute: ‘Europe’ was renamed to ‘Eurosceptic’ to denote the sentiment against the European Union.
- The values of atribute ‘gender’ was changed. The value ‘male’ has been converted to 0 and ‘female’ has been converted to 1.
- The values of attribute ‘vote’ was changed. The value ‘Labour’ has been converted to 0 and ‘Conservative’ has been converted to 1.

- The attribute: ‘unnamed: 0’ was dropped from the dataset as it was not making any contribution to the analysis.

Is scaling necessary for the dataset provided?

- Out of the 9 columns, 8 columns are categorical in nature, including the dependent variable: ‘vote’. So, no scaling is required for these.
- Since in the machine learning models (Ex: KNN) that are going to be used in subsequent analysis and predictions, scaling is required on the ‘age’ column.
- Before dividing the data into train and test, min-max scaling has been performed on the ‘Age’ column to reshape the data and change the range of Age between 0 and 1.

Split the data into train and test in the ration of 70:30:

- Using the sklearn.model_selection, the data was split into train and test data with random_state =1.
- The code to perform the same is present in the code file.

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

Logistic Regression:

- The data set was split into training and test data in the ration of 70:30.
- Sklearn.metrics package was applied for training the model and making further predictions.
- Following metrics are obtained for training data:

```
model_score: 0.8406747891283973
Confusion Matrix:
 [[667  68]
 [102 230]]
Classification Report:
               precision    recall  f1-score   support

           0       0.87      0.91      0.89       735
           1       0.77      0.69      0.73       332

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

Accuracy Score: 0.8406747891283973
ROC AUC Score: 0.8001270387673141
```

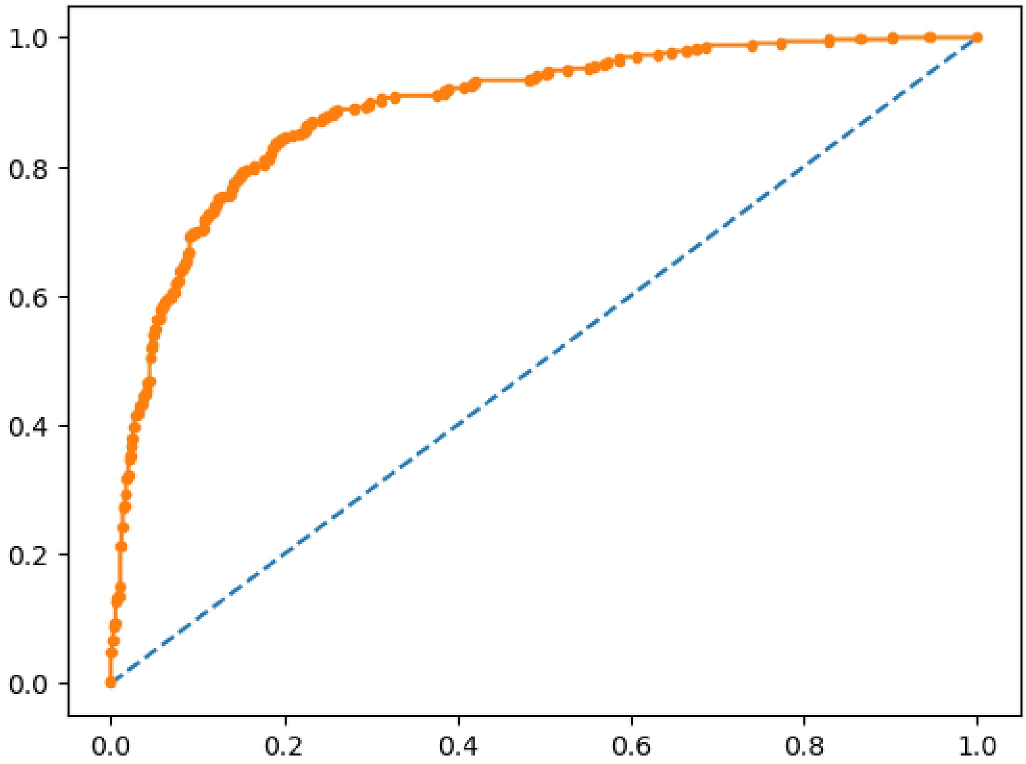

Figure 4 Logistic Regression ROC_AUC Curve - training data

- The following metrics are obtained for test data:

```
Confusion Matrix:
 [[292  36]
 [ 45  85]]
Classification Report:
               precision    recall  f1-score   support

           0       0.87      0.89      0.88       328
           1       0.70      0.65      0.68       130

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

Accuracy Score: 0.8231441048034934
ROC AUC Score: 0.7720450281425891
```

- The model has performed very well with both train and test data for vote = 0 (i.e, for voting for Labor party). For the vote=1, the recall is average at best.
- Same inference can be drawn from the f1 score as well.

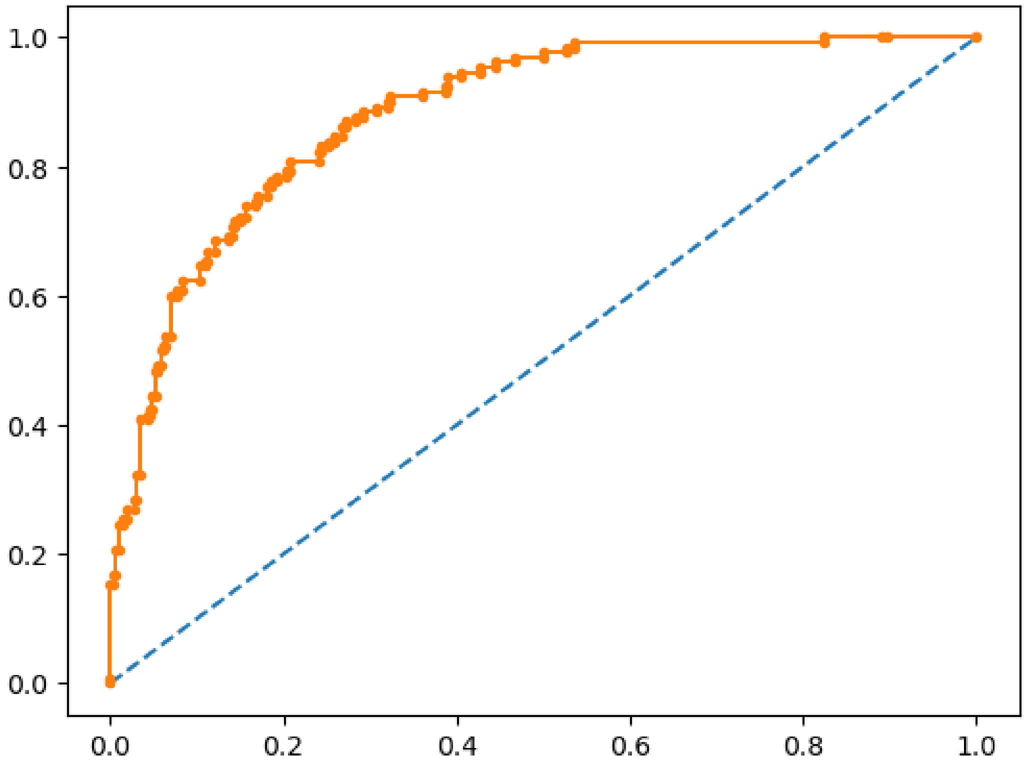

Figure 5 Logistic Regression ROC_AUC Curve – test data

LDA:

- The Linear Discriminant Analysis was performed on the dataset.
- The data was split into training and test data as explained above.
- Metrics obtained from the training data are described below:

```
Confusion Matrix:
 [[660  75]
 [ 99 233]]
Classification Report:
               precision    recall  f1-score   support

           0       0.87      0.90      0.88       735
           1       0.76      0.70      0.73       332

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

Accuracy Score: 0.8369259606373008
ROC AUC Score: 0.799883206294566
```

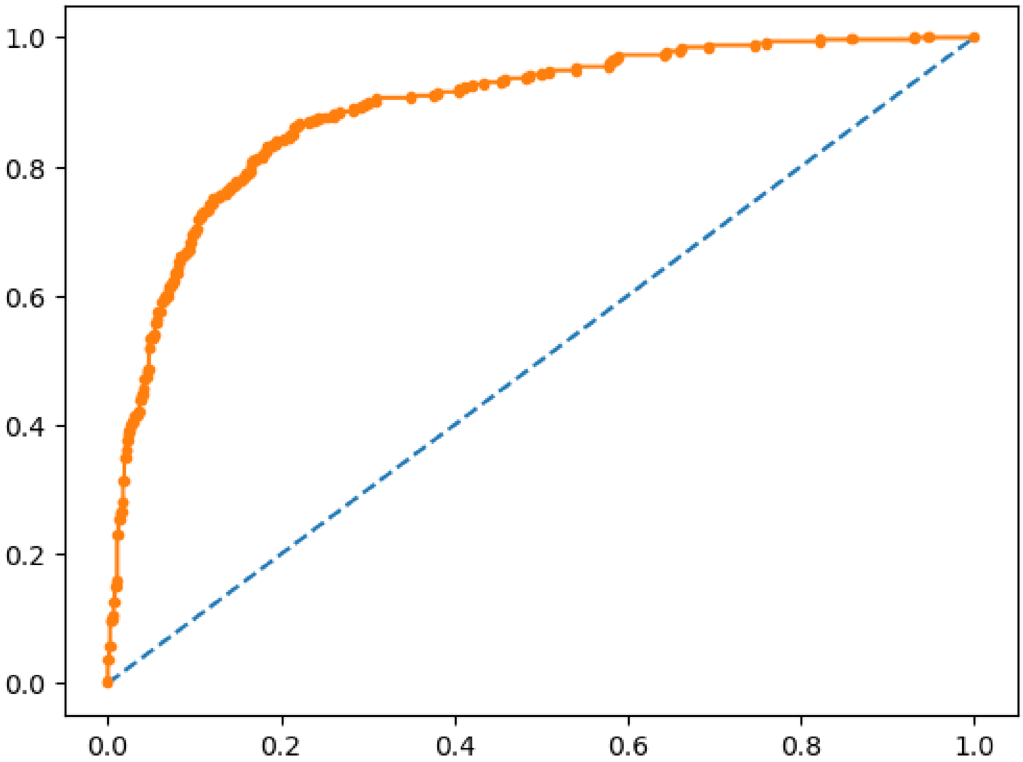

Figure 6 LDA - ROC_AUC Curve – training data

- After training the model on training data, the same was evaluated on test data. Following are the metrics for the same:

```
Confusion Matrix:
 [[289  39]
 [ 44  86]]
Classification Report:
               precision    recall  f1-score   support

           0       0.87      0.88      0.87       328
           1       0.69      0.66      0.67       130

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

Accuracy Score: 0.8187772925764192
ROC AUC Score: 0.7713180112570356
```

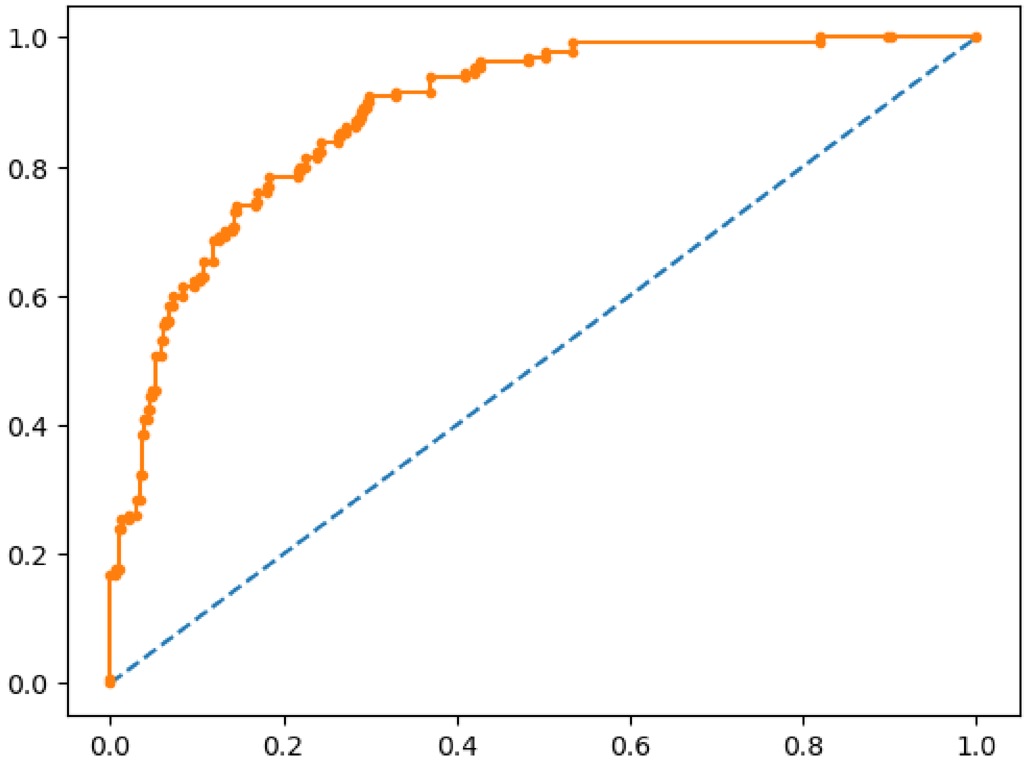

Figure 7 LDA - ROC_AUC Curve – test data

- The LDA model's overall accuracy is marginally better than that of Logistic regression.
- But the recall for vote =1 is marginally higher in LDA than Logistic regression.

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

Naïve Bayes model:

- Naïve Bayes model was applied on the train and then test data. Following are the metrics for train and test data respectively:
- Train data:

```
Confusion Matrix:
 [[649  86]
 [ 92 240]]
Classification Report:
               precision    recall  f1-score   support

           0       0.88      0.88      0.88       735
           1       0.74      0.72      0.73       332

    accuracy                           0.83      1067
```

```
   macro avg       0.81      0.80      0.80      1067
weighted avg       0.83      0.83      0.83      1067

Accuracy Score: 0.8331771321462043
ROC AUC Score: 0.802942381771986
```

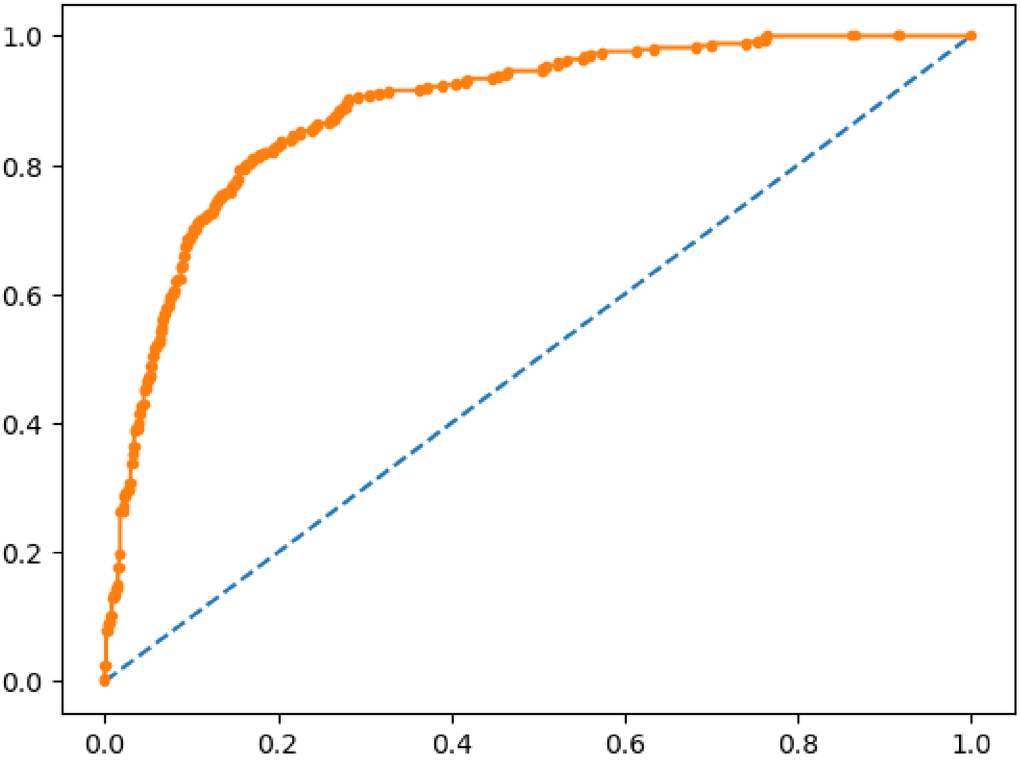

Figure 8 Naive Bayes - ROC AUC Curve - training data

- Following are the scores for test data using Naïve Bayes model:

```
Confusion Matrix:
 [[292  36]
 [ 39  91]]
Classification Report:
               precision    recall  f1-score   support

           0       0.88      0.89      0.89       328
           1       0.72      0.70      0.71       130

    accuracy                           0.84       458
   macro avg       0.80      0.80      0.80       458
weighted avg       0.84      0.84      0.84       458

Accuracy Score: 0.8362445414847162
ROC AUC Score: 0.7951219512195122
```

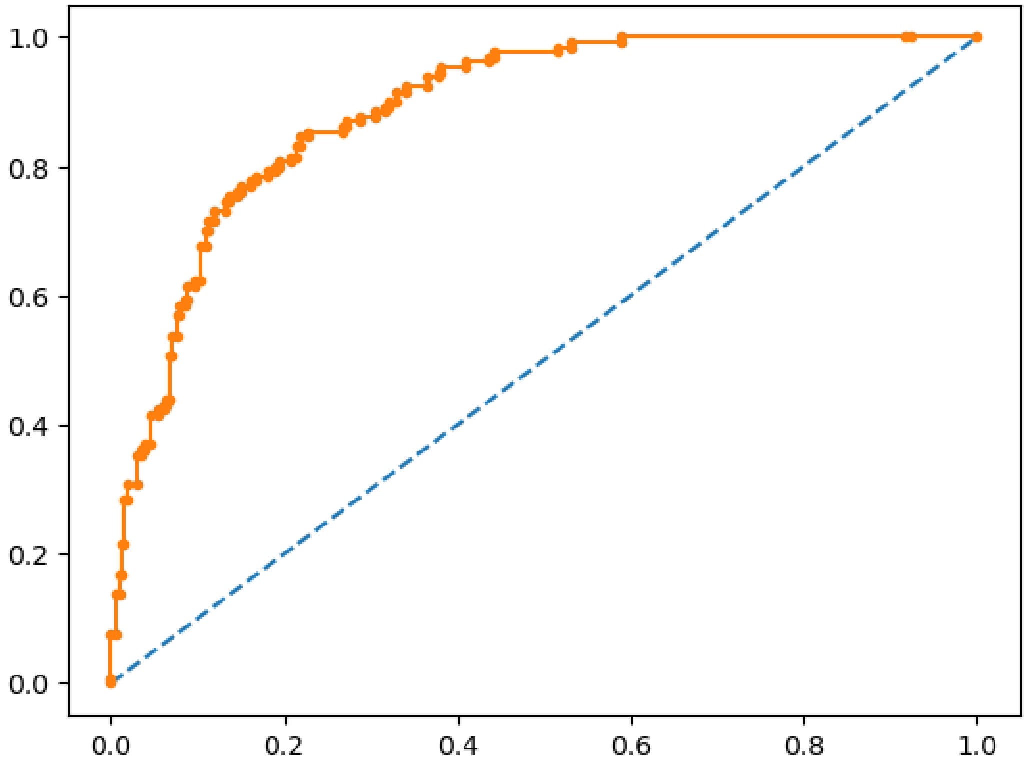

Figure 9 Naive Bayes - ROC AUC Curve - test data

- The accuracy of the Naïve Bayes model is 84%.
- The recall for vote=0 is 89% whereas that of vote=1 is 70%.
- Interpretation: For the test data, the prediction is much more accurate on the votes for the votes=0, i.e, the predictions for the votes when a voter polls on Labour party is higher than that of Conservative party.

KNN model:

- The K-nearest neighbour model also was implemented in the similar way, i.e, fit into training data ad then evaluate using test data.
- Once the model is fit into training data, following are the metrics:

```
Confusion Matrix:
 [[669  66]
 [ 83 249]]
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.91      0.90       735
           1       0.79      0.75      0.77       332
```

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

Accuracy Score: 0.8603561387066542
ROC AUC Score: 0.8301020408163265
```

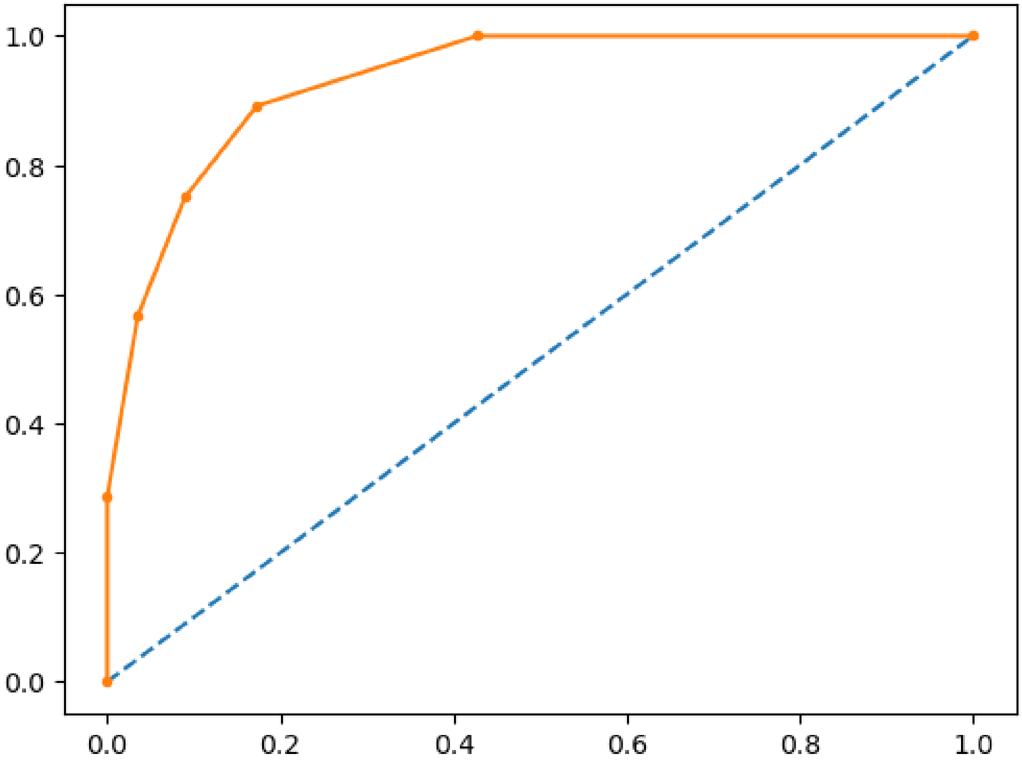

Figure 10 KNN - ROC AUC Curve - training data

- When the same model is applied onto the test data, the metrics obtained is:

```
Confusion Matrix:
 [[275  53]
 [ 39  91]]
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.84      0.86       328
           1       0.63      0.70      0.66       130

    accuracy                           0.80       458
   macro avg       0.75      0.77      0.76       458
weighted avg       0.81      0.80      0.80       458

Accuracy Score: 0.7991266375545851
```

```
ROC AUC Score: 0.7692073170731707
```

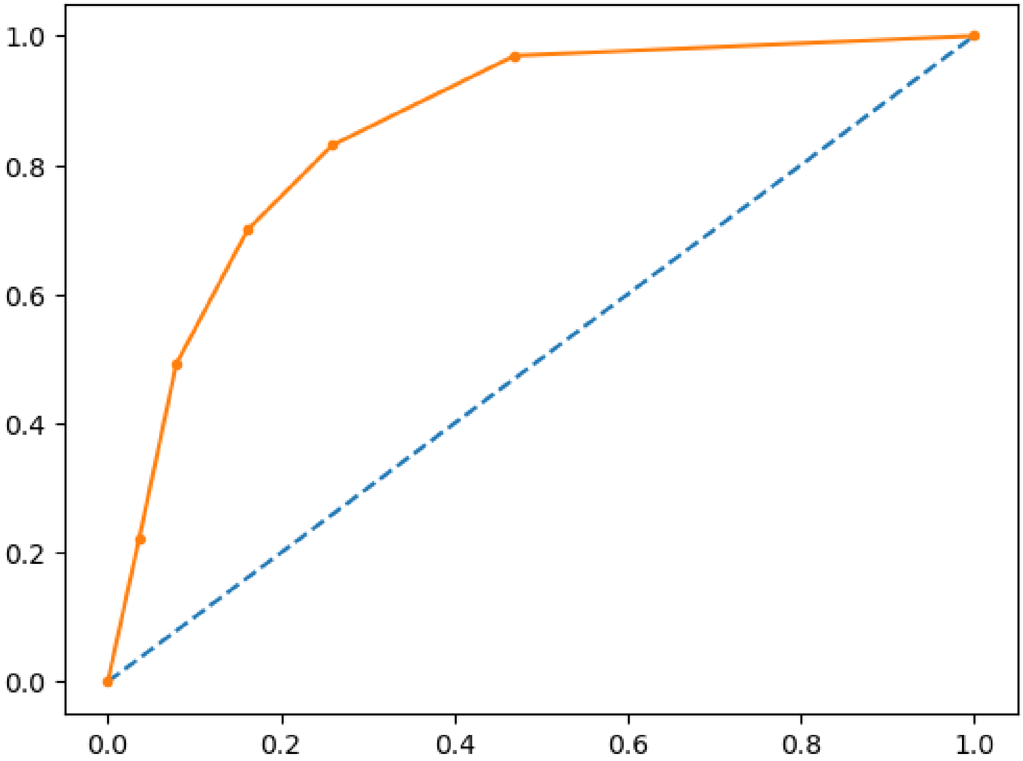

Figure 11 KNN - ROC AUC - test data

- The KNN model is showing as ~80% accurate.
- It is more likely to predict an inaccurate vote for Labour party(Vote = 0) than the Naïve Bayes model. This is evident from the false positive with KNN(53) compared to NB(36)
- Both the models are predicting the False negatives with the same level of inaccuracy.
- Overall, Naïve Bayes is more accurately predicting the dependent variable for the election data.
- All the models above, except KNN have performed well on test data as compared with training data.

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

- Model Tuning: As shown in the Figure 3, there is weak correlation between the independent variables and dropping them will not have any significant impact on the performance of the models. For KNN model, optimum value of K, however, is required to be identified to tune it.
- The optimum value was identified to be k=14.
- With K=14, the classification report and confusion matrices are mentioned below:

```
Confusion Matrix:
 [[284  44]
 [ 41  89]]
Classification Report:
               precision    recall  f1-score   support

           0       0.87      0.87      0.87       328
           1       0.67      0.68      0.68       130

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

- The accuracy of KNN increased by 1% after identifying and applying the optimum value of K.

Bagging and boosting:

- Random forest was implemented and the accuracy was obtained as 81.66%
- Confusion matrix:

```
[[286  42]
 [ 42  88]]
```

- Classification report:

```
               precision    recall  f1-score   support

           0       0.87      0.87      0.87       328
           1       0.68      0.68      0.68       130

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

- The performance of random forest was found to be almost identical to KNN model after tuning.

Decision Tree:

- **Decision Tree Classifier** is utilized for prediction. Following are the metrics obtained:
-

```
Confusion Matrix:
 [[265  63]
 [ 55  75]]
Classsification Report:
               precision    recall  f1-score   support

           0       0.83      0.81      0.82       328
           1       0.54      0.58      0.56       130

    accuracy                           0.74       458
```

```
   macro avg       0.69      0.69      0.69       458
weighted avg       0.75      0.74      0.74       458
```

- The same model when used with **bagging** classifier, provides the below result:

```
Confusion Matrix:
 [[287  41]
 [ 47  83]]
Classsification Report:
               precision    recall  f1-score   support

           0       0.86      0.88      0.87       328
           1       0.67      0.64      0.65       130

    accuracy                           0.81       458
   macro avg       0.76      0.76      0.76       458
weighted avg       0.81      0.81      0.81       458
```

- The above metrics, when compared, shows that the accuracy increased from 74% to 81%.
- The other parameters also showed slight improvement as shown in the classification report.
- Moreover, the feature importance provides the following list with importance of each column as displayed below:

```
                              Imp
age                      0.036732
economic.cond.national   0.022153
economic.cond.household  0.019490
Blair                    0.157607
Hague                    0.426457
Eurosceptic              0.216291
political.knowledge      0.121270
gender                   0.000000
```

- It indicates that the column gender has got no significance in the prediction of vote. This column can be dropped as well.

Boosting:

- **Ada boosting** technique was used on **Decision Tree Classifier**.
- The metrics obtained is as below:

```
Confusion Matrix:
 [[269  59]
 [ 45  85]]
Classification Report:
               precision    recall  f1-score   support

           0       0.86      0.82      0.84       328
           1       0.59      0.65      0.62       130
```

```
    accuracy                           0.77       458
   macro avg       0.72      0.74      0.73       458
weighted avg       0.78      0.77      0.78       458
```

- After using Ada boosting, the accuracy increased from 74% to 77%.
- Similarly, other parameters also showed minor improvements.

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model, classification report

- The plots ROC_AUC curve, Confusion matrix, Classification Report, Accuracy, and ROC_AUC score were shown in the questions 1.4 – 1.7 above.
- The comparison of all the confusion matrices obtained from the different models:

Logistic Regression	[[292 36]
	[45 85]]
LDA	[[289 39]
	[44 86]]
Naïve Bayes	[[292 36]
	[39 91]]
KNN	[[275 53]
	[39 91]]
KNN after tuning	[[284 44]
	[41 89]]
Random Forest	[[286 42]
	[42 88]]
Decision Tree	[[265 63]
	[55 75]]
DT - bagging	[[287 41]
	[47 83]]
DT - Ada boost	[[269 59]
	[45 85]]

- Of all the models, Logistic Regression has the highest True Positive whereas Naïve Bayes has the best True Negative numbers.
- Models like Decision Tree, Random Forest along with various bagging and boosting techniques fail to provide the accuracy of Logistic Regression for the data set provided.
- The classification Report of all the models deployed are consolidated below:

	Precision	Recall	f1-score	support
	Logistic Regression			
0 0.87 0.89 0.88 328				
1 0.70 0.65 0.68 130				
accuracy 0.82 458				
macro avg 0.78 0.77 0.78 458				
weighted avg 0.82 0.82 0.82 458				
	LDA			
0 0.87 0.88 0.87 328				
1 0.69 0.66 0.67 130				
accuracy 0.82 458				
macro avg 0.78 0.77 0.77 458				
weighted avg 0.82 0.82 0.82 458				
	Naïve Bayes			
0 0.88 0.89 0.89 328				
1 0.72 0.70 0.71 130				
accuracy 0.84 458				
macro avg 0.80 0.80 0.80 458				
weighted avg 0.84 0.84 0.84 458				
	KNN			
0 0.88 0.84 0.86 328				
1 0.63 0.70 0.66 130				
accuracy 0.80 458				
macro avg 0.75 0.77 0.76 458				
weighted avg 0.81 0.80 0.80 458				
	KNN with model tuning			
0 0.87 0.87 0.87 328				
1 0.67 0.68 0.68 130				
accuracy 0.81 458				

macro avg 0.77 0.78 0.77 458				
weighted avg 0.82 0.81 0.82 458				
	Random Forest			
0 0.87 0.87 0.87 328				
1 0.68 0.68 0.68 130				
accuracy 0.82 458				
macro avg 0.77 0.77 0.77 458				
weighted avg 0.82 0.82 0.82 458				
	Decision Tree			
0 0.83 0.81 0.82 328				
1 0.54 0.58 0.56 130				
accuracy 0.74 458				
macro avg 0.69 0.69 0.69 458				
weighted avg 0.75 0.74 0.74 458				
	Decision Tree with Bagging			
0 0.86 0.88 0.87 328				
1 0.67 0.64 0.65 130				
accuracy 0.81 458				
macro avg 0.76 0.76 0.76 458				
weighted avg 0.81 0.81 0.81 458				
	Decision Tree with ADA boost			
0 0.86 0.82 0.84 328				
1 0.59 0.65 0.62 130				
accuracy 0.77 458				
macro avg 0.72 0.74 0.73 458				
weighted avg 0.78 0.77 0.78 458				

- Based on the above table, following are the inferences:
    - The Logistic regression, LDA and Random forest provided the highest accuracy overall
    - Naïve bayes has the highest f1-score for both vote =0 and vote =1

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

Based on the analysis performed, following are the recommendations:

- Higher the Eurosceptic sentiment, more the likelihood of the voter voting for Conservatives: As the Eurosceptic sentiment moves from 0 to 11, the probability of a voter voting for Conservatives increases. So, the Exit/opinion pol questions should focus on the voter's Eurosceptic sentiment.
- When the prediction is made of a voter voting for Labour party, it has a high degree of accuracy(range: 75-88%), compared of the accuracy of them voting for Conservative Party.
- Also, the economic sentiment(both household and national) has an impact on the vote. If the economic condition is better, then the voting tends to happen for Labour Party and vice versa. So, focus should be to understand the people's(who are surveyed) economic condition and his/her view of the national economic condition.
- The voting is inclined more with the party(Labour/Conservative) and less with the candidate(Hague/Blair). So, the questions for exit poll should be more party-centric and less candidate centric.

Problem 2:

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

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

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

2.1 Find the number of characters, words, and sentences for the mentioned documents. – 3 Marks

2.2 Remove all the stopwords from all three speeches. – 3 Marks

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) – 3 Marks

2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks [refer to the End-to-End Case Study done in the Mentored Learning Session]

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

Total number of characters before and after processing is shown in the table below:

	Name	char_count	word_count	full_stop_count	totalwords
0	Roosevelt	7651	1323	69	691
1	Kennedy	7673	1364	56	744
2	Nixon	10106	1769	70	866

Note:

- char_count: total number of characters
- word_count: total number of words before removal of stopwords
- totalwords: total number of words after removal of stopwords
- full_stop_count: total number of sentences.

2.2 Remove all the stopwords from all three speeches

Snippet of the speech before and after the removal of stop-words:

Speech	stopwords	word_count	totalwords
On each national day of inauguration since 178...	On national day inauguration since 1789, peopl...	1323	691
Vice President Johnson, Mr. Speaker, Mr. Chief...	Vice President Johnson, Speaker, Chief Justice...	1364	744
Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	Vice President, Speaker, Chief Justice, Senato...	1769	866

Note:

- Word_count is the count of words before removal of stopwords
- Totalwords is the count of words after removal of stopwords.

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)

Below is the word count that occurred the most number of times for each president:

President	Word	Count
Roosevelt	us	44
Kennedy	new	25
Nixon	let	18

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

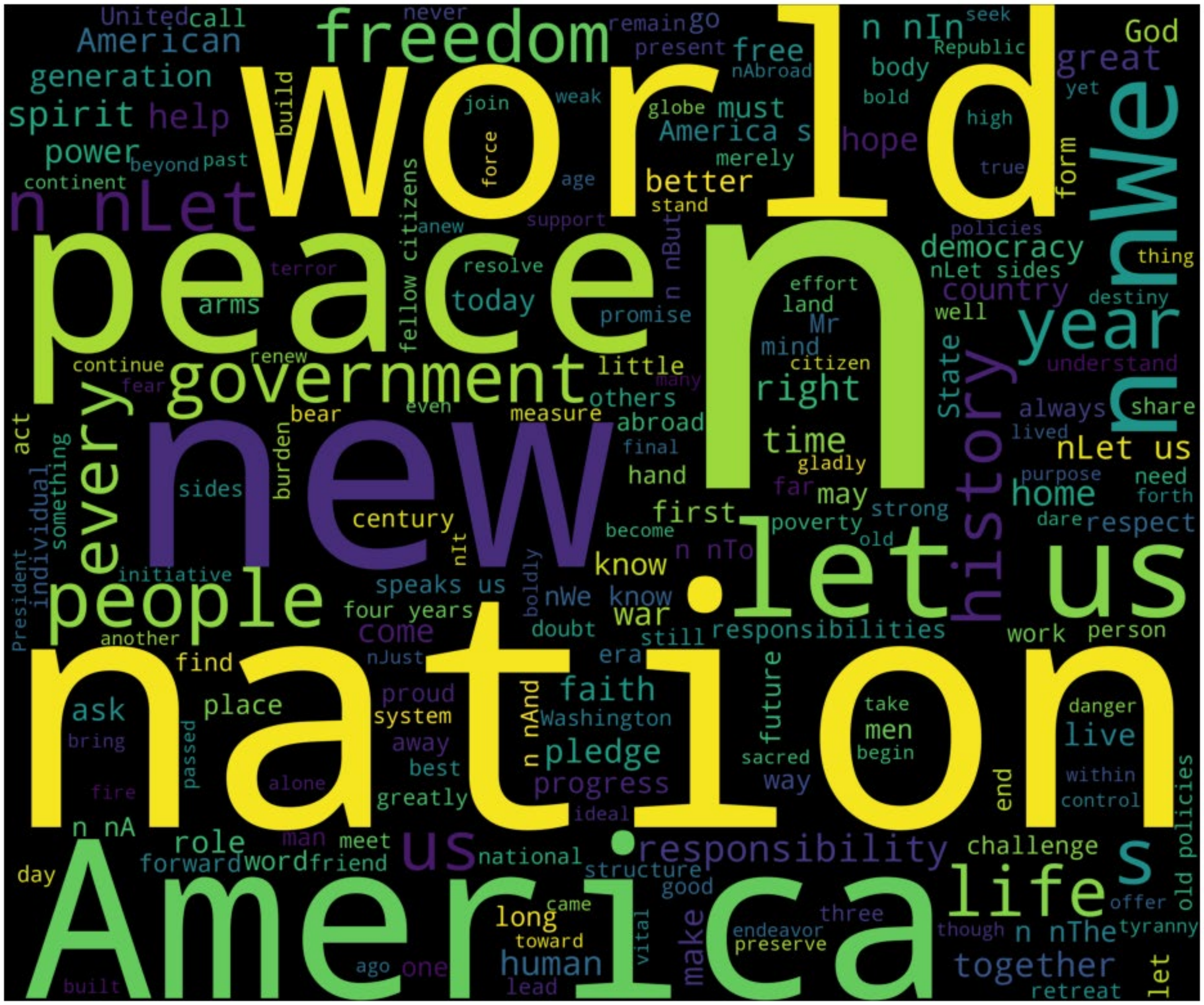

Figure 12 Word Cloud - Presidents' speeches