Project

Machine Learning

Vinay Santosh

3rd October 2021

DSBA March 21 B G7

Overview:

This project applies various Ensemble Techniques used in Machine Learning. It provides analysis of the datasets which are used to draw insights and recommendations. The project also touches upon the fundamentals of text analysis.

Goals:

- To perform exploratory data analysis on the given dataset.
- Perform data ingestion, encoding and build various models.
- To visualize the results in the form of tables and plots.
- Perform text analysis to identify word frequencies/ patterns.
- Draw insights and provide business recommendations from the analysis.

Contents:

List of Figures:

List of Tables:

Problem #1: Predicting Election Results

The problem statement demands a prediction of the exit poll based on the voter information at hand.

The voters choose between two candidates, Tony Blair of the Labour party and William Hauge of the Conservative party.

	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

1.1 Sample Dataset

1.1 Descriptive Analysis:

- The given dataset contains a total of 1525 entries with 9 different variables which provide information of each voter such as voter age, vote cast, gender, assessment of candidates by the voter, etc.
- There are 7 features that are of integer datatype and 2 of object datatype.
- There are no null values present in the dataset and it contains 8 duplicate entries which are removed for further analysis.

```
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)
```

1.2 Data Info

- We observe that the majority of votes are cast for the Labour party ie 70%.
- The number of male and female voters are closely balanced with 53% of entries are female voters.

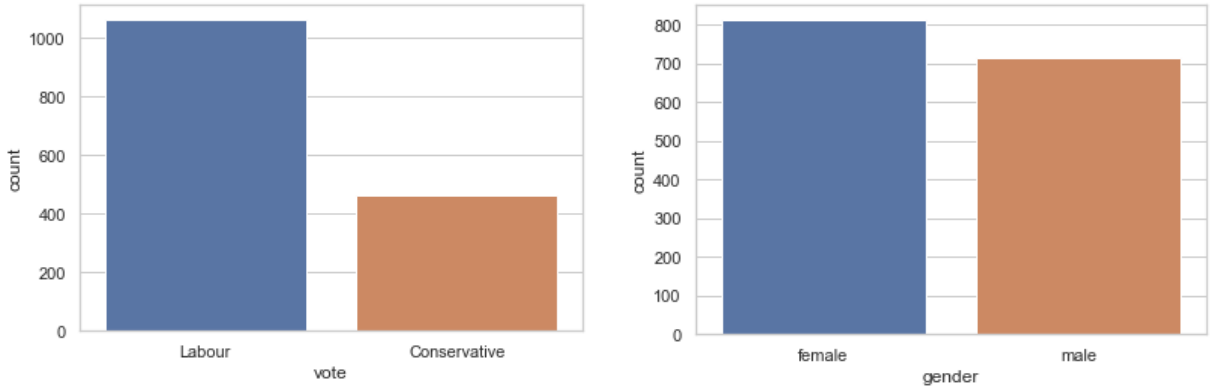


Fig 1.1 Count Plot of Catagorical data

```
VOTE :  2
Conservative     462
Labour          1063
Name: vote, dtype: int64


Labour          0.697049
Conservative    0.302951
Name: vote, dtype: float64


GENDER :  2
male      713
female    812
Name: gender, dtype: int64


female    0.532459
male      0.467541
Name: gender, dtype: float64
```

1.3 Counts of Categorical Variables

- The minimum age of the voters in 24 and the maximum is 93. The median age of the voters is 53 years.
- The average assessment of economical condition on both national and household level is 3 on a scale of 1 to 5.

	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

1.4 Summary of the dataset

1.2 Exploratory Data Analysis:

- The average assessment of Tony Blair is higher (3.3) as compared to that of William Hague (2.7).
- We can see that on an average the voters scored 6.7 on a 11 point scale when asked about their attitude on European integration. However, from the distribution, plot we can see that a large number of voters scored 10.
- The average knowledge of the voters when it comes to the partie's stance on European integration is 1.5 out of 3. From the distribution plot, we observe that a large majority of the voters have a high score which indicates that they caste their votes based on the party's stance on the European integration while for a group of voters this knowledge does not matter.

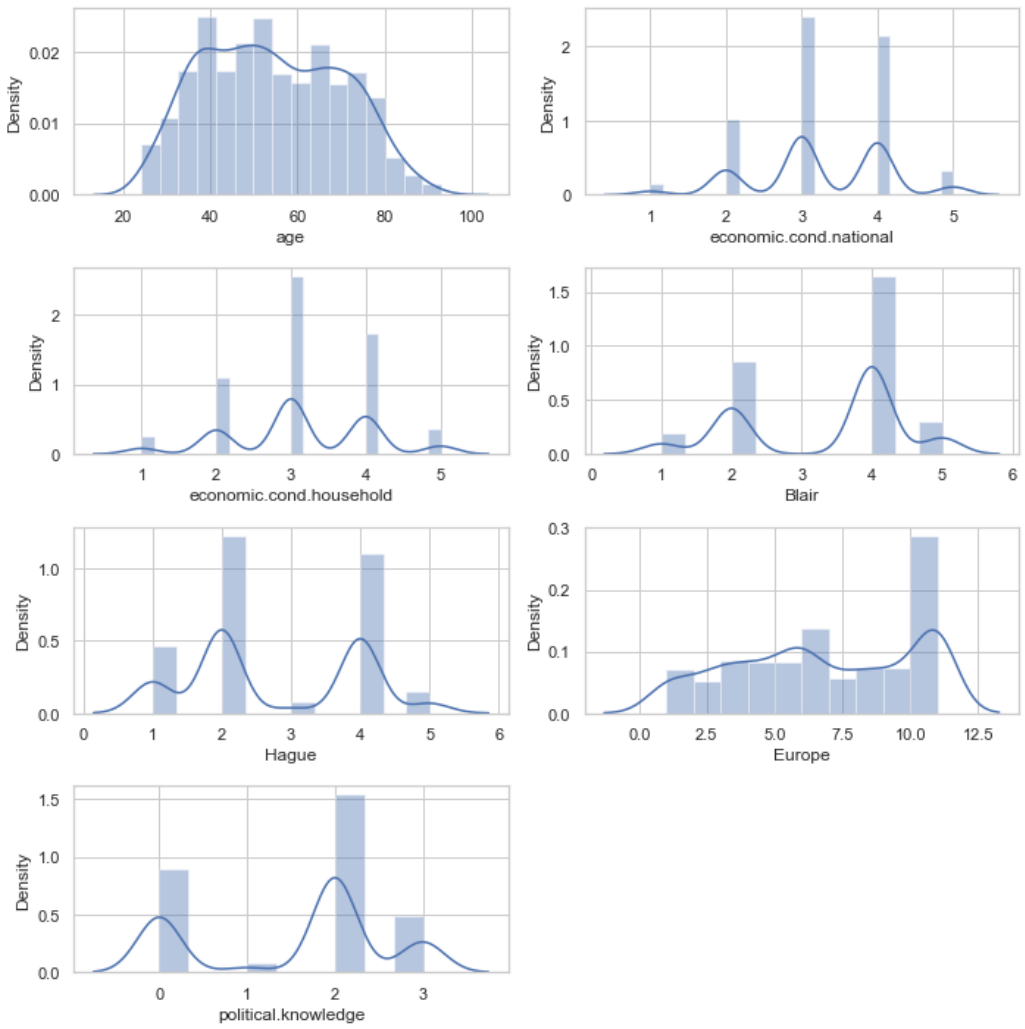


Fig 1.2 Univariate Distribution

- The dataset contains a few outliers and upon removing them we can observe the boxplot of the numeric data given in the figure below.

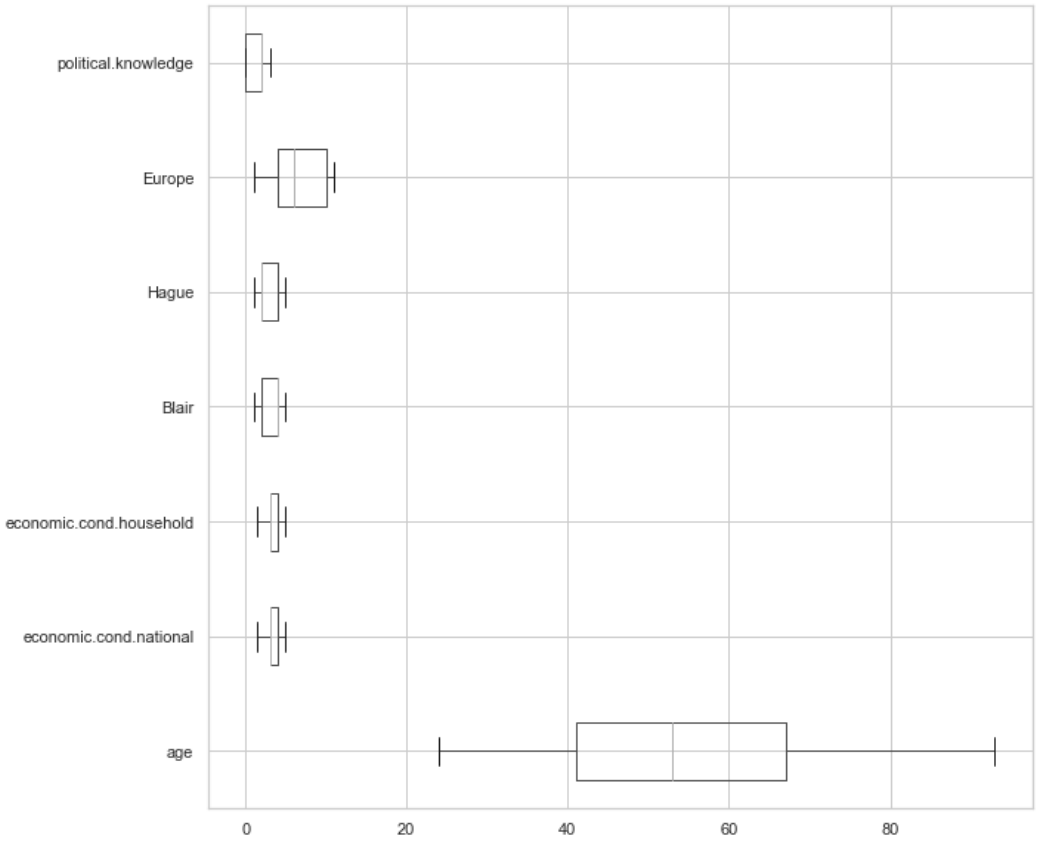


Fig 1.3 Boxplot of Numeric Data

- The heatmap below shows that there is no strong correlation between the features.

Fig 1.4 Correlation HeatMap

1.3 Data Preperation:

- The features of object datatype are converted into numeric for further analysis and model building.
- The various numeric data are in different scales. Age is stored as a whole number, assessment of various economic conditions and candidates are scored on a 5 point scale.
- The feature “Europe” is scaled on a 11 point scale and “political knowledge” is scaled from 0 to 3.
- The scaling of data will centralize the features to 0 and the standard deviation/ variance will be close to 1.
- Scaling is required for certain distance based algorithms like KNN. Models like Logistic Regression and LDA performs better on scaled data.

- However for models like Naive Bayes and boosting techniques, scaling does not affect the accuracy scores.
- Bagging and Boosting models are not affected by scaling and performs the same.
- We split the dataset into Train and Test sets in the ratio of 70:30. The resulting target variable “vote” have an approximate majority class (Labour votes) of 70% and minority class (Conservative votes). Dealing with the imbalance is not required in this case as the minority class is not below 5%.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	1	43	3.0	3.0	4	1	2	2	0
1	1	36	4.0	4.0	4	4	5	2	1
2	1	35	4.0	4.0	5	2	3	2	1
3	1	24	4.0	2.0	2	1	4	0	0
4	1	41	2.0	2.0	1	1	6	2	1

1.5 Encoded Dataset

Model Building, Tuning and Comparison:

- Using the training data, we build models which help predict results of future predictions.
- The test set is used to check the models performance on accuracy, precision and recall. We can make decisions about the model based on these metrics.
- In Ensamble Techniques, we train multiple models using different algorithms and evaluate it’s performances on the test sets. The optimal model is then choses to address the given problem.
- For the given problem statement, we have to build a model that best predicts the choice of candidate the voters will choose based on certain information. Thus the ideal model should be highly accurate inorder to for us to predict the outcome of the election.

1.4 Logistic Regression and Linear Discriminant Analysis:

- We build the models with the default parameters for both algorithms and check it’s performance on both train and test sets.

	Train Set Score	Test Set Score
Log R	0.840675	0.820961
LDA	0.839738	0.818777

- The train score for both the tests are similar. The logistic regression model performs better on the test set.

- We can not tune the models to better the accuracy scores. Upon performing the grid searches on both the models, we can determine the best parameters.
- By passing the best parameters to the algorithms, we derive the following scores for both the train and test sets.

	Best Score on Train Set	Best Score on Test Set
Log R	0.839738	0.820961
LDA	0.839738	0.839738

- We observe that the LDA model performs better on the test set after model tuning. Both these models perform well and do not overfit or underfit the data.

1.5 KNN and NB Models:

- We further build the K-nearest neighbours and Naive Bayes models but training them and observing it's performances.
- Since KNN is a distance based algorithm, we scale the data using the z score technique and build the model.
- Naive Bayes is a probabilistic model that is used for classification problems. It is based on Bayes theorem and assumes that the features in the dataset are independent of each other.

	Train Set Score	Test Set Score
KNN	0.864105	0.818777
NB	0.832240	0.823144

- We observe that the KNN model performs best on the train set but the Naive bayes model outperforms on the test set.

	Best Score on Train Set	Best Score on Test Set
KNN	0.840675	0.816594
NB	0.839738	0.818777

- We choose the optimum number of neighbours by plotting the following misclassification error graph below and conclude that the value 11 is the ideal number of neighbours.

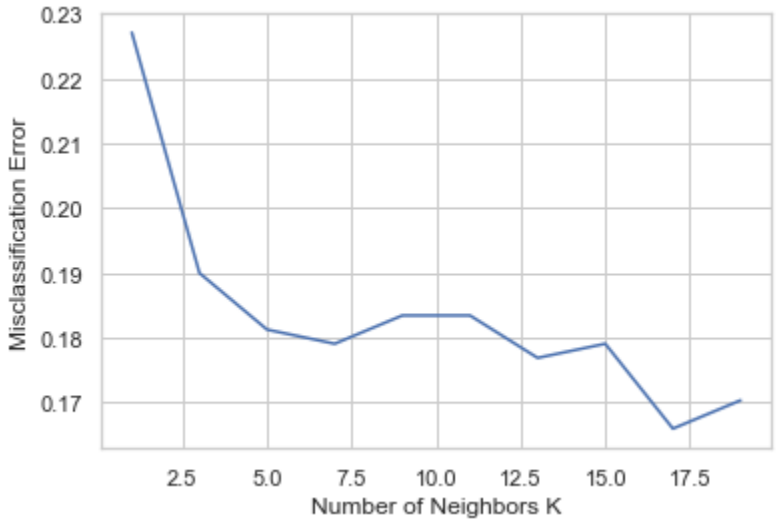


Fig 1.5 Mis-classification Error Graph

- By tuning the models and choosing the appropriate hyper parameters based on the dataset, both these output similar scores with the Naive Bayes model doing slightly better.

Best Parameters:

- By changing the hyperparameters from it's default values, we are able to optimize the models and better it's accuracy in most cases.
- The table below shows the best parameters used in each of the models.

Models:	Best Parameters:
Logistic Regression	'penalty': 'l2', 'solver': 'lbfgs', 'tol': 0.0001
LDA	'solver': 'svd', 'tol': 0.0001
KNN	'algorithm': 'ball_tree', 'leaf_size': 30, 'n_neighbors': 13, 'weights': 'uniform'
NB	'priors': None, 'var_smoothing': 1e-09

1.6 Best Parameters

1.6 Bagging and Boositng:

- Bagging is an ensemble technique that is done in parallel where as boosting is done in a sequential manner.
- Bagging runs multiple samples from the same dataset which are subsets of columns and rows.
- We use two types of boosting algorithms, Ada boost and gradient boost on the dataset.

	Train Set Score	Test Set Score
Bagging	0.981256	0.814410
Ada Boost	0.842549	0.823144
Gradient Boost	0.873477	0.836245

- From the model scores, we find that bagging overfits the data as it performs exceptionally well on the train set but poorly on the test.
- Gradient boosting gives us the best result in accuracy compared to all other models.

1.7 Performance Metrices of models:

- We compare the various models' performances by calculation the different metrics like precision, recall and accuracy.
- We can observe that the f1 scores and accuracy is highest for the Gradient Boosting model.

Logistic Regression:

Train:

	precision	recall	f1-score
0	0.77	0.69	0.73
1	0.87	0.91	0.89
accuracy			0.84

Test:

	precision	recall	f1-score
0	0.70	0.65	0.67
1	0.87	0.89	0.88
accuracy			0.82

LDA:

Train:

	precision	recall	f1-score
0	0.76	0.71	0.73
1	0.87	0.90	0.89
accuracy			0.84

Test:

	precision	recall	f1-score
0	0.69	0.66	0.67
1	0.87	0.88	0.87
accuracy			0.82

KNN:

Train:

	precision	recall	f1-score
0	0.76	0.70	0.73
1	0.87	0.90	0.89
accuracy			0.84

Test:

	precision	recall	f1-score
0	0.67	0.70	0.69
1	0.88	0.87	0.87
accuracy			0.82

Naive Bayes:

Train:

	precision	recall	f1-score
0	0.76	0.65	0.70
1	0.85	0.91	0.88
accuracy			0.83

Test:

	precision	recall	f1-score
0	0.68	0.72	0.70
1	0.89	0.86	0.87
accuracy			0.82

Bagging:

Train:

	precision	recall	f1-score
0	0.96	0.98	0.97
1	0.99	0.98	0.99
accuracy			0.98

Test:

	precision	recall	f1-score
0	0.66	0.70	0.68
1	0.88	0.86	0.87
accuracy			0.81

Ada Boosting:

Train:

	precision	recall	f1-score
0	0.77	0.71	0.74
1	0.87	0.90	0.89
accuracy			0.84

Test:

	precision	recall	f1-score
0	0.69	0.69	0.69
1	0.88	0.88	0.88
accuracy			0.82

Gradient Boosting:

Train:

	precision	recall	f1-score
0	0.84	0.79	0.81
1	0.91	0.93	0.92
accuracy			0.89

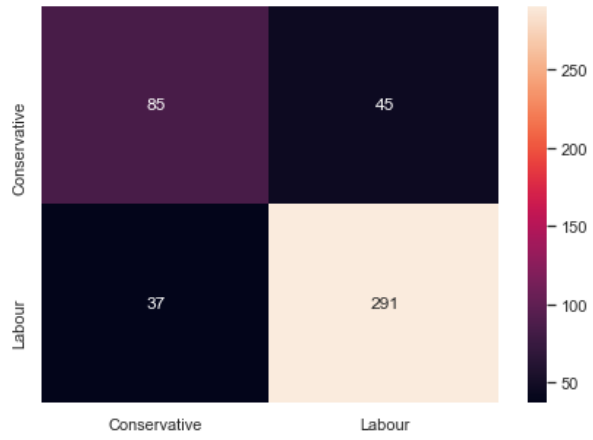
Test:

	precision	recall	f1-score
0	0.69	0.74	0.71
1	0.89	0.87	0.88
accuracy			0.83

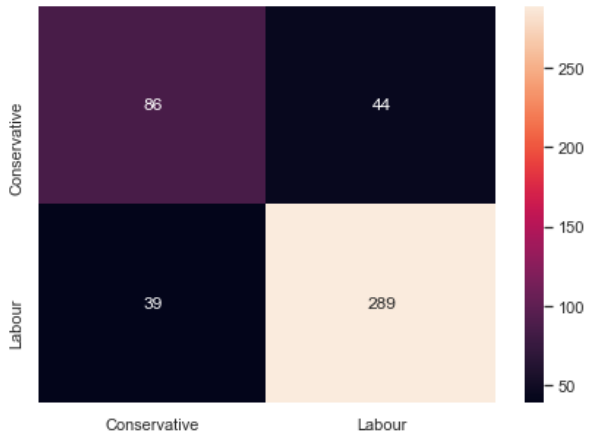
1.7 Performance Metrices

Confusion Matrices on Test data:

Logistic Regression:

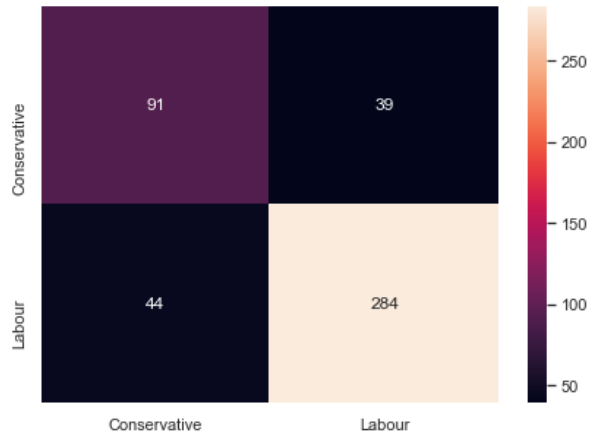


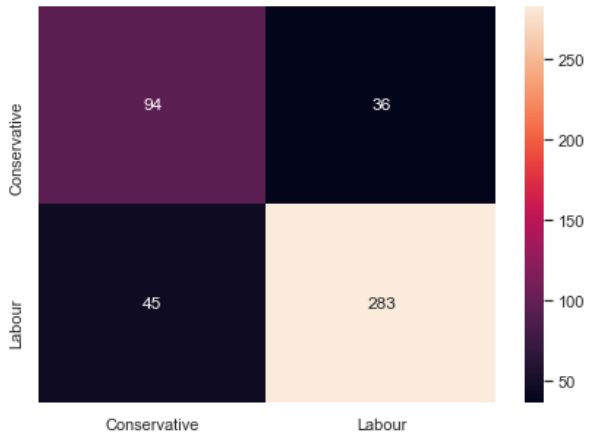
LDA:



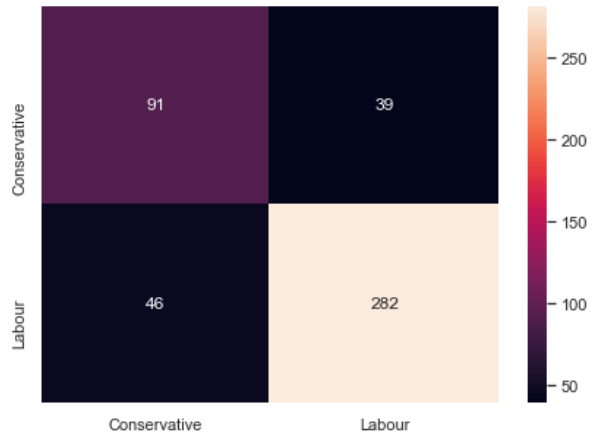
KNN:

Naive Bayes:

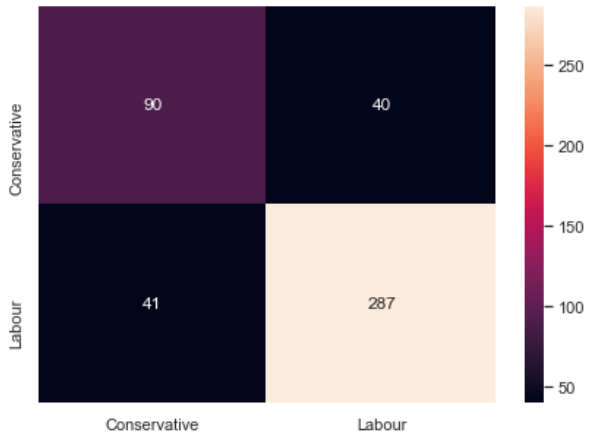




Bagging:



Ada Boost:



Gradient Boost:

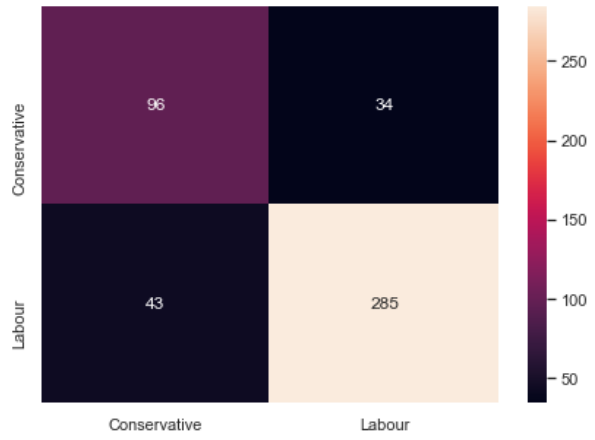


Fig 1.6 Confusion Matrices

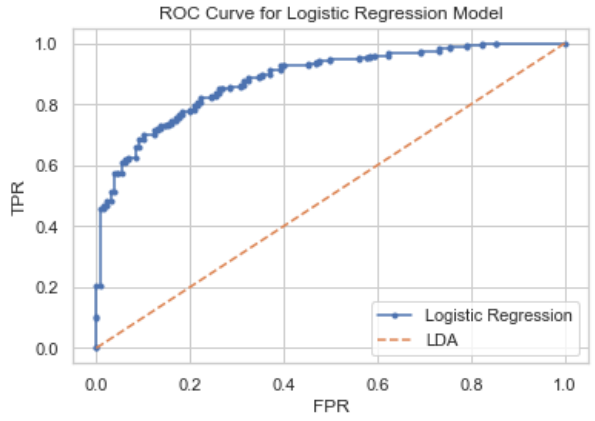
- From the matrices above, we can see the number of right and wrong predictions of each class by all the models built.
- The Ada boost has the minimum amount of wrong predictions of the “Labour” class and Gradient boost has the minimum number of wrong predictions of the “Conservative” class.

AUC and ROC:

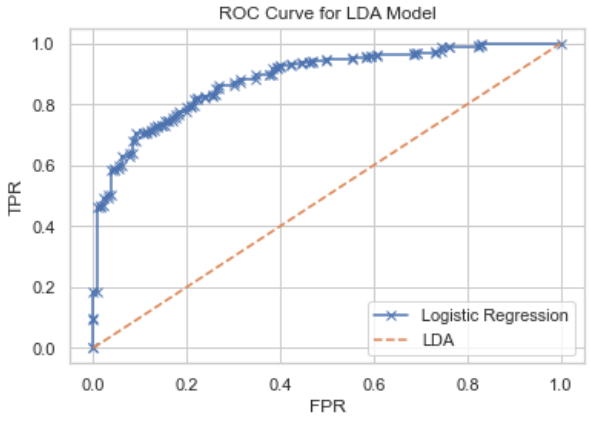
- We calculate the Area Under the curve scores for all the models and plot the Receiver Operating Characteristic Cure for the same.
- The gradient boosting model has the highest AUC score of 90.4 which indicates that it performs across all thresholds.

```
Area under the curve for Logistic Regression Model is 0.8833489681050657
Area under the curve for LDA Model is 0.8841932457786117
Area under the curve for KNN Model is 0.8686913696060037
Area under the curve for NB Model is 0.875375234521576
Area under the curve for NB Model is 0.875375234521576
Area under the curve for Bagging Model is 0.8661819887429644
Area under the curve for AdaBoosting Model is 0.8801946529080675
Area under the curve for Gradient Boosting Model is 0.9043386491557224
```

Logistic Regression:

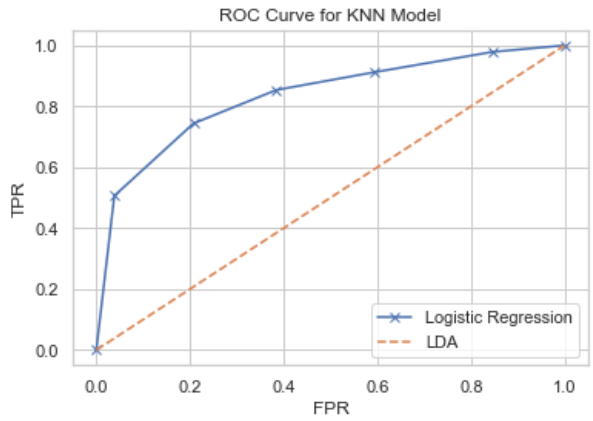


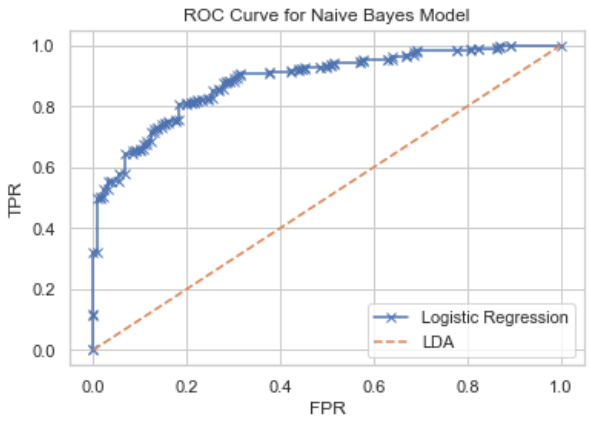
LDA:



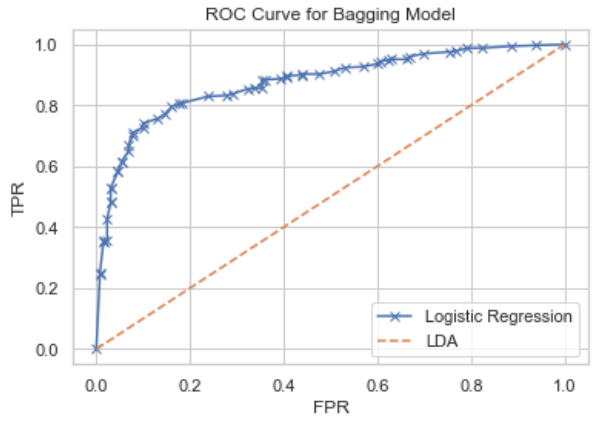
KNN:

Naive Bayes:

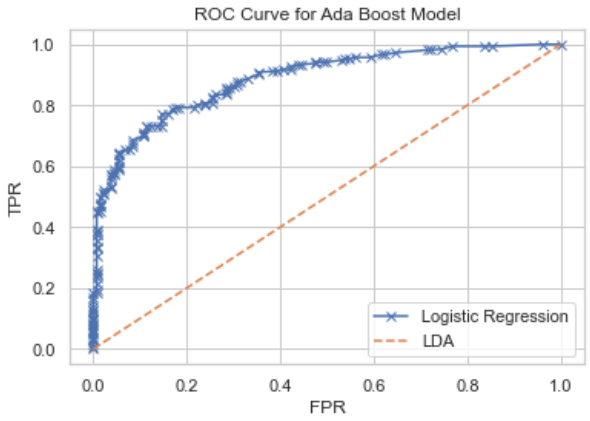




Bagging:



Ada Boosting:



Gradient Boosting:

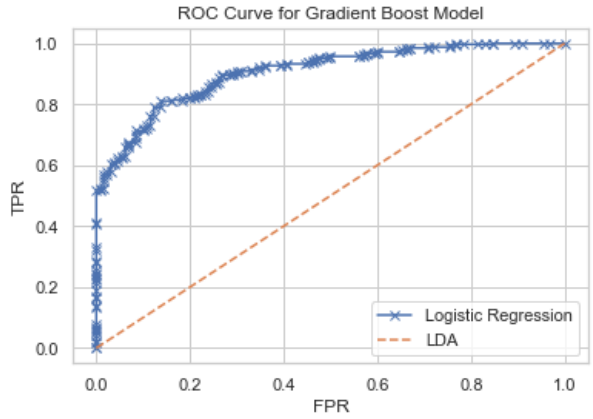


Fig 1.7 ROC Curves

1.8 Insights and Recommendations:

- From the above analysis, we can implement a model to help predict the election results based on the provided voter information.
- The majority of the voters favor the Labour party and the candidate Tony Blair who is more likely to gain more number of seats.
- The candidates view towards European integration seems to play a key role in the voter's choice.
- We will ideally require more datapoints to make a more accurate prediction of the results.

Problem 2: Text Analysis

- The following analysis is done on three presidential inauguration speeches by former American Presidents - Franklin D. Roosevelt, John F. Kennedy and Richard Nixon.

2.1 Word, Character and Sentence Counts:

- To count the number of words in a text document, we use the .words() function and the .raw() for finding out the length of the characters. To count the number of sentences we use the .sents() function.

```
The number of words in the Roosevelt Speech:  1536
The number of words in the Kennedy Speech:  1546
The number of words in the Nixon Speech:  2028
```

```
The number of sentences in the Roosevelt Speech:  68
The number of sentences in the Kennedy Speech:  52
The number of sentences in the Nixon Speech:  69
```

```
The number of characters in the Roosevelt Speech:  7571
The number of characters in the Kennedy Speech:  7618
The number of characters in the Nixon Speech:  9991
```

2.2 Stopwords Removal:

- Stopwords are a list of commonly used words in the English language. To clean a text document, we remove all the stopwords and punctuations present in it.
- A cleaned text makes it easier for us to do further analysis.
- We can add more words to the list that we wish to remove from the document by using the .extend() function.
- The following is a sample of the list of stopwords available in the library:

```
[i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
 'himself']
```

1.8 Stopwords Examples

2.3 Word Frequency:

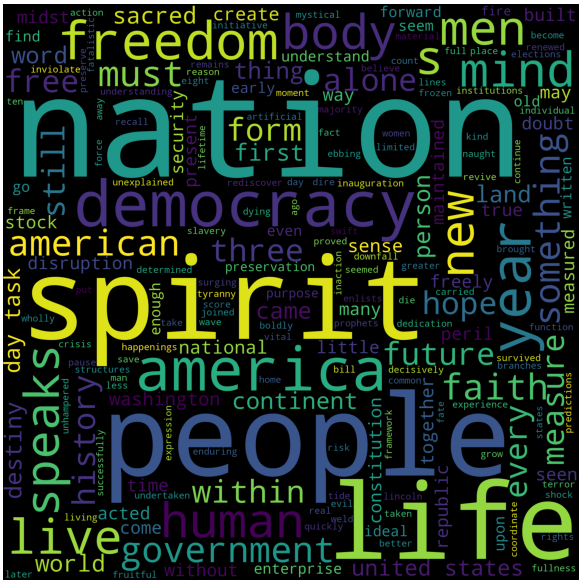
- To find out the word frequencies, we tokenize the data such that each word is converted into a token.
- Since two of the speeches have “let” and “us” with high opccurances and are pretty common words, we add them to the stopwords list.
- The most top 3 most frequently used words in Roosevelt's speech are : ['america', 'peace', 'world']
- The most top 3 most frequently used words in Kennedy's speech are : ['america', 'peace', 'world']

- The most top 3 most frequently used words in Nixon's speech are : ['america', 'peace', 'world']

2.4 Word Counts:

- To extract the word cloud, we add the cleaned text to a corpus to create a bag of words.
- The spaces are removed and is added to the word cloud function.

Roosevelt's Word Cloud:



Kennedy's Word Cloud:

Nixon's Word Cloud:

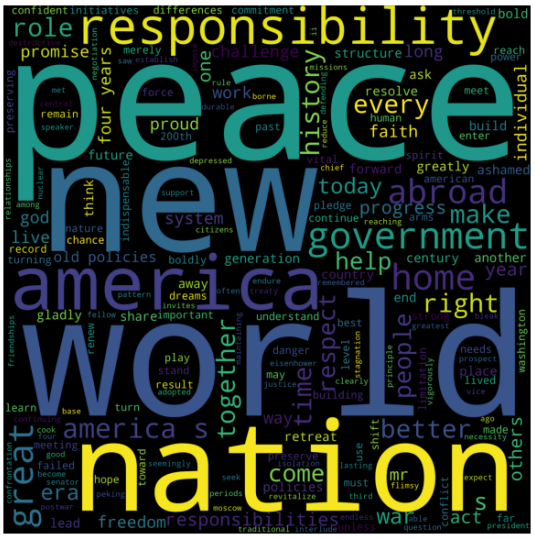


Fig 2.1 Word Clouds