PROBLEM 1

Election Data Analysis

Executive Summary:

A leading news channel, CNBE, have conducted a survey on voters in order to analyze the recent elections. They would like to have an exit poll, wherein it can be predicted whether people will vote for the Labour Party or the Conservative party.

1.1 Reading the dataset:

The dataset had 15,250 entries, with 1525 rows and 10 columns. One of the columns (Unnamed: 0) was non-essential to the analysis, so was dropped right in the beginning, leaving 9 columns in the dataset.

The target variable here is the column ‘vote’.

The first 5 rows of the resulting dataset are given in the table below:

Table: 1.1

	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

Data description:

A brief description of the variables that were used for the analysis:

Table: 1.2

Variable	Description
Vote	Party choice: Conservative or Labour
Age	Age in years
economic.cond. national	Assessment of current national economic conditions, 1 to 5.
economic.cond. household	Assessment of current household economic conditions, 1 to 5.
Blair	Assessment of the Labour leader, 1 to 5.
Hague	Assessment of the Conservative leader, 1 to 5.
Europe	An 11-point scale that measures respondents' attitudes toward European integration. (High scores represents 'Eurosceptic' sentiment)
political. knowledge	Knowledge of parties' positions on European integration, 0 to 3.
gender	Female or male.

Checking datatypes:

The data types of the dataset was checked. The following data types were present in the dataset:

Table: 1.3

Data type	Variables
float64	age, economic.cond.national, economic.cond.household, Blair, Hague, Europe, political.knowledge
object	vote, gender

Statistical summary of the data:

Post imputation of missing values, the statistical summary of the data was checked using the describe() function.

Table: 1.4

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
count	1525	1525	1525	1525	1525	1525	1525	1525	1525
unique	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2
top	Labour	NaN	NaN	NaN	NaN	NaN	NaN	NaN	female
freq	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN	812
mean	NaN	54.182	3.246	3.140	3.334	2.747	6.729	1.542	NaN
std	NaN	15.711	0.881	0.930	1.175	1.231	3.298	1.083	NaN
min	NaN	24	1	1	1	1	1	0	NaN
25%	NaN	41	3	3	2	2	4	0	NaN
50%	NaN	53	3	3	4	2	6	2	NaN
75%	NaN	67	4	4	4	4	10	2	NaN
max	NaN	93	5	5	5	5	11	3	NaN

Inferences:

- A whooping 1063 voters (out of 1525) have voted for the Labour party, which is approx. 69.7% of the people who have voted.
- Most voters are middle-aged persons (mean age of the voters is around 54 years), the youngest being 24 years old, and the oldest voter being 93 years old!
- Most of the voters have a relatively high regard about the economic conditions of the nation (they rate it about 3.246 out of 5), while they feel somewhat lesser about the economic conditions of households

(3.14 out of 5). On the whole, their outlook towards economic conditions is above average.

- In terms of the two party leaders, leader of the Labour Party, Tony Blair, was rated more liberally by the voters (scoring approx. 3.33 / 5). The Conservative party leader, William Hague scored only about 2.75 /5. This is indicative that Blair is a far more popular leader than Hague.
- Regarding their outlook towards European integration, the voters have shown a mean score of 6.73 out of 11, which is more of a negative attitude towards the union. This implies that majority of the voters are sceptic about their country integrating into Europe.
- Concerning the voters' knowledge of the stand/attitude of political parties on the subject of European integration, they are fairly uninformed. A score of approx. 1.5 / 3 shows that they only have moderate information regarding the parties' outlook toward the integration.
- In terms of gender, there are more female voters, but the proportion is fairly divided (about 53% females and 47% males).

Checking for duplicate values:

The function duplicated() was used to check for duplicate values - it revealed that there were 8 duplicate rows present in the dataset. Since the proportion of duplicate rows is very minuscule (8 / 1525 = 0.52 %), we have opted to delete the duplicate rows entirely.

Checking for null values:

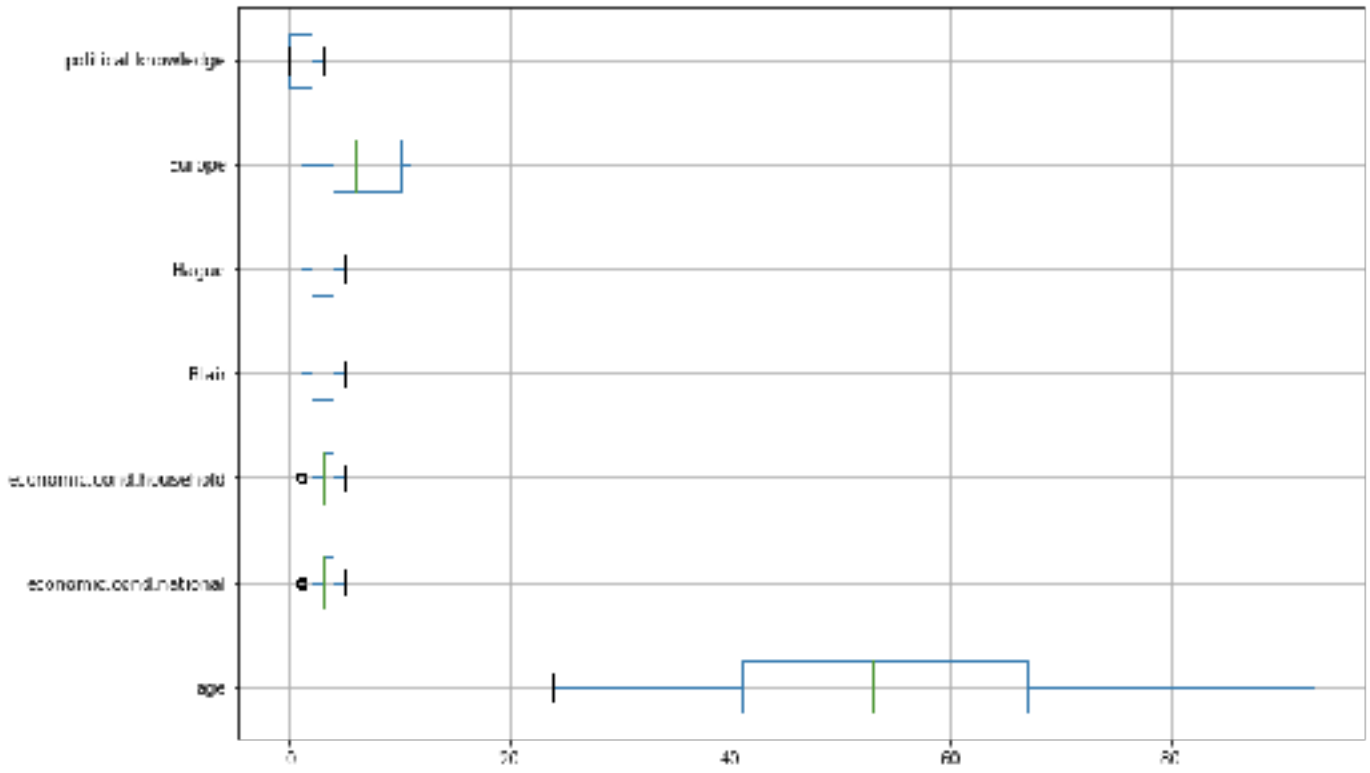
The dataset was checked for null values using the isnull() function - no null values were found.

1.2 Exploratory data analysis:

Univariate analysis:

Boxplots were used to check the spread of the numerical variables.

Figure: 1.1



Observations:

Figure below shows that ‘age’ variable shows far greater spread than the others because all other numerical variables are points belong to a rating scale.

For the categorical variables (‘gender’ and ‘vote’), count plots were used to check their distribution, as shown in the figures below:

Figure: 1.2

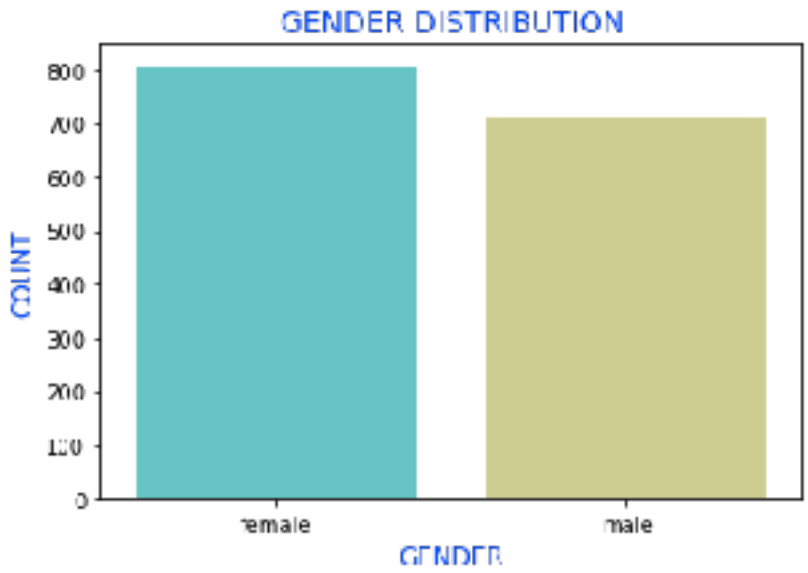
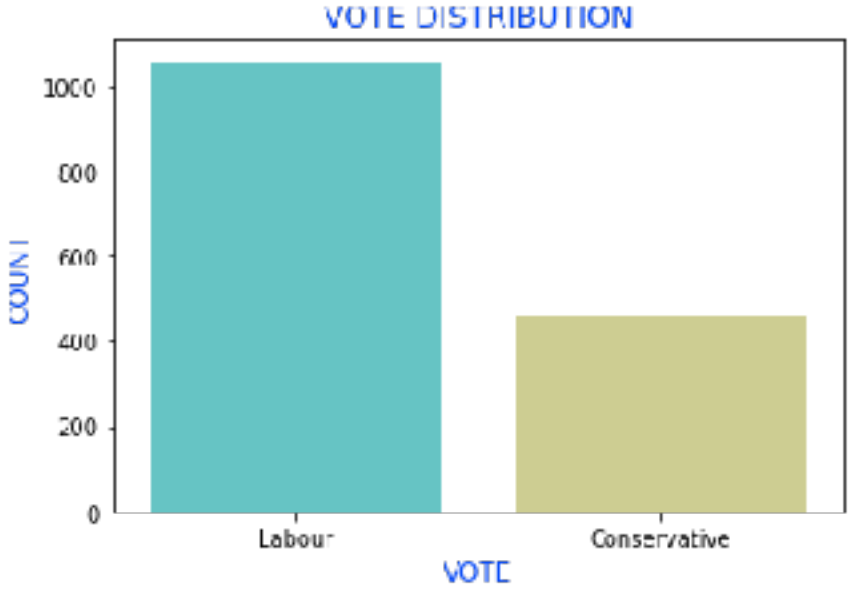


Figure: 1.3



Observations:

- Although there are more females in the voter base, the gender distribution is not very unequal.
- The distribution of votes, however, shows high inequality. The Conservative party got less than half the votes that the Labour Party has got.

Bivariate analysis:

A pair plot was used to check correlation between variables, as shown below:

Figure: 1.4

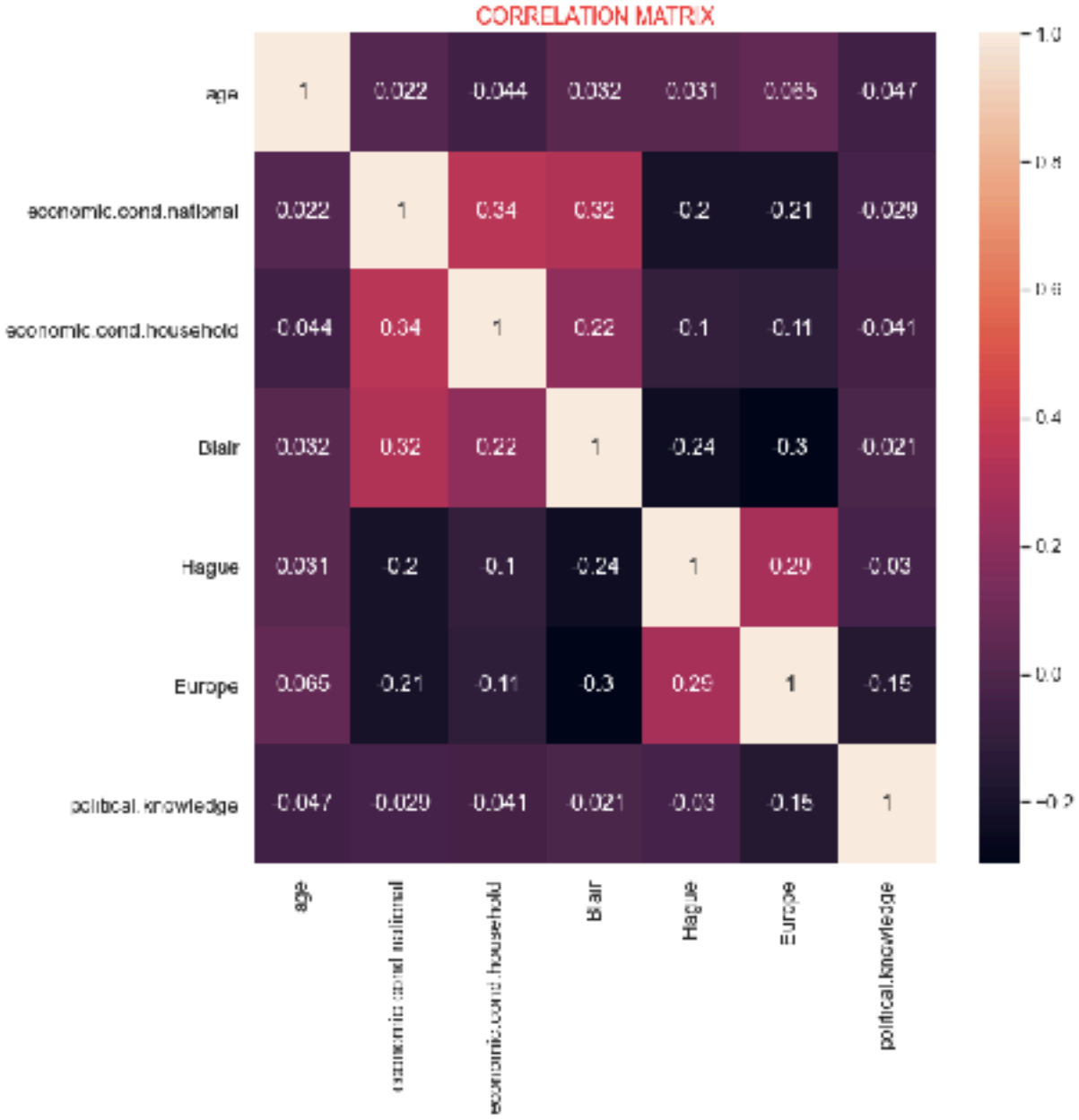
Observations:

- The heat map clearly shows that there is not collinearity among the variables in the dataset.

- Also, ‘age’ is the only variable that shows a somewhat normal distribution, if not a perfect one.

Further, correlation among variables was checked using a heatmap, so that a fair idea of the interdependence of variables could be derived.

Figure: 1.5

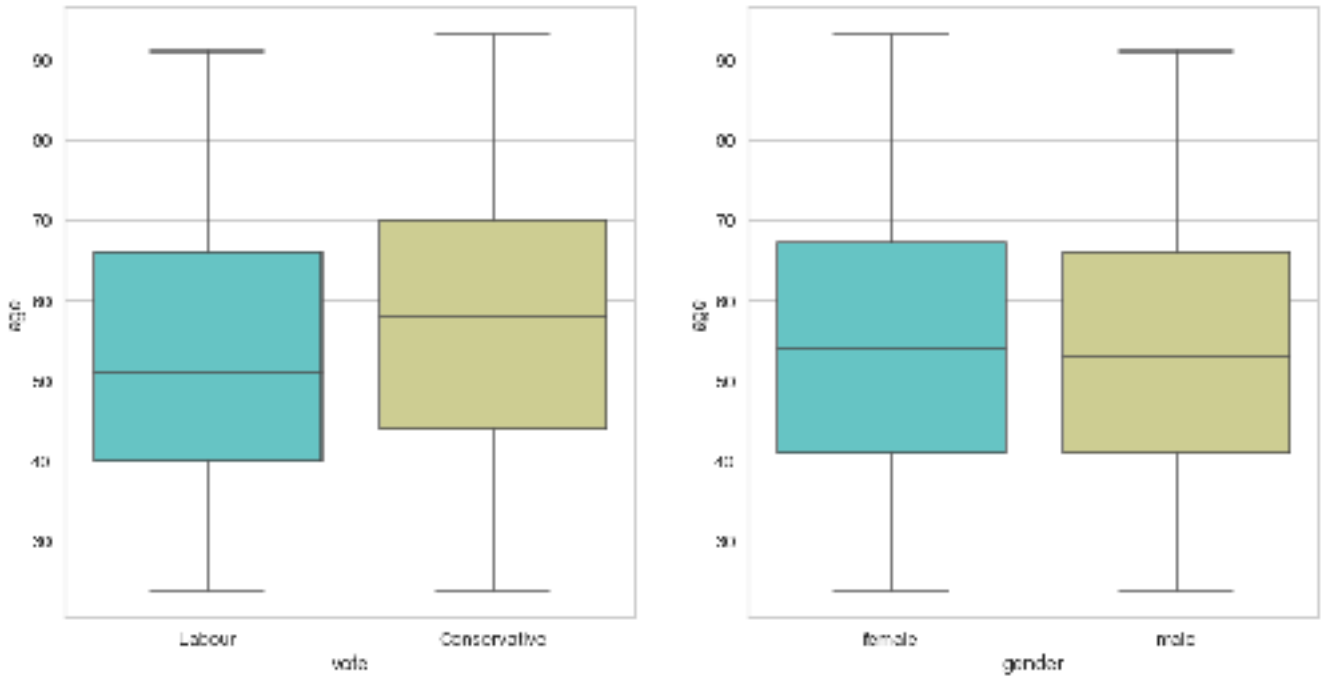


Observations:

- The heat map decisively shows that there is very little correlation between the variables in this dataset, meaning that the independent variables do not have interdependence amounts themselves

To have greater understanding of nature of the variables, box plots were used to distribution of ‘vote’ and ‘gender’ with respect to age. The figure below shows the result.

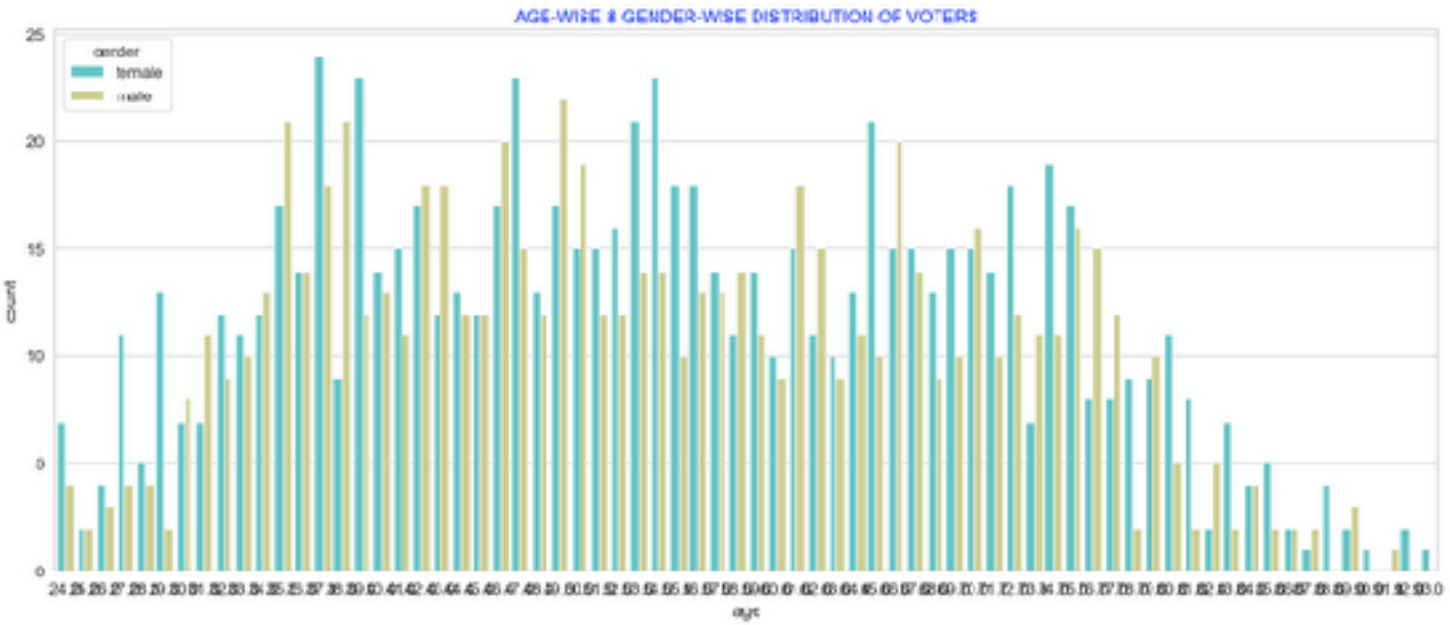
Figure: 1.6



Observations:

- The mean age of people who have voted for the Conservative party (approx. 57) is greater than the mean age of Labour Party supporters (approx. 52).
- Not only are female voters greater in number, but the mean age of the female voters is slightly more than the mean age of the male voters.

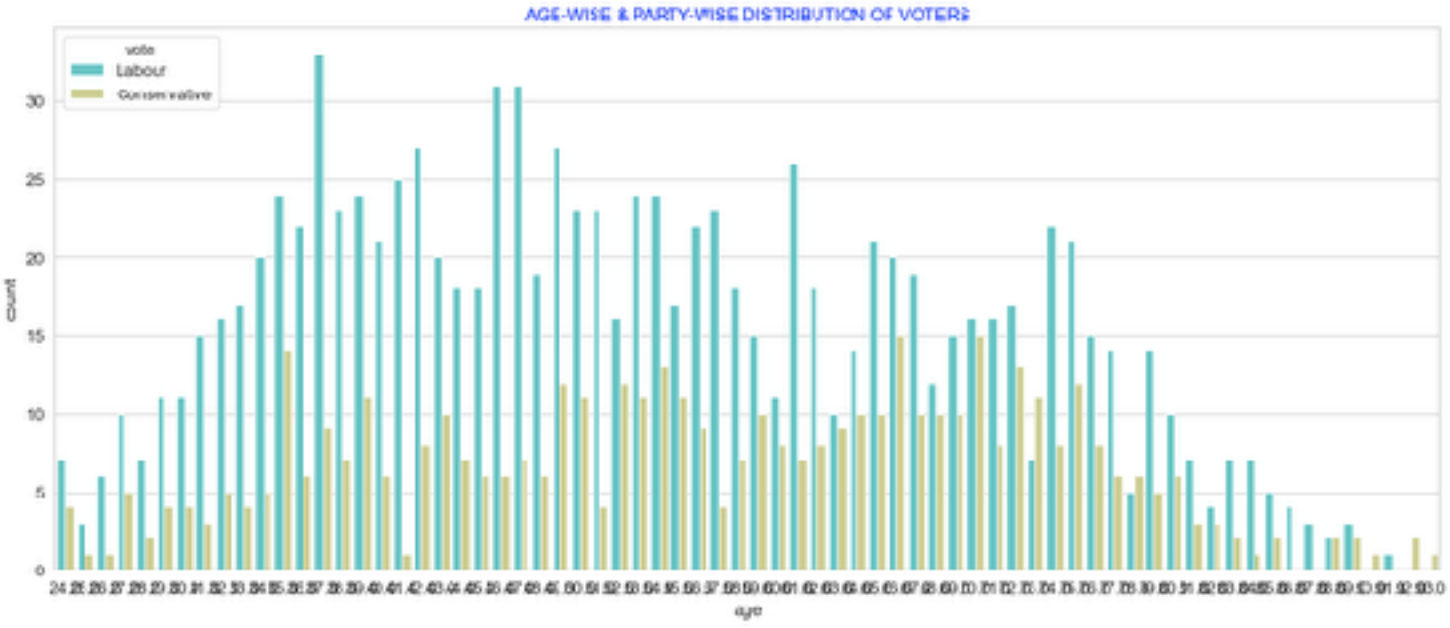
Figure: 1.7 : Count plot of ‘age’ and ‘gender’-wise distribution voters:



Observations:

- It is evident from the distribution that the highest number of votes come from people in their late 30s, but there is consistency in voting in people in their 50s, especially women.

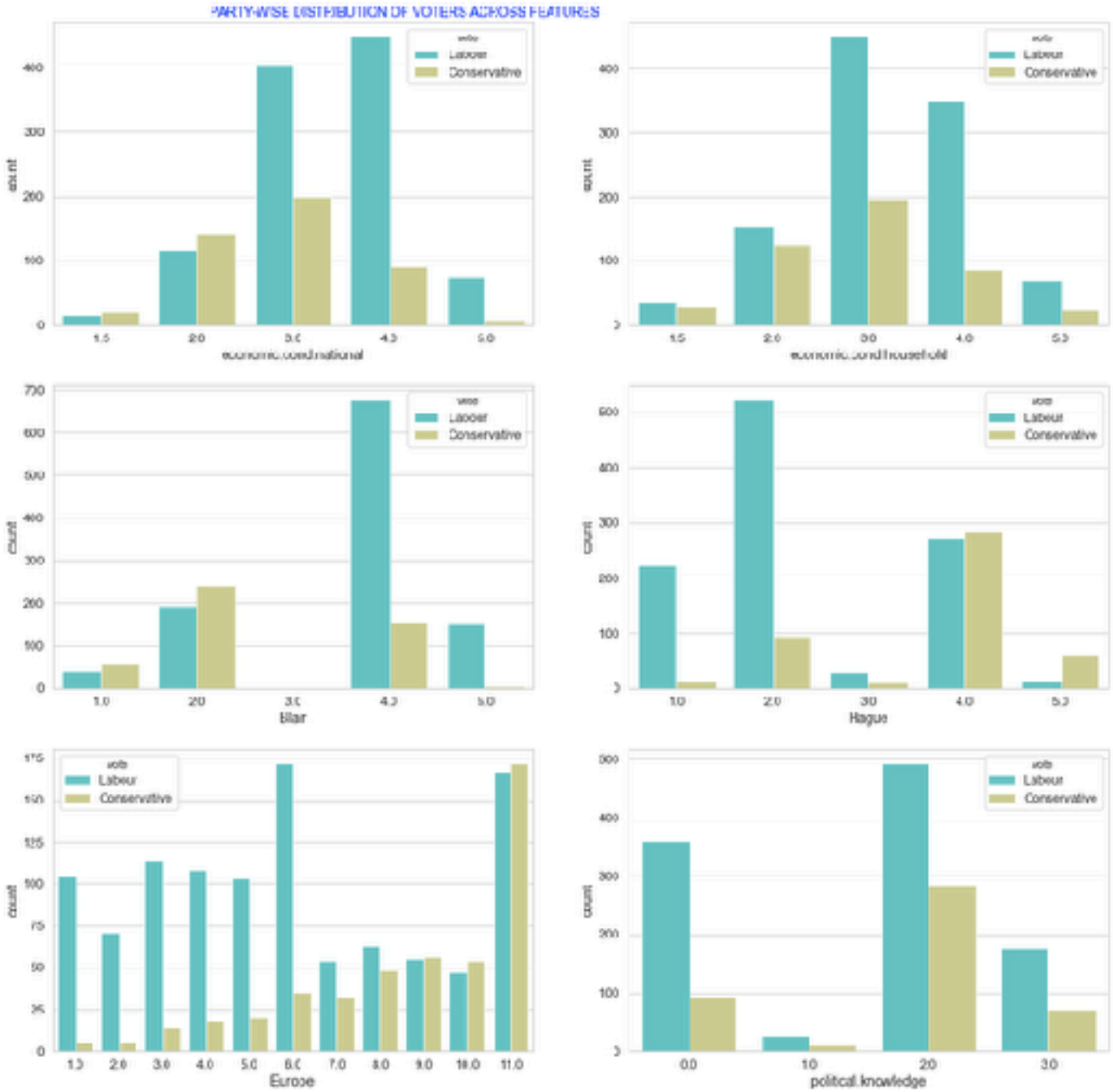
Figure: 1.8 : Count plot of ‘age’ and ‘party’-wise distribution of voters:



Observations:

- It is clear that across all ages, the Labour party has acquired more votes than the Conservative party (except at the very end of the age bracket - late 80s and early 90s).

Figure: 1.9: Party-wise distribution of voters across variables



Observations:

- Voters generally have a fairly positive viewpoint of the economic conditions prevailing in the country as well as at the individual level. Majority of them have rated these two features as 3 and 4 on a 5-point scale. This is true for supporters of both the parties.
- Labour Party supporters have a very positive viewpoint of the economic condition prevailing at the national level (majority have given this attribute a rating of 4 out of 5). Most Conservative party voters have given this attribute a ranking of 3.
- In terms of the economic condition of households, majority of both party supporters have given a ranking of 3 out of 5, demonstrating a rather moderate viewpoint about it.
- It is natural that majority of the Labour Party voters have given a very high ranking (4 out of 5) to Tony Blair, their party leader.
- Similarly, majority of the Conservative party voters have given a high ranking (4 out of 5) to William Hague, who is their party leader.
- It is interesting that a vast majority of Conservative party supporters are totally against the integration of UK with Europe (11 points on the ranking scale - the highest).
- However, an almost equal number of Labour Party supporters are divided on this idea (they have given 6 and 11 points on rating scale). 6 points means that they are slightly sceptic about the integration, and 11 means they are highly against it. Among the Labour Party voters, there is a consistent number of people who are in favour of the integration as well (rating 1, 3 , 4, 5 out of 11).
- With regard to the people's knowledge of their party's stand on the integration issue, voters are divided. A vast majority feels they know fairly well what their party is thinking in terms of the European integration (rating 2 out of 3), while a large number feel that they have absolutely no idea about how their party feels (rating 0 out of 3).

Multivariate analysis:

Crosstab of 'vote' v/s 'Europe' and 'age'

Table: 1.5

vote	Conservative	Labour	All
Europe			
1.00	61.40	61.34	61.34
2.00	62.00	51.82	52.61
3.00	58.86	53.73	54.29
4.00	55.44	49.27	50.15
5.00	55.75	48.76	49.89
6.00	58.60	51.20	52.45
7.00	52.59	46.37	48.69
8.00	53.42	50.41	51.71
9.00	53.27	51.85	52.57
10.00	54.69	55.21	54.93
11.00	59.87	58.30	59.10
All	56.84	53.11	54.24

Observations:

- The mean age of voters is higher for Conservative party (56.84 years), while it is slightly younger for Labour Party (53.11 years).
- The youngest age bracket for voters of both parties are very sceptic on the issue of the European integration (have given a 7 point rating).
- People in their 60s (from both parties) are very highly positive about the European integration (rating 1 out of 11).

Crosstab of 'vote' v/s 'gender' and 'age'

Table: 1.6

gender	female	male	All
vote			
Conservative	58.25	55.05	56.84
Labour	52.76	53.49	53.11
All	54.50	53.94	54.24

Crosstab of count of 'vote' v/s' 'gender'

Table: 1.7

gender	female	male	All
vote			
Conservative	257	203	460
Labour	551	506	1057
All	808	709	1517

Observations:

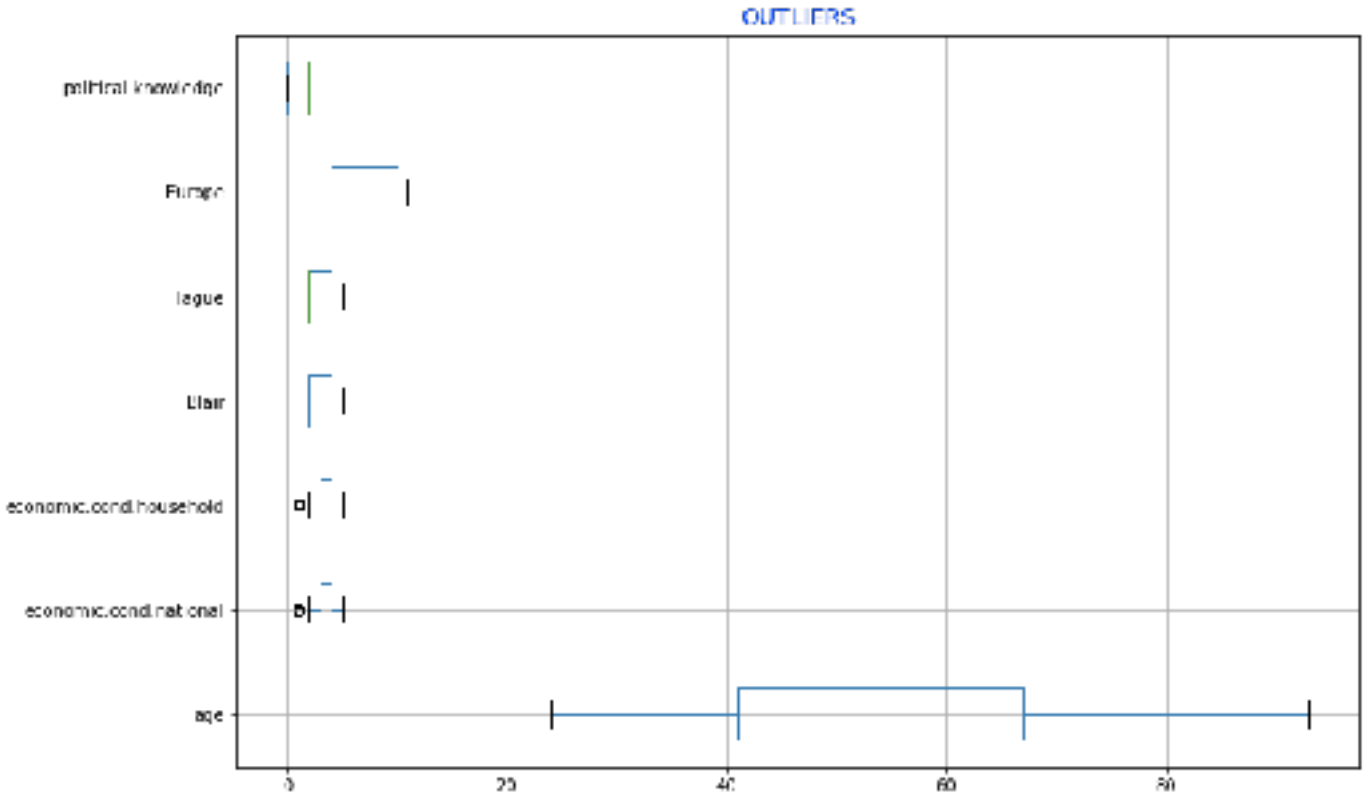
- The average age for voters of the Conservative party is slightly higher that voters of the Labour Party (53.11).
- Within the genders, older people vote for the Conservative party, while younger ones opt for the Labour Party.
- Out of the 1517 voters, 1057 have voted for the Labour Party, while only about 30% (460) have opted for the Conservative party.
- Labour Party seems to be the clear winner.

Outlier checking and treatment:

Outliers in the data were checked using box-plot in the Seaborn package, which revealed the presence of few outliers in the columns 'economic.cond.national' and 'economic.cond.household' (shown in figure below).

These outliers were treated using Inter-quartile range (IQR) method. It was imperative to treat the outliers since certain machine learning models used here are very sensitive to outliers.

Figure: 1.10



1.3. Data preprocessing:

Variable encoding for categorical variables (object type):

In order to convert object type variables to integer values, the columns ‘vote’ and ‘gender’ were converted to numerical codes using categorical codes. The following codes were used:

Table: 1.8

Variable	Ordinal codes
Vote	Labour’ : 0, ‘Conservative’ : 1
Gender	female’ : 0, ‘male’ : 1

Scaling of data:

The given dataset has a lot of variables with scaled data already (where observations are in the form of rating scales). The only column that does not have a rating scale value was ‘age’. However, the observations were neither too varied in scale nor magnitude, so it was decided that scaling of data was not necessary.

Data splitting for model building:

The dataset was split into training and test set in the ratio 70:30 before building of the machine learning models using the ‘train_test_split' function of the ‘sklearn.model_selection’ package.

Post data splitting, the shape of the training and test sets were as follows:

```
Shape of train dataset: (1061, 8) (1061, 1)
Shape of test dataset: (456, 8) (456, 1)
```

1.4. Logistic Regression and LDA:

Building the Logistic Regression (LR) model:

The model was built using the 'LogisticRegression()' function from the 'sklearn.linear_model' package. The parameters used for this model were:

- max_iterations = 10,000
- n_jobs = 2
- solver = 'newton-cg'

Building the Linear Discriminant Analysis (LDA) model:

This model was built using the 'LinearDiscriminantAnalysis()' function from the 'sklearn.discriminant_analysis' package, using default parameters (cut-off value = 0.5).

The model was later tuned using GridSearchCV, where cut-off values from 0.1 to 0.9 were tried and the cut-off yielding the highest Accuracy and F1 score was selected to build the tuned model. The cut-off yielding the best metrics was 0.3. This model, however, did not perform as well as the base model.

1.5. K-Nearest Neighbors (KNN) and Naive-Bayes' model:

Building the KNN model:

This model was built using the 'KNeighborsClassifier' function from the 'sklearn.KNeighbors' package.

The model was built on default parameters, like:

- n_neighbors = 5
- Weights = 'distance'

- The KNN model was also built on a scaled model to compare the results, to see if it enhanced the effectiveness of the model. However, it was observed that
- Later, the model was tuned using various values of k (number of neighboring points). Odd values of k between 1-19 were tried out and their MCE (miscalculation errors) were compared. The least MCE was derived by k=7, hence the tuned KNN model was built with the k value.
- The KNN model was also rebuilt using GridSearchCV with the following parameters:
  - algorithm = 'auto'
  - leaf_size = 30
  - n_neighbors = 10
  - Weights = 'distance'

Building the Gaussian Naive-Bayes' (NB) model:

This model was built using the 'GaussianNB' function from the 'sklearn.naive_bayes' package, using default parameters.

The model was rebuilt using GridSearchCV and cross validation =10 and np.logspace(0,-9,num=100).

Post tuning results yielded similar metrics as the base model.

1.6. Bagging and Boosting:

Bagging algorithm:

The 'BaggingClassifier' function from the 'sklearn.ensemble' package was used to build this ensemble learning model. The parameters were as follows:

- base_estimator = RandomForestClassifier()
- max_features = 7
- n_estimators = 40

AdaBoosting model:

The AdaBoosting Ensemble Learning model was built using the 'AdaBoostClassifier' from the 'sklearn.ensemble' package. The parameter used was n_estimators=30 and default base_estimator.

Gradient Boosting model:

The Gradient Boosting ensemble learning model was built with the following parameters:

- n_estimators = 70
- learning_rate = 0.1
- Tol = 0.0001

1.7. Performance metrics:

- In this dataset, both classes (voting for Labour Party as well as voting for Conservative party) has equal importance in our eyes.
- Hence, the best best performance metrics to evaluate the machine learning model would be Accuracy and F1 score, because it is equally importance here to predict the voter preference correctly, as well as keep the false predictions to a minimum (in order to have an accurate exit poll).

LR (Logistic Regression) Model evaluation:

To evaluate the LR model, various metrics were used like the AUC, ROC curve, Confusion matrix and the Classification report. The following ROC curve and Confusion matrix was derived:

Figure: 1.11

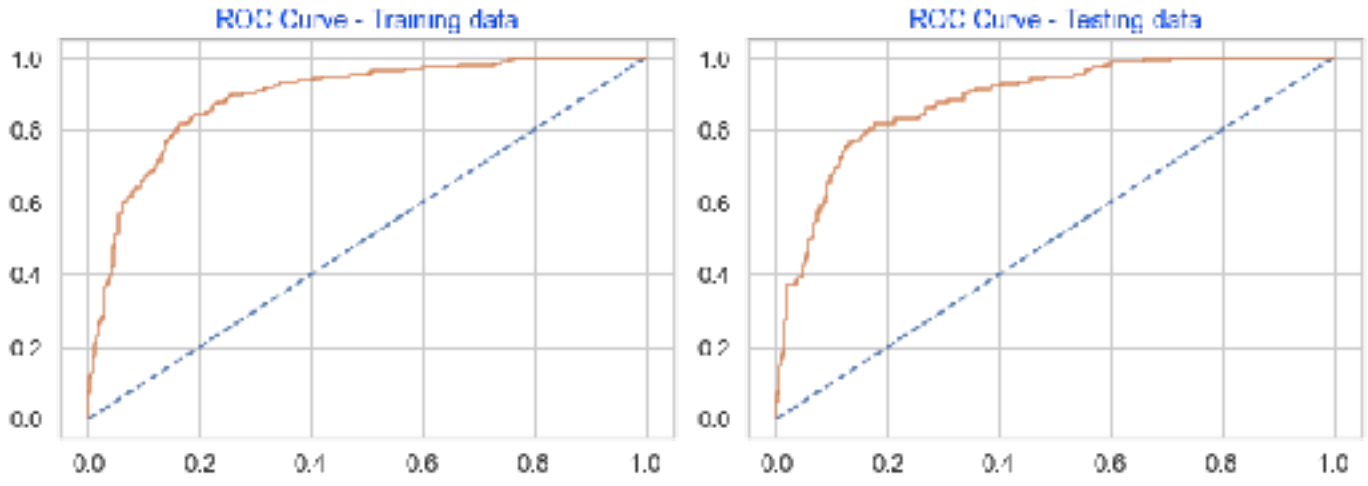


Figure: 1.12

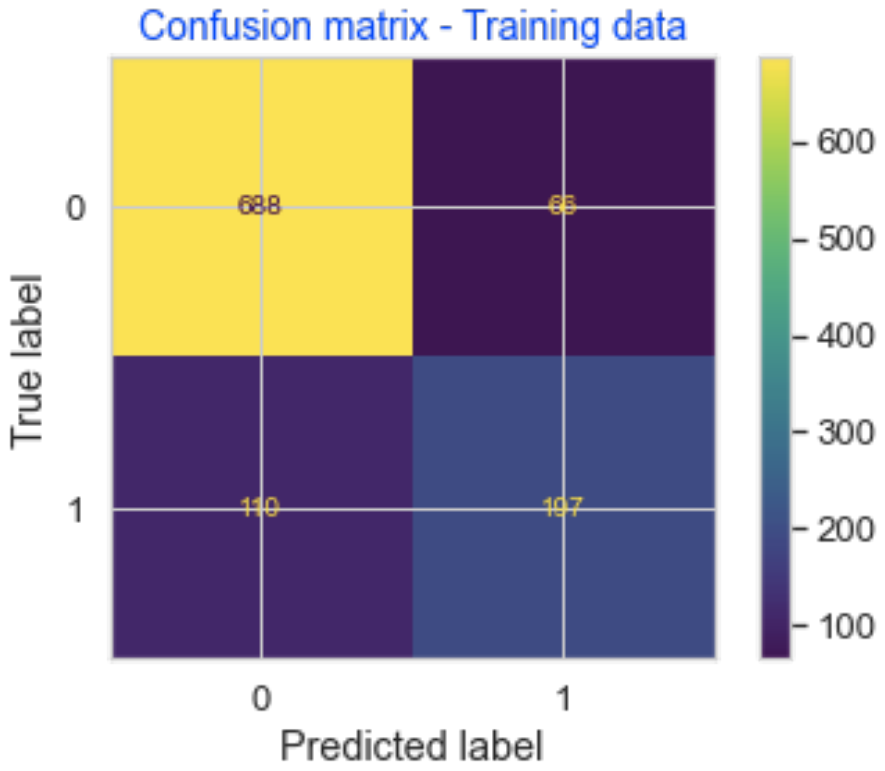
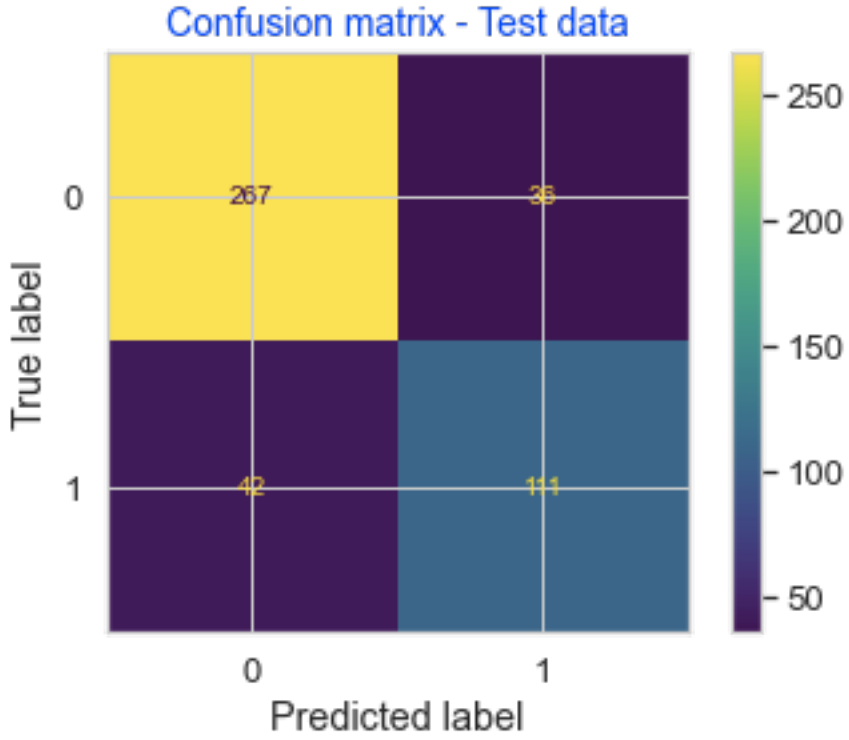


Figure: 1.13



Based on the Classification report for Logistic Regression model, the following metrics were derived, which helped to evaluate the model:

Table: 1.9

	Log-reg Train	Log-reg Test
AUC	0.89	0.88
Accuracy	0.83	0.83
Recall	0.64	0.73
Precision	0.75	0.76
F1 Score	0.69	0.74

Observations:

- We can see that model accuracy is very high for both training as well as test data - this is an indication of an effective model.
- Also, the metrics are more or less similar across train and test data, which means that there is no case of over-fitting or under-fitting. Model is performing almost similarly on both sets of data.

Later, the model was tuned using GridSearchCV and the following parameter:

- max_iterations = 15,000
- n_jobs = 5
- solver = 'sag'

The resulting tuned model, however, yielded the same results, as follows:

Table: 1.10

	Best_LR Train	Best_LR Test
AUC	0.89	0.88
Accuracy	0.84	0.83
Recall	0.65	0.73
Precision	0.75	0.76
F1 Score	0.69	0.74

Evaluating the LDA model:

To evaluate the LDA model, various metrics were used like the AUC, ROC curve, Confusion matrix and the Classification report. The following ROC curve and Confusion matrix was derived:

Figure: 1.13

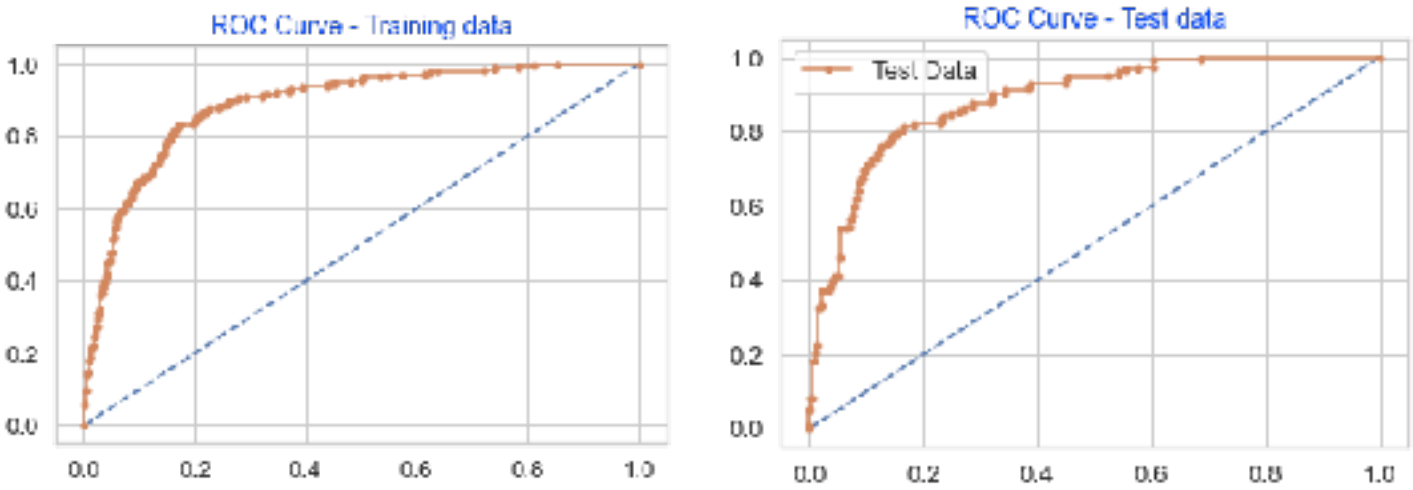


Figure: 1.14

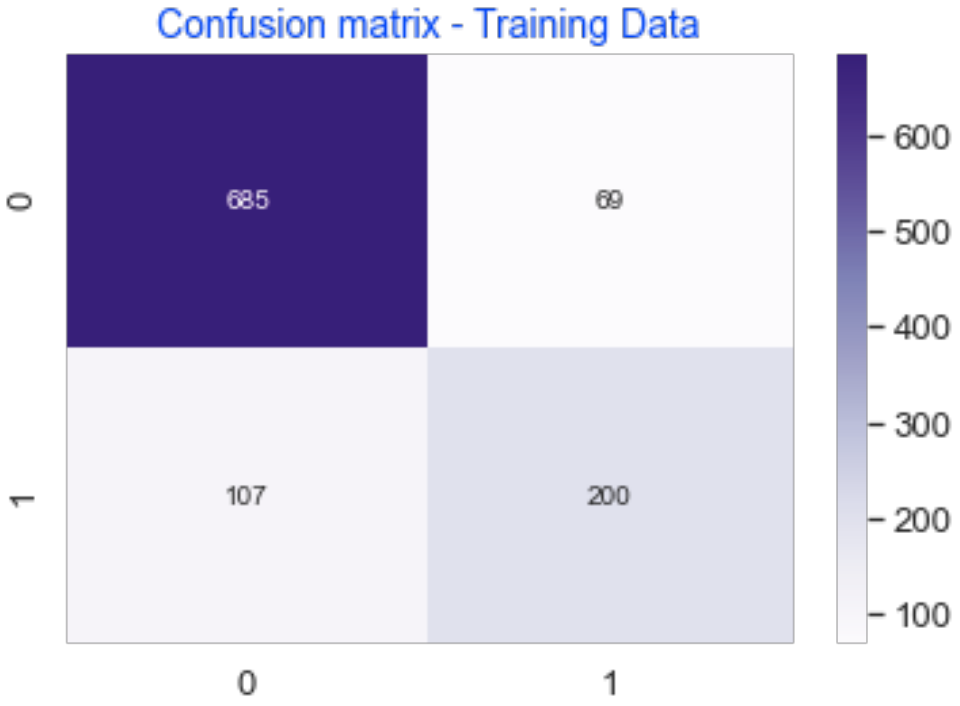
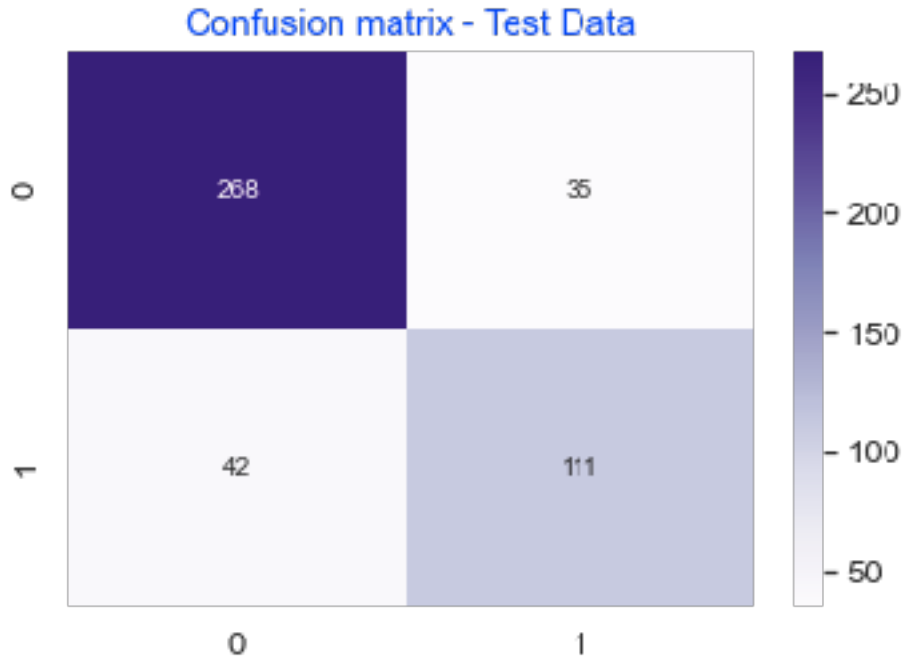


Figure: 1.15



Based on the Classification report, the following metrics were derived, which helped to evaluate the model:

Table: 1.11

	LDA Train	LDA Test
Accuracy	0.83	0.83
AUC	0.89	0.88
Recall	0.65	0.73
Precision	0.74	0.76
F1 Score	0.69	0.74

Observations:

- We can again see a very robust model here - metrics reveal high values.
- And there is evidently no case of over-fitting or under-fitting - metrics on both train and test data sets are almost similar, so model is behaving similarly on both sets of data.

The LDA model was later tuned using GridSearchCV, where cut-off values from 0.1 to 0.9 were tried and the cut-off yielding the highest Accuracy and F1 score was selected to build the tuned model. The cut-off yielding the best metrics was 0.3. This model, however, did not perform as well as the base model, as is evident in the table of metrics below:

Table: 1.12

	LDA Tuned Train	LDA Tuned Test
Accuracy	0.83	0.81
AUC	0.89	0.88
Recall	0.82	0.82
Precision	0.67	0.67
F1 Score	0.74	0.74

Comparing metrics in tables 1.11 and 1.12, we can see that recall has improved drastically post the GridSearch tuning. However, test accuracy has decreased and so has precision, so we can conclude that the base model was better performing.

Evaluating the KNN model:

The KNN model was evaluated based on the following metrics:

Table: 1.13

	KNN Train	KNN Test
Model score	1	0.81
Recall	1	0.65
Precision	1	0.76
F1 Score	1	0.7

We can see that the model is an overfitting model, doing exceeding well on the training data, while performing lower on the test data.

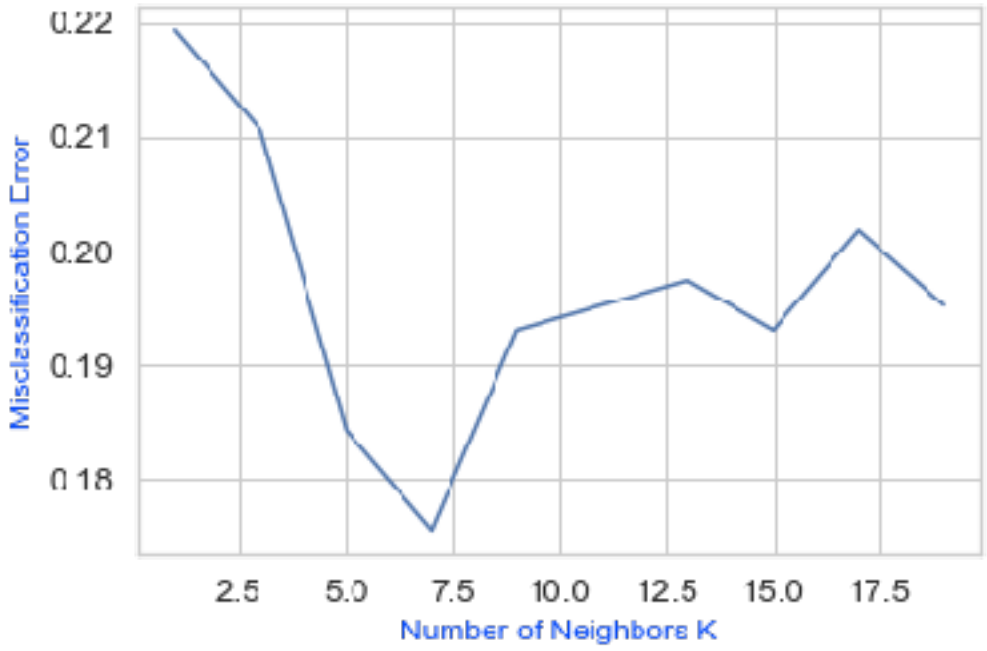
Further, training data was scaled to see if it impacted the performance of the model, but the results were not very different, as shown below:

Table: 1.14

	Scaled-KNN Train	Scaled-KNN Test
Model score	1	0.82
Recall	1	0.69
Precision	1	0.76
F1 Score	1	0.73

Tuning was done using various values of K (number of neighbors) and the best model was built on the k-value that gave least MCE (misclassification error). Here, the k-value with least MCE was k=7, as shown in graph below:

Figure: 1.16



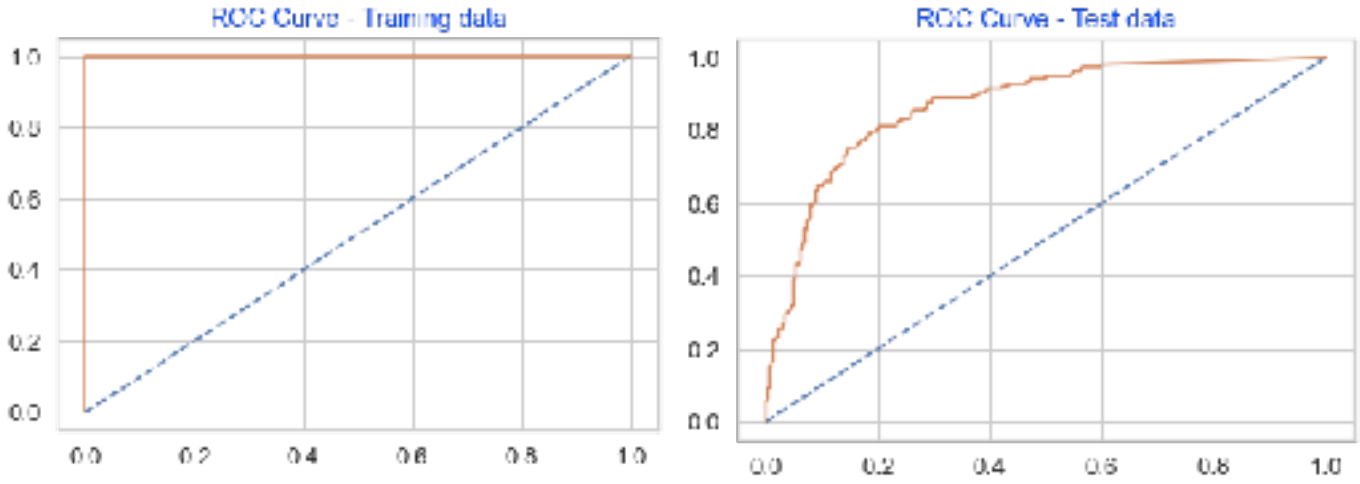
The performance metrics for the tuned KNN model with k=7, is as given below. This tuned model seems like a robust model with good accuracy and F1 score. There is also no case of over or under fitting here, unlike the earlier KNN models.

Table: 1.14

	Tuned_KNN Train	Tuned_KNN Test
Model score	0.85	0.82
Recall	0.66	0.65
Precision	0.78	0.79
F1 Score	0.72	0.71

The KNN model was also tuned using GridSearchCV also gave an overfitting model, as shown in the ROC curves below, hence it was not considered:

Figure: 1.17



Evaluating the Gaussian Naive Bayes (GNB) model:

This model was evaluated based on the following metrics:

Table: 1.15

	GNB Train	GNB Test
Model score	0.83	0.82
Recall	0.69	0.73
Precision	0.72	0.74
F1 Score	0.71	0.73

Using GridSearchCV, the GNB model was tuned with cross validation = 10, and the resulting metrics are given below.

Table: 1.16

	Best_GNB Train	Best_GNB Test
Model score	0.84	0.82
Recall	0.68	0.73
Precision	0.73	0.74
F1 Score	0.71	0.74

Both models have performed more or less the same, so we can select the one which has more consistent performance on both training and test set, i.e. the tuned model (table 1.16).

Evaluating the Ensemble learning- Bagging model:

The bagging model using RandomForest as base estimator yielded good results, as shown by the confusion matrix below:

Figure: 1.18

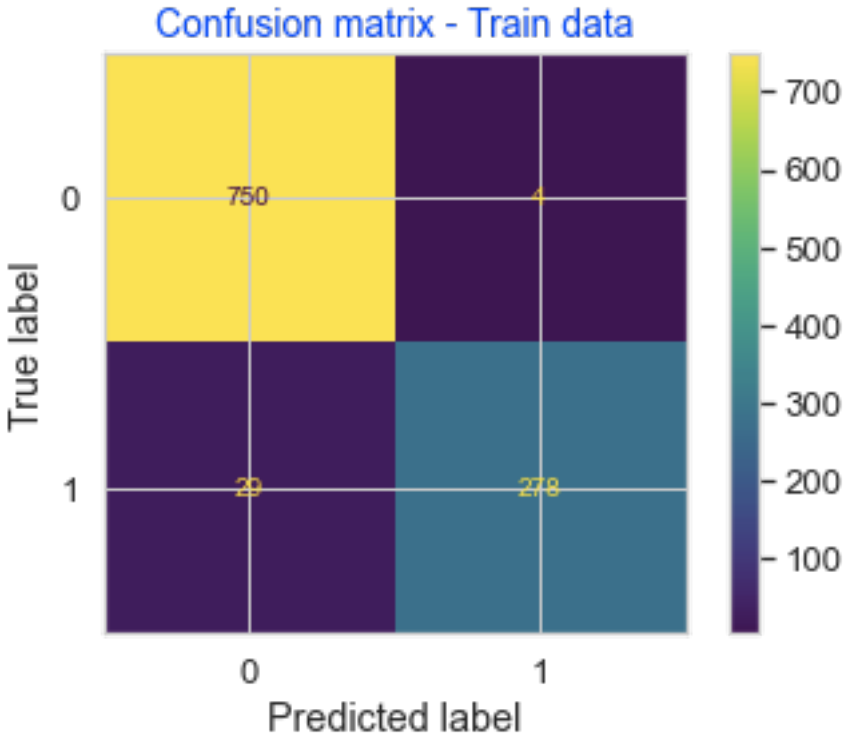
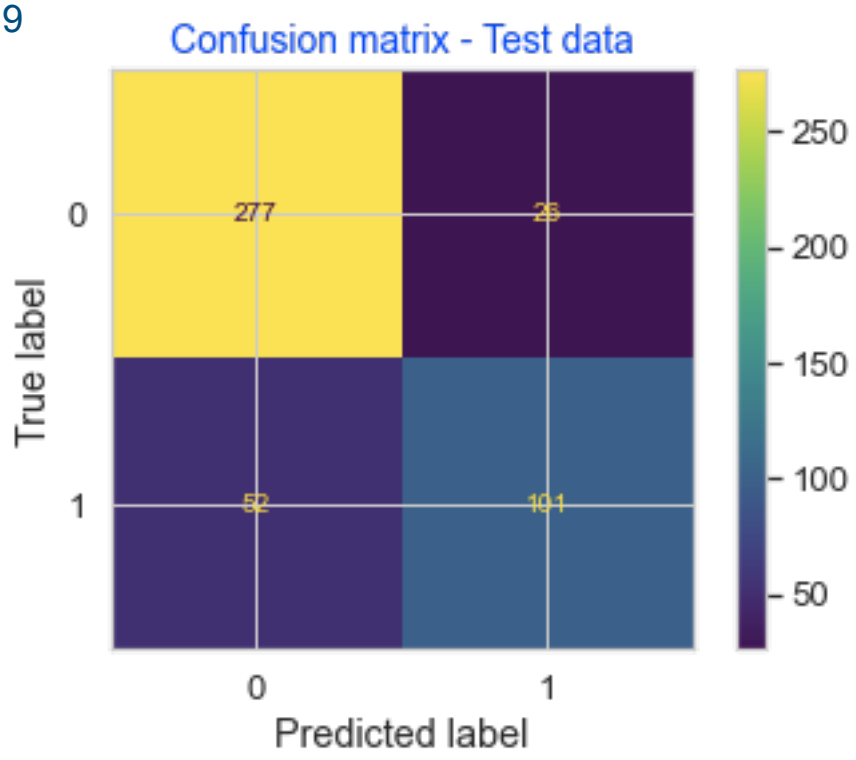


Figure: 1.19



The performance metrics for the Bagging model is given below. This model is slightly over fitting, as it is performing exceedingly well on the training data but performing lesser on the testing data.

Table: 1.17

	Bagging Train	Bagging Test
Model score	0.97	0.83
Recall	0.91	0.66
Precision	0.97	0.8
F1 Score	0.94	0.72

Evaluating the Ensemble learning- AdaBoosting model:

The confusion matrix for this model on training and test sets are given below:

Figure: 1.20

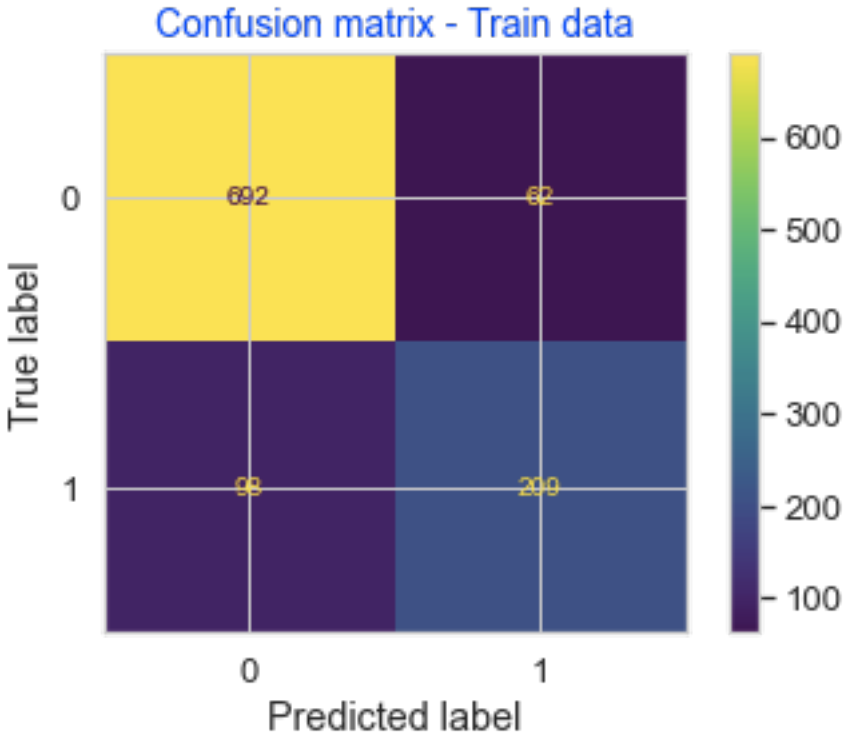
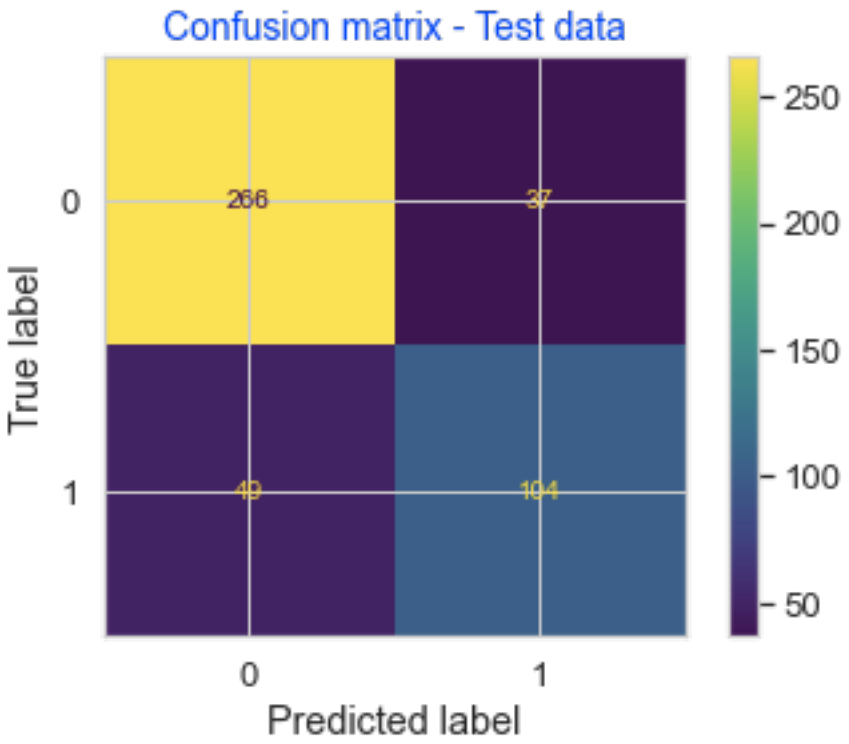


Figure: 1.21



The performance metrics for the AdaBoosting model is given below. This is a robust model, with very similar performance between train and test data sets, and high enough accuracy and F1 scores.

Table: 1.18

	Ada_Boost Train	Ada_Boost Test
Model score	0.85	0.81
Recall	0.68	0.68
Precision	0.77	0.74
F1 Score	0.72	0.71

Evaluating the Ensemble learning-Gradient Boosting model:

This ensemble learning model gave fairly high performance metrics, as given below:

Figure: 1.22

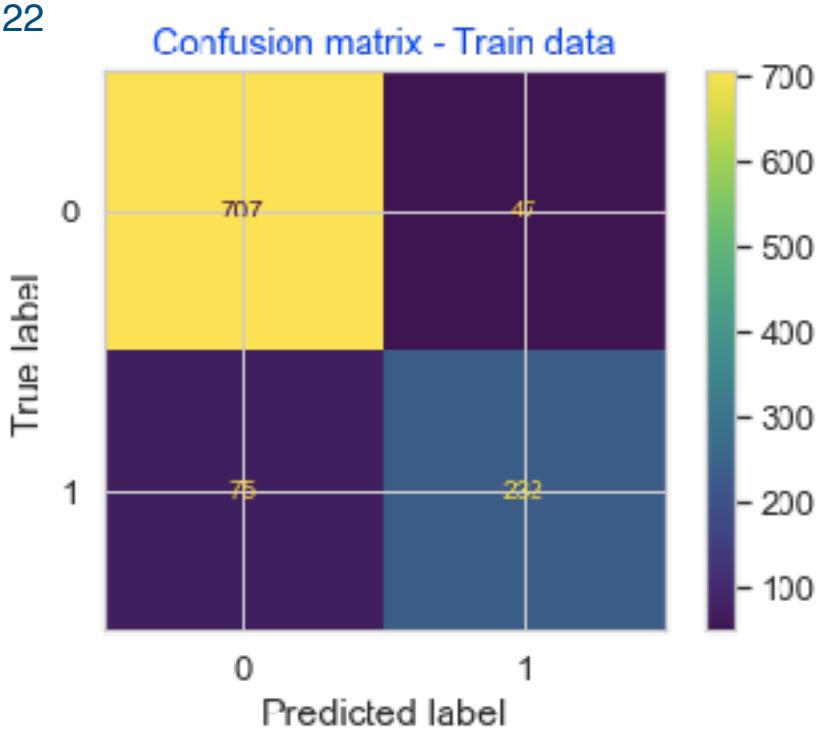
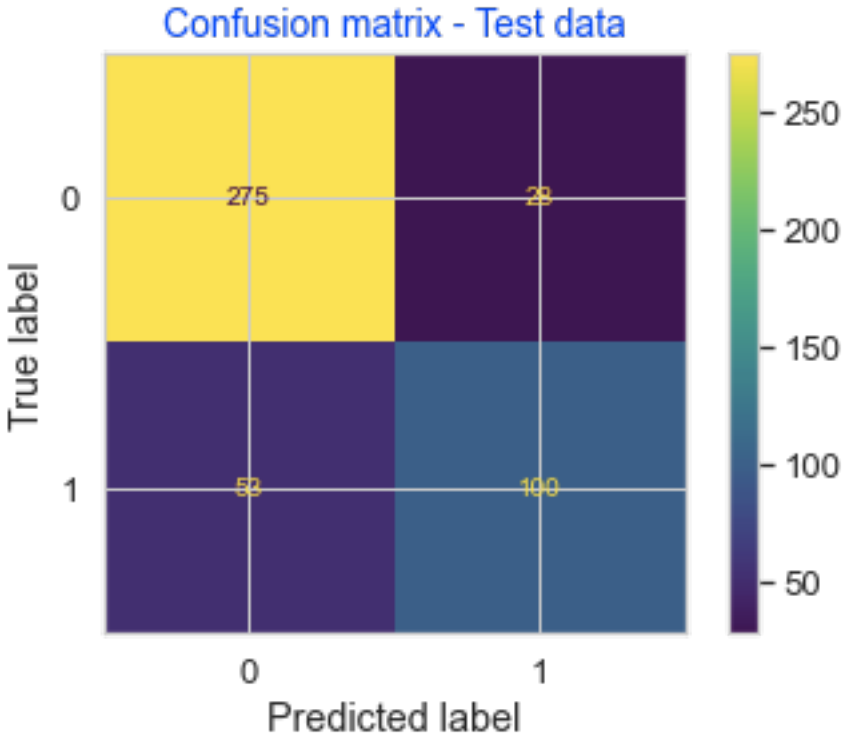


Figure: 1.23



The performance metrics for the Gradient Boosting model is given below. This is also a model performing well with accuracy of 82% and F1score of 71%.

Table: 1.18

	Grad_Boost Train	Grad_Boost Test
Model score	0.89	0.82
Recall	0.76	0.65
Precision	0.83	0.78
F1 Score	0.79	0.71

Best model:

Based on the performance metrics of the various models, we can say that all models have performed well, with accuracies more than 80%. However, it is evident that the Logistic Regression model (tuned with GridSearchCV) is the best of all. The metrics of this model are given below:

	Best_LR Train	Best_LR Test
AUC	0.89	0.88
Accuracy	0.84	0.83
Recall	0.65	0.73
Precision	0.75	0.76
F1 Score	0.69	0.74

1.8. Insights

- The analysis of the election data has given some very meaningful insights, that can determine voter behaviour.
- The younger section of the voters have favored the Labour party, so the Conservative party needs to bring in more policies that favor the young, like more employment opportunities, educational subsidies, etc.
- Similarly, the Labour Party could bring in more policies for the betterment of the elderly voters, like social security benefits, health care, etc. to attract votes from the older population.
- The women are a very large chunk of the voter base. So both parties should bring in more women friendly changes and policies that can benefit women, like child-care benefits, equitable income policies for women, etc.
- In general, the outlook of voters is somewhat negative towards the European integration, hence both parties must try to downplay the integration issue. More open and clear communication about the European integration issue should be targeted towards the voters.
- A majority of the voters are also not clear what the stand of the political parties is about the European integration issue. Each party could issue a statement / hold a press conference to clarify their stand on the issue and bring clarity to their voter base. This will help them connect better with the voters.
- Blair and Hague both need to establish better rapport with the elderly voter. Blair has a good enough image in the eyes of the Conservative supporters, but Hague needs to work on his public image to boost votes for the Conservative party.

- The outlook of the voters on national economic condition is above average, while they have a lesser impression about the economic condition of individual households. Both parties need to work in this direction. They can bring in legislation and policies that will benefit people financially and improve their life money-wise, like bringing down inflation, reducing loan rates, making utilities and household expenses less costly, making education more affordable, etc. This will give them more confidence on the parties.

PROBLEM 2

Speech Analysis

Executive Summary:

The task involves analyzing 3 speeches by former US Presidents. The three speeches are part of the inaugural corpora from nltk in Python. The three speeches are the following:

1. President Franklin D. Roosevelt in 1941 ('1941-Roosevelt.txt')
2. President John F. Kennedy in 1961 ('1961-Kennedy.txt')
3. President Richard Nixon in 1973 ('1973-Nixon.txt')

2.1 Character, word, sentence count:

- For the purpose of finding the count of characters, words and sentences in the speeches, firstly all the text files were combined into and written to a new text file named 'newfile.txt', saved in the 'inaugural' folder.
- The first few lines of the text file is shared below:

————————

```
On each national day of inauguration since 1789, the
people have renewed their sense of dedication to the
United States.

In Washington's day the task of the people was to
create and weld together a nation.

In Lincoln's day the task of the people was to preserve
that Nation from disruption from within.
```

```
In this day the task of the people is to save that
Nation and its institutions from disruption from
without.

To us there has come a time, in the midst of swift
happenings, to pause for a moment and take stock -- to
recall what our place in history has been, and to
rediscover what we are and what we may be.
                        ————————
```

- The function ‘len()' was used to derive the number of characters, words and sentences in the three documents. The results are given below:

```
Total number of characters in the documents: 25183
Total number of words in the documents: 4569
Total number of sentences in the documents: 188
```

2.2 Removal of stopwords:

- Stopwords are commonly used words in the English language that do not really make an impact on the analysis of the text like ‘when’, ‘you’, ‘if’, etc. For text analysis, we use the list ‘stopwords.words(‘english’)’ containing the list of common stop words used in the English language.
- Stopword removal is an important preprocessing step in the analysis of text.
- Stopword removal was done using the ‘stopwords’ function from ‘nltk.corpus’ package.

- A prerequisite of stop word removal was tokenization of the text, i.e., transforming the text file into a series of tokens (words) separated by commas. To do this, the ‘word_tokenize()' function from ‘nltk.tokenize' package was utilized.

- The first few lines of the text file (post stopword removal) is shared below:

```
['On', 'national', 'day', 'inauguration', 'since',
'1789', 'people', 'renewed', 'sense', 'dedication',
'United', 'States', 'In', 'Washington', "'s", 'day',
'task', 'people', 'create', 'weld', 'together',
'nation', 'In', 'Lincoln', "'s", 'day', 'task',
'people', 'preserve', 'Nation', 'disruption',
'within', 'In', 'day', 'task', 'people', 'save',
'Nation', 'institutions', 'disruption', 'without',
'To', 'us', 'come', 'time', 'midst', 'swift',
'happenings',
```

- Later in the analysis, the stopword.extend() was used to remove a few additional words that did not seem essential to the analysis and meaning of the text.

2.3 Word frequencies:

The task involved finding the 3 most frequently occurring words (after stop word removal) in the 3 speeches individually. The ‘FreqDist’ function from the ‘nltk.probability' package was used.

The results were as follows:

```
--------

The 3 most frequently used words in President
Roosevelt's speech are:

      nation    12
      know      10
      spirit     9



The 3 most frequently used words in President
Kennedy's speech are:

      let       16
      us        12
      world      8


The 3 most frequently used words in President
Nixon's speech are:

      us          26
      let         22
      america     21
                  --------
```

2.4 Word clouds:

- Word clouds are the visual presentation of the words in a text that are arrange in randomized order.
- The ‘WordCloud()' function from the ‘wordcloud' package was used.