Data Analysis to Predict Election Results and Text Analysis of the Speeches of the Presidents of the USA

Name: Aishwariya Hariharan
PGP-DSBA Online September’ 23
Date: 17/03/2024

Contents

List of Figures

List of Tables

Problem 1

Business Context

CNBE, a prominent news channel, is gearing up to provide insightful coverage of recent UK 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

The data contains different attributes of voters and their opinions. The detailed data dictionary is given below.

Data Dictionary

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.

Sample of the Dataset:

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

Table 1: Sample dataset

Data Types

```
 #   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
dtypes: int64(8), object(2)
memory usage: 119.3+ KB
```

Figure 1: Data types

- Total number of rows = 1525
- Total number of columns = 10
    - 8 columns of integer type
    - 2 columns of object type
- From the above results, we can see that there is no missing value present in the dataset.
- The data was also checked for duplicate rows. There are no duplicate rows in the dataset.

Statistical Summary of the Dataset

	Unnamed: 0	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525.00000	1525.00000	1525.00000	1525.00000	1525.00000	1525.00000	1525.00000	1525.00000
mean	763.00000	54.18230	3.24590	3.14033	3.33443	2.74689	6.72852	1.54230
std	440.37389	15.71121	0.88097	0.92995	1.17482	1.23070	3.29754	1.08331
min	1.00000	24.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.00000
25%	382.00000	41.00000	3.00000	3.00000	2.00000	2.00000	4.00000	0.00000
50%	763.00000	53.00000	3.00000	3.00000	4.00000	2.00000	6.00000	2.00000
75%	1144.00000	67.00000	4.00000	4.00000	4.00000	4.00000	10.00000	2.00000
max	1525.00000	93.00000	5.00000	5.00000	5.00000	5.00000	11.00000	3.00000

Table 2: Statistical summary

Exploratory Data Analysis

Univariate Analysis

- **Distribution of the age of voters**

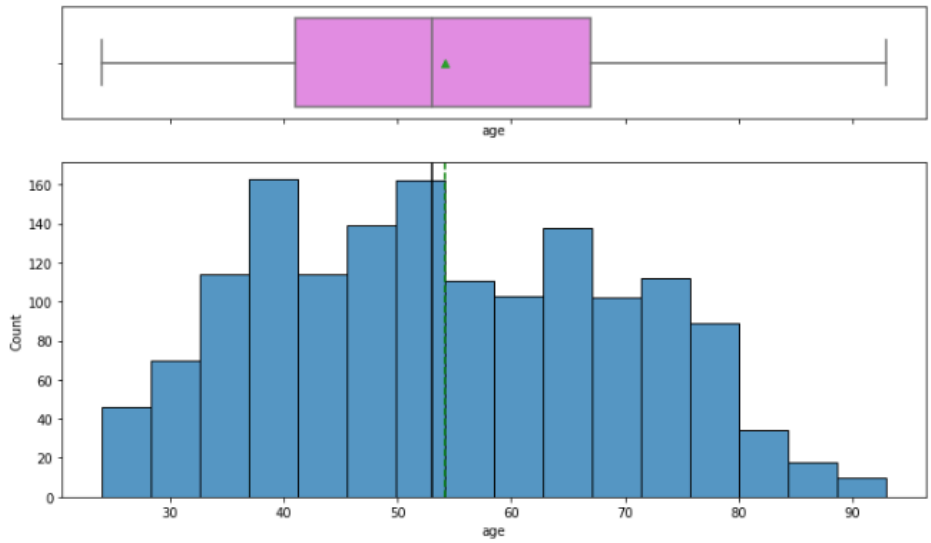


Figure 2: Distribution of age of voters

➔ The average age of the voters is about 55 years.

- **Observation on vote cast**

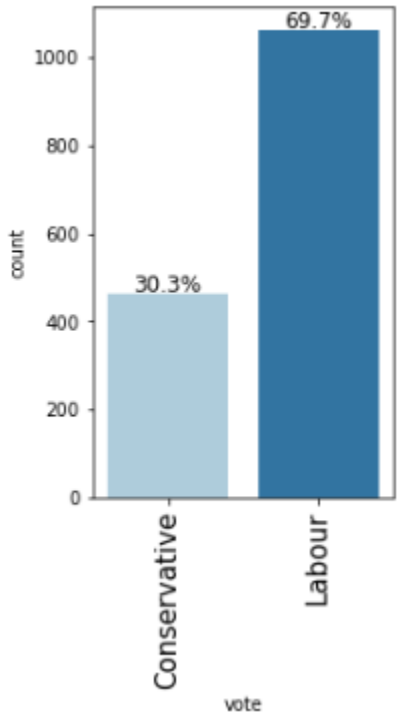


Figure 3: Vote distribution

➔ The majority of the votes, i.e. a whopping 70%, is in favour of the Labour Party.

- **Assessment of current national economic conditions**

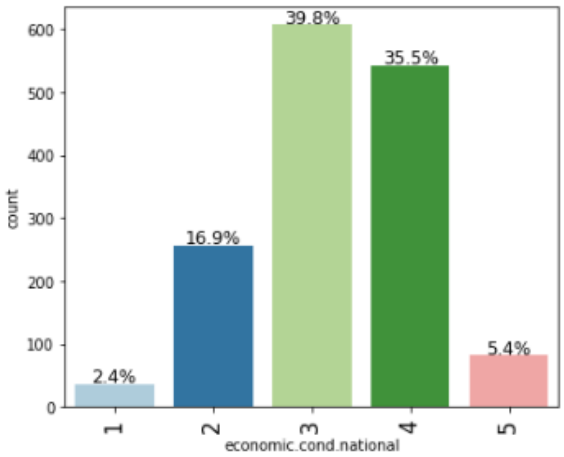


Figure 4: Assessment of current national economic conditions

➔ A rating of 1 indicates that the voters' opinion on the current national economic conditions is poor and a rating of 5 indicates that it is excellent.

➔ About 40% of the people surveyed have rated 3, indicating that they think the nation's current economic condition is average, neither too good nor too bad.
➔ This good be due to a lack of awareness, which made them choose the middle ground, or it could also be that the nation is doing just okay economically.

- **Assessment of the Labour Leader Tony Blair**

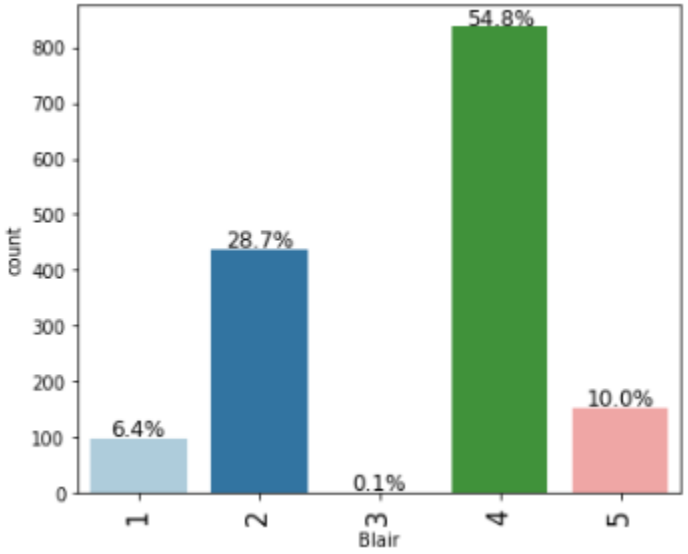


Figure 5: Assessment of the Labour Leader Tony Blair

➔ 55% of the voters have given Tony Blair a rating of 4
➔ It indicates that the voters have a lot of confidence in the leader

- **Assessment of the Conservative Leader William Hague**

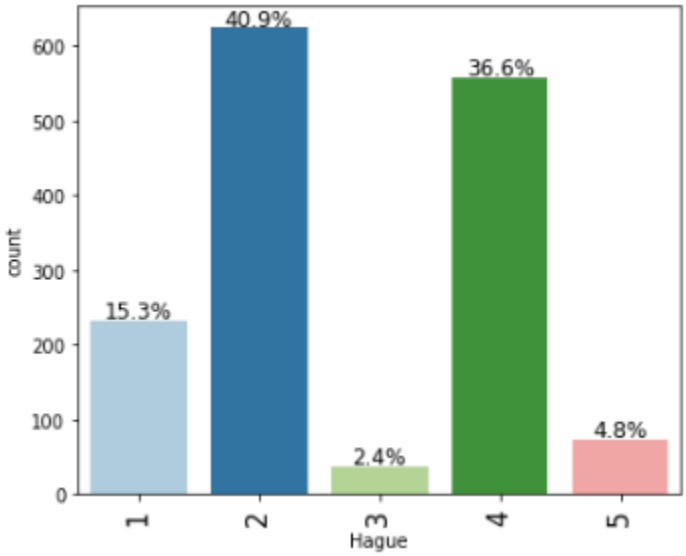


Figure 6: Assessment of the Conservative Leader William Hague

➔ 41% of the people have given Hague a rating of 2 and 36% a rating of 4.

➔ People seem to be highly divided in their opinions about the leader.

- **Voters' attitude towards Europs's integration**

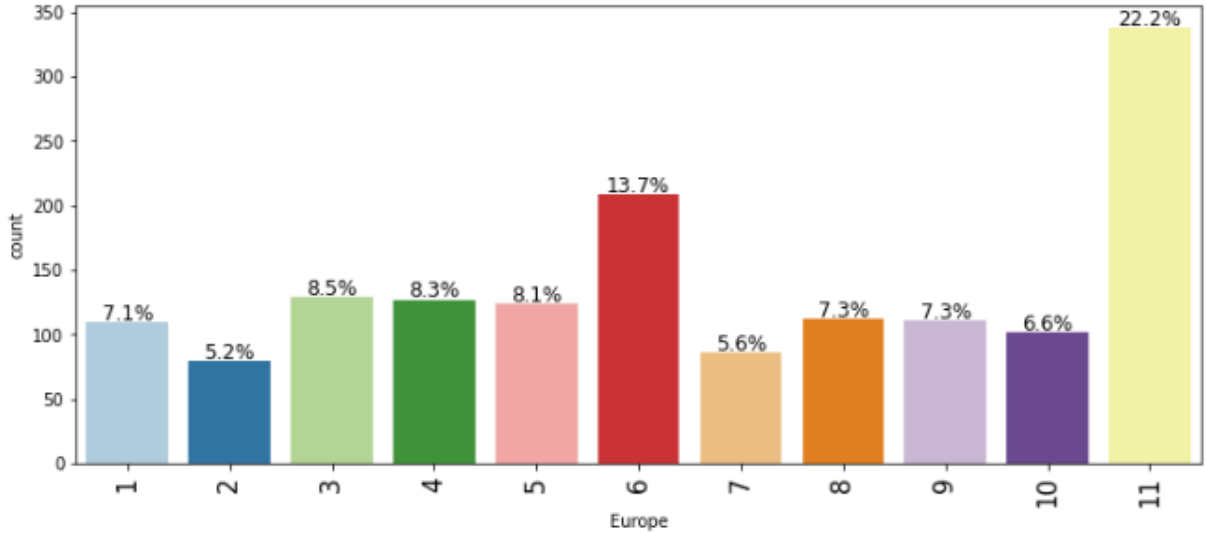


Figure 7: Voters' attitude towards Europs's integration

➔ The rating of 11 has the maximum number of votes, which implies a Eurosceptic sentiment.

- **Knowledge of parties' positions on European integration**

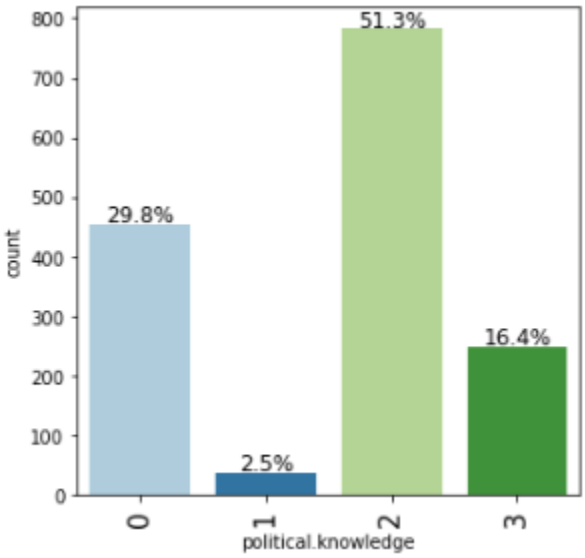


Figure 8: Knowledge of parties' positions on European integration

➔ A majority rating of 2 indicates reasonable knowledge of the parties' positions on European integration

Multivariate Analysis

- **Correlation**

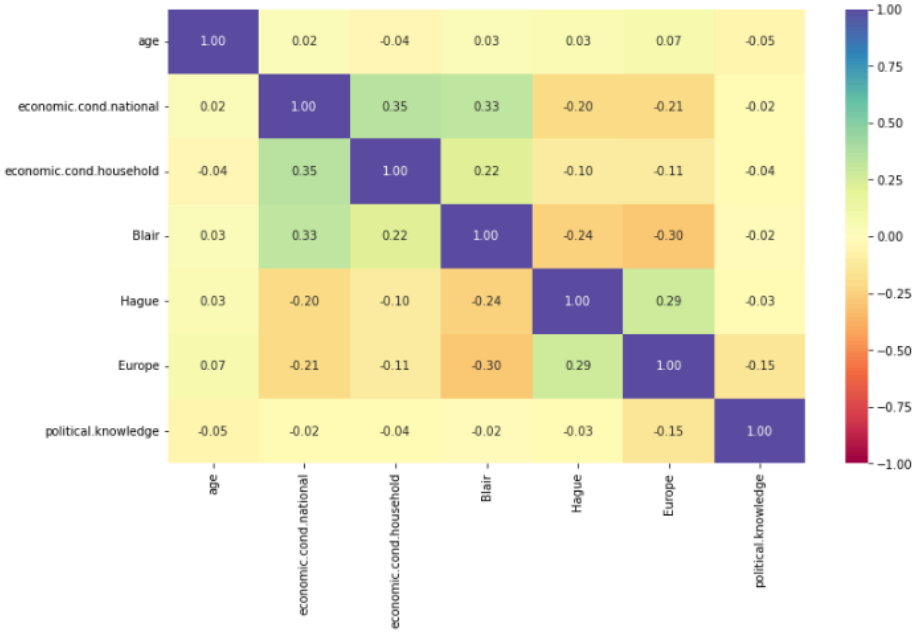
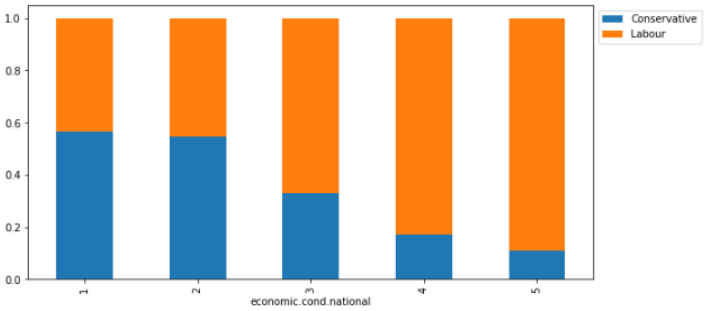


Figure 9: Correlation plot

➔ There's a high correlation between the economic conditions of the household and the nation and also between voters' confidence in Blair and the nation's economic conditions.

- **Assessment of national economic conditions vs Party that gets the vote**



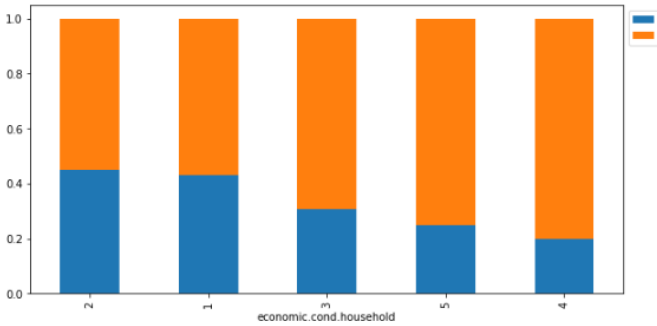
vote	Conservative	Labour	All
economic.cond.national			
All	462	1063	1525
3	200	407	607
2	140	117	257
4	92	450	542
1	21	16	37
5	9	73	82

Figure 10: Assessment of national economic conditions vs Party that gets the vote

➔ Of all the 82 people who have rated the national economic condition as 5, 73 are the voters in favour of the Labour Party. It is the opposite for the rating of 1.

➔ Obsv: The voters of Labour Party carry a positive view about the nation's economic condition.

- **Assessment of household economic conditions vs Party that gets the vote**

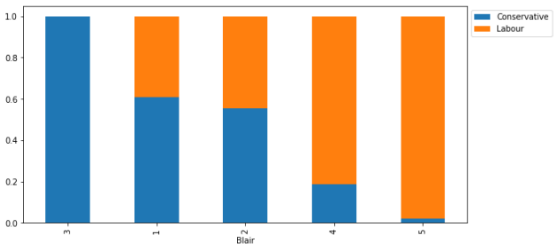


vote	Conservative	Labour	All
economic.cond.household			
All	462	1063	1525
3	198	450	648
2	126	154	280
4	87	353	440
1	28	37	65
5	23	69	92

Figure 11: Assessment of household economic conditions vs Party that gets the vote

➔ Voters of the Labour Party carry a positive view about the household economic conditions.

- **Assessment of the labour leader Blair vs Party that gets the vote**

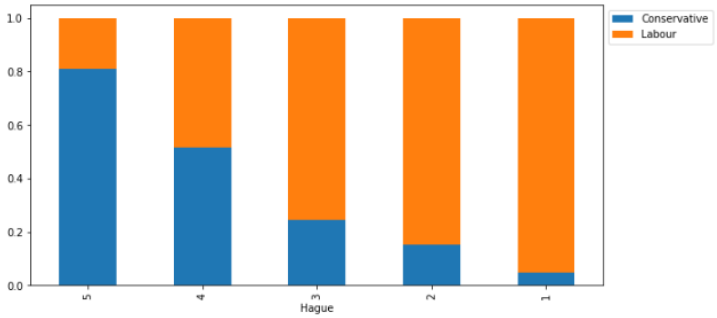


vote	Conservative	Labour	All
Blair			
All	462	1063	1525
2	242	196	438
4	157	679	836
1	59	38	97
5	3	150	153
3	1	0	1

Figure 12: Assessment of the labour leader Blair vs Party that gets the vote

➔ Clearly, people who have a good understanding (rating of 5) of the labour leader Blair have voted for the Labour Party and people who have little or no inclination towards Blair (rating of 1) have voted for the conservative party.

- **Assessment of the conservative leader Hague vs Party that gets the vote**



vote	Conservative	Labour	All
Hague			
All	462	1063	1525
4	287	271	558
2	96	528	624
5	59	14	73
1	11	222	233
3	9	28	37

Figure 13: Assessment of the conservative leader Hague vs Party that gets the vote

➔ Majority of the people who have voted for the conservative party have a good understanding (rating 5) of the party leader Hague.

- **Voters' attitude towards Europe's integration vs Party that gets the vote**

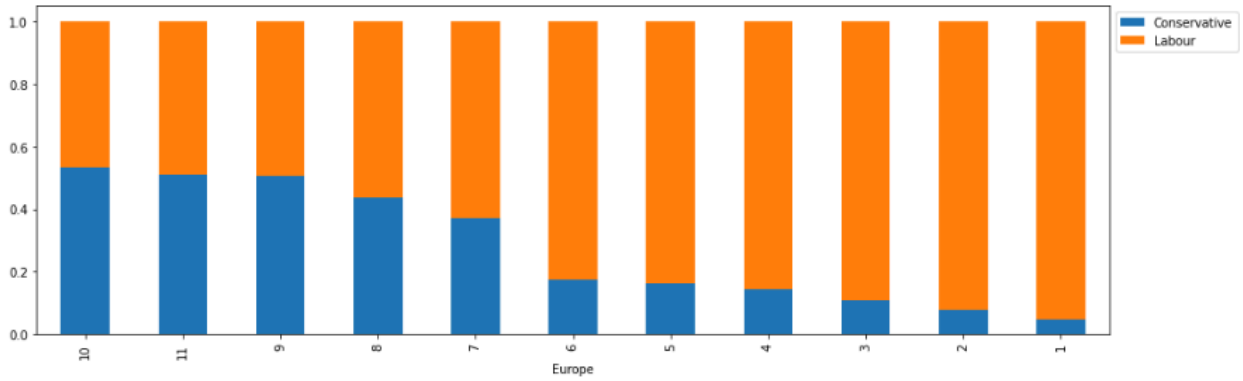


Figure 14: Voters' attitude towards Europe's integration vs Party that gets the vote

➔ The highest point 11 indicates high Euroscepticsm, i.e they have reservations about the European Union and European integration. This may be in light of the BREXIT issue. The proportion of votes for either parties seem divided

➔ However, among the voters who have rated 1, i.e. pro EU, the proportion of voters who have voted for the labour party is higher than that of the conservative party.

- **Political Knowledge vs Party that gets the vote**

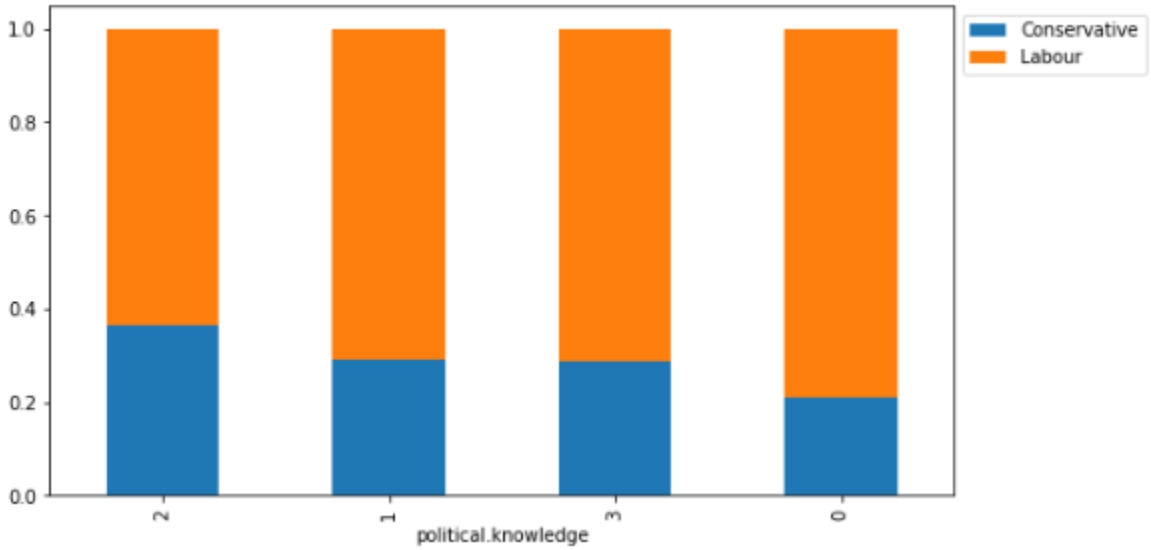
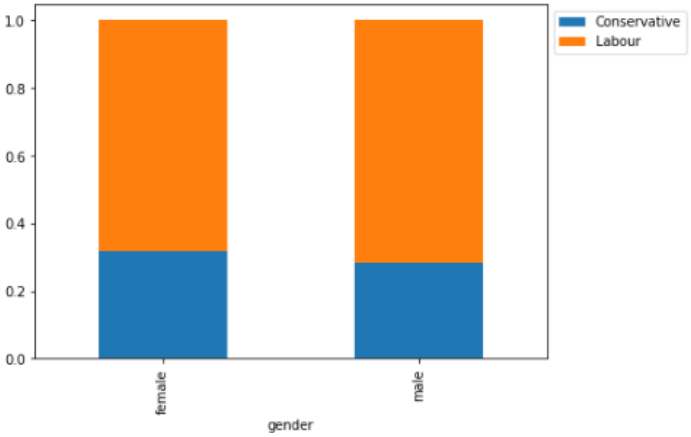


Figure 15: Political Knowledge vs Party that gets the vote

- **Gender vs Party that gets the vote**



vote	Conservative	Labour	All
gender			
All	462	1063	1525
female	259	553	812
male	203	510	713

Figure 16: Gender vs Party that gets the vote

➔ Gender does not play a major role in the voting behaviour. We can see that nearly the same number of male and female voters have voted for the Labour Party and the Conservative Party.

- **Age of voters vs Party that gets the vote**

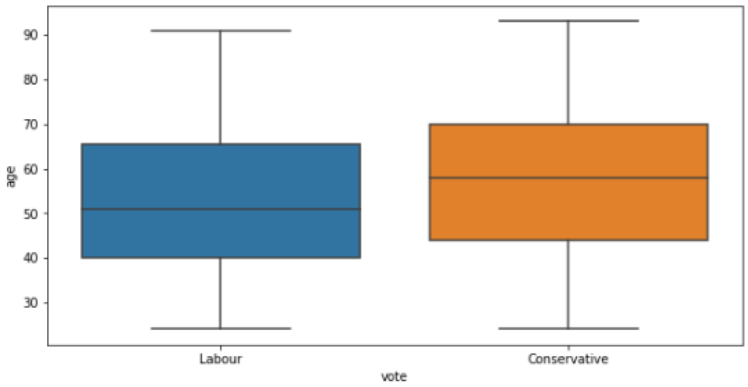


Figure 17: Age of voters vs Party that gets the vote

- The median age of Labour Party voters is about 50 and that of the Conservative Party is about 60.
- Age seems to be a determining factor in voting behaviour.

- **Age of voters vs Assessment of nation's economic condition**

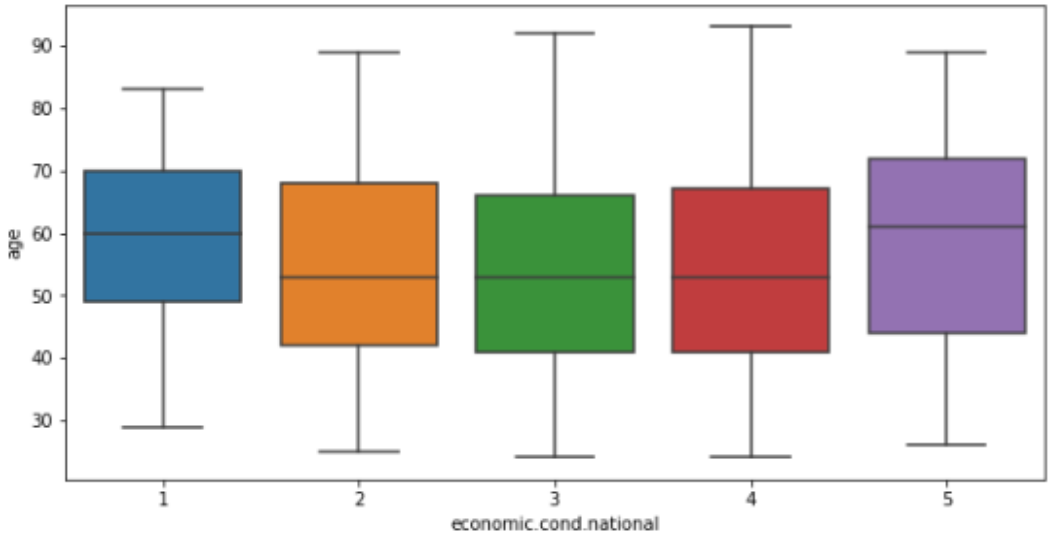


Figure 18: Age of voters vs Assessment of nation's economic condition

- The median age of the extreme ratings of 1 and 5 for the assessment of the nation's economic condition is about 60. Perhaps, as people grow older, they are highly critical of the elections and government.

- **Age of voters vs Assessment of household economic condition**

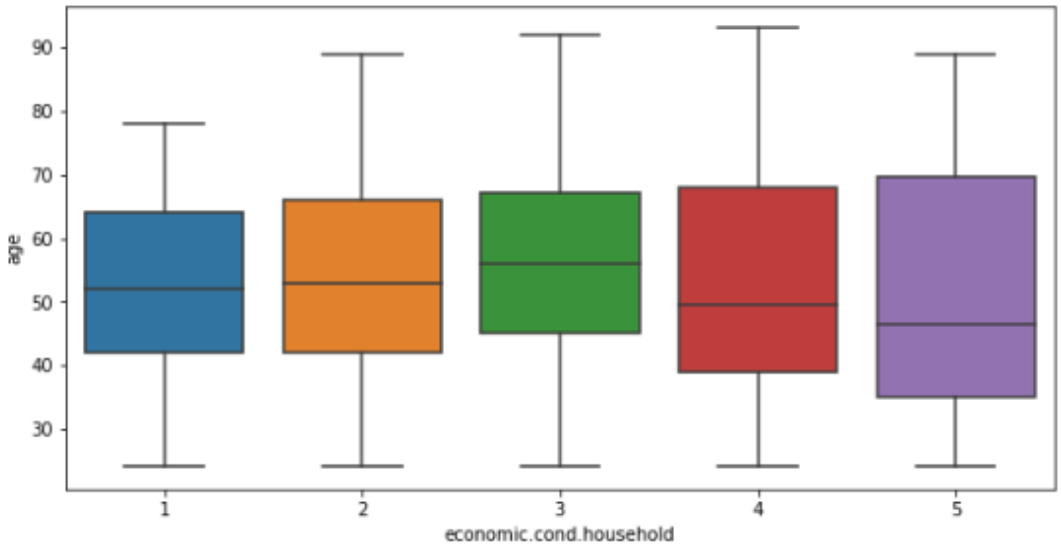


Figure 19: Age of voters vs Assessment of household economic condition

- The age of voters' perception of the household economic condition is well distributed for the rating of 5. This is also reflected in the vote distribution, i.e. 70% of the votes are for the Labour Party.

- **Age of voters vs Assessment of Blair**

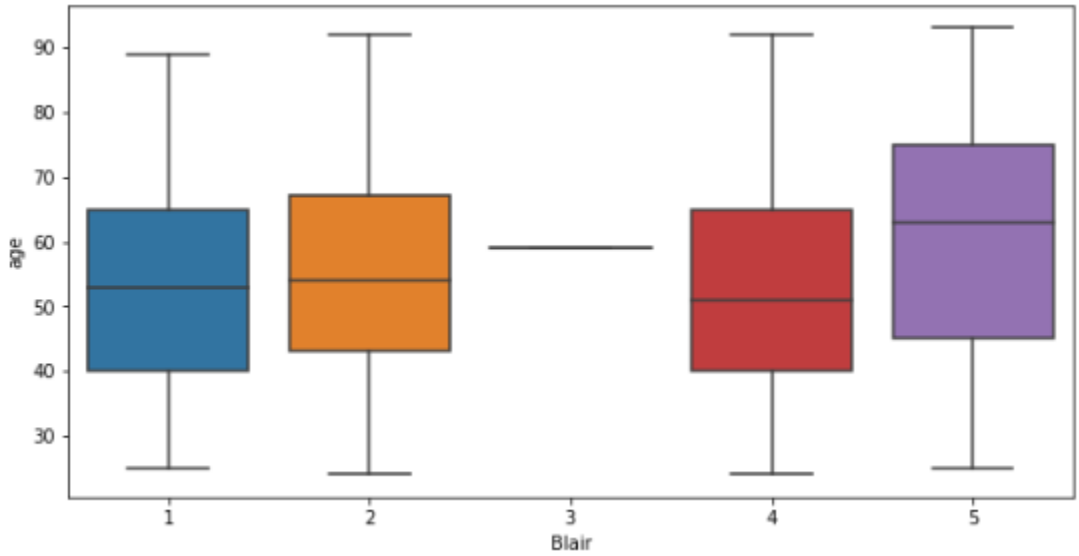


Figure 20: Age of voters vs Assessment of Blair

- The median age of voters who have a positive opinion (rating) of the Labour Party leader Blair is over 60.

- **Age of voters vs Assessment of Hague**

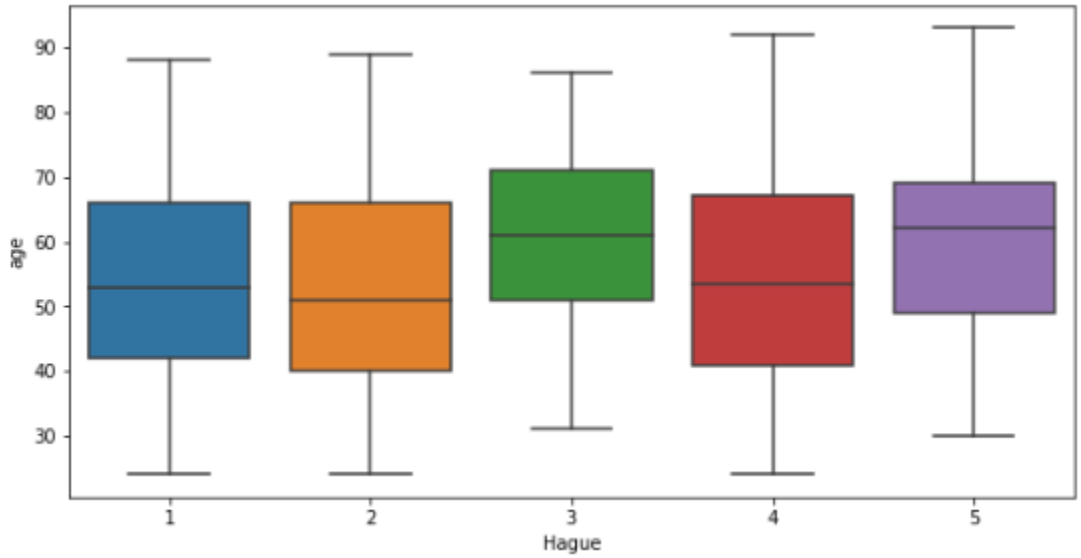


Figure 21: Age of voters vs Assessment of Hague

- The median age of voters who have a positive opinion (rating) of the Conservative Party leader Hague is also over 60.
- Clearly, it is the older people that take active interest in politics.

- **Age of voters vs Voters' attitude towards Europs's integration**

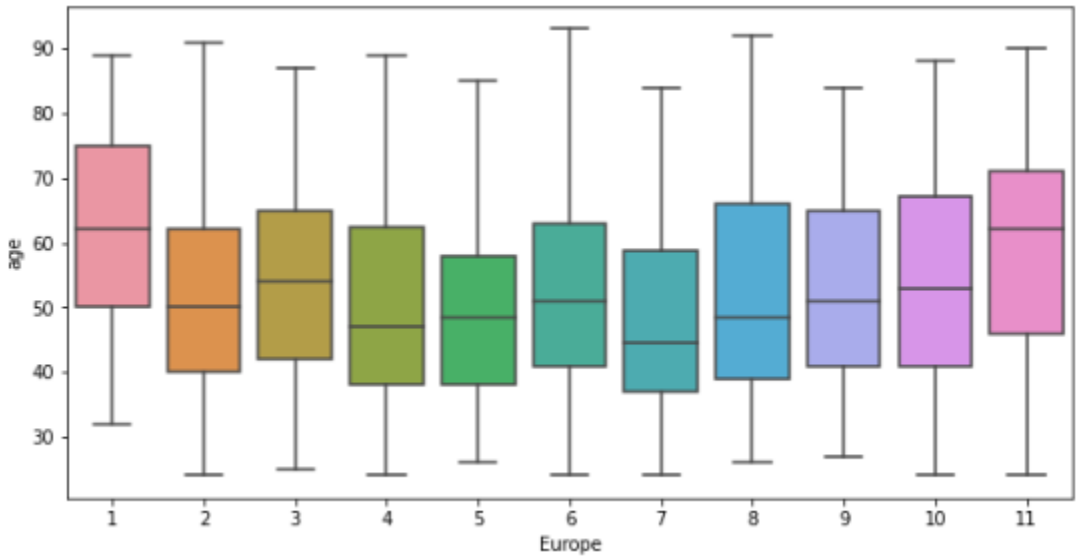


Figure 22: Age of voters vs Voters' attitude towards Europe's integration

- Pro-EU (rating 1) and anti-EU (rating 11) sentiments among voters across all age groups is divided.
- Hence, we cannot tell that age changes voters' opinion about the EU.

- **Age of voters vs Political knowledge**

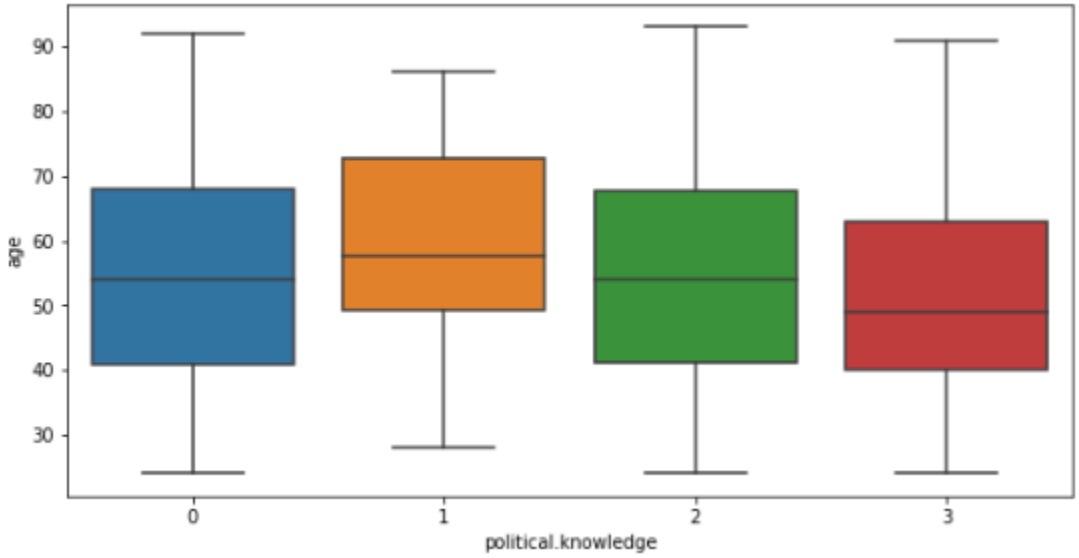


Figure 23: Age of voters vs Political knowledge

- The voters' knowledge of parties' position on EU integration is more or less the same across the age groups.

- **Age of voters vs Gender**

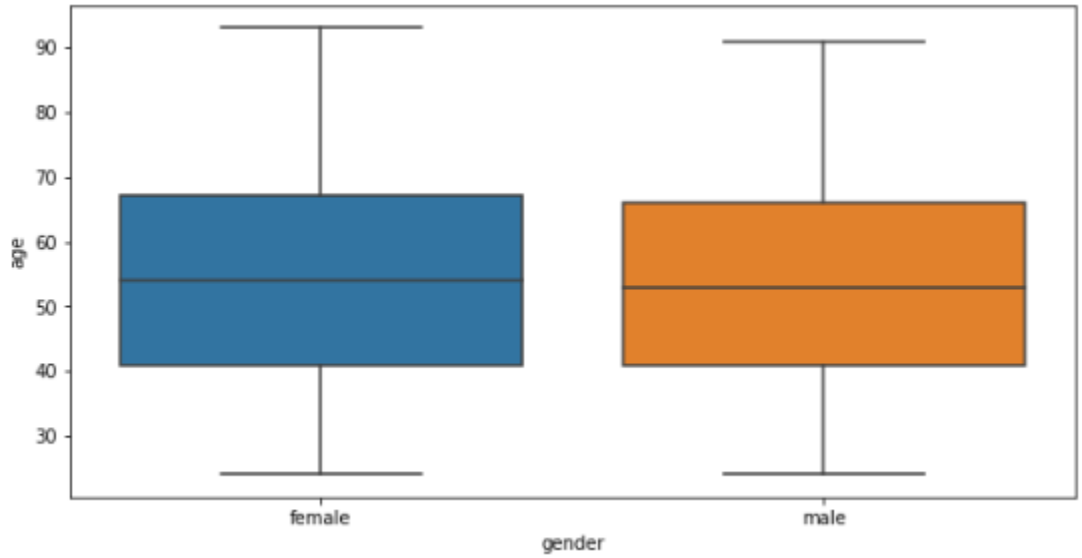


Figure 24: Age of voters vs Gender

- **Gender vs Political knowledge**

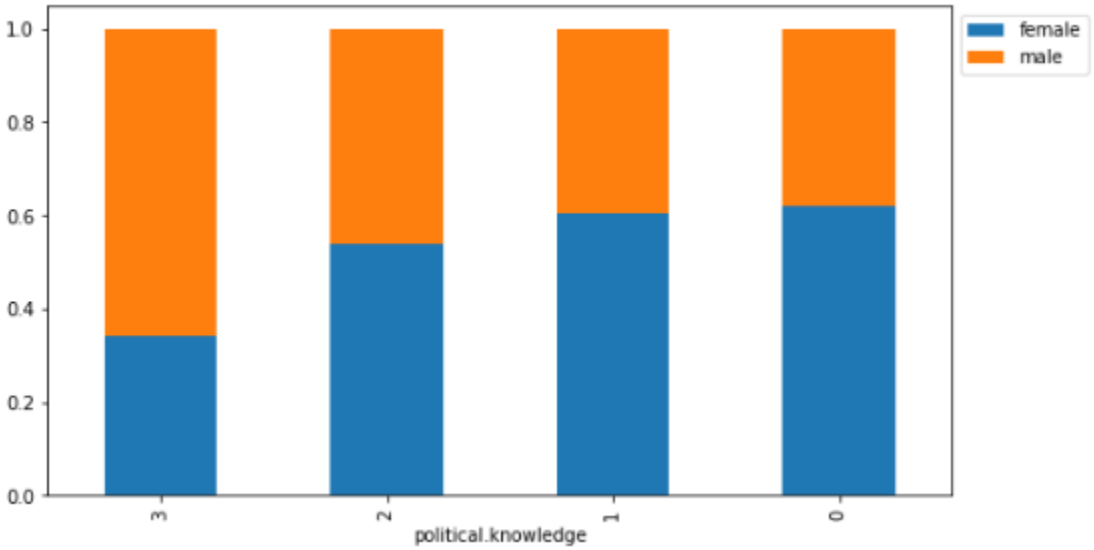


Figure 25: Gender vs Political knowledge

- Men seem to have more stronger understanding (rating 3) about the parties' position on EU integration than women.

- **Voters' attitude towards Europe's integration vs Gender**

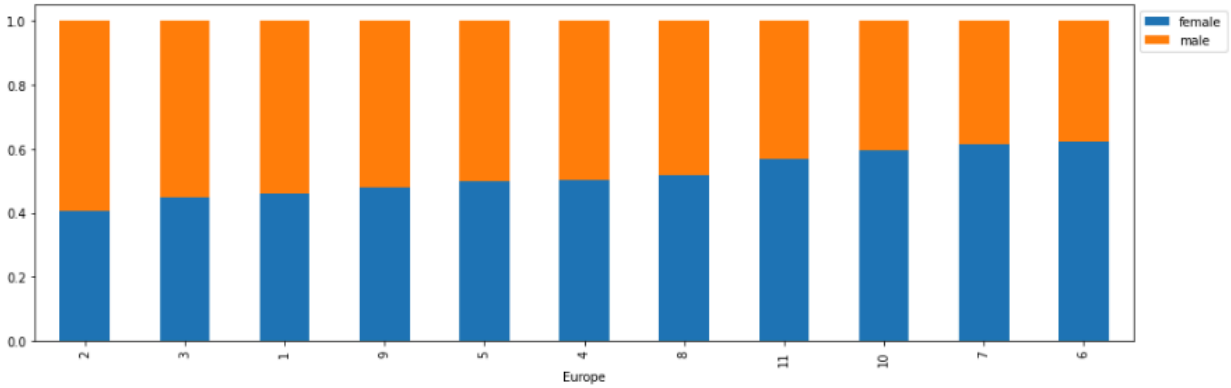


Figure 26: Voters' attitude towards Europe's integration vs Gender

- **Gender vs Assessment of Hague**

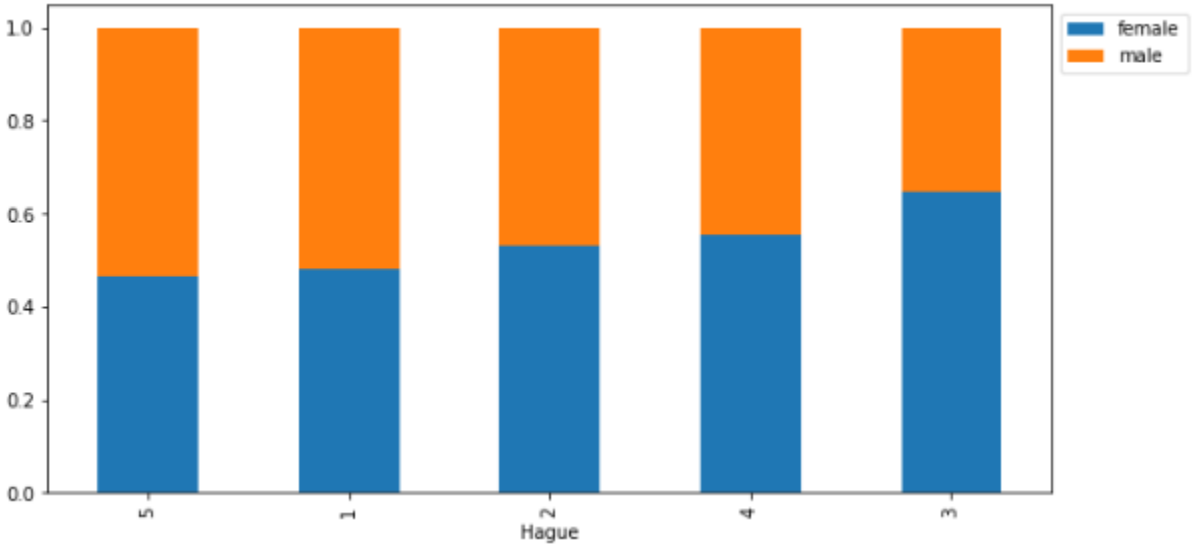


Figure 27: Gender vs Assessment of Hague

- **Gender vs Assessment of Blair**

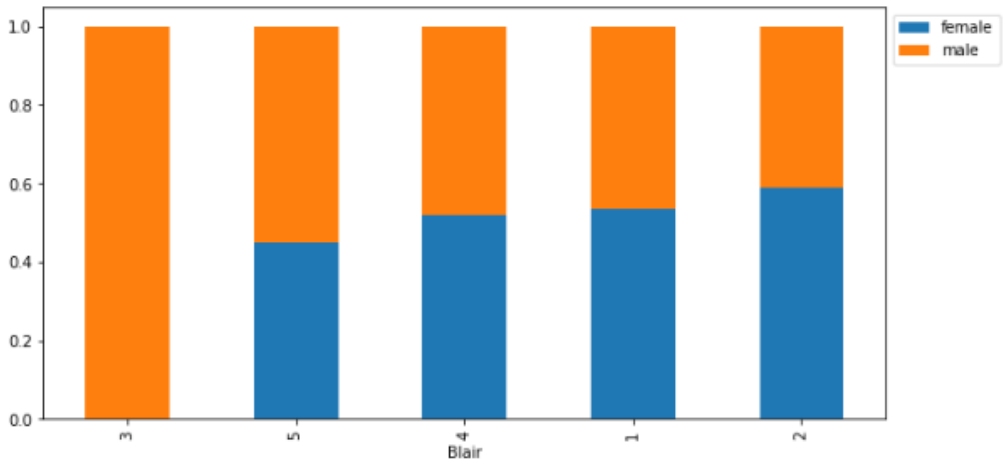


Figure 28: Gender vs Assessment of Blair

- **Gender vs Assessment of household economic conditions**

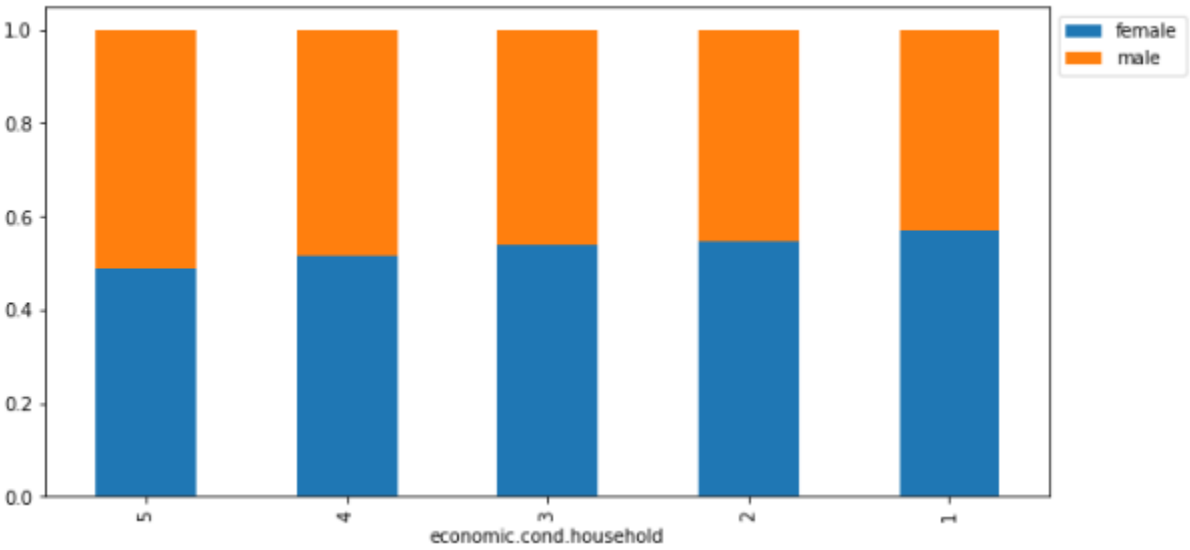


Figure 29: Gender vs Assessment of household economic conditions

- **Gender vs Assessment of nation's economic conditions**

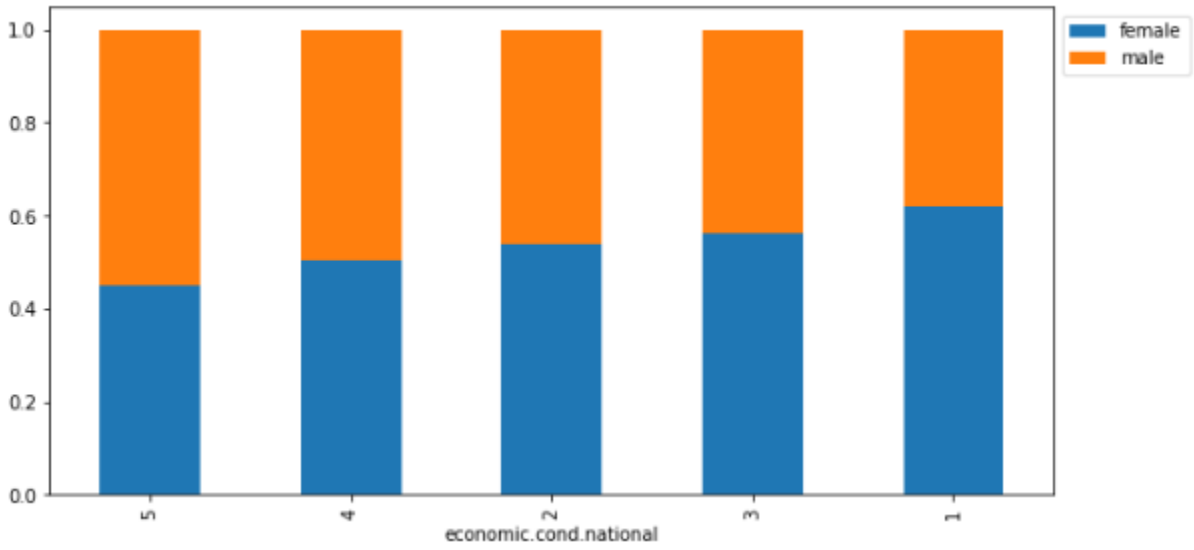


Figure 30: Gender vs Assessment of nation's economic conditions

- **Pair plot**

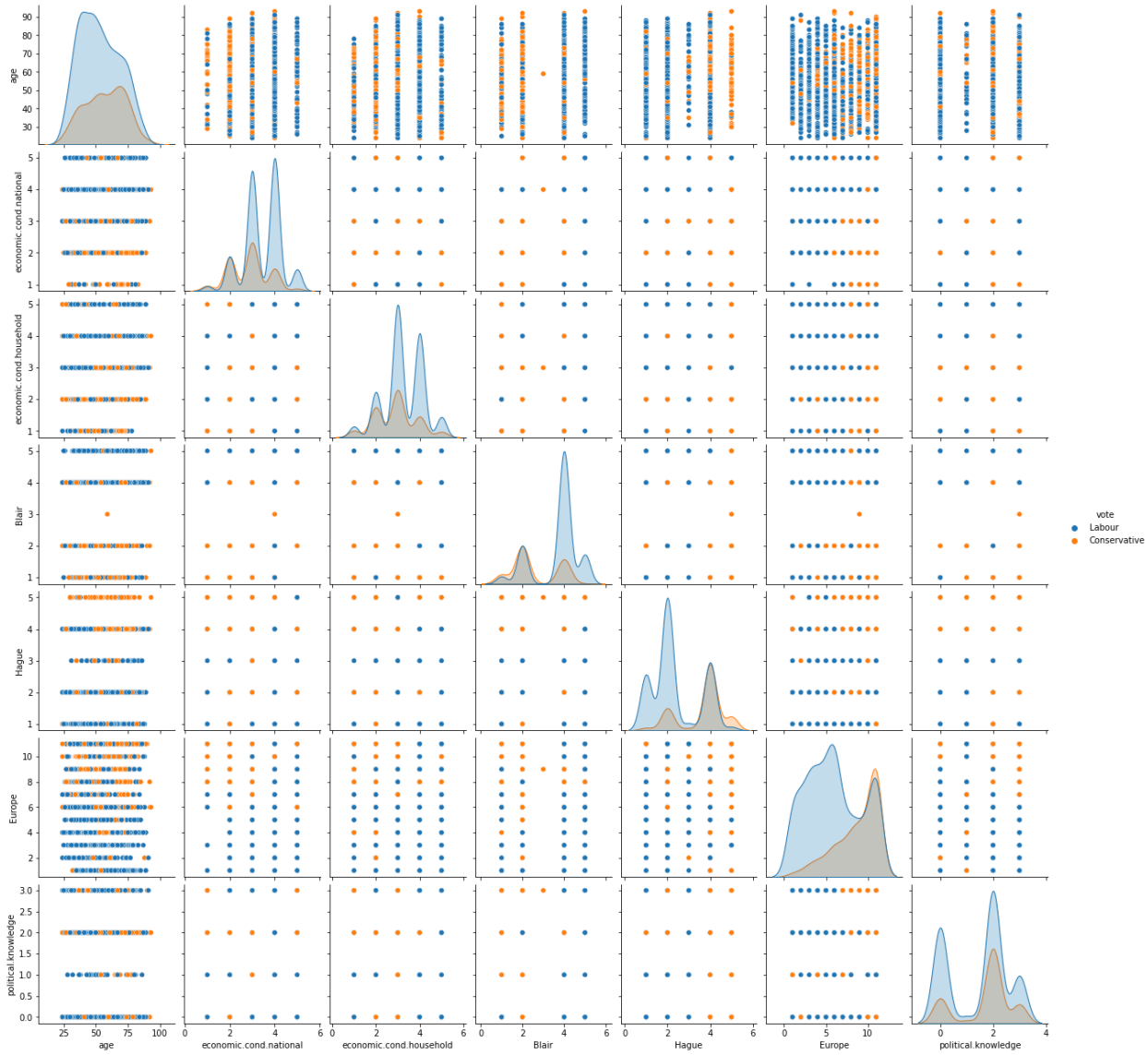


Figure 31: Pair plot

Data Preparation for Modeling

Checking for outliers

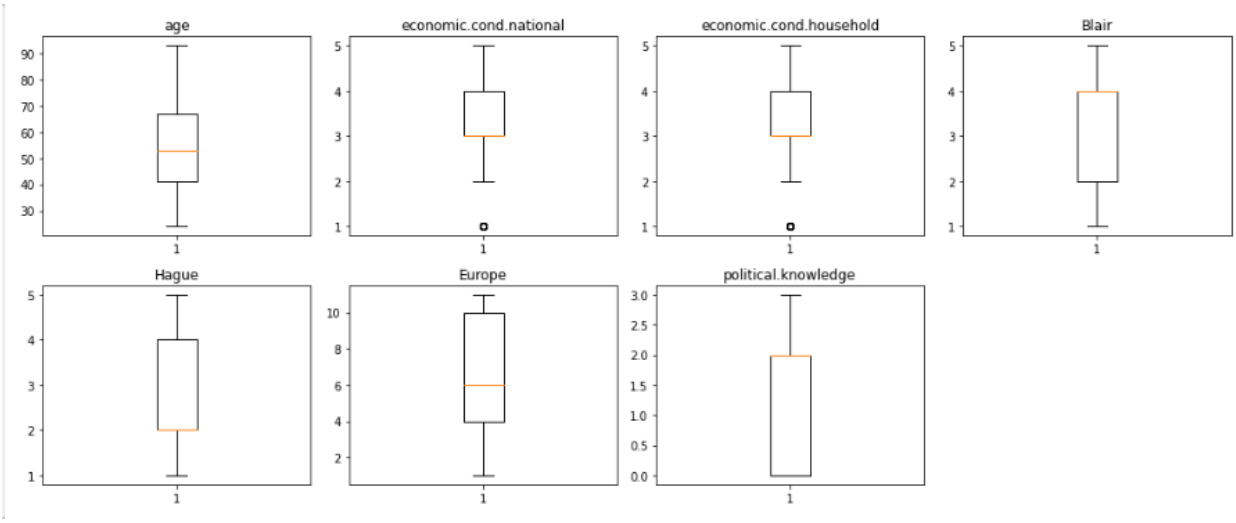


Figure 32: Outliers

The data doesn't have significant outliers. Hence, we need not treat for outliers.

- ★ We assign the Labour Party 0 and the Conservative Party 1.
- ★ Using One Hot Encoding, the gender column is encoded.
- ★ This makes all the columns integer data type.
- ★ We split the data into train and test in a 70:30 ratio to evaluate the model that we build on the train data.
    - Shape of Training set : (1067, 8)
    - Shape of test set : (458, 8)

Model Building and Model Evaluation

A model can make wrong predictions as:

- Model predicts which political party a voter is likely to support but in reality, the voter does not support the predicted political party.
- Model predicts which political party a voter is unlikely to support but in reality, the voter supports the political party.

Which case is more important?

- Both cases are important to get an accurate exit poll result.

How to reduce the losses?

- In the context of predicting election results between Labour and Conservative parties, we can consider prioritizing metrics like precision, recall, or the F1 score, since the election outcome seems highly imbalanced (**70% voted for Labour party and 30% for Conservative Party**).

Let's look at the evaluation metrics for different models - KNN, Naive Bayes, Bagging, and Boosting - for both training and testing datasets.

Key:

True positive: Actual conversion of leads into paid customers

False positive: Incorrect classification that the leads get converted into paid customers

True negative: Actual representation of leads who do not get converted into paid customers

False negative: Leads actually get converted to paid customers but are incorrectly classified as non-convertees.

KNN Model

Training data

Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
KNN_model	0.85	0.88	0.91	0.89	0.78	0.71	0.74

Table 3: KNN model metrics (training data)

- Accuracy is 85%. However, it's not a reliable measure for skewed data.
- Recall score is 91%, so the model is 91% good at minimizing false negatives, meaning it rarely misses instances of the positive class.
- F1 score is about 89%, which is not a very high score. The model is fairly good.
- Precision score is about 78%, i.e. the model is 78% good at making relatively few false positive predictions compared to the total number of positive predictions it has made.

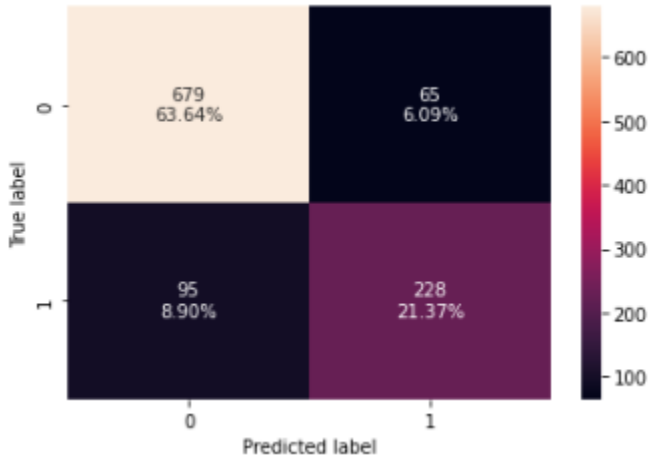


Figure 33: Confusion matrix KNN model (training data)

- The model has predicted 63% of the data correctly where the voters have voted for the Labour Party, and 6% incorrectly.
- For the Conservative party, it has predicted only 21% of the data correctly and 8% incorrectly.

Testing data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
KNN_test_model		0.83	0.86	0.9	0.88	0.75	0.68	0.71

Table 4: KNN model metrics (testing data)

- There's an overall decrease by about 2% across the metric scores for the testing data vis-a-vis the training data

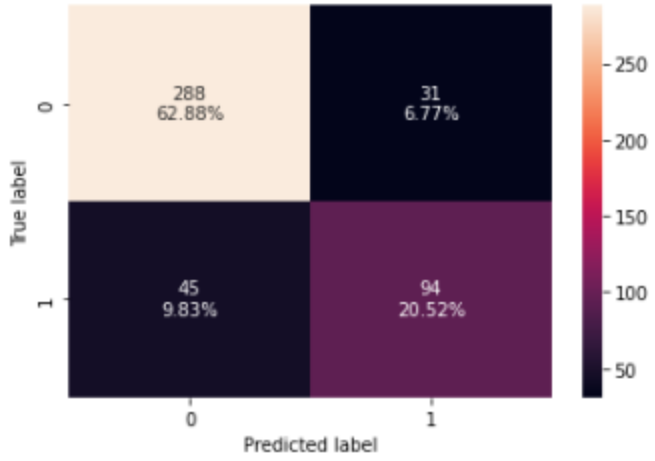


Figure 34: Confusion matrix KNN model (testing data)

ROC-AUC Curve

```
Train ROC-AUC score is : 0.9175592563001432

Test ROC-AUC score is : 0.8649782368462595
```

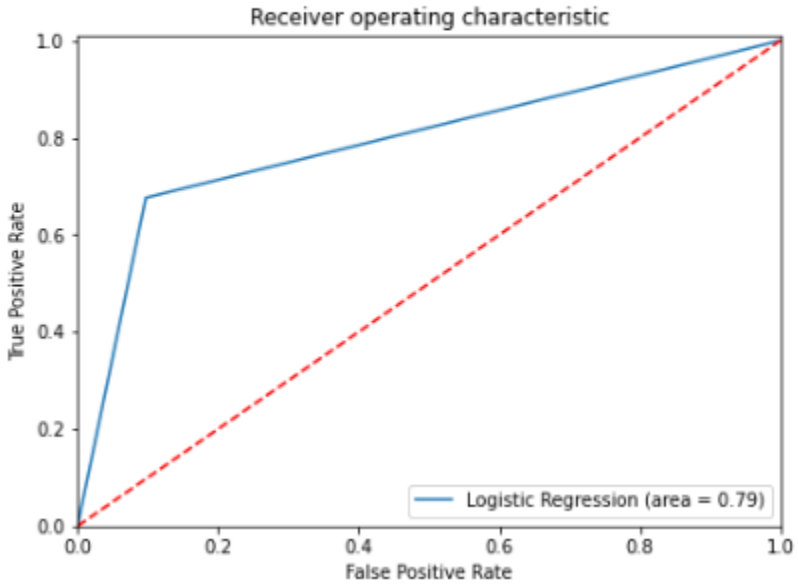


Figure 35: ROC-AUC for KNN model

- The ROC-AUC score for the training data is good, whereas for the test data it is not good enough.

Naive Bayes Model

Training data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
NB_train_model		0.82	0.87	0.88	0.87	0.71	0.69	0.7

Table 5: Naive Bayes model metrics (training data)

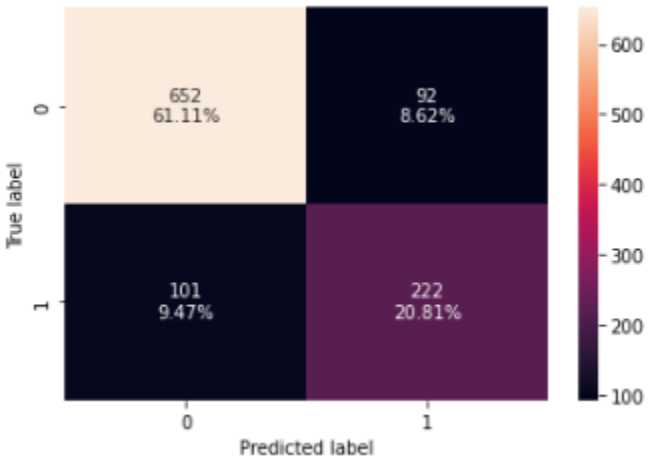


Figure 36: Confusion matrix Naive Bayes model (training data)

Testing data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
NB_test_model		0.86	0.89	0.92	0.9	0.79	0.73	0.76

Table 6: Naive Bayes model metrics (testing data)

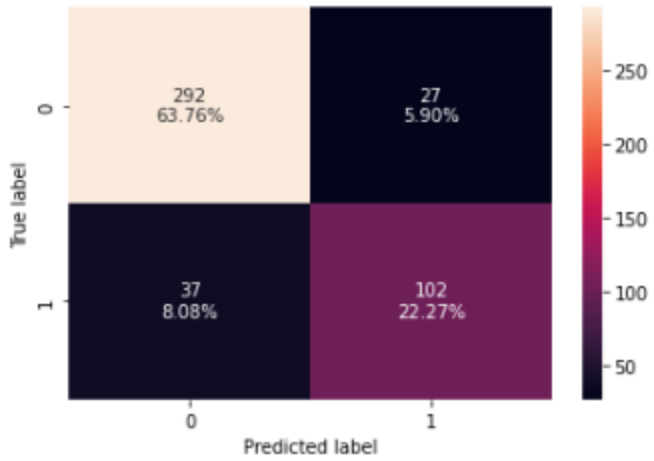


Figure 37: Confusion matrix Naive Bayes model (testing data)

ROC-AUC Curve

```
Train ROC-AUC score is : 0.8751851759379474

Test ROC-AUC score is : 0.9102185336370403
```

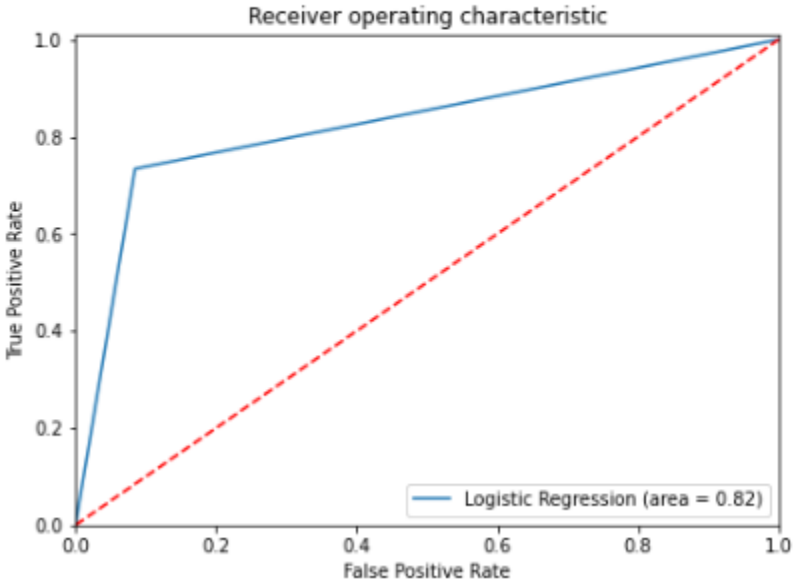


Figure 38: ROC-AUC for Naive Bayes model

- For the naive bayes model, there seems to be a slight improvement from the training data to the testing data, which was other way round for the KNN model.

Bagging Model

Training data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
Bagging_train_model		1	1	1	1	1	1	1

Table 7: Bagging model metrics (training data)

- The perfect score (unrealistic) of 1 for all metrics could be due to overfitting to a small dataset.

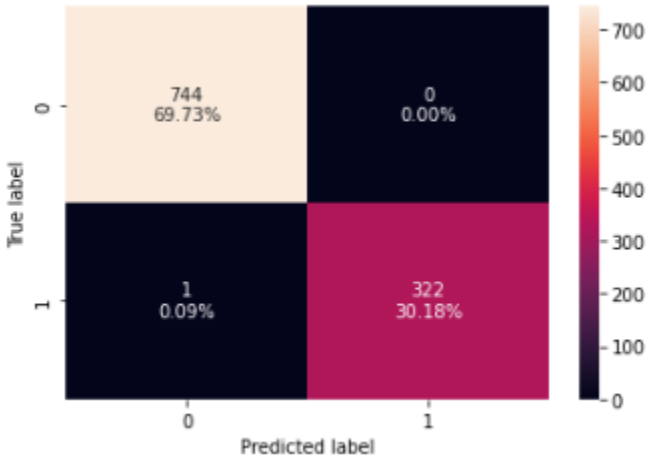


Figure 39: Confusion matrix Bagging model (training data)

Testing data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
Bagging_test_model		0.83	0.87	0.9	0.88	0.75	0.68	0.71

Table 8: Bagging model metrics (testing data)

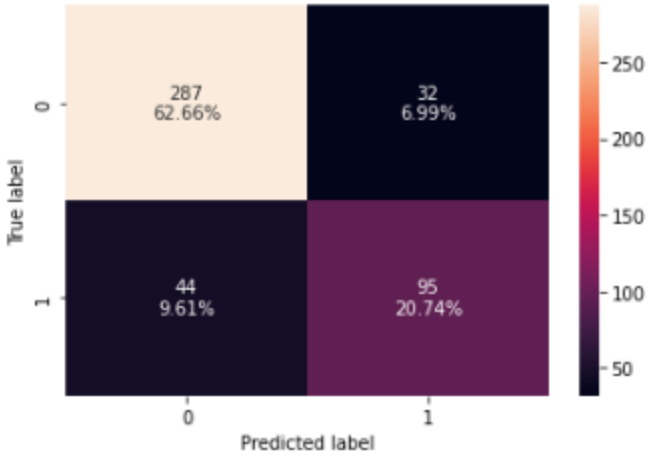


Figure 40: Confusion matrix Bagging model (testing data)

- The metric scores for the test data seem realistic, however they are not better than the KNN or Naive Bayes models.

ROC-AUC Curve

```
Train ROC-AUC score is : 0.9999895968574187

Test ROC-AUC score is : 0.8951534697007284
```

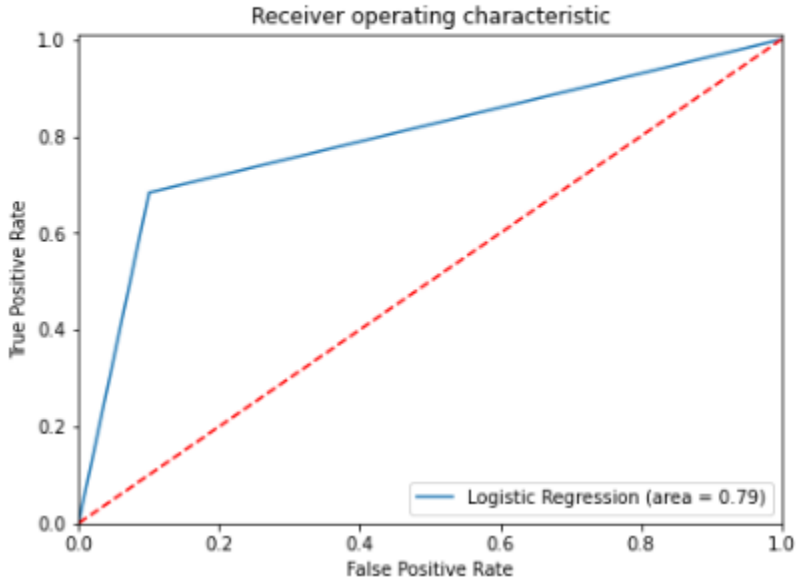


Figure 41: ROC-AUC curve for Bagging model

Boosting Models

We built 3 ensemble models here - AdaBoost Classifier, Gradient Boosting Classifier and XGBoost Classifier.

AdaBoost Classifier

Training data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
abc_train_model		0.84	0.87	0.91	0.89	0.76	0.67	0.72

Table 9: AdaBoost Classifier model metrics (training data)

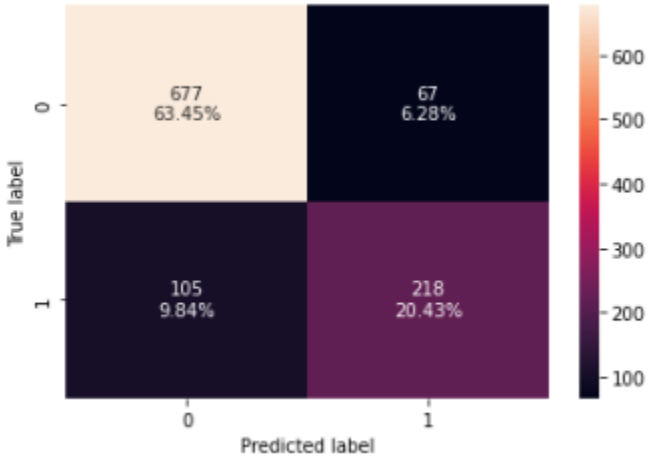


Figure 42: Confusion matrix AdaBoost Classifier model (training data)

Testing data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
abc_test_model		0.86	0.86	0.94	0.9	0.84	0.66	0.74

Table 10: AdaBoost Classifier model metrics (testing data)

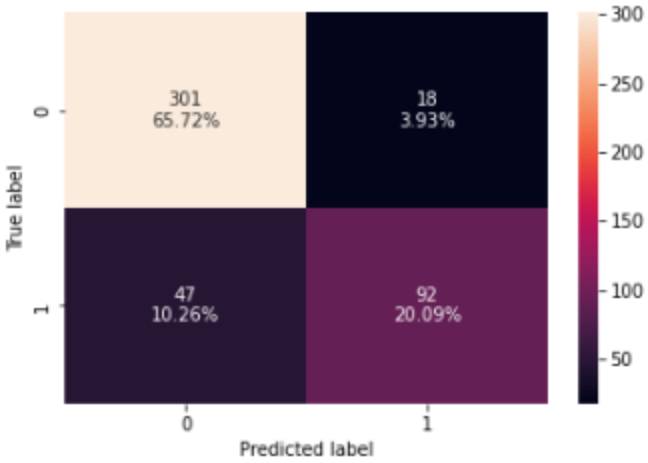


Figure 43: Confusion matrix AdaBoost Classifier model (testing data)

ROC-AUC Curve

```
Train ROC-AUC score is : 0.8987815839408769

Test ROC-AUC score is : 0.9105793734918022
```

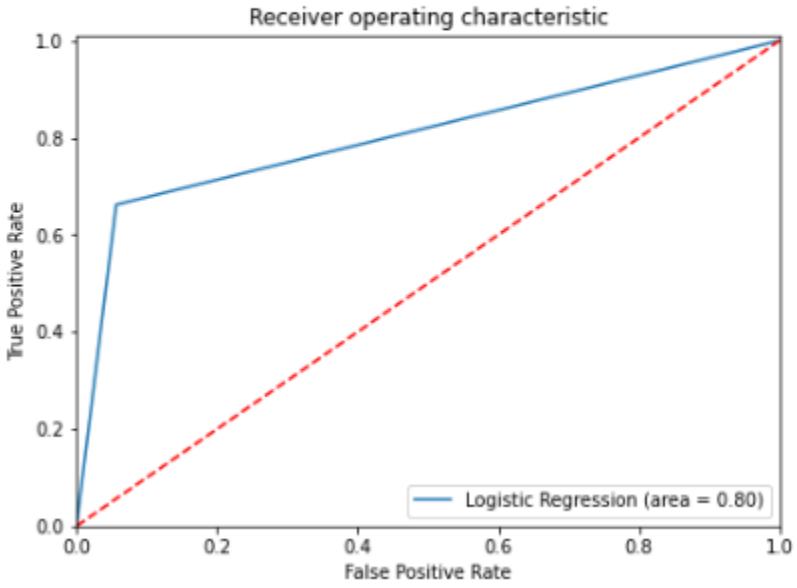


Figure 44: ROC-AUC curve for AdaBoost Classifier model

- The improvement in model performance from the training data to the testing data is very meagre.

Gradient Boosting Classifier

Training data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
gbc_train_model		0.87	0.89	0.93	0.91	0.82	0.73	0.77

Table 11: Gradient Boosting Classifier model metrics (training data)

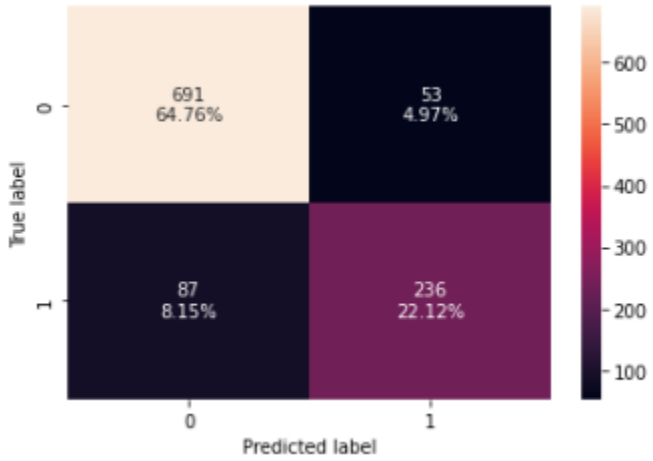


Figure 45: Confusion matrix Gradient Boosting Classifier (training data)

Testing data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
gbc_test_model		0.87	0.88	0.94	0.91	0.84	0.7	0.76

Table 12: Gradient Boosting Classifier model metrics (testing data)

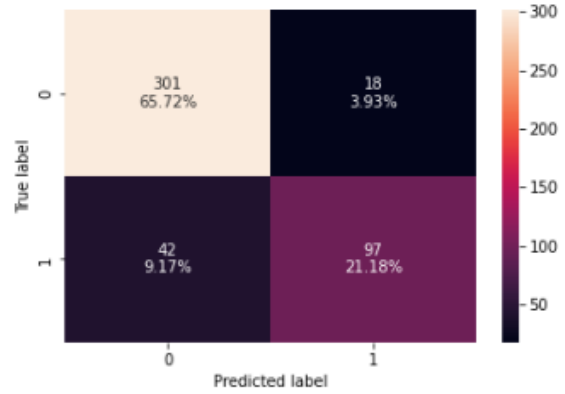


Figure 46: Confusion matrix Gradient Boosting Classifier (testing data)

ROC-AUC Curve

```
Train ROC-AUC score is : 0.9424560571257365

Test ROC-AUC score is : 0.9263773933830991
```

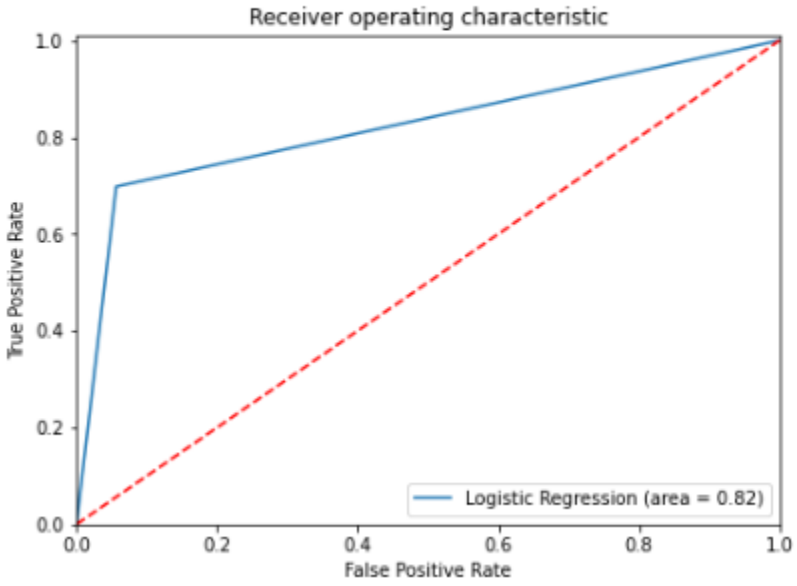


Figure 47: ROC-AUC curve for Gradient Boosting Classifier

- The model performance from the training to testing data has dropped by 2%.

XGBoost Classifier

Training data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
xgb_train_model		0.995	0.995	0.999	0.997	0.997	0.988	0.992

Table 13: XGBoost Classifier model metrics (training data)

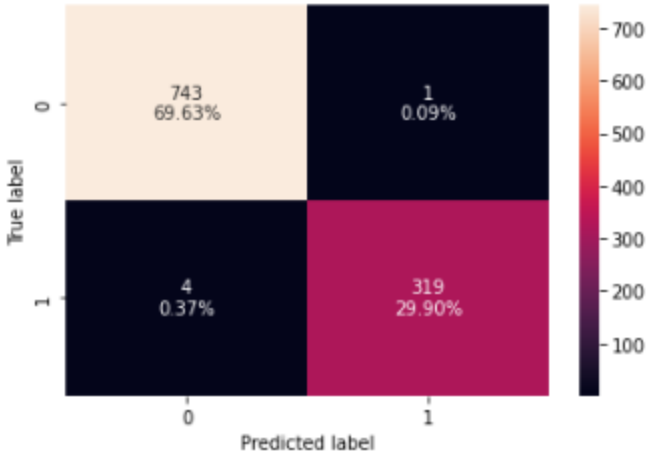


Figure 48: Confusion matrix XGBoost Classifier (training data)

Testing data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
xgb_test_model		0.849	0.877	0.912	0.894	0.778	0.705	0.74

Table 14: XGBoost Classifier model metrics (testing data)

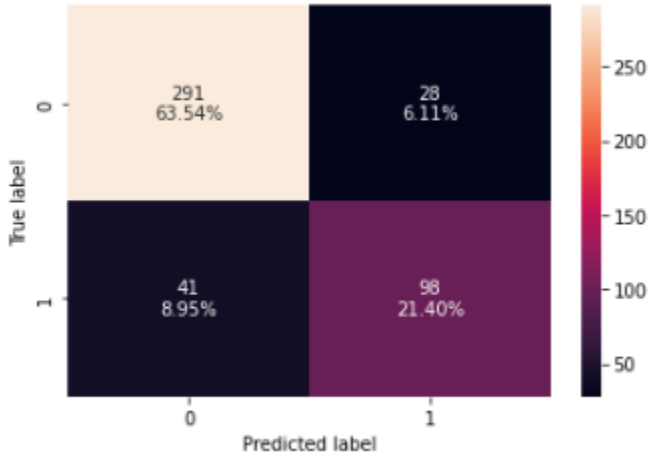


Figure 49: Confusion matrix XGBoost Classifier (testing data)

ROC-AUC Curve

```
Train ROC-AUC score is : 0.999785695262825

Test ROC-AUC score is : 0.9006111725040031
```

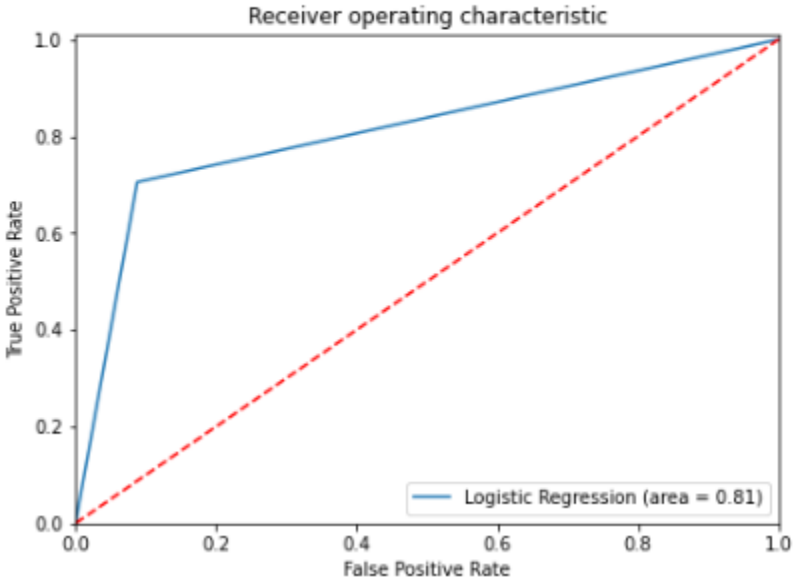


Figure 50: ROC-AUC curve for XGBoost Classifier

- Of all the models built so far, the XGBoost Classifier has the best metric scores.
- Although the performance from the training to testing data has dropped significantly, we can check this model by tuning it a little.

Model Performance Imporvement

We improve the model performance of bagging and boosting models by tuning the model.

Bagging Classifier

Training data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	\
tuned_bagging_train_model		0.962	1	1	1	

Precision_1	Recall_1	F1-Score_1
1	1	1

Table 15: Tuned Bagging model metrics (training data)

Testing data

Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
tuned_bagging_test_model	0.956	1	1	1	0.947	1	0.973

Table 16: Tuned Bagging model metrics (testing data)

- There's a significant increase in the metric values after tuning, which is good.

AdaBoost Classifier

Training data

Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
abc_tuned_train_model	0.956	1	1	1	0.912	0.969	0.939

Table 17: Tuned AdaBoost Classifier model metrics (training data)

Testing data

Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
abc_tuned_test_model	0.956	1	1	1	1	1	1

Table 18: Tuned AdaBoost Classifier model metrics (testing data)

- There's a significant increase in the metric values after tuning, which is good.
- The testing data shows better precision, recall, and F1 scores, indicating improved performance.

Features of importance

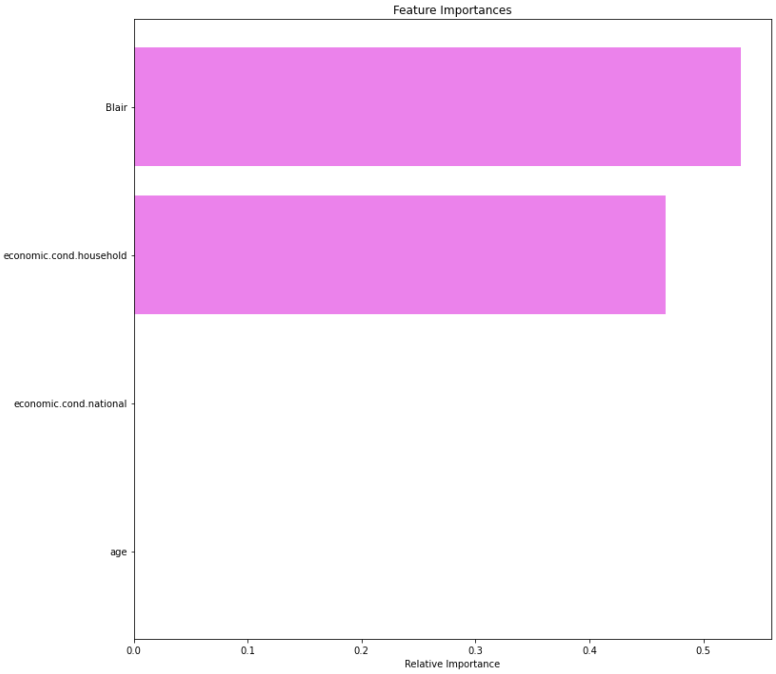


Figure 51: Feature of importance using AdaBoost Classifier

- Voters' assessment of the Labour Party leader Blair and their assessment of the household economic conditions seem to be the features of importance, i.e., features that influence voters' decisions.

Gradient Boosting Classifier

Training data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
gbc_tuned_train_model		1	1	1	1	1	1	1

Table 19: Tuned Gradient Boosting Classifier model metrics (training data)

Testing data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
gbc_tune_test_model		0.956	1	1	1	0.944	0.944	0.944

Table 20: Tuned Gradient Boosting Classifier model metrics (testing data)

Features of importance

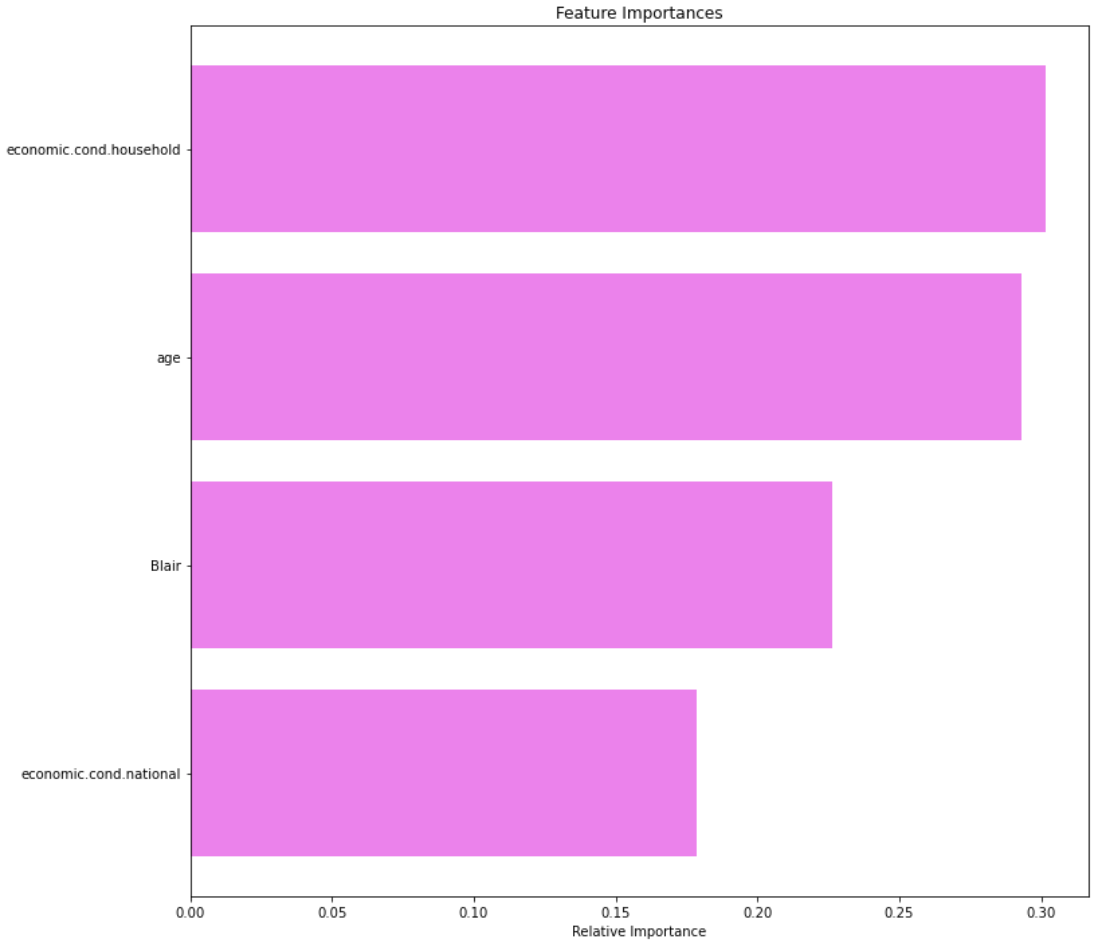


Figure 52: Feature of importance using Gradient Boosting Classifier

- The model has good metric scores. However, it performs better on training data than on testing data.
- The features of importance that impact voters' decisions are the economic conditions of the household, the age of voters, their assessment of the Labour Party leader Blair, and their assessment of the nation's economic conditions - in the same order of importance and weightage.

XGBoost Classifier

Training data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
xgb_tuned_train_model		0.962	1	1	1	0.938	0.938	0.938

Table 21: Tuned XGBoost Classifier model metrics (training data)

Testing data

	Model	Accuracy	Precision_0	Recall_0	F1-Score_0	Precision_1	Recall_1	F1-Score_1
xgb_tuned_test_model		0.933	1	1	1	0.941	0.889	0.914

Table 22: Tuned XGBoost Classifier model metrics (testing data)

Features of importance

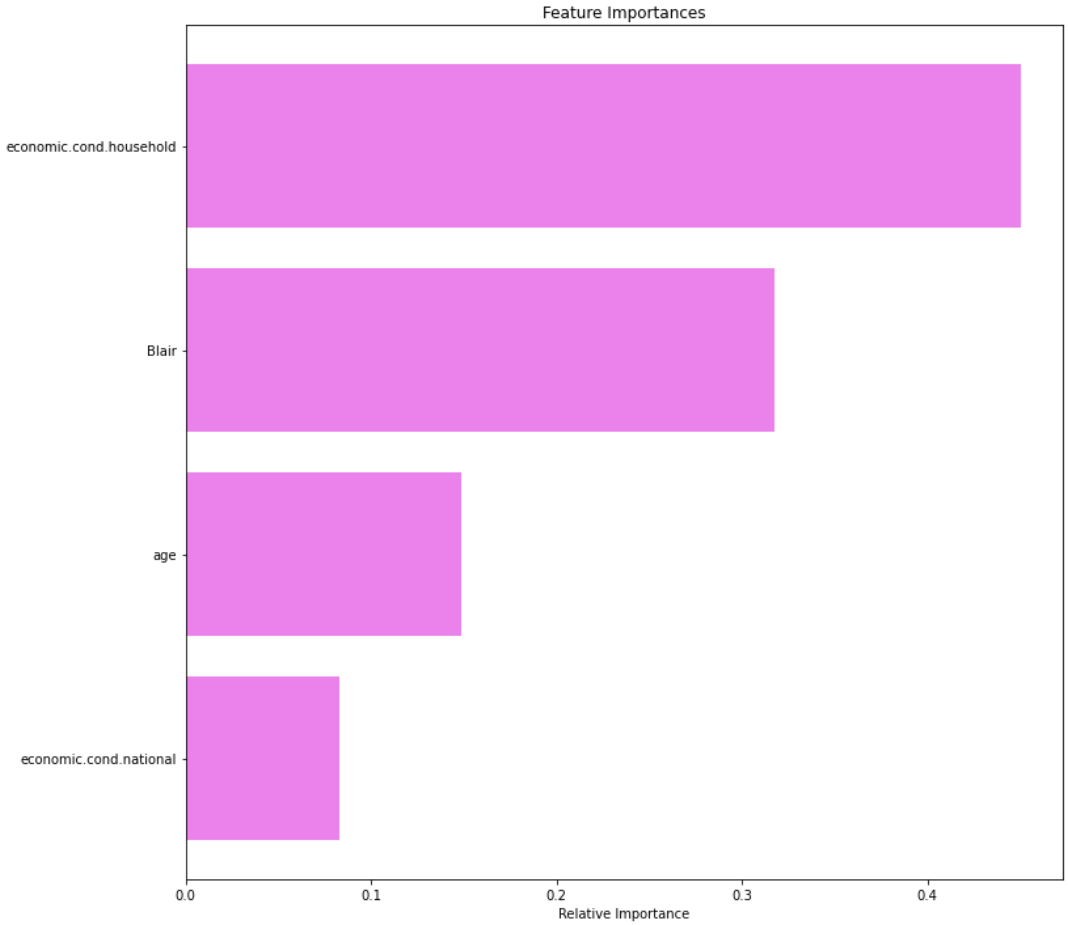


Figure 53: Feature of importance using XGBoost Classifier model

- The XGBoost classifier is very similar to the Gradient Boost Classifier in terms of metric scores and model performance.
- However, we can see differences in the importance assigned to the features.

Final Model Selection

Comparing all the models built so far

- For training data

```
Training performance comparison:
```

	KNN	Naive Bayes	Bagging	Ada Boosting	Gradient Boosting Classifier	XG Boost Classifier
Model	KNN_train_model	NB_train_model	Bagging_train_model	abc_train_model	gbc_train_model	xgb_train_model
Accuracy	0.85	0.82	1	0.84	0.87	0.995
Precision_0	0.88	0.87	1	0.87	0.89	0.995
Recall_0	0.91	0.88	1	0.91	0.93	0.999
F1-Score_0	0.89	0.87	1	0.89	0.91	0.997
Precision_1	0.78	0.71	1	0.76	0.82	0.997
Recall_1	0.71	0.69	1	0.67	0.73	0.988
F1-Score_1	0.74	0.7	1	0.72	0.77	0.992

Table 23: Training performance comparison of models

- We will not consider the Bagging model as the metrics are overly idealistic for the training data due to overfitting, and given the size of the dataset.
- Evidently, the XG Boost Classifier has the best scores.

- For testing data

```
Testing performance comparison:
```

	KNN	Naive Bayes	Bagging	Ada Boosting	Gradient Boosting Classifier	XG Boost Classifier
Model	KNN_test_model	NB_test_model	Bagging_test_model	abc_test_model	gbc_test_model	xgb_test_model
Accuracy	0.83	0.86	0.83	0.86	0.87	0.849
Precision_0	0.86	0.89	0.87	0.86	0.88	0.877
Recall_0	0.9	0.92	0.9	0.94	0.94	0.912
F1-Score_0	0.88	0.9	0.88	0.9	0.91	0.894
Precision_1	0.75	0.79	0.75	0.84	0.84	0.778
Recall_1	0.68	0.73	0.68	0.66	0.7	0.705
F1-Score_1	0.71	0.76	0.71	0.74	0.76	0.74

Table 24: Testing performance comparison of models

- For the testing data, we can use any of the boosting models as they have higher scores for most metrics.

Comparing the tuned models built so far

- For training data

	Bagging train	AdaBoost train	GradientBoost train	XGBoost train
Model	tuned_bagging_train_model	abc_tuned_train_model	gbc_tuned_train_model	xgb_tuned_train_model
Accuracy	0.962	0.956	1	0.962
Precision_0	1	1	1	1
Recall_0	1	1	1	1
F1-Score_0	1	1	1	1
Precision_1	1	0.912	1	0.938
Recall_1	1	0.969	1	0.938
F1-Score_1	1	0.939	1	0.938

Table 25: Training performance comparison of tuned models

- For testing data

	Bagging test	AdaBoost test	GradientBoost test	XGBoost test
Model	tuned_bagging_test_model	abc_tuned_test_model	gbc_tune_test_model	xgb_tuned_test_model
Accuracy	0.956	0.956	0.956	0.933
Precision_0	1	1	1	1
Recall_0	1	1	1	1
F1-Score_0	1	1	1	1
Precision_1	0.947	1	0.944	0.941
Recall_1	1	1	0.944	0.889
F1-Score_1	0.973	1	0.944	0.914

Table 26: Testing performance comparison of tuned models

  - After tuning, we can observe that the Gradient Boost model has performed perfectly well for the training data and reasonably well for the testing data as well.

Actionable Insights & Recommendations

- After comparing different models before and after tuning, we can conclude that the Gradient Boosting Model is a better choice.
- The Gradient Boosting Model has a perfect score for the training set and scores above 0.9 (values close to 1) for the test dataset.
- On determining the features of importance, it does a good job of assigning reasonable weightage to the features that influence voting behaviour.

- However, the dataset we have analysed is highly skewed. 70% of the voters have voted for the Labour Party. We may not have enough information about voters who voted for the Conservative Party. Hence, we may need to run the models for a larger dataset.

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:

- President Franklin D. Roosevelt in 1941
- President John F. Kennedy in 1961
- President Richard Nixon in 1973

nltk is the Natural Language Toolkit is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language.

Roosevelt’s speech

- Number of characters in Roosevelt's speech: 7571
- Number of words in Roosevelt's speech: 1360
- Number of sentences in Roosevelt's speech: 70

Kennedy’s speech

- Number of characters in Kennedy's speech: 7618
- Number of words in Kennedy's speech: 1390
- Number of sentences in Kennedy's speech: 57

Nixon’s speech

- Number of characters in Nixon's speech: 9991
- Number of words in Nixon's speech: 1819
- Number of sentences in Nixon's speech: 74

Text cleaning

We remove stopwords such as a, an, the, is, etc., and stem the words, i.e. removing prefixes and suffixes from words to obtain their root form.

We find the 3 most common words used in all three speeches.

Roosevelt's speech
('nation', 17), ('know', 10), ('peopl', 9)

Kennedy's speech
('let', 16), ('us', 12), ('power', 9)

Nixon's speech
('us', 26), ('let', 22), ('america', 21)

Word cloud of all three speeches

- Roosevelt's speech

Figure 54: Word cloud for Roosevelt's speech

- Kennedy's speech

Figure 55: Word cloud for Kennedy’s speech

- Nixon’s speech

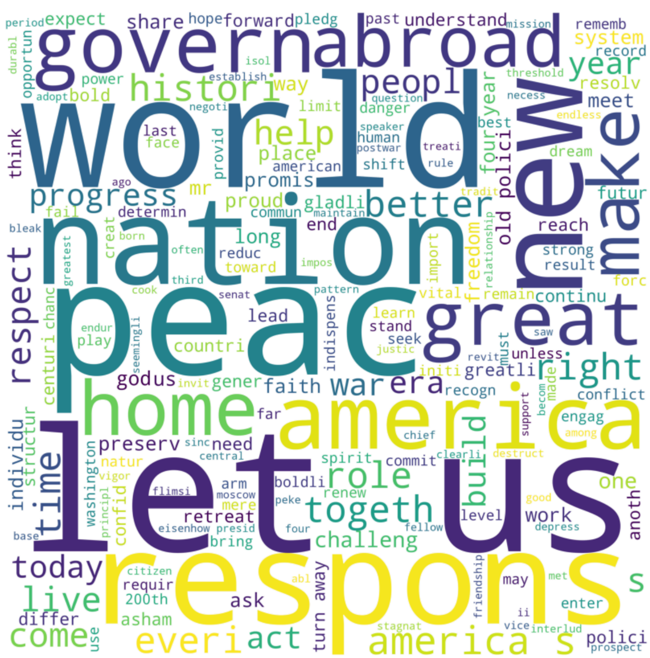


Figure 56: Word cloud for Kennedy’s speech