

Machine Learning -2

Coded Project

SRINIVASAN T

Table of Content:

S.NO.	TITLE	PAGE NO.
	Problem 1	
1.1	Define the problem and perform Exploratory Data Analysis	8
1.1.1	Problem definition	8
1.1.2	Loading the Dataset	8
1.1.3	Check shape	8
1.1.4	Data types	8
1.1.5	Statistical summary	9
1.1.6	Univariate analysis	10
1.1.6.1	Numerical Variable	10
1.1.6.2	Categorical Variable	10
1.1.7	Multivariate analysis	11
1.1.7.1	Numerical (all) vs Categorical (Vote)	11
1.1.7.2	Numerical (all) vs Categorical (Gender)	12
1.1.8	Patterns and insights - Key meaningful observations	12
1.2	Data Pre-processing	13
1.2.1	Outlier Detection(treat, if needed)	13
1.2.1.1	Before Outlier Treatment	13
1.2.1.2	After Outlier Treatment	14
1.2.2	Encode the data	14
1.2.3	Data split	15
1.2.4	Scale the data (and state your reasons for scaling the features)	15
1.3	Model Building	16
1.3.1	Model Building (KNN, Naive bayes, Bagging, Boosting)	16
1.3.1.1	Train Test Split	16
1.3.2	Metrics of Choice (Justify the evaluation metrics)	16
1.4	Model Performance evaluation	16
1.4.1	Check the confusion matrix	16
1.4.2	classification metrics for all the models (for both train and test dataset)	17
1.4.3	ROC-AUC score and plot the curve	20
1.4.3.1	ROC-AUC Train data	20
1.4.3.2	ROC-AUC Test data	21
1.4.4	Comment on all the model performance	23

1.5	Model Performance improvement	23
1.5.1	Improve the model performance of bagging and boosting models by tuning the model	23
1.5.1.1	Hypertuning – Bagging	23
1.5.1.2	Hypertuning – AdaBoosting	24
1.5.1.3	Hypertuning - Gradient Boosting	25
1.5.1.4	Hypertuning - XG Boosting	25
1.5.2	Comment on the model performance improvement on training and test data	26
1.6	Final Model Selection	27
1.6.1	Compare all the model built so far	27
1.6.2	Select the final model with the proper justification	27
1.6.3	Check the most important features in the final model and draw inferences	28
1.7	Actionable Insights & Recommendations	28
1.7.1	Compare all four models	28
1.7.2	Conclude with the key takeaways for the business	28
	Problem 2	
2.1	Define the problem and Perform Exploratory Data Analysis	29
2.1.1	Problem Definition	29
2.1.2	Find the number of Character, words & sentences in all three speeches	29
2.2	Text cleaning	29
2.2.1	Stopword removal	29
2.2.2	Stemming	31
2.2.3	Find the 3 most common words used in all three speeches	32
2.3	Plot Word cloud of all three speeches	33
2.3.1	Show the most common words used in all three speeches in the form of word clouds	33

List of Tables:

TABLE.NO.	TITLE	PAGE NO.
	Problem 1	
1.1	Loading the dataset and checking the first 5 rows	8
1.2	Statistical Summary of the numerical variables	9
1.3	Statistical Summary of the categorical variables	9
1.4	Independent variables after data split	15
1.5	Dependent variables after data split	15
1.6	Data after Zscore scaling	15
1.7	Model Building with Metrics	16
1.8	Model Building after Hypertuning	26
1.9	Model Comparison after Hypertuning	27
1.10	Model Comparison	28
	Problem 2	
2.1	Char, Word, Sent. Count of all three speeches	29

List of Figures:

FIGURE.NO.	TITLE	PAGE NO.
	Problem 1	
1.1	Loading the dataset and checking the first 5 rows	8
1.2	Univariate analysis – numerical variable	10
1.3	Univariate analysis – categorical variable - Vote	10
1.4	Univariate analysis – categorical variable - Gender	11
1.5	Multivariate analysis – numerical vs categorical	11
1.6	Multivariate analysis – numerical (all) vs categorical (gender)	12
1.7	Before Outlier Treatment	13
1.8	After Outlier Treatment	14
1.9	Datatypes after encoding	14
1.10	ROC-AUC GaussianNB Train	20
1.11	ROC-AUC KNeighborsClassifier Train	20
1.12	ROC-AUC BaggingClassifier Train	21
1.13	ROC-AUC AdaBoostClassifier Train	21
1.14	ROC-AUC GradientBoostClassifier Train	21
1.15	ROC-AUC XGBoostClassifier Train	21
1.16	ROC-AUC GaussianNB Test	22
1.17	ROC-AUC KNeighborsClassifier Test	22
1.18	ROC-AUC BaggingClassifier Test	22
1.19	ROC-AUC AdaBoostClassifier Test	22
1.20	ROC-AUC GradientBoostClassifier Test	22
1.21	ROC-AUC XGBoostClassifier Test	22
1.22	Hypertuning - Bagging	23
1.23	Feature Importance - Bagging	24
1.24	Hypertuning – AdaBoosting	24
1.25	Feature Importance - AdaBoosting	24
1.26	Hypertuning – Gradient Boosting	25
1.27	Feature Importance – Gradient Boosting	25
1.28	Hypertuning – XG Boosting	25
1.29	Feature Importance – XG Boosting	26
1.30	Feature Importance – XG Boosting Tuned	28
	Problem 2	
2.1	Most used words in Roosevelt Speech	33
2.2	Most used words in Kennedy Speech	33
2.3	Most used words in Nixon Speech	34

Scoring guide (Rubric) - ML Project - Coded

Criteria	Points
Define the problem and perform Exploratory Data Analysis - Problem definition - Check shape, Data types, and statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables	6
Data Pre-processing Prepare the data for modelling: - Outlier Detection(treat, if needed)) - Encode the data - Data split - Scale the data (and state your reasons for scaling the features)	2
Model Building - Metrics of Choice (Justify the evaluation metrics) - Model Building (KNN, Naive bayes, Bagging, Boosting)	10
Model Performance evaluation - Check the confusion matrix and classification metrics for all the models (for both train and test dataset) - ROC-AUC score and plot the curve - Comment on all the model performance	8
Model Performance improvement - Improve the model performance of bagging and boosting models by tuning the model - Comment on the model performance improvement on training and test data	9
Final Model Selection - Compare all the model built so far - Select the final model with the proper justification - Check the most important features in the final model and draw inferences.	4
Actionable Insights & Recommendations - Compare all four models - Conclude with the key takeaways for the business	6
Problem 2 - Define the problem and Perform Exploratory Data Analysis -Problem Definition - Find the number of Character, words & sentences in all three speeches	3
Problem 2 - Text cleaning - Stopword removal - Stemming - find the 3 most common words used in all three speeches	3
Problem 2 - Plot Word cloud of all three speeches - Show the most common words used in all three speeches in the form of word clouds	3
Business Report Quality -Adhere to the business report checklist	6
	Points 60

Problem Statement - ML Project - Coded

Problem 1

Context

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

Objective

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

Data Description

1. **vote**: Party choice: Conservative or Labour
2. **age**: in years
3. **economic.cond.national**: Assessment of current national economic conditions, 1 to 5.
4. **economic.cond.household**: Assessment of current household economic conditions, 1 to 5.
5. **Blair**: Assessment of the Labour leader, 1 to 5.
6. **Hague**: Assessment of the Conservative leader, 1 to 5.
7. **Europe**: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
8. **political.knowledge**: Knowledge of parties' positions on European integration, 0 to 3.
9. **gender**: female or male.


Problem 2

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

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

Code Snippet to extract the three speeches:

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

If the above code doesn't work, use data: [Speeches](#) 

Problem 1

1.1 Define the problem and perform Exploratory Data Analysis

1.1.1 Problem definition

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

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.

1.1.2 Loading the Dataset

Loading the dataset using the read function and checking whether it is properly loaded or not using the head function.

	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 1.1: Loading the dataset and checking the first 5 rows

1.1.3 Check shape

The no. of rows is 1525

The no. of columns is 10

1.1.4 Data types

The datatypes are identified using the info function.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 columns):
#   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 1.1: Loading the dataset and checking the first 5 rows

- The variables vote and gender are object type and other variables are int
- The two object variables can be encoded to numeric variables
- There are no null values

1.1.5 Statistical summary

The statistical summary can be derived from the describe function

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	1525.0	763.000000	440.373894	1.0	382.0	763.0	1144.0	1525.0
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

Table 1.2: Statistical Summary of the numerical variables

- All the variables except Blair are almost normally distributed (Mean almost equal to Median)
- The age variable has outliers
- The median age is 53 as the voters are older age people

	count	unique	top	freq
vote	1525	2	Labour	1063
gender	1525	2	female	812

Table 1.3: Statistical Summary of the categorical variables

- Almost 70% of the voters are Labour party and independent variable is skewed towards one party
- SMOTE technique need to be used to balance the data before modelling

1.1.6 Univariate analysis

1.1.6.1 Numerical Variable

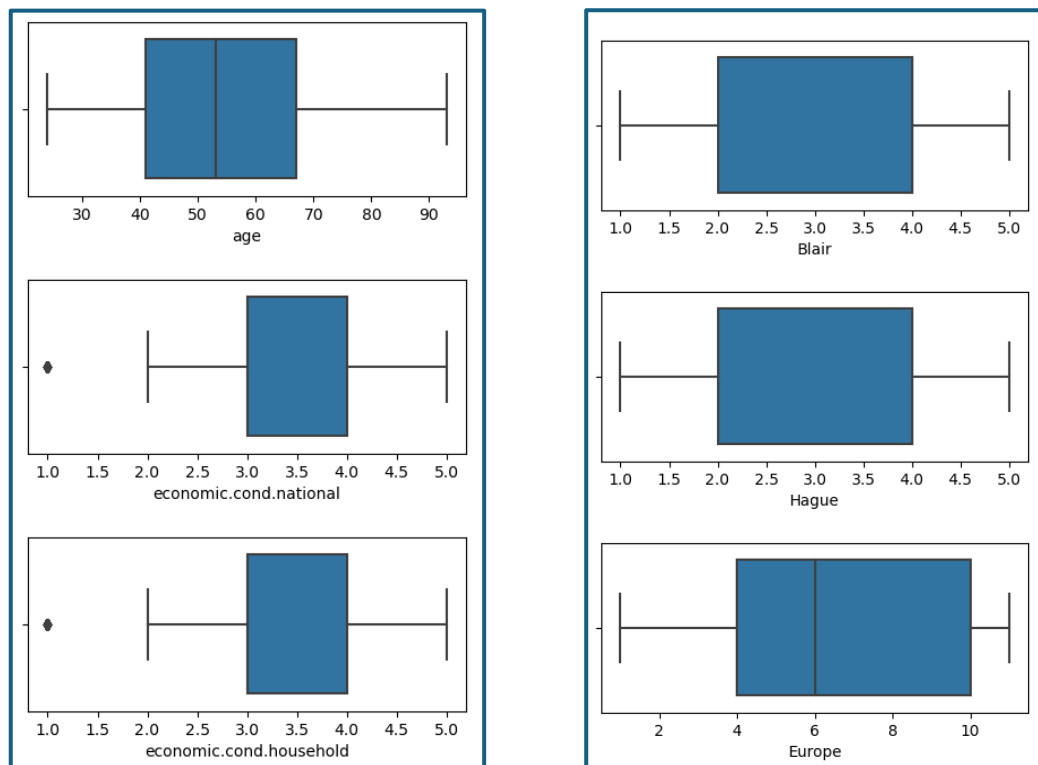


Figure 1.2: Univariate analysis – numerical variable

- Expect variables 'economic.cond.national' and 'economic.cond.household' all the other variables does not have a outlier
- Variable political knowledge does not have lower limit - as there are people with 0 political knowledge
- Variables 'economic.cond.national', 'economic.cond.household', 'Blair' and 'Hauge' have the median and 25% as similar values
- Variable Europe Mean is greater than median which is right/positively skewed

1.1.6.2 Categorical Variable

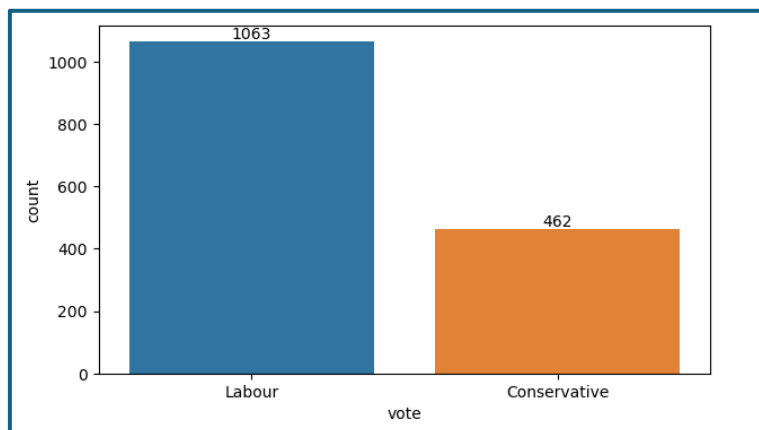


Figure 1.3: Univariate analysis – categorical variable - Vote

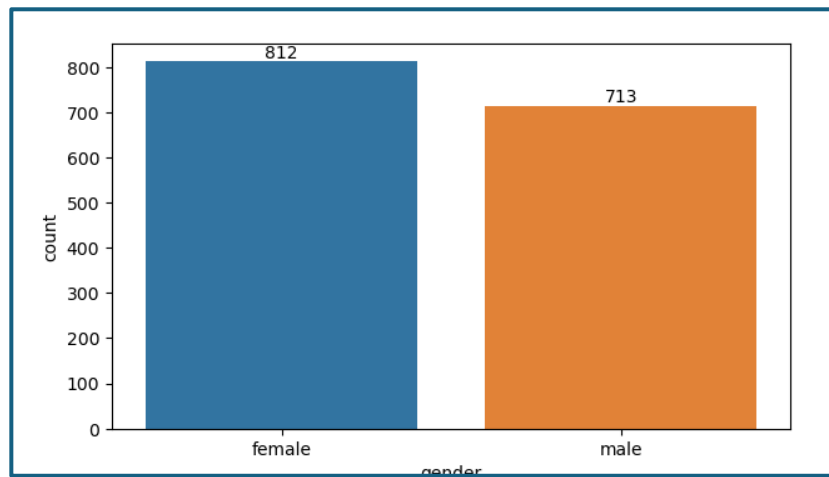


Figure 1.4: Univariate analysis – categorical variable - Gender

- The labour party voters are higher up to 70%
- Gender ratio is almost the same

1.1.7 Multivariate analysis

1.1.7.1 Numerical (all) vs Categorical (Vote)

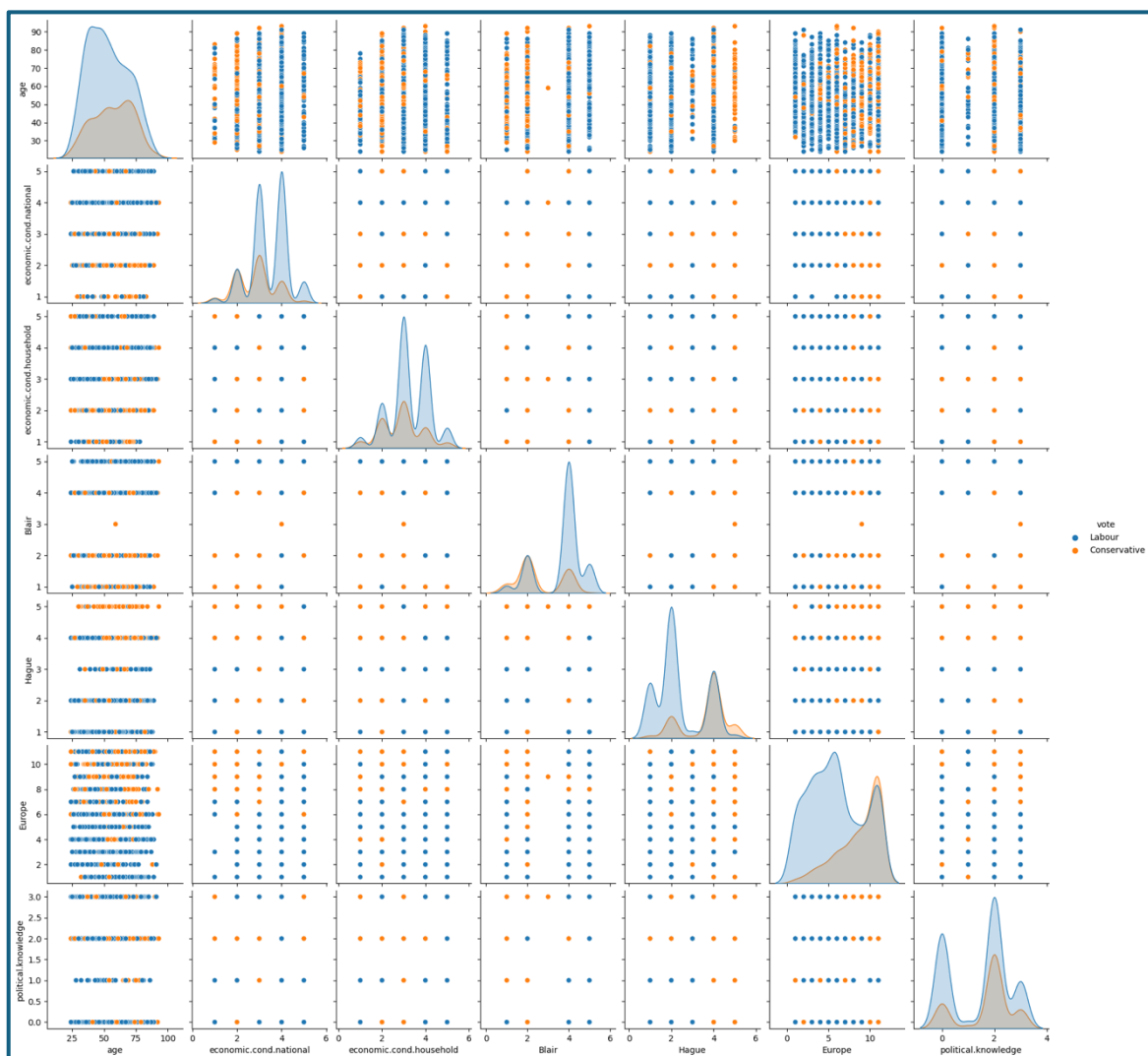


Figure 1.5: Multivariate analysis – numerical vs categorical

1.1.7.1 Numerical (all) vs Categorical (Gender)

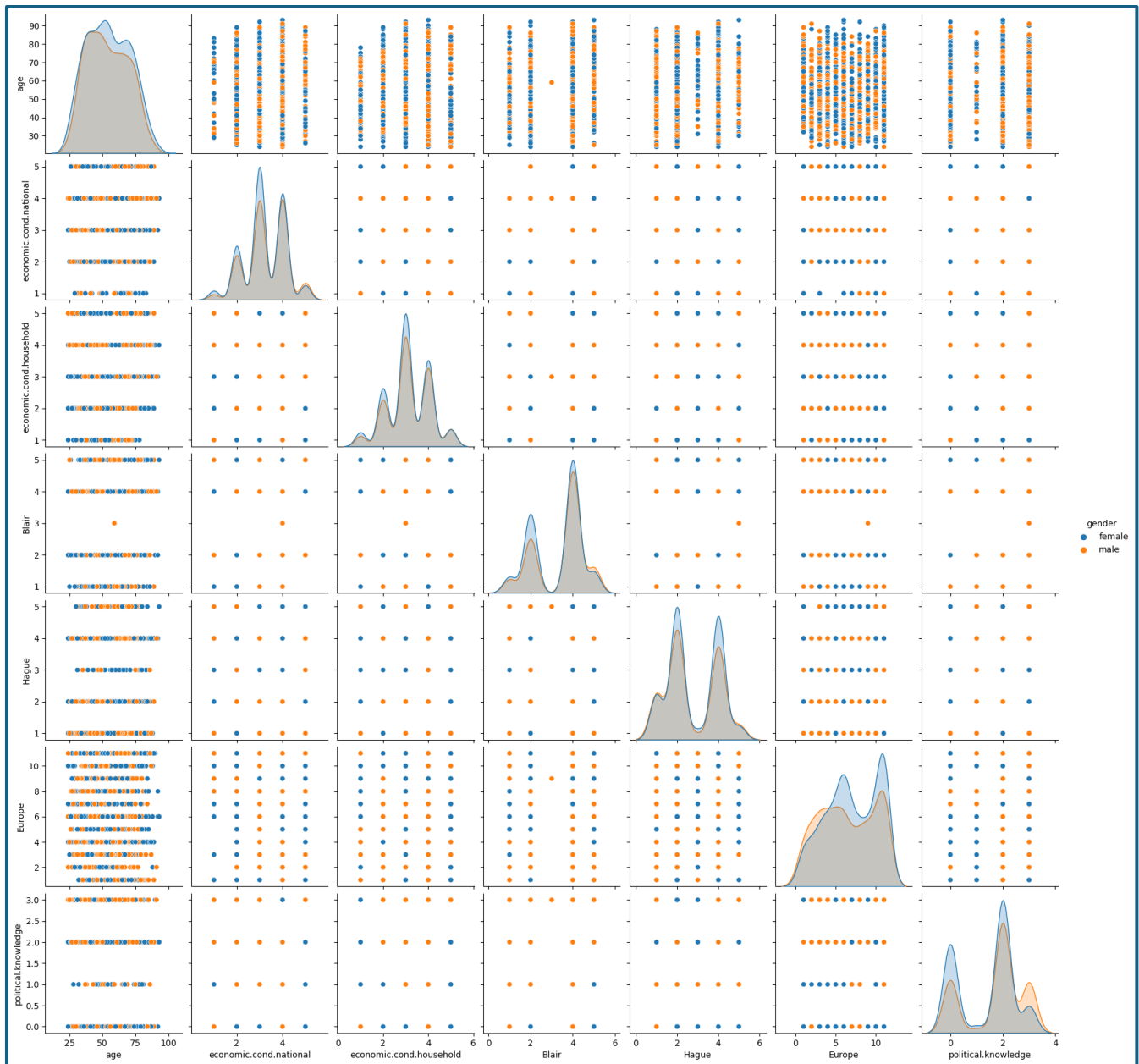


Figure 1.6: Multivariate analysis – numerical (all) vs categorical (gender)

1.1.8 Patterns and insights - Key meaningful observations on individual

Univariate analysis:

- Expect variables 'economic.cond.national' and 'economic.cond.household' all the other variables does not have a outlier
- Variable political knowledge does not have lower limit - as there are people with 0 political knowledge
- Variables 'economic.cond.national', 'economic.cond.household', 'Blair' and 'Hauge' have the median and 25% as similar values
- Variable Europe Mean is greater than median which is right/positively skewed

- The labour party voters are higher up to 70%
- Gender ratio is almost the same

Multivariate analysis:

- As there are not patterns recognized
- In gender as a huge there is not significant difference on the variables to differentiate
- Labour & Conservative voters can be differentiated/classified for all the variables significantly

1.2 Data Pre-processing

1.2.1 Outlier Detection(treat, if needed)

1.2.1.1 Before Outlier Treatment

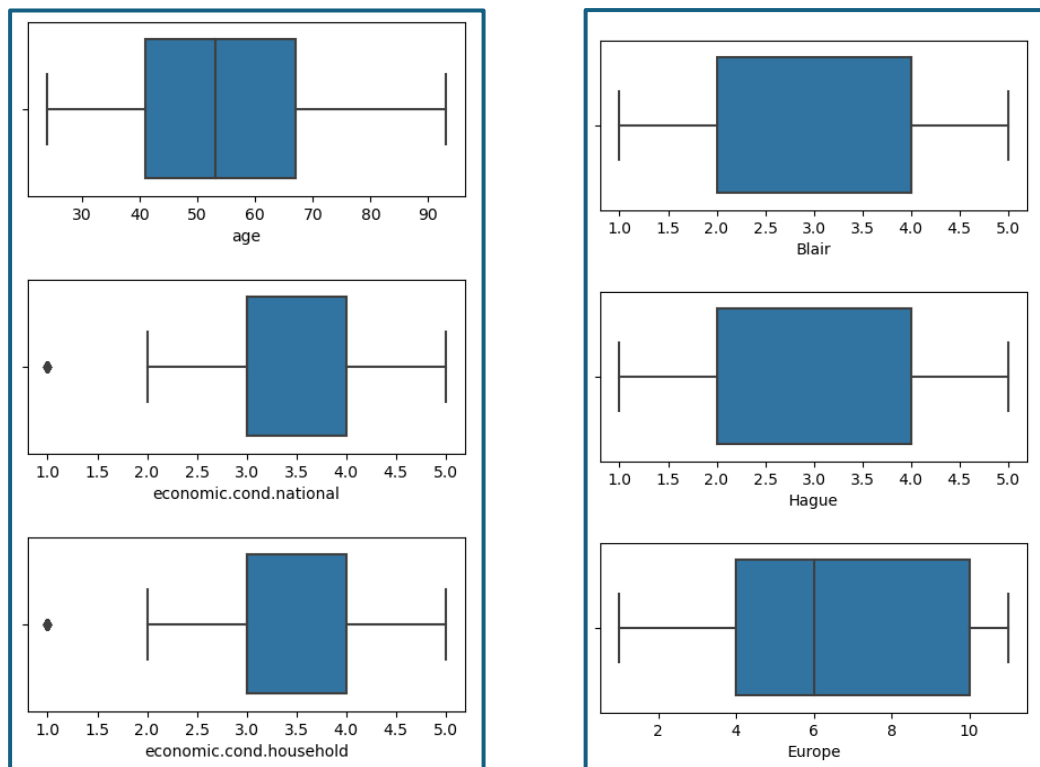


Figure 1.7: Before Outlier Treatment

1.2.1.2 After Outlier Treatment

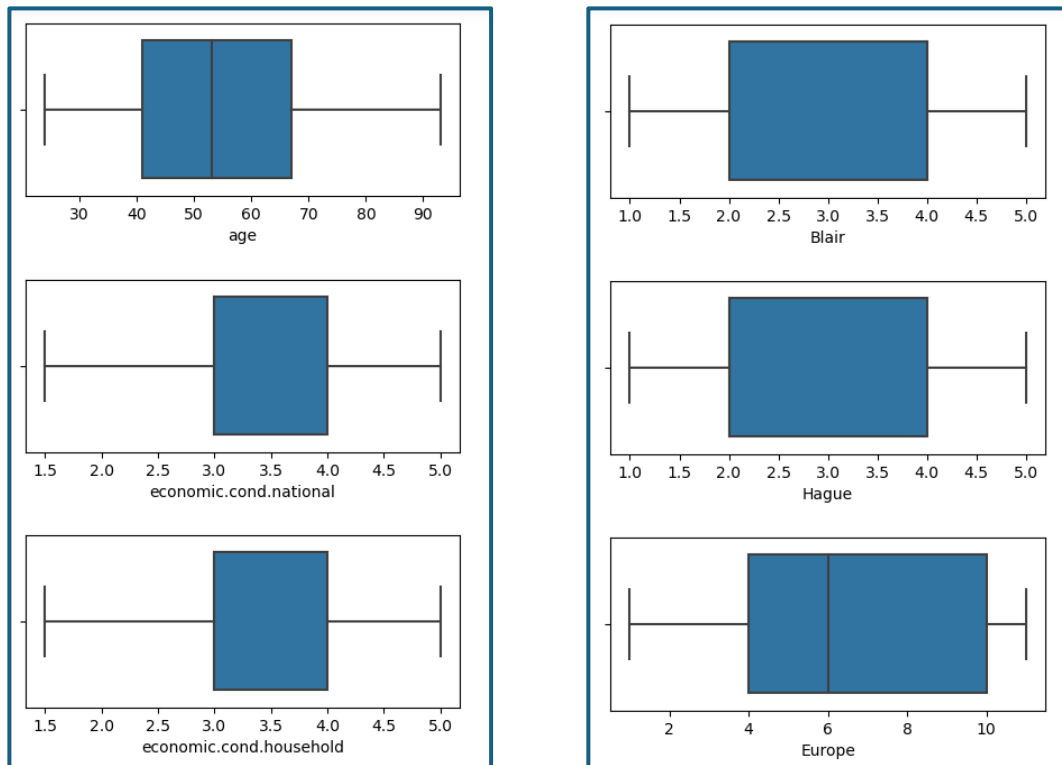


Figure 1.8: After Outlier Treatment

1.2.2 Encode the data

The object variables are encode to 0 & 1 with the datatype as categorical

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                -
0   vote                                  1525 non-null   int8
1   age                                  1525 non-null   float64
2   economic.cond.national               1525 non-null   float64
3   economic.cond.household              1525 non-null   float64
4   Blair                                1525 non-null   float64
5   Hague                                1525 non-null   float64
6   Europe                                1525 non-null   float64
7   political.knowledge                   1525 non-null   float64
8   gender                                1525 non-null   int8
dtypes: float64(7), int8(2)
memory usage: 86.5 KB
```

Figure 1.9: Datatypes after encoding

1.2.3 Data split

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	43.0	3.0	3.0	4.0	1.0	2.0	2.0	0
1	36.0	4.0	4.0	4.0	4.0	5.0	2.0	1
2	35.0	4.0	4.0	5.0	2.0	3.0	2.0	1
3	24.0	4.0	2.0	2.0	1.0	4.0	0.0	0
4	41.0	2.0	2.0	1.0	1.0	6.0	2.0	1

Table 1.4: Independent variables after data split

```

0    1
1    1
2    1
3    1
4    1
Name: vote, dtype: int8

```

Table 1.5: Dependent variables after data split

1.2.4 Scale the data (and state your reasons for scaling the features)

The variables age and Europe is not in the same scale of other variables. So, scaling data will ensure all the variables are in the same range.

	count	mean	std	min	25%	50%	75%	max
age	1525.0	1.013397e-16	1.000328	-1.921698	-0.839313	-0.075276	0.816100	2.471512
economic.cond.national	1525.0	8.386734e-17	1.000328	-2.061826	-0.302622	-0.302622	0.870182	2.042985
economic.cond.household	1525.0	-1.258010e-16	1.000328	-1.877568	-0.182644	-0.182644	0.947305	2.077254
Blair	1525.0	1.677347e-16	1.000328	-1.987695	-1.136225	0.566716	0.566716	1.418187
Hague	1525.0	1.164824e-17	1.000328	-1.419886	-0.607076	-0.607076	1.018544	1.831354
Europe	1525.0	-1.327900e-16	1.000328	-1.737782	-0.827714	-0.221002	0.992422	1.295778
political.knowledge	1525.0	-8.153769e-17	1.000328	-1.424148	-1.424148	0.422643	0.422643	1.346038
gender	1525.0	-5.125226e-17	1.000328	-0.937059	-0.937059	-0.937059	1.067169	1.067169

Table 1.6: Data after Zscore scaling

1.3 Model Building

1.3.1 Model Building (KNN, Naive bayes, Bagging, Boosting)

1.3.1.1 Train Test Split

The data is split into train and test using the train test split function with test size as 0.30

	Train Precision	Test Precision	Train Recall	Test Recall	Train F1_score	Test F1_score	Train Accuracy	Test Accuracy
Naive Bayes	0.88	0.89	0.88	0.86	0.88	0.87	0.83	0.82
KNN Classifier	0.89	0.88	0.92	0.87	0.90	0.87	0.86	0.82
Bagging	0.99	0.88	0.99	0.84	0.99	0.86	0.98	0.80
AdaBoosting	0.87	0.88	0.90	0.88	0.89	0.88	0.84	0.82
Gradient Boosting	0.91	0.89	0.93	0.87	0.92	0.88	0.89	0.83
XG Boosting	0.99	0.87	1.00	0.85	0.99	0.86	0.99	0.80

Table 1.7: Model Building with Metrics

1.3.2 Metrics of Choice (Justify the evaluation metrics)

Recall:

- Since the data is not a balanced one, as the target variable is skewed towards one result, we can go with Precision & recall instead of Accuracy.
- Since Precision and Recall are inversely proportional to each other we can choose F1 score for the evaluation metrics
- High **recall** would be important since the primary concern is to ensure that all supporters of the party are captured accurately in the exit poll results

1.4 Model Performance evaluation

1.4.1 Check the confusion matrix

Confusion Matrix for GaussianNB():

```
[[ 94  36]
 [ 45 283]]
```

Confusion Matrix for KNeighborsClassifier():

```
[[ 91  39]
 [ 44 284]]
```

Confusion Matrix for BaggingClassifier():

```
[[ 86  44]
 [ 56 272]]
```


Confusion Matrix for AdaBoostClassifier():

```
[[ 90  40]
 [ 41 287]]
```

Confusion Matrix for GradientBoostingClassifier():

```
[[ 96  34]
 [ 43 285]]
```

Confusion Matrix for XGBClassifier(base_score=None, booster=None, callbacks=None, colsample_bylevel=None, colsample_bynode=None, colsample_bytree=None, device=None, early_stopping_rounds=None, enable_categorical=False, eval_metric=None, feature_types=None, gamma=None, grow_policy=None, importance_type=None, interaction_constraints=None, learning_rate=None, max_bin=None, max_cat_threshold=None, max_cat_to_onehot=None, max_delta_step=None, max_depth=None, max_leaves=None, min_child_weight=None, missing=nan, monotone_constraints=None, multi_strategy=None, n_estimators=None, n_jobs=None, num_parallel_tree=None, random_state=None, ...):

```
[[ 88  42]
 [ 50 278]]
```

1.4.2 classification metrics for all the models (for both train and test dataset)

Classification Report for GaussianNB(): Train Data

	precision	recall	f1-score	support
0	0.73	0.72	0.73	332
1	0.88	0.88	0.88	735
accuracy			0.83	1067
macro avg	0.80	0.80	0.80	1067
weighted avg	0.83	0.83	0.83	1067

Classification Report for GaussianNB(): Test Data

	precision	recall	f1-score	support
0	0.68	0.72	0.70	130
1	0.89	0.86	0.87	328
accuracy			0.82	458
macro avg	0.78	0.79	0.79	458
weighted avg	0.83	0.82	0.82	458

Classification Report for KNeighborsClassifier(): Train Data

	precision	recall	f1-score	support
0	0.80	0.75	0.77	332
1	0.89	0.92	0.90	735
accuracy			0.86	1067
macro avg	0.85	0.83	0.84	1067
weighted avg	0.86	0.86	0.86	1067

Classification Report for KNeighborsClassifier(): Test Data

	precision	recall	f1-score	support
0	0.67	0.70	0.69	130
1	0.88	0.87	0.87	328
accuracy			0.82	458
macro avg	0.78	0.78	0.78	458
weighted avg	0.82	0.82	0.82	458

Classification Report for BaggingClassifier(): Train Data

	precision	recall	f1-score	support
0	0.97	0.97	0.97	332
1	0.99	0.99	0.99	735
accuracy			0.98	1067
macro avg	0.98	0.98	0.98	1067
weighted avg	0.98	0.98	0.98	1067

Classification Report for BaggingClassifier(): Test Data

	precision	recall	f1-score	support
0	0.63	0.71	0.67	130
1	0.88	0.84	0.86	328
accuracy			0.80	458
macro avg	0.76	0.77	0.76	458
weighted avg	0.81	0.80	0.80	458

Classification Report for AdaBoostClassifier(): Train Data

	precision	recall	f1-score	support
0	0.77	0.71	0.74	332
1	0.87	0.90	0.89	735
accuracy			0.84	1067
macro avg	0.82	0.81	0.81	1067
weighted avg	0.84	0.84	0.84	1067

Classification Report for AdaBoostClassifier(): Test Data

	precision	recall	f1-score	support
0	0.69	0.69	0.69	130
1	0.88	0.88	0.88	328
accuracy			0.82	458
macro avg	0.78	0.78	0.78	458
weighted avg	0.82	0.82	0.82	458

Classification Report for GradientBoostingClassifier(): Train Data

	precision	recall	f1-score	support
0	0.84	0.79	0.81	332
1	0.91	0.93	0.92	735
accuracy			0.89	1067
macro avg	0.87	0.86	0.87	1067
weighted avg	0.89	0.89	0.89	1067

Classification Report for GradientBoostingClassifier(): Test Data

	precision	recall	f1-score	support
0	0.69	0.74	0.71	130
1	0.89	0.87	0.88	328
accuracy			0.83	458
macro avg	0.79	0.80	0.80	458
weighted avg	0.84	0.83	0.83	458

Classification Report for XGBClassifier(base_score=None, booster=None, callbacks=None,

```
colsample_bylevel=None, colsample_bynode=None,
colsample_bytree=None, device=None, early_stopping_rounds=None,
enable_categorical=False, eval_metric=None, feature_types=None,
gamma=None, grow_policy=None, importance_type=None,
interaction_constraints=None, learning_rate=None, max_bin=None,
max_cat_threshold=None, max_cat_to_onehot=None,
max_delta_step=None, max_depth=None, max_leaves=None,
min_child_weight=None, missing=nan, monotone_constraints=None,
multi_strategy=None, n_estimators=None, n_jobs=None,
num_parallel_tree=None, random_state=None, ...): Train Data
```

	precision	recall	f1-score	support
0	0.99	0.98	0.98	332
1	0.99	1.00	0.99	735
accuracy			0.99	1067
macro avg	0.99	0.99	0.99	1067
weighted avg	0.99	0.99	0.99	1067

Classification Report for XGBClassifier(base_score=None, booster=None, callbacks=None,

```

colsample_bylevel=None, colsample_bynode=None,
colsample_bytree=None, device=None, early_stopping_rounds=None,
enable_categorical=False, eval_metric=None, feature_types=None,
gamma=None, grow_policy=None, importance_type=None,
interaction_constraints=None, learning_rate=None, max_bin=None,
max_cat_threshold=None, max_cat_to_onehot=None,
max_delta_step=None, max_depth=None, max_leaves=None,
min_child_weight=None, missing=nan, monotone_constraints=None,
multi_strategy=None, n_estimators=None, n_jobs=None,
num_parallel_tree=None, random_state=None, ...): Test Data

```

	precision	recall	f1-score	support
0	0.64	0.68	0.66	130
1	0.87	0.85	0.86	328
accuracy			0.80	458
macro avg	0.75	0.76	0.76	458
weighted avg	0.80	0.80	0.80	458

1.4.3 ROC-AUC score and plot the curve

1.4.3.1 ROC-AUC Train data

```

AUC : 0.887 GaussianNB()
AUC : 0.930 KNeighborsClassifier()
AUC : 0.999 BaggingClassifier()
AUC : 0.910 AdaBoostClassifier()
AUC : 0.950 GradientBoostingClassifier()
AUC : 1.000 XGBClassifier

```

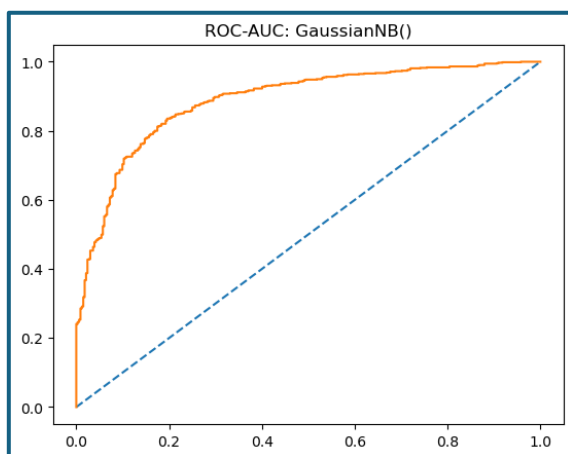


Figure 1.10: ROC-AUC GaussianNB

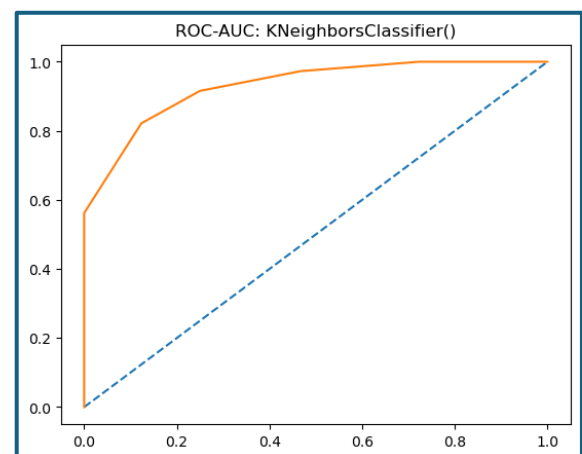


Figure 1.11: ROC-AUC KNeighborsClassifier

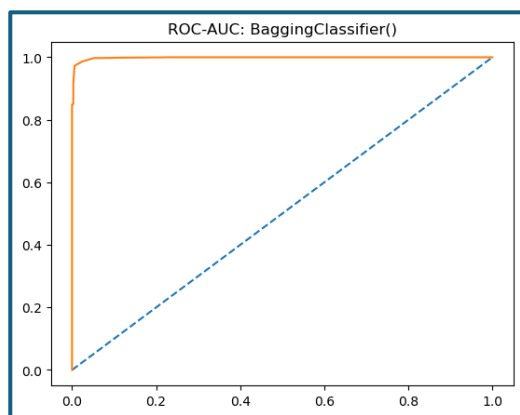


Figure 1.12: ROC-AUC BaggingClassifier

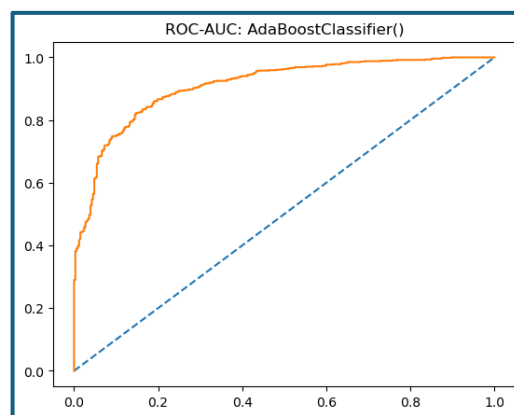


Figure 1.13: ROC-AUC AdaBoostClassifier

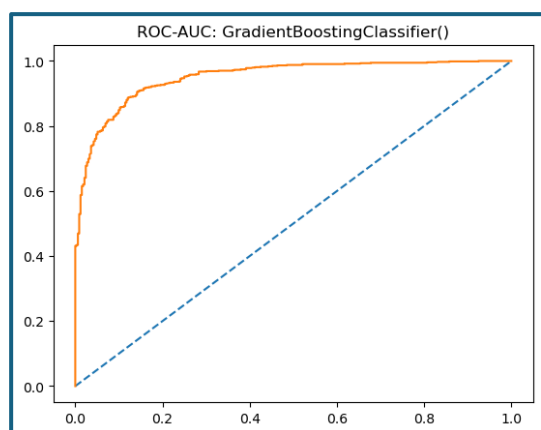


Figure 1.14: ROC-AUC GradientBoostingClassifier

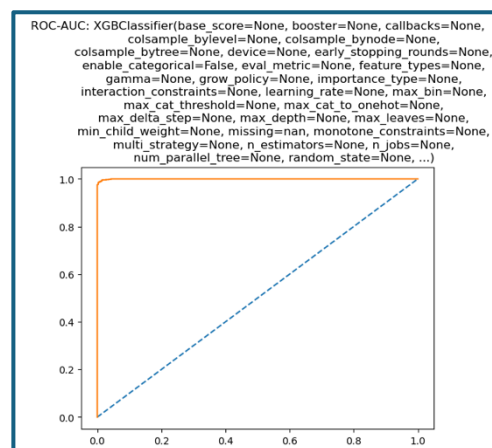


Figure 1.15: ROC-AUC XGBClassifier

1.4.3.2 ROC-AUC Test data

```
AUC : 0.885 GaussianNB()
AUC : 0.869 KNeighborsClassifier()
AUC : 0.850 BaggingClassifier()
AUC : 0.880 AdaBoostClassifier()
AUC : 0.904 GradientBoostingClassifier()
AUC : 0.863 XGBClassifier
```

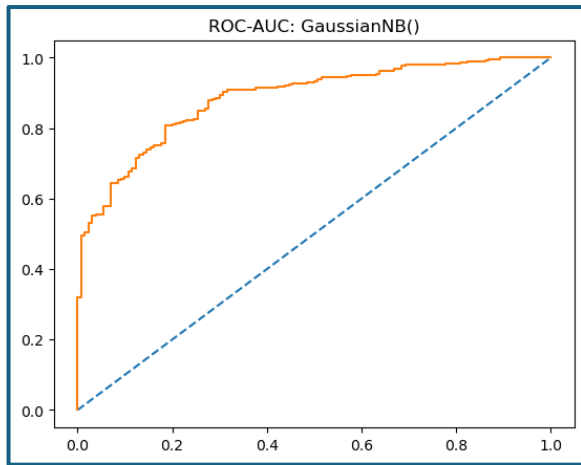


Figure 1.16: ROC-AUC GaussianNB

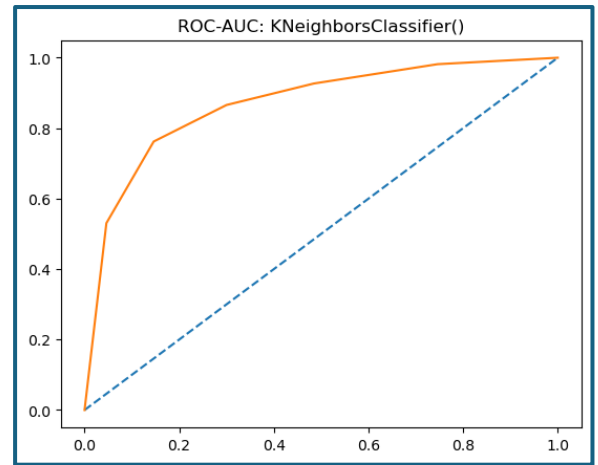


Figure 1.17: ROC-AUC KNeighborsClassifier

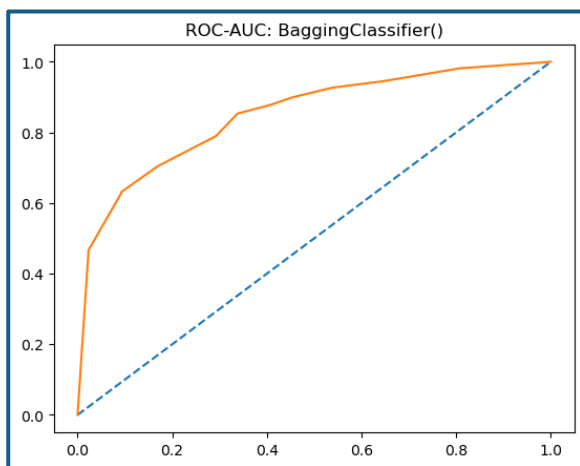


Figure 1.18: ROC-AUC BaggingClassifier

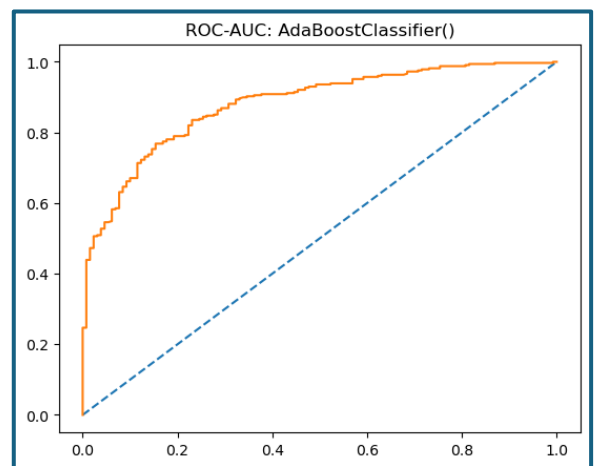


Figure 1.19: ROC-AUC AdaBoostClassifier

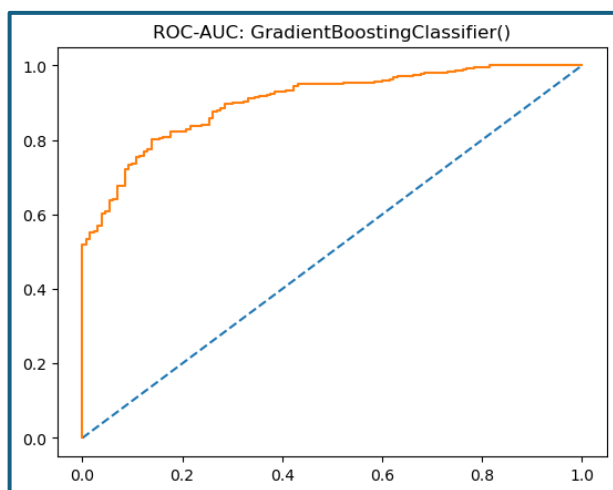


Figure 1.20: ROC-AUC GradientBoostClassifier

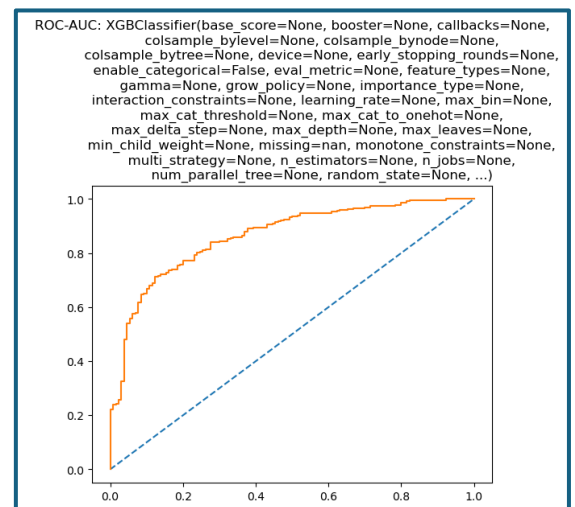


Figure 1.21: ROC-AUC XGBoostClassifier

1.4.4 Comment on all the model performance

Naive Bayes:

The recall difference the train(0.88) and test(0.86) data is 0.02. Which is one of the best model in terms of lesser recall difference. The F1 score difference is 0.01. 45 labour votes are identified as conservative (0.16) The AUC score difference is 0.002

KNN Classifier

The recall difference the train(0.88) and test(0.92) data is 0.04. The F1 score difference is 0.03. 44 labour votes are identified as conservative(0.16) The AUC score difference is 0.061

Bagging

The recall difference the train(0.99) and test(0.86) data is 0.13. The F1 score difference is 0.04. 52 labour votes are identified as conservative(0.19) The AUC score difference is 0.154

AdaBoosting

The recall difference the train(0.90) and test(0.88) data is 0.02. Which is one of the best model in terms of lesser recall difference. The F1 score difference is 0.04. 41 labour votes are identified as conservative(0.14) The AUC score difference is 0.065

Gradient Boosting

The recall difference the train(0.93) and test(0.87) data is 0.06. The F1 score difference is 0.04. 43 labour votes are identified as conservative(0.15) The AUC score difference is 0.046

XG Boosting

The recall difference the train(1.0) and test(0.85) data is 0.15. The F1 score difference is 0.13. 50 labour votes are identified as conservative(0.18) The AUC score difference is 0.137

1.5 Model Performance improvement

1.5.1 Improve the model performance of bagging and boosting models by tuning the model

1.5.1.1 Hypertuning – Bagging

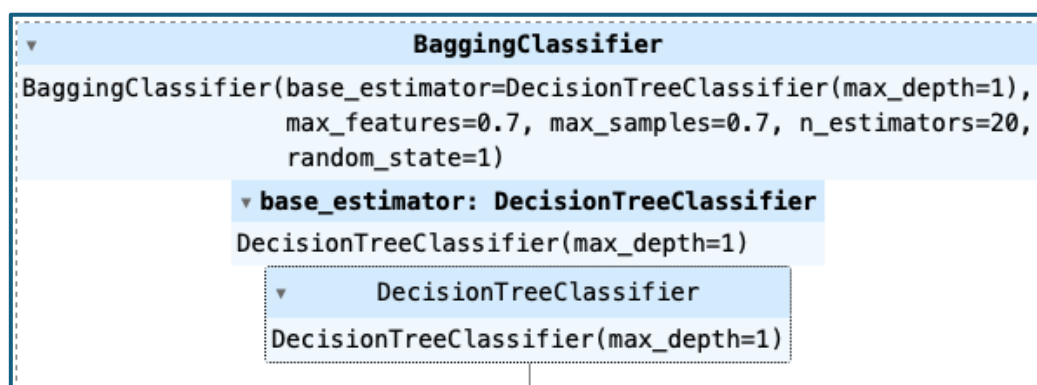


Figure 1.22: Hypertuning - Bagging

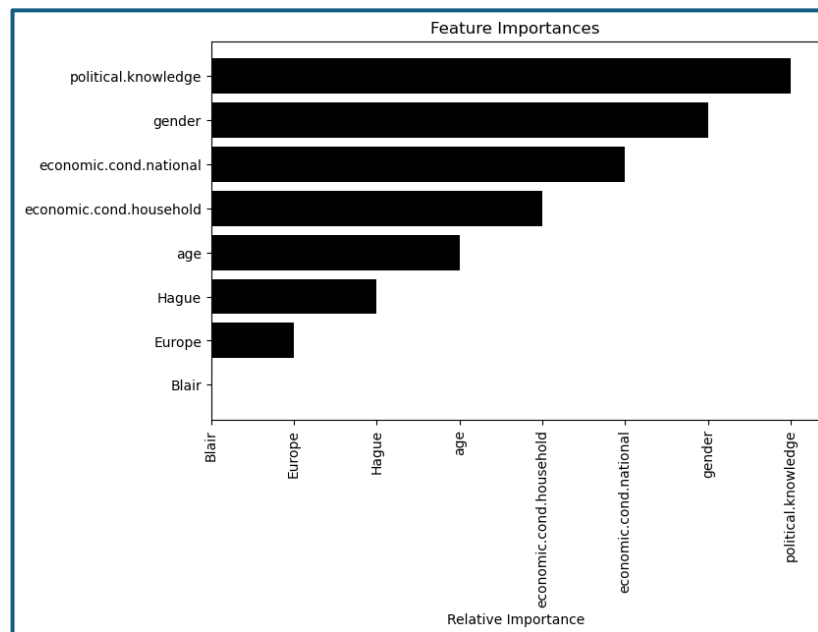


Figure 1.23: Feature Importance - Bagging

1.5.1.2 Hypertuning – AdaBoosting

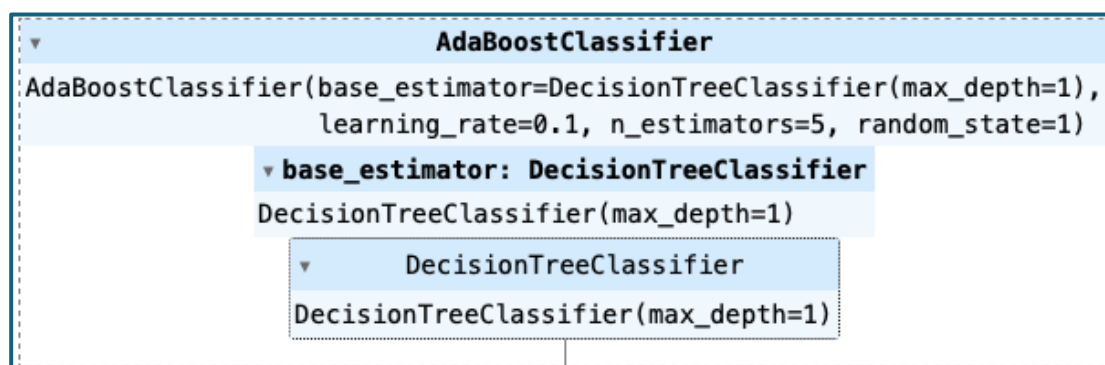


Figure 1.24: Hypertuning – AdaBoosting

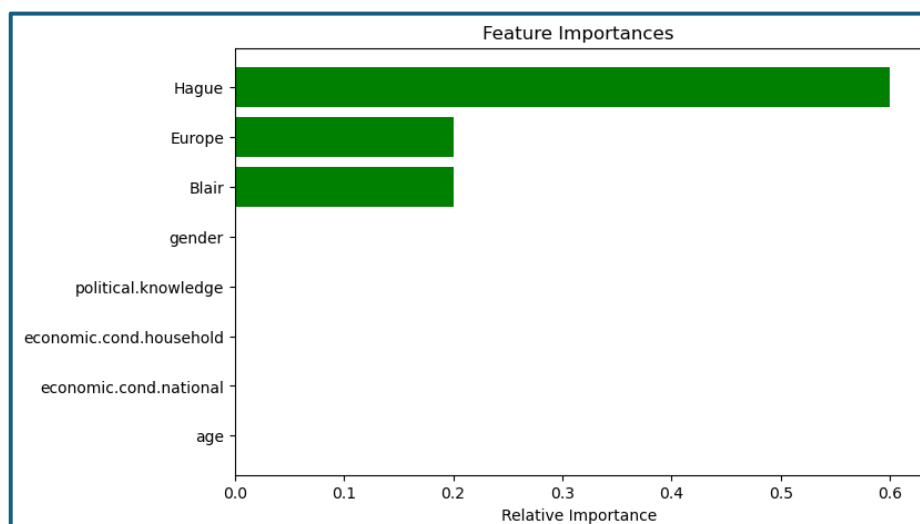


Figure 1.25: Feature Importance - AdaBoosting

1.5.1.3 Hypertuning - Gradient Boosting

```
▼ GradientBoostingClassifier
GradientBoostingClassifier(init=AdaBoostClassifier(random_state=1),
                           max_features=1, n_estimators=20, random_state=1,
                           subsample=0.8)
  ▼ init: AdaBoostClassifier
  AdaBoostClassifier(random_state=1)
    ▼ AdaBoostClassifier
    AdaBoostClassifier(random_state=1)
```

Figure 1.26: Hypertuning – Gradient Boosting

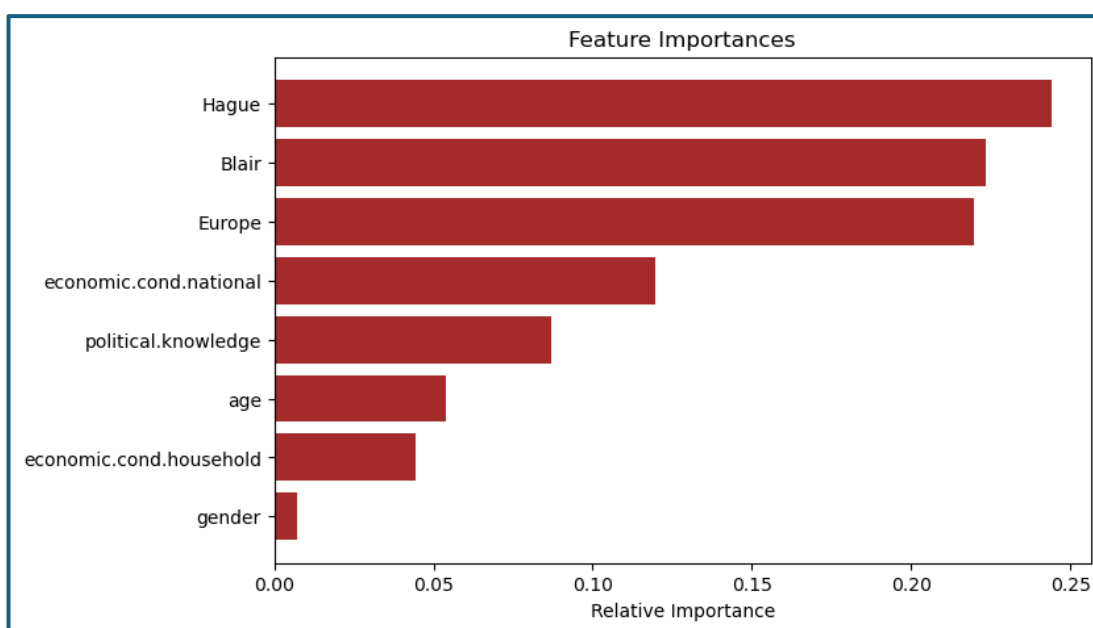


Figure 1.27: Feature Importance – Gradient Boosting

1.5.1.4 Hypertuning - XG Boosting

```
▼ XGBClassifier
colsample_bylevel=0.5, colsample_bynode=None,
colsample_bytree=0.5, device=None, early_stopping_rounds=None,
enable_categorical=False, eval_metric=None, feature_types=None,
gamma=0, grow_policy=None, importance_type=None,
interaction_constraints=None, learning_rate=0.01, max_bin=None,
max_cat_threshold=None, max_cat_to_onehot=None,
max_delta_step=None, max_depth=None, max_leaves=None,
min_child_weight=None, missing=nan, monotone_constraints=None,
multi_strategy=None, n_estimators=10, n_jobs=None,
num_parallel_tree=None, random_state=1, ...)
```

Figure 1.28: Hypertuning – XG Boosting

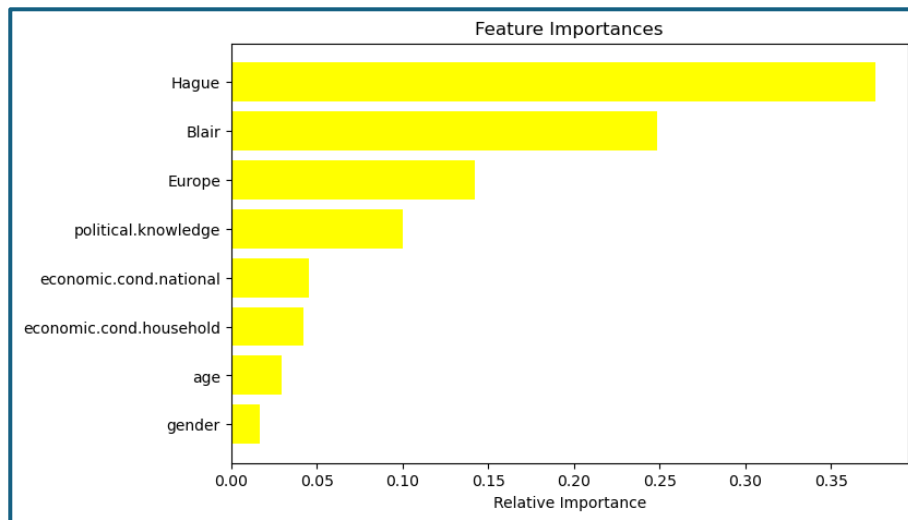


Figure 1.29: Feature Importance – XG Boosting

	Train Precision	Test Precision	Train Recall	Test Recall	Train F1_score	Test F1_score	Train Accuracy	Test Accuracy
Bagging	0.78	0.82	0.95	0.95	0.86	0.88	0.79	0.81
AdaBoosting Tuned	0.78	0.81	0.96	0.96	0.86	0.88	0.78	0.81
Gradient Boosting Tuned	0.83	0.84	0.96	0.92	0.89	0.88	0.84	0.82
XG Boosting Tuned	0.69	0.72	1.00	1.00	0.82	0.83	0.69	0.72

Table 1.8: : Model Building after Hypertuning

1.5.2 Comment on the model performance improvement on training and test data

Bagging:

The recall for train decreased from 0.99 to 0.95 and test increased from 0.86 to 0.95 and the recall difference for the train & test is 0.00 but before tuning it was 0.02. The hyperparameter tuning for Bagging increased the model performance

AdaBoosting:

The recall for train increased from 0.90 to 0.96 and test increased from 0.88 to 0.96 and the recall difference for the train & test is 0.00 but before tuning it was 0.02. The hyperparameter tuning for AdaBoosting increased the model performance

Gradient Boosting:

The recall for train increased from 0.93 to 0.96 and test increased from 0.87 to 0.92 and the recall difference for the train & test is 0.06 but before tuning it was 0.05. The hyperparameter tuning for Gradient Boosting does not show significant improvement in the model performance

XG Boosting:

The recall for train remain same as 1.00 and test increased from 0.85 to 1.00 and the recall difference for the train & test is 0.00 but before tuning it was 0.15. The hyperparameter tuning for XG Boosting increased the model performance

1.6 Final Model Selection

1.6.1 Compare all the model built so far

	Train Precision	Test Precision	Train Recall	Test Recall	Train F1_score	Test F1_score	Train Accuracy	Test Accuracy
Naive Bayes	0.88	0.89	0.88	0.86	0.88	0.87	0.83	0.82
KNN Classifier	0.89	0.88	0.92	0.87	0.90	0.87	0.86	0.82
Bagging	0.99	0.86	0.99	0.85	0.99	0.86	0.98	0.80
Bagging Tuned	0.78	0.82	0.95	0.95	0.86	0.88	0.79	0.81
AdaBoosting	0.87	0.88	0.90	0.88	0.89	0.88	0.84	0.82
AdaBoosting Tuned	0.78	0.81	0.96	0.96	0.86	0.88	0.78	0.81
Gradient Boosting	0.91	0.89	0.93	0.87	0.92	0.88	0.89	0.83
Gradient Boosting Tuned	0.83	0.84	0.96	0.92	0.89	0.88	0.84	0.82
XG Boosting	0.99	0.87	1.00	0.85	0.99	0.86	0.99	0.80
XG Boosting Tuned	0.69	0.72	1.00	1.00	0.82	0.83	0.69	0.72

Table 1.9: : Model Comparison after Hypertuning

1.6.2 Select the final model with the proper justification

Confusion Matrix for XGBCL tuned:

```
[[ 0 130]
 [ 0 328]]
```

Classification Report for XGBCL tuned: Test Data

	precision	recall	f1-score	support
0	0.00	0.00	0.00	130
1	0.72	1.00	0.83	328
accuracy			0.72	458
macro avg	0.36	0.50	0.42	458
weighted avg	0.51	0.72	0.60	458

XG Boosting Tuned:

Before Tuning:

The recall difference between the train(1.00) and test(0.85) data is 0.15. The F1 score difference is 0.13. 50 labour votes are identified as conservative(0.18) The AUC score difference is 0.137

After Tuning:

The recall for train remain same as 1.00 and test increased from 0.85 to 1.00 and the recall difference for the train & test is 0.00 but before tuning it was 0.15. The hyperparameter tuning for XG Boosting increased the model performance significantly

1.6.3 Check the most important features in the final model and draw inferences

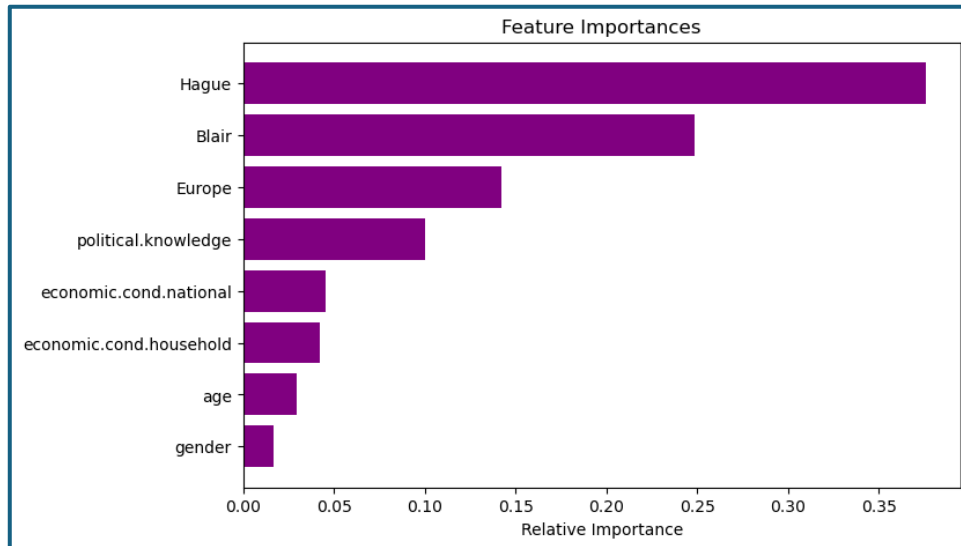


Figure 1.30: Feature Importance – XG Boosting Tuned

1.7 Actionable Insights & Recommendations

1.7.1 Compare all four models

	Train Precision	Test Precision	Train Recall	Test Recall	Train F1_score	Test F1_score	Train Accuracy	Test Accuracy
Naive Bayes	0.88	0.89	0.88	0.86	0.88	0.87	0.83	0.82
KNN Classifier	0.89	0.88	0.92	0.87	0.90	0.87	0.86	0.82
Bagging Tuned	0.78	0.82	0.95	0.95	0.86	0.88	0.79	0.81
AdaBoosting Tuned	0.78	0.81	0.96	0.96	0.86	0.88	0.78	0.81
Gradient Boosting Tuned	0.83	0.84	0.96	0.92	0.89	0.88	0.84	0.82
XG Boosting Tuned	0.69	0.72	1.00	1.00	0.82	0.83	0.69	0.72

Table 1.10: : Model Comparison

1.7.2 Conclude with the key takeaways for the business

- Across all the models features the Hague > Blair > Europe variable list at the top 3. The voters are more concerned about the leaders and the European Integration sentiment.
- Since the dataset is skewed towards the labours output with 70% then all the predictions will tend to skew towards the same.
- As the age factor plays an important role in the exit polls. The mean age is 54 and most of them are the voters of labour party. Which indicates more the age group more change for the labour party to win
- Voters with political Knowledge plays an important role on choosing the right party who can impact on the economic conditions on the nation and the household

Problem 2

2.1 Define the problem and Perform Exploratory Data Analysis

2.1.1 Problem Definition

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

2.1.2 Find the number of Character, words & sentences in all three speeches

	Speech	Characters	Words	Sentences
0	On each national day of inauguration since 178...	7651	1323	69
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	7673	1364	56
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10106	1769	70

Table 2.1: Char, Word, Sent. Count of all three speeches

2.2 Text cleaning

2.2.1 Stopword removal

Importing stops words from the necessary library

Roosevelt Speech:

```
['national', 'day', 'inauguration', 'since', '1789', 'people', 'renewed', 'sense',  
'dedication', 'united', 'statesnnin', 'washingtons', 'day', 'task', 'people', 'crea  
te', 'weld', 'together', 'nationnnnin', 'lincolns', 'day', 'task', 'people', 'preser  
ve', 'nation', 'disruption', 'withinnnin', 'day', 'task', 'people', 'save', 'nation  
, 'institutions', 'disruption', 'withoutnnto', 'us', 'come', 'time', 'midst', 'swi  
ft', 'happenings', 'pause', 'moment', 'take', 'stock', 'recall', 'place', 'history'  
, 'rediscover', 'may', 'risk', 'real', 'peril', 'inactionnnlives', 'nations', 'dete  
rmined', 'count', 'years', 'lifetime', 'human', 'spirit', 'life', 'man', 'threescor  
e', 'years', 'ten', 'little', 'little', 'less', 'life', 'nation', 'fullness', 'meas  
ure', 'livennthere', 'men', 'doubt', 'men', 'believe', 'democracy', 'form', 'govern  
ment', 'frame', 'life', 'limited', 'measured', 'kind', 'mystical', 'artificial', 'f  
ate', 'unexplained', 'reason', 'tyranny', 'slavery', 'become', 'surging', 'wave', '  
future', 'freedom', 'ebbing', 'tidennbut', 'americans', 'know', 'truenneight', 'yea  
rs', 'ago', 'life', 'republic', 'seemed', 'frozen', 'fatalistic', 'terror', 'proved  
, 'true', 'midst', 'shock', 'acted', 'acted', 'quickly', 'boldly', 'decisivelynnth  
ese', 'later', 'years', 'living', 'years', 'fruitful', 'years', 'people', 'democrac  
y', 'brought', 'us', 'greater', 'security', 'hope', 'better', 'understanding', 'lif  
es', 'ideals', 'measured', 'material', 'thingsnnmost', 'vital', 'present', 'future'  
, 'experience', 'democracy', 'successfully', 'survived', 'crisis', 'home', 'put', '  
away', 'many', 'evil', 'things', 'built', 'new', 'structures', 'enduring', 'lines',
```

'maintained', 'fact', 'democracynnfor', 'action', 'taken', 'within', 'threeway', 'framework', 'constitution', 'united', 'states', 'coordinate', 'branches', 'government', 'continue', 'freely', 'function', 'bill', 'rights', 'remains', 'inviolable', 'freedom', 'elections', 'wholly', 'maintained', 'prophets', 'downfall', 'american', 'democracy', 'seen', 'dire', 'predictions', 'come', 'naughtnndemocracy', 'dyingnnwe', 'know', 'seen', 'reviveand', 'grownnwe', 'know', 'cannot', 'die', 'built', 'unhindered', 'initiative', 'individual', 'men', 'women', 'joined', 'together', 'common', 'enterprise', 'enterprise', 'undertaken', 'carried', 'free', 'expression', 'free', 'majoritynnwe', 'know', 'democracy', 'alone', 'forms', 'government', 'enlists', 'full', 'force', 'mens', 'enlightened', 'willnnwe', 'know', 'democracy', 'alone', 'constructed', 'unlimited', 'civilization', 'capable', 'infinite', 'progress', 'improvement', 'human', 'lifennwe', 'know', 'look', 'surface', 'sense', 'still', 'spreading', 'every', 'continent', 'humane', 'advanced', 'end', 'unconquerable', 'forms', 'human', 'societynna', 'nation', 'like', 'person', 'bodya', 'body', 'must', 'fed', 'clothed', 'housed', 'invigorated', 'rested', 'manner', 'measures', 'objectives', 'time', 'nation', 'like', 'person', 'mind', 'mind', 'must', 'kept', 'informed', 'alert', 'must', 'know', 'understands', 'hopes', 'needs', 'neighbors', 'nations', 'live', 'within', 'narrowing', 'circle', 'worldnnand', 'nation', 'like', 'person', 'something', 'deeper', 'something', 'permanent', 'something', 'larger', 'sum', 'parts', 'something', 'matters', 'future', 'calls', 'forth', 'sacred', 'guarding', 'presentnnit', 'thing', 'find', 'difficult', 'even', 'impossible', 'hit', 'upon', 'single', 'simple', 'wordnnand', 'yet', 'understand', 'spirit', 'faith', 'america', 'product', 'centuries', 'born', 'multitudes', 'came', 'many', 'lands', 'high', 'degree', 'mostly', 'plain', 'people', 'sought', 'early', 'late', 'find', 'freedom', 'freelynnthe', 'democratic', 'aspiration', 'mere', 'recent', 'phase', 'human', 'history', 'human', 'history', 'permeated', 'ancient', 'life', 'early', 'peoples', 'blazed', 'anew', 'middle', 'ages', 'written', 'magna', 'chartannin', 'americas', 'impact', 'irresistible', 'america', 'new', 'world', 'tongues', 'peoples', 'continent', 'newfound', 'land', 'came', 'believed', 'could', 'create', 'upon', 'continent', 'new', 'life', 'life', 'new', 'freedomnnits', 'vitality', 'written', 'mayflower', 'compact', 'declaration', 'independence', 'constitution', 'united', 'states', 'gettysburg', 'addressnnthose', 'first', 'came', 'carry', 'longings', 'spirit', 'millions', 'followed', 'stock', 'sprang', 'moved', 'forward', 'constantly', 'consistently', 'toward', 'ideal', 'gained', 'stature', 'clarity', 'generationnnthe', 'hopes', 'republic', 'cannot', 'forever', 'tolerate', 'either', 'undeserved', 'poverty', 'selfserving', 'wealthnnwe', 'know', 'still', 'far', 'go', 'must', 'greatly', 'build', 'security', 'opportunity', 'knowledge', 'every', 'citizen', 'measure', 'justified', 'resources', 'capacity', 'landnnbut', 'enough', 'achieve', 'purposes', 'alone', 'enough', 'clothe', 'feed', 'body', 'nation', 'instruct', 'inform', 'mind', 'also', 'spirit', 'three', 'greatest', 'spiritnnwithout', 'body', 'mind', 'men', 'know', 'nation', 'could', 'livennbut', 'spirit', 'america', 'killed', 'even', 'though', 'nations', 'body', 'mind', 'constricted', 'alien', 'world', 'lived', 'america', 'know', 'would', 'perishednnthat', 'spirit', 'faith', 'speaks', 'us', 'daily', 'lives', 'ways', 'often', 'unnoticed', 'seem', 'obvious', 'speaks', 'us', 'capital', 'nation', 'speaks', 'us', 'processes', 'governing', 'sovereignties', '48', 'states', 'speaks', 'us', 'counties', 'cities', 'towns', 'villages', 'speaks', 'us', 'nations', 'hemisphere', 'across', 'seas', 'enslaved', 'well', 'free', 'sometimes', 'fail', 'hear', 'heed', 'voices', 'freedom', 'us', 'privilege', 'freedom', 'old', 'old', 'storynnthe', 'destiny', 'america', 'proclaimed', 'words', 'prophecy', 'spoken', 'first', 'president', 'first', 'inaugural', '1789', 'words', 'almost', 'directed', 'would', 'seem', 'year', '1941', 'preservation', 'sacred', 'fire', 'liberty', 'destiny', 'republican', 'model', 'government', 'justly', 'considered', 'deeply', 'finally', 'staked', 'experiment', 'intrusted', 'hands', 'american', 'peoplennif', 'lose', 'sacred', 'fireif', 'let', 'smothered', 'doubt', 'fear', 'shall', 'reject', 'destiny', 'washington', 'strove', 'valiantly', 'triumphantly', 'establish', 'preservation', 'spirit', 'faith', 'nation', 'furnish', 'highest', 'justification', 'every', 'sacrifice', 'may', 'make', 'cause', 'national', 'defensennin', 'face', 'great', 'perils', 'never', 'encountered', 'strong', 'purpose', 'protect', 'perpetuate', 'integrity', 'democracynnfor', 'musterr', 'spirit', 'america', 'faith', 'americannwe', 'retreat', 'content', 'stand', 'still', 'americans', 'go', 'forward', 'service', 'country', 'godn']

2.2.2 Stemming

The stemming action is done using the porter stemmer function

Roosevelt Speech:

['nation', 'day', 'inaugur', 'sinc', '1789', 'peopl', 'renew', 'sens', 'dedic', 'un
it', 'statesnnin', 'washington', 'day', 'task', 'peopl', 'creat', 'weld', 'togeth',
'nationnnnin', 'lincoln', 'day', 'task', 'peopl', 'preserv', 'nation', 'disrupt', 'w
ithinnnin', 'day', 'task', 'peopl', 'save', 'nation', 'institut', 'disrupt', 'witho
utnnto', 'us', 'come', 'time', 'midst', 'swift', 'happen', 'paus', 'moment', 'take',
, 'stock', 'recal', 'place', 'histori', 'rediscover', 'may', 'risk', 'real', 'peril',
'inactionnnnl', 'nation', 'determin', 'count', 'year', 'lifetim', 'human', 'spirit',
'life', 'man', 'threescore', 'year', 'ten', 'littl', 'littl', 'less', 'life', 'natio
n', 'full', 'measur', 'livennther', 'men', 'doubt', 'men', 'believ', 'democraci',
form', 'govern', 'frame', 'life', 'limit', 'measur', 'kind', 'mystic', 'artifici',
'fate', 'unexplain', 'reason', 'tyranni', 'slaveri', 'becom', 'surg', 'wave', 'futu
r', 'freedom', 'eb', 'tidennbut', 'american', 'know', 'truenneight', 'year', 'ago',
'life', 'republ', 'seem', 'frozen', 'fatalist', 'terror', 'prove', 'true', 'midst',
'shock', 'act', 'act', 'quickli', 'boldli', 'decisivelynnthes', 'later', 'year', 'l
ive', 'year', 'fruit', 'year', 'peopl', 'democraci', 'brought', 'us', 'greater', 's
ecur', 'hope', 'better', 'understand', 'life', 'ideal', 'measur', 'materi', 'things
nnmost', 'vital', 'present', 'futur', 'experi', 'democraci', 'success', 'surviv',
'crisi', 'home', 'put', 'away', 'mani', 'evil', 'thing', 'built', 'new', 'structur',
'endur', 'line', 'maintain', 'fact', 'democracynnfor', 'action', 'taken', 'within',
'threeway', 'framework', 'constitut', 'unit', 'state', 'coordin', 'branch', 'govern
, 'continu', 'freeli', 'function', 'bill', 'right', 'remain', 'inviol', 'freedom',
'elect', 'wholli', 'maintain', 'prophet', 'downfal', 'american', 'democraci', 'seen
, 'dire', 'predict', 'come', 'naughtnndemocraci', 'dyingnnw', 'know', 'seen', 'rev
iveand', 'grownnw', 'know', 'cannot', 'die', 'built', 'unhamp', 'initi', 'individu
, 'men', 'women', 'join', 'togeth', 'common', 'enterpris', 'enterpris', 'undertaken
, 'carri', 'free', 'express', 'free', 'majoritynnw', 'know', 'democraci', 'alon',
'form', 'govern', 'enlist', 'full', 'forc', 'men', 'enlighten', 'willnnw', 'know',
'democraci', 'alon', 'construct', 'unlimit', 'civil', 'capabl', 'infiniti', 'progres
s', 'improv', 'human', 'lifennw', 'know', 'look', 'surfac', 'sens', 'still', 'sprea
d', 'everi', 'contin', 'human', 'advanc', 'end', 'unconquer', 'form', 'human', 'soc
ietynna', 'nation', 'like', 'person', 'bodya', 'bodi', 'must', 'fed', 'cloth', 'hou
s', 'invigor', 'rest', 'manner', 'measur', 'object', 'timenna', 'nation', 'like', 'p
erson', 'mind', 'mind', 'must', 'kept', 'inform', 'alert', 'must', 'know', 'unders
tand', 'hope', 'need', 'neighbor', 'nation', 'live', 'within', 'narrow', 'circl', 'w
orldnnand', 'nation', 'like', 'person', 'someth', 'deeper', 'someth', 'perman', 's
ometh', 'larger', 'sum', 'part', 'someth', 'matter', 'futur', 'call', 'forth', 'sac
r', 'guard', 'presentnnit', 'thing', 'find', 'difficult', 'even', 'imposs', 'hit',
'upon', 'singl', 'simpl', 'wordnnand', 'yet', 'understand', 'spirit', 'faith', 'ame
rica', 'product', 'centuri', 'born', 'multitud', 'came', 'mani', 'land', 'high', 'd
egre', 'mostli', 'plain', 'peopl', 'sought', 'earli', 'late', 'find', 'freedom', 'f
reelynnth', 'democrat', 'aspir', 'mere', 'recent', 'phase', 'human', 'histori', 'hu
man', 'histori', 'permeat', 'ancient', 'life', 'earli', 'peopl', 'blaze', 'anew', 'm
iddl', 'age', 'written', 'magna', 'chartannin', 'america', 'impact', 'irresist', 'a
merica', 'new', 'world', 'tongu', 'peopl', 'contin', 'newfound', 'land', 'came', 'b
eliev', 'could', 'creat', 'upon', 'contin', 'new', 'life', 'life', 'new', 'freedom
nnit', 'vital', 'written', 'mayflow', 'compact', 'declar', 'independ', 'constitut',
'unit', 'state', 'gettysburg', 'addressnnthos', 'first', 'came', 'carri', 'long', 's
pirit', 'million', 'follow', 'stock', 'sprang', 'move', 'forward', 'constantli', 'c
onsist', 'toward', 'ideal', 'gain', 'statur', 'clariti', 'generationnnth', 'hope',
'republ', 'cannot', 'forev', 'toler', 'either', 'undeserv', 'poverti', 'selfserv',
'wealthnnw', 'know', 'still', 'far', 'go', 'must', 'greatli', 'build', 'secur', 'op
portun', 'knowledg', 'everi', 'citizen', 'measur', 'justifi', 'resourc', 'capac', 'l
andnnbut', 'enough', 'achiev', 'purpos', 'alon', 'enough', 'cloth', 'feed', 'bodi
, 'nation', 'instruct', 'inform', 'mind', 'also', 'spirit', 'three', 'greatest', 's

piritnnwithout', 'bodi', 'mind', 'men', 'know', 'nation', 'could', 'livennbut', 'sp
 irit', 'america', 'kill', 'even', 'though', 'nation', 'bodi', 'mind', 'constrict',
 'alien', 'world', 'live', 'america', 'know', 'would', 'perishednnthat', 'spirit', '
 faith', 'speak', 'us', 'daili', 'live', 'way', 'often', 'unnot', 'seem', 'obviou',
 'speak', 'us', 'capit', 'nation', 'speak', 'us', 'process', 'govern', 'sovereignti'
 , '48', 'state', 'speak', 'us', 'counti', 'citi', 'town', 'villag', 'speak', 'us',
 'nation', 'hemispher', 'across', 'sea', 'enslav', 'well', 'free', 'sometim', 'fail'
 , 'hear', 'heed', 'voic', 'freedom', 'us', 'privileg', 'freedom', 'old', 'old', 'st
 orynnth', 'destini', 'america', 'proclaim', 'word', 'propheci', 'spoken', 'first',
 'presid', 'first', 'inaugur', '1789', 'word', 'almost', 'direct', 'would', 'seem',
 'year', '1941', 'preserv', 'sacr', 'fire', 'liberti', 'destini', 'republican', 'mod
 el', 'govern', 'justli', 'consid', 'deepli', 'final', 'stake', 'experi', 'intrust',
 'hand', 'american', 'peoplennif', 'lose', 'sacr', 'fireif', 'let', 'smother', 'doub
 t', 'fear', 'shall', 'reject', 'destini', 'washington', 'strove', 'valiantli', 'tri
 umphantli', 'establish', 'preserv', 'spirit', 'faith', 'nation', 'furnish', 'highes
 t', 'justif', 'everi', 'sacrific', 'may', 'make', 'caus', 'nation', 'defensennin',
 'face', 'great', 'peril', 'never', 'encount', 'strong', 'purpos', 'protect', 'perpe
 tu', 'integr', 'democracynnfor', 'muster', 'spirit', 'america', 'faith', 'americann
 w', 'retreat', 'content', 'stand', 'still', 'american', 'go', 'forward', 'servic',
 'countri', 'godn']

2.2.3 Find the 3 most common words used in all three speeches

The most common words used in Roosevelt speech are [('nation', 16), ('kn
 ow', 10), ('peopl', 8)]

The most common words used in Kennedy speech are [('let', 11), ('us', 11
), ('power', 9)]

The most common words used in Nixon speech are [('us', 26), ('america',
 19), ('respons', 16)]

2.3.1 Show the most common words used in all three speeches in the form of word clouds



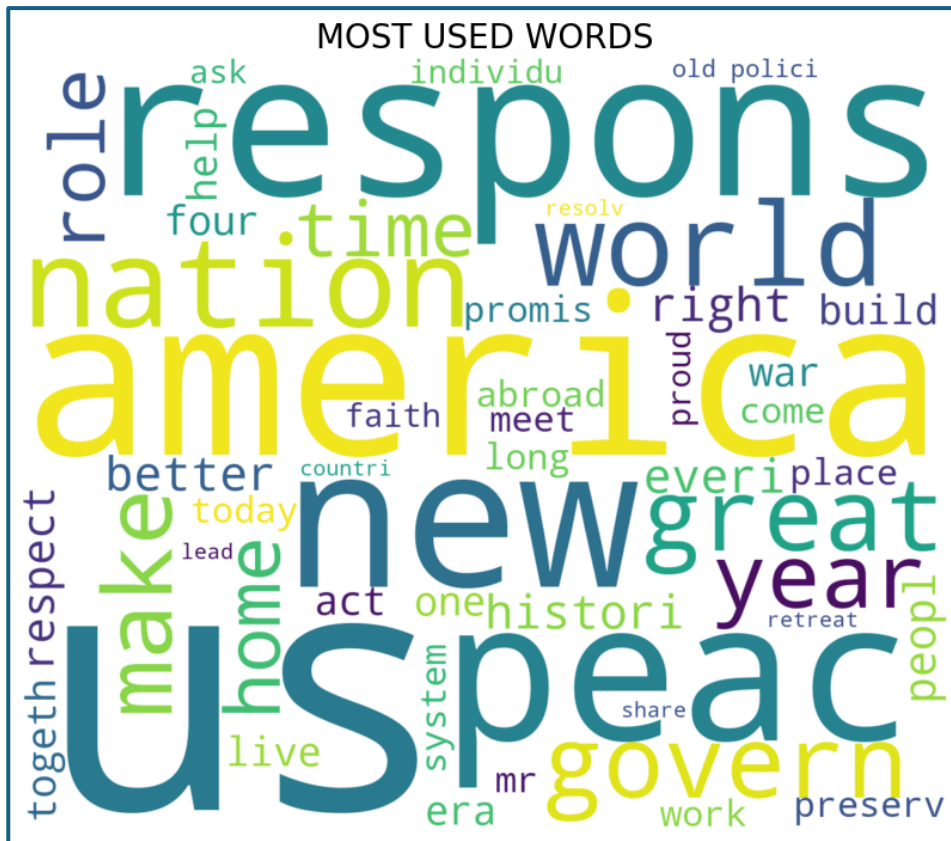


Figure 2.1: Most used words in Nixon Speech