greatlearning
Learning for Life

ML-2 PROJECT REPORT

DSBA

Rohit Nagarahalli

Table of Content

List of Tables

List of Figures

Problem 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.

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.

Shape, Data-types and Statistical summary

The raw data consists of 1525 records and 10 variables which includes an unnamed column of unique records probably an Id variable which shall be dropped for further analysis. The data consists of two Object data type and 8 int64 data types.

Of the remaining 9 columns 8 are identified to be categorical and 1 being continuous.

count	1525.000000
mean	54.182295
std	15.711209
min	24.000000
25%	41.000000
50%	53.000000
75%	67.000000
max	93.000000

Table 1 Statistical Summary Age

Univariate Analysis

The below pie chart indicates the percentage of voters preferring their party of choice. Voters highest preferred is Labour accounting to approximately 70% of the votes and the least being Conservative accounting up to 30% of the total votes. This also being the target variable we can say the target is not highly oversampled yet there is a domination of Labour over Conservative

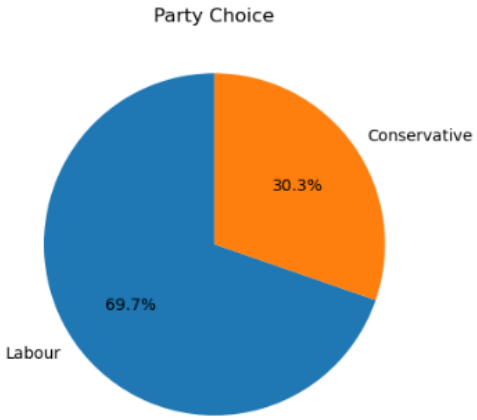

Figure 1 Party Choice

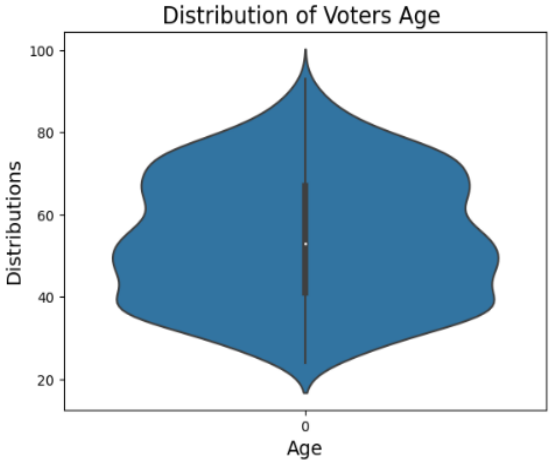

Figure 2 Distribution of Age

In the figure 2, the distribution of the variable Age says that the data is normally distributed except that 3 small peaks can be seen at the 40, 50 and 70. The skewness is absent in the data and there are no outliers present.

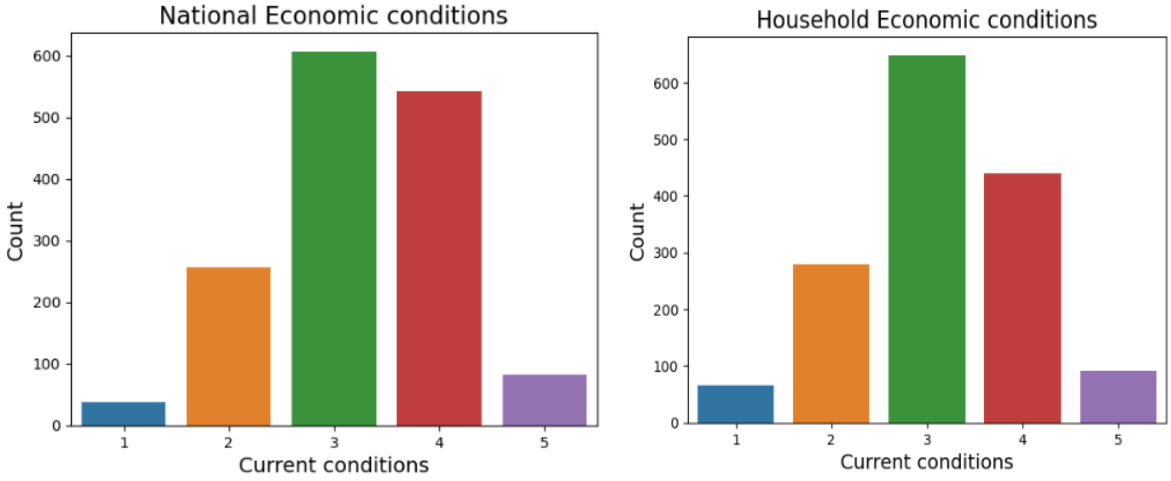

Figure 3 National and Household Economic conditions

Figure 3 represents the National and Economic Conditions. The current economic conditions of the national and the household appears to be almost similar both indicated Condition 3 is the highest amount the voters and 1 being the least. This analysis could also imply that both the variables are displaying same details and dropping one could bring no harm to the model. However, a further analysis needs to be performed to evaluate the same.

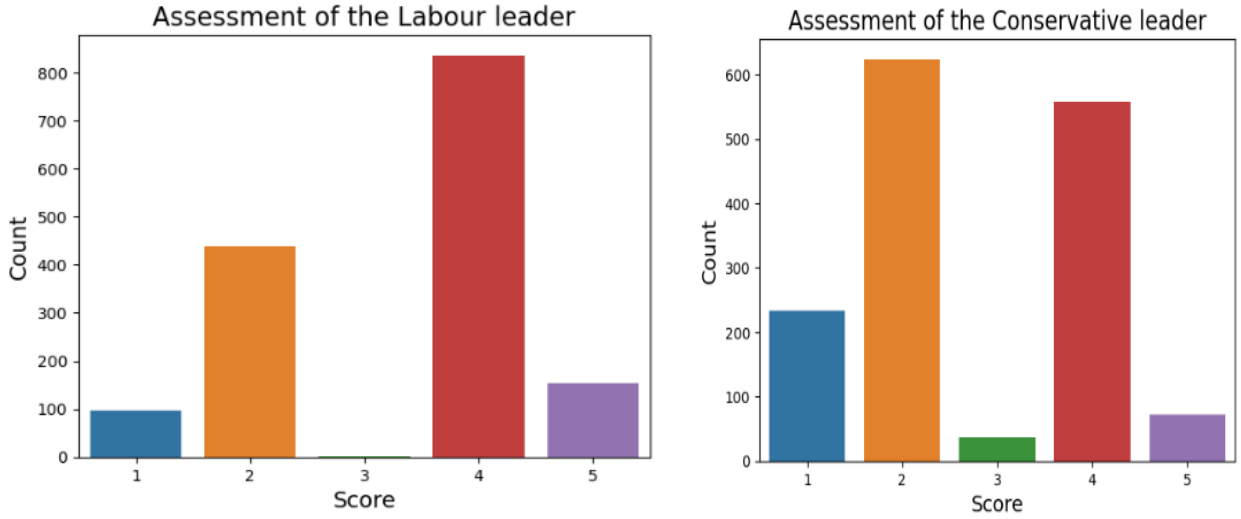

Figure 4 Assessment of Labour and Conservative leader

The assessment of the Labour and the Conservative Leader comes out quite interesting since the Labour leader has the highest assessment at point 4 which is the second best after 5 and the conservative leader has the highest assessment at point 2 which is the second worst after 1. The conservative leader also has a score at 4 in the most number of occasions after score at 2.

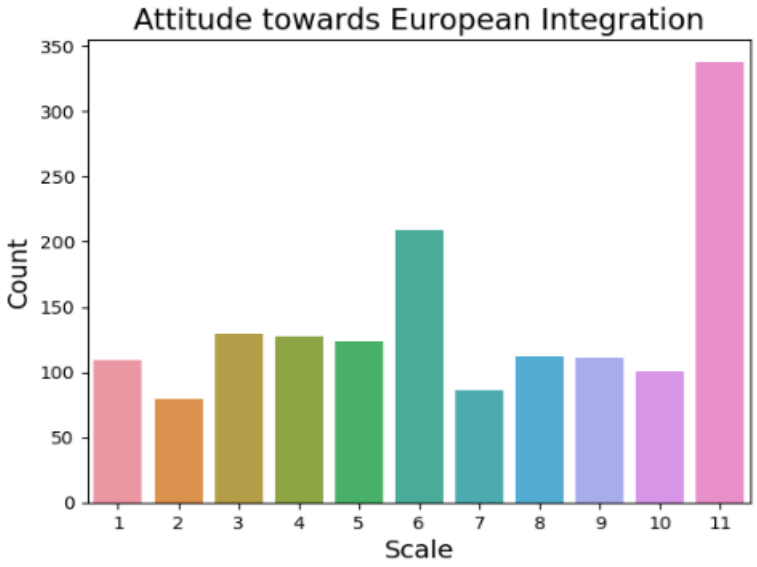

Figure 5 European Integration

Highest number of voters in a scale of 1-11 have shown Eurosceptic sentiment as the scale of 11 has the maximum count followed by 6. The plot also says majority of the voters have shown Eurosceptic sentiment since the scale above 5 appears to have more counts compared to the scale 1-5.

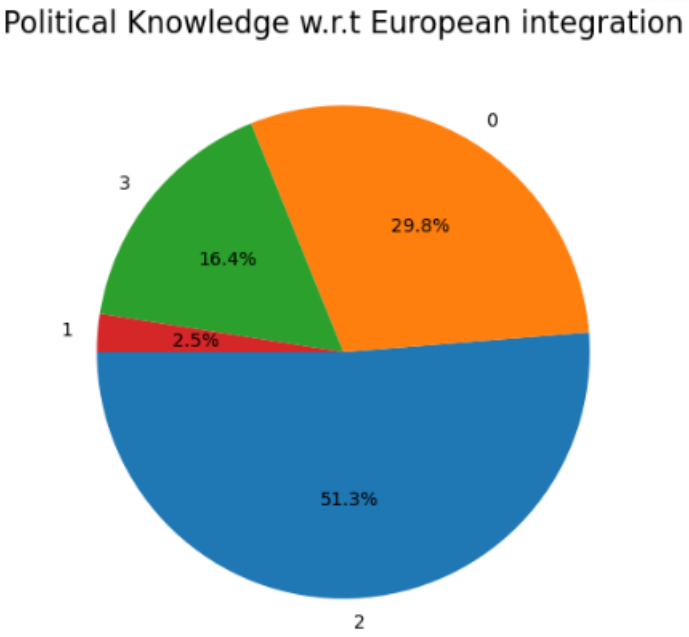

Figure 6 Political knowledge w.r.t European Integration

51.3 % percentage of voters had a knowledge of party's positions on European integration. Whereas, approximately 30% of the voters did not have a knowledge of party's positions on European integration. Approximately 16% of the voters had a full knowledge of European integration

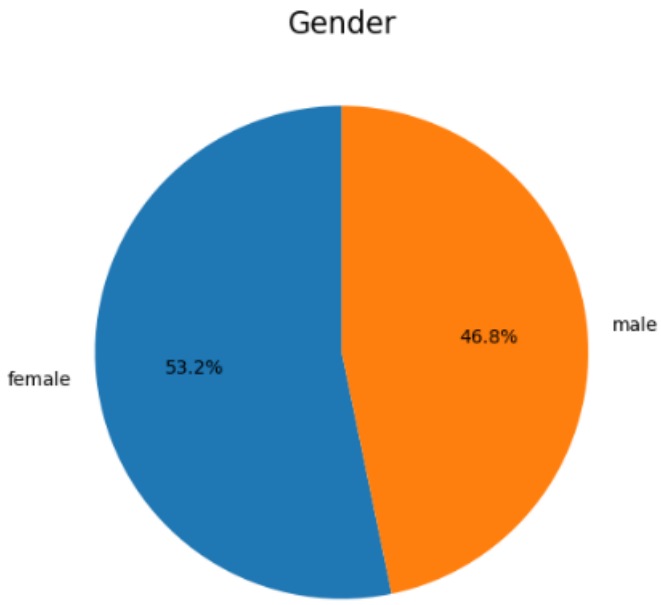

Figure 7 Gender

The sampling of the gender variable is almost same and imbalance between the class is not observed. However, majority of the voters are female owing to 53.2% (812) of the total number of voters in the data provided

Multivariate Analysis

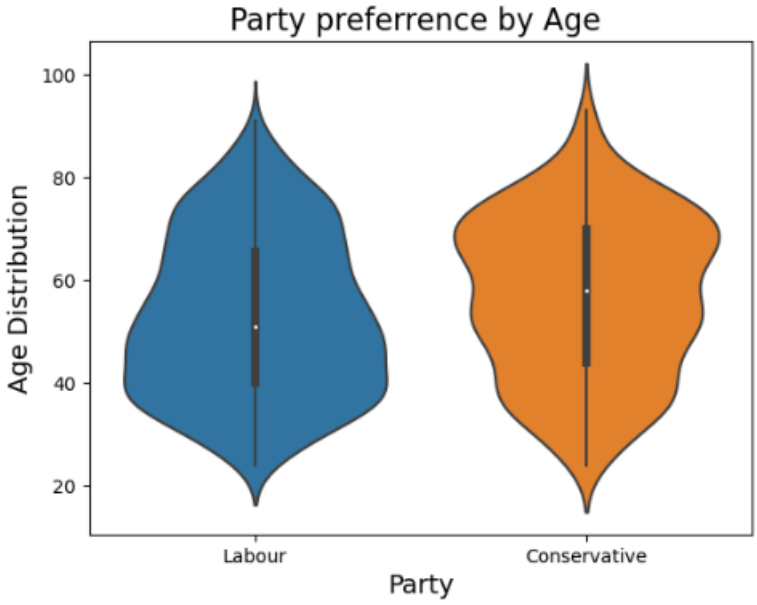

Figure 8 Party Preference by Age

The median of the Age for the Labour Party voters closes in at approximately 50 whereas the Age for the conservative party voters closes in at 60. The Labour party is slightly skewed towards its right indicating most of its voters are distributed in the lower range (age). The Conservative party is slightly skewed towards its left indicating most of its voters are piled in the upper range (age).

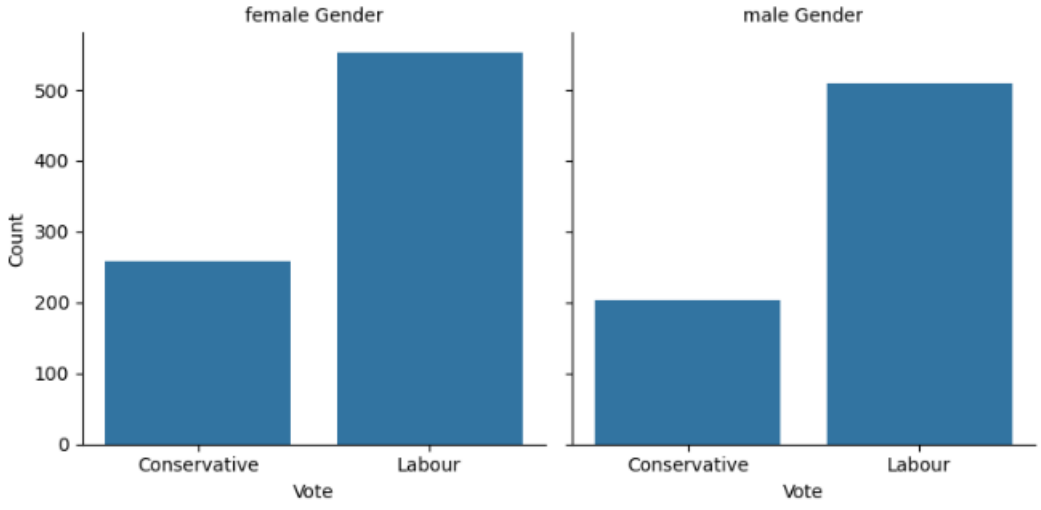

Figure 9 Part Preference by gender

From the Figure 9, the gender female tends to prefer Labour party over Conservative and same goes for male. Although nothing much can be inferred out of it except that the count of Labour voters is high compared to Conservative which might indicate mildly oversampled Labour class.

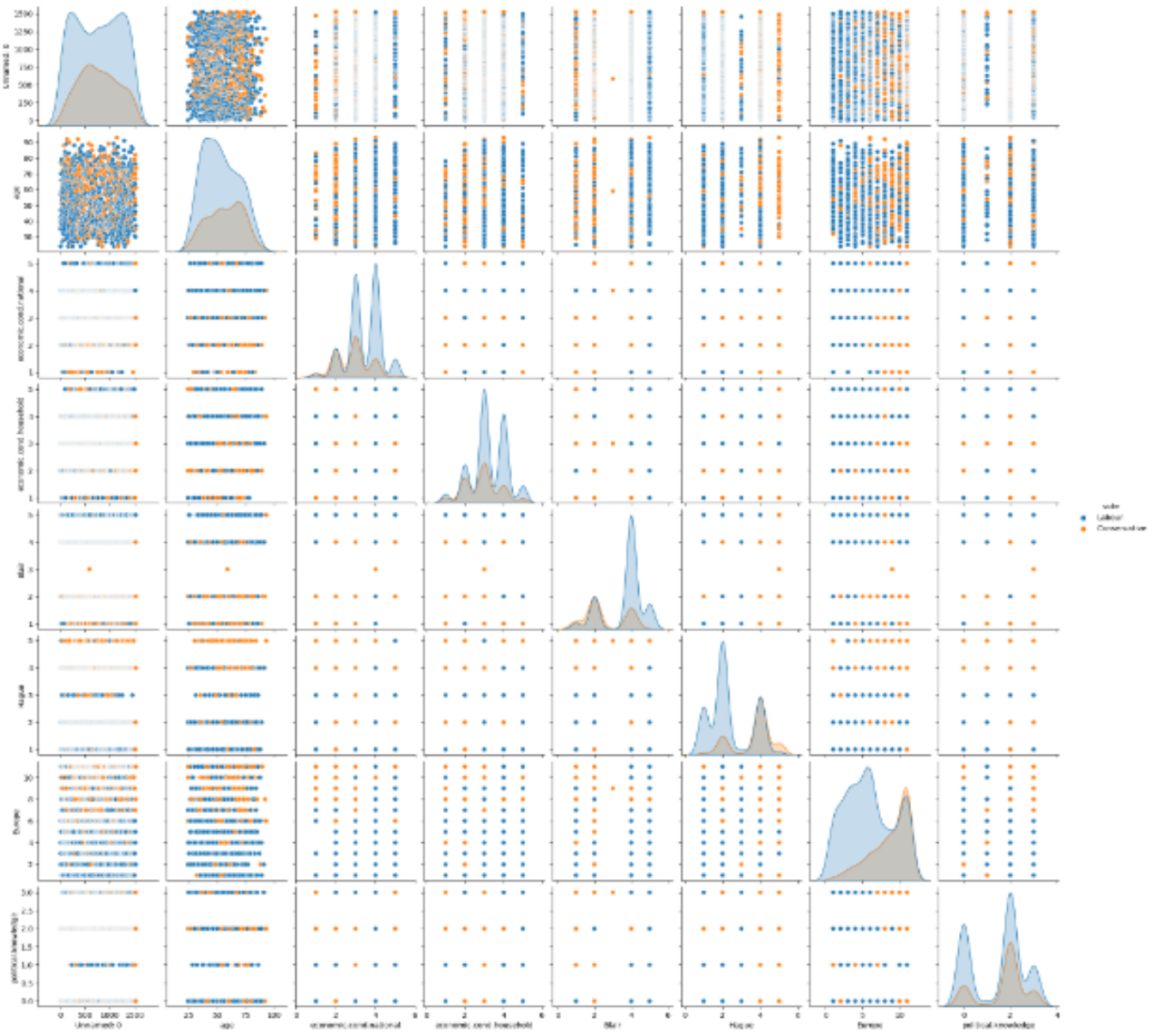

Figure 10 Pair plot

The distribution of the Labour and conservative class across other independent variables can be visualized through the Pair plot in the figure 10. It can be seen that majority of the variables such as political knowledge and Economic conditions appears to have similar patter overlapping by looking at the KDE plot. The peaks for such overlapping variables are low for Labour class considering its imbalance.

The overlapping of vote variable class among other independent variables could imply harm to model building but the importance is unknown hence ruling out the option of dropping variables.

We will build the model with all possible variables given by the team except the unnamed identity column which is of no use. The importance of features post model building will have a direct relation if we come back and look at the pair plot.

Key Meaningful Observations

- As per the data provided voters highest preferred is Labour party accounting to approximately 70% of the votes and the least being Conservative accounting up to 30% of the total votes. Though this might not indicate under sampled data, a better result could be provided if a greater number of votes were taken into consideration.

- The Age variable appeared to have distributed almost normally with 3 minor peaks observed and did not possess any outliers and skewness.

- The economic conditions of the national and household indicated almost the same result. Condition 3 being the highest among the voters and 1 being the least. The plots also indicated having one of the either instead of two would do no harm to the data since both indicated almost same most of the cases. However, a domain approach could help solve if or not the two variables resembled same.

- The assessment of the Labour and the Conservative Leader comes out quite interesting since the Labour leader has the highest assessment at point 4 which is the second best after 5 and the conservative leader has the highest assessment at point 2 which is the second worst after 1. The conservative leader also has a score at 4 in the most number of occasions after score at 2.

- The attitude towards the European Integration indicates highest number of voters in a scale of 1-11 have shown Eurosceptic sentiment as the scale of 11 has the maximum count followed by 6. The plot also says majority of the voters have shown Eurosceptic sentiment since the scale above 5 appears to have more counts compared to the scale 1-5.

- 51.3 % percentage of voters had a knowledge of party's positions on European integration. Whereas, approximately 30% of the voters did not have a knowledge of party's positions on European integration. Approximately 16% of the voters had a full knowledge of European integration.

- The median of the Age for the Labour Party voters closes in at approximately 50 whereas the Age for the conservative party voters closes in at 60. The Labour party is slightly skewed towards its right indicating most of its voters are distributed in the lower range (age). The Conservative party is slightly skewed towards its left indicating most of its voters are piled in the upper range (age).

Data Preprocessing

Outlier Detection

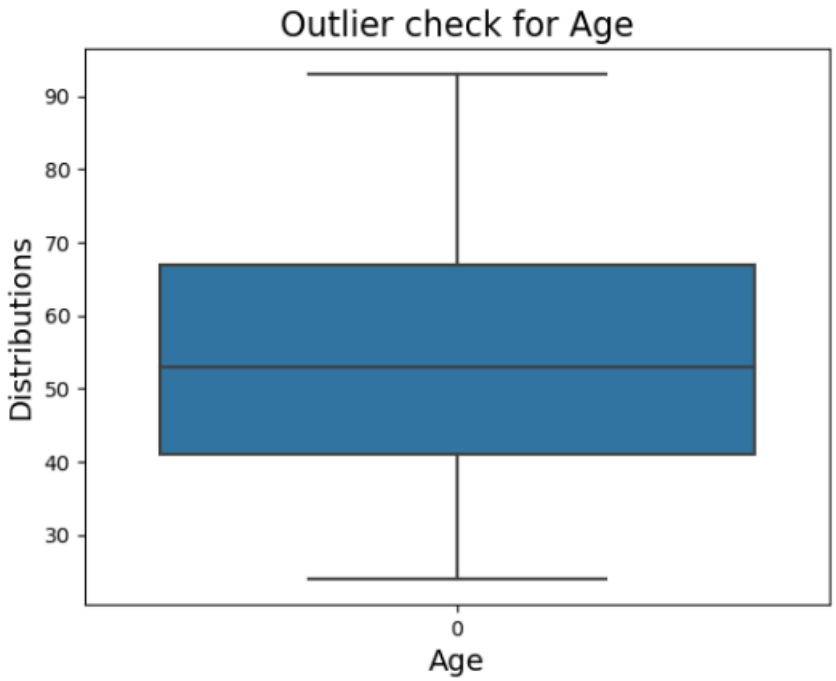

Figure 11 Outlier Check

The Age was the only continuous variable among the group and rest all being categorical. Outliers are absent for the Age variable and doesn't need any further treatment.

Data Encode

Vote and Gender variables needs to be encoded and simple mapping with 0 and 1 for the category is applied since they are binary variable. The other variables being ordinal we will keep the order as it is.

The Labour was set as 0 and the Conservative to 1 since the goal was not to focus on variable of interest instead it was to predict who could win the election by securing majority of seats.

Data Split and Scaling

KNN algorithm relies on distance measures to make predictions. The distance between data points is a key factor in KNN, and features with larger scales can dominate the distance calculations. Hence scaling of the age variable was performed using standard scaler. However, scaling was performed only for dealing with KNN algorithm. Rest other algorithms such as Naïve Bayes, Bagging and Boosting techniques were robust to different scales of data and outliers.

Data split for all model building was split with considering 30% of the data for testing and a random state of 7 across all the models.

This is a kind of problem that focuses on overall predictions by the model since there is no variable of interest but both the outcomes. The problem focuses on predicting which party is more likely to win the elections our main **metrics of choice** would be F1, Precision, Recall and Confusion Matrix.

Model Building, Performance Evaluation and Improvement

KNN

Evaluation Metrics

Training

	precision	recall	f1-score	support
0	0.90	0.91	0.90	729
1	0.80	0.79	0.79	338
accuracy			0.87	1067
macro avg	0.85	0.85	0.85	1067
weighted avg	0.87	0.87	0.87	1067

```
[[661  68]
 [ 72 266]]
```

Table 2 KNN Training

Testing

	precision	recall	f1-score	support
0	0.88	0.87	0.88	334
1	0.66	0.67	0.67	124
accuracy			0.82	458
macro avg	0.77	0.77	0.77	458
weighted avg	0.82	0.82	0.82	458

```
[[292  42]
 [ 41  83]]
```

Table 3 Testing KNN

Observations:

Train Score: 0.8687910028116214

Test Score: 0.8187772925764192

f1 Score (Labour): 0.87

f1 Score (Conservative): 0.67

The **confusion matrix** is present in the above table 2 and 3 for training and testing respectively. The support for predicting Conservative is less than half of support for Labour further influencing the f1 score for the same. The overall Training and Testing scores are not overfitting largely.

A cross-fold **CV** was also performed to check if the model could improve further and the observations were noted down. However, the accuracy did not improve largely and the model was classified stable and predicting better with an accuracy of 82%.

```
AUC for the Training Data: 0.939
AUC for the Test Data: 0.831
```

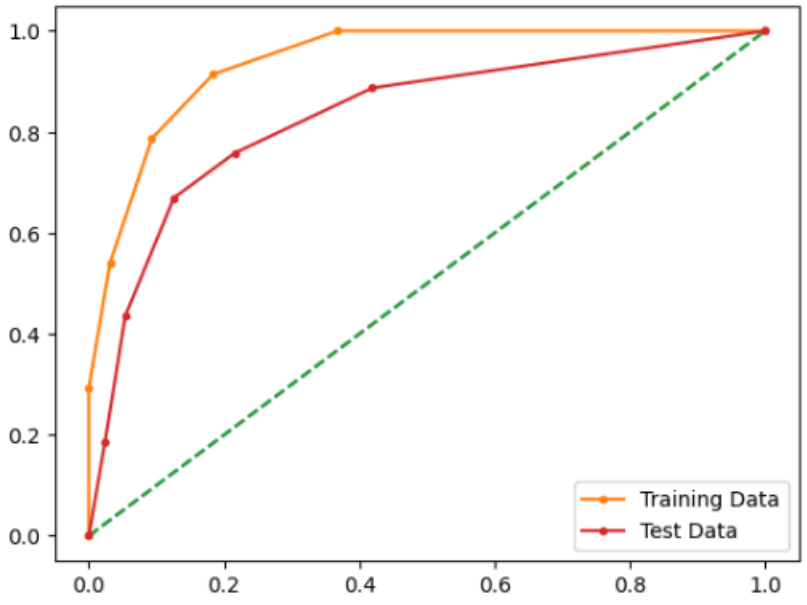

Figure 12 ROC-AUC for KNN model

The AUC score for both the training and testing data comes out to be 93% and 83% respectively. Both of them implying well above the diagonal line and not too far from the edge (1) and also a slight overfitting could be observed as there is at least 10% variation between the training and the testing score. Any further decision shall be taken post domain interference.

Naïve Bayes

Gaussian Naïve bayes was the algorithm that was used to predict since the data had a mixture of Continuous as well as Categorical data types. The model was built and observations were noted down.

Training

	precision	recall	f1-score	support
0	0.88	0.89	0.88	729
1	0.75	0.74	0.75	338
accuracy			0.84	1067
macro avg	0.82	0.81	0.81	1067
weighted avg	0.84	0.84	0.84	1067

```
[[648  81]
 [ 89 249]]
```

Table 4 Naive Bayes Training

Testing

	precision	recall	f1-score	support
0	0.88	0.87	0.88	334
1	0.66	0.67	0.67	124
accuracy			0.82	458
macro avg	0.77	0.77	0.77	458
weighted avg	0.82	0.82	0.82	458

```
[[292  42]
 [ 41  83]]
```

Table 5 Naive Bayes Testing

Observations:

Train Score: 0.8406747891283973

Test Score: 0.8187772925764192

f1 Score (Labour): 0.88

f1 Score (Conservative): 0.67

The Naïve bayes model appears to perform better than the KNN considering the gap between the Training and Testing is bridged and reduced. Although the overall model accuracy of the model remains same but the prediction appears to be smooth.
The confusion matrix for the training and testing is present in the table 4 and 5 for reference.

```
AUC for the Training Data: 0.894
AUC for the Test Data: 0.866
```

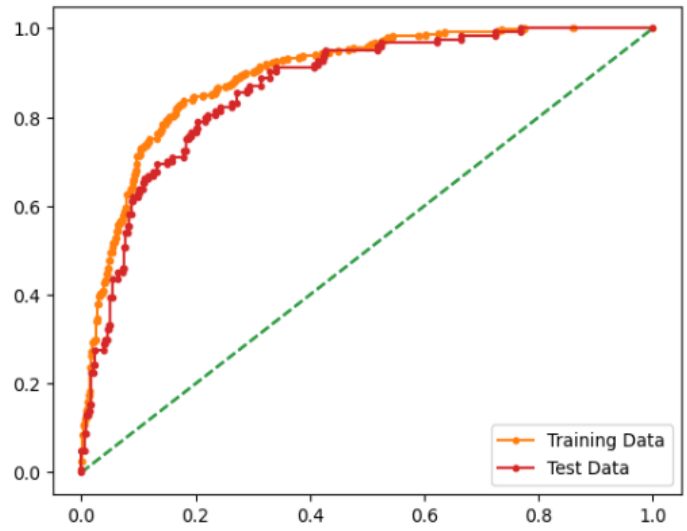

Figure 13ROC-AUC Naive Bayes

The AUC-ROC appears to be better than the KNN as the distance between the training and testing looks marginally small and farther away from the diagonal.

Before Proceeding to the Ensemble Techniques model was built using Decision trees and pruned Decision trees. This also gives an enhanced reasons to not just restrict to decision trees but to also check with Ensemble techniques

Decision Trees

Initially a normal Decision tree was built without tuning it and the observations were noted down. Usually, Decision trees are prone to overfit which was the case here as well. The raining score came out to be 100% but the testing score was 76%. This was evidently overfitting.

Now, the decision tree was pruned with criterion being Gini and a max depth of 4. Post doing this the model performed better with a better Training and Testing Scores.

Training

	precision	recall	f1-score	support
0	0.88	0.89	0.89	729
1	0.76	0.73	0.75	338
accuracy			0.84	1067
macro avg	0.82	0.81	0.82	1067
weighted avg	0.84	0.84	0.84	1067

```
[[651  78]
 [ 91 247]]
```

Table 6 Decision Tree Training

Testing

	precision	recall	f1-score	support
0	0.88	0.88	0.88	334
1	0.67	0.69	0.68	124
accuracy			0.83	458
macro avg	0.78	0.78	0.78	458
weighted avg	0.83	0.83	0.83	458

```
[[293  41]
 [ 39  85]]
```

Table 7 Decision Tree Testing

The Training and Testing Scores came out to be 84% and 83% indicating elimination of Overfitting. The Precision and Recall performed better in predicting the class 0 (Labour) compared to the other class.

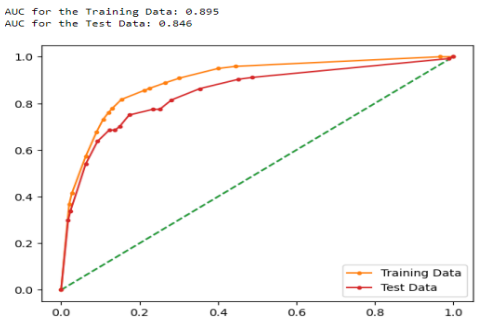

Figure 14 AUC ROC Decision Trees Pruned

The AUC was quite good and well above the diagonal. The Overfitting was also notably absent.

Ensemble Methods

Bagging

With a base estimator as decision tree and no other hyper parameters tuning a base model was built and following observations were noted

Training

	precision	recall	f1-score	support
0	0.97	1.00	0.99	729
1	1.00	0.94	0.97	338
accuracy			0.98	1067
macro avg	0.99	0.97	0.98	1067
weighted avg	0.98	0.98	0.98	1067

```
[[728   1]
 [ 19 319]]
```

Table 8 Bagging training

Testing

	precision	recall	f1-score	support
0	0.85	0.90	0.87	334
1	0.67	0.57	0.62	124
accuracy			0.81	458
macro avg	0.76	0.73	0.74	458
weighted avg	0.80	0.81	0.80	458

```
[[299  35]
 [ 53  71]]
```

Table 9 Bagging testing

From the above tables it can be observed that the model is overfitting. But the model has done slightly better than the decision tree (without pruning) but the concept of overfitting has not been reduced yet. The concept of Random Forest in bagging could be an optimal solution.

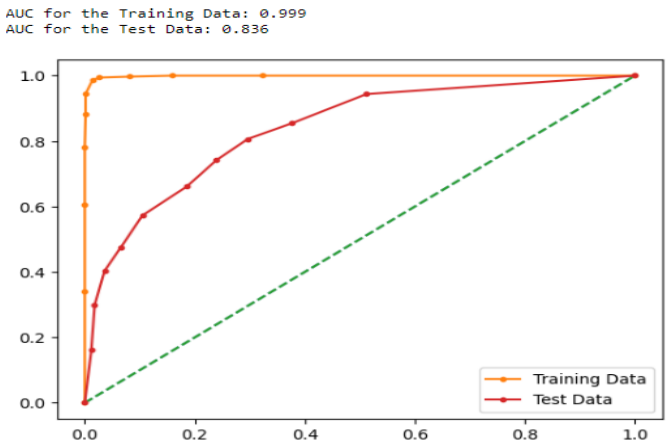

Figure 15 ROC-AUC Bagging

Random Forest

A hyper tuned parameters were used for Random Forest classifier. Max depth 10, min sample leaf 2, min sample split 2 and n estimators 100 was used. The training and testing scores were observed

Training

	precision	recall	f1-score	support
0	0.94	0.95	0.95	729
1	0.90	0.88	0.89	338
accuracy			0.93	1067
macro avg	0.92	0.92	0.92	1067
weighted avg	0.93	0.93	0.93	1067

```
[[696  33]
 [ 41 297]]
```

Table 10 Training Random Forest

Testing

	precision	recall	f1-score	support
0	0.87	0.90	0.88	334
1	0.70	0.65	0.67	124
accuracy			0.83	458
macro avg	0.79	0.77	0.78	458
weighted avg	0.83	0.83	0.83	458

```
[[300  34]
 [ 44  80]]
```

Table 11 Testing Random Forest

We can observe an improvement in the model than the basic bagging model. Although the model still appears to overfit other techniques such as boosting could further improve the performance.

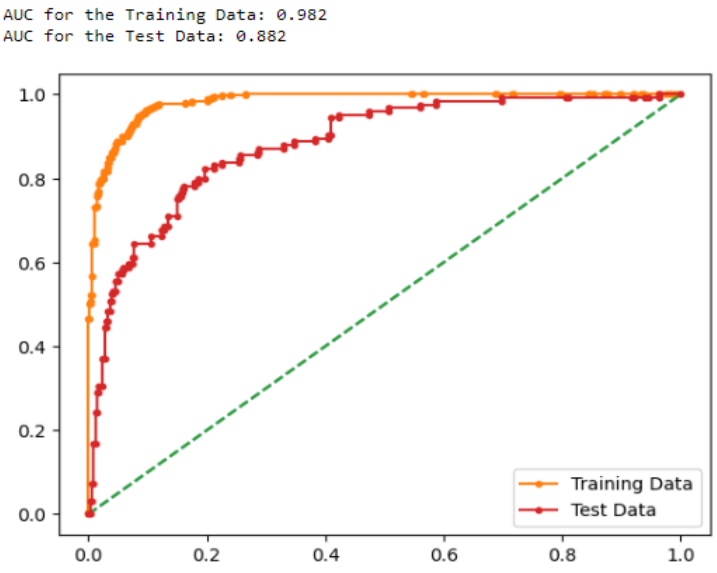

Figure 16 ROC-AUC Random Forest

Boosting

AdaBoost

Training

	precision	recall	f1-score	support
0	0.88	0.90	0.89	729
1	0.78	0.73	0.75	338
accuracy			0.85	1067
macro avg	0.83	0.82	0.82	1067
weighted avg	0.85	0.85	0.85	1067

```
[[657  72]
 [ 90 248]]
```

Table 12 Ada-Boost Training

Testing

	precision	recall	f1-score	support
0	0.87	0.90	0.89	334
1	0.71	0.65	0.68	124
accuracy			0.83	458
macro avg	0.79	0.77	0.78	458
weighted avg	0.83	0.83	0.83	458

```
[[302  32]
 [ 44  80]]
```

Table 13 Ada-Boost Testing

Precision, Recall and f1-score all align with the training as well as testing. Overfitting is not observed so does the underfitting. Hence, a perfect model with better accuracy.

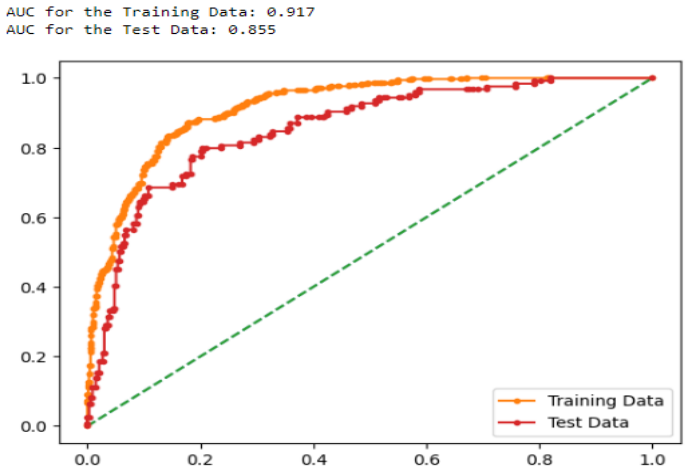

Figure 17 ROC-AUC Ada-Boost

The AUC score seems fairly good and curve well above the diagonal and close to the edge. Ada-Boost could prove to be a good classification model for the data provided. However, we would try with Gradient and XGBoost model to check for further betterment of the model and scores.

Gradient Boosting

Training

	precision	recall	f1-score	support
0	0.91	0.93	0.92	729
1	0.84	0.80	0.82	338
accuracy			0.89	1067
macro avg	0.88	0.87	0.87	1067
weighted avg	0.89	0.89	0.89	1067

```
[[677  52]
 [ 66 272]]
```

Table 14 Gradient Boosting Training

Testing

	precision	recall	f1-score	support
0	0.87	0.90	0.89	334
1	0.70	0.65	0.68	124
accuracy			0.83	458
macro avg	0.79	0.78	0.78	458
weighted avg	0.83	0.83	0.83	458

```
[[300  34]
 [ 43  81]]
```

Table 15 Gradient Boosting Testing

The training and testing appear to not overfit on higher ranges yet a gap of 6% between f1-score can be observed and 15% gap in the Recall of predicting Conservative between training and testing.

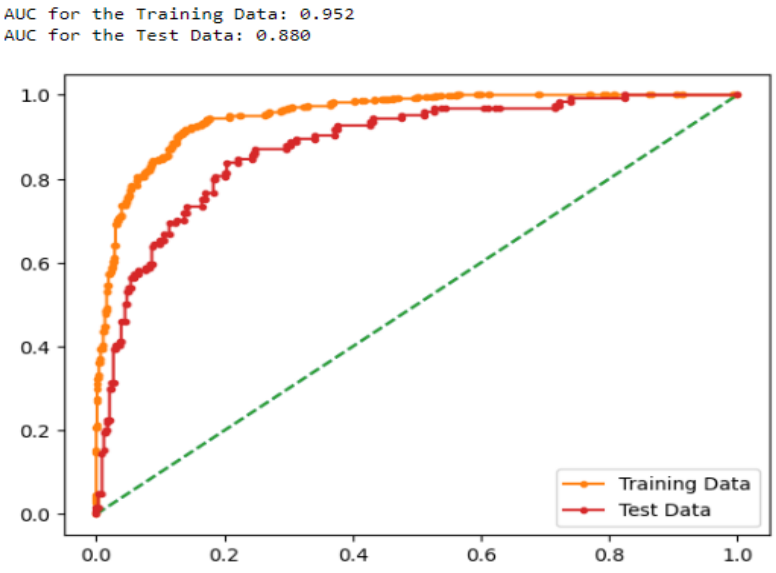

Figure 18 ROC-AUC Gradient Boosting

The ROC-AUC for the training and testing is minimal and well above the diagonal.

Gradient Boosting Hyper tuned

Training

	precision	recall	f1-score	support
0	0.92	0.93	0.92	729
1	0.85	0.81	0.83	338
accuracy			0.90	1067
macro avg	0.88	0.87	0.88	1067
weighted avg	0.89	0.90	0.89	1067

```
[[680  49]
 [ 63 275]]
```

Table 16 Tuned GB Training

Testing

	precision	recall	f1-score	support
0	0.88	0.91	0.89	334
1	0.72	0.65	0.69	124
accuracy			0.84	458
macro avg	0.80	0.78	0.79	458
weighted avg	0.83	0.84	0.84	458

```
[[303  31]
 [ 43  81]]
```

Table 17 Tuned GB Testing

Post tuning the hyper parameters the model accuracy has improved slightly. A rigorous tuning in the future could prove model accuracy improvement.

```
AUC for the Training Data: 0.953
AUC for the Test Data: 0.881
```

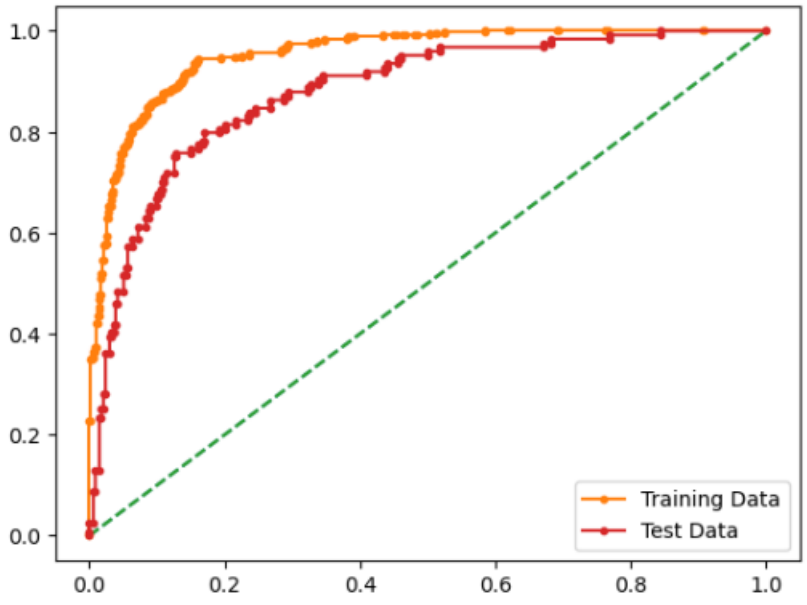

Figure 19 ROC-AUC Score Hyper tuned GB

The ROC-AUC too has improved with curve well above the diagonal and close to 1 edge.

Xtreme Gradient Boosting (Base Model)

Training

	precision	recall	f1-score	support
0	1.00	1.00	1.00	729
1	0.99	0.99	0.99	338
accuracy			1.00	1067
macro avg	1.00	1.00	1.00	1067
weighted avg	1.00	1.00	1.00	1067

```
[[727   2]
 [  2 336]]
```

Table 18 XGB Training

Testing

	precision	recall	f1-score	support
0	0.87	0.86	0.87	334
1	0.64	0.65	0.65	124
accuracy			0.81	458
macro avg	0.75	0.76	0.76	458
weighted avg	0.81	0.81	0.81	458

```
[[288  46]
 [ 43  81]]
```

Table 19 XGB Testing

From the tables above the model is clearly overfitting as there is a huge gap between the training and testing. However, a hyper tuned parameters could possibly reduce overfitting.

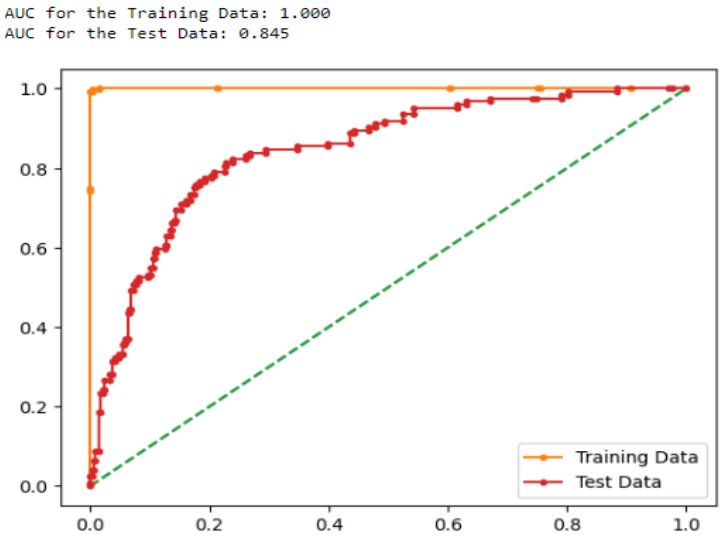

Figure 20 ROC-AUC XGB base model

AUC score alongside the curve indicates the model is clearly overfitting. A hyper tuned model could bridge the gap further reducing overfitting.

Note: For all the models confusion matrix is present in the bottom of the table for train and test.

XGBoost Hyper tuned parameters

Training

	precision	recall	f1-score	support
0	0.89	0.92	0.90	729
1	0.82	0.75	0.78	338
accuracy			0.87	1067
macro avg	0.85	0.83	0.84	1067
weighted avg	0.86	0.87	0.87	1067

```
[[673  56]
 [ 86 252]]
```

Table 20 XGB tuned Training

Testing

	precision	recall	f1-score	support
0	0.88	0.90	0.89	334
1	0.72	0.66	0.69	124
accuracy			0.84	458
macro avg	0.80	0.78	0.79	458
weighted avg	0.83	0.84	0.84	458

```
[[302  32]
 [ 42  82]]
```

Table 21 XGB tuned Testing

The hyper tuned parameters model indicates the model has performed better than the base XGBoost. The overfitting now is eliminated and the accuracy has improved drastically indicating hyper tuning has significant impact on the model performance.

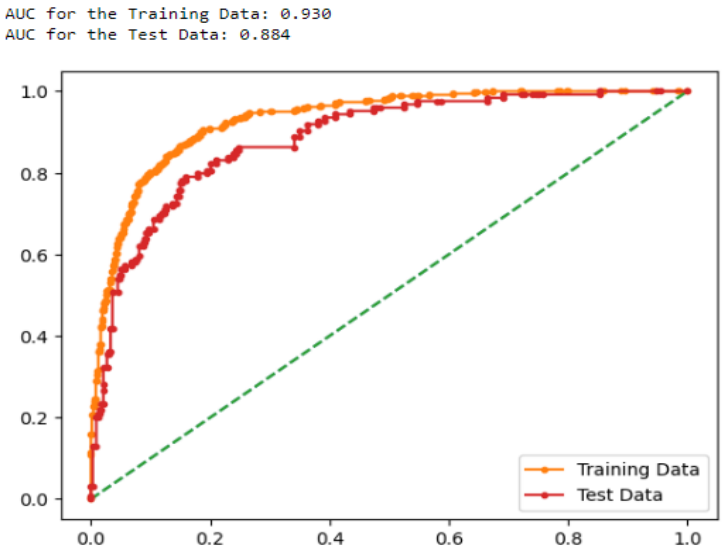

Figure 21 ROC-AUC XGB tuned

Comparing the base model alongside the tuned shows the drastic change. The gap between the training and testing has been bridged and the curve is visibly well above the diagonal.

Final Model Selection

Model	Training	Testing	Precision Train & Test	Recall Train & Test	AUC Train	AUC Test
KNN	87%	82%	0 - 90% & 88% 1- 80% & 77%	0 - 91% & 87% 1- 79% & 67%	94%	83%
Naïve Bayes	84%	82%	0 - 88% & 88% 1- 75% & 66%	0 - 89% & 87% 1- 74% & 67%	89%	87%
Decision tree Pruned	84%	83%	0 - 88% & 88% 1- 76% & 67%	0 - 89% & 88% 1- 73% & 69%	89%	85%
Bagging (Base)	98%	80%	0 - 97% & 85% 1- 100% & 67%	0 - 100% & 90% 1- 94% & 58%	99%	84%
Bagging Tuned(RF)	93%	83%	0 - 94% & 87% 1- 90% & 70%	0 - 95% & 90% 1- 88% & 65%	98%	88%
Ada Boost	85%	83%	0 - 88% & 87% 1- 78% & 71%	0 - 90% & 90% 1- 73% & 65%	91%	85%
GBM	88%	83%	0 - 91% & 87% 1- 84% & 70%	0 - 93% & 90% 1- 80% & 65%	95%	88%
XGB (Base)	100%	80%	0 - 100% & 87% 1- 99% & 64%	0 - 100% & 86% 1- 99% & 65%	100%	84%
XGB Tuned	86%	83%	0 - 89% & 88% 1- 82% & 72%	0 - 92% & 90% 1- 75% & 66%	93%	88%

Table 22 Model comparison

Since it has been asked to check for feature importance only for the final model. We have decided that Xtreme gradient boosting tuned model to be the final model and the important features have been plotted in a bar graph in the Figure 22 below.

The Xtreme Gradient Boosting post tuning was observed to have performed exceptionally well compare to other models.

Note: The 0 and 1 in the Precision and Recall represents the Labour and Conservative party respectively.

Key Takeaways

Of all the models, Xtreme Gradient Boosting has performed better in terms of accuracy, precision, recall and ROC-AUC curve. Although algorithms such as KNN, Naïve bayes, pruned Decision trees and Ada-Boost have performed exceptionally well it was Xtreme Gradient boosting which outperformed rest all other good performing algorithms. However, models such as Decision trees without pruning and bagging without tuning the parameters resulted in Overfitting and to overcome that pruning and hyper tuning parameters were employed and the results were outstanding as seen in the Table 22.

Most important features

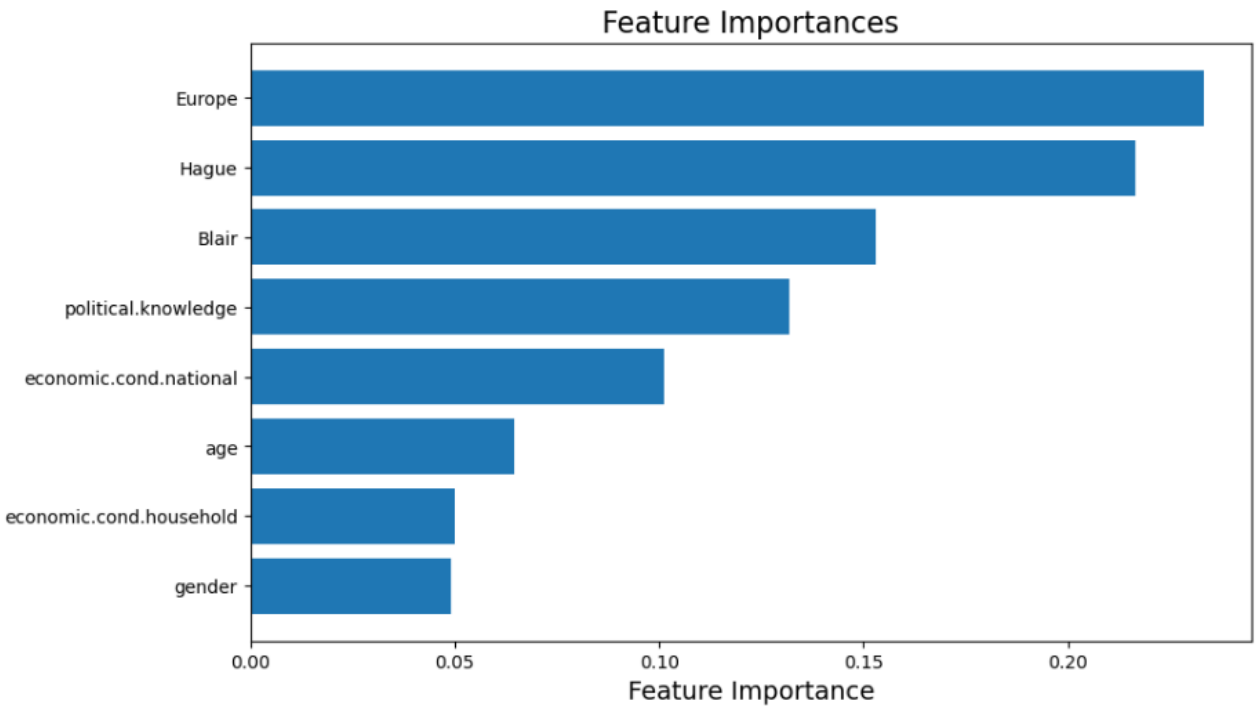

Figure 22 XGB Tuned model Feature Importance

It can be observed that the top important feature is Europe which is the respondent's attitude towards European Integration representing Eurosceptic sentiments. This is followed by the Hague, assessment of the conservative leader. The least important features are gender and economic household conditions.

Any further decisions with respect to dropping unimportant features can be taken post consulting domain and further improve the model prediction and so.

Note: If any model exists that could have been tune but wasn't, the possible reasons could be either the model was not over/under fitting or/and the accuracy was excellent and tuning did no further improvement. Tuning from the Table 22 can be observed that was performed on Bagging model (Random Forest), Boosting and tree was pruned in Decision trees.

Problem-2

Problem Definition

Analyzing the speeches of presidents of the United States of America. A detail corpus available in the NLTK was downloaded and the following president speeches were observed.

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

Number of characters, words and sentences in all three speeches.

Roosevelt Speech:

Number of Characters: 7571

Number of Words: 1526

Number of Sentences: 68

Kennedy Speech:

Number of Characters: 7618

Number of Words: 1543

Number of Sentences: 52

Nixon Speech:

Number of Characters: 9991

Number of Words: 2006

Number of Sentences: 68

Text Cleaning

Stop words were eliminated from the corpus and Stemming method was employed to reduce inflections down to common base root words. Three most common words were extracted from all three speeches.

Three most common words

The 3 most common words in Roosevelt speech are [('nation', 17), ('know', 10), ('peopl', 9)]

The 3 most common words in Kennedy speech are [('let', 16), ('us', 12), ('power', 9)]

The 3 most common words in Nixon speech are [('us', 26), ('let', 22), ('america', 21)]

Three most common words in all three speeches after preprocessing:

us: 46 times

nation: 40 times

let: 39 times

Word cloud for all three speeches and combined

Figure 23 Roosevelt Speech Word cloud

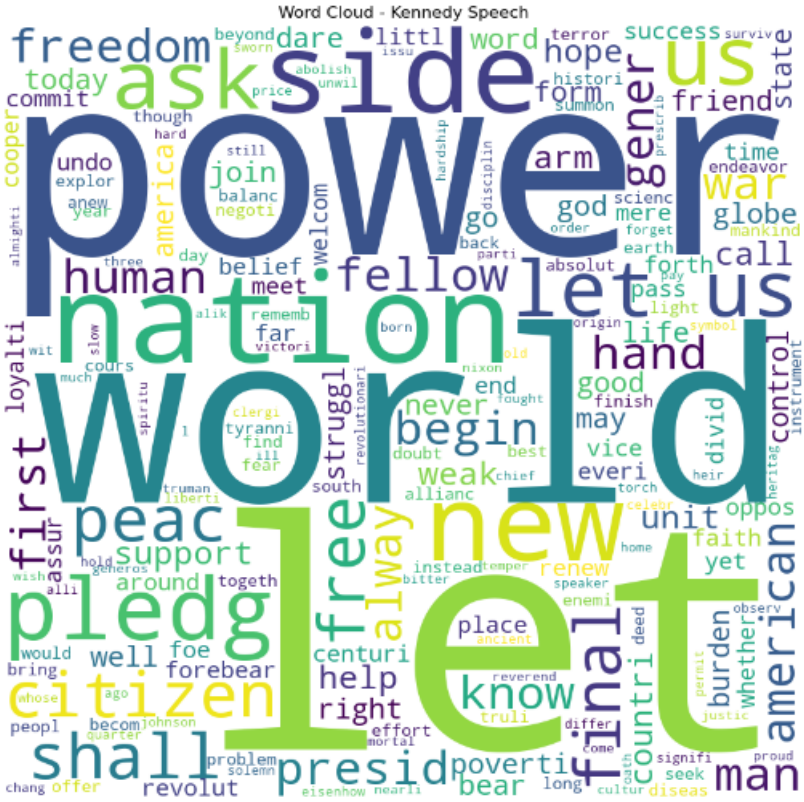

Figure 24 Kennedy Speech Word cloud

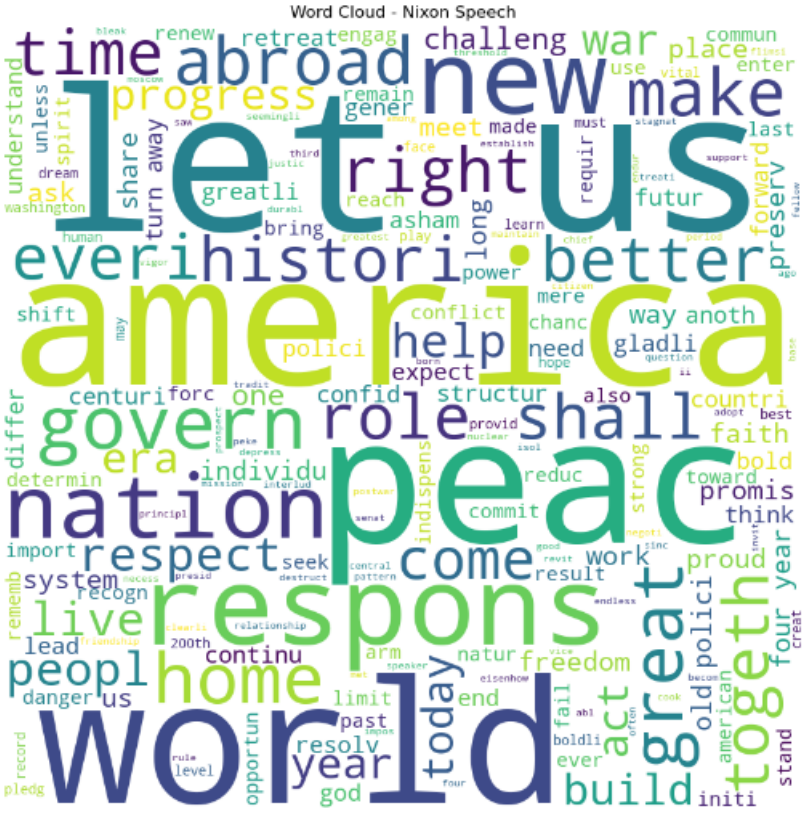

Figure 25 Nixon Speech Word cloud

A word cloud for all the 3 speeches combined for most common words

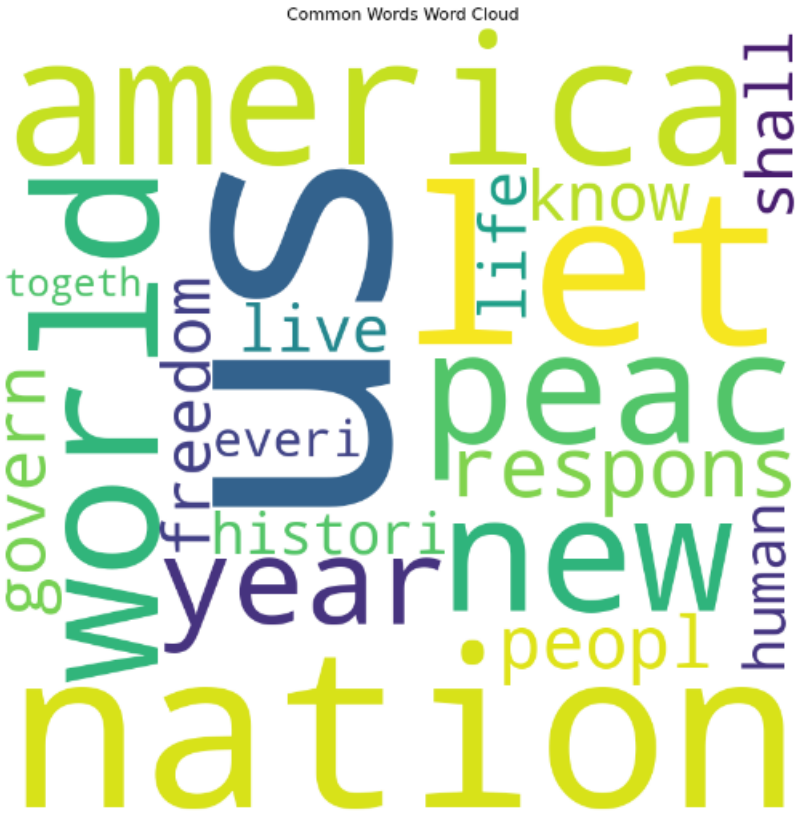

Figure 26 Common Words combined speeches Word cloud

- Most common words for Individual Speeches by the President of America have been identified and a word cloud has been plotted.

- Most common words for all the 3 speeches by the President of United states of America combined have been identified and noted down. A word cloud for the same has been plotted in the Figure 26.

- By this both the Problem statements have been solved. Multiple approaches have been employed especially for problem 1 and have been compared for better performing model.