# Machine Learning - 2

Coded Project

Anirudh Sardiwal

April 28, 2024

# Table of Contents

# 1.0 Voting Prediction

## 1.1 Problem Definition & EDA

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

### 1.1.1 Basics

- The data consists of 1525 rows and 9 columns.
- There are no ‘null’ values in the data. ‘
- Age’ is the only continuous variable. ‘
- Vote’ and ‘Gender’ are categorical and the rest are Ordinal.

### 1.1.2 Data Description

Various features are as follows:

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.

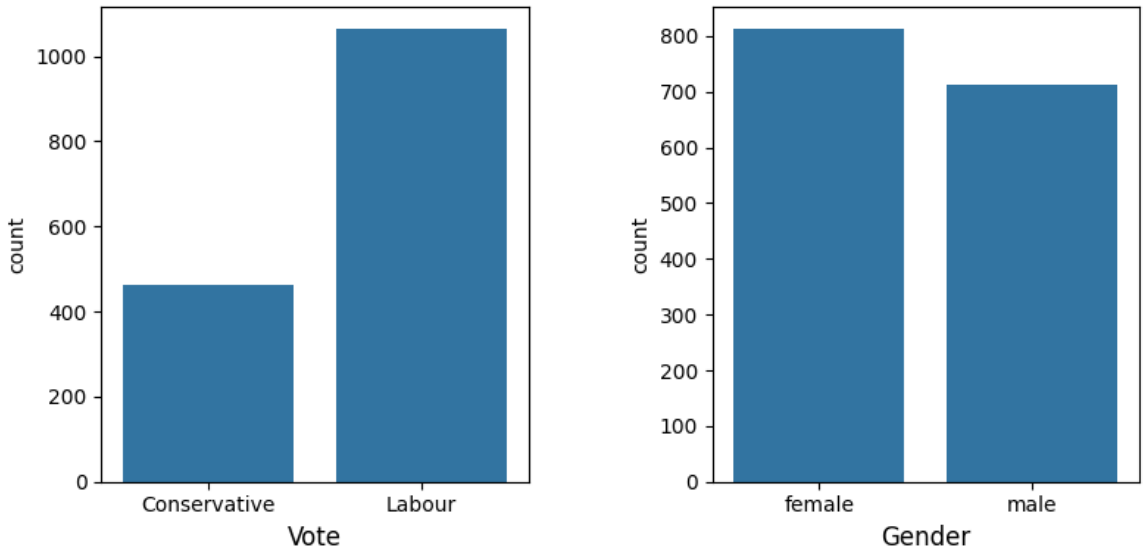
| | vote | age | economic.cond.national | economic.cond.household | Blair | Hague | Europe | political.knowledge | gender |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Labour | 43 | 3 | 3 | 4 | 1 | 2 | 2 | female |
| **1** | Labour | 36 | 4 | 4 | 4 | 4 | 5 | 2 | male |
| **2** | Labour | 35 |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

1.1 - First few rows of dataset

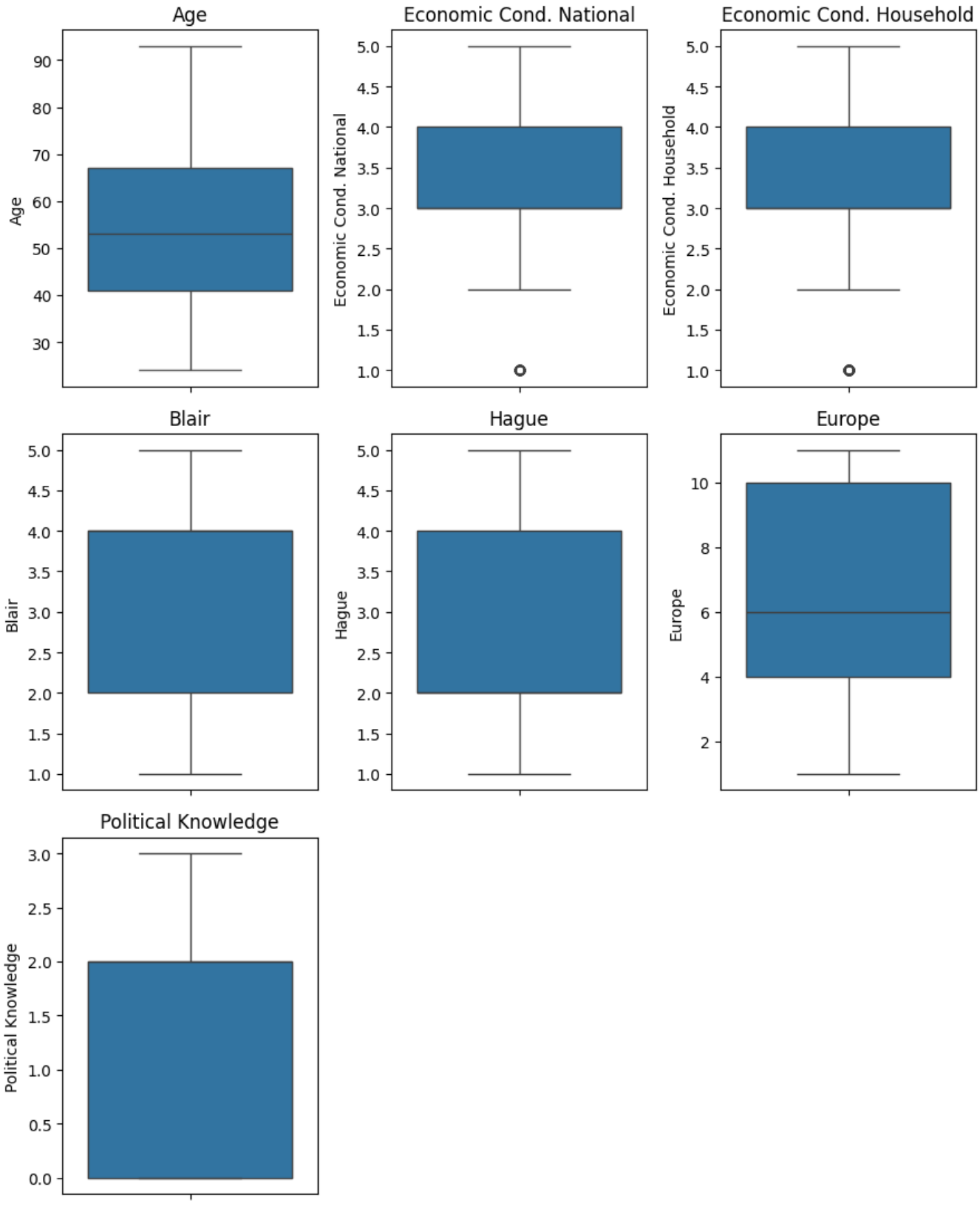
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

1.2 - Data types of features

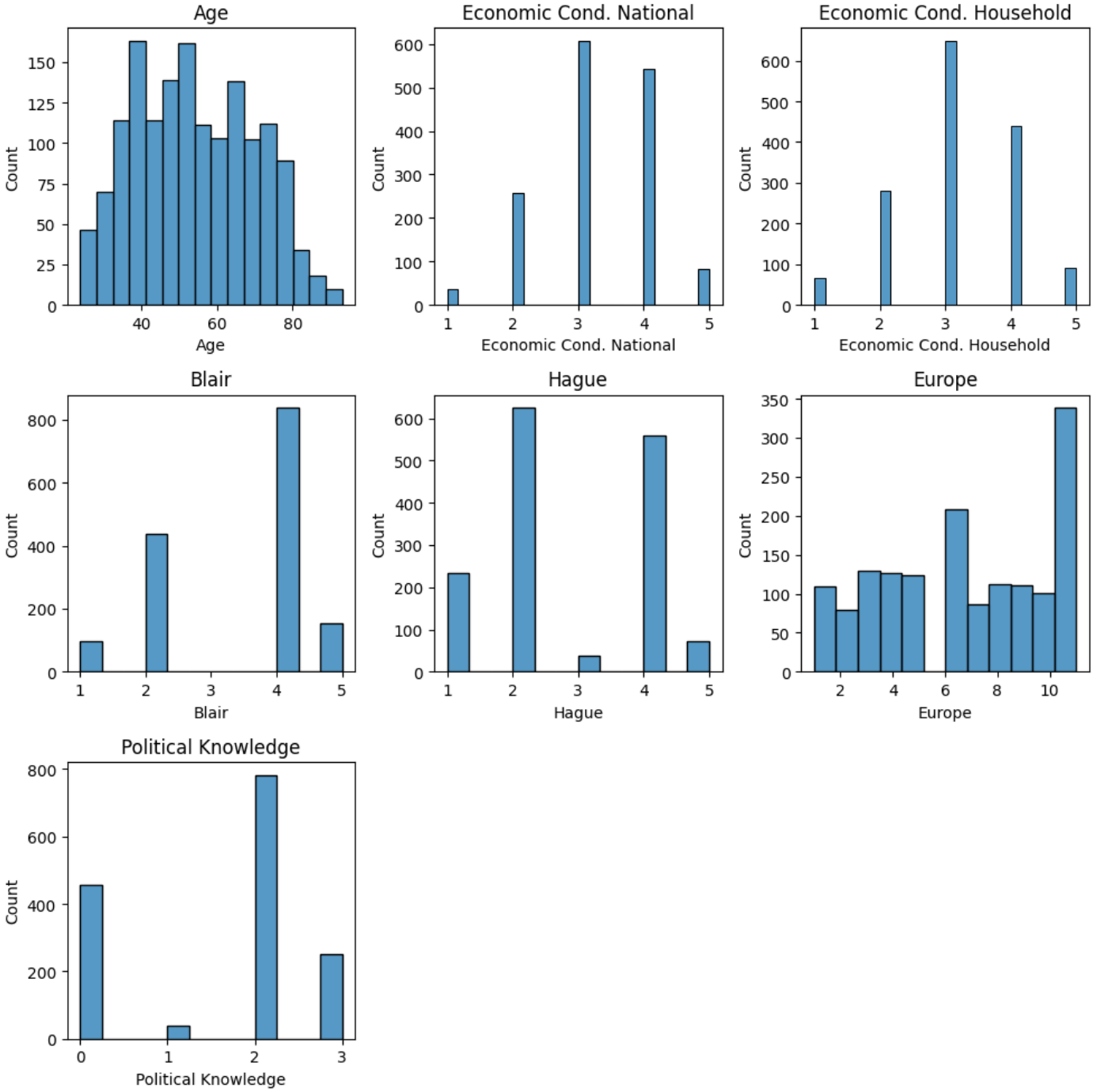
## 1.1.3 Univariate Analysis




1.3 - Countplots of Vote and Gender columns

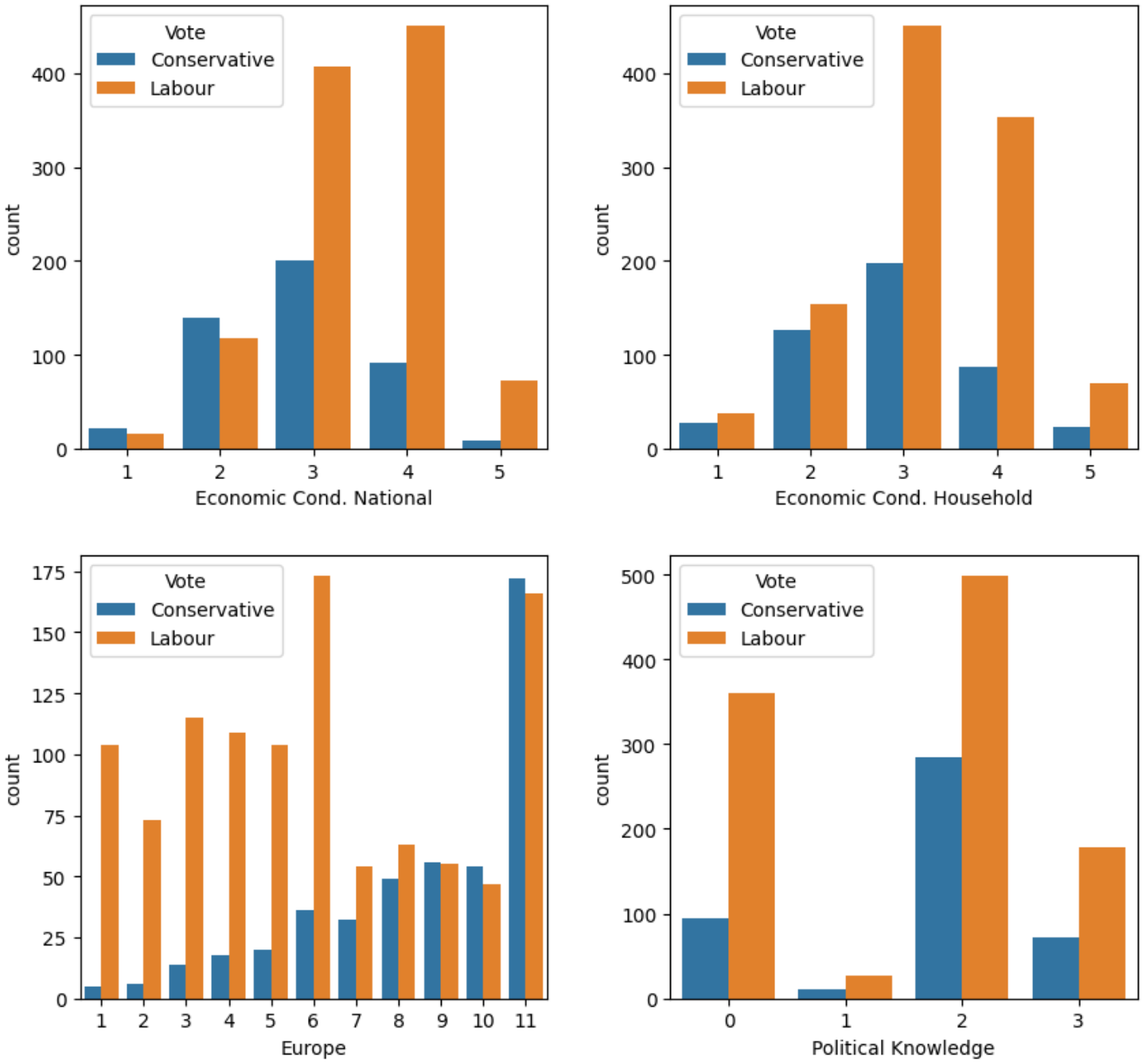



1.4 - Boxplots of Numerical features

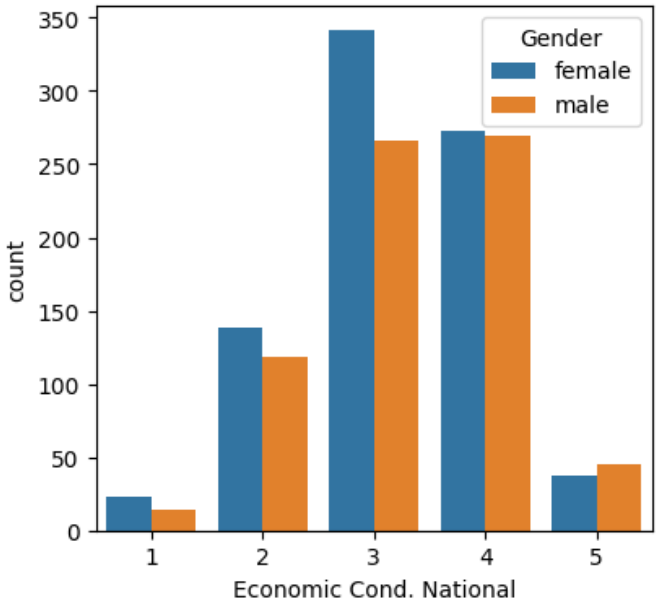


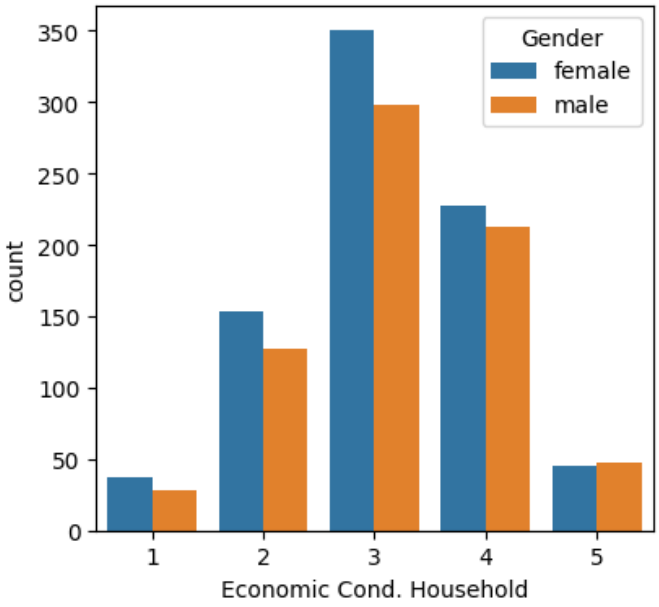

1.5 - Histograms of Numerical features

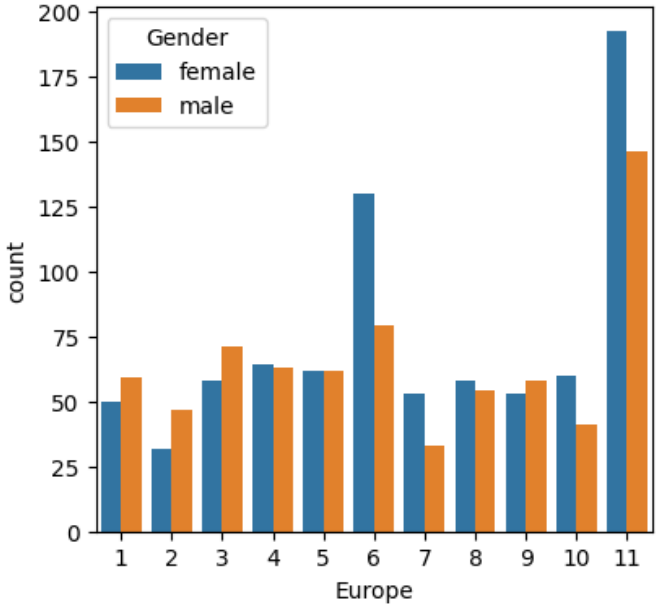
## 1.1.4 Multivariate Analysis

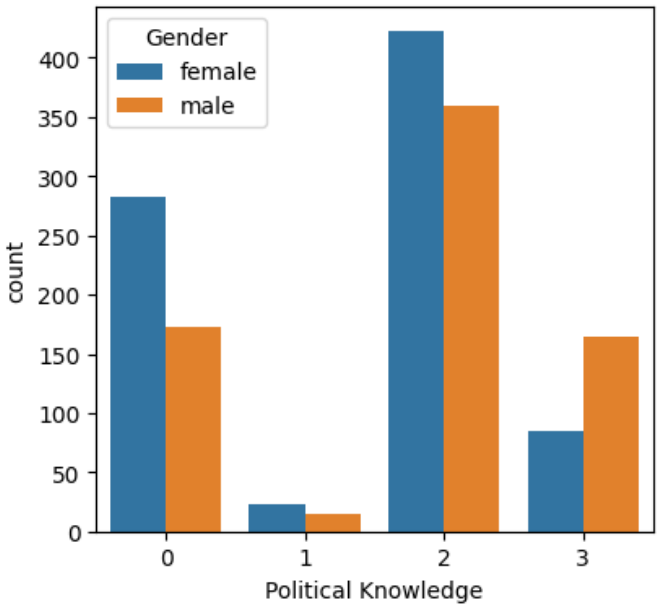



1.6 - Countplots of various features by Vote

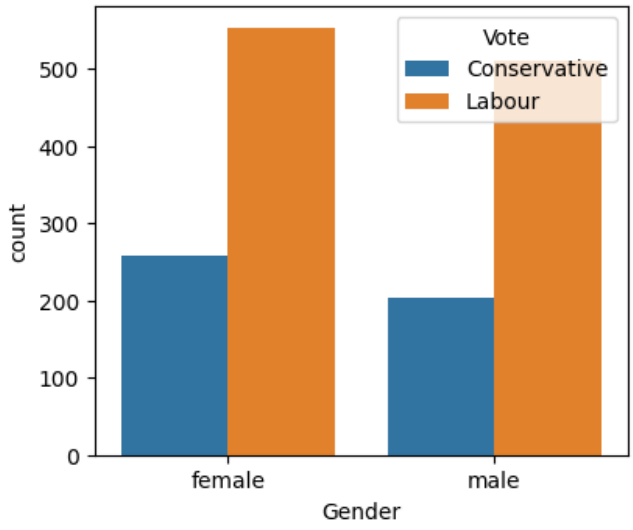


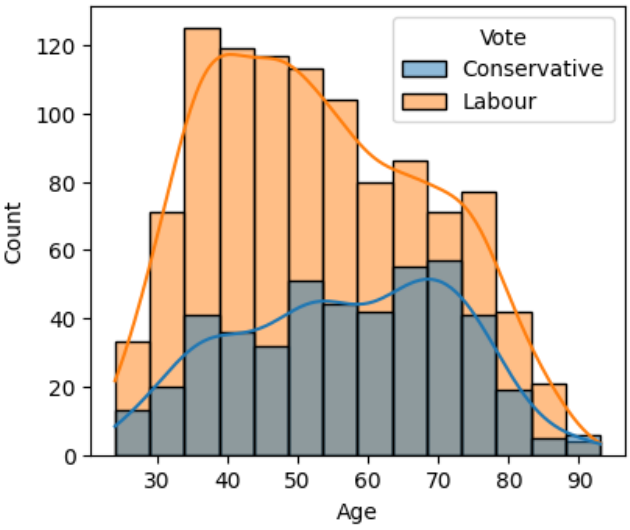



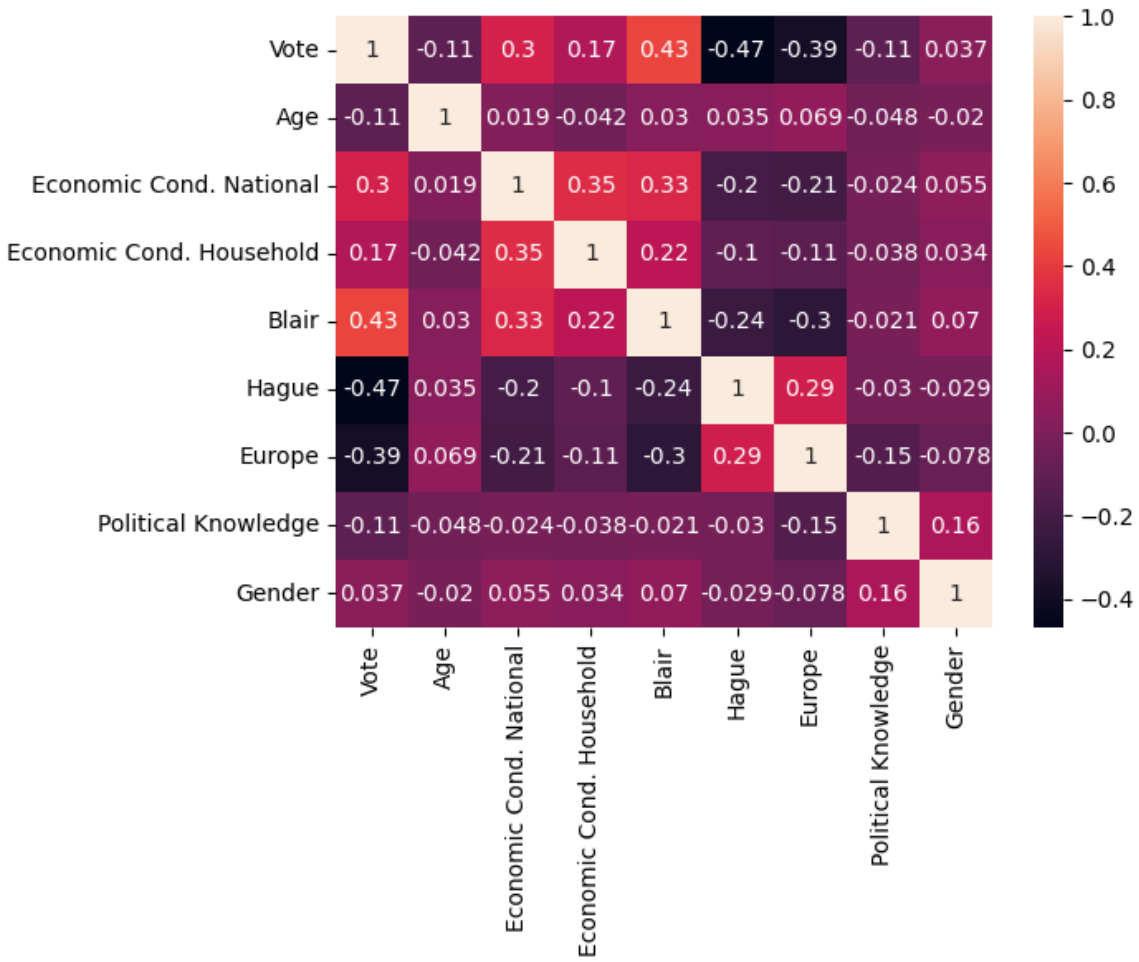




1.7 - Countplots of various features by Gender




1.8 - Votes by Gender and Age




1.9 - Correlation matrix of all features

## 1.1.5 Inferences

- None of the features are correlated with each other.
- Labour Party is clearly more popular.
- There are slightly more female voters. Gender is not having much impact on party selection.
- Majority of voters are between 41 and 67 years of age. Younger people are supporting Labour Party while Older age groups are supporting Conservative party.
- Sentiment is a bit more towards Eurosceptic side.
- Majority of Labour party supporters think National and Household Economic Conditions are normal to good.
- Labour Party supporters favor European integration.

## 1.2 Model Building, Performance Evaluation & Tuning

### 1.2.1 Metrics of Choice

- Overall Accuracy of model is important. We would like to get maximum accuracy as possible. Accuracy of Train and Test Data should be within 10% of each other.
- Precision and Recall - We want Precision of Class 0 to be equal to Precision of Class 1 and same for Recall as well, because we want predictions towards both parties to be equally correct.
- AUC of Train and Test data should be high and within 10% of each other.

### 1.2.2 KNN

Feature encoding is as follows:

- Labour = 1, Conservative = 0
- Male = 1, Female = 0

Scaling is important in KNN because the calculation is based on distance.

| | Vote | Age | Economic Cond. National | Economic Cond. Household | Blair | Hague | Europe | Political Knowledge | Gender |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 43 | 3 | 3 | 4 | 1 | 2 | 2 | 0 |
| **1** | 1 | 36 | 4 | 4 | 4 | 4 | 5 | 2 | 1 |
| **2** | 1 | 35 | 4 | 4 | 5 | 2 | 3 | 2 | 1 |
| **3** | 1 | 24 | 4 | 2 | 2 | 1 | 4 | 0 | 0 |
| **4** | 1 | 41 | 2 | 2 | 1 | 1 | 6 | 2 | 1 |

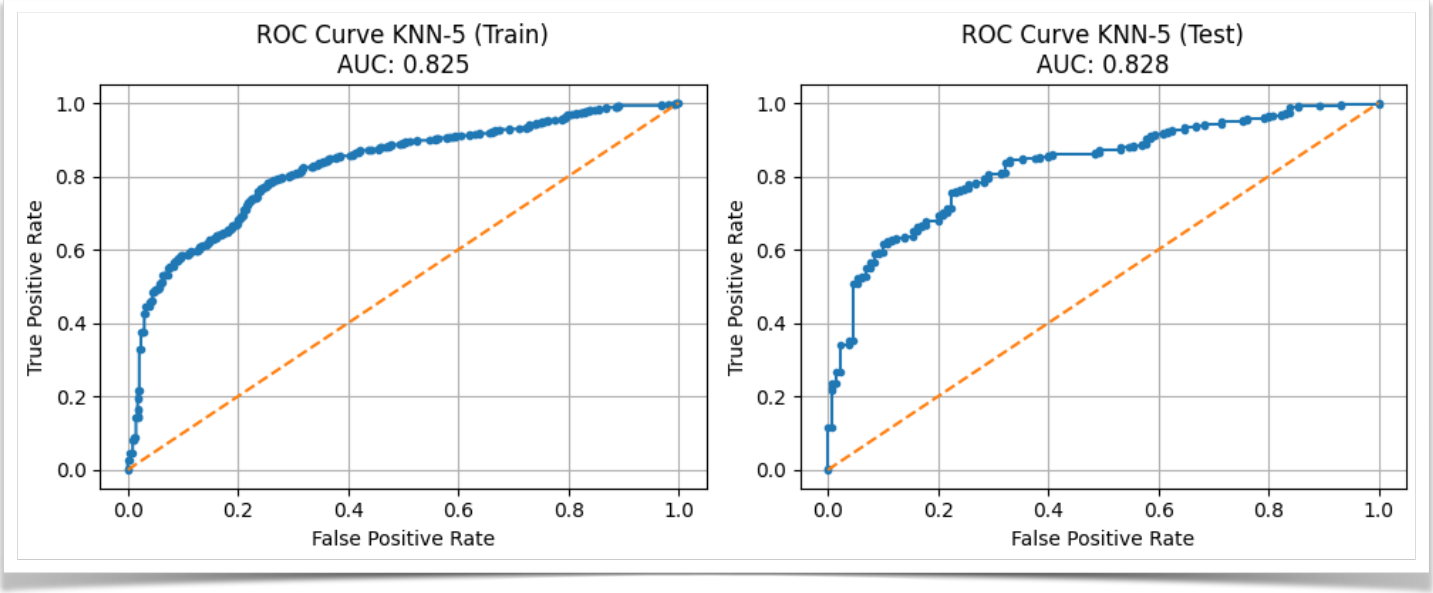
1.2.1 - Head of encoded dataframe

| | Age | Economic Cond. National | Economic Cond. Household | Blair | Hague | Europe | Political Knowledge | Gender |
|---|---|---|---|---|---|---|---|---|
| **0** | -0.711973 | -0.279218 | -0.150948 | 0.566716 | -1.419886 | -1.434426 | 0.422643 | -0.937059 |
| **1** | -1.157661 | 0.856268 | 0.924730 | 0.566716 | 1.018544 | -0.524358 | 0.422643 | 1.067169 |
| **2** | -1.221331 | 0.856268 | 0.924730 | 1.418187 | -0.607076 | -1.131070 | 0.422643 | 1.067169 |
| **3** | -1.921698 | 0.856268 | -1.226625 | -1.136225 | -1.419886 | -0.827714 | -1.424148 | -0.937059 |
| **4** | -0.839313 | -1.414704 | -1.226625 | -1.987695 | -1.419886 | -0.221002 | 0.422643 | 1.067169 |

1.2.2 - Head of scaled dataframe

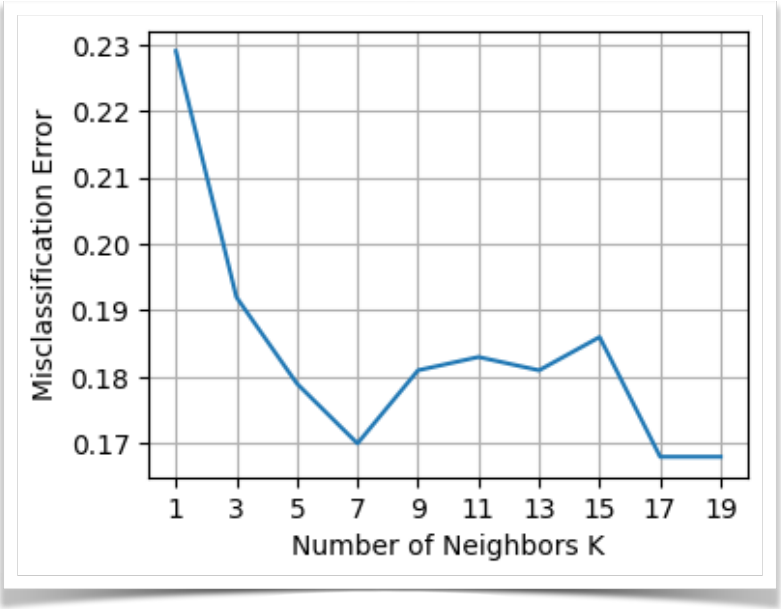
```
Model Score for KNN-5 Train: 0.74

Confusion Matrix:
[[261  71]
 [209 526]]
              precision    recall  f1-score   support

           0       0.56      0.79      0.65       332
           1       0.88      0.72      0.79       735

    accuracy                           0.74      1067
   macro avg       0.72      0.75      0.72      1067
weighted avg       0.78      0.74      0.75      1067
```

```
Model Score for KNN-5 Test: 0.73

Confusion Matrix:
[[103  27]
 [ 98 230]]
              precision    recall  f1-score   support

           0       0.51      0.79      0.62       130
           1       0.89      0.70      0.79       328

    accuracy                           0.73       458
   macro avg       0.70      0.75      0.70       458
weighted avg       0.79      0.73      0.74       458
```




1.2.3 - ROC curves of KNN with 5 neighbors

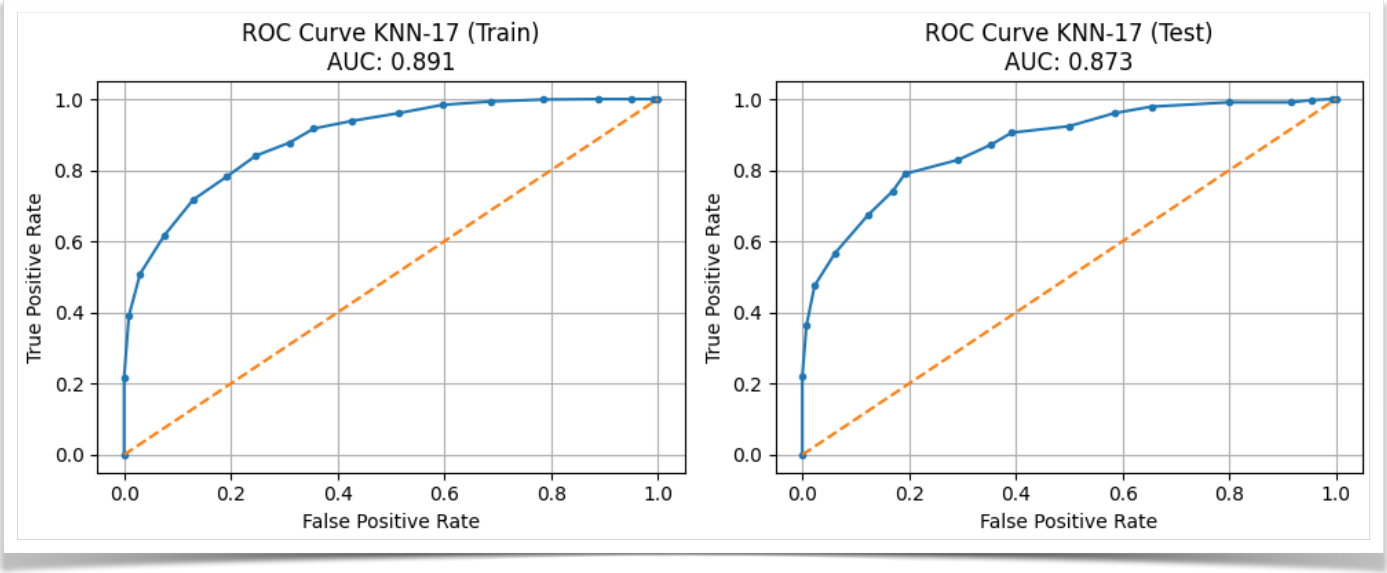
## KNN Model Tuning




1.2.3 - Determining the optimum number of K-neighbors

```
Model Score for KNN-17 Train: 0.7

Confusion Matrix:
[[ 22 310]
 [  6 729]]
              precision    recall  f1-score   support

           0       0.79      0.07      0.12       332
           1       0.70      0.99      0.82       735

    accuracy                           0.70      1067
   macro avg       0.74      0.53      0.47      1067
weighted avg       0.73      0.70      0.60      1067
```

```
Model Score for KNN-17 Test: 0.74

Confusion Matrix:
[[ 12 118]
 [  2 326]]
              precision    recall  f1-score   support

           0       0.86      0.09      0.17       130
           1       0.73      0.99      0.84       328

    accuracy                           0.74       458
   macro avg       0.80      0.54      0.51       458
weighted avg       0.77      0.74      0.65       458
```



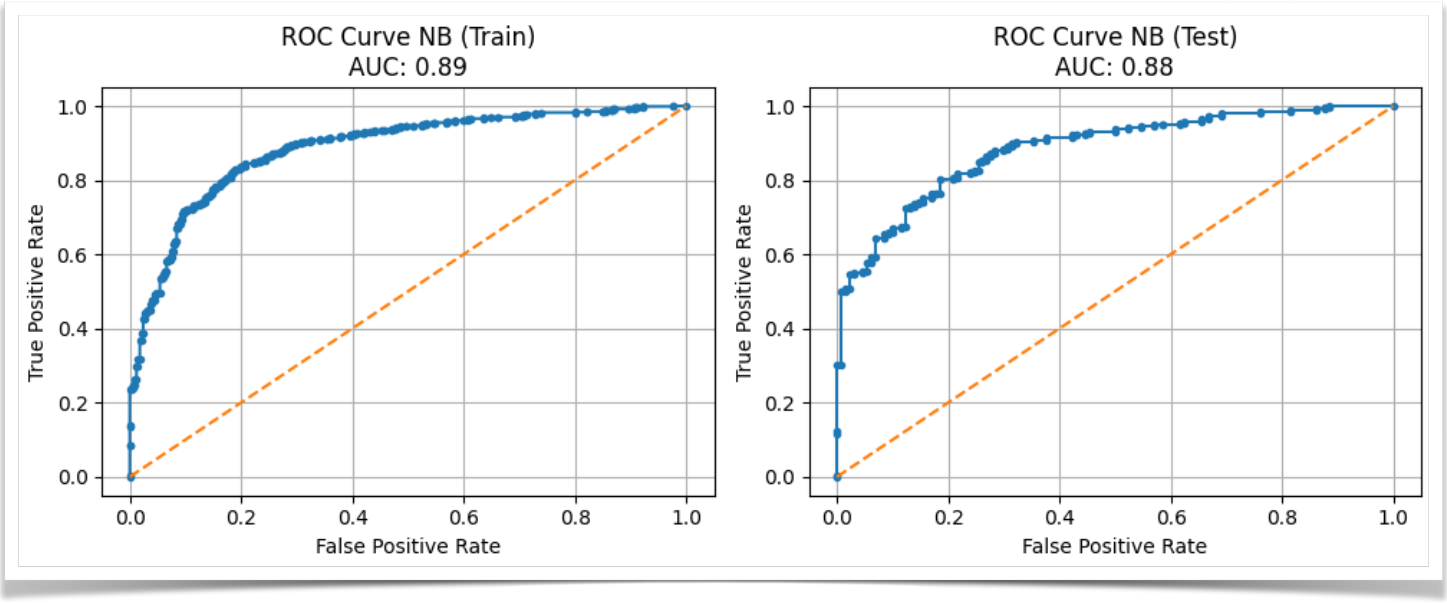

1.2.4 - ROC curves of KNN with 17 neighbors

## 1.2.3 Naive Bayes

Rest of the models other than KNN will not use scaled data.

```
Model Score for Naive Bayes Train: 0.83

Confusion Matrix:
[[240  92]
 [ 86 649]]
              precision    recall  f1-score   support

           0       0.74      0.72      0.73       332
           1       0.88      0.88      0.88       735

    accuracy                           0.83      1067
   macro avg       0.81      0.80      0.80      1067
weighted avg       0.83      0.83      0.83      1067
```

```
Model Score for Naive Bayes Test: 0.83

Confusion Matrix:
[[ 94  36]
 [ 44 284]]
              precision    recall  f1-score   support

           0       0.68      0.72      0.70       130
           1       0.89      0.87      0.88       328

    accuracy                           0.83       458
   macro avg       0.78      0.79      0.79       458
weighted avg       0.83      0.83      0.83       458
```

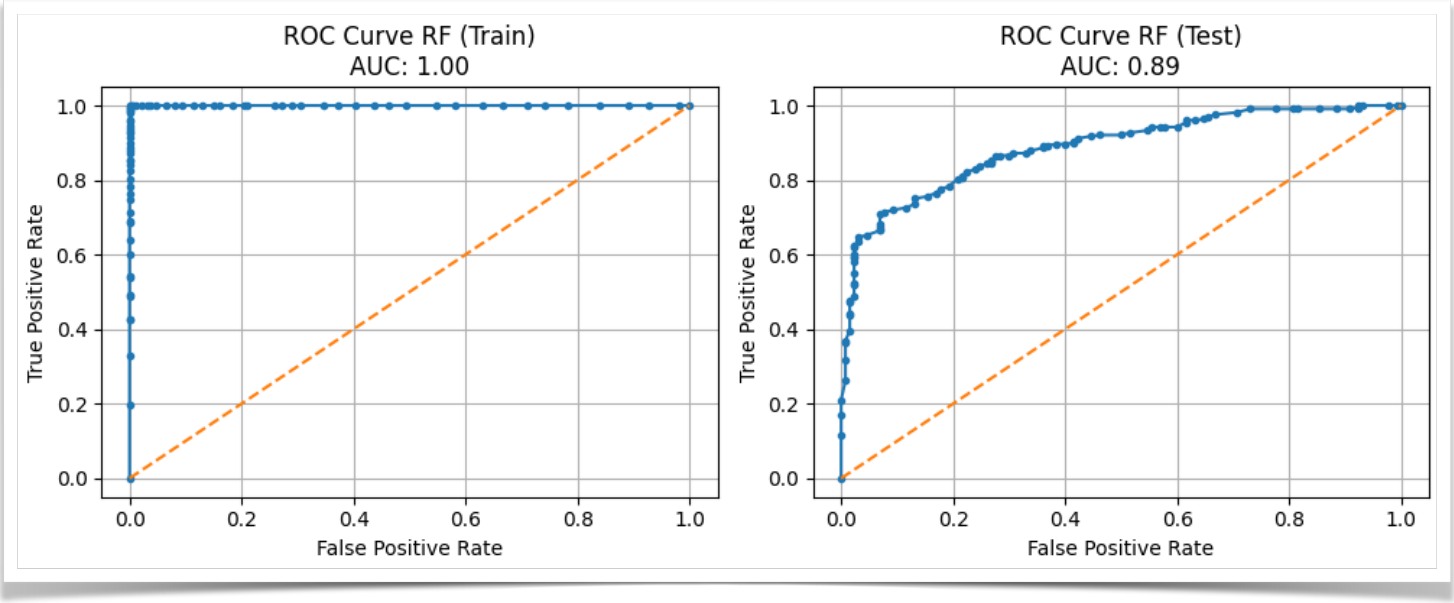



1.2.5 - ROC curves of Naive Bayes

## 1.2.4 Bagging - Random Forest

```
Model Score for Random Forest (Train): 1.0

Confusion Matrix:
[[331   1]
 [  0 735]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       332
           1       1.00      1.00      1.00       735

    accuracy                           1.00      1067
   macro avg       1.00      1.00      1.00      1067
weighted avg       1.00      1.00      1.00      1067
```

```
Model Score for Random Forest (Test): 0.82

Confusion Matrix:
[[ 90  40]
 [ 42 286]]
              precision    recall  f1-score   support

           0       0.68      0.69      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
```




1.2.6 - ROC curves of Random Forest

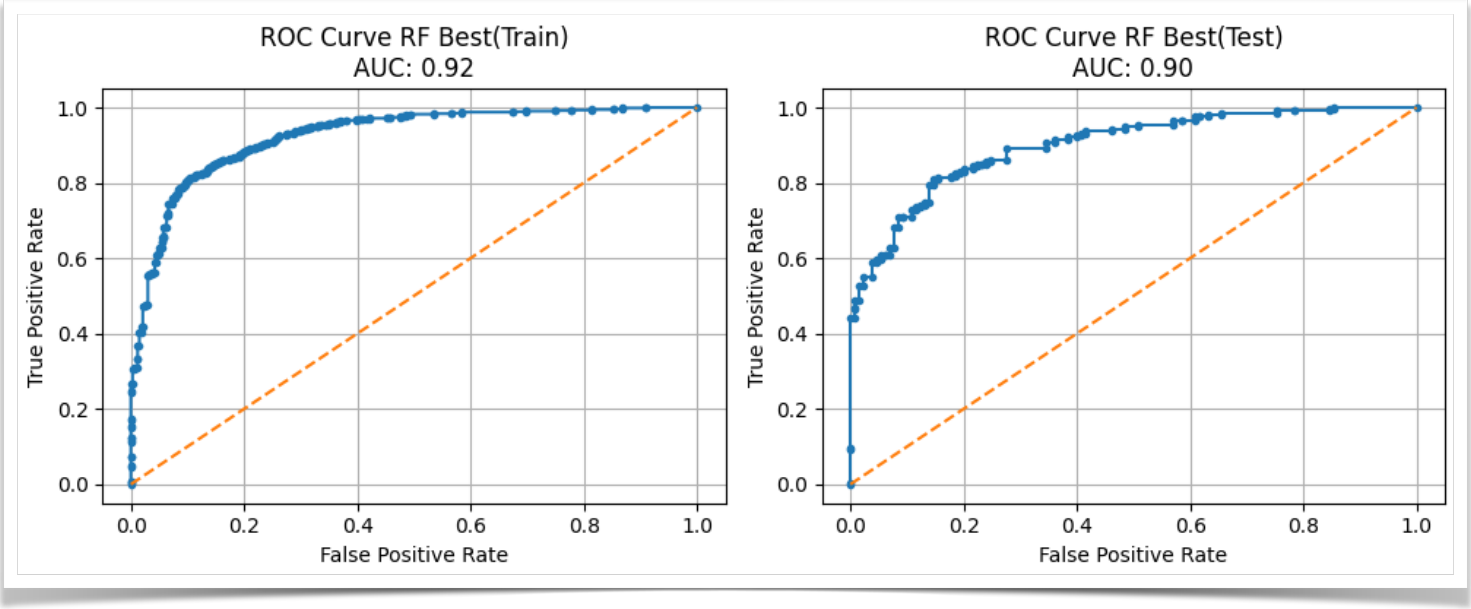
## Random Forest Model Tuning

Using RandomizedSearchCV, the Best Parameters are:

- 'n_estimators': 100,
- 'min_samples_split': 2,
- 'min_samples_leaf': 2,
- 'max_features': 'log2',
- 'max_depth': 5

```
Model Score for Random Forest Tuned (Train): 0.87

Confusion Matrix:
[[241  91]
 [ 51 684]]
              precision    recall  f1-score   support

           0       0.83      0.73      0.77       332
           1       0.88      0.93      0.91       735

    accuracy                           0.87      1067
   macro avg       0.85      0.83      0.84      1067
weighted avg       0.86      0.87      0.86      1067
```

```
Model Score for Random Forest Tuned (Test): 0.83

Confusion Matrix:
[[ 87  43]
 [ 34 294]]
              precision    recall  f1-score   support

           0       0.72      0.67      0.69       130
           1       0.87      0.90      0.88       328

    accuracy                           0.83       458
   macro avg       0.80      0.78      0.79       458
weighted avg       0.83      0.83      0.83       458
```




1.2.7 - ROC curves of Random Forest Tuned

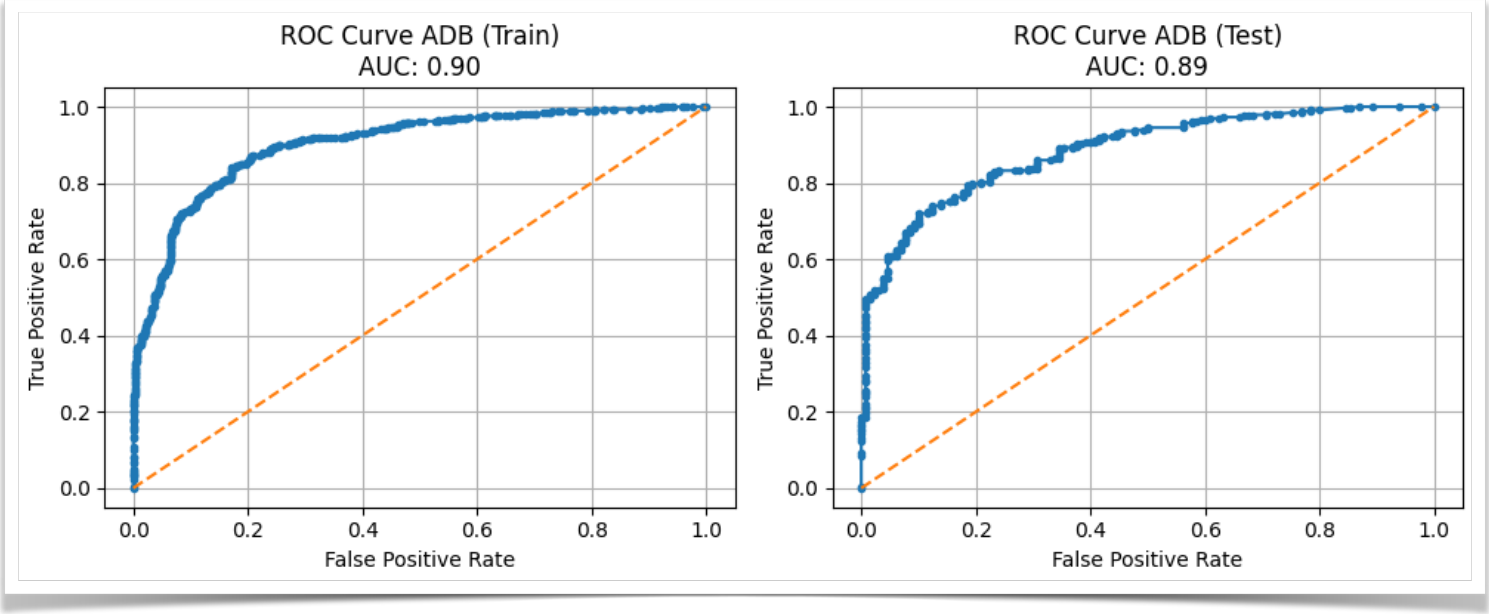
| | Imp |
|---|---|
| Hague | 0.30 |
| Europe | 0.23 |
| Blair | 0.18 |
| Political Knowledge | 0.09 |
| Economic Cond. National | 0.08 |
| Age | 0.07 |
| Economic Cond. Household | 0.04 |
| Gender | 0.01 |

1.2.8 - Feature importances - Random Forest Tuned

## 1.2.5 Boosting - AdaBoost

```
Model Score for AdaBoost (Train): 0.85

Confusion Matrix:
[[234  98]
 [ 64 671]]
              precision    recall  f1-score   support

           0       0.79      0.70      0.74       332
           1       0.87      0.91      0.89       735

    accuracy                           0.85      1067
   macro avg       0.83      0.81      0.82      1067
weighted avg       0.85      0.85      0.85      1067
```

```
Model Score for AdaBoost (Test): 0.81

Confusion Matrix:
[[ 85  45]
 [ 40 288]]
              precision    recall  f1-score   support

           0       0.68      0.65      0.67       130
           1       0.86      0.88      0.87       328

    accuracy                           0.81       458
   macro avg       0.77      0.77      0.77       458
weighted avg       0.81      0.81      0.81       458
```




1.2.9 - ROC curves of AdaBoost model

```
                            Imp
Blair                      0.27
Age                        0.21
Europe                     0.17
Hague                      0.16
Political Knowledge        0.10
Economic Cond. National    0.06
Economic Cond. Household   0.02
Gender                     0.01
```

1.2.10 - Feature Importances - AdaBoost

## AdaBoost Model Tuning

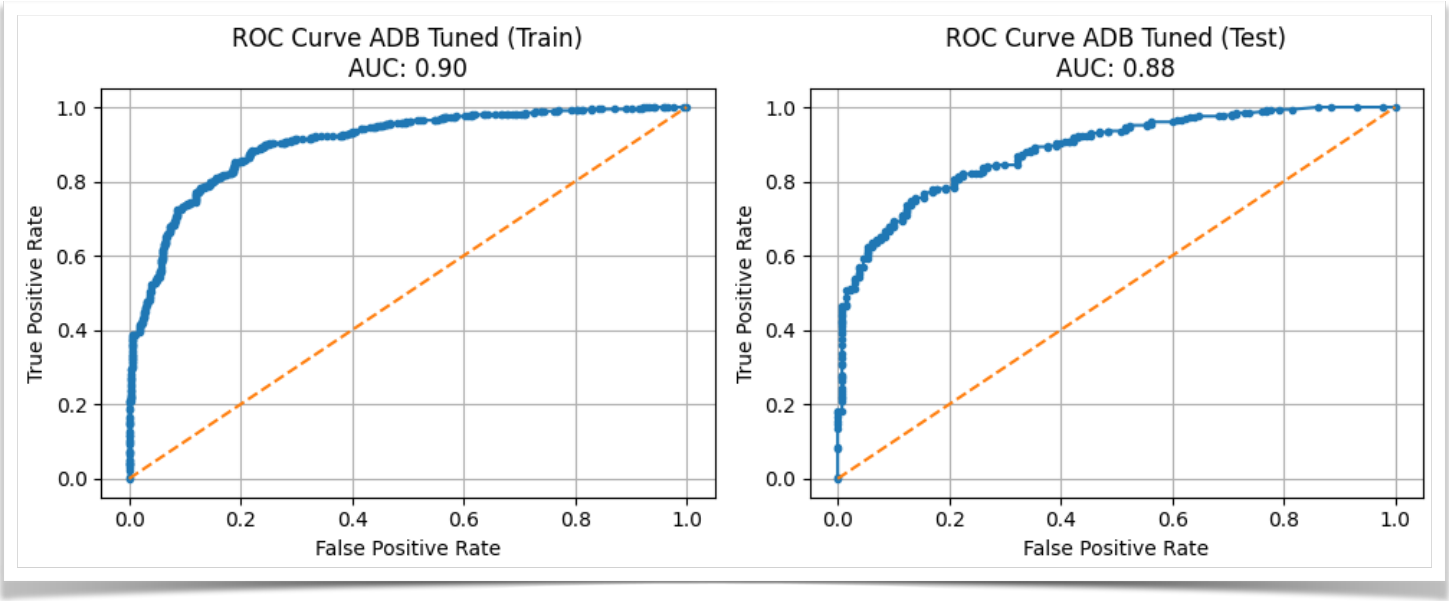
Using RandomizedSearchCV, best parameters:

'learning_rate': 1.0,

'n_estimators': 149

```
Model Score for AdaBoost, Tuned (Train): 0.85

Confusion Matrix:
[[233  99]
 [ 66 669]]
              precision    recall  f1-score   support

           0       0.78      0.70      0.74       332
           1       0.87      0.91      0.89       735

    accuracy                           0.85      1067
   macro avg       0.83      0.81      0.81      1067
weighted avg       0.84      0.85      0.84      1067
```

```
Model Score for AdaBoost, Tuned (Test): 0.81

Confusion Matrix:
[[ 85  45]
 [ 40 288]]
              precision    recall  f1-score   support

           0       0.68      0.65      0.67       130
           1       0.86      0.88      0.87       328

    accuracy                           0.81       458
   macro avg       0.77      0.77      0.77       458
weighted avg       0.81      0.81      0.81       458
```




1.2.11 - ROC curves of AdaBoost model, Tuned

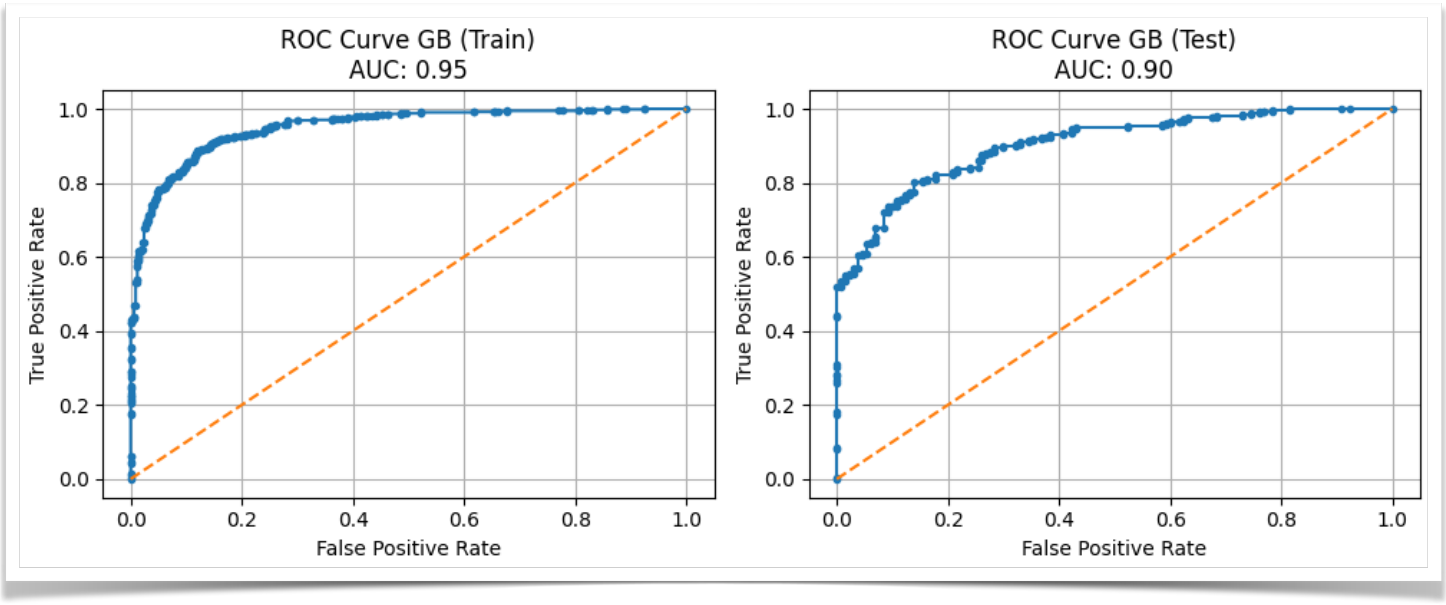
| | Imp |
|---|---|
| Blair | 0.26 |
| Age | 0.25 |
| Europe | 0.16 |
| Hague | 0.15 |
| Political Knowledge | 0.10 |
| Economic Cond. National | 0.05 |
| Economic Cond. Household | 0.02 |
| Gender | 0.01 |

1.2.12 - Feature Importances - AdaBoost, Tuned

## 1.2.6 Boosting - Gradient Boost

```
Model Score for Gradient Boost (Train): 0.89

Confusion Matrix:
[[262  70]
 [ 51 684]]
              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
```

```
Model Score for Gradient Boost (Test): 0.83

Confusion Matrix:
[[ 96  34]
 [ 43 285]]
              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
```




1.2.13 - ROC curves of Gradient Boost model

| | Imp |
|---|---|
| Hague | 0.35 |
| Europe | 0.19 |
| Blair | 0.18 |
| Age | 0.11 |
| Political Knowledge | 0.11 |
| Economic Cond. National | 0.04 |
| Economic Cond. Household | 0.02 |
| Gender | 0.00 |

1.2.14 - Feature Importances - Gradient Boost model

## Gradient Boost Model Tuning

Using RandomizedSearchCV, best Hyperparameters:

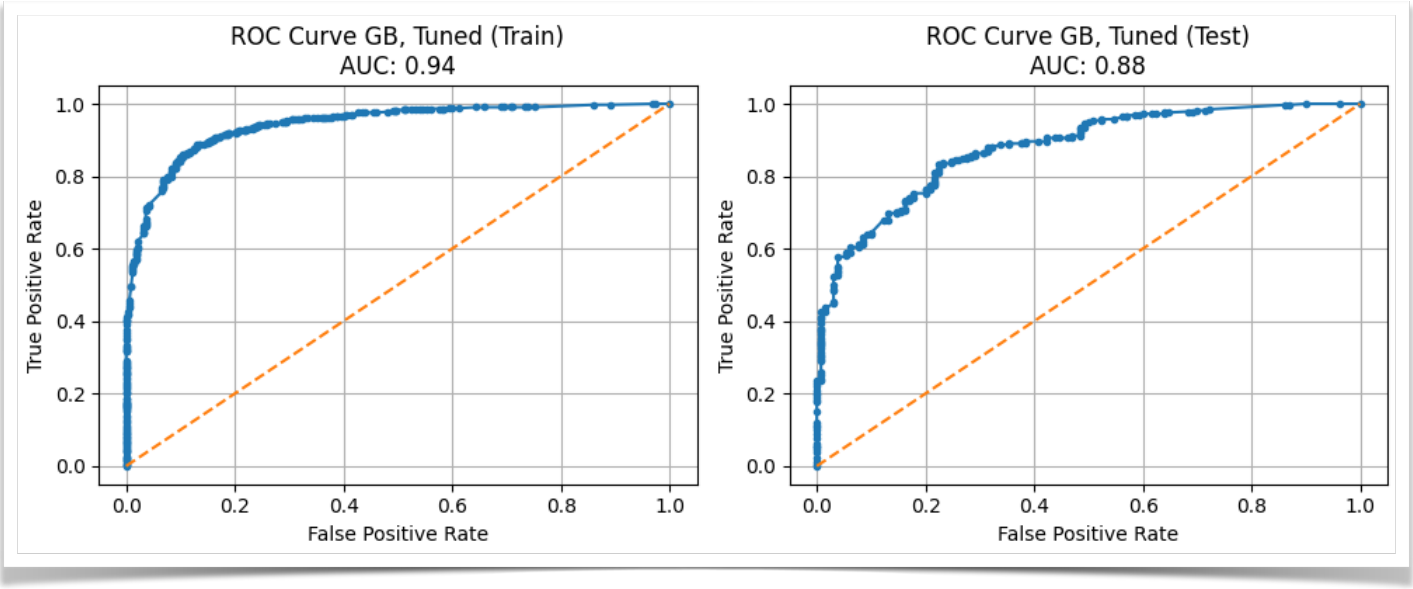
'learning_rate': 0.01,

'max_depth': 5,

'n_estimators': 108

```
Model Score for Gradient Boost, Tuned (Train): 0.88

Confusion Matrix:
[[236  96]
 [ 37 698]]
              precision    recall  f1-score   support

           0       0.86      0.71      0.78       332
           1       0.88      0.95      0.91       735

    accuracy                           0.88      1067
   macro avg       0.87      0.83      0.85      1067
weighted avg       0.87      0.88      0.87      1067
```

```
Model Score for Gradient Boost, Tuned (Test): 0.81

Confusion Matrix:
[[ 81  49]
 [ 36 292]]
              precision    recall  f1-score   support

           0       0.69      0.62      0.66       130
           1       0.86      0.89      0.87       328

    accuracy                           0.81       458
   macro avg       0.77      0.76      0.76       458
weighted avg       0.81      0.81      0.81       458
```




1.2.15 - ROC curves of Gradient Boost model, Tuned

| | Imp |
|---|---|
| Hague | 0.38 |
| Europe | 0.18 |
| Blair | 0.16 |
| Political Knowledge | 0.12 |
| Age | 0.09 |
| Economic Cond. National | 0.04 |
| Economic Cond. Household | 0.02 |
| Gender | 0.01 |

1.2.16 - Feature Importances - Gradient Boost model, Tuned

## 1.3 Final Model Selection

| Model | Accuracy | | Train - Precision | | Train - Recall | | Test - Precision | | Test - Recall | | AUC | | Important Features |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | Train | Test | |
| KNN - 5 neighbors | 0.74 | 0.73 | 0.56 | 0.88 | 0.79 | 0.72 | 0.51 | 0.89 | 0.79 | 0.70 | 0.83 | 0.83 | None |
| KNN - 17 neighbors | 0.70 | 0.74 | 0.79 | 0.70 | 0.07 | 0.99 | 0.86 | 0.73 | 0.09 | 0.99 | 0.89 | 0.87 | None |
| Naive Bayes | 0.83 | 0.83 | 0.74 | 0.88 | 0.72 | 0.88 | 0.68 | 0.89 | 0.72 | 0.87 | 0.89 | 0.88 | None |
| Random Forest | 1.0 | 0.82 | 1.0 | 1.0 | 1.0 | 1.0 | 0.68 | 0.88 | 0.69 | 0.87 | 1.0 | 0.89 | Age, Europe, Hague |
| Random Forest, Tuned | 0.87 | 0.83 | 0.83 | 0.88 | 0.73 | 0.93 | 0.72 | 0.87 | 0.67 | 0.90 | 0.92 | 0.90 | Hague, Europe, Blair |
| Ada Boosting | 0.85 | 0.81 | 0.79 | 0.87 | 0.70 | 0.91 | 0.68 | 0.86 | 0.65 | 0.88 | 0.90 | 0.89 | Blair, Age, Europe, Hague |
| Ada Boosting, Tuned | 0.85 | 0.81 | 0.78 | 0.87 | 0.70 | 0.91 | 0.68 | 0.86 | 0.65 | 0.88 | 0.90 | 0.88 | Blair, Age, Europe, Hague |
| Gradient Boosting | 0.89 | 0.83 | 0.84 | 0.91 | 0.79 | 0.93 | 0.69 | 0.89 | 0.74 | 0.87 | 0.95 | 0.9 | Hague, Europe, Blair |
| Gradient Boosting, Tuned | 0.88 | 0.81 | 0.86 | 0.88 | 0.71 | 0.95 | 0.69 | 0.86 | 0.62 | 0.89 | 0.94 | 0.88 | Hague, Europe, Blair |

1.3.1 - Model comparison matrix

### 1.3.1 Commentary

1. **KNN, 5 Neighbors** — Accuracy is low, average AUC, has over 30% difference in Precision values of both Train and Test data.
2. **KNN, 17 Neighbors** — poor Accuracy, very poor Recall for Class 0 in both Train and Test data.

3. **Naive Bayes** — Above average Accuracy, high AUC, lesser difference between Precision and Recall values of both classes.
4. **Random Forest, Default** — Overfitted
5. **Random Forest, Tuned** — Above average Accuracy, high AUC, average difference between Precision and Recall values of both classes.
6. **Ada Boost, Default** — Above average Accuracy, high AUC, average difference between Precision and Recall values of both classes.
7. **Ada Boost, Tuned** — Above average Accuracy, high AUC, average to low difference between Precision and Recall values of both classes. Precision and Recall for Class 0 is less probably because of lesser number of samples.
8. **Gradient Boost, Default** — Best Precision and Recall values on Test data, AUC and Accuracy differences <10%.
9. **Gradient Boost, Tuned** — As compared to Gradient Boost, Default, higher differences between Recall on Train and Test data, poor Test Recall for Class 0

### 1.3.2 Selected Model

Final model selected is **No. 8 - Gradient Boost, Default**, because:

- 83% Accuracy on Test data - highest among all
- 90% AUC score on Test data - highest among all
- Difference between Precision on Test data between both classes : 20%, which is average as compared
- Difference between Recall on Test data between both classes : 13%, which is 2nd best among all

## 1.4 Actionable Insights & Recommendations

1. **Age Range** — Make sure data is captured from voters between ages 41 and 67, because age distribution on captured data shows majority of voters from this age range. Gender is irrelevant.
2. **Important Features** — Make sure data is captured on the topics of Hague, Blair, and Europe, because these are features of highest importance and will therefore have highest predictive power.
3. **Accuracy** — The model achieves an Accuracy of 83%, that means it's important to keep a margin of error while predicting the final result. For example if the Labour Party has 1000 votes and Conservative has 900, we cannot be completely sure that Labour Party will win because we can be sure of only 830 votes.
4. **Precision** — Precision for class 0 (Conservative party) is 69% and for class 1 (Labour party) is 89%, suggesting that the model is better at predicting votes for the Labour party. We can trust only 69% of the predictions for Conservative party.
5. **Discrimination** — The AUC score of 0.90 indicates that the model has a strong ability to differentiate between the two classes.
6. **Focus on Labour Party** — Focus on leveraging the model's strengths in predicting votes for the Labour Party (class 1), as it demonstrates higher precision and recall for this class.
7. **Predictive performance for Conservative party** — Need to improve the model's performance for predicting votes for the Conservative party (class 0) by collecting additional data.

# 2.0 Presidential Speech Analysis

## 2.1 Problem Definition & EDA

There are 3 presidential speeches which need to be analyzed for the kind of words used and the expression arising thereby.

```
Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and
good country we share together:

When we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroad and
of destructive conflict at home.

As we meet here today, we stand on the threshold of a new era of peace in the world.

The central question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not
be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invi
tes new danger abroad.

Let us resolve that this will be what it can become: a time of great responsibilities greatly borne, in which we renew the
spirit and the promise of America as we enter our third century as a nation.
```

2.1.1 - Part of text from a speech

| Metric | Nixon | Kennedy | Roosevelt |
|---|---|---|---|
| **Word Count** | 1,819 | 1,390 | 1,360 |
| **Character Count (incl. Spaces)** | 9,991 | 7,618 | 7,571 |
| **Character Count (excl. Spaces)** | 8,223 | 6,255 | 6,249 |
| **Sentence Count** | 68 | 52 | 68 |
| **Avg. Word Length** | 4.47 | 4.46 | 4.54 |

## 2.2 Text Cleaning

### 2.2.1 Removal of Punctuation Marks

```
Mr Vice President Mr Speaker Mr Chief Justice Senator Cook Mrs Eisenhower and my fellow citizens of this great and good cou
ntry we share together

When we met here four years ago America was bleak in spirit depressed by the prospect of seemingly endless war abroad and o
f destructive conflict at home

As we meet here today we stand on the threshold of a new era of peace in the world

The central question before us is How shall we use that peace Let us resolve that this era we are about to enter will not b
e what other postwar periods have so often been a time of retreat and isolation that leads to stagnation at home and invite
s new danger abroad

Let us resolve that this will be what it can become a time of great responsibilities greatly borne in which we renew the sp
irit and the promise of America as we enter our third century as a nation

This past year saw farreaching results from our new policies for peace By continuing to revitalize our traditional friendsh
```

2.2.1 - After removal of punctuation marks

## 2.2.2 Removal of Stopwords

```
['mr', 'vice', 'president', 'mr', 'speaker', 'mr', 'chief', 'justice', 'senator', 'cook', 'mrs', 'eisenhower', 'fellow', 'c
itizens', 'great', 'good', 'country', 'share', 'together', 'met', 'four', 'years', 'ago', 'america', 'bleak', 'spirit', 'de
pressed', 'prospect', 'seemingly', 'endless', 'war', 'abroad', 'destructive', 'conflict', 'home', 'meet', 'today', 'stand',
'threshold', 'new', 'era', 'peace', 'world', 'central', 'question', 'us', 'shall', 'use', 'peace', 'let', 'us', 'resolve',
'era', 'enter', 'postwar', 'periods', 'often', 'time', 'retreat', 'isolation', 'leads', 'stagnation', 'home', 'invites', 'n
ew', 'danger', 'abroad', 'let', 'us', 'resolve', 'become', 'time', 'great', 'responsibilities', 'greatly', 'borne', 'rene
w', 'spirit', 'promise', 'america', 'enter', 'third', 'century', 'nation', 'past', 'year', 'saw', 'farreaching', 'results',
'new', 'policies', 'peace', 'continuing', 'revitalize', 'traditional', 'friendships', 'missions', 'peking', 'moscow', 'abl
e', 'establish', 'base', 'new', 'durable', 'pattern', 'relationships', 'among', 'nations', 'world', 'americas', 'bold', 'in
itiatives', '1972', 'long', 'remembered', 'year', 'greatest', 'progress', 'since', 'end', 'world', 'war', 'ii', 'toward', '
lasting', 'peace', 'world', 'peace', 'seek', 'world', 'flimsy', 'peace', 'merely', 'interlude', 'wars', 'peace', 'endure',
'generations', 'come', 'important', 'understand', 'necessity', 'limitations', 'americas', 'role', 'maintaining', 'peace', '
unless', 'america', 'work', 'preserve', 'peace', 'peace', 'unless', 'america', 'work', 'preserve', 'freedom', 'freedom', 'l
et', 'us', 'clearly', 'understand', 'new', 'nature', 'americas', 'role', 'result', 'new', 'policies', 'adopted', 'past', 'f
our', 'years', 'shall', 'respect', 'treaty', 'commitments', 'shall', 'support', 'vigorously', 'principle', 'country', 'righ
```

2.2.2 - After removal of Stopwords

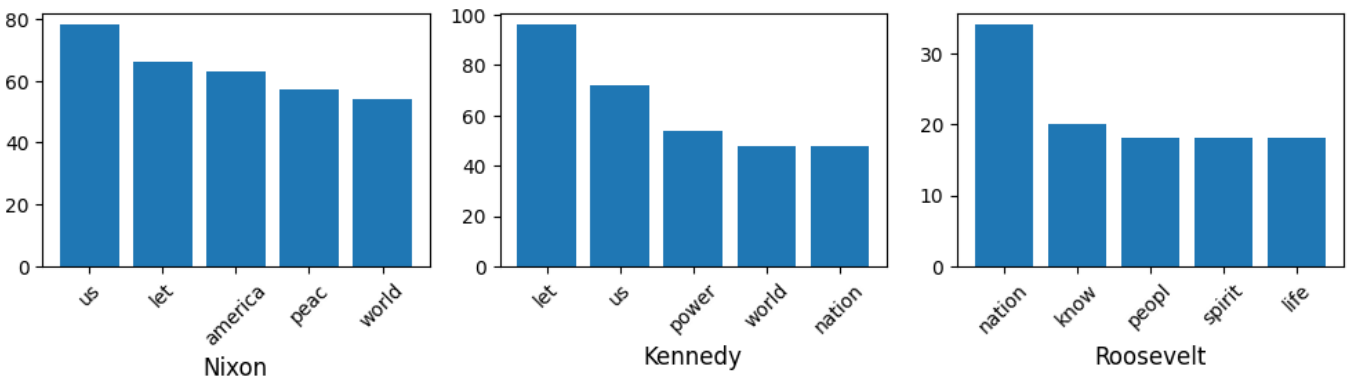
## 2.2.3 Stemming

Stemming was done using PorterStemmer().

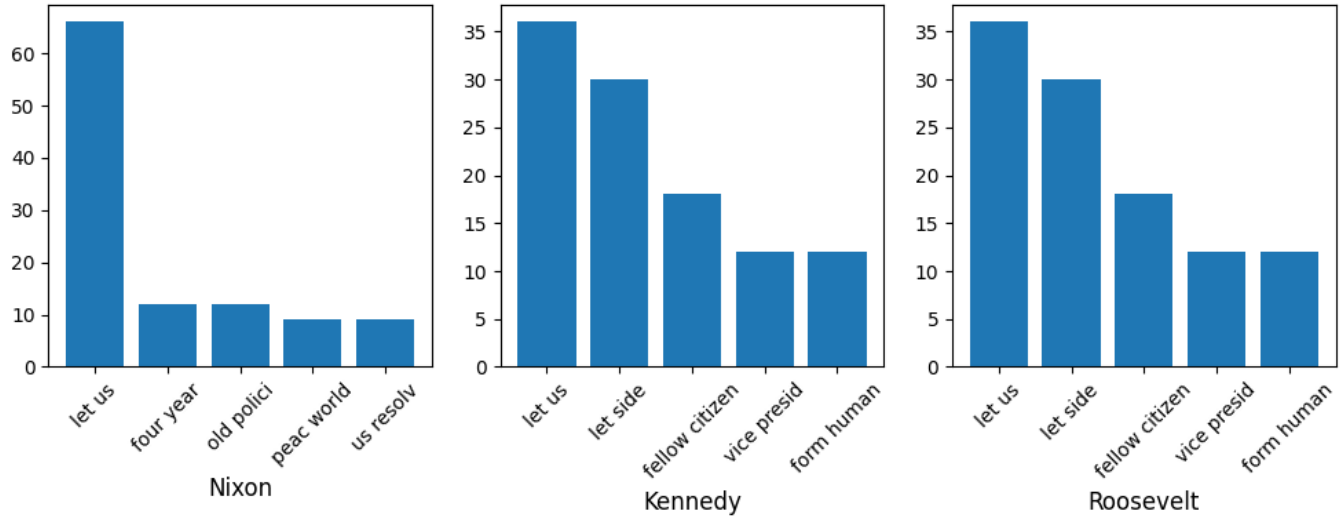
```
['mr', 'vice', 'presid', 'mr', 'speaker', 'mr', 'chief', 'justic', 'senat', 'cook', 'mr', 'eisenhow', 'fellow', 'citizen',
'great', 'good', 'countri', 'share', 'togeth', 'met', 'four', 'year', 'ago', 'america', 'bleak', 'spirit', 'depress', 'pros
pect', 'seemingli', 'endless', 'war', 'abroad', 'destruct', 'conflict', 'home', 'meet', 'today', 'stand', 'threshold', 'ne
w', 'era', 'peac', 'world', 'central', 'question', 'us', 'shall', 'use', 'peac', 'let', 'us', 'resolv', 'era', 'enter', 'po
stwar', 'period', 'often', 'time', 'retreat', 'isol', 'lead', 'stagnat', 'home', 'invit', 'new', 'danger', 'abroad', 'let',
'us', 'resolv', 'becom', 'time', 'great', 'respons', 'greatli', 'born', 'renew', 'spirit', 'promis', 'america', 'enter', 't
hird', 'centuri', 'nation', 'past', 'year', 'saw', 'farreach', 'result', 'new', 'polici', 'peac', 'continu', 'revit', 'trad
it', 'friendship', 'mission', 'peke', 'moscow', 'abl', 'establish', 'base', 'new', 'durabl', 'pattern', 'relationship', 'am
ong', 'nation', 'world', 'america', 'bold', 'initi', '1972', 'long', 'rememb', 'year', 'greatest', 'progress', 'sinc', 'en
d', 'world', 'war', 'ii', 'toward', 'last', 'peac', 'world', 'peac', 'seek', 'world', 'flimsi', 'peac', 'mere', 'interlud',
'war', 'peac', 'endur', 'gener', 'come', 'import', 'understand', 'necess', 'limit', 'america', 'role', 'maintain', 'peac',
'unless', 'america', 'work', 'preserv', 'peac', 'peac', 'unless', 'america', 'work', 'preserv', 'freedom', 'freedom', 'le
t', 'us', 'clearli', 'understand', 'new', 'natur', 'america', 'role', 'result', 'new', 'polici', 'adopt', 'past', 'four', '
year', 'shall', 'respect', 'treati', 'commit', 'shall', 'support', 'vigor', 'principl', 'countri', 'right', 'impos', 'rul
```

2.2.3 - After Stemming

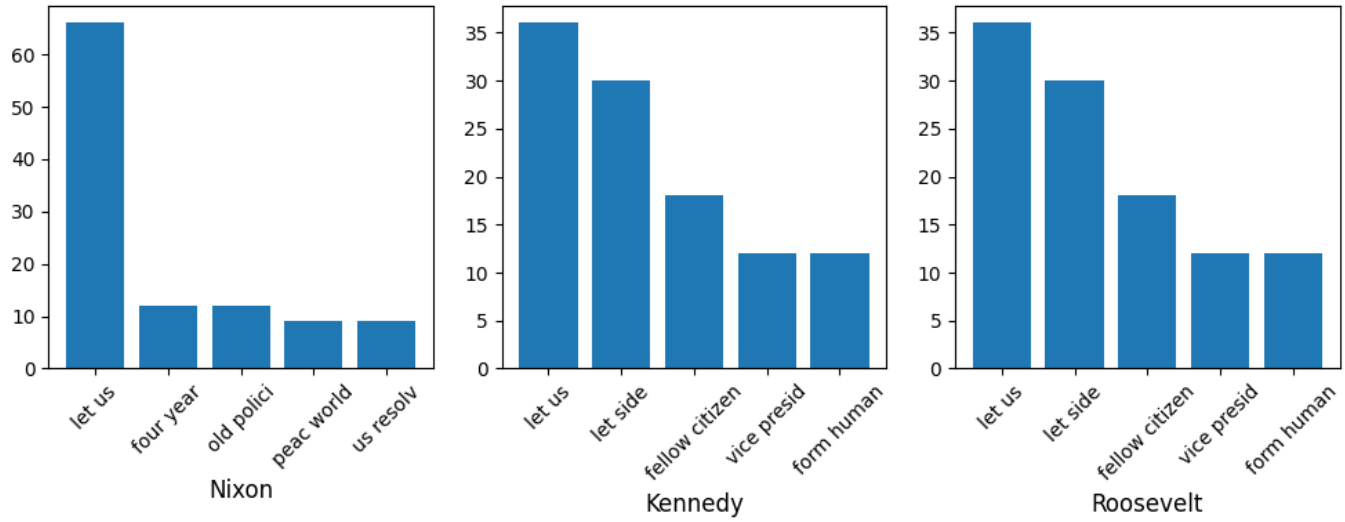
# 2.3 Frequency Distribution




2.3.1 - Frequency distribution of Monograms

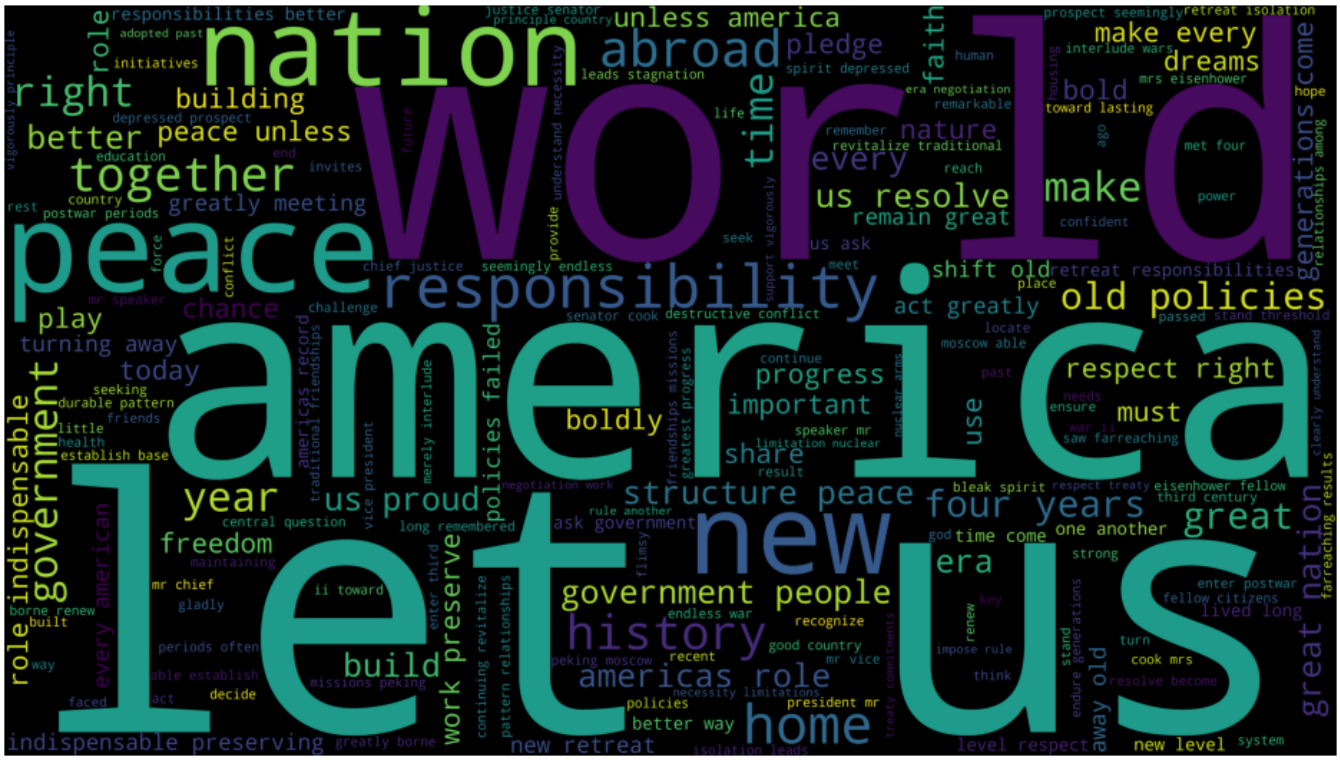



2.3.2 - Frequency distribution of Bigrams

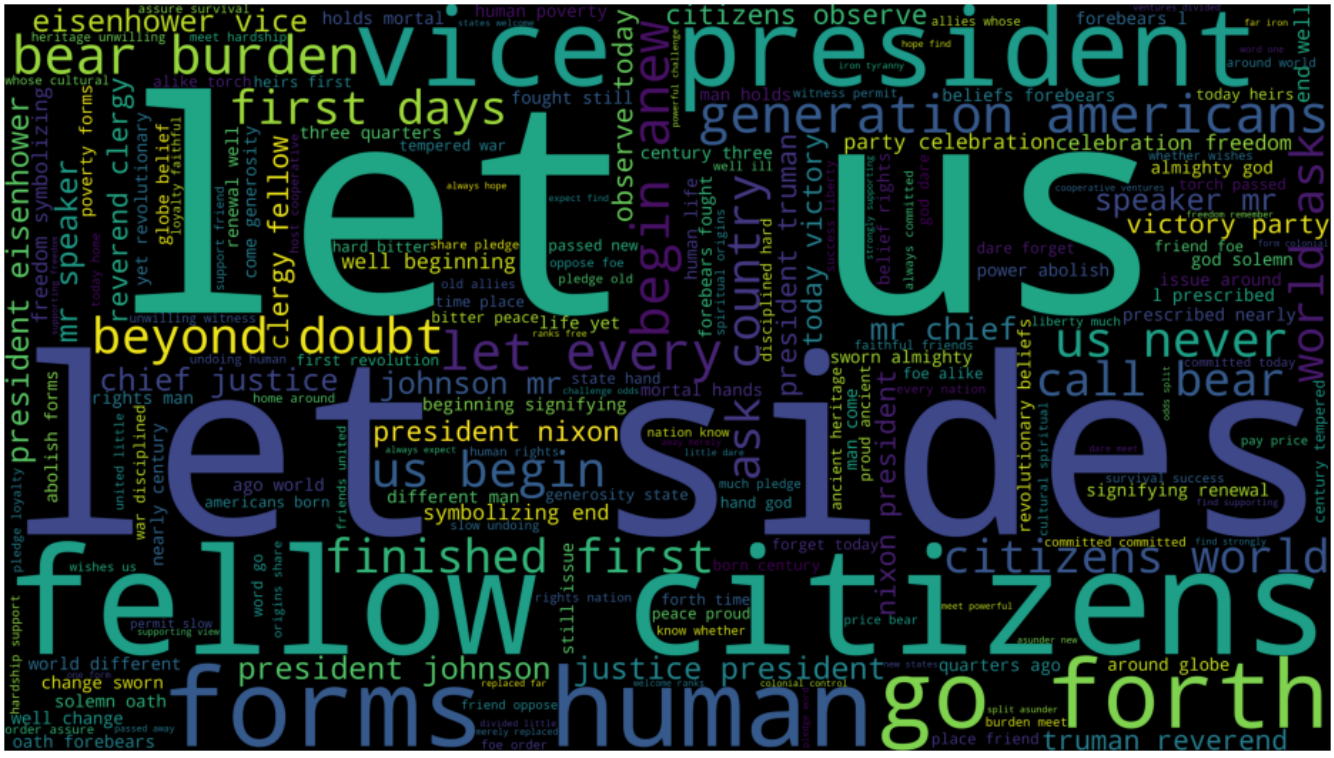



2.3.3 - Frequency distribution of Trigrams

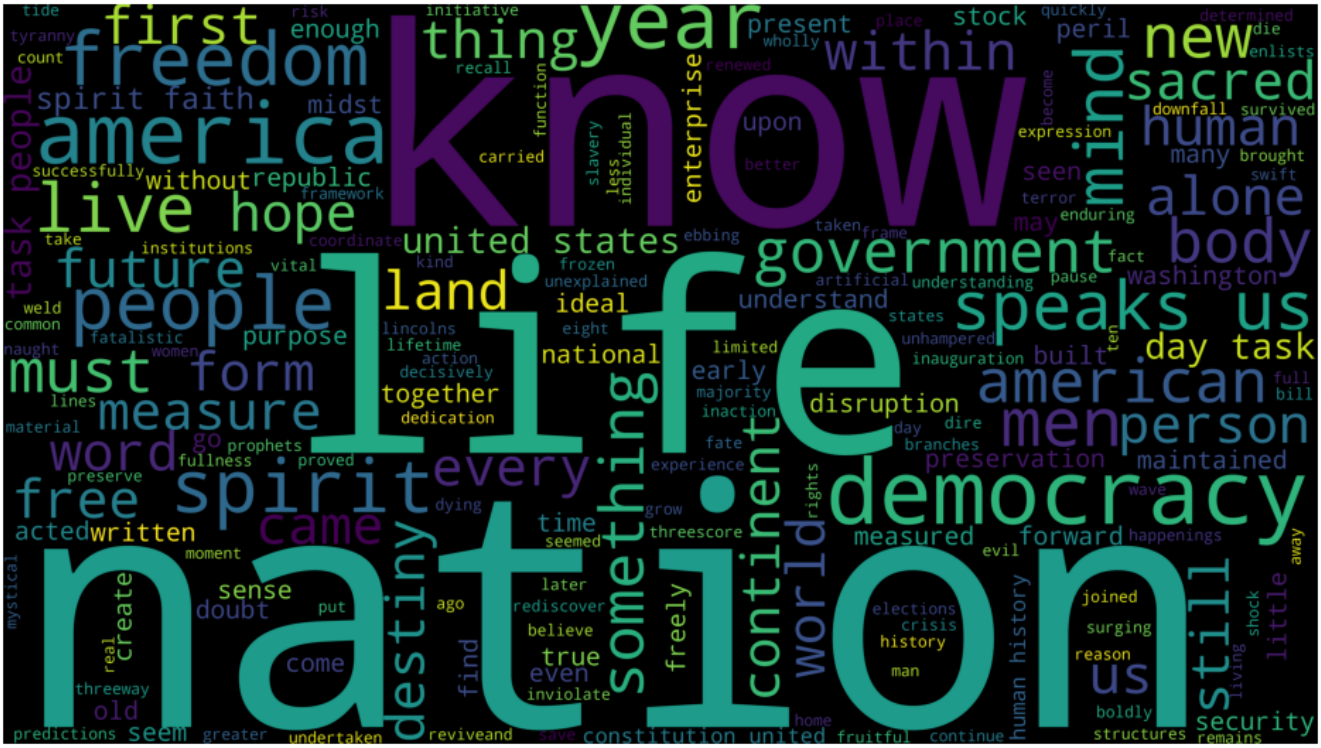
## 2.4 Wordclouds




2.4.1 - Nixon Wordcloud




2.4.2 - Kennedy Wordcloud




2.4.3 - Roosevelt Wordcloud