

# Election Results Prediction

By  
Mercy Jhansi Bai





# Data Types

## Independent Features

- 1 continuous numerical feature - age
- 1 nominal categorical feature - gender
- 6 ordinal categorical features - economic.cond.national, economic.cond.household, Blair, Hague, Europe, political.knowledge

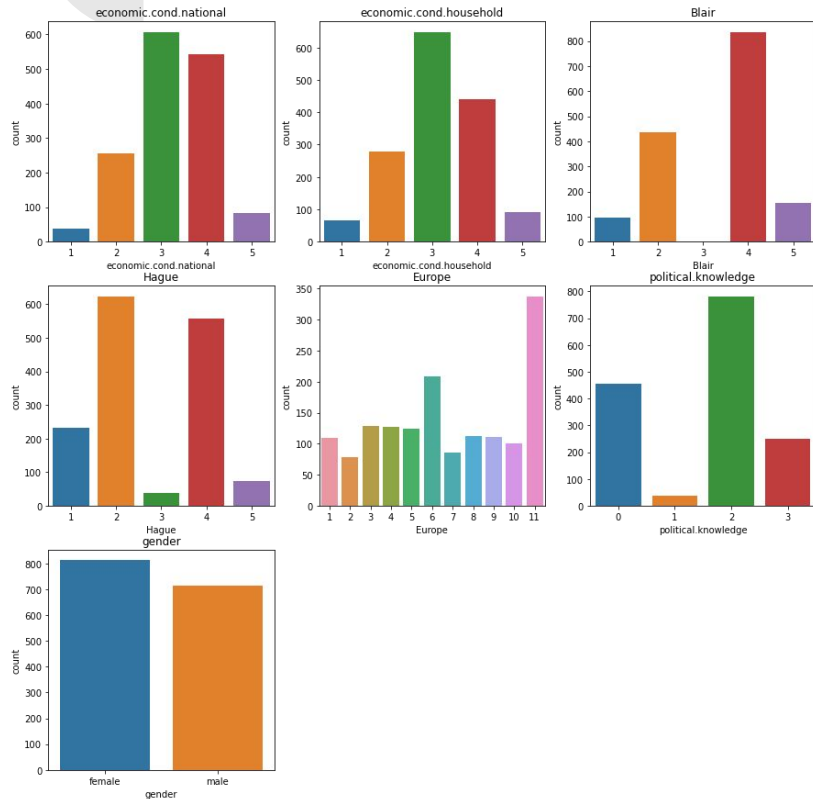
## Dependent Feature

- nominal categorical

Shape: (1525, 9)

	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

# Detecting Outliers using Countplot



**economic.cond.national,  
economic.cond.household**

- 1 & 5 are outliers as they have very low percentage. We can replace 1 with 2 and 5 with 4.

**Blair**

- 1 & 3 are outliers. We can replace 1 with 2 and 3 with 4.

**Hague**

- 3 & 5 are outliers. We can replace 5 with 4 and 3 with 2.

**Political.knowledge**

- 1 is a outlier. We can replace 1 with 2.

# Distribution & Box plot of Numerical Features



## Outliers Percentage

economic.cond.national - 7.80%

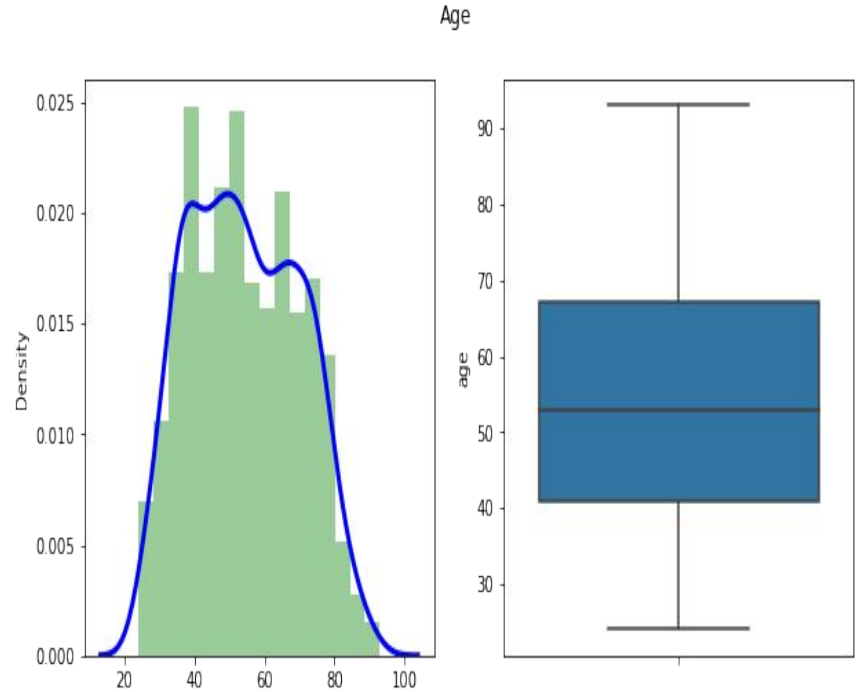
economic.cond.household - 7.74%

Blair - 4.92%

Hague - 5.11%

political.knowledge - 1.77%

**27.3%** of Outliers are detected in the dataset. We are not handling them as Decision Tree and Random Forest are not sensitive to outliers.



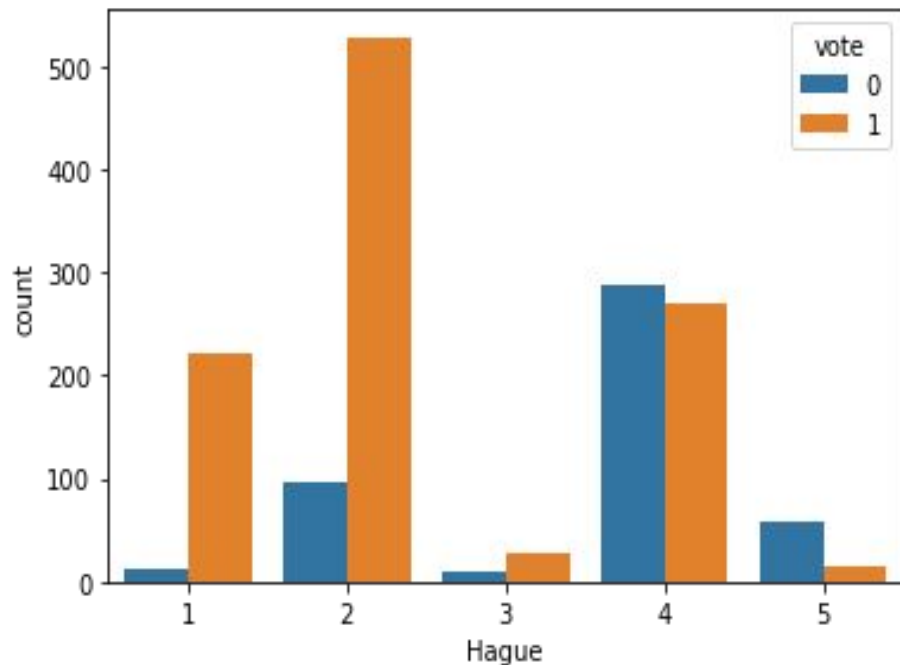


# Descriptive Statistics

- The assessment of current national economic condition is either 3 or 4 for 75% of the data.
- The assessment of current national household condition is either 3 or 4 for 70% of the data.
- The assessment of labour leader is either 2 or 4 for around 80% of the data.
- The assessment of conservative leader is either 2 or 4 for around 75% of the data.
- Around 22% of the data is very eurosceptic with a measure of 11 on a 1 to 11 scale.
- The knowledge of the parties positions for around 50% of the data is 2 and is 0 for around 30% of the data.
- We have almost balanced records for both male and female.
- Age ranges from 24 to 93 with a mean of 54 and a standard deviation of 16. The mean and median are close, hence the distribution may not be skewed.



# Hague Feature Importance

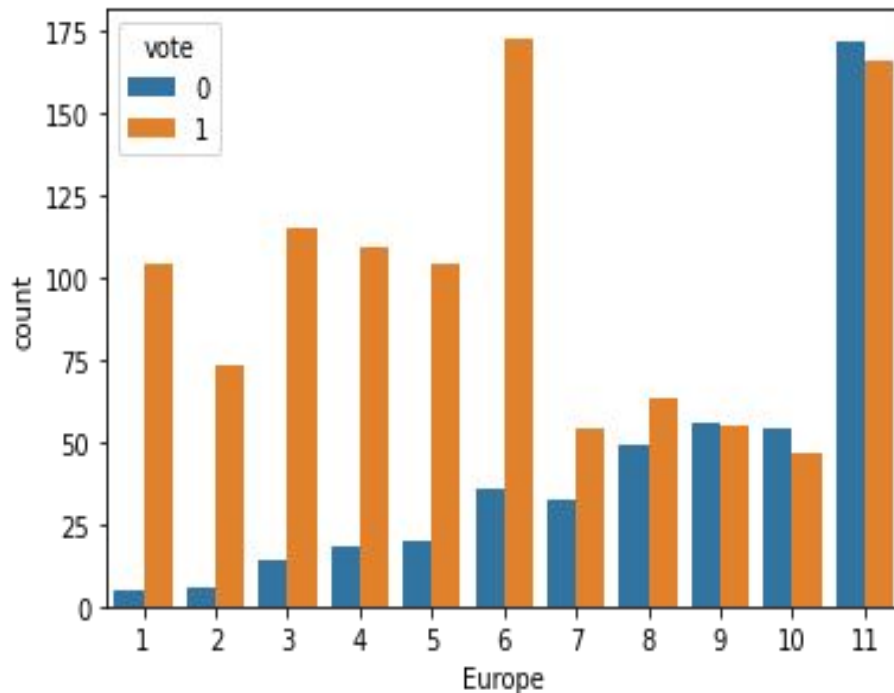


Except 4, all others count values differ heavily for Labour and Conservative. Hence we could say Hague is a good feature for building a model to classify the target variable.



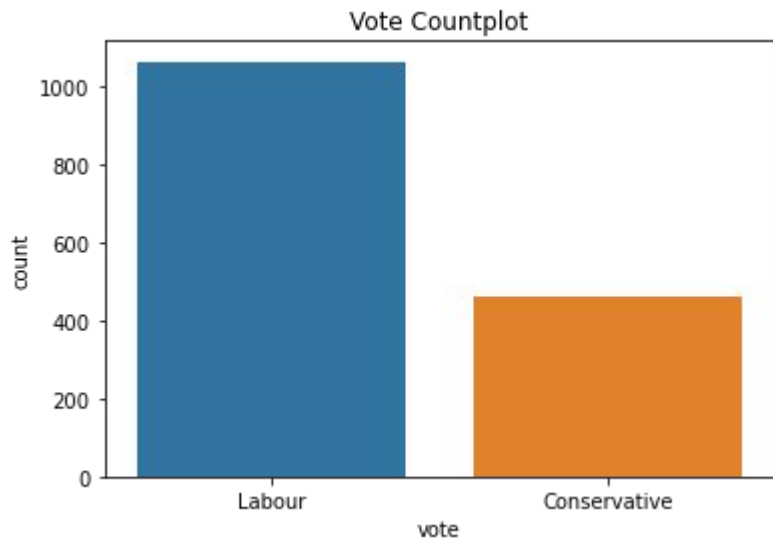
# Europe Feature Importance

There are heavy differences in the bar length for Labour and Conservative from 1 to 8. Hence we could say Europe is a good feature for building the model.





# Data Pre-processing Steps

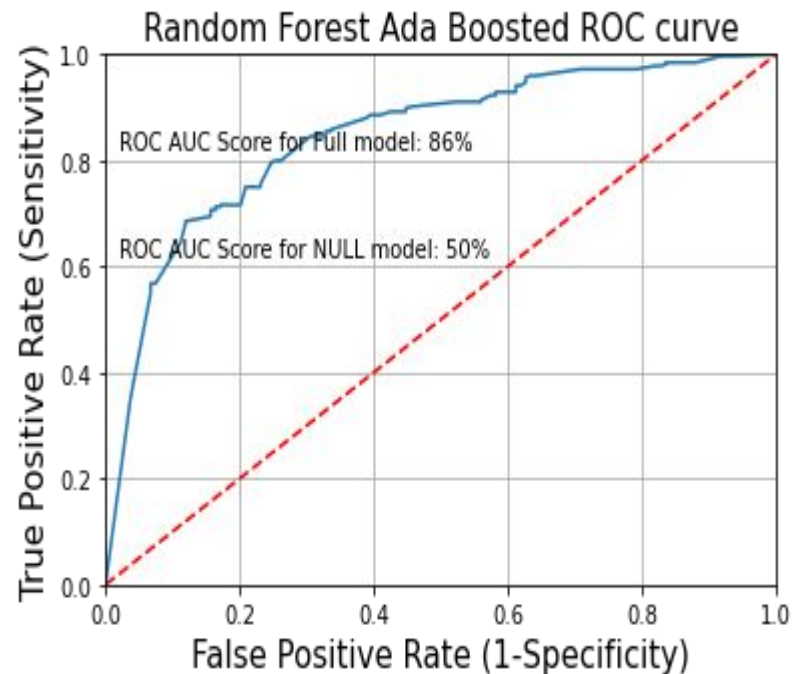


- Balanced dataset (70-30%)
  - Performance Metric: ROC AUC Score
- Encoding Techniques
  - Vote & Gender: LabelEncoder
- Scaling is not performed as scaling is not necessary for Decision Tree & Ensemble Algorithms
- Train-Test Split Ratio: 70:30



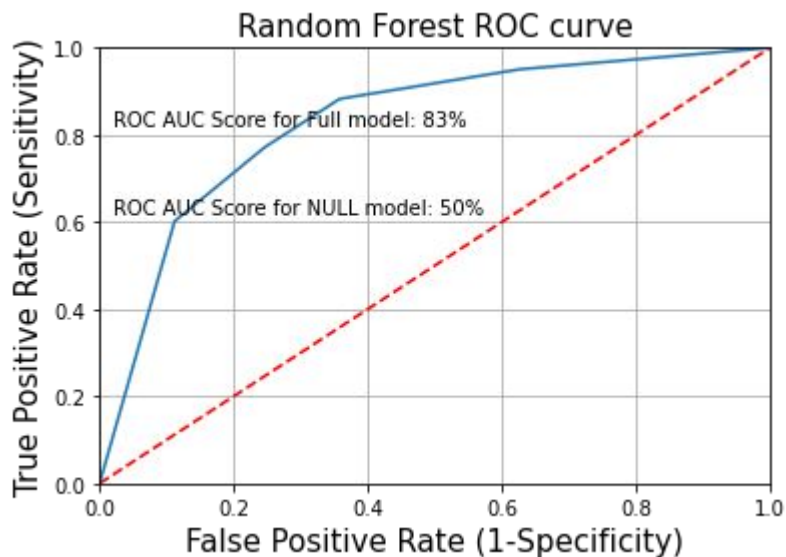
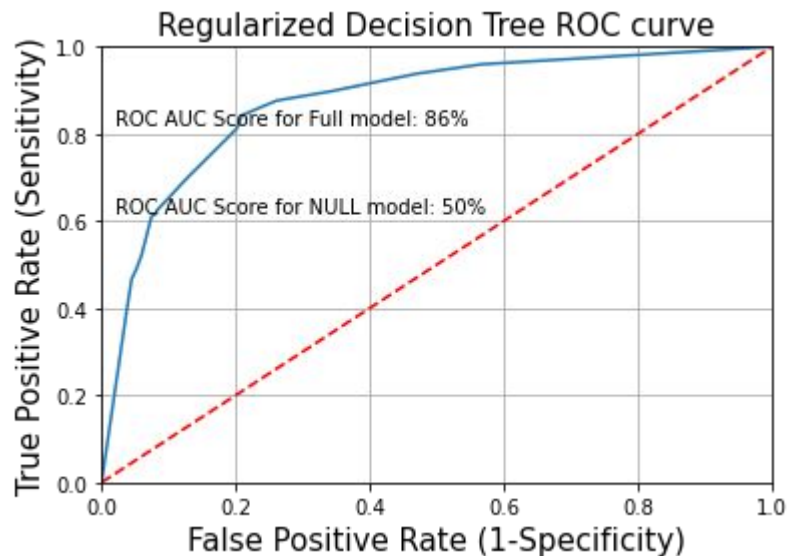
# ML Models Used

- **Fully Grown Decision Tree**
  - Avg ROC AUC Score: **71.7%**
- **Regularized Decision Tree**
  - Best Params: Criterion - Gini & Max Depth - 4
  - Avg ROC AUC Score: **86.3%**
- **Random Forest**
  - Minimum Variance Error: 0.92
  - Best Params: NEstimator - 4, Criterion - Gini
  - Avg ROC AUC Score: **83.2%**
- **Random Forest Ada Boosted**
  - Minimum Bias Error: 14.33%
  - Best Params: NEstimator - 3
  - Avg ROC AUC Score: **85.7%**





# ROC Curves





# Classification Report - RF Ada Boosted

	precision	recall	f1-score	support
0	0.66	0.66	0.66	134
1	0.86	0.86	0.86	324
accuracy			0.80	458
macro avg	0.76	0.76	0.76	458
weighted avg	0.80	0.80	0.80	458

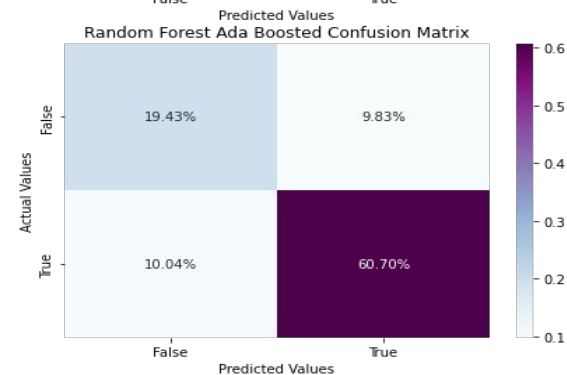
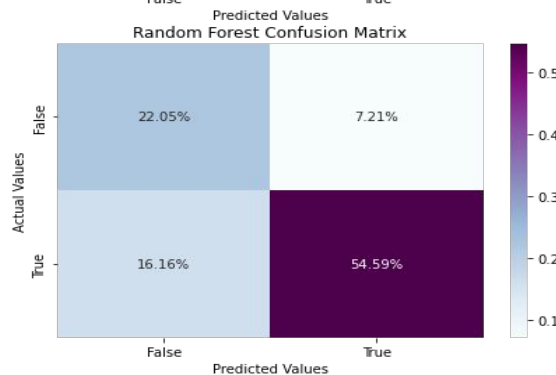
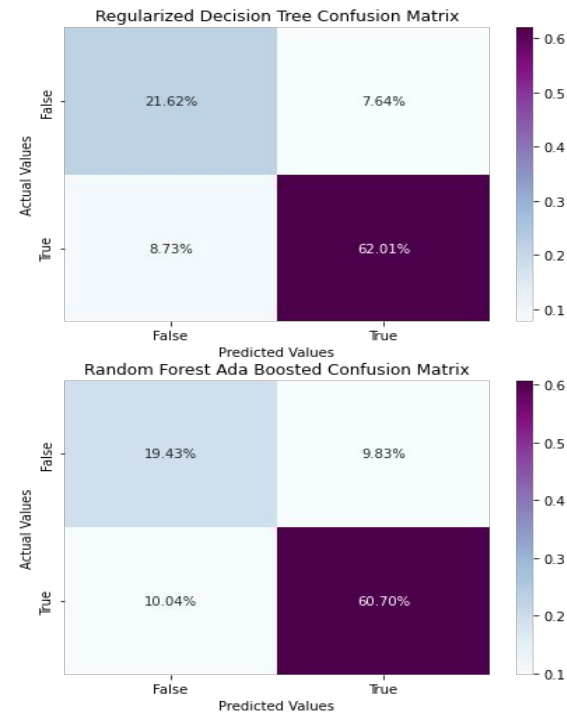
- The weighted average F1 score is 80% for Random Forest Ada Boosted, which is a good score. Hence we could say the model built is not underfitting.
- The precision and recall scores for Labour class are higher than Conservative Class by 20%.



# Confusion Matrices

## Random Forest Ada Boosted

- 60.7% of the data is correctly predicted as Labour
- 19.43% of data is correctly predicted as Conservative
- 10.04% of data is wrongly predicted as Conservative
- 9.83% of data is wrongly predicted as Labour.





# Inference

- Though the Regularized Decision Tree has the highest average score of 86.3%, the variance error is higher for it.
- Hence we have used Random Forest to reduce the variance error to .99% from 3.26%, for which we had to sacrifice on the bias error little bit of around 3%.
- Still to compensate this sacrifice in the bias error, we decided to Boost the Random Forest model using the Ada Boost Classifier, and we successfully achieved almost a similar Average Score of Regularized Decision Tree upto 85.7%.
- Hence we were able to successfully build a Random Forest Ada Boosted Machine Learning model with 85.67% Average ROC AUC Score and a variance error of .99% in predicting the overall win as Labour with 71% of votes.