Problem 1: Vote Prediction

You are hired by one of the leading news channels CNBE who wants to analyse recent elections. This survey was conducted on 1525 voters with 9 variables. You must build a model, to predict which party a voter will vote for based on the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

**Data Inferences**

1. Features Vote and Gender are of Object Type and Categorical in nature.
2. Age feature provides information about the voter’s age.
3. National and Household economy are provided in the features “economic.cond.national” and “economic.cond.household”.
4. Blair feature will give information about the Voter’s assessment of the Labour Leader, rating will be between 1 and 5, latter being better.
5. Hague feature will give information about the Voter’s assessment of the Conservative Leader, rating will be between 1 and 5, latter being better.
6. Europe feature will explain the sentiment of Eurosceptic among the voters, range beings at 1 and ends at 11, where in the latter is maximum Eurosceptic.
7. Feature “political.knowledge” gives information about the knowledge of Parties on the integration of US policies with Europe, range starts from 0 to 3, where in the latter is of the best knowledge.

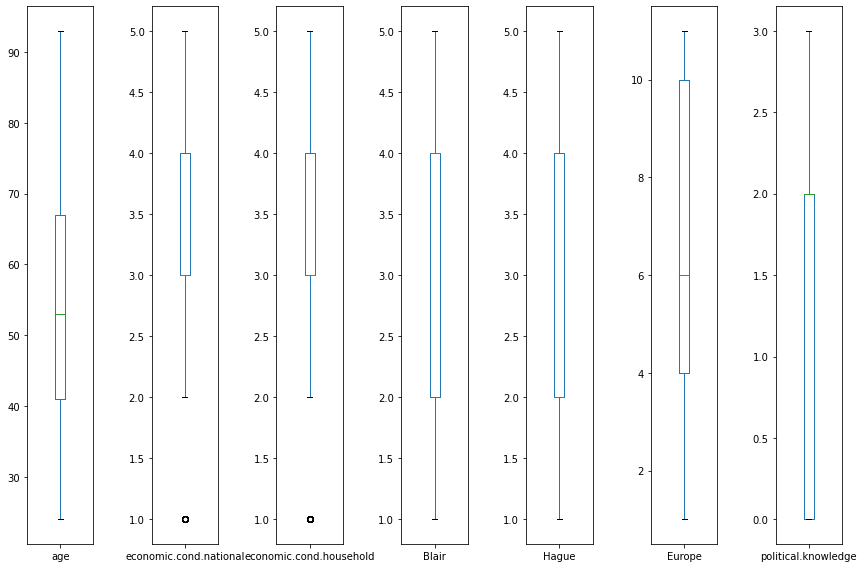
**Data Shape**

1. Data set contains 9 Features and 1525 Voter details.

**Data Validation**

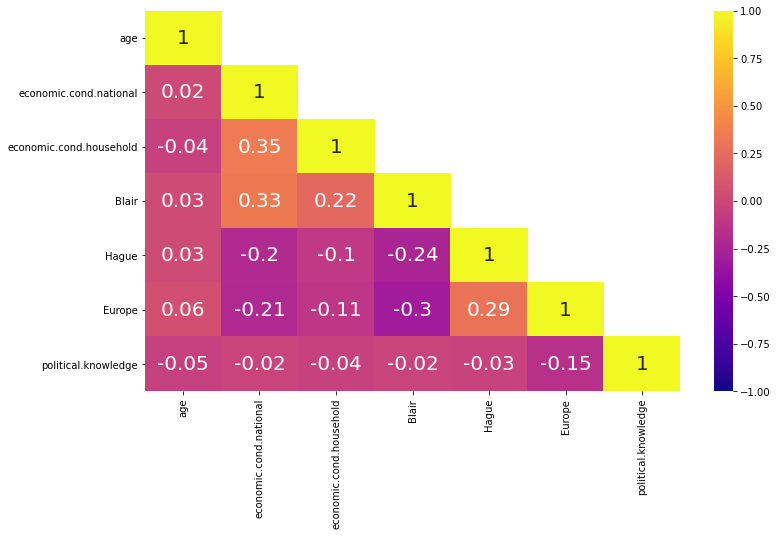
1. No nulls are seen in the data set.
2. Total duplicates found are 8 in the data set, will be dropping them as their count is insignificant, considering data set size.
3. Vote feature has two Categories “Labour” and “Conservative” and has no anomalies.
4. Gender feature has two Categories “Female” and “Male” and has no anomalies.

**Outliers**



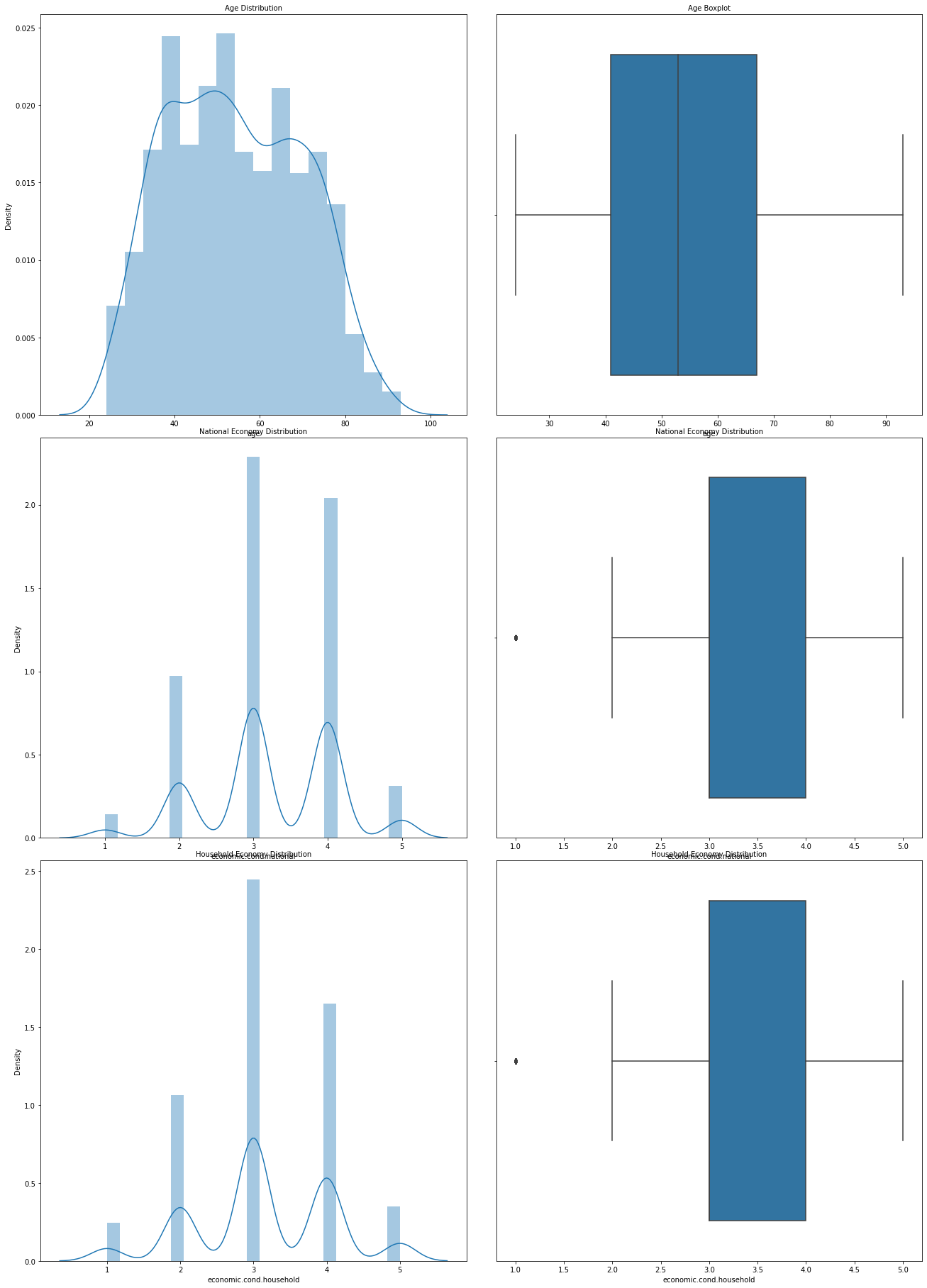
1. National and Household economy features have very insignificant Q1 outliers.
2. Other columns have no outliers.

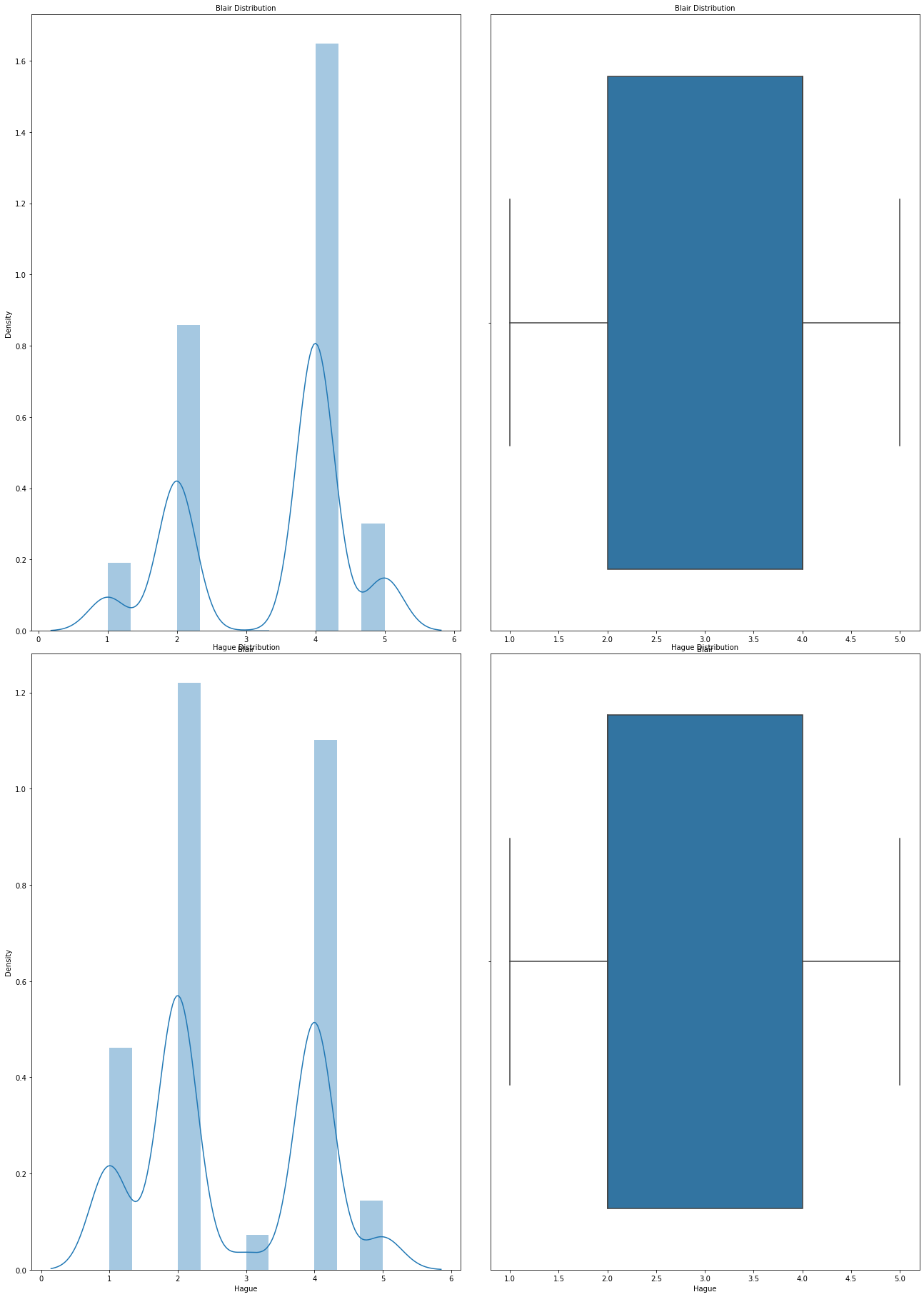
**Correlation Analysis**

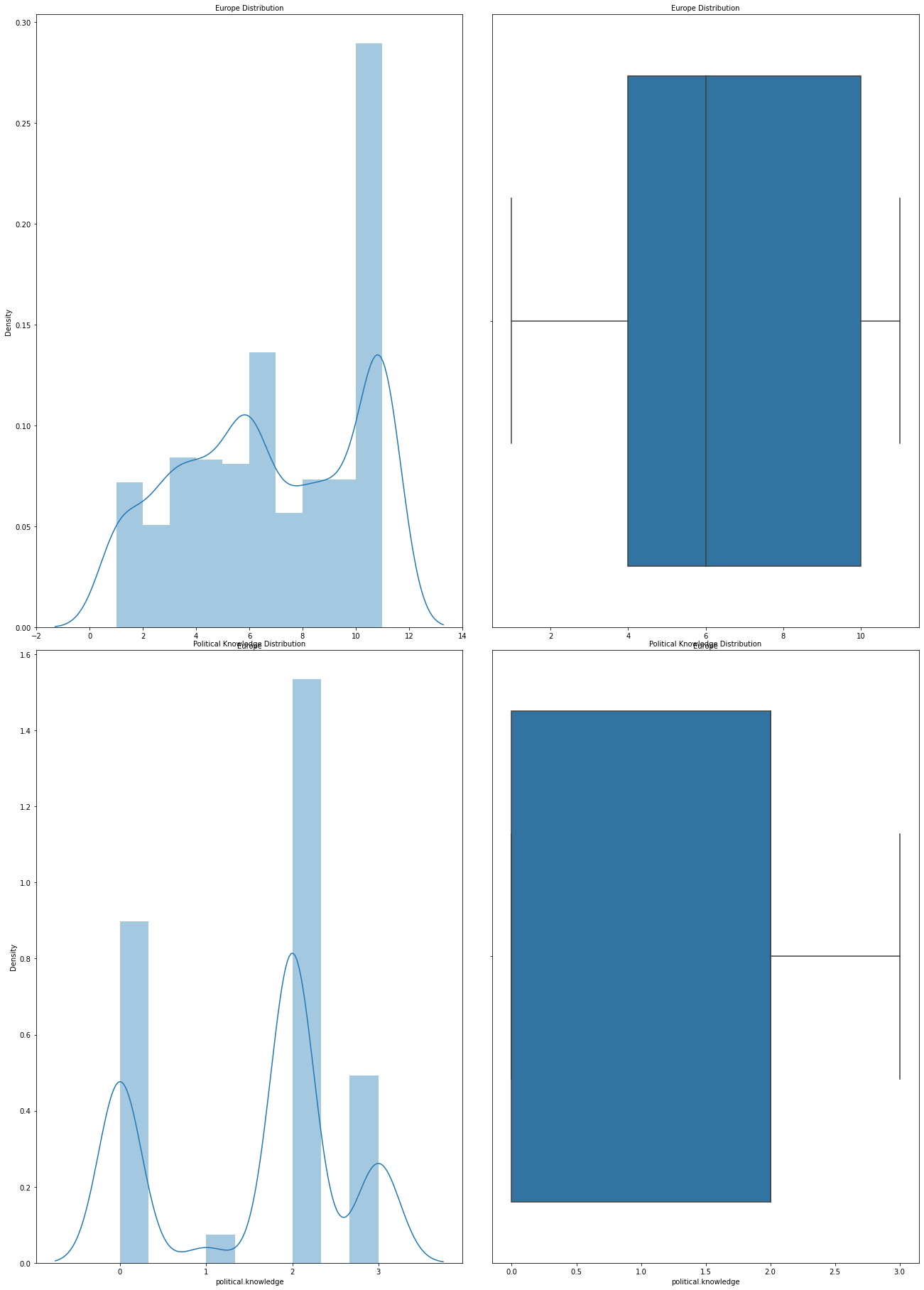


1. Age feature has no correlation with any other features of the dataset, it means, age has no effect in the voter’s understanding on the other features of the dataset.
2. National Economy has a weak-positive-correlation with household economy and Blair-Labour Leader’s rating.
3. National Economy has a weak-negative-correlation with Eurosceptic sentiment. Other features will not have any affect.
4. Labour Leader’s rating-Blair has a negative correlation with Conservative Leader’s rating-Hague and vice versa.
5. Eurosceptic rating has a negative correlation with the political knowledge of the voter.

**Uni-Variate Analysis**







1. Age feature has a slight right skewed distribution, younger age group has significant effect in voting.
2. National Economy distribution is a normal distribution with evenly spaced classes and frequency of class3 is higher than others.
3. Household Economy distribution is a normal distribution with evenly spaced classes and frequency of class3 is higher than others.
4. Blair and Hague distribution are left skewed and right skewed distributions with Classes 2 and 4 have maximum count of voters.
5. Europe distribution is a left skewed with peak count at the right. Median is however on the left skew.
6. Political knowledge is a left skewed distribution with maximum count at the category 2.

**Vote vs Other Features**

Chart, box and whisker chart

Description automatically generated

Conservative party voters have a higher age in comparison to the age of Labour party voters. He min and max age is alike in both parties.

Chart, box and whisker chart

Description automatically generated

National economic condition of the voters, for the Labour party is higher in comparison to the Conservative Party.

Chart, box and whisker chart

Description automatically generated

Household economic condition of the voters, for the Labour party is higher in comparison to the Conservative Party.

Chart, box and whisker chart

Description automatically generated

Labour party voters have a similar rating for their leader Blair. Conservative party voters have rated Blair from 2 to 4.

Chart, box and whisker chart

Description automatically generated

Conservative party voters have a similar rating for their leader Hague. Conservative party voters have rated Hague from 2 to 4.

Chart, box and whisker chart

Description automatically generated

Conservative party voters are majorly not supporting the Eurosceptic sentiment in comparison to the Labour party voters.

Chart, box and whisker chart

Description automatically generated

Conservative Party voters have a high political knowledge in comparison with the Labour Party voters.

**Categorical column analysis**

Chart, bar chart

Description automatically generated

Male and Female percentages are 50 and 50 between the Conservative and Labour Party Voters.

**Pair plot**

**A picture containing shoji, crossword puzzle, window

Description automatically generated**

There is no correlation between any features of the data set.

**Data Encoding**

Table

Description automatically generated

Features Labour and Gender were an Object Type. Converted them to Numeric using pandas get dummies.

**Data Scaling**

Table

Description automatically generated

Age column was Scaled to set all the features into single digit numeric scale. With a higher scaled feature, distance based algorithms like, Random Forest and Decision Tree will have questionable results.

**Train and Test split**

 **and** 

Train and Test split is at 70 and 30 percent. Both the sets have 70 and 30 percentage of voters who vote of Labour Party and other vote for Conservative Party.

**Logistic Regression**

LogisticRegression(max\_iter=10000, n\_jobs=2, penalty='none', random\_state=123,solver='newton-cg', verbose=True)

Logistic Regression model is created to run for an 10000 iterations and penalty is set as none, random\_state is set to 123 and solver that is being using is newton\_cg. No penalization will be applied for insignificant features as the Penalty is set to None. We have a smaller data set of 1500 rows, so newton\_cg would give affirmative results.

**Test Prediction probabilities**

Graphical user interface, text, application

Description automatically generated

Top of the prediction probabilities after applying the model, anything above 0.5 by default is 1.

**Accuracy Score – Training data : 0.83**

Training model is 83 percent accurate in predicting the vote result.

**ROC Curve – Training A picture containing graphical user interface

Description automatically generated**

With AUC at 88%, there is 88 percent chance the model will be able to distinguish between Labour Party voter and Conservative Party voter.

**Accuracy Score – Testing Data: 0.85**

Testing model accuracy is at 85 percent. It will be 85 percent accurate in predicting the result. Considering the Training Model accuracy, test accuracy is slightly high, mentioning that the model is **slightly Overfit**.

**ROC Curve – Testing Data**

**A picture containing graphical user interface

Description automatically generated**

AUC stands at 91%, test model will be is 91 percent chance the model will be able to distinguish between Labour Party voter and Conservative Party voter. However, the test model is overfit in comparison to the Training model, there will be high chances of wrong predictions.

**Confusion Matrix and Classification Report – Training data**

**Chart, treemap chart

Description automatically generated**

Table

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 88%
2. Accuracy of predicting the right result is at 83%.
3. Recall score and Precision are at 90% and 86%, model’s ability to predict minority and majority class looks positive.

**Confusion Matrix and Classification Report – Testing dataChart, treemap chart

Description automatically generated**

Table

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 86%.
3. Recall score and Precision are at 93% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is overfit, we may have a wrong minority class prediction.**
5. **Tuning: Using all the algorithms for Solvers and Penalty, I was not able to push the Accuracy Scores of both training and testing data.**

**GridSearchCV for Logistic Regression**

Text, letter

Description automatically generated

GridSearchCV model is set with hyper parameters; Cross Validation of 3, LogisticRegression will be the model, number of iterations to run will be 10K, l1\_ratio range will be from 0.25-0.75, penalty and solvers all algorithms were specified, very low tolerance of 0.0001 and 0.00005 is selected to have accurate model.

**Best Model**



Best model selected by the GridSearchCv has l1\_ratio of 0.25, solver and penalty will be Saga and L1.

**Test probabilities for the best model**

Graphical user interface, text, application, chat or text message

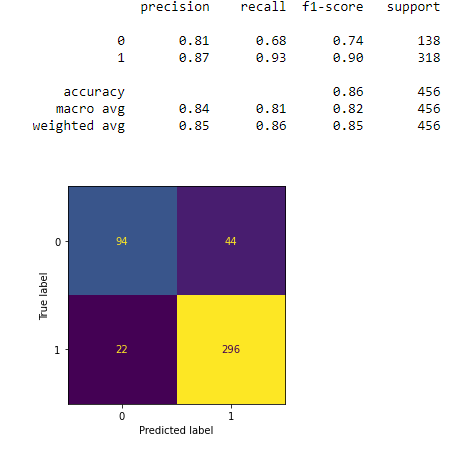
Description automatically generated

**Confusion matrix and Classification report – Training data**

Chart, treemap chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 88%
2. Accuracy of predicting the right result is at 83%.
3. Recall score and Precision are at 90% and 86%, model’s ability to predict minority and majority class looks positive.

**Confusion matrix and Classification report – Testing data**

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 86%.
3. Recall score and Precision are at 93% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is overfit, we may have a wrong minority class prediction.**
5. **Tuning: Using all the algorithms for Solvers and Penalty, I was not able to push the Accuracy Scores of both training and testing data. The scores of GridSearchCV match with the Logistic regression model.**

**Linear Discriminant Analysis**

Linear Discriminant Analysis will help us to identify the important features of vote Prediction. Default parameters were used for LinearDiscriminantAnalaysis algorithm.

**Training Scores: Accuracy: 83% and Recall: 88.7%**

**Testing Scores: Accuracy: 85% and Recall: 92%**

From the above scores the model is overfit and chance of having minority class wrong predictions is possible.

**Confusion matrix and Classification report – Training data**

Chart, treemap chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 87%
2. Accuracy of predicting the right result is at 82%.
3. Recall score and Precision are at 89% and 86%, model’s ability to predict minority and majority class looks positive.

**Confusion matrix and Classification report – Testing data**Chart, treemap chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 85%.
3. Recall score and Precision are at 92% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is overfit, we may have a wrong minority class prediction.**

**AUC and ROC curve – Training and Testing data**A picture containing graphical user interface

Description automatically generated

Test AUC is 91% and Train AUC 88% chance of predicting the right result. Test model is overfit, and it has a chance to predict incorrect results.

**Important Features**

Chart, waterfall chart

Description automatically generated

Affiliation towards Blair and Hague are the important features, followed with political knowledge and national economic status, we can drop the household economy and gender columns to make better prediction model.

**GridSearchCV – Linear Discriminant Analysis**

Text

Description automatically generated

GridSearchCV provided with best params model to work with all solver algorithms with high tolerance. Scoring is selected as F1.

**Classification Report - Training**

Table

Description automatically generated

Chart, treemap chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 87%
2. Accuracy of predicting the right result is at 82%.
3. Recall score and Precision are at 89% and 86%, model’s ability to predict minority and majority class looks positive.

**Classification Report - Testing**

Table

Description automatically generated

Chart, treemap chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 85%.
3. Recall score and Precision are at 92% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is overfit, we may have a wrong minority class prediction.**
5. **In comparison to LDA, classification report values match with GridSearchCV as well.**

**Naïve Bayes**

**Training data**

Table

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 87%
2. Accuracy of predicting the right result is at 82%.
3. Recall score and Precision are at 87% and 87%, model’s ability to predict minority and majority class looks positive.

**Testing data**

Chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 86%.
3. Recall score and Precision are at 92% and 88%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is overfit, we may have a wrong minority class prediction.**
5. Test AUC is 91% and Train AUC 88% chance of predicting the right result. Test model is overfit, and it has a chance to predict incorrect results.

**GridSearchCV- Naïve Bayes**

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 87%
2. Accuracy of predicting the right result is at 82%.
3. Recall score and Precision are at 87% and 87%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 86%.
3. Recall score and Precision are at 92% and 88%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is overfit, we may have a wrong minority class prediction.**

**KNN**

**KNN-5 Training data**

Table

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 87%
2. Accuracy of predicting the right result is at 82%.
3. Recall score and Precision are at 87% and 87%, model’s ability to predict minority and majority class looks positive.

**KNN-5 Testing data**

Chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 86%.
3. Recall score and Precision are at 92% and 88%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is overfit, we may have a wrong minority class prediction.**
5. Test AUC is 91% and Train AUC 88% chance of predicting the right result. Test model is overfit, and it has a chance to predict incorrect results.

**KNN-7 Training data**

Table

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 89%
2. Accuracy of predicting the right result is at 89%.
3. Recall score and Precision are at 91% and 88%, model’s ability to predict minority and majority class looks positive.

**KNN-7 Testing data**

Chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 85% which is same as the Train model score.
3. Recall score and Precision are at 91% and 88%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is best fit.**
5. Test AUC is 91% and Train AUC 88% chance of predicting the right result. Test model is slightly overfit, and it has a chance to predict incorrect results.

**GridSearchCV-KNN**

**KNN-7**

Text, letter

Description automatically generated

Hyper parameters with uniform and distance weights are selected with 7 as K value and Cross validation as 5.

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 89%
2. Accuracy of predicting the right result is at 89%.
3. Recall score and Precision are at 91% and 88%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 85% which is same as the Train model score.
3. Recall score and Precision are at 91% and 88%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is best fit.**
5. Test AUC is 91% and Train AUC 88% chance of predicting the right result. Test model is slightly overfit, and it has a chance to predict incorrect results.

**Decision Tree**

**Training Metrics**

Chart, treemap chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 100%
2. Accuracy of predicting the right result is at 100%.
3. Recall score and Precision are at 100% and 100%, model’s ability to predict minority and majority class looks positive.
4. The above results are incorrect, need to be checked for best parameters in GridSearchCV.

**Testing Metrics**

Chart, treemap chart

Description automatically generated

1. F1 Score mentions the overall reliability of the model and it stands at 84%
2. Accuracy of predicting the right result is at 78% which is less than the Train model score.
3. Recall score and Precision are at 86% and 83%, model’s ability to predict minority and majority class looks positive.
4. Looking at the accuracy scores the Test model could be underfit.
5. Test AUC is 91% and Train AUC 88% chance of predicting the right result. Test model is slightly overfit, and it has a chance to predict incorrect results.

**GridSearchCV - Decision Tree**

Calendar

Description automatically generated

Chart, treemap chart

Description automatically generated

DecisionTree with max\_leaf\_nodes produced a desirable result.

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 90%
2. Accuracy of predicting the right result is at 85%.
3. Recall score and Precision are at 92% and 88%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 88%
2. Accuracy of predicting the right result is at 83% which is same as the Train model score.
3. Recall score and Precision are at 91% and 86%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is slightly underfit.**

**Random Forest**Chart, treemap chart

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 100%
2. Accuracy of predicting the right result is at 100%.
3. Recall score and Precision are at 100% and 100%, model’s ability to predict minority and majority class looks positive.
4. The above results are incorrect, need to be checked for best parameters in GridSearchCV..

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 89%
2. Accuracy of predicting the right result is at 84%.
3. Recall score and Precision are at 91% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is underfit, chance of having wrong predictions for Majority class is possible.**

**GridSearchCV-Random Forest**

A picture containing text

Description automatically generated

Chart

Description automatically generated with medium confidence

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 100%
2. Accuracy of predicting the right result is at 100%.
3. Recall score and Precision are at 100% and 100%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 89%
2. Accuracy of predicting the right result is at 84%.
3. Recall score and Precision are at 91% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the Random Forest is not the right model for Vote Prediction.**

**Bagging**

**Metrics on Training data**

Chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 100%
2. Accuracy of predicting the right result is at 100%.
3. Recall score and Precision are at 100% and 100%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 89%
2. Accuracy of predicting the right result is at 84%.
3. Recall score and Precision are at 91% and 87%, model’s ability to predict minority and majority class looks positive.

**GridSearchCV-Bagging**

Calendar

Description automatically generated with low confidence

Chart

Description automatically generated

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 98%
2. Accuracy of predicting the right result is at 97%.
3. Recall score and Precision are at 100% and 96%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 88%
2. Accuracy of predicting the right result is at 83% which is same as the Train model score.
3. Recall score and Precision are at 94% and 83%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is heavily underfit and susceptible to wrong majority and minority class predictions.**

**Ada Boost**

Chart, treemap chart

Description automatically generated

Chart

Description automatically generated

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 89%
2. Accuracy of predicting the right result is at 85%.
3. Recall score and Precision are at 91% and 88%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 88%
2. Accuracy of predicting the right result is at 84% which is same as the Train model score.
3. Recall score and Precision are at 90% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is underfit and susceptible to wrong majority and minority class predictions.**

**GridSearchCV - Ada Boost**

A picture containing table

Description automatically generated

Chart, treemap chart

Description automatically generated

**Training Metrics**

1. Estimators were set at 200 by GridSearchCV to get optimal result.
2. F1 Score mentions the overall reliability of the model and it stands at 89%
3. Accuracy of predicting the right result is at 84%.
4. Recall score and Precision are at 91% and 87%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 88%
2. Accuracy of predicting the right result is at 83% which is same as the Train model score.
3. Recall score and Precision are at 90% and 86%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is slightly underfit and susceptible to wrong majority and minority class predictions.**

**Gradient Boost**

Chart, treemap chart

Description automatically generated

Chart

Description automatically generated

**Training Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 92%
2. Accuracy of predicting the right result is at 89%.
3. Recall score and Precision are at 93% and 91%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 89%
2. Accuracy of predicting the right result is at 84%.
3. Recall score and Precision are at 91% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is underfit and susceptible to wrong majority and minority class predictions.**

**GridSearchCV – Gradient Boost**

Calendar

Description automatically generated

Chart

Description automatically generated

**Training Metrics**

1. Estimators were set at 200 by GridSearchCV and max\_features is set to ‘sqrt’ to get optimal result.
2. F1 Score mentions the overall reliability of the model and it stands at 93%
3. Accuracy of predicting the right result is at 91%.
4. Recall score and Precision are at 95% and 92%, model’s ability to predict minority and majority class looks positive.

**Testing Metrics**

1. F1 Score mentions the overall reliability of the model and it stands at 89%
2. Accuracy of predicting the right result is at 84% which is same as the Train model score.
3. Recall score and Precision are at 90% and 87%, model’s ability to predict minority and majority class looks positive.
4. **By the above scores it can be said that the test model is underfit and susceptible to wrong majority and minority class predictions.**

**Comparison Between Models**

Table

Description automatically generated

**In the above list of models, KNN shows promising recall scores in test and train with slight underfit on the test. I would recommend KNN to predict the Voter for Conservative and Labour Parties.**

**Recommendations**

1. **Features Household Economy, Gender have no significance, we can remove them for better predictability of the models.**
2. **Hauge and Blair have inverse coefficients, they directly affect the Voter sentiment from Labour and Conservative Parties. Good campaigning about Euroscepticity will help Conservative party to win the election as their voters are highly Eurosceptic.**
3. **National Economy has a positive impact on the Voter Confidence, party with a promise on Voter Confidence will win the election.**
4. **Many voters from both parties are found to have less political knowledge, majority are from the Labour Party.**

**Problem 2**

**In this 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**

**Speech Analysis for President Roosvelt**

**Total sentences in President Roosvelt speech is:** 68

**Total words in President Roosvelt speech is:** 1526

**Top 3 words of president Roosvelt are**: ['national', 'day', 'inauguration']

**Without stop words:** national day inauguration since people renewed sense dedication united states washington day task people create weld together nation lincoln day task people preserve nation disruption within day task people save

**A picture containing text, newspaper

Description automatically generated**

**Words nation, life, people, know and democratic, spirit, America are more frequently used by President Roosvelt.**

**Speech Analysis for President Kennedy**

**Total characters in President Kennedy speech is:** 7618

**Total sentences in President Kennedy speech is:** 52

**Total words in President Kennedy speech is:** 1543

**Top 3 words of president Kennedy are:** ['vice', 'president', 'johnson']

**After stop words removal:** vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying

**--------------------------------------------------**

**Text

Description automatically generated**

**Words power, world, let, nation and new more frequently used by President Kennedy.**

**Speech Analysis for President Nixon**

**Total characters in President Nixon speech is:** 9991

**Total sentences in President Nixon speech is:** 68

**Total words in President Nixon speech is:** 2006

**Top 3 words of president Nixon are:** ['mr', 'vice', 'president']

**After stop words removal:** mr vice president mr speaker mr chief justice senator cook mrs eisenhower fellow citizens great good country share together met four years ago america bleak spirit depressed prospect seemingly endless

**A picture containing text, newspaper

Description automatically generated**

**Words power, world, let, peace, America, new and nation were more frequently used by President Nixon.**