* 1. **Read the dataset. Do the descriptive statistics and do null value condition check.**

Below is the data structure for the given election data set with vote as the dependent target variable with rest of them as independent variables.

# 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

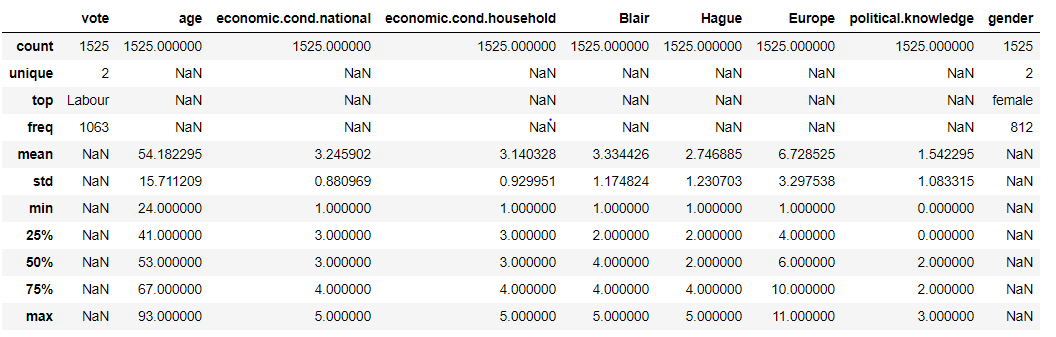
5 Hague 1525 non-null int64

6 Europe 1525 non-null int64

7 political.knowledge 1525 non-null int64

8 gender 1525 non-null object

Below is the description of the dataset across categorical and continuous variables.



It could be noticed that there isn’t much outlier considering the spread on the either side of median for continuous variables and that the target variable vote has only 2 distinct values indicating only two classes from classification perspective with Labour holding higher frequency. Also it could be noticed that gender has 2 unique values with female holding higher frequency of rows.

Below gives the exact distribution of categorical variables:

1. VOTE : 2

Conservative 462

Labour 1063

Name: vote, dtype: int64

1. GENDER : 2

male 713

female 812

Name: gender, dtype: int64

Also there are no null values present in any of the variables in the dataset as listed below:

vote

age

economic.cond.national

economic.cond.household

Blair

Hague

Europe

political.knowledge

gender

Dataset contains 1525 rows and 9 columns.

**1.2. Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts)**

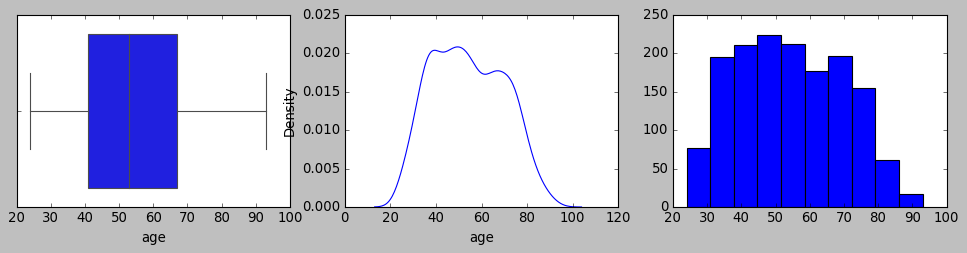
**Univariate Analysis for unscaled dataset**

1. Univariate analysis for age

Mean is 54.182295, Median is 53.000000, Mode(s) are 37.0000

Column age does not have outliers

word will result in an error or misinterpretation.

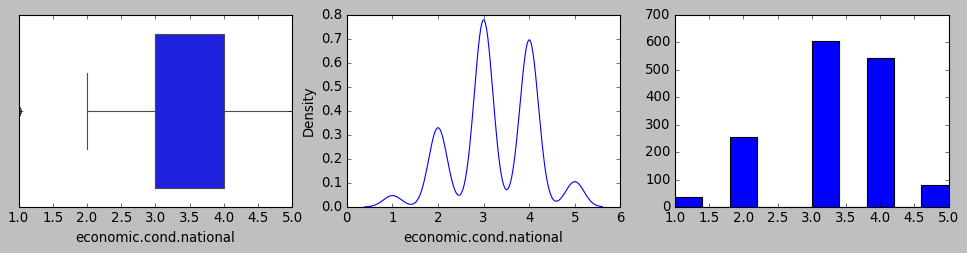


Column age is not normally distributed

2. Univariate analysis for economic.cond.national

Mean is 3.245902, Median is 3.000000, Mode(s) are 3.0000

Column economic.cond.national has outliers

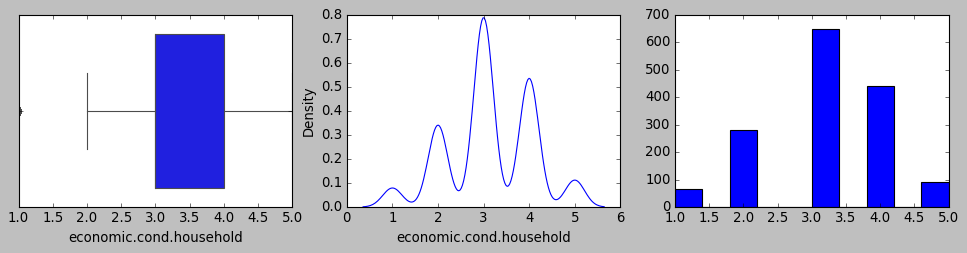


Column economic.cond.national is not normally distributed

3. Univariate analysis for economic.cond.household

Mean is 3.140328, Median is 3.000000, Mode(s) are 3.0000

Column economic.cond.household has outliers

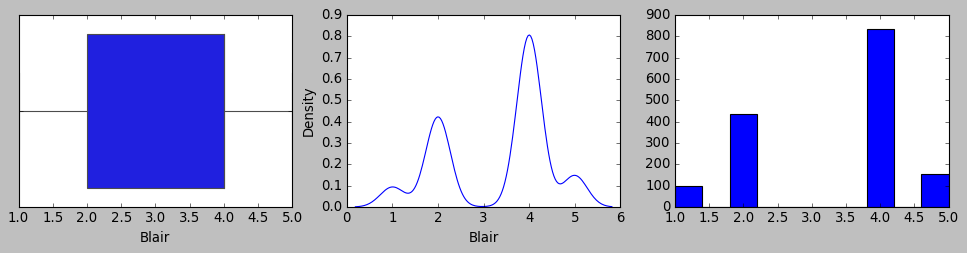


Column economic.cond.household is not normally distributed

4. Univariate analysis for Blair

Mean is 3.334426, Median is 4.000000, Mode(s) are 4.0000

Column Blair does not have outliers

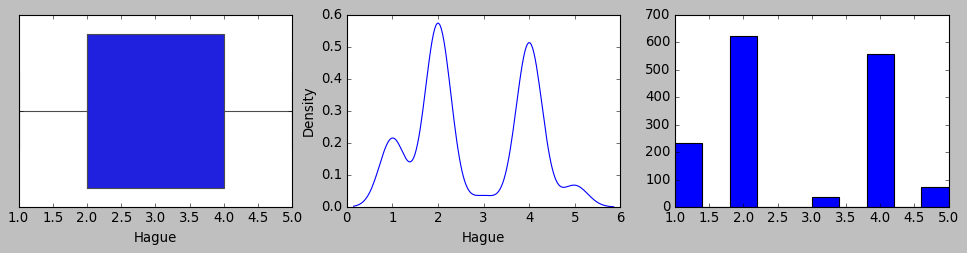


Column Blair is not normally distributed

5. Univariate analysis for Hague

Mean is 2.746885, Median is 2.000000, Mode(s) are 2.0000

Column Hague does not have outliers

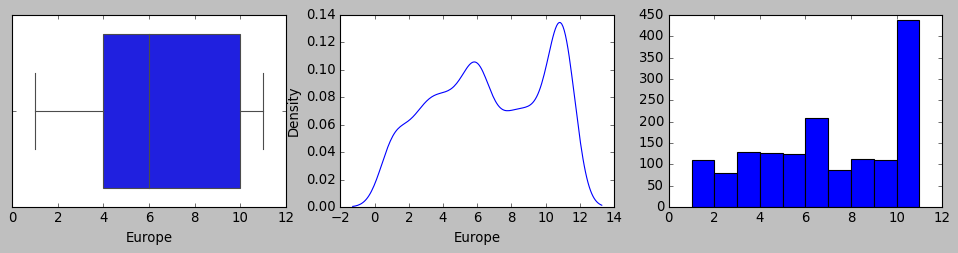


Column Hague is not normally distributed

6. Univariate analysis for Europe

Mean is 6.728525, Median is 6.000000, Mode(s) are 11.0000

Column Europe does not have outliers

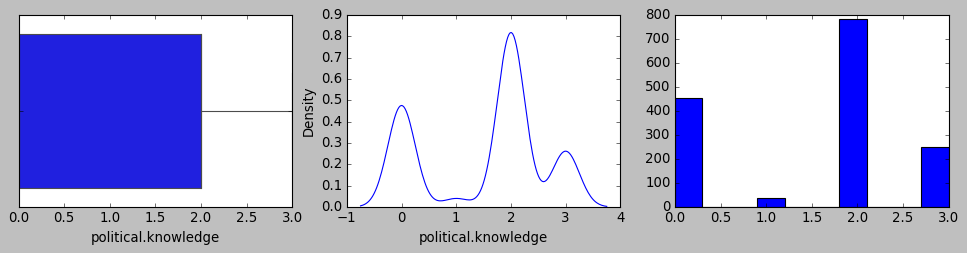


Column Europe is not normally distributed

7. Univariate analysis for political.knowledge

Mean is 1.542295, Median is 2.000000, Mode(s) are 2.0000

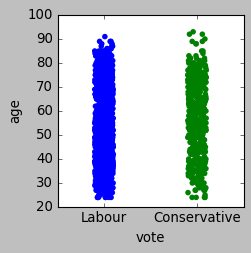
Column political.knowledge does not have outliers



Column political.knowledge is not normally distributed

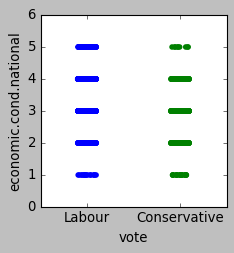
Bivariate analysis

1. bivariate analysis for age



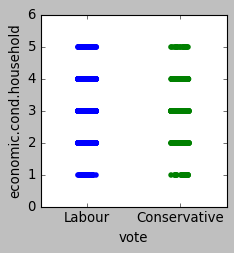
Data reveals that lower age group tend to vote for Labour while Labour also retains better vote share for higher age group comparatively.

2. bivariate analysis for economic.cond.national



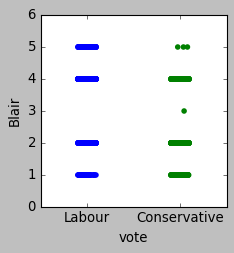
Trend suggests that voter among higher economic condition at national level tends to vote Labour and comparatively lower economic levels in the same category also tends to vote Labour higher than Conservative.

3. bivariate analysis for economic.cond.household



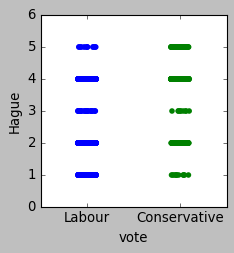
No noticeable difference across both voter groups across the levels of household economic condition.

4. bivariate analysis for Blair



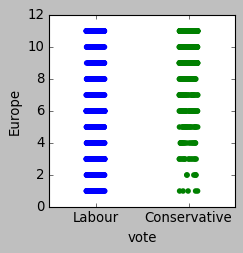
Trend suggests that Blair has received better higher scale assessment ratings among Labour voters compared to Conservative.

5. bivariate analysis for Hague



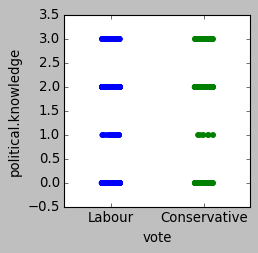
Trend suggests that Hague received higher scale ratings from Conservative voters while increased frequency in lower scale ratings from Labour

6. bivariate analysis for Europe



Trend suggests that conservative voters have comparatively lesser number of voters with low eurosceptic sentiments.

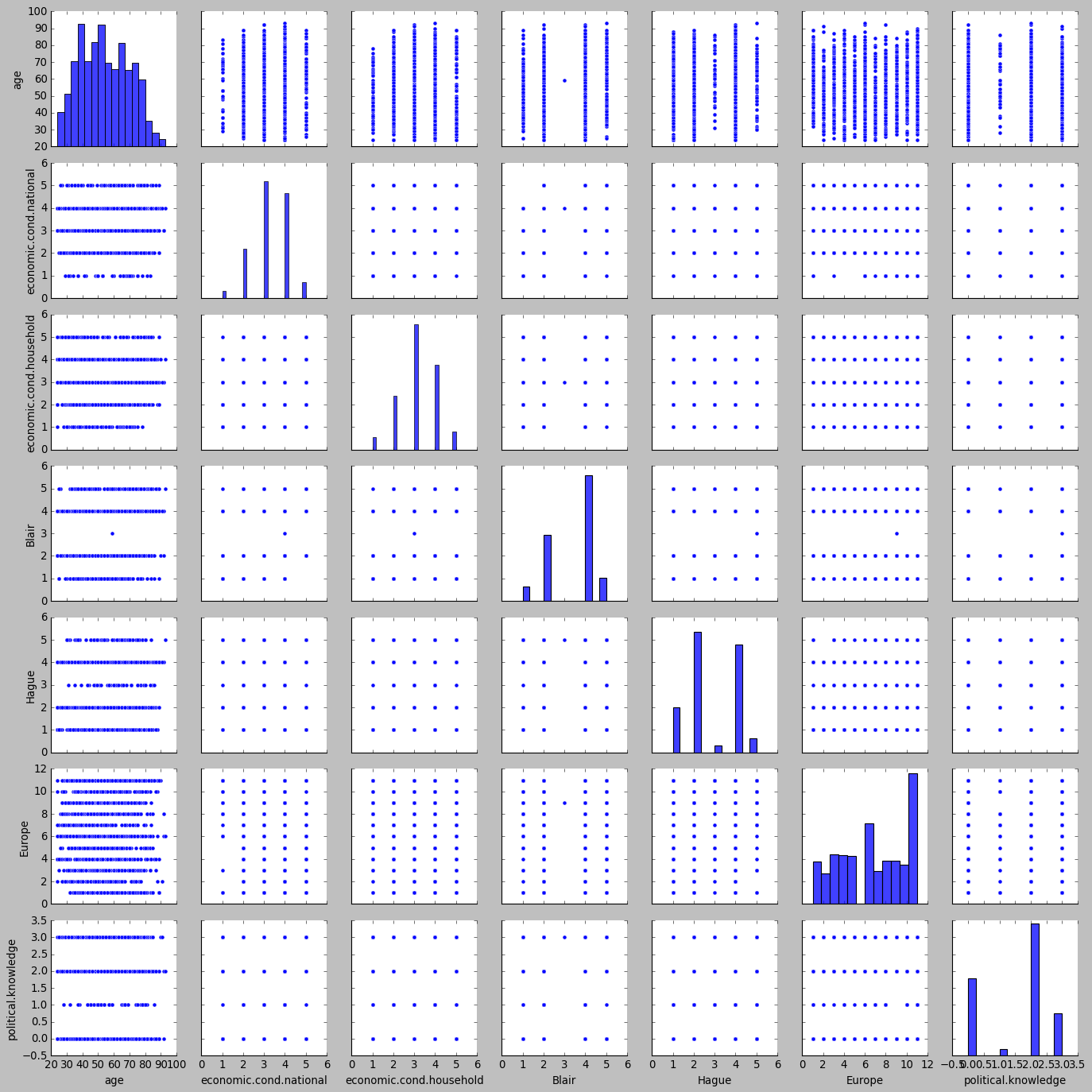
7. bivariate analysis for political.knowledge



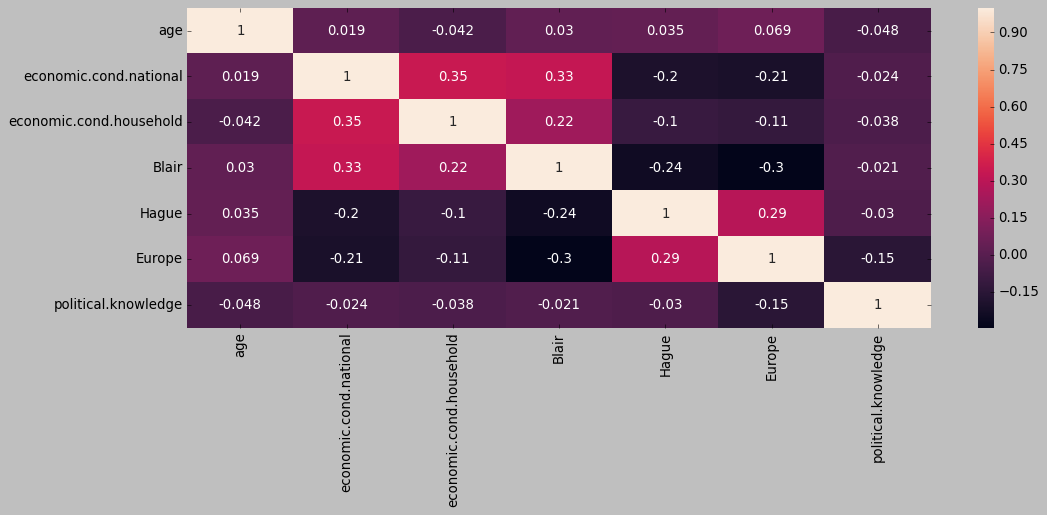
Lesser number of voters for Conservative falls under lower scale knowledge regarding party position in european integrations.

Multivariate analysis:

Pairplot for unscaled dataset



Heatmap for unscaled dataset:



It is noticed that at lower age group is are less in frequency within lower scale range for national level economic condition. Also folks with highest political knowledge tends to dip with highest age group.

* 1. **Encode the data (having string values) for Modelling. Is Scaling necessary here or not?( 3 pts), Data Split: Split the data into train and test (70:30) (2 pts).**

**Scaling:**

For logistic regression since it goes by linear equations scaling needs to be assessed.

Also since KNN goes by distance computation (Euclidean) between data points it is highly dependent on scaling needs.

For LDA, scaling standardizes the co efficients of independent variables which helps in clear separation of classes as comparison of coefficients happens on standardized data.

However Naïve bayes is unaffected by scaling. Going by the fact that independent variables have different units, scaling becomes necessary to remove the units the variables are associated with so that the linear equation can be formed on the independent variables post standardization of them.

Z scoring based scaling of data would change the coefficient, neutralize/remove the intercept while the accuracy score remains the same before and after. MSE would get scaled too.

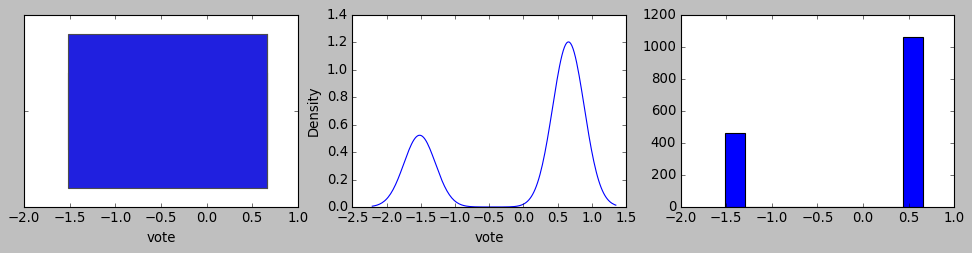
Accordingly, below is the univariate analysis for the dataset that is standardized on z scores

Univariate analysis for scaled and outlier treated dataset:

1. Univariate analysis for vote

Mean is -0.000000, Median is 0.659256, Mode(s) are 0.6593

Column vote does not have outliers

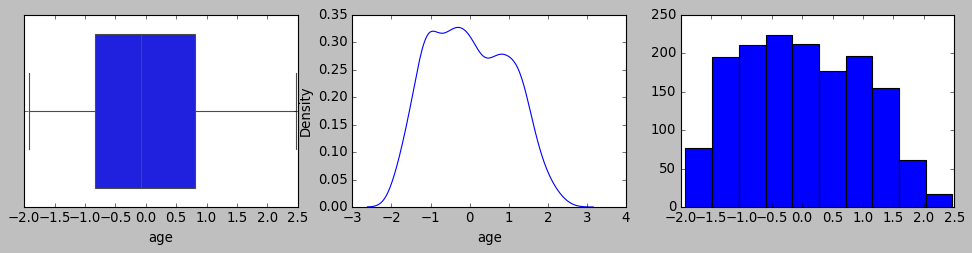


Column vote is not normally distributed

2. Univariate analysis for age

Mean is 0.000000, Median is -0.075276, Mode(s) are -1.0940

Column age does not have outliers

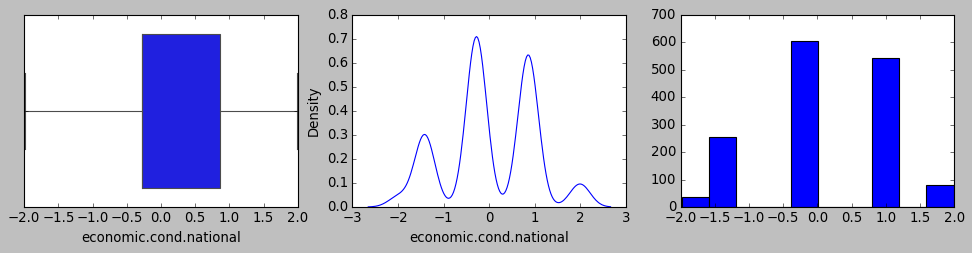


Column age is not normally distributed

3. Univariate analysis for economic.cond.national

Mean is 0.013775, Median is -0.279218, Mode(s) are -0.2792

Column economic.cond.national does not have outliers

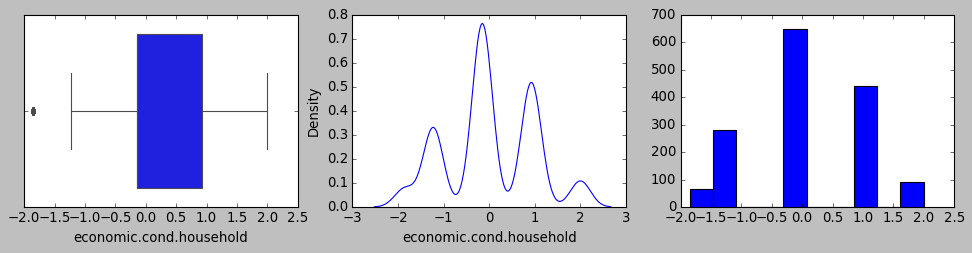


Column economic.cond.national is not normally distributed

4. Univariate analysis for economic.cond.household

Mean is 0.019100, Median is -0.150948, Mode(s) are -0.1509

Column economic.cond.household has outliers

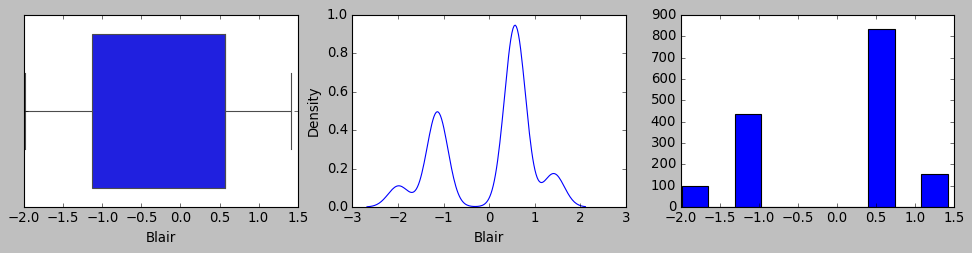


Column economic.cond.household is not normally distributed

5. Univariate analysis for Blair

Mean is 0.000000, Median is 0.566716, Mode(s) are 0.5667

Column Blair does not have outliers

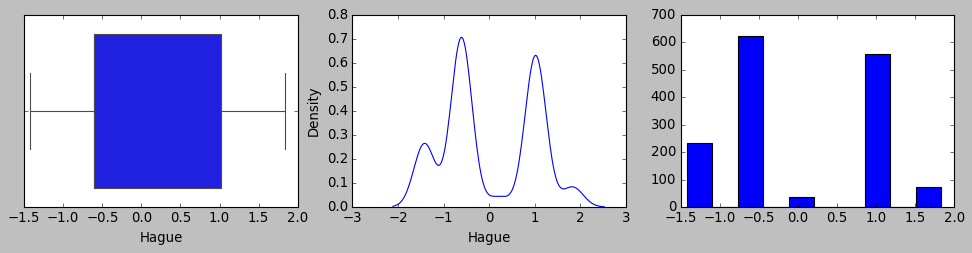


Column Blair is not normally distributed

6. Univariate analysis for Hague

Mean is -0.000000, Median is -0.607076, Mode(s) are -0.6071

Column Hague does not have outliers

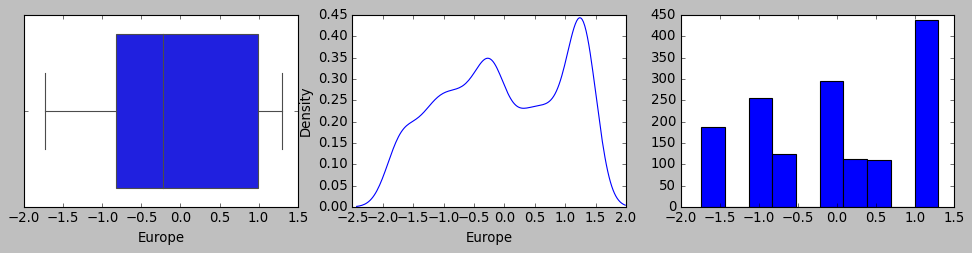


Column Hague is not normally distributed

7. Univariate analysis for Europe

Mean is -0.000000, Median is -0.221002, Mode(s) are 1.2958

Column Europe does not have outliers

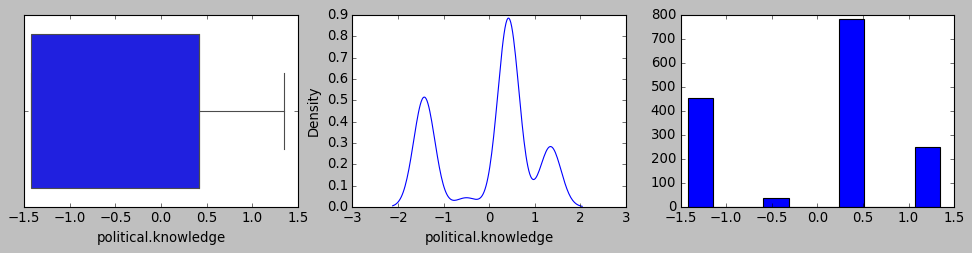


Column Europe is not normally distributed

8. Univariate analysis for political.knowledge

Mean is -0.000000, Median is 0.422643, Mode(s) are 0.4226

Column political.knowledge does not have outliers



Column political.knowledge is not normally distributed

9. Univariate analysis for gender\_male

Mean is -0.000000, Median is -0.937059, Mode(s) are -0.9371

Column gender\_male does not have outliers

Column gender\_male is not normally distributed

**Encoding the categorical variables:**

Below gives the exact distribution of categorical variables:

1. VOTE : 2

Conservative 462

Labour 1063

Name: vote, dtype: int64

1. GENDER : 2

male 713

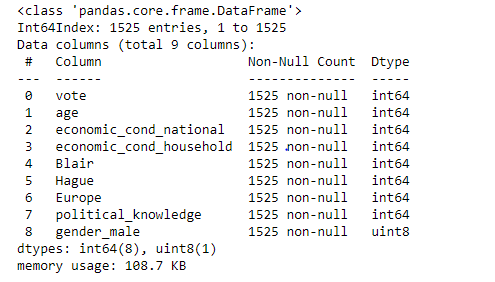
female 812

Name: gender, dtype: int64

Target variable vote has been encoded with Labour=1 and Conservative=0

One hot encoding has been applied to gender to create binary variable depicting the binary status for each of the gender (Male/Female) and to optimize the first one has been dropped as the 0 or 1 in the other is suffice to represent both.

Below is the data structure after encoding the dataset.



**Data split into training and testing by 70:30 ratio.**

Based on the splitting of training and testing data below are the respective dataset properties for its shape and 5 number summary depicting the split.

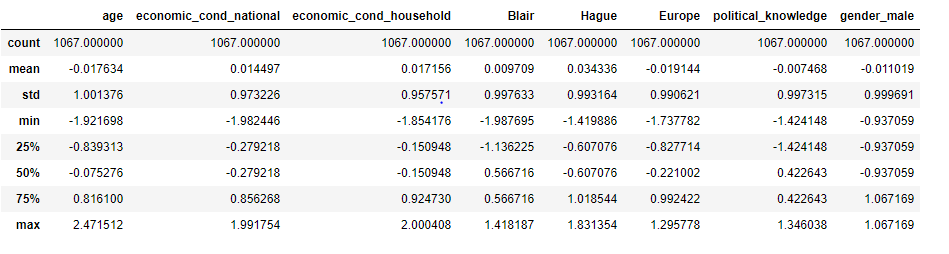
Number of rows and columns of the training set for the independent variables: (1067, 8)

Number of rows and columns of the training set for the dependent variable: (1067,)

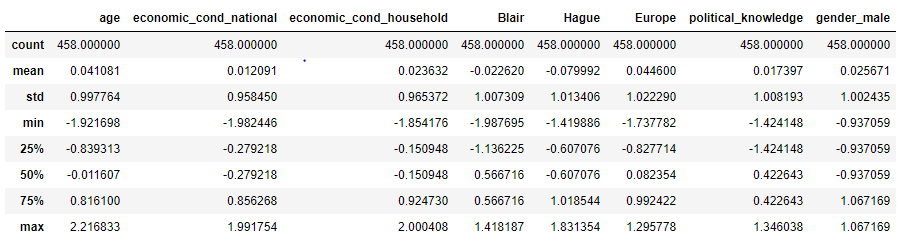
Number of rows and columns of the test set for the independent variables: (458, 8)

Number of rows and columns of the test set for the dependent variable: (458,)

5 number summary for training data set

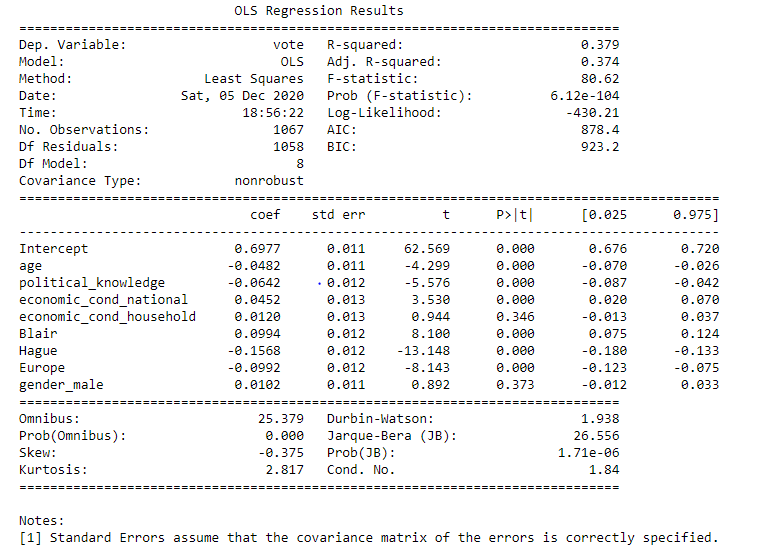


5 number summary for testing data set



* 1. **Apply Logistic Regression and LDA (Linear Discriminant Analysis) (3 pts). Interpret the inferences of both model s (2 pts)**

Based on the regression model that has been built based on the above training and testing data set below are the statistical summary using the statsmodel library.



Based on the null hypothesis that states that the independent variables have no co relationship with target variable from the universe, please note the coefficients of the predictor variables from the sample should be having p value lesser than the significance value of 5% . Based on the stats model based statistical observation all the predictor variables have their p value is much lesser than 5%. And hence the null hypothesis will be rejected for each of the predictor variables and all of the predictors will be part of the linear equation.

Subsequently VIF has been evaluated to check multi collinearity among predictor variables. Any multi collinearity among predictor variables would mean that despite prediction itself need not be impacted with any low accuracy, the interpretation of the coefficients would be wrong without treating multi collinearity. Any VIF score between 1 to 5 is an acceptable range to retain the predictor variables in linear equation. Please find below the VIF for all the predictor variables.

VIF for age is 1.0139198118441526

VIF for gender\_male is 1.0340695795969062

VIF for political\_knowledge is 1.0584898802222538

VIF for Hague is 1.1370817306128767

VIF for economic\_cond\_household is 1.1556640850266833

VIF for Europe is 1.2096030388419405

VIF for Blair is 1.2372034397053333

VIF for economic\_cond\_national is 1.2565812563307652

Based on the above observation none of the independent variables have correlation with other independent variables and hence there is no multi collinearity.

Please find below the accuracy metrics/score:

* R Squared: 0.379 (Same as adjusted R Squared)

Accordingly based on the scaled (normalized) data followed by evaluation of tuning opportunities of the model looking at sample vs universe hypothesis testing, it could be narrowed down the below formula goes well towards building both Logistic regression as well as LDA models.

**vote ~ age+political\_knowledge+economic\_cond\_national+economic\_cond\_household+Blair+Hague+Europe+gender\_male**

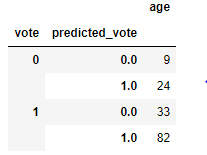
Accordingly, both models has been built based on below linear equation that depicts individual coefficients of independent variables along with intercept value.

(0.7) \* Intercept + (-0.05) \* age + (-0.06) \* political\_knowledge + (0.05) \* economic\_cond\_national + (0.01) \* economic\_cond\_household + (0.1) \* Blair + (-0.16) \* Hague + (-0.1) \* Europe + (0.01) \* gender\_male

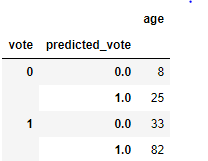
Based on the model built, below are the count comparison of prediction of test samples against the original classification.

Note: Vote refers to original classification in the dataset and predicted vote refers to model output. Age column refers to count grouped by vote and predicted vote to give a comparison. It enables us to see the real mapping of classification across original and prediction in terms of accuracy by each model.

For Logistical regression:



For LDA:



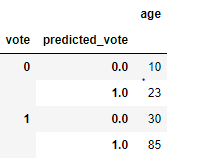
* 1. **Apply KNN Model and Naïve Bayes Model(5 pts). Interpret the inferences of each model (2 pts)**

KNN model is a distance based mechanism mostly using Euclidean measure. And in this case we are looking for a classification need and KNN can help building classification based on feature similarity. Since KNN continues to retain the training dataset instead of learning to create a model using the learning, it is a lazy learner and simplest of all models for classification. K in KNN refers to number of nearest neighbors. A small value of K indicates higher influence of noise over result while larger value is cost heavy for compute. So usually a standardized approach is to adopt K=sqrt(N)/2 where N is the size of training data set. Also, K has to be odd number to avoid tie between predicting classes.

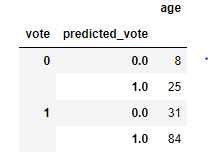
Bayes theorem is an extension of conditional probability that is based on the knowledge of prior probability values of something that has occurred. It provides a mechanism to compute posterior probability of class for the given predictor based on prior probability of the given predictor for the same class. Idea is to factor all available evidence in form of predictors into naïve Bayes rule to obtain more accurate probability for class prediction. Naïve Bayes classifier works on the principle of Bayes theorem with the assumption that input features are independent of each other. Usually not ideal for data sets with large number of numerical attributes, however, does well with noisy and missing data even with low training samples. For those independent variables that are continuous it is assumed to be distributed normally and the estimates go by mean and standard deviation of continuous variable.

Based on the model built below are the count comparison of prediction of test samples against the original classification.

For Naïve Bayes:



For KNN:



* 1. **Model Tuning (2 pts) , Bagging ( 2.5 pts) and Boosting (2.5 pts).**

For KNN,

Based on the standardized approach to adopt K=sqrt(N)/2 where N is the size of training data set and that K has to be odd number to avoid tie between predicting classes we could with K value as 39 for KNN.

For Logistical regression and LDA we have made our investigations to arrive at below conclusions:

Based on the null hypothesis that states that the independent variables have no co relationship with target variable from the universe, please note the coefficients of the predictor variables from the sample should be having p value lesser than the significance value of 5% . Based on the stats model based statistical observation all the predictor variables have their p value is much lesser than 5%. And hence the null hypothesis will be rejected for each of the predictor variables and all of the predictors will be part of the linear equation.

Subsequently VIF has been evaluated to check multi collinearity among predictor variables. Any multi collinearity among predictor variables would mean that despite prediction itself need not be impacted with any low accuracy, the interpretation of the coefficients would be wrong without treating multi collinearity. Any VIF score between 1 to 5 is an acceptable range to retain the predictor variables in linear equation. Please find below the VIF for all the predictor variables.

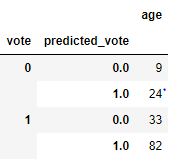
Also the max\_iteration for Logistical regression has been kept at 100 to strike a balance to optimize as well as control over fitment by increasing to more iterations with “newton-cg” solver to arrive at optimal prediction for train and test convergence.

Bagging:

For Bagging we are using random forest that utilizes bootstrapping and ensemble methods to create multiple trees out which final classification goes by voting mechanism across the models.

And in this case we are also tuning the model in terms of number of estimators (trees) to be limited at 50 followed by maximum features to 50% of total features which is 4. Accordingly bagging spawns many trees as strong learners in parallel and hence over fitment of individual tree should not be that much of a concern if in rare case as the overall classification goes by voting across trees and the individual datasets across multiple such models running parallelly are only a subset sample.

Accordingly based on the model execution below are the comparative count between original classification in the dataset versus predicted vote by the random forest classifier.



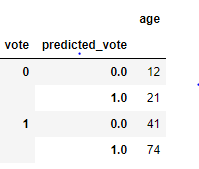
Boosting:

In contrast to Bagging, boosting trains large number of weak learners and that too sequentially by focusing on misclassified data by improving their weights further towards training it for accuracy. Boosting then combines all weak learners into a single strong learner to provide a overall classification and hence it has increased aggregate complexity while using simple low complex individual models.

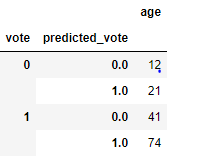
In the project we have used over sampling technique using SMOTE (synthetic minority oversampling technique) towards improving the under represented class of the target variable to avoid false positive predictions. SMOTE creates more samples of the minority class, however not by replicating the existing data points but by creating new data points within the range of possibility.

Accordingly based on the model execution below are the comparative count between original classification in the dataset versus predicted vote on the test dataset for each of the four models which clearly depicts increased accuracy and prevent false positives:

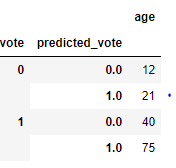
For logistical regression:



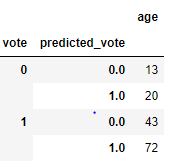
For LDA:



For Naïve Bayes:



For KNN



* 1. **Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model (4 pts) Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized (3 pts)**

**Accuracy metrics below:**

|  |  |  |
| --- | --- | --- |
|  | **Without boosting** | |
| **Model** | **Training score** | **Test score** |
| Logistical regression | 0.830365511 | 0.849344978 |
| LDA | 0.826616682 | 0.847161572 |
| Naïve Bayes | 0.826616682 | 0.847161572 |
| KNN | 0.999062793 | 0.858078603 |

|  |  |  |
| --- | --- | --- |
|  | **Boosting** | |
| **Model** | **Training score** | **Test score** |
| Logistical regression | 0.819892473 | 0.825328 |
| LDA | 0.819892473 | 0.825328 |
| Naïve Bayes | 0.8125 | 0.825328 |
| KNN | 0.891801075 | 0.790393 |

**Performance accuracy**

|  |  |  |
| --- | --- | --- |
|  | **Original** | |
| **Model** | **Training score** | **Test score** |
| Logistical regression | 0.830365511 | 0.849344978 |
| LDA | 0.826616682 | 0.847161572 |
| Naïve Bayes | 0.826616682 | 0.847161572 |
| KNN | 0.999062793 | 0.858078603 |
| Random Forest classifier | 0.999062793 | 0.829694323 |

|  |  |  |
| --- | --- | --- |
|  | **Boosting** | |
| **Model** | **Training score** | **Test score** |
| Logistical regression | 0.819892473 | 0.825327511 |
| LDA | 0.819892473 | 0.825327511 |
| Naïve Bayes | 0.8125 | 0.825327511 |
| KNN | 0.891801075 | 0.790393013 |

**Confusion matrix:**

Logistic regression:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Without boosting** | | |  | **Without boosting** | | |
| **Logistic regression/Training** | **Predicted Negative** | **Predicted Positive** |  | **Logistic regression/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 212 | 111 |  | Actual Negative | 94 | 45 |
| Actual Postive | 70 | 674 |  | Actual Postive | 24 | 295 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Boosting** | | |  | **Boosting** | | |
| **LDA/Training** | **Predicted Negative** | **Predicted Positive** |  | **LDA/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 610 | 134 |  | Actual Negative | 114 | 25 |
| Actual Postive | 134 | 610 |  | Actual Postive | 55 | 264 |

LDA:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Without boosting** | | |  | **Without boosting** | | |
| **LDA/Training** | **Predicted Negative** | **Predicted Positive** |  | **LDA/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 217 | 106 |  | Actual Negative | 98 | 41 |
| Actual Postive | 79 | 665 |  | Actual Postive | 29 | 290 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Boosting** | | |  | **Boosting** | | |
| **LDA/Training** | **Predicted Negative** | **Predicted Positive** |  | **LDA/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 610 | 134 |  | Actual Negative | 114 | 25 |
| Actual Postive | 134 | 610 |  | Actual Postive | 55 | 264 |

Naïve Bayes:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Without boosting** | | |  | **Without boosting** | | |
| **Naïve Bayes/Training** | **Predicted Negative** | **Predicted Positive** |  | **Naïve Bayes/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 226 | 97 |  | Actual Negative | 102 | 37 |
| Actual Positive | 88 | 656 |  | Actual Positive | 33 | 286 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Boosting** | | |  | **Boosting** | | |
| **Naïve Bayes/Training** | **Predicted Negative** | **Predicted Positive** |  | **Naïve Bayes/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 600 | 144 |  | Actual Negative | 112 | 27 |
| Actual Positive | 135 | 609 |  | Actual Positive | 53 | 266 |

KNN:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Without boosting** | | |  | **Without boosting** | | |
| **KNN/Training** | **Predicted Negative** | **Predicted Positive** |  | **KNN/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 323 | 0 |  | Actual Negative | 99 | 40 |
| Actual Positive | 1 | 743 |  | Actual Positive | 25 | 294 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Boosting** | | |  | **Boosting** | | |
| **KNN/Training** | **Predicted Negative** | **Predicted Positive** |  | **KNN/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 715 | 29 |  | Actual Negative | 114 | 25 |
| Actual Positive | 132 | 612 |  | Actual Positive | 71 | 248 |

Bagging through Random Forest:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Bagging** | | |  | **Bagging** | | |
| **Random Forest/Training** | **Predicted Negative** | **Predicted Positive** |  | **Random Forest/Testing** | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | 323 | 0 |  | Actual Negative | 95 | 44 |
| Actual Positive | 1 | 743 |  | Actual Positive | 34 | 285 |

AUC score:

|  |  |  |
| --- | --- | --- |
| **AUC** | **Without boosting** | |
| **Model** | **Training score** | **Test score** |
| Logistical regression | 0.781 | 0.801 |
| LDA | 0.877 | 0.916 |
| Naïve Bayes | 0.875 | 0.91 |
| KNN | 1 | 0.917 |
| Random Forest classifier | 1 | 0.899 |

|  |  |  |
| --- | --- | --- |
| **AUC** | **Boosting** | |
| **Model** | **Training score** | **Test score** |
| Logistical regression | 0.884 | 0.914 |
| LDA | 0.884 | 0.914 |
| Naïve Bayes | 0.886 | 0.908 |
| KNN | 0.965 | 0.863 |

ROC curve by model without boosting.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Logistic (without boosting) | |  |  |  |  |  | Logistic(with boosting) | |  |  |  |  |  |  |  |  |
| Training |  |  |  |  |  |  | Training |  |  |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  | |  | | --- | |  | |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | Testing |  |  |  |  |  |  |  |  |  |
| Testing-Logistic (without boosting) | |  |  |  |  |  | Testing-Logistic(with boosting) | | |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  | | | | | | | | | |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| LDA(Without boosting) |  |  |  |  |  |  | LDA( boosting) | |  |  |  |  |  |  |  |  |
| Training |  |  |  |  |  |  | Training |  |  |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  | | | | | | | | | |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Testing-LDA(Without boosting) | |  |  |  |  |  | Testing-LDA( boosting) | |  |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  | | | | | | | | | |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Naïve Bayes (Without boosting) | |  |  |  |  |  | Naïve Bayes ( boosting) | |  |  |  |  |  |  |  |  |
| Training |  |  |  |  |  |  | Training |  |  |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  | | | | | | | | | |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Testing-Naïve Bayes (Without boosting) | |  |  |  |  |  | Testing-Naïve Bayes ( boosting) | | |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  | | | | | | | | | |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| KNN (without boosting) |  |  |  |  |  |  | KNN ( boosting) | |  |  |  |  |  |  |  |  |
| Training |  |  |  |  |  |  | Training |  |  |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  | | | | | | | | | |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Testing(KNN without boosting) | |  |  |  |  |  | Testing(KNN with boosting) | | |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  | | | | | | | | | |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Random forest classifier |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Training |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Testing(Random forest) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Classification matrix is as below:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Logistic regression without boosting** | | | |  |  |  |  | **Logistic regression with boosting** | | | |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  |  | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **LDA without boosting** | | |  |  |  |  |  | **LDA with boosting** | |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  |  | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Naïve Bayes without boosting** | | |  |  |  |  |  | **Naïve Bayes with boosting** | | |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  |  | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| **KNN without boosting** | | |  |  |  |  |  | |  | | --- | | **KNN with boosting** | |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Random Forest** | |  |  |  |  |  |  |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

* 1. **Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.**

Since the classification has to focus on the probability of favorable votes across both the parties of labour and conservative, the business question is to find who will the vote (True positives) more than who did not get the vote (True negative). And hence it is important to strike a balance on the true positives across prediction metrics (precision) and actual metrics(recall). That leads to the fact both Logistic regression and LDA has done better without boosting even though naïve bayes has better scores under boosting wherein the training and testing convergence is much better in case of earlier case of LDA and logistic regression.

Also based on the model we could derive below linear equation where we can see that Blair assessment has most significant impact on the voting pattern, meaning every unit of increase in his assessment results in better prediction of voting by 10%.

(0.7) \* Intercept + (-0.05) \* age + (-0.06) \* political\_knowledge + (0.05) \* economic\_cond\_national + (0.01) \* economic\_cond\_household + (0.1) \* Blair + (-0.16) \* Hague + (-0.1) \* Europe + (0.01) \* gender\_male

**2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)**

Number of characters:

Total characters in the speech from President Roosevelt:7571

Total characters in the speech from President Kennedy:7571

Total characters in the speech from President Nixon:7571

Number of words**:**

Total words spoken by President Roosevelt:1360

Total words spoken by President Kennedy:1390

Total words spoken by President Nixon:1819

Number of sentences:

Total sentences in the speech from President Roosevelt:38

Total sentences in the speech from President Kennedy:27

Total sentences in the speech from President Nixon:51

**2.2) Remove all the stopwords from the three speeches.**

**Stop words are common words that are not useful in providing value or context and hence would have to be processed to eliminate them as a part of text cleaning. Ex: ‘the’, ‘an’, ‘in’**

**Following are the top 10 common words by it’s frequency spoken in president Roosevelt’s speech**

[('the', 104),

('of', 81),

('and', 41),

('to', 35),

('in', 30),

('a', 28),

('is', 24),

('--', 22),

('we', 22),

('that', 21)]

After stop words have been removed with words converted to lower case following are the top 10 common words spoken in president Roosevelt’s speech by it’s frequency.

[('us', 25),

('let', 22),

('--', 17),

('new', 15),

('peace', 11),

('great', 9),

('america', 9),

('world.', 8),

("america's", 8),

('shall', 7)]

Following are the top 10 common words by it’s frequency spoken in president Kennedy’s speech

[('the', 86),

('of', 65),

('to', 42),

('and', 41),

('we', 30),

('a', 29),

('in', 26),

('--', 24),

('our', 21),

('not', 19)]

After stop words have been removed with words converted to lower case following are the top 10 common words spoken in president Kennedy ‘s speech by it’s frequency.

[('--', 24),

('let', 16),

('us', 11),

('new', 7),

('pledge', 7),

('sides', 7),

('shall', 5),

('ask', 5),

('president', 4),

('fellow', 4)]

Following are the top 10 common words by it’s frequency spoken in president Nixon’s speech

[('the', 83),

('of', 68),

('to', 65),

('in', 58),

('and', 50),

('we', 47),

('a', 35),

('that', 33),

('our', 32),

('for', 32)]

After stop words have been removed with words converted to lower case following are the top 10 common words spoken in president Nixon’s speech by it’s frequency.

[('us', 25),

('let', 22),

('--', 17),

('new', 15),

('peace', 11),

('great', 9),

('america', 9),

('world.', 8),

("america's", 8),

('shall', 7)]

**2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)**

After stop words have been removed following are the top 3 words spoken in President Roosevelt’s speech by it’s frequency.

[('us', 25),

('let', 22),

('--', 17)]

Subsequently after removing the punctuations and performing stemming the top 3 words stands corrected as below for President Roosevelt.

[('nation', 10),

('know', 9),

('us', 8)]

After stop words have been removed following are the top 10 common words spoken in President Kennedy ‘s speech by it’s frequency.

[('--', 24),

('let', 16),

('us', 11)]

Subsequently after removing the punctuations and performing stemming the top 3 words stands corrected as below for President Kennedy.

[('let', 16),

('us', 11),

('power', 7)]

After stop words have been removed following are the top 10 common words spoken in President Nixon’s speech by it’s frequency.

[('us', 25),

('let', 22),

('--', 17)]

Subsequently after removing the punctuations and performing stemming the top 3 words stands corrected as below for President Nixon.

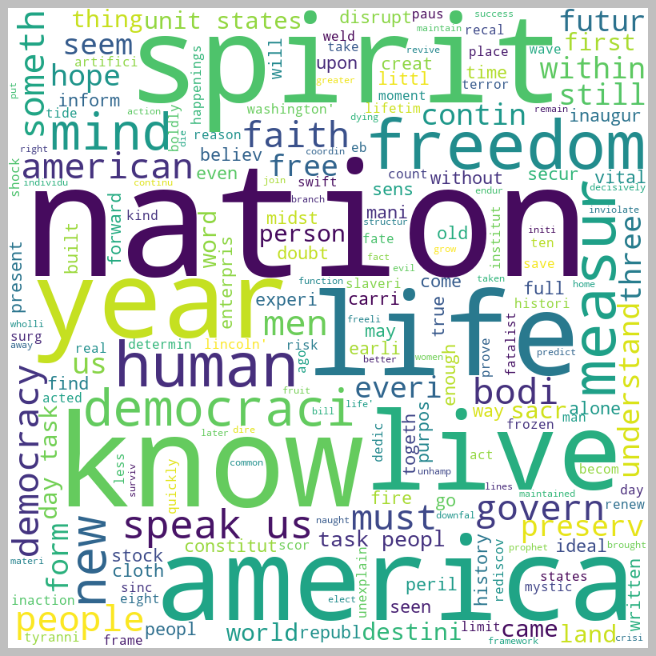
[('us', 25),

('let', 22),

('new', 15)]

**2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)**

Following is the word cloud from President Roosevelt’s speech which can be correlated to the top words spoken after stop words and punctuation removals and stemming based on the font sizing.



Following is the word cloud from President Kennedy’s speech which can be correlated to the top words spoken after stop words and punctuation removals and stemming based on the font sizing..



Following is the word cloud from President Nixon’s speech which can be correlated to the top words spoken after stop words and punctuation removals and stemming based on the font sizing..

