MACHINE LEARNING PROJECT

PGP-DSBA Online November'21

07 Jun 2022

Abstract

Questionnaire based analysis for two different business problems

Venkadasubramanian Jayakumar

Venkadasubramanian_j@outlook.com

Table of Contents

List of Figures

List of Tables

List of Equation

1. Problem: Election Exit Poll Analysis

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 have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Data Dictionary:

Variable Name	Description	Data Types
Vote	Party choice: Conservative or Labour	Object
Age	In years	Int 64
Economic.cond.national	Assessment of current national economic conditions, 1 to 5.	Int 64
Economic.cond.household	Assessment of current household economic conditions, 1 to 5.	Int 64
Blair	Assessment of the Labour leader, 1 to 5	Int 64
Hague	Assessment of the Conservative leader, 1 to 5.	Int 64
Europe	an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.	Int 64
Political.knowledge	Knowledge of parties' positions on European integration, 0 to 3.	Int 64
gender	Female or male.	Object

Table 1 Data Dictionary

There are total of 1525 rows and 9 columns out of which 7 columns are int, 2 object data type.

Sample of Dataset

	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

Table 2 Sample of Dataset

1.1) Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.

Summary Stats:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
vote	1525	2	Labour	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	1525.0	NaN	NaN	NaN	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	NaN	NaN	NaN	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	NaN	NaN	NaN	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	NaN	NaN	NaN	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	NaN	NaN	NaN	2.746885	1.230703	1.0	2.0	2.0	4.0	5.0
Europe	1525.0	NaN	NaN	NaN	6.728525	3.297538	1.0	4.0	6.0	10.0	11.0
political.knowledge	1525.0	NaN	NaN	NaN	1.542295	1.083315	0.0	0.0	2.0	2.0	3.0
gender	1525	2	female	812	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 3 Summary Stats

- Average age of voter is 54 years. We have the youngest voter as 24 years and oldest voter as 93 years old.
- Male and Female voters are equally present in the dataset.

Checking of Null Values:

There are no null values in the dataset.

Check for Duplicates

Number of Duplicated Rows: 8
We have 8 rows of duplicate observation, which shall be removed from the original dataset for further processing.

Number of Rows after removing duplicates: 1517
Number of Columns after removing duplicates: 9

Skewness

Variable	Skewness
Blair	-0.5389805841647254
Europe	-0.1417506103835579
Hague	0.14604675166469203
Age	0.1396615989084527
Economic.cond.household	-0.144005097351352
Economic.cond.national	-0.23823834819079348
Politica.knowledge	-0.42250931746800596

Table 4 Skewness of Variables

Skewness for all the variables are close to zero, thus all the variables are normally distributed.

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) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

Univariate Analysis

- Assessment score of both the leaders **Hague** and **Blair** has very small number of vote who chose '**3**'
- Male and female voters are almost equally represented.
- Most of the people represented are politically knowledge.
- National and household economic condition are similarly distributed.

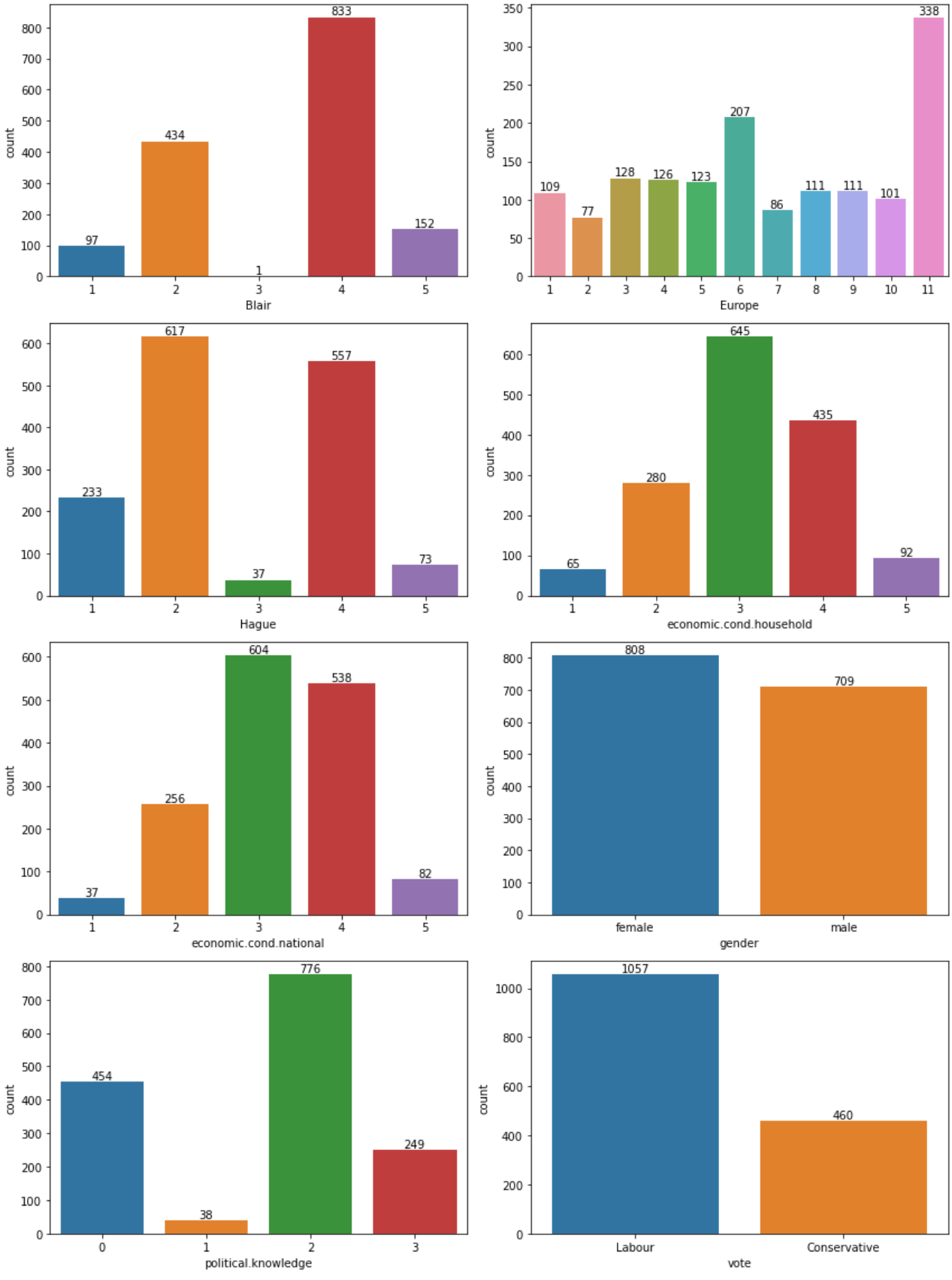


Figure 1 Univariate Analysis of Categorical variables

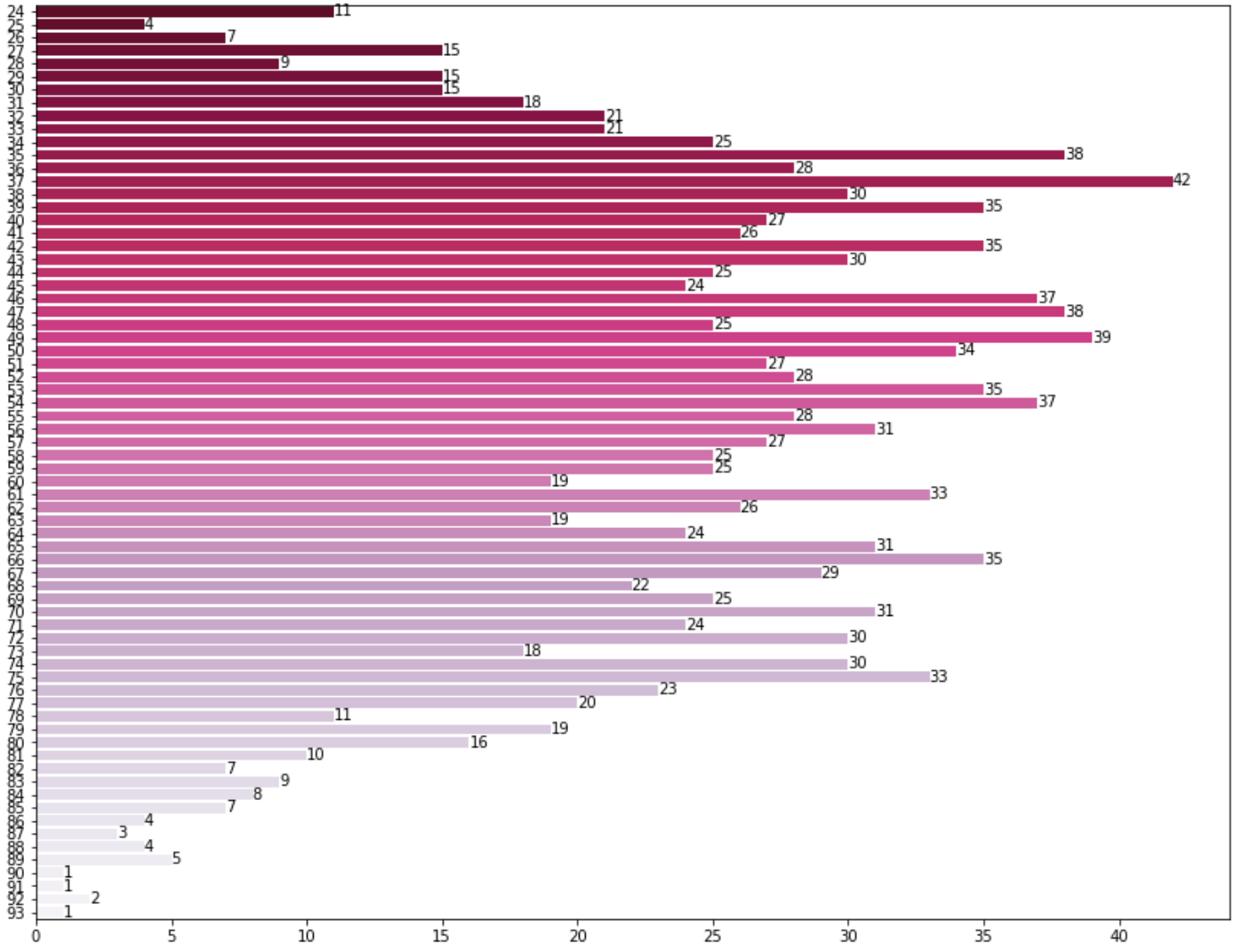


Figure 2 Spread of Voters among Age groups

- As we can see that most of the voter are between the age of 30 to 75 years.
- We can also notice that there is a more number of voters clustered among the age of 60-70 years.

Labour	**0.6967**
Conservation	**0.3032**

Table 5 Class Balance ratio

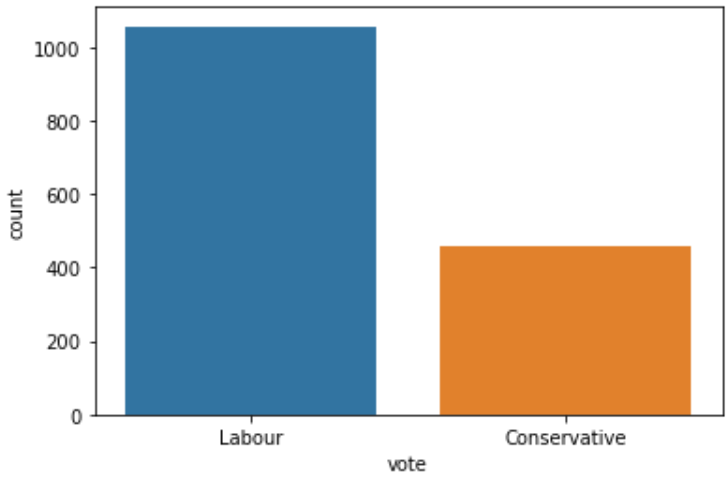


Figure 3 Class Balance ratio

As far the class representation, we have a very decent ratio of 70:30 of both classes. Thus we may not need for any class balance technique.

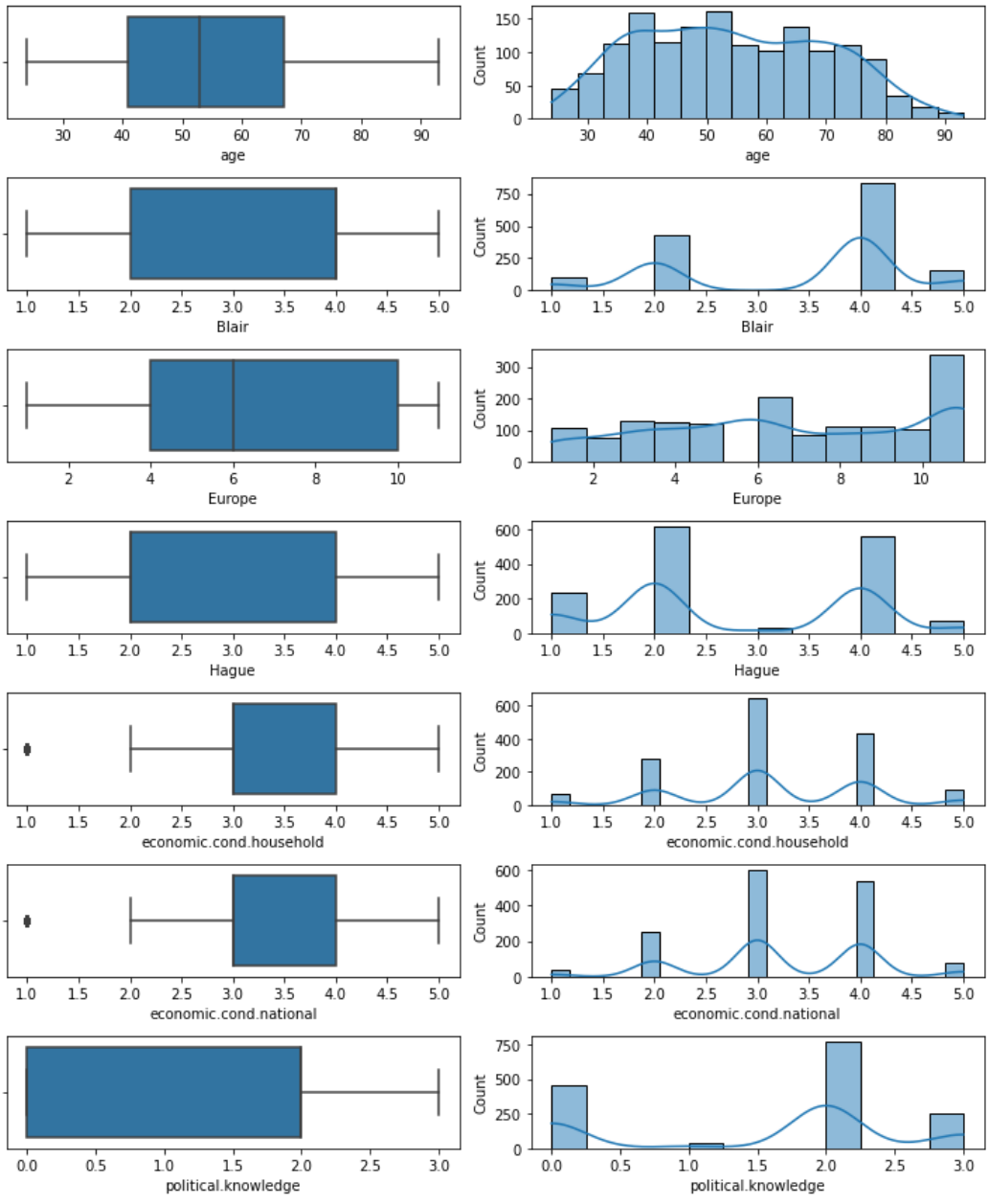


Figure 4 Boxplots and Histograms

Age is the only continuous variable, which has no outliers. Thus no outlier treatment is necessary.

Bivariate & Multivariate

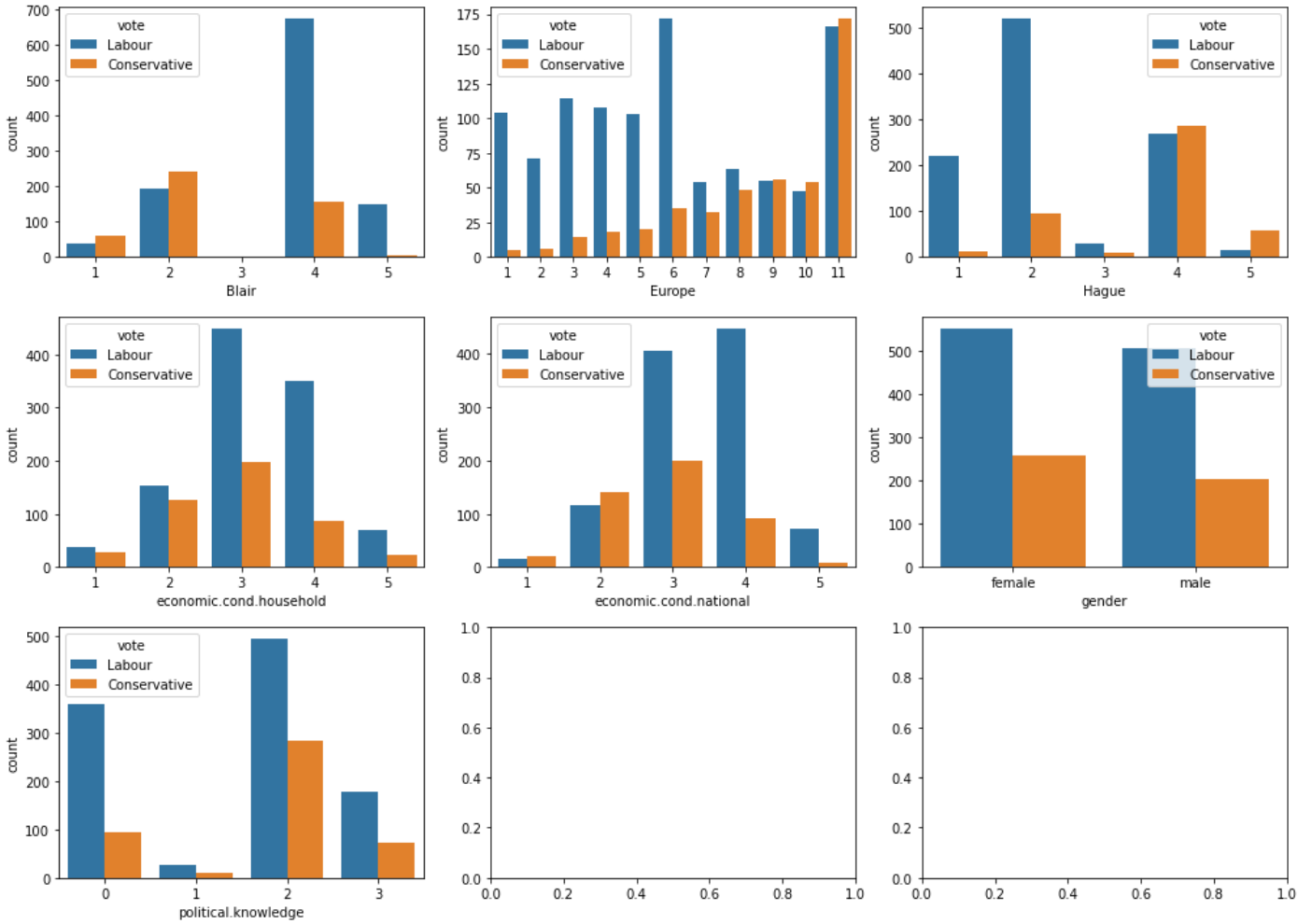


Figure 5 Bivariate Analysis of Categorical Variables

- It's evident that when the assessment of the Leader from each party is higher than they are more likely to vote for that party. For example, When assessment score for ***Blair*** is more than 3 then they are more likely to vote for **Labor Party**. Meanwhile when assessment score for ***Hague*** is more than 3 then they are more likely to vote for **Conservative Party**.
- Person with more 'Eurosceptic' sentiment (>9 score) are more likely to vote for **Conservative Party**.
- Higher the ***economic.cond.nation***, more likely to vote for **Labour Party**.
- ***Gender*** does not have significance in voting for either party.
- Level of ***political.knowledge*** also does not help in identifying the voters.

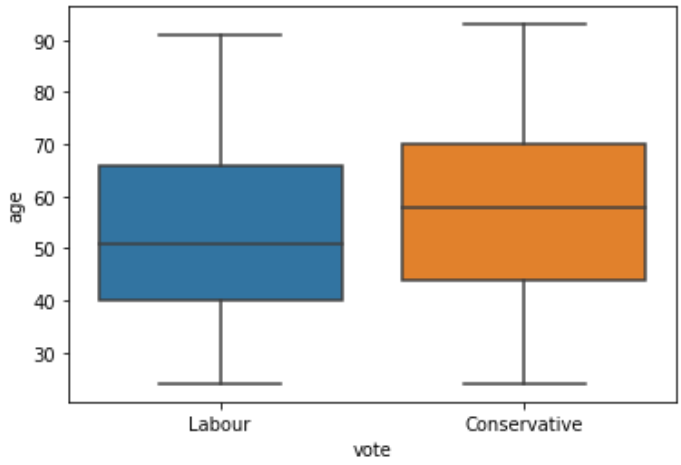


Figure 6 Average Age between two classes

Median Age of Labour Party voters is slightly lesser than that of Conservative Party voters. Thus more of young voters are tend to vote for Labour party compared to Conservative part

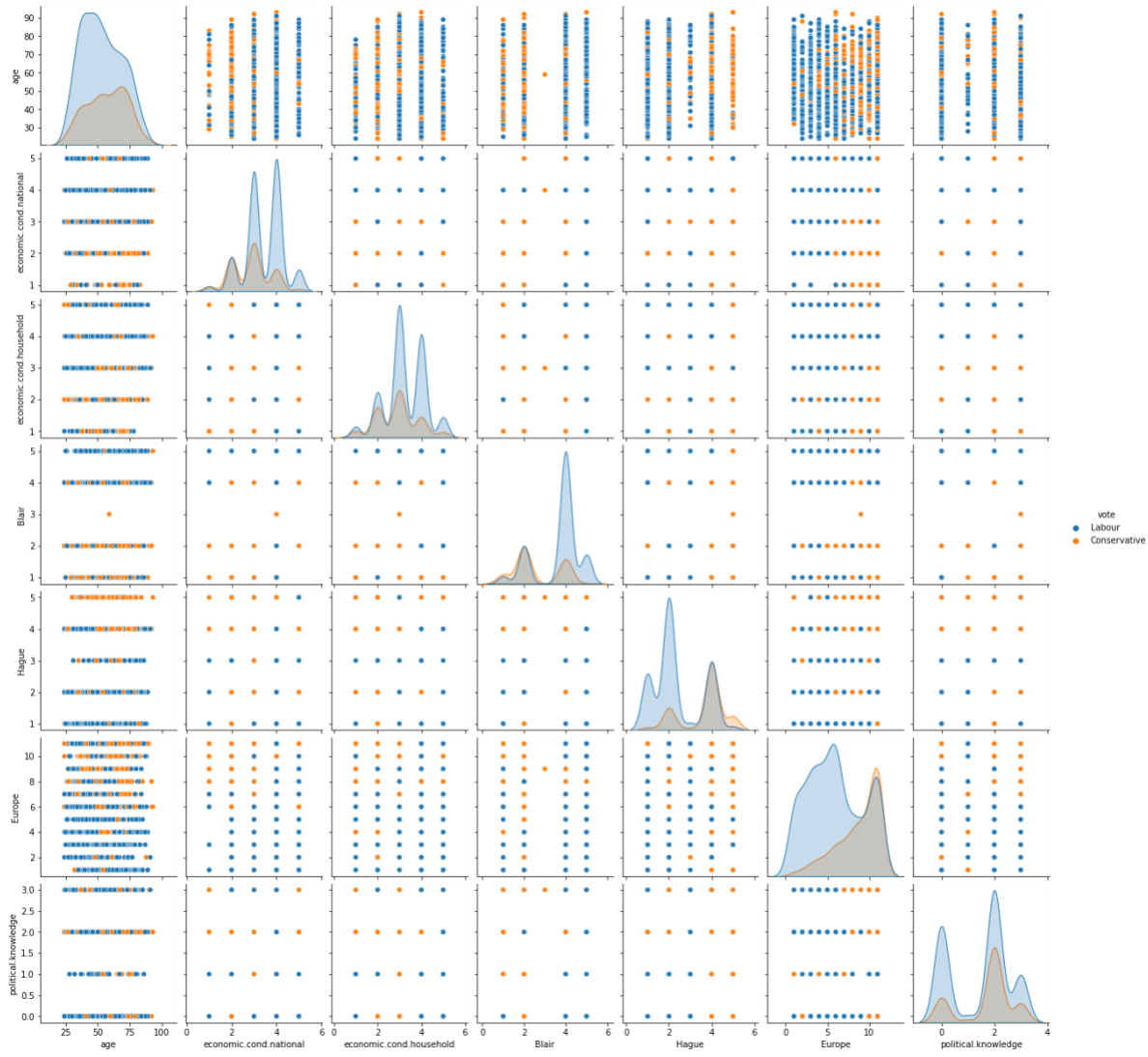


Figure 7 Pairplot

- Pair plot show that no variable has any type of correlation with any variable

From the correlation matrix we can see that none of the variable is significantly correlated with any other variables.

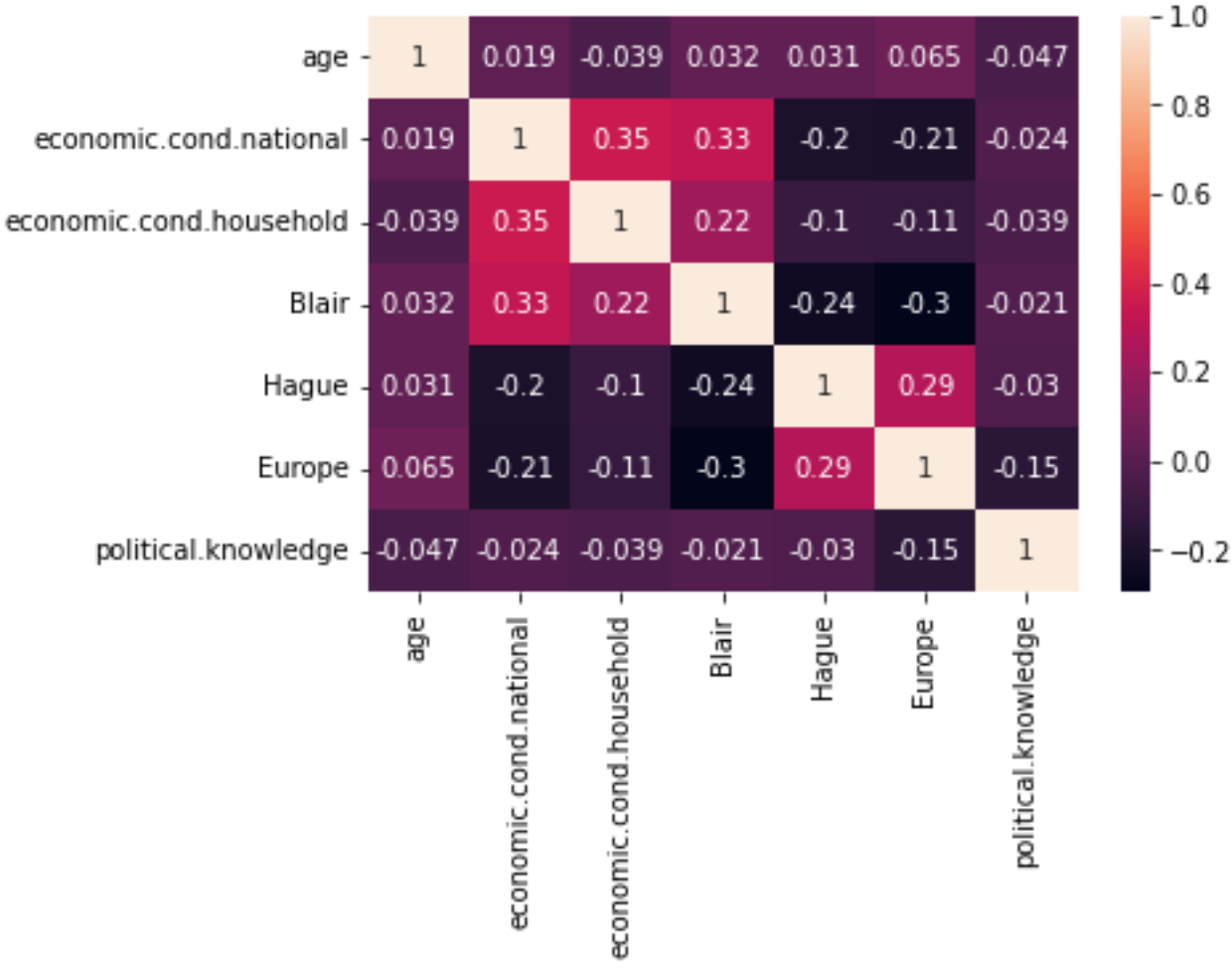


Figure 8 Correlation plot as Heatmap

1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed.

As most of the data is encoded already we only need to encode the variable '***Gender***'. We shall use **pd.get_dummies(drop_first=True)** for encoding.

- Column '**gender**' has been encoded with dummy variable. When it is 0 then it represents female and male when it is 1.

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

Table 6 Gender Variable Encoded

Why do we need Scaling?

When different variables are in different scales and thus having a different variance which in trun makes variables with higher scale of difference has more weightage. Thus it is necessary to scale all the variables to have mean of 0 and standard deviation of 1. The standard score of a sample x is calculated as:

$$z = \frac{(x - \bar{x})}{s}$$

Equation 1 Standard Scaler

OR

$$\frac{(X - X_{min})}{(X_{max} - X_{min})}$$

Equation 2 Min-Max Scalar

$\bar{x}$ is mean of the feature and s is the standard deviation. Scaling happens independently on each feature in the dataset. Mean and Standard Deviation are stored to use later for transform the data. We can also use the Min-Max scaler as here since many of the variables are ordinal variables thus it makes sense to use Min-Max scaler.

However, most of the variables are categorical in nature except age. And only KNN model (distance based) benefits from scaling. Thus we use scaled variable only for KNN model. Other model doesn't benefit from scaling.

Variance Before Scaling:

age	246.544655
economic.cond.national	0.777558
economic.cond.household	0.866890

Blair	1.380089
Hague	1.519005
Europe	10.883687
political.knowledge	1.175961
gender	0.249099

Variance After Scaling:

age	0.051784
economic.cond.national	0.048597
economic.cond.household	0.054181
Blair	0.086256
Hague	0.094938
Europe	0.108837
political.knowledge	0.130662
gender	0.249099

- Variance in the dataset is minimized almost same for all the features. Even though we don't need scaling for Logistic Regression, LDA, Naive Bayes, Bagging or Boosting.
- KNN is sensitive to scale of the each variables as it is a distance based model. Thus we have scaled the variable to a common scale between 0 to 1.

Split the Independent variable and dependent variable as X and y respectively. And then divide further into train and test data. As a general practice we shall consider 70:30 with 30% as test data and 70% as train data.

Shape of X_train set: (1061, 8)
Shape of X_test set: (456, 8)
Shape of Y_train set: (1061,)
Shape of Y_test set: (456,)

By using '***stratify'*** we shall maintain the same response ratio in both train and test labels.

Reason for Data Split:

Every data set must be split into train set and test set data in order to train the model first using train data and then test the performance of the model on the unseen data using test data. By doing then we can access the generalisation of the model for any unseen data, if in case the model is performing well on the train set but not on the test set of data, then it's evident that the model is overfit and generalised.

It's common practice to use 1/4 th of the data for test and the remaining for training purpose. And by using Stratify in train_test_split we shall maintain the same response ratio (Response ratio: ratio of response of class of interest against other class)

1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both models (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

Logistic Regression:

Let us initialisation Logistic Regression model with default parameters.

Logistic Regression model Train Accuracy: 0.8284637134778511
Logistic Regression model Test Accuracy: 0.8552631578947368

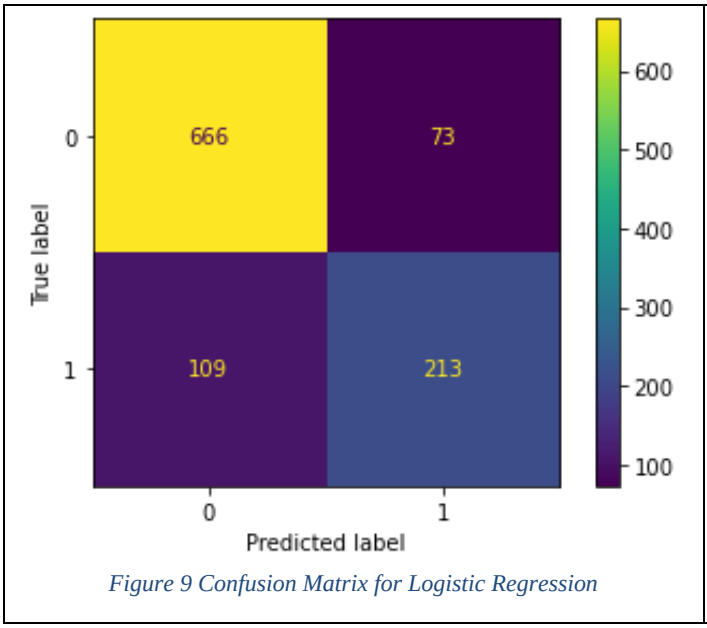
CLASSIFICATION REPORT FOR TRAIN

	precision	recall	f1-score	support
0	0.86	0.90	0.88	739
1	0.74	0.66	0.70	322
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.82	0.83	0.83	1061

CLASSIFICATION REPORT FOR TEST

	precision	recall	f1-score	support
0	0.87	0.93	0.90	318
1	0.81	0.68	0.74	138
accuracy			0.86	456
macro avg	0.84	0.81	0.82	456
weighted avg	0.85	0.86	0.85	456

Table 7 Classification report for Logistic Regression



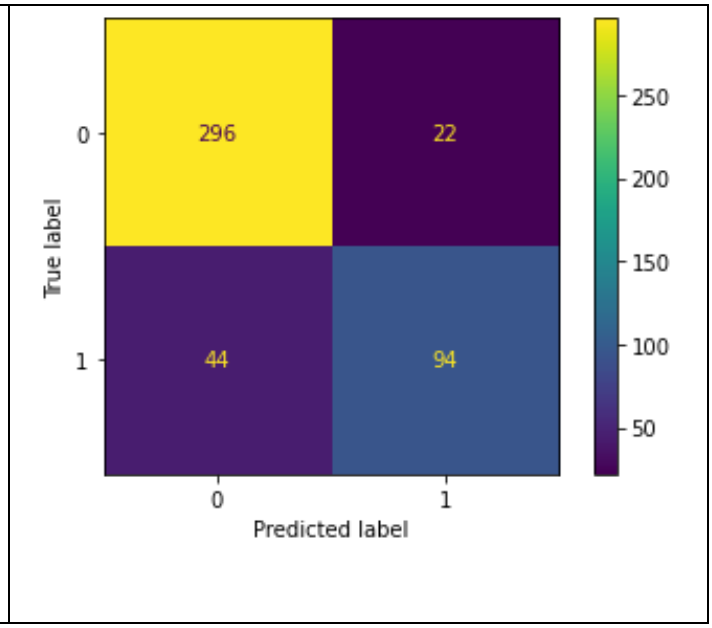


Figure 9 Confusion Matrix for Logistic Regression

- Model becomes overfit when the test performance (Accuracy) is 10% below the train performance (Accuracy).
- Here the train accuracy is ~83% whereas test accuracy is ~86% which higher than the train accuracy also the difference is not vast as there is only 5% difference. Thus the model is neither overfit nor underfit it's decent fit and a good generalisation on unknow data.

Linear Discriminant Analysis (LDA)

Let us initialise the LDA using the default parameters.

LDA model Train Accuracy: 0.822808671065033
LDA model Test Accuracy: 0.8530701754385965

CLASSIFICATION REPORT FOR TRAIN

	precision	recall	f1-score	support
0	0.86	0.89	0.87	739
1	0.72	0.67	0.70	322
accuracy			0.82	1061
macro avg	0.79	0.78	0.79	1061
weighted avg	0.82	0.82	0.82	1061

CLASSIFICATION REPORT FOR TEST

	precision	recall	f1-score	support
0	0.87	0.92	0.90	318
1	0.80	0.69	0.74	138
accuracy			0.85	456
macro avg	0.84	0.81	0.82	456
weighted avg	0.85	0.85	0.85	456

Table 8 Classification Report for LDA

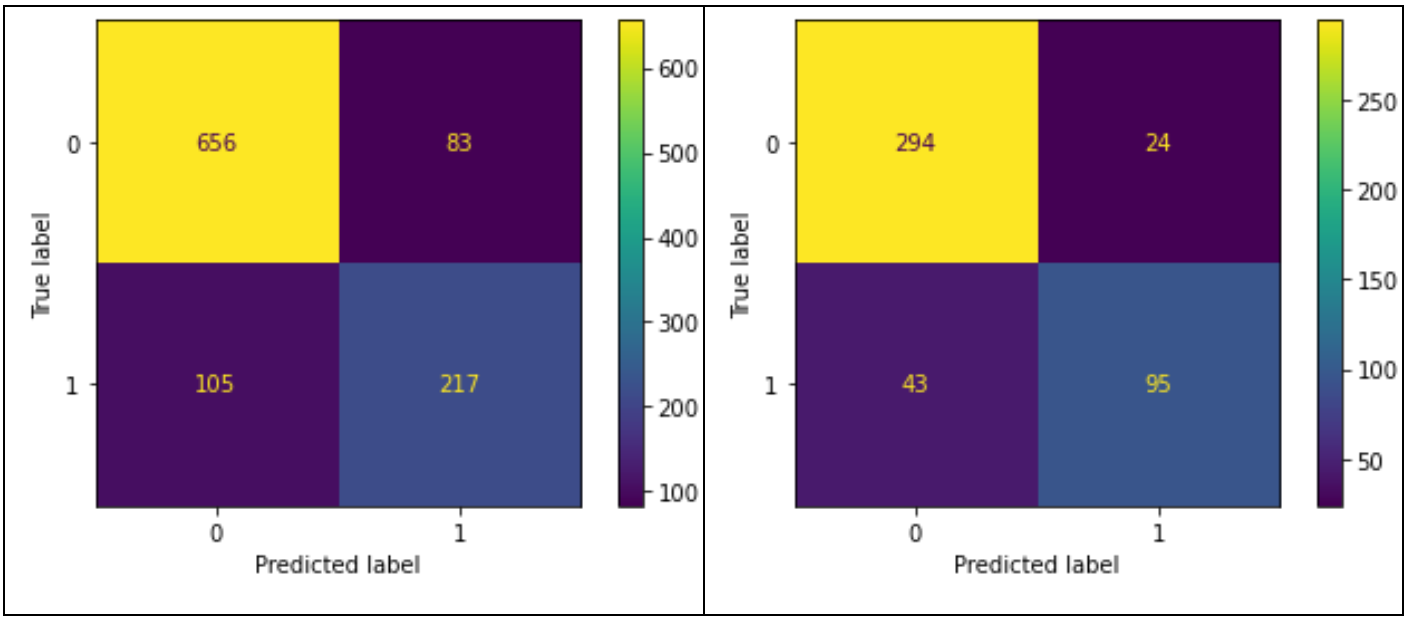


Figure 10 Confusion Matrix for LDA

- Model becomes overfit when the test performance (Accuracy) is 10% below the train performance (Accuracy).
- Here the train accuracy is ~82% whereas test accuracy is ~85% which higher than the train accuracy also the difference is not vast as there is only 5% difference. Thus the model is neither overfit nor underfit it's decent fit and a good generalisation on unknow data.

1.5) Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

KNN

Let us initialise the KNN model using default parameters.

Model without scaling

KNN model Train Accuracy score: 0.8482563619227145
KNN model Test Accuracy score: 0.8070175438596491

Model with scaling

KNN model Train Accuracy score: 0.8680490103675778
KNN model Test Accuracy score: 0.8442982456140351

As we can see that the scaled variables help in increasing the performance of the model. We shall use scaled variables for KNN model.

Naïve Bayes

Let us initialise the Naïve Bayes using default parameters.

Naive Bayes model Train Accuracy score: 0.8199811498586239
Naive Bayes model Test Accuracy score: 0.8574561403508771

1.6) Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.

Bagging

Let us initialise the Bagging with default model using base_estimator as Decision Tree.

Bagging Accuracy on Train Accuracy: 0.9868049010367578
Bagging Accuracy on Test Accuracy: 0.8377192982456141

Boosting

Let us initialise the AdaBoot and GradiantBoost model with default parameters

AdaBoosting Accuracy on Train Accuracy: 0.8397737983034873
AdaBoosting Accuracy on Train Accuracy: 0.8355263157894737

GradiantBoosting model Train Accuracy: 0.885956644674835
GradiantBoosting model Test Accuracy: 0.8399122807017544

Hyper Parameter Tuning

Using GridSearchCV, we shall find the best parameters among the different parameters,

KNN with best parameters

Number of Neighbors = (1 to 21 with increment of 2)
Distance metric = (Euclidean, Chebyshev, Manhattan)

Best parameters = {metric='manhattan', n_neighbors=17}

LDA with Best Parameters

Solver = (svd, lsqr, eigen)
Tolerance = (0.001,0.0001)

Best parameters = {solver='svd', tol=0.001}

Logistic Regression with Best parameters

Penalty = (L2 (ridge regression), None)
Solver = (Newton-cg, lbfgs)
Tolerance = (0.0001,0.00001,0.000001)
Maximum iteration = (1000,10000)

Best parameters = {max_iter = 1000, penalty = l2, solver = netwon-cg, tol = 0.0001}

Bagging with Best Parameters

Number of estimators =
Maximum Samples =
Maximum Features =

Best parameters = { max_features = 0.6, max_samples = 0.6, n_estimators = 201}

AdaBoost with Best Parameters

Number estimators = (101 to 202 with increment of 2)
Learning rate = (0.01, 0.001, 0.0001)

Best Parameters = {learning_rate = 0.01, n_estimators = 175}

GradiantBoosting with Best Parameters

Number estimators = (101 to 202 with increment of 2)
Learning rate = (0.01, 0.001, 0.0001)

Best Parameters = {learning_rate = 0.01, n_estimators = 201}

Model	Train_Accuracy	Test_Accuracy	5perc of train_accu	Model_performance
Logistic	0.8285	0.8553	0.7456	Good Fit
LDA	0.8228	0.8531	0.7405	Good Fit
KNN	0.8256	0.8443	0.7431	Good Fit
Naïve_Bayes	0.82	0.8575	0.738	Good Fit
Bagging	0.9472	0.8355	0.8525	Overfitting
AdaBoost	0.7823	0.807	0.7041	Good Fit
GradBoost	0.8464	0.8487	0.7617	Good Fit

Table 9 Generalization of Model on both Test and Train Data

- After tuning the hyper parameters we can see that all the model has good train and test score except Bagging model which is overfitting as the Train score is more than 10% higher than test score.
- Naive Bayes model has the higher Test Accuracy score followed by Logistic Regression and LDA.

	Imp_Ada	Imp_Grad
Hague	0.44	0.4673
Blair	0.28	0.1802
Europe	0.28	0.1559
political.knowledge	0	0.123
age	0	0.0401
economic.cond.national	0	0.0307
economic.cond.household	0	0.0029
gender	0	0

Table 10 Variable of Importance

- According to feature importance we can see that the most two important variable is **Hague** and **Blair**. Thus, voters assessment of Candidates of both parties helps in predicting the voters choice of party to vote.
- Next to that is **Eurosceptic sentiment,** based on the score we can able to predict the likelihood of voters choosing Labour or Conservative.

1.7 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, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)

AUC Score for All models

	AUC_Train	AUC_Test
Logistic	0.877	0.913
LDA	0.877	0.914
KNN	0.892	0.899
Naive Bayes	0.873	0.912
Bagging	0.993	0.901
AdaBoost	0.853	0.875
GradBoost	0.913	0.912

Table 11 AUC Scores of All models

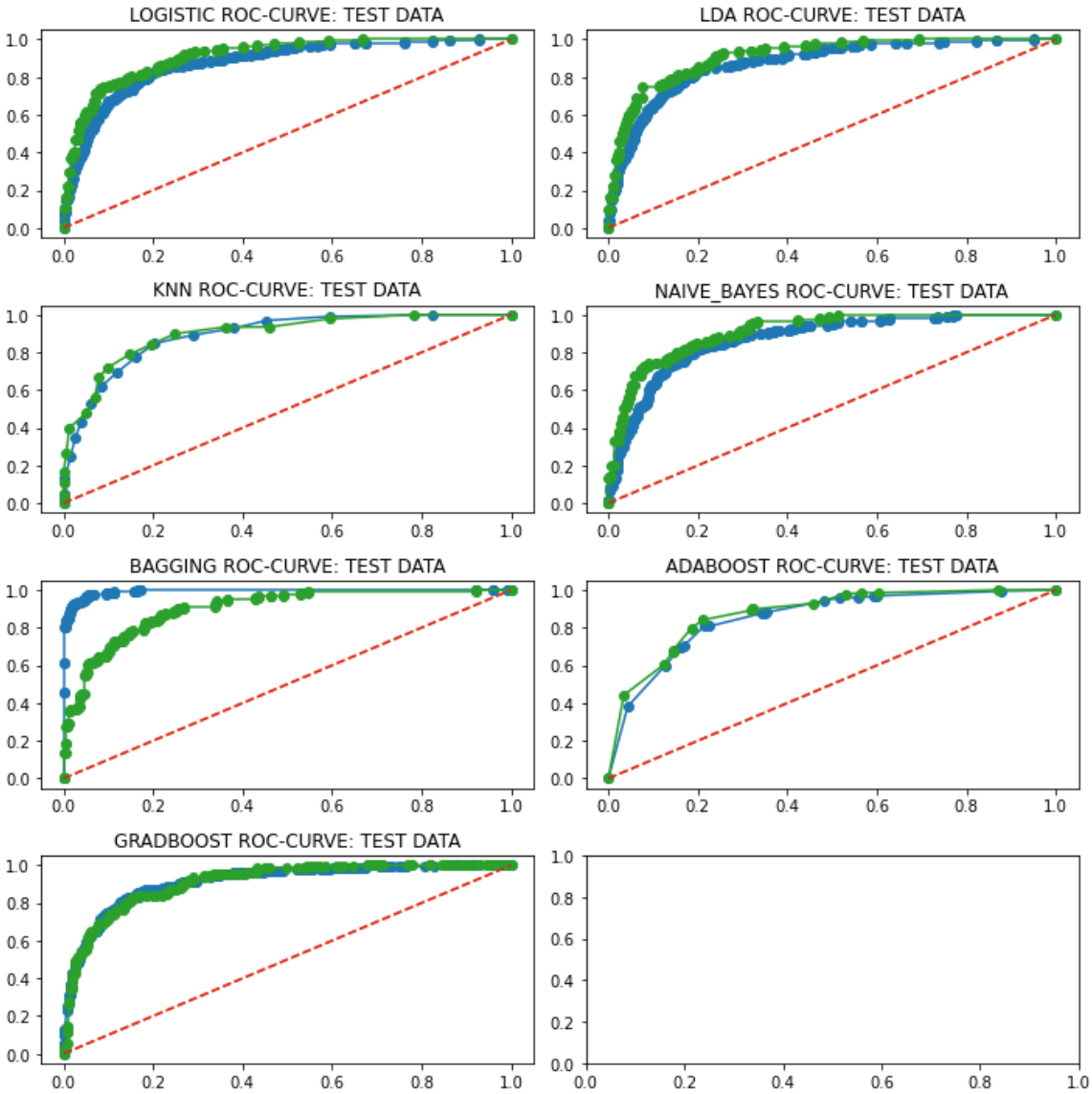


Figure 11 ROC Curves for All Models

Performance Metrics of all models

Model	Accuracy	Precision	Recall	f1 Score
Logistic	0.855263	0.8103	0.6812	0.7402
LDA	0.85307	0.7983	0.6884	0.7393
KNN	0.844298	0.7863	0.6667	0.7216
Naive_Bayes	0.835526	0.8247	0.5797	0.6809
Bagging	0.848684	0.7899	0.6812	0.7315
AdaBoost	0.807018	0.8472	0.442	0.581
GradBoost	0.857456	0.7874	0.7246	0.7547

Table 12 Performance Metrics of All Models

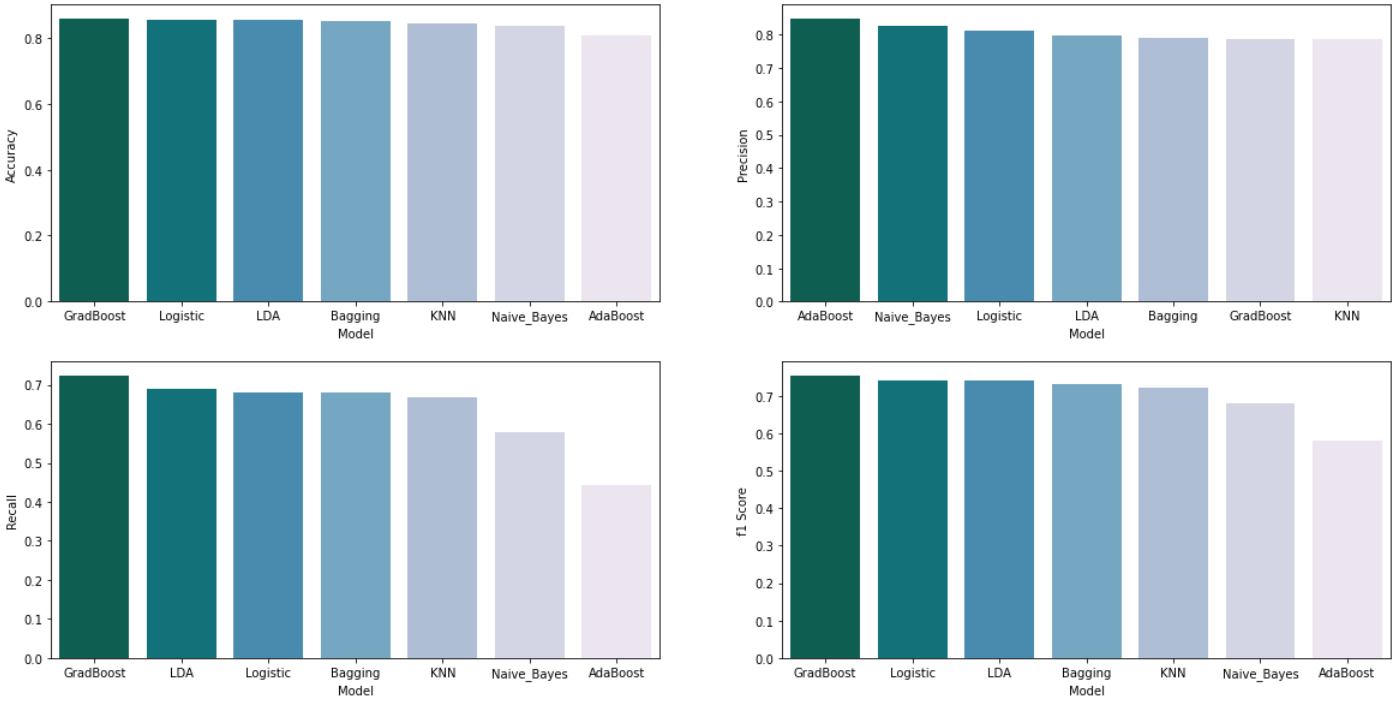


Figure 12 Performance Metrics Visualised

By comparing all the metrics of various Machine Learning models, we can see that Gradient Boosting model performs well with good Accuracy, Recall and F1 score.
Since, this problem doesn't has any class of interest. We shall look for a better F1 score of a model as both precision and recall will be more important in evaluating the model.

1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

Based on the EDA and Model Evaluation we can conclude the following recommendation,

- Major contribution for classifying the voter's choice is their assessment on the party's leaders. Thus every party must concentrate on it in order to improve their scores.
- Another important variable is 'Europe' which is intensity of Opposing the EU policies and in support of coming out of it i.e. Eurosceptic mind side. A voter with higher their score more interest in coming out of European Union
- Also voter who are more politically knowledge are tends to vote for Labour Party.
- Average age of voters who vote for Conservative party is little higher over the Labour party. Thus we can say that more young voter are more likely to vote for Labour party where as older voters are likely to vote for Conservative party.

2 Problem: Text Analysis

In this particular 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 (Hint: use .words(), .raw(), .sent() for extracting counts)

Download three corpus files and store it as a three separate variable as we are going to analysis each files individually.

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

Each document is stored in a dataframe format

	Speech-text
0	On each national day of inauguration since 178...
1	Vice President Johnson, Mr. Speaker, Mr. Chief...
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...

Table 13 Document Dataframe

Total number of Characters: 25183
Total number of Words: 5110
Total number of Sentence: 189

	Speech-text	characters	Words	Sentence
0	On each national day of inauguration since 178...	7572	1536	68
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	7619	1546	52
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9992	2028	69

Table 14 Number of Characters, Words and Sentences

2.2) Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.

Remove Stopwords

Remove all the stopwords using nltk's stopwords for English language.
Before removing stopword we must convert everything into lowercase format as the stopword list only checks with lowercase words.

	Speech-text	Speech_StopRem
0	on each national day of inauguration since 178...	national day inauguration since 1789, people r...
1	vice president johnson, mr. speaker, mr. chief...	vice president johnson, mr. speaker, mr. chief...
2	mr. vice president, mr. speaker, mr. chief jus...	mr. vice president, mr. speaker, mr. chief jus...

Table 15 Stopwords Removed

All words before stopwords: 4569
All words after stopwords: 2257

Remove Punctuation

After removing the stopwords now we shall remove the punctuation as it doesn't provide any meaningful context.

	Speech-text	Speech_StopRem	Speech_PuncRem
0	on each national day of inauguration since 178...	national day inauguration since 1789, people r...	national day inauguration since 1789 people re...
1	vice president johnson, mr. speaker, mr. chief...	vice president johnson, mr. speaker, mr. chief...	vice president johnson mr speaker mr chief jus...
2	mr. vice president, mr. speaker, mr. chief jus...	mr. vice president, mr. speaker, mr. chief jus...	mr vice president mr speaker mr chief justice ...

Table 16 Punctuations Removed

Stemming Words

After removing stopwords and punctuation, now we shall drop all the additional characters in way that even after removing it we shall be able to understand the context of words. For which we shall use Snowball Stemmer library.

	Speech-text	Speech_StopRem	Speech_PuncRem	Speech_Stem
0	on each national day of inauguration since 178...	national day inauguration since 1789, people r...	national day inauguration since 1789 people re...	nation day inaugur sinc 1789 peopl renew sens ...
1	vice president johnson, mr. speaker, mr. chief...	vice president johnson, mr. speaker, mr. chief...	vice president johnson mr speaker mr chief jus...	vice presid johnson mr speaker mr chief justic...
2	mr. vice president, mr. speaker, mr. chief jus...	mr. vice president, mr. speaker, mr. chief jus...	mr vice president mr speaker mr chief justice ...	mr vice presid mr speaker mr chief justic sena...

Table 17 Words are Stemmed

After cleansing the words we shall now get the final word counts in each document.

	Speech-text	Speech_StopRem	Speech_PuncRem	Speech_Stem	Final_WordCount
0	on each national day of inauguration since 178...	national day inauguration since 1789, people r...	national day inauguration since 1789 people re...	nation day inaugur sinc 1789 peopl renew sens ...	644
1	vice president johnson, mr. speaker, mr. chief...	vice president johnson, mr. speaker, mr. chief...	vice president johnson mr speaker mr chief jus...	vice presid johnson mr speaker mr chief justic...	705
2	mr. vice president, mr. speaker, mr. chief jus...	mr. vice president, mr. speaker, mr. chief jus...	mr vice president mr speaker mr chief justice ...	mr vice presid mr speaker mr chief justic sena...	844

Table 18 Word count after cleaning the Text

Visualising the number of words used by each president in their speech, and we can see that Kennedy's speech as more number of words compared to others, followed by Nixon and Roosevelt.

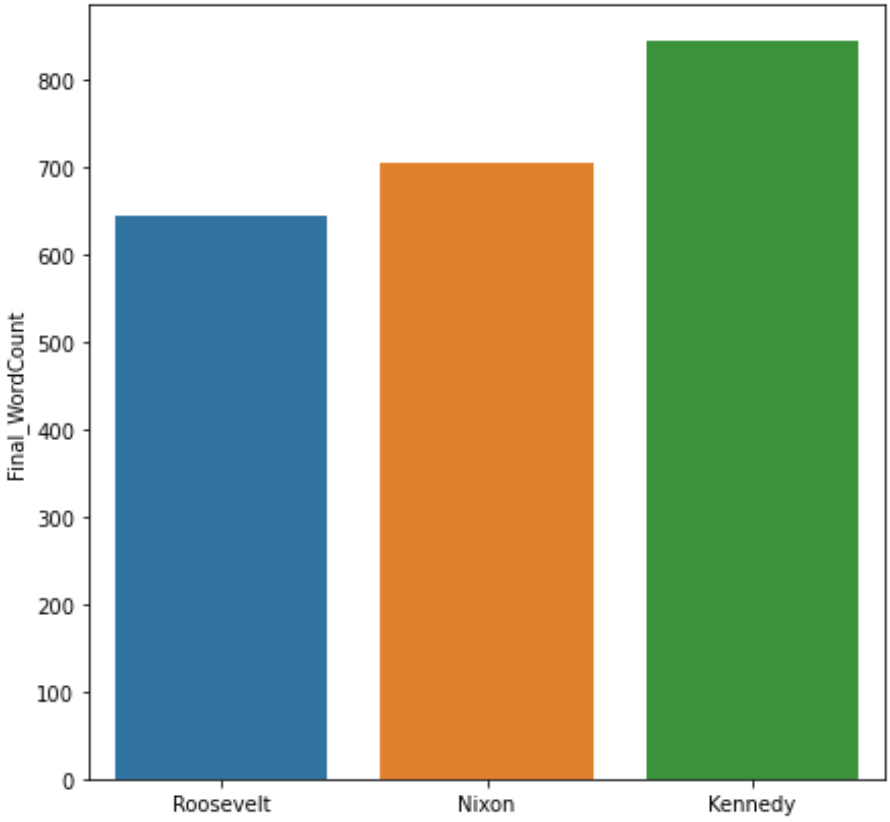


Figure 13 Word count as per Documents

president_name	Speech-text	Speech_StopRem	Speech_PuncRem	Speech_Stem	Final_WordCount	n_stopwords
Roosevelt	on each national day of inauguration since 178...	national day inauguration since 1789, people r...	national day inauguration since 1789 people re...	nation day inaugur sinc 1789 peopl renew sens ...	644	694
Nixion	vice president johnson, mr. speaker, mr. chief...	vice president johnson, mr. speaker, mr. chief...	vice president johnson mr speaker mr chief jus...	vice presid johnson mr speaker mr chief justic...	705	660
Kennedy	mr. vice president, mr. speaker, mr. chief jus...	mr. vice president, mr. speaker, mr. chief jus...	mr vice president mr speaker mr chief justice ...	mr vice presid mr speaker mr chief justic sena...	844	958

Table 19 Stopwords count of each documents

From the number of stopwords we can see that in all the three speech we have close to 50% of the words are stopwords only.

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)

President Roosevelt Speech

	Words	Frequency
0	nation	17
87	know	10
65	democraci	9
52	life	9
5	peopl	9

Table 20 Top 5 Words - President Roosevelt

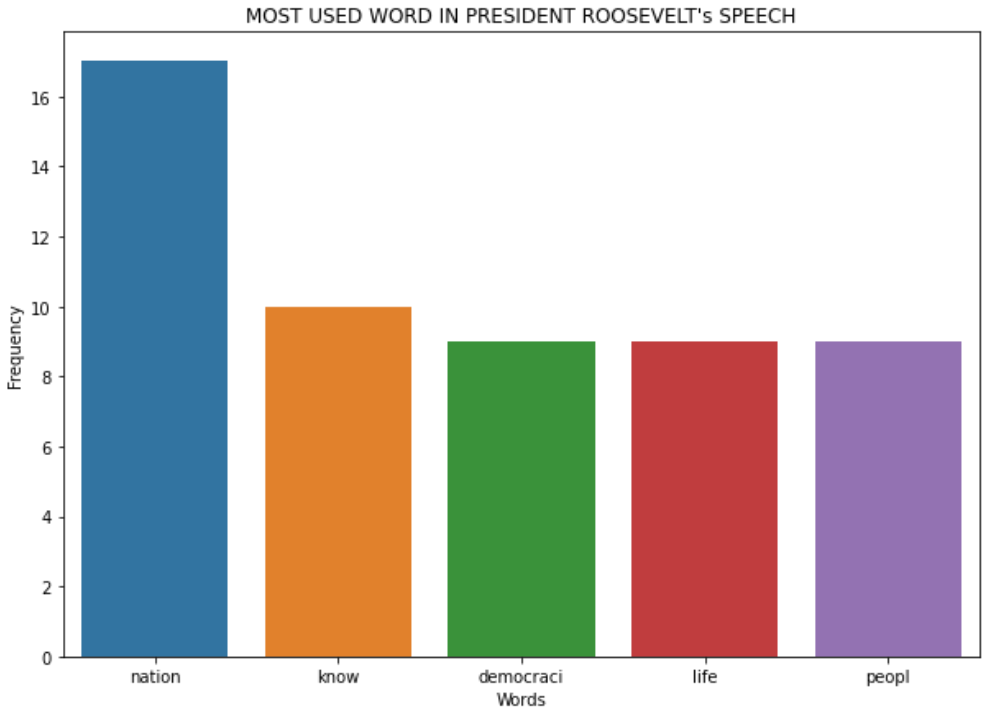


Figure 14 Top 5 Words President Roosevelt

- The top 3 words used in Roosevelt's Speech are Nation, know, democratic.
- If we ignore the word Know, then we can include life into top 3 followed by People

President Nixon Speech

	Words	Frequency
70	let	16
107	us	12
47	power	9
40	world	8
250	side	8

Table 21 Top 5 Words President Nixon

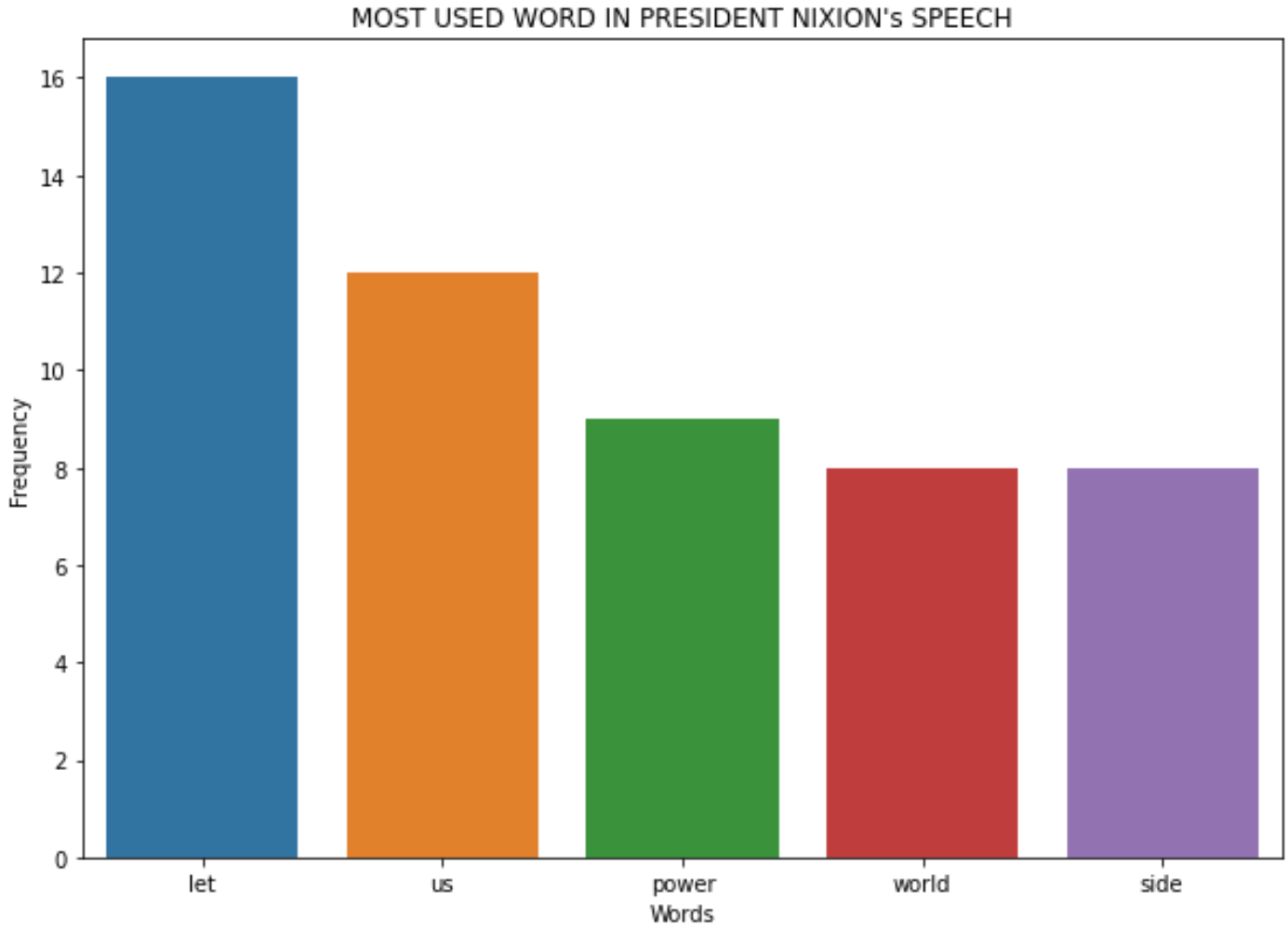


Figure 15 Top 5 Words President Nixon

- The top 3 words used in Nixon Speech are let, us, power.
- If we drop let and us, then we have power, world and side as top 3 frequent words.

President Kennedy Speech

	Words	Frequency
43	us	26
47	let	22
21	america	21
39	peac	19
40	world	18

Table 22 Top 5 Words President Kennedy

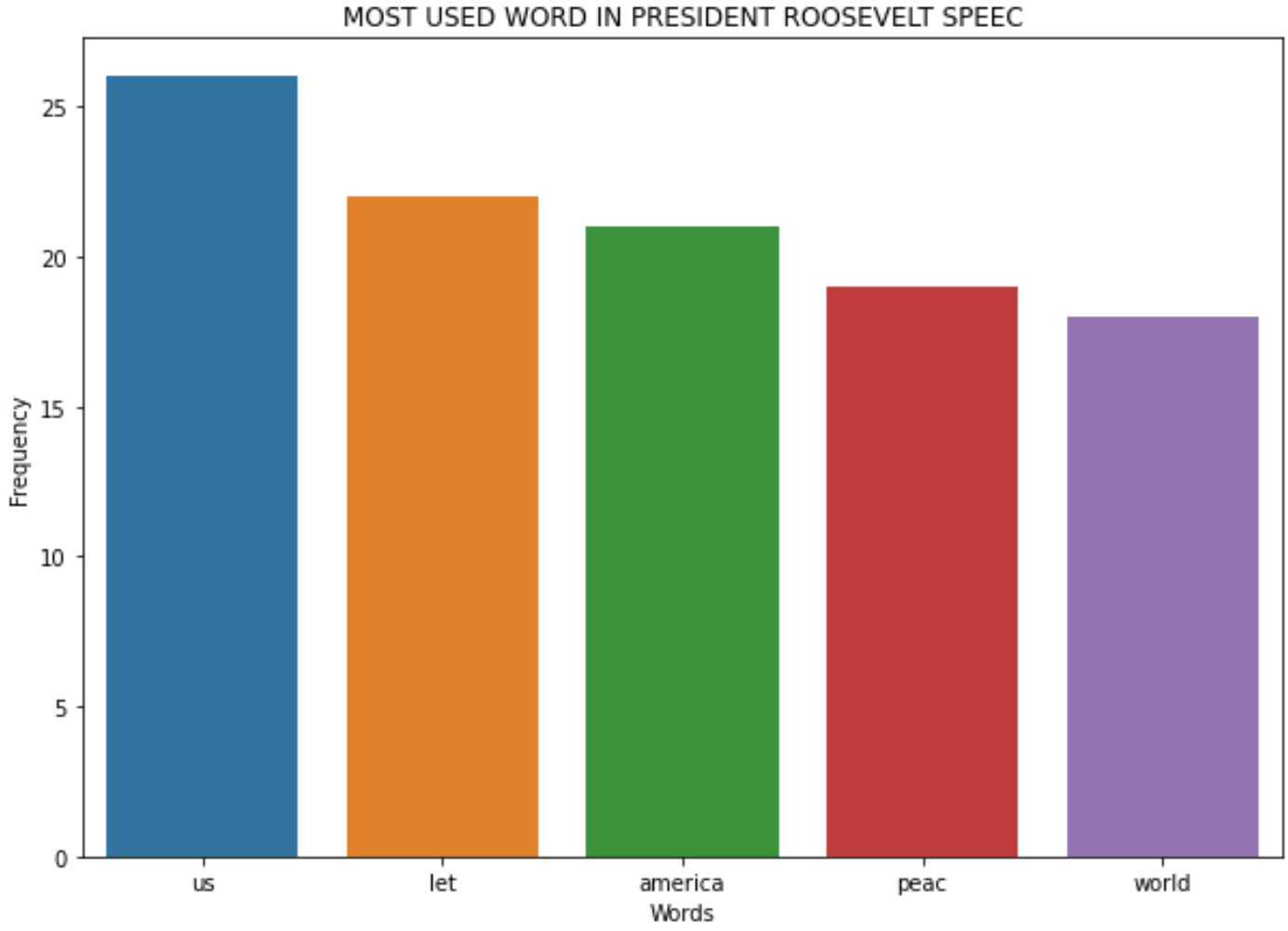


Figure 16 Top 5 Words President Kennedy

- The top 3 frequently used words in Kennedy Speech are us, let, America.
- If we drop us and let we have America, peace and world as most frequent words.

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

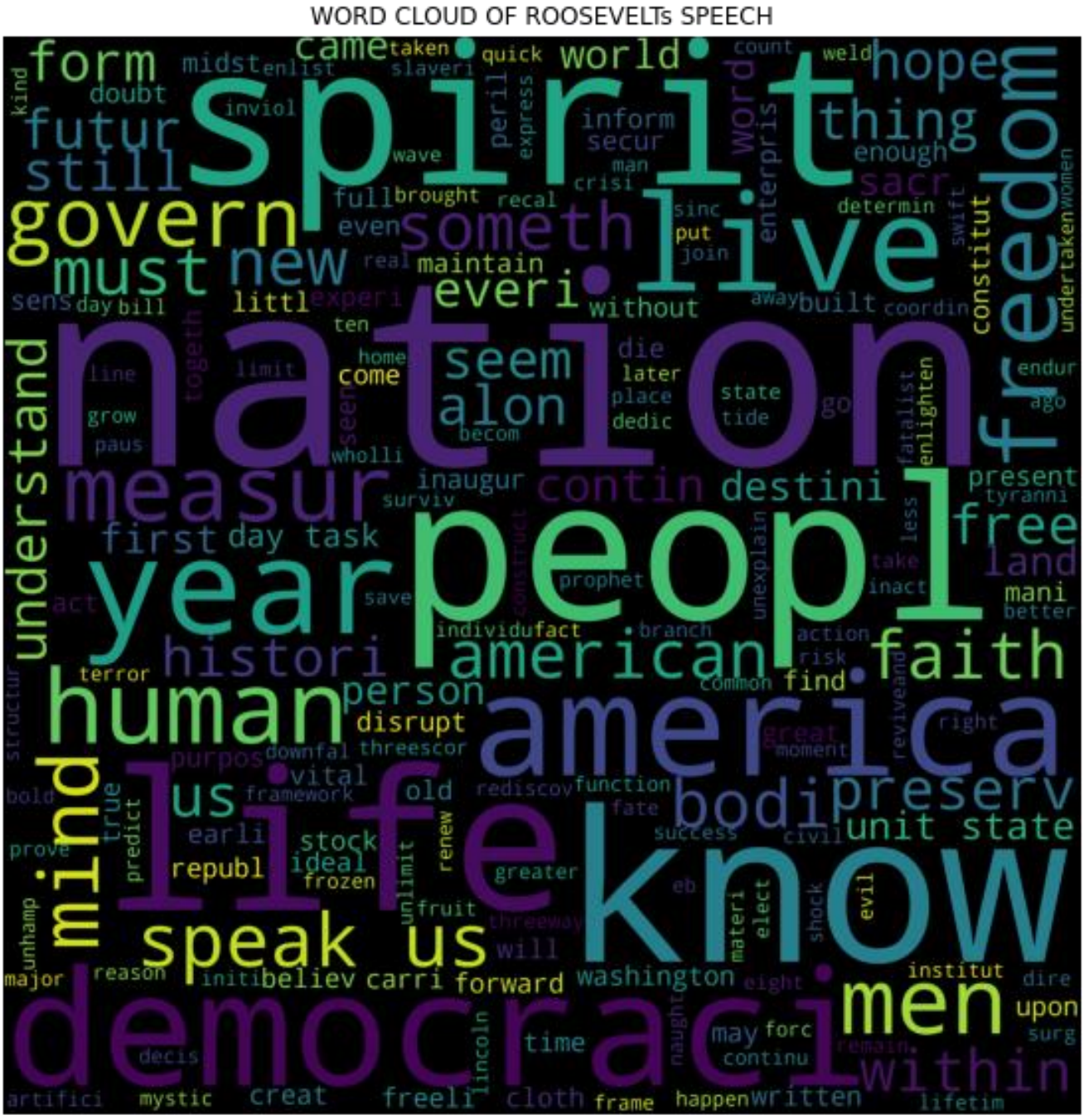


Figure 17 Word Cloud - President Roosevelt

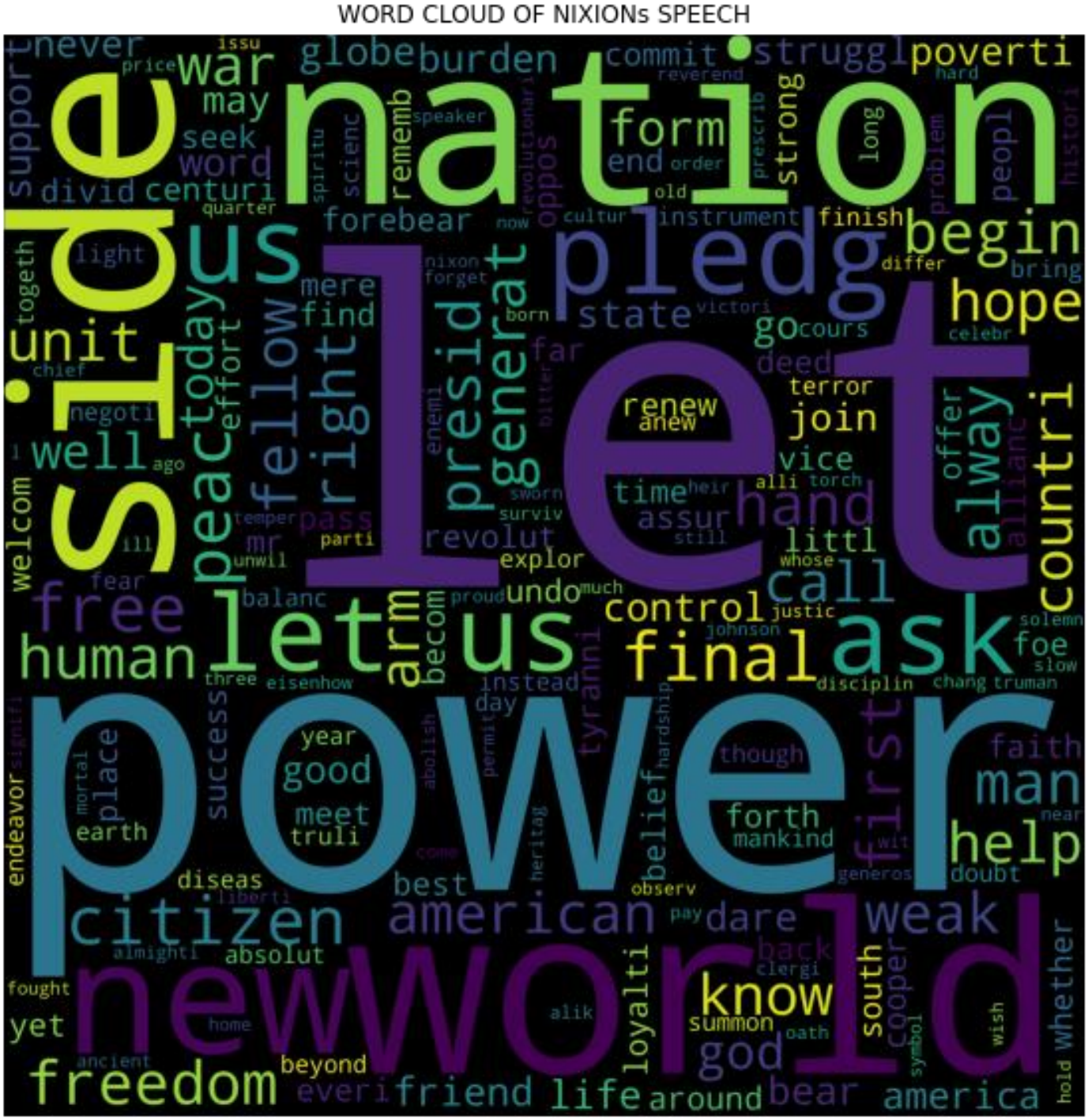


Figure 18 Word Cloud - President Nixon

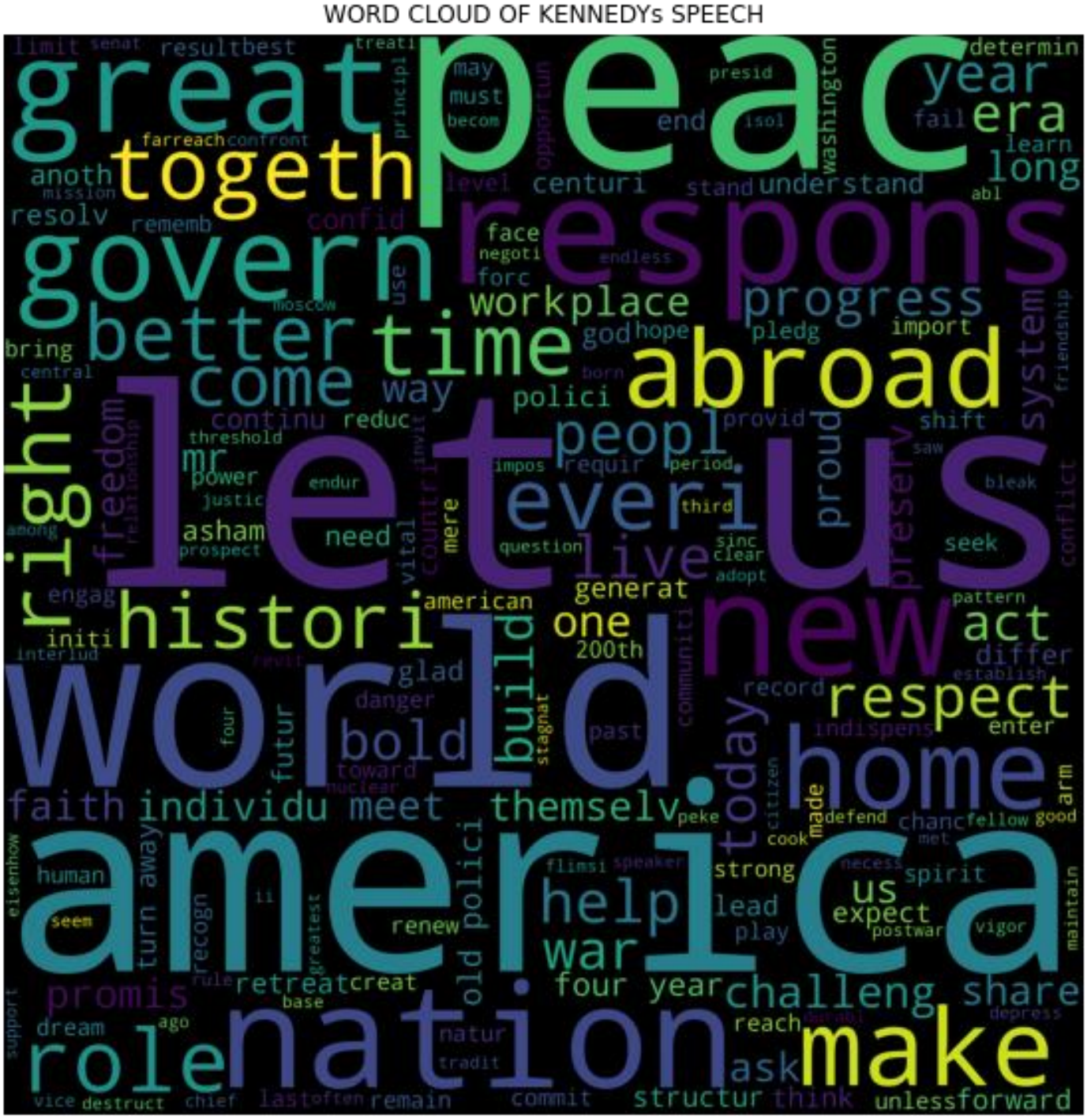


Figure 19 Word Cloud - President Kennedy

- From the word cloud as well we can see the most frequent word used in every president's speeches are same as before using Frequency Distribution.
- All three president's speech one commonly used frequent word "**nation**".
- **Nixon** and **Kennedy's** speech has **world** in common.