Machine Learning Project Report

Preetam Sarmah

PGPDSBA.O.JAN23.A
January' 23
Date: 02/07/2023

Table of Contents

List of Figures

References PreetamSarmah_02-July-2023.ipynb

Problem 1

Executive Summary

You are hired by one of the leading news channels CNBE who wants to analyze 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.

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.

There are 1525 rows and 9 columns in the dataset

The dataset has 1 dependent variable - vote which can take value of either Labour or Conservative and 8 predictor variables as follows age, economic.cond.national, economic.cond.household, Blair , Hague, Europe , Political Knowledge and gender.

```
Int64Index: 1525 entries, 1 to 1525
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
```

The Dataset has 7 numerical variables (int64) and 2 object variables as datatypes

Overview of first 5 records and last 5 records

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

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1521	Conservative	67	5	3	2	4	11	3	male
1522	Conservative	73	2	2	4	4	8	2	male
1523	Labour	37	3	3	5	4	2	2	male
1524	Conservative	61	3	3	1	4	11	2	male
1525	Conservative	74	2	3	2	4	11	0	female

	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

From the above data description its clear that there a slight positive skew in age criterion, as the mean is slightly greater than the median

The Other numerical variables are essentially, ratings on a scale of 1-5 , 1-11 and 1-3 etc

The people who vote for Labour party are dominant with 1063 of the 1525 entries and female voters are slightly greater than male voters (821/1525).

Also the dataset has no Null values , but has 8 duplicated records, which are removed as a part of data cleaning.

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. 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.

There are 7 int64 datatype variables and 2 object datatype variables

There are 1525 non null entries with 8 duplicated records and 9 columns

Univariate Analysis

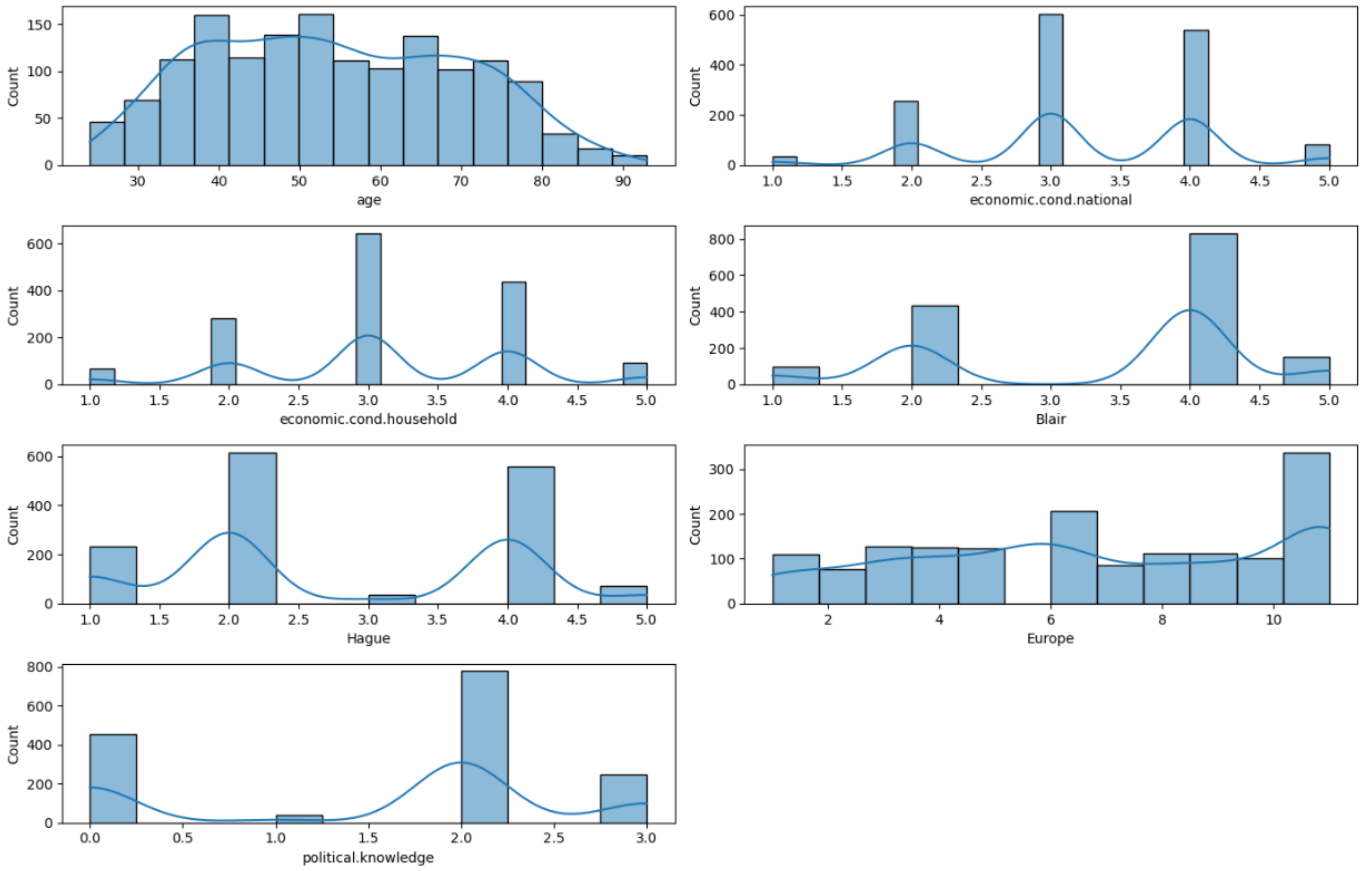

Fig 1 Univariate Analysis Numerical

- Most of the voters are in the age group of 41 to 67 year olds
- assessment of current national economic conditions and current household economic conditions follows a tri modal distribution with most giving an assessment score of 3/5 followed by 4/5 and 2/5 .
- Assessment for Labour Leader Blair is mostly 4/5 followed by 2/5
- Assessment for Conservative Leader Hague is mostly 2/5 followed by 4/5
- Respondents' attitudes toward European integration, is slightly on the positive side with about 50% of repondants rating (6/11)
- About 75 % of the respondents have 2/3 political knowledge

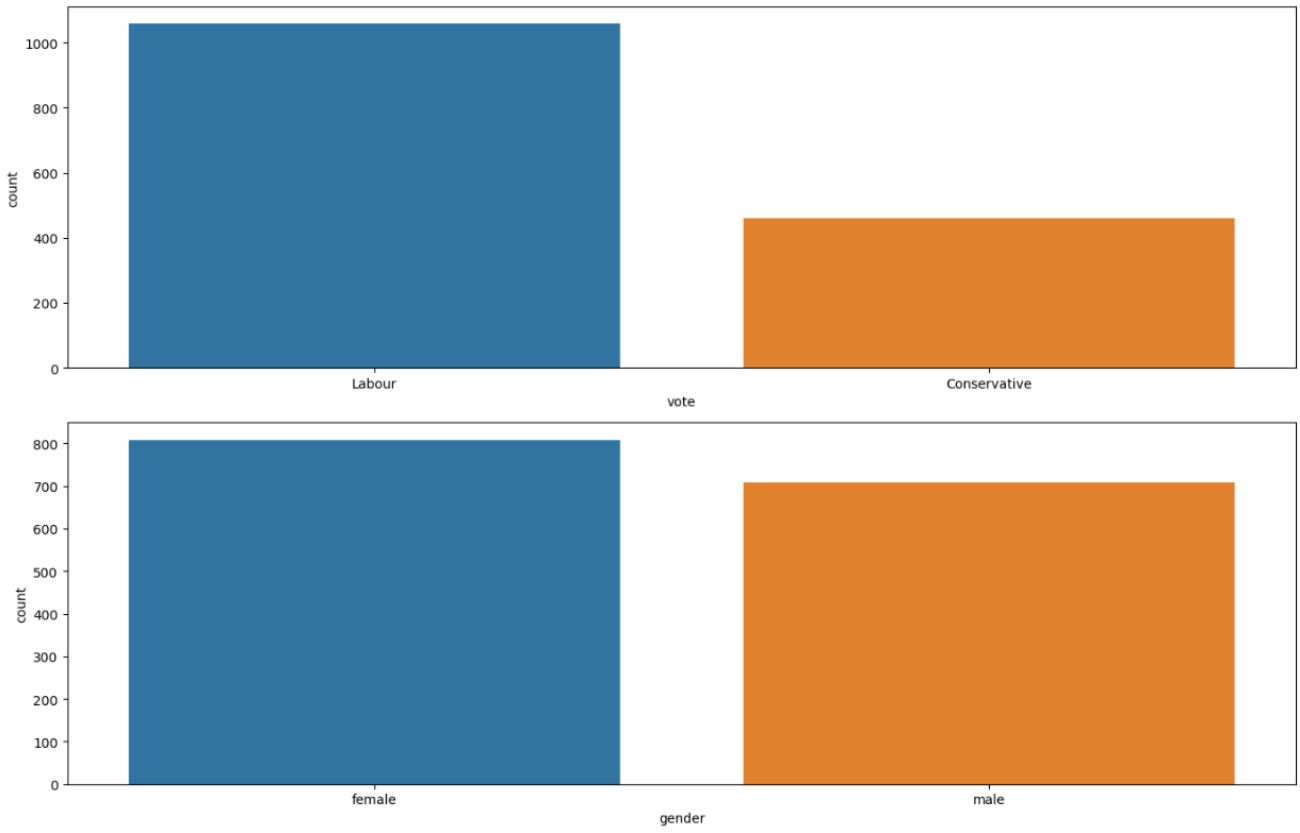

Fig 2 Univariate Analysis Categorical

1. Most respondants have party choice Labour and the number of female respondents is slightly more than number of male respondents.

BiVariate and Multivariate Analysis

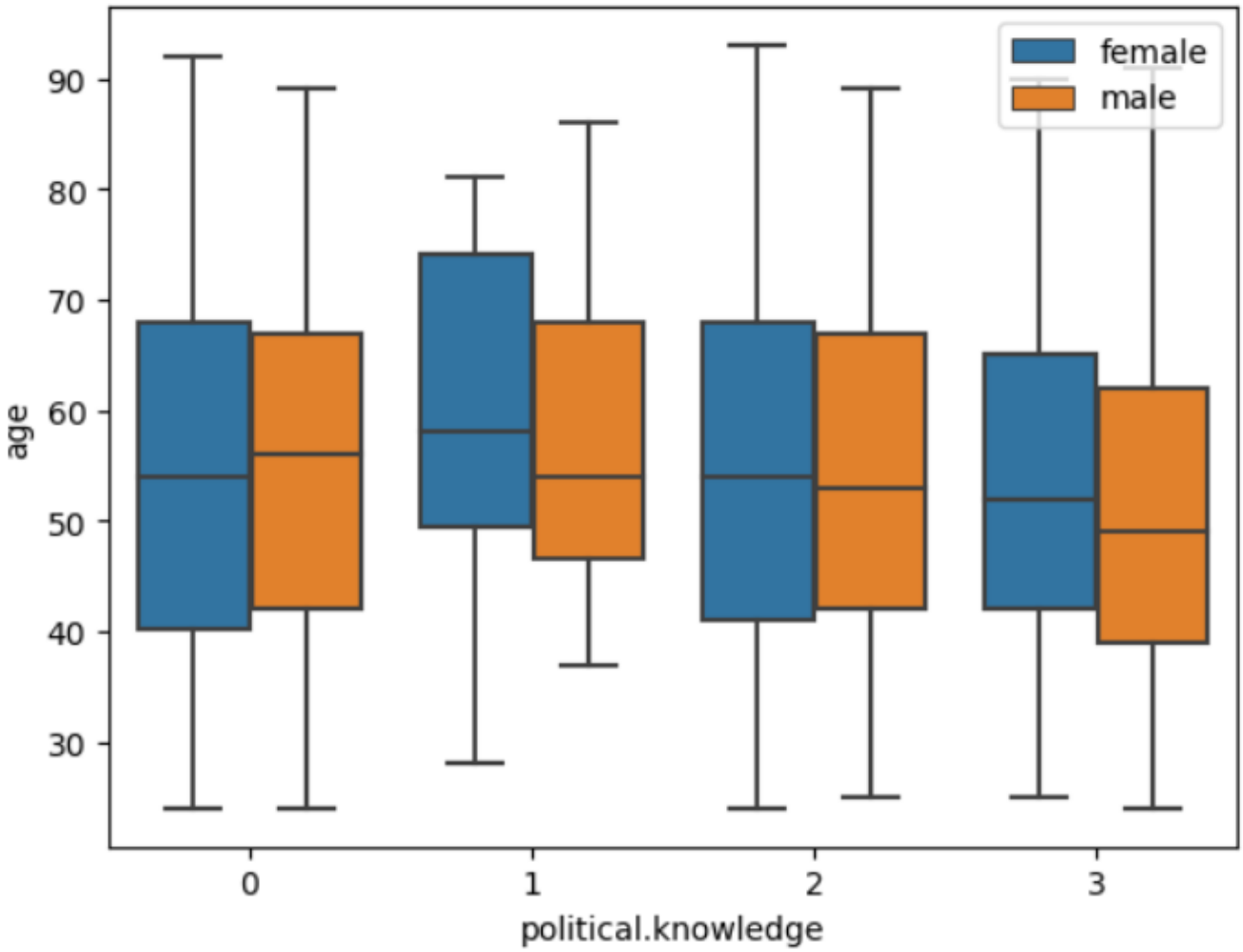

Fig 2 Multivariate Analysis

1. The age groups 41-67 show both 0/3 and 3/3 in terms of political knowlegde
2. Its also interesting to note that the youngest male respondants display the highest political knowledge.
3. Political knowledge is highest in respondant in late 40s and lowest in repondants in their mid 50s. male and female respondants show similar political knowledge in the age group of early 50s, while in general women seem to have more political knowledge across the board.

assessment of current national economic conditions and current household economic conditions have 1 as outliers , since this is a rating from 1-5 with 1 being a valid value , these are genuine outliers. **So no outlier treatment is necessary**

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.

1. Scaling is needed only for distance based algorithms such as KNN so i choose to have the original data frame as well as a scaled dataframe with the age attribute, scaled for the KNN algorithm.
2. I use z score based scaling which is $z = (x – \mu) / \sigma$, which gives an idea of how far from the mean a data point is.
3. I decide to go with dummy encoding as vote and gender are nomial
4. The Dataset is split such that 70% data split into train data and 30% into test data, we train the model on train data and predict on the test data
5. Age has a mean of around 54 and std dev 15 , so some respondent with age 55 will have a z score of 0.067

1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting).

Logistic Regression

Test Data Accuracy 0.8530701754385965

Train Data Accuracy 0.8284637134778511

	precision	recall	f1-score	support
0	0.81	0.67	0.74	138
1	0.87	0.93	0.90	318
accuracy			0.85	456
macro avg	0.84	0.80	0.82	456
weighted avg	0.85	0.85	0.85	456

Default parameters are used with the Logistic Regression model

The model has very good precision > 0.80

The difference between Test Data Accuracy and Train Data accuracy are under 3% which implies no significant Under Fitting or Over Fitting

The recall is a bit bad for Conservative party voters class at 0.67 , where as its very good for Labour party voters at 0.93.

LDA

Test Data Accuracy 0.8530701754385965

Train Data Accuracy 0.822808671065033

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

Default parameters are used with the LDA model

The model has very good precision > 0.80

The difference between Test Data Accuracy and Train Data accuracy are under 3% which implies no significant Under Fitting or Over Fitting

The recall is a bit bad for Conservative party voters class at 0.69 , where as its very good for Labour party voters at 0.92.

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 should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting).

KNN

Train Data Accuracy 0.8294062205466541

Test Data Accuracy 0.8552631578947368

	precision	recall	f1-score	support
0	0.79	0.72	0.75	138
1	0.88	0.92	0.90	318
accuracy			0.86	456
macro avg	0.83	0.82	0.82	456
weighted avg	0.85	0.86	0.85	456

1. As the difference between train and test accuracies is less than 10%, it is a valid model, no over fitting or under fitting.
2. Number of Neighbours is set to 31 as it results in the least missclassification error and weights is chosen as uniform which is default as when distance is used as weight metric, there is an issue with overfitting.
3. KNN being a distance based algorithmn, scaled dataset is used.
4. The Recall for conservative party voters have slightly improved to 79, and precision is good as well

Naive Bayes

Train Data Accuracy 0.8199811498586239

Test Data Accuracy 0.8574561403508771

	precision	recall	f1-score	support
0	0.79	0.72	0.75	138
1	0.88	0.92	0.90	318
accuracy			0.86	456
macro avg	0.84	0.82	0.83	456
weighted avg	0.86	0.86	0.86	456

1. As the difference between train and test accuracies is less than 10%, it is a valid model, no over fitting or under fitting.
2. The recall for conservative voters have slightly degraded .
3. The precision and recall for the Labour party voters is very good , close to 0.90.

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. 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.

Model Tuning

LDA

Best Parameters : {'shrinkage': None, 'solver': 'svd', 'store_covariance': True}

Test Data accuracy 0.8530701754385965

Logistic Regression

Best Parameters : {'C': 100, 'penalty': 'l1', 'solver': 'liblinear'}

Test Data accuracy 0.8552631578947368

KNN

Best Parameters : {'metric': 'manhattan', 'n_neighbors': 14, 'weights': 'uniform'}

Test Data accuracy 0.8618421052631579

Random Forest Classifier

Best Parameters : {'max_features': 'log2', 'n_estimators': 10}

Test Data accuracy 0.8267543859649122 and train data accuracy is around 0.985, hence it indicates that there is over fitting , ie the model learned from the noise in data also.

Bagging

Train Data Accuracy 0.9679547596606974

Test Data Accuracy 0.8508771929824561

	precision	recall	f1-score	support
0	0.79	0.69	0.74	138
1	0.87	0.92	0.90	318
accuracy			0.85	456
macro avg	0.83	0.80	0.82	456
weighted avg	0.85	0.85	0.85	456

Boosting

Ada Boost

Train Data Accuracy 0.8397737983034873

Test Data Accuracy 0.8355263157894737

	precision	recall	f1-score	support
0	0.76	0.67	0.71	138
1	0.86	0.91	0.88	318
accuracy			0.84	456
macro avg	0.81	0.79	0.80	456
weighted avg	0.83	0.84	0.83	456

Gradient Boosting

Train Data Accuracy 0.885956644674835

Test Data Accuracy 0.8421052631578947

	precision	recall	f1-score	support
0	0.77	0.69	0.73	138
1	0.87	0.91	0.89	318
accuracy			0.84	456
macro avg	0.82	0.80	0.81	456
weighted avg	0.84	0.84	0.84	456

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)

Naive Bayes

Train Data confusion matrix

[[226 96]

[95 644]]

Test Data Confusion matrix

[[100 38]

 [27 291]]

AUC for the Training Data: 0.873

AUC for the Test Data: 0.912

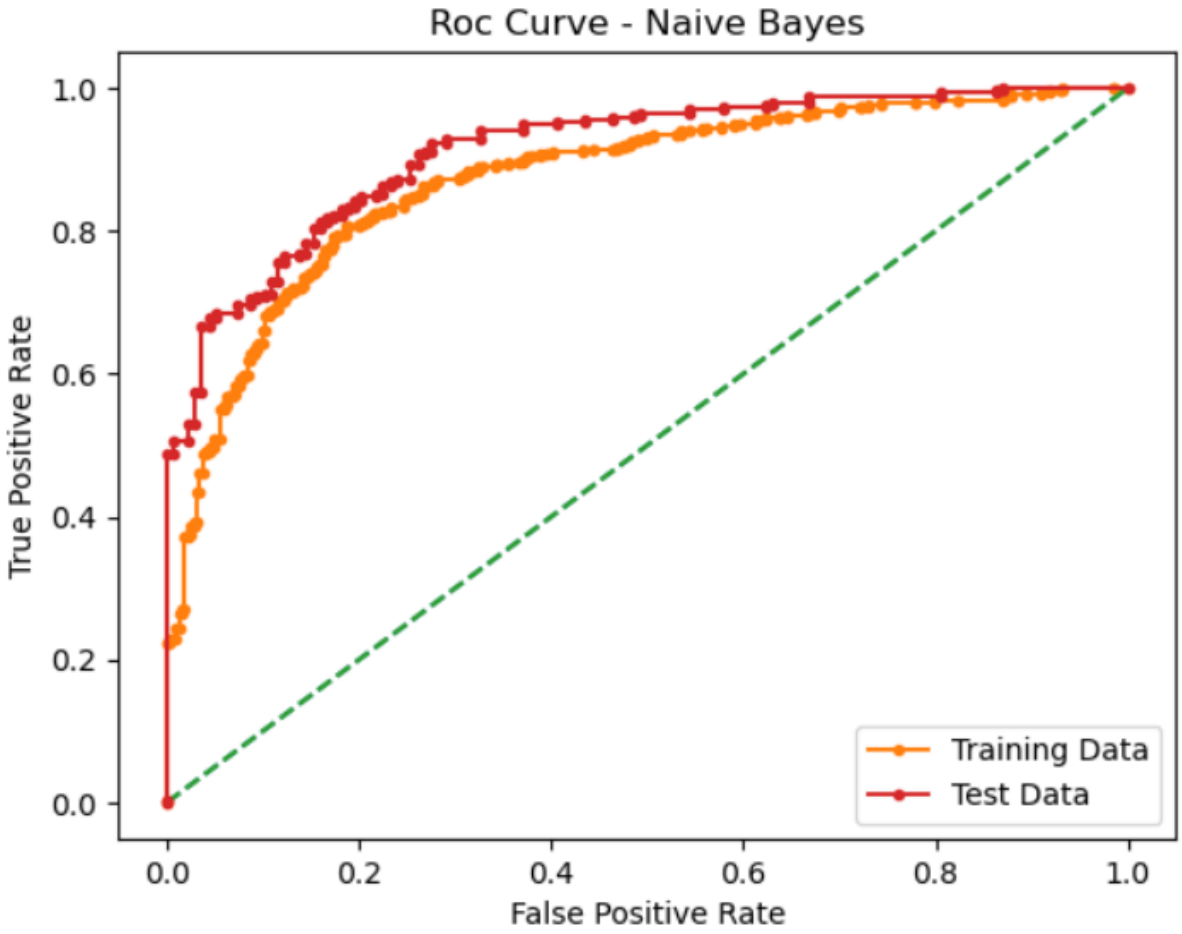

Fig 4 ROC Curve Naive Bayes

KNN

Train Data Confusion matrix

[[235 87]

 [82 657]]

Test Data Confusion matrix

[[108 30]

 [33 285]]

AUC for the Training Data: 0.905

AUC for the Test Data: 0.907

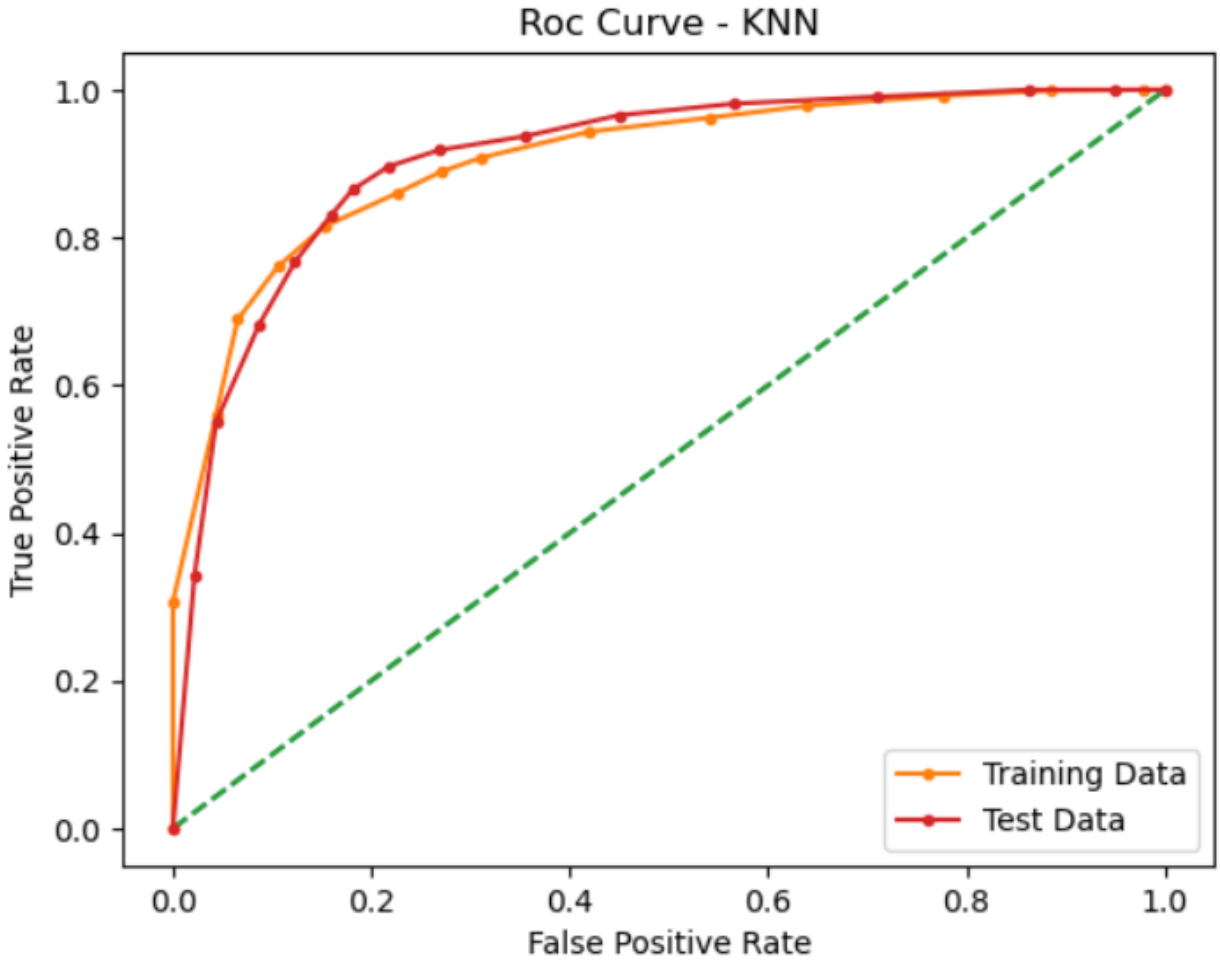

Fig 5 ROC Curve KNN

LDA

Train Data confusion matrix

[[217 105]

 [83 656]]

Test Data confusion matrix

[[95 43]

[24 294]]

AUC for the Training Data: 0.877

AUC for the Test Data: 0.914

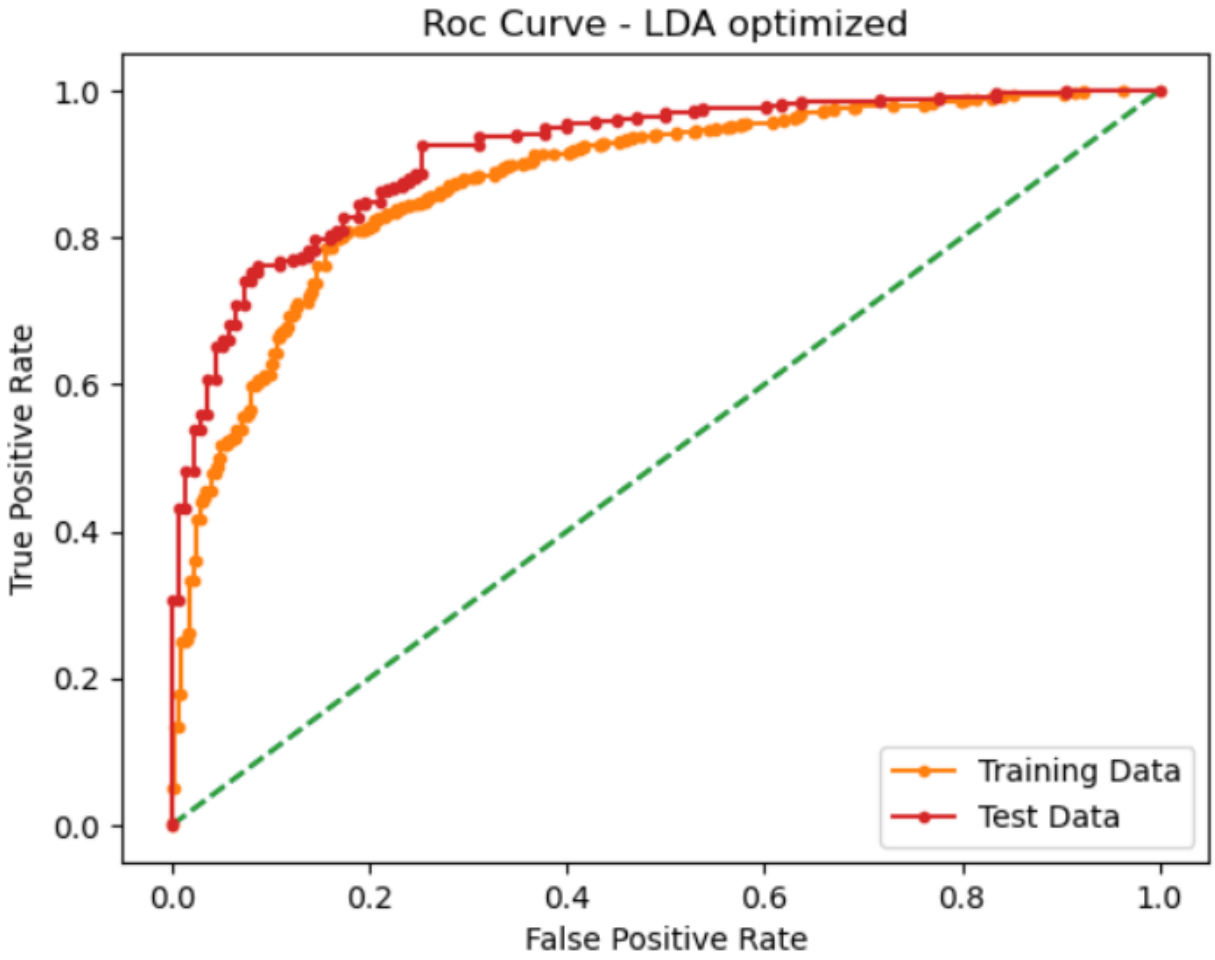

Fig 6 ROC Curve LDA

Logistic Regression

Train Data Confusion Matrix

[[213 109]

[73 666]]

Test Data Confusion Matrix

[[93 45]

[22 296]]

AUC for the Training Data: 0.877

AUC for the Test Data: 0.913

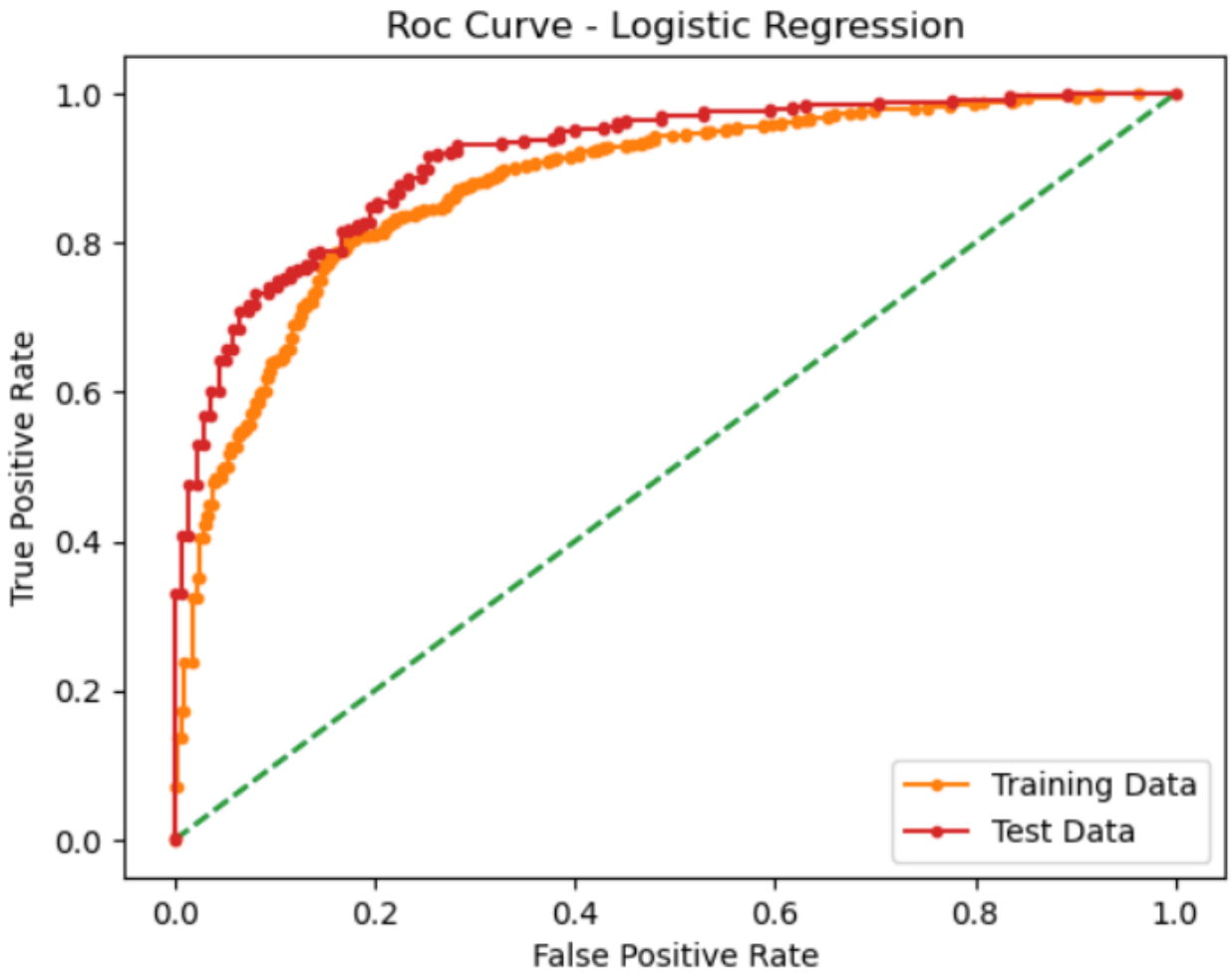

Fig 7 ROC Curve Logistic Regression

Train Data Summary

	Accuracy	Precision	Recall	F1 score	AUC
Naive Bayes Train	0.819981	0.870000	0.870000	0.870000	0.870000
Logistic Regression Train	0.828464	0.860000	0.900000	0.880000	0.880000
KNN Train	0.840716	0.880000	0.890000	0.890000	0.910000
LDA Train	0.822809	0.860000	0.890000	0.870000	0.880000

- The Accuracy,precision,AUC score,F1 score is best for KNN model

- Recall is the best for Logistic Regression model

Test Data Summary

	Accuracy	Precision	Recall	F1 score	AUC
Naive Bayes Test	0.857456	0.880000	0.920000	0.900000	0.910000
Logistic Regression Test	0.853070	0.870000	0.930000	0.900000	0.910000
KNN Test	0.861842	0.900000	0.900000	0.900000	0.910000
LDA Test	0.853070	0.870000	0.920000	0.900000	0.910000

- KNN model has the best accuracy , precision
- F1 score nd AUC score is pretty much same across all the models
- Logistic Regression model has the best recall

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.

- The younger age group tends to have high political knowledge
- The female voters who have high ‘Eurosceptic’ sentiment tend to vote for conservatives, so Hague as better chance
- The male voters , tend to have lower eurosceptic sentiment and tend to vote for Labour party, and Blair
- Blair is definitely the more popular candidate with a majority giving him 4/5 ratings , where as Hague is equally likely to receive 2/5 and 4/5 rating.

Problem 2

Introduction

inaugural corpora has the following speeches of the Presidents of the United States of America:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

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

President Franklin D. Roosevelt in 1941 speech

- Number of Words : 1536
- Number of Characters : 6174
- Number of Sentences : 68

President John F. Kennedy in 1961 speech

- Number of Words : 1546
- Number of Characters : 6202
- Number of Sentences : 52

President Richard Nixon in 1973 speech

- Number of Words : 2028
- Number of Characters : 8122
- Number of Sentences : 69

inaugural. words(), takes into consideration the total count of all words, but the split functions take the spaces count also into consideration, so the count may differ. We choose to use inaugural. words(),

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.

	filename	text	WordCount_Before	WordCount_After
0	1941-Roosevelt.txt	national day inauguration since 1789, people r...	1360	644
1	1961-Kennedy.txt	vice president johnson, mr. speaker, mr. chief...	1390	705
2	1973-Nixon.txt	mr. vice president, mr. speaker, mr. chief jus...	1819	844

Sample Sentence (after stopwords removal)

1941-Roosevelt speech : ' national day inauguration since 1789, people renewed sense dedication united states. washington's day task people create weld together nation. lincoln's day task people preserve nation disruption within. day task people save nation institutions disruption without. us come time, midst swift happenings, pause moment take stock recall place history been, rediscover may be. not, risk real peril inaction. lives nations determined count years, lifetime human spirit. life man three-score years ten: little more, little less. life nation fullness measure live. men doubt this. men believe democracy, form government frame life, limited measured kind mystical artificial fate that, unexplained reason, tyranny slavery become surging wave future freedom ebbing tide. americans know true. eight years ago, life republic seemed frozen fatalistic terror, proved true. midst shock acted. acted quickly, boldly, decisively. later years living years fruitful years people democracy. brought us greater security and, hope, better understanding life's ideals measured material things. vital present future experience democracy successfully survived crisis home; put away many evil things; built new structures enduring lines; and, all, maintained fact democracy. action taken within three-way framework constitution united states. coordinate branches government continue freely function. bill rights remains inviolate. freedom elections wholly maintained. prophets downfall american democracy seen dire predictions come naught. democracy dying. know seen revive--and grow. know cannot die built unhampered initiative individual men women joined together common enterprise enterprise

undertaken carried free expression free majority. know democracy alone, forms government, enlists full force men's enlightened will. know democracy alone constructed unlimited civilization capable infinite progress improvement human life. know because, look surface, sense still spreading every continent humane, advanced, end unconquerable forms human society. nation, like person, body--a body must fed clothed housed, invigorated rested, manner measures objectives time. nation, like person, mind mind must kept informed alert, must know itself, understands hopes needs neighbors nations live within narrowing circle world. nation, like person, something deeper, something permanent, something larger sum parts. something matters future calls forth sacred guarding present. thing find difficult even impossible hit upon single, simple word. yet understand spirit faith america. product centuries. born multitudes came many lands high degree, mostly plain people, sought here, early late, find freedom freely. democratic aspiration mere recent phase human history. human history. permeated ancient life early peoples. blazed anew middle ages. written magna charta. americas impact irresistible. america new world tongues, peoples, continent new-found land, came believed could create upon continent new life life new freedom. vitality written mayflower compact, declaration independence, constitution united states, gettysburg address. first came carry longings spirit, millions followed, stock sprang moved forward constantly consistently toward ideal gained stature clarity generation. hopes republic cannot forever tolerate either undeserved poverty self-serving wealth. know still far go; must greatly build security opportunity knowledge every citizen, measure justified resources capacity land. enough achieve purposes alone. enough clothe feed body nation, instruct inform mind. also spirit. three, greatest spirit. without body mind, men know, nation could live. spirit america killed, even though nation's body mind, constricted alien world, lived on, america know would perished. spirit faith speaks us daily lives ways often unnoticed, seem obvious. speaks us capital nation. speaks us processes governing sovereignties 48 states. speaks us counties, cities, towns, villages. speaks us nations hemisphere, across seas enslaved, well free. sometimes fail hear heed voices freedom us privilege freedom old, old story. destiny america proclaimed words prophecy spoken first president first inaugural 1789 words almost directed, would seem, year 1941: "the preservation sacred fire liberty destiny republican model government justly considered deeply, finally, staked experiment intrusted hands american

people." lose sacred fire--if let smothered doubt fear shall reject destiny washington strove valiantly triumphantly establish. preservation spirit faith nation does, will, furnish highest justification every sacrifice may make cause national defense. face great perils never encountered, strong purpose protect perpetuate integrity democracy. muster spirit america, faith america. retreat. content stand still. americans, go forward, service country, god.',

Steps:

1. Get the prior word count using string.split()
2. Conversion of the speech text to lower case
3. The Stopwords include Stopwords corpus + list of punctions of the String class eg: on , a, the, ; , ! etc
4. The text is stemmed, ie its trimmed upto the rootword
5. The word count post the above cleanup.

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).

top 3 words for 1941-Roosevelt speech

Word	Frequency
know	9
us	8
life	6

top 3 words for 1961-Kennedy speech

Word	Frequency
let	16
us	11
new	7

top 3 words for 1973-Nixon speech

Word	Frequency
us	25
let	22
new	15

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

World cloud is a visual representation to skim the important words , buzz words of a certain text content

Word cloud for the president speech are as follows …..

Word Cloud for 1941-Roosevelt speech

Fig 8 Word Cloud for 1941-Roosevelt speech

Word Cloud for 1961-Kennedy speech

Fig 9 Word Cloud for 1961-Kennedy speech

Word Cloud for 1973-Nixon speech

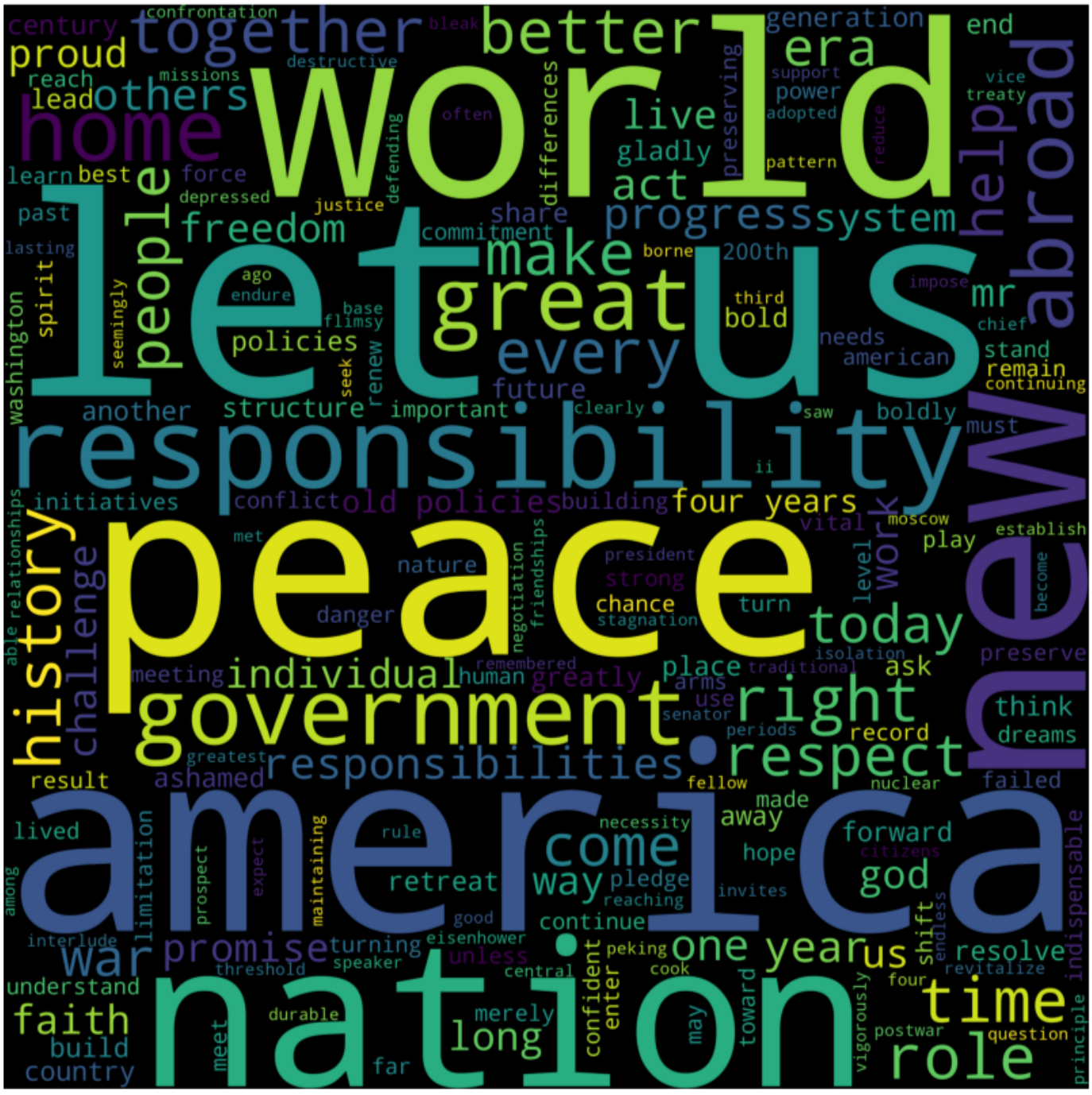

Fig 10 Word Cloud for 1973-Nixon speech