MACHINE LEARNING
PROJECT - 6 REPORT

Piyush Kumar Singh
PGP – DSBA Online
May-21 Batch

Date: 5/12/2021

Table of Contents

List of Figures

List of Tables

Problem 1:

Executive Summary

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

Data Dictionary
1. vote: Party choice: Conservative or Labour
2. age: in years
3. economic.cond.national: Assessment of current national economic conditions, 1 to 5.
4. economic.cond.household: Assessment of current household economic conditions, 1 to 5.
5. Blair: Assessment of the Labour leader, 1 to 5.
6. Hague: Assessment of the Conservative leader, 1 to 5.
7. Europe: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
8. political.knowledge: Knowledge of parties' positions on European integration, 0 to 3.
9. gender: female or male.

Table.1

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.

Dataset Sample

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

Table. 2

- I am dropping 'Unnamed:0' column here itself as it is just index numbers and will not contribute to anything in model analysis.

Dataset Info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
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)
memory usage: 107.4+ KB
```

Table. 3

- The given dataset has now 9 columns out of following datatypes:
  - Int64 type - 7 column
  - Object type - 2 column
- Dataset has 1525 rows and 9 column

Summary of Dataset

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

Table. 4

- We can see that there are 2 object type columns so they need to encoded to numerical type for analysis
- Now as per dataset dictionary only ‘age’ column needs to be treated as numeric rest all the ‘int64’ type columns are numeric representation of categorical values.

Checking for missing values in dataset

```
vote                        0
age                         0
economic.cond.national      0
economic.cond.household     0
Blair                       0
Hague                       0
Europe                      0
political.knowledge         0
gender                      0
dtype: int64
```

Table. 5

- There are no missing values in the dataset.

Checking for entries’ with 0 in dataset

```
Column Name: vote - 0
Column Name: age - 0
Column Name: economic.cond.national - 0
Column Name: economic.cond.household - 0
Column Name: Blair - 0
Column Name: Hague - 0
Column Name: Europe - 0
Column Name: political.knowledge - 455
Column Name: gender - 0
```

Table. 6

- Here we are checking the dataset for entries of value = 0. As per data dictionaries of the dataset values for column ‘political.knowledge’ can take 0 value as it is a category type representation of this column.
- So we are ignoring this as only ‘political.knowledge’ column has this value.

Checking for duplicates in dataset

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
67	Labour	35	4	4	5	2	3	2	male
626	Labour	39	3	4	4	2	5	2	male
870	Labour	38	2	4	2	2	4	3	male
983	Conservative	74	4	3	2	4	8	2	female
1154	Conservative	53	3	4	2	2	6	0	female
1236	Labour	36	3	3	2	2	6	2	female
1244	Labour	29	4	4	4	2	2	2	female
1438	Labour	40	4	3	4	2	2	2	male

Table. 7

- There are 8 duplicate entries in the dataset, I am dropping all of these duplicate records.

Skewness of columns :

Numeric Type Columns – age

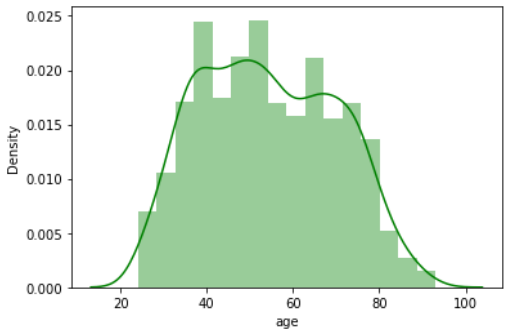


Fig. 1

As per the histogram age column is normally distributed.

Categorical Type Columns - economic.cond.national, economic.cond.household, Blair, Hague, Europe, political.knowledge

All these are categorical columns with numeric representation of their categorical levels so there is no point in checking the distribution of such columns.

We can use count plot to check the count of each levels.

Object Type Columns – vote, gender

These are object type columns and need to be encoded to numeric values before model building can be done on them

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

Exploratory Data Analysis

Univariate Analysis

We perform the univariate analysis on the data set and display distplot to check distribution for each continuous type column and use boxplot to check for outliers if any.

Continuous Type Columns

age –

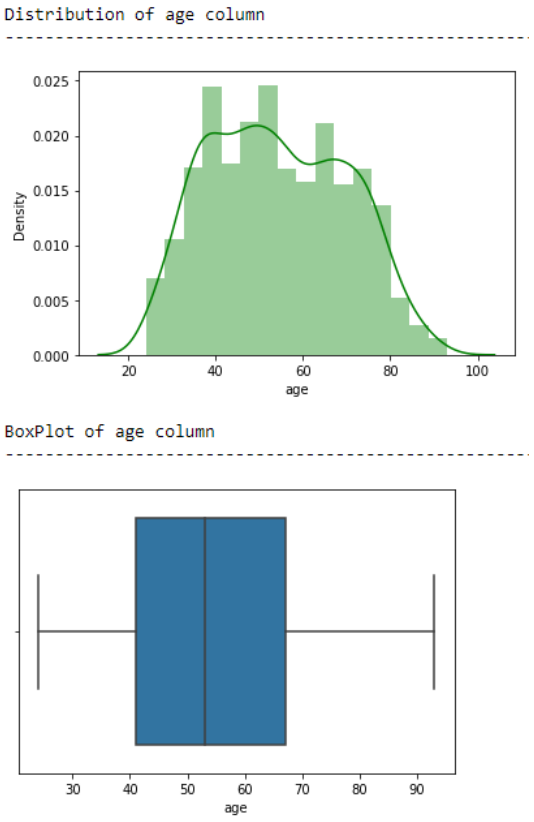


Fig.2

In our dataset only 'age' column is numerical.

- The mean number of age is 54.18 whereas the SD is 15.71.
- The maximum value is 93 and min value is 24
- The distribution is normal.
- Outliers are not present.

Categorical Type Columns

economic.cond.national -

```
3    604
4    538
2    256
5     82
1     37
Name: economic.cond.national, dtype: int64
------------------------------------------------
```

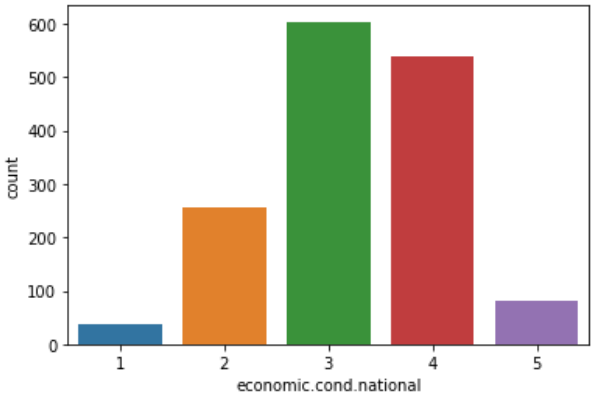


Fig.3

- Current national economic conditions has 5 levels. 1 being lowest and 5 being highest.
- People have assessed Level 3 most number of times and Level 1 lowest.

economic.cond.household -

```
3    645
4    435
2    280
5     92
1     65
Name: economic.cond.household, dtype: int64
------------------------------------------------
```

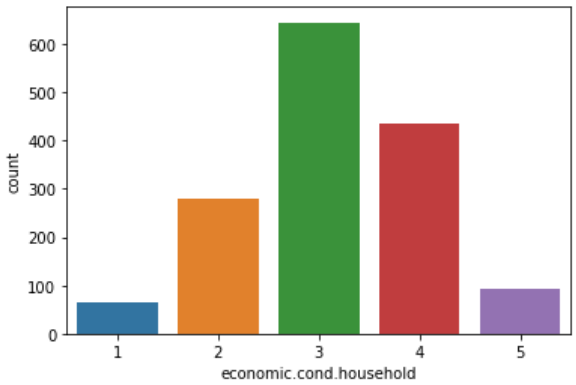


Fig.4

- Current household economic conditions has 5 levels. 1 being lowest and 5 being highest.
- People have assessed Level 3 most number of times and Level 1 lowest.

Blair -

```
4    833
2    434
5    152
1     97
3      1
Name: Blair, dtype: int64
------------------------------------------------
```

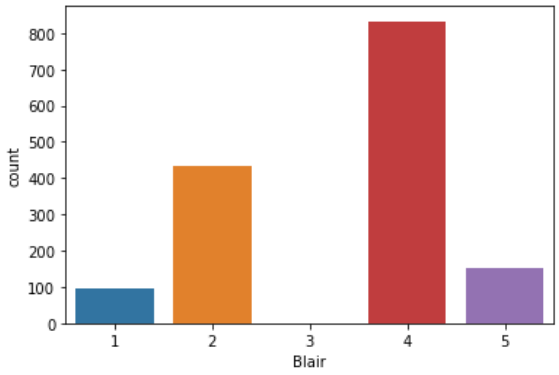


Fig.5

- Assessment of Blair as a labour leader has 5 levels. 1 being lowest and 5 being highest.
- People have assessed Blair at Level 4 most number of times and Level 3 lowest.

Hague -

```
2    617
4    557
1    233
5     73
3     37
Name: Hague, dtype: int64
------------------------------------------------
```

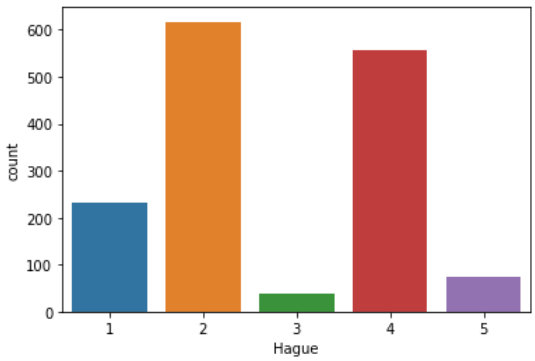


Fig.6

- Assessment of Hague as a Conservative leader has 5 levels. 1 being lowest and 5 being highest.
- People have assessed Hague at Level 2 most number of times and Level 3 lowest.

Europe -

```
11    338
6     207
3     128
4     126
5     123
9     111
8     111
1     109
10    101
7      86
2      77
Name: Europe, dtype: int64
------------------------------------------------
```

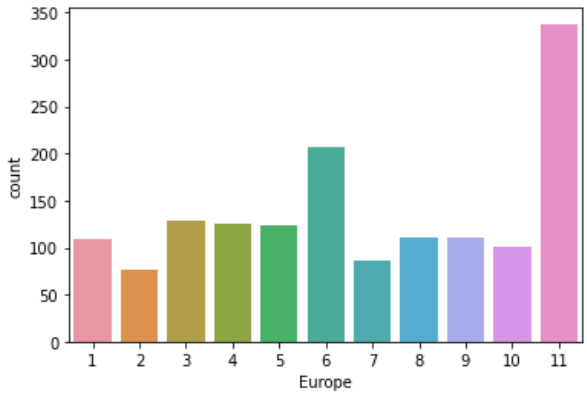


Fig.7

- An 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment.
- Most number of people have ‘Eurosceptic’ sentiment as seen by our count graph. 338 people have given a scale of 11.
- Only 77 people have selected scale of 2 which is lowest among all other scales, for their attitudes toward European integration.

political.knowledge -

```
2    776
0    454
3    249
1     38
Name: political.knowledge, dtype: int64
------------------------------------------------
```

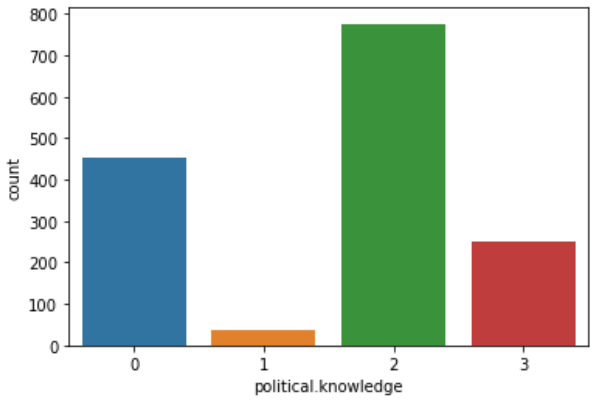


Fig.8

- Knowledge of parties' positions on European integration has 4 levels.
- People with level 2 are highest in number and level 1 are lowest.

gender –

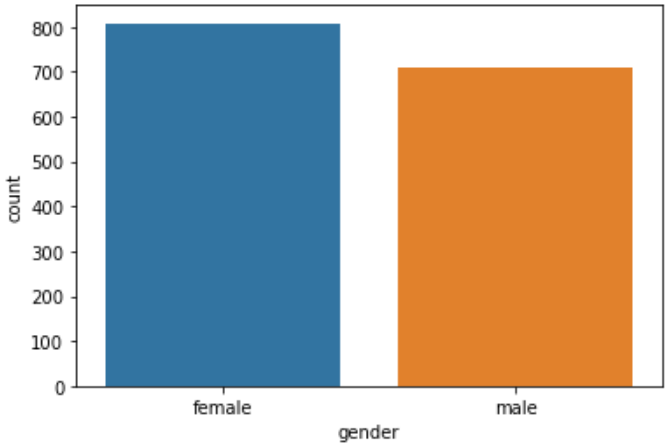


Fig.9

- Number of females are more than the number of males in our dataset.

vote – Our Target Variable

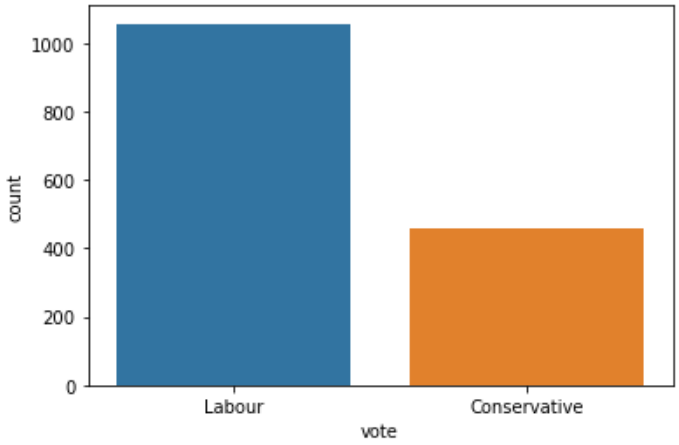


Fig.10

- Vote is our target variable to predict. There are two parties to vote for Labour and Conservative.
- People who have voted for Labour party are more than people who voted for Conservative party.
- Percentage of people who voted for Labour party – 69.67%
  Percentage of people who voted for Conservative party – 30.32%
- Our target variable is not well balanced and we can ask for more records to balance the variable. For the given dataset the probability of predicting Labour for an unknown record is higher.
- I have opted not to perform SMOTE technique to balance our target variable because even by applying SMOTE I was not getting significant improvement in the prediction results of different models that I have performed.

Checking for Outliers

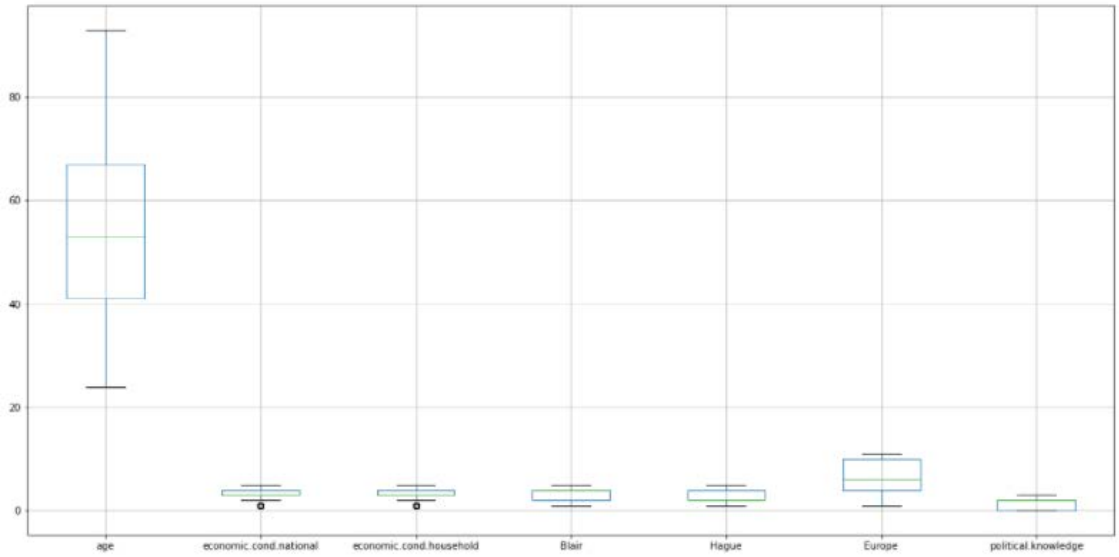


Fig.11

- Only 2 columns has outliers but we are ignoring them because these columns are categorical in nature. Outlier treatment is not suggested for categorical columns for any type machine learning analysis.

Bi/Multivariate Analysis

For Bi variant analysis we compare our target variable with every other variable in our dataset to find out hidden details in the dataset.

For our dataset Vote is our target variable.

Categorical Vs Numeric Analysis:

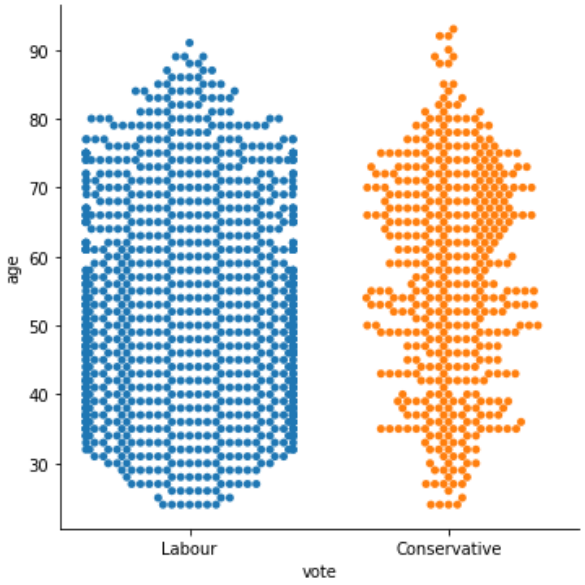


Fig.12

- Here we are comparing between vote and age variable.
- Most people have opted for Labour party across various age groups.
- Only pattern that I can see is that people above the age of 80 yrs. are opting for Labour party.

Categorical Vs Categorical Analysis:

For this type analysis I am using crosstab to find distribution of different levels in each category with respect to the party people are voting for.

Vote & economic.cond.national

economic.cond.national vote	1	2	3	4	5	All
Conservative	21	140	199	91	9	460
Labour	16	116	405	447	73	1057
All	37	256	604	538	82	1517

Table.8

- People from both the party are opting for level 3 mostly.
- People have opted level 1 least number of times.

Vote & economic.cond.household

economic.cond.household vote	1	2	3	4	5	All
Conservative	28	126	197	86	23	460
Labour	37	154	448	349	69	1057
All	65	280	645	435	92	1517

Table.9

- People from both the party are opting for level 3 mostly.
- People have opted level 1 least number of times.

Vote & Blair

Blair vote	1	2	3	4	5	All
Conservative	59	240	1	157	3	460
Labour	38	194	0	676	149	1057
All	97	434	1	833	152	1517

Table.10

- People from both the party are opting for level 4 mostly.
- Only 1 person from conservative party has chosen level 3, and none one has chosen level 3 from labour party.

Vote & Hague

Hague vote	1	2	3	4	5	All
Conservative	11	95	9	286	59	460
Labour	222	522	28	271	14	1057
All	233	617	37	557	73	1517

Table.11

- People from both the party are opting for level 2 mostly.
- People have opted level 3 least number of times.

Vote & Europe

Europe vote	1	2	3	4	5	6	7	8	9	10	11	All
Conservative	5	6	14	18	20	35	32	48	56	54	172	460
Labour	104	71	114	108	103	172	54	63	55	47	166	1057
All	109	77	128	126	123	207	86	111	111	101	338	1517

Table.12

- As per attitudes toward European integration represented by an 11-point scale, most people from both the party has opted for a scale of 11.
- Scale of level 2 has least number of people.

Vote & political.knowledge

political.knowledge vote	0	1	2	3	All
Conservative	94	11	283	72	460
Labour	360	27	493	177	1057
All	454	38	776	249	1517

Table.13

- On scale of political knowledge about the party most people has opted 2 level.
- Second highest is at level 0 and least number of people has opted for level 1.

Vote & gender

gender vote	female	male	All
Conservative	257	203	460
Labour	551	506	1057
All	808	709	1517

Table.14

- There are more female than male in our present dataset.
- Most female have opted for Labour party.
- Most male have opted for labour party as well.

For Multivariate Analysis we use pair plot and Heatmap:

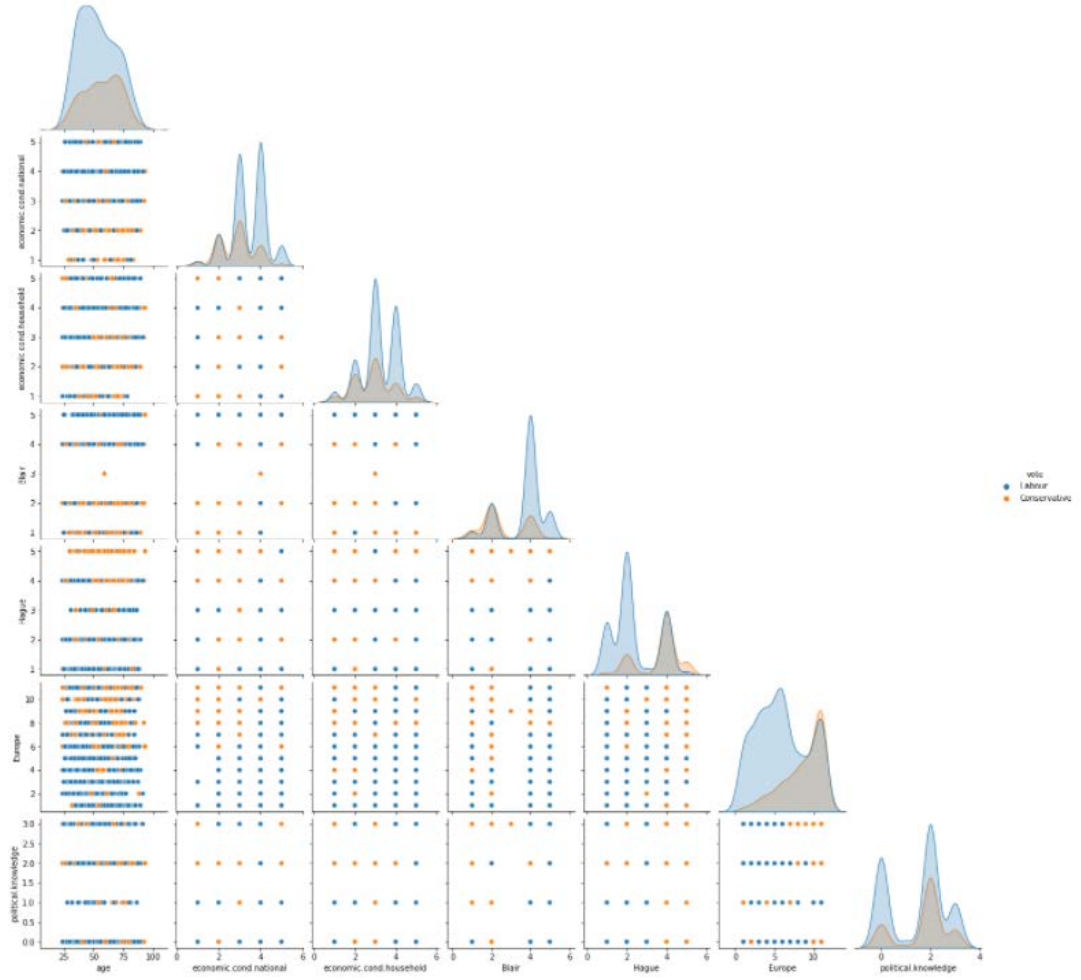


Fig.13

- As we can see visually no correlation exists between any variable pair in the dataset.

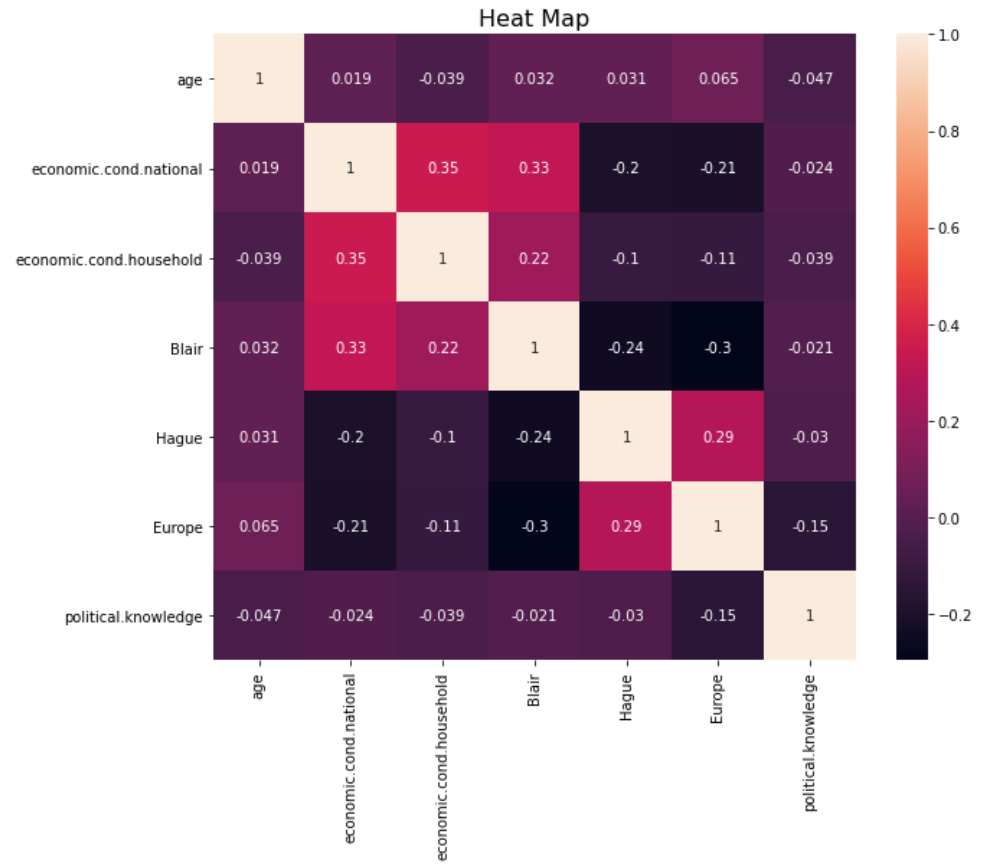


Fig.14

- No variable pair has any kind of significant correlation between them as expected as most of the columns as categorical not related with one another.

1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).

- Only 'vote' and 'gender' columns need to be encoded rest are all already in numeric form. I have used normal encoding to replace the labels of columns with unique numerical values.

Dataset after performing label encoding for the object type variable in dataset.

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

Table.15

Dataset info after label encoding:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1517 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1517 non-null   int32
 1   age                     1517 non-null   int64
 2   economic.cond.national  1517 non-null   int64
 3   economic.cond.household 1517 non-null   int64
 4   Blair                   1517 non-null   int64
 5   Hague                   1517 non-null   int64
 6   Europe                  1517 non-null   int64
 7   political.knowledge     1517 non-null   int64
 8   gender                  1517 non-null   int32
dtypes: int32(2), int64(7)
memory usage: 146.7 KB
```

Table.16

- Since now all the variables are in numeric form we can apply n number of models on it for analysis as per our need.

Scaling:

- Once we have performed label encoding we can observe that out of all the 9 columns only 1 columns (age) is numeric and all the other 8 columns are categorical type columns, so scaling the data does not make sense.
  Also mean number of age is 54.18 whereas the SD is 15.71. So all variable values in our dataset are in 10's range only, hence normalisation will not give any significant improvement in the results of our models.

- Now using 'vote' as our target variable. We split the data into train and test set into 70:30 percent size ratio respectively.
- Sample in train set are 1061 and test set are 456 respectively.

1.4 Apply Logistic Regression and LDA (linear discriminant analysis).

Logistic Regression Model -

Here we are building a Logistic Regression Model with hyper parameter tuning using Grid Search CV.

Hyperparameters chosen for tuning:

Penalty - Specify the norm of the penalty.

Tol - Tolerance for stopping criteria.

Solver - algorithm to use in the optimization problem

- Train Set

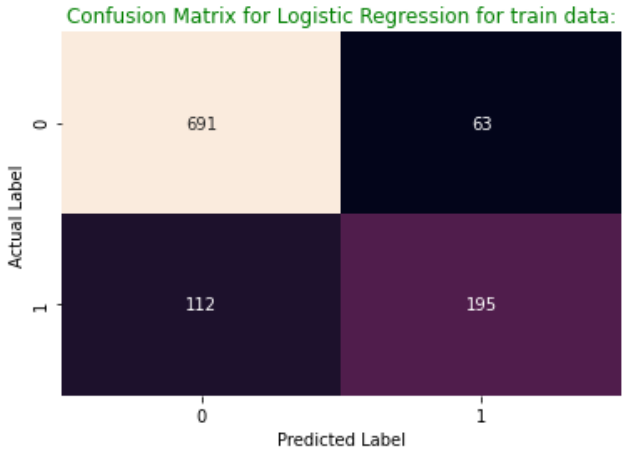
```
Accuracy for Logistic Regression model for train data is: 0.8350612629594723


Classification report for Logistic Regression model for train data is:
              precision    recall  f1-score   support

           0       0.86      0.92      0.89       754
           1       0.76      0.64      0.69       307

    accuracy                           0.84      1061
   macro avg       0.81      0.78      0.79      1061
weighted avg       0.83      0.84      0.83      1061


Confusion Matrix for Logistic Regression model for train data is:
```



AUC Score of Train set is – 0.88

Coefficient of Model

```
The intercept for the model is-1.840138043920511
The coefficient for age is 0.013271015512599137
The coefficient for economic.cond.national is -0.6533293734713137
The coefficient for economic.cond.household is -0.07530819015596574
The coefficient for Blair is -0.6127410039232611
The coefficient for Hague is 0.8035063959216324
The coefficient for Europe is 0.20465447528494551
The coefficient for political.knowledge is 0.30027129240613004
The coefficient for gender is 0.1404633184044687
```

None of the features has strong influence on target variable alone, as also observed in EDA.

- **Test Set**

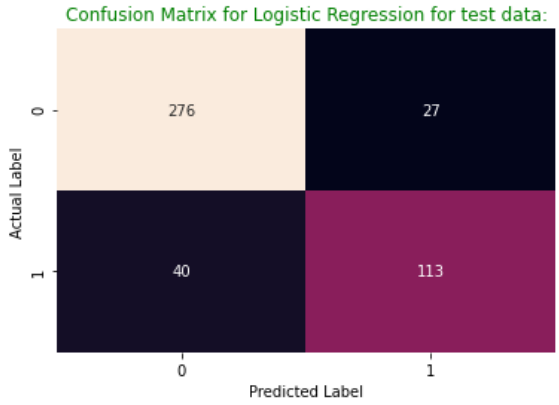
```
Accuracy for Logistic Regression model for test data is: 0.8530701754385965


Classification report for Logistic Regression model for test data is:
              precision    recall  f1-score   support

           0       0.87      0.91      0.89       303
           1       0.81      0.74      0.77       153

    accuracy                           0.85       456
   macro avg       0.84      0.82      0.83       456
weighted avg       0.85      0.85      0.85       456


Confusion Matrix for Logistic Regression model for test data is:
```



AUC Score of Test set is – 0.9

- **ROC Curve for Train and Test set – Logistic Regression Model**

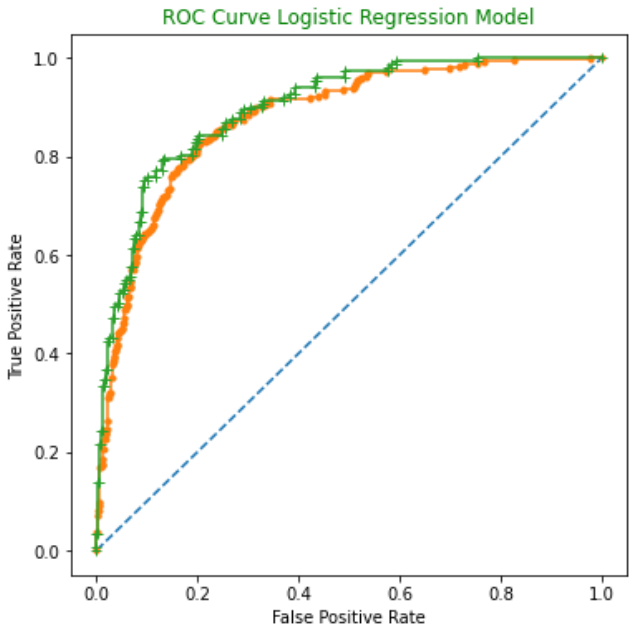


Fig.15

Inferences:

- Logistic Regression model is performing well in both train and test set and there is no problem of over-fitting or under-fitting, hence it's a valid model.
  AUC score for train is 0.88
  AUC score for test is 0.9
- For test set for both 0 and 1 value's:
  This is model is performing well in test data and is able to correctly identify the labels for test set with acceptable values of recall and precision.

Linear Discriminant Analysis -

Here we are building a LDA Model with hyper parameter tuning using Grid Search CV.

Hyperparameters chosen for tuning:

Solver - algorithm to use in the optimization problem.

Tol - Tolerance for stopping criteria.

- Train Set

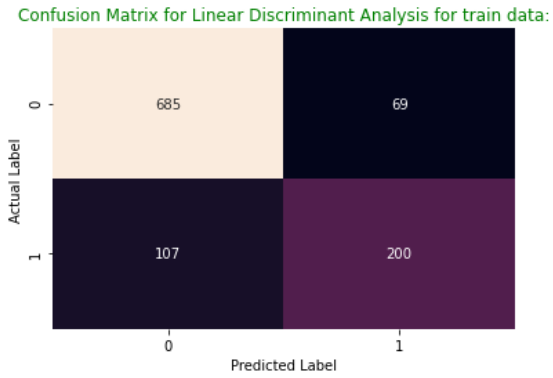
```
Accuracy for Linear Discriminant Analysis model for train data is: 0.8341187558906692


Classification report for Linear Discriminant Analysis model for train data is:
              precision    recall  f1-score   support

           0       0.86      0.91      0.89       754
           1       0.74      0.65      0.69       307

    accuracy                           0.83      1061
   macro avg       0.80      0.78      0.79      1061
weighted avg       0.83      0.83      0.83      1061


Confusion Matrix for Linear Discriminant Analysis model for train data is:
```



AUC Score of Train set is – 0.88

- Test Set

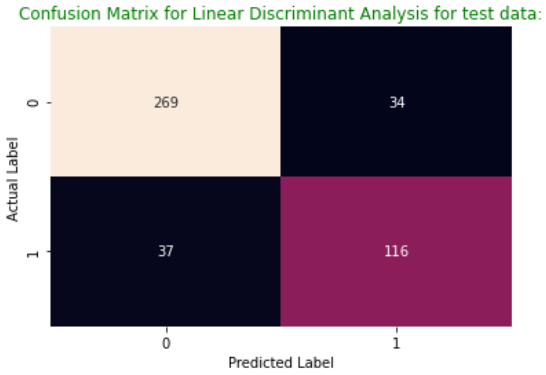
```
Accuracy for Linear Discriminant Analysis model for test data is: 0.8442982456140351


Classification report for Linear Discriminant Analysis model for test data is:
              precision    recall  f1-score   support

           0       0.88      0.89      0.88       303
           1       0.77      0.76      0.77       153

    accuracy                           0.84       456
   macro avg       0.83      0.82      0.82       456
weighted avg       0.84      0.84      0.84       456


Confusion Matrix for Linear Discriminant Analysis model for test data is:
```



AUC Score of Test set is – 0.9

- ROC Curve for Train and Test set – LDA Model

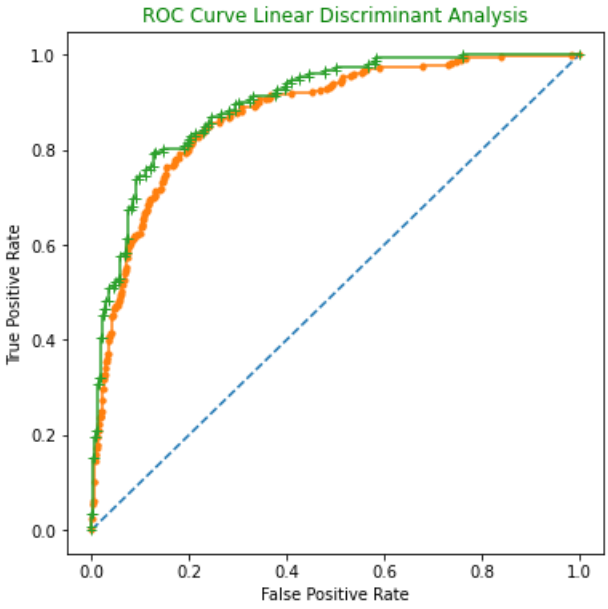


Fig.16

Inferences:

- LDA model is performing well in both train and test set and there is no problem of over-fitting or under-fitting, hence it's a valid model.
  AUC score for train is 0.88
  AUC score for test is 0.9
- For test set for both 0 and 1 value's:
  This is model is performing well in test data and is able to correctly identify the labels for test set with acceptable values of recall and precision.

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.

Generally, good KNN model expects that variables similarly scaled and centered.

Secondly, due to the distinct natures of categorical and numerical data, we usually need to standardize the numerical variables, such as the contributions to the Euclidean distances from a numerical variable and a categorical variable are basically on the same level.

Hence, for KNN we are specially using a scaled copy of dataset and splitting the scaled dataset in 70:30 train test ratio respectively.

KNN Model -

Here we are building a KNN Model with hyper parameter tuning using Grid Search CV.

Hyperparameters chosen for tuning:

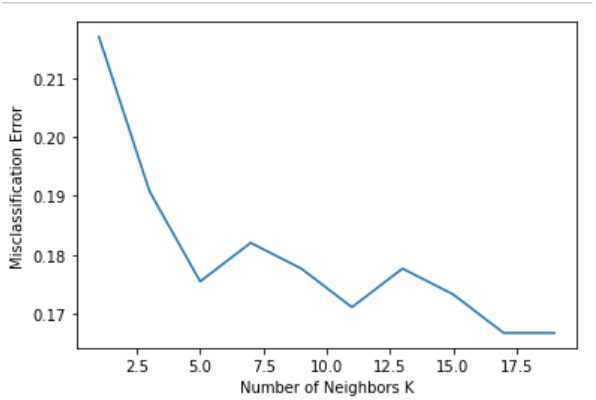
Metric - The distance metric to use for the tree such as Euclidean, Manhattan , Minkowski.

Weights - Weight function used in prediction.

- uniform : uniform weights. All points in each neighbourhood are weighted equally.
- distance : weight points by the inverse of their distance.

Before performing the Grid-Search CV we have calculated the Miss Classification Error (MSE) at different values of k so that we can choose good k-value for our KNN model.

```
[0.2171052631578947,
 0.1907894736842105,
 0.17543859649122806,
 0.18201754385964908,
 0.17763157894736847,
 0.17105263157894735,
 0.17763157894736847,
 0.17324561403508776,
 0.16666666666666663,
 0.16666666666666663]
```



At k = 9 we are getting the least MSE value after which the there is no change in the value of MSE.

So while building the KNN model we are using n_neighbours value as 9 and then applying Grid-Search CV to the model.

- Train Set

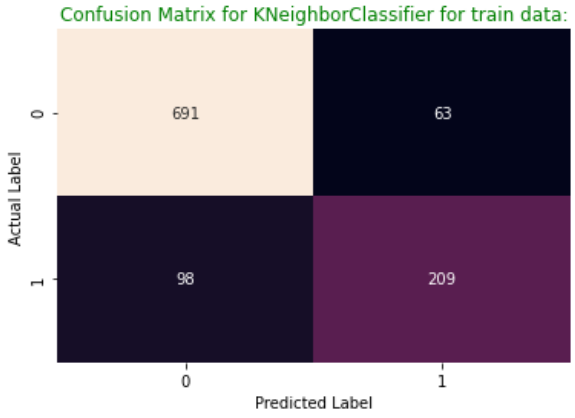
```
Accuracy for KNeighborClassifier model for train data is: 0.8482563619227145


Classification report for KNeighborClassifier model for train data is:
              precision    recall  f1-score   support

           0       0.88      0.92      0.90       754
           1       0.77      0.68      0.72       307

    accuracy                           0.85      1061
   macro avg       0.82      0.80      0.81      1061
weighted avg       0.84      0.85      0.85      1061


Confusion Matrix for KNeighborClassifier model for train data is:
```



AUC Score of Train set is – 0.87

- Test Set

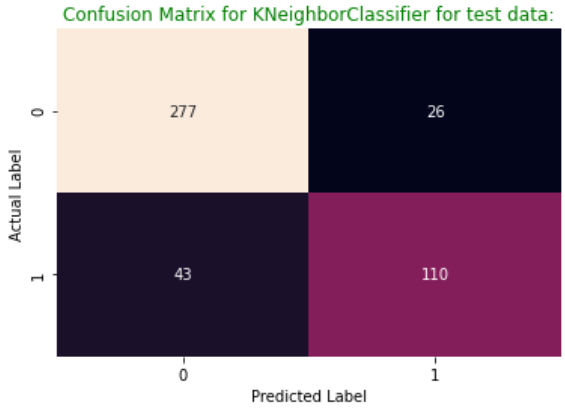
```
Accuracy for KNeighborClassifier model for test data is: 0.8486842105263158


Classification report for KNeighborClassifier model for test data is:
              precision    recall  f1-score   support

           0       0.87      0.91      0.89       303
           1       0.81      0.72      0.76       153

    accuracy                           0.85       456
   macro avg       0.84      0.82      0.83       456
weighted avg       0.85      0.85      0.85       456


Confusion Matrix for KNeighborClassifier model for test data is:
```



AUC Score of Test set is – 0.92

- ROC Curve for Train and Test set – KNN Model

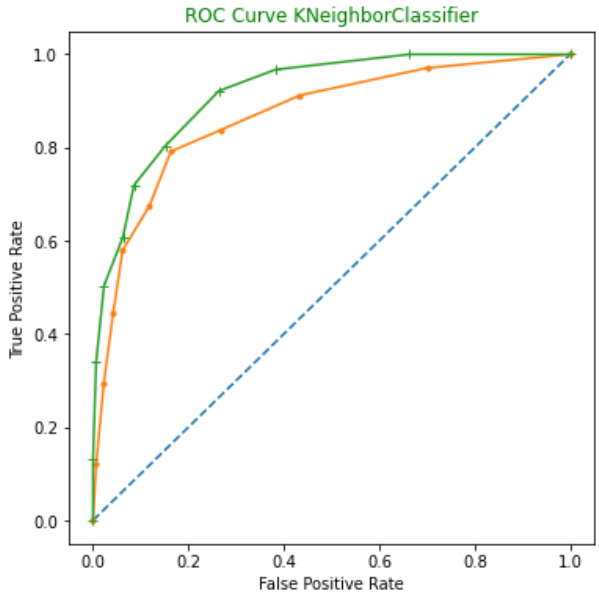


Fig.17

Inferences:

- KNN model is performing well in both train and test set and there is no problem of over-fitting or under-fitting, hence it's a valid model.
  AUC score for train is 0.87
  AUC score for test is 0.92
- For test set for both 0 and 1 value's:
  This is model is performing well in test data and is able to correctly identify the labels for test set with acceptable values of recall and precision.

Gaussian Naïve Bayes Model -

Here we are building a Gaussian Naïve Bayes Model normally, as this model do not have any hyper parameters to tune.

- Train Set

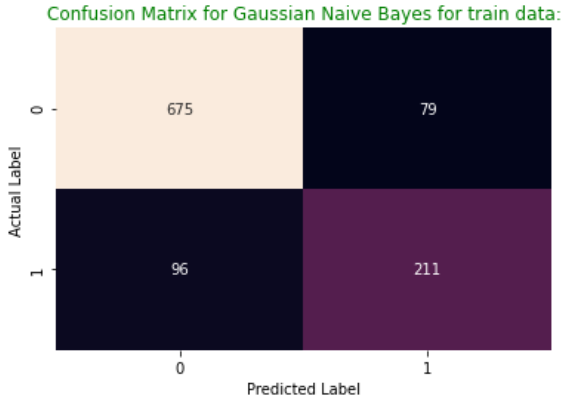
```
Accuracy for Gaussian Naive Bayes model for train data is: 0.8350612629594723


Classification report for Gaussian Naive Bayes model for train data is:
              precision    recall  f1-score   support

           0       0.88      0.90      0.89       754
           1       0.73      0.69      0.71       307

    accuracy                           0.84      1061
   macro avg       0.80      0.79      0.80      1061
weighted avg       0.83      0.84      0.83      1061


Confusion Matrix for Gaussian Naive Bayes model for train data is:
```



AUC Score of Train set is – 0.89

- Test Set

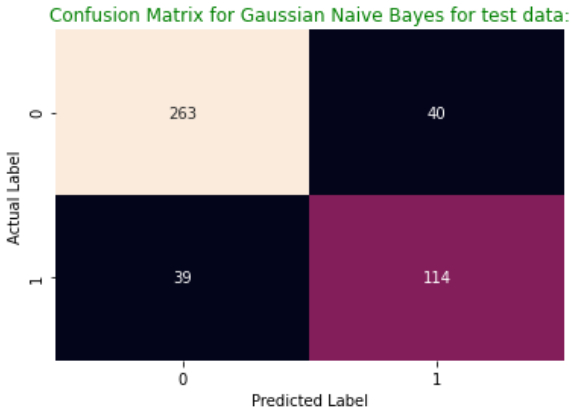
```
Accuracy for Gaussian Naive Bayes model for test data is: 0.8267543859649122


Classification report for Gaussian Naive Bayes model for test data is:
              precision    recall  f1-score   support

           0       0.87      0.87      0.87       303
           1       0.74      0.75      0.74       153

    accuracy                           0.83       456
   macro avg       0.81      0.81      0.81       456
weighted avg       0.83      0.83      0.83       456


Confusion Matrix for Gaussian Naive Bayes model for test data is:
```



AUC Score of Test set is – 0.88

- ROC Curve for Train and Test set – Gaussian Naïve Bayes Model

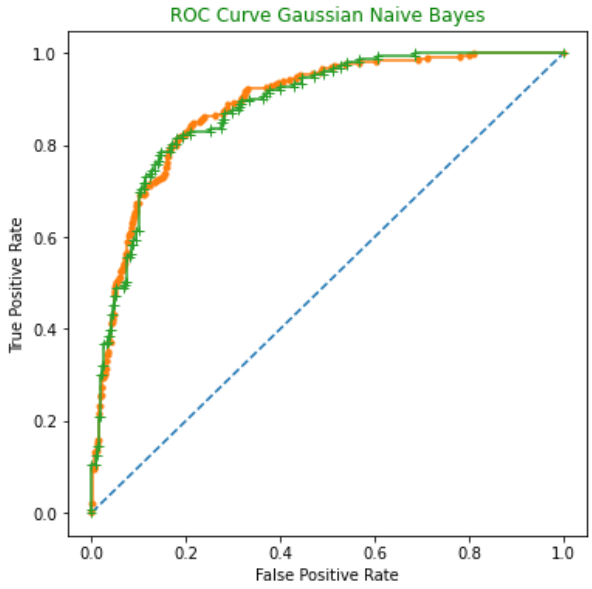


Fig.18

Inferences:

- Gaussian Naïve Bayes model is performing well in both train and test set and there is no problem of over-fitting or under-fitting, hence it's a valid model.
  AUC score for train is 0.89
  AUC score for test is 0.88
- For test set for both 0 and 1 value's:
  This is model is performing well in test data and is able to correctly identify the labels for test set with acceptable values of recall and precision.

1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.

Bagging Model – Using Random Forest Classifier

Here we are building a Random Forest Classifier Model with hyper parameter tuning using Grid Search CV.

Hyperparameters chosen for tuning:

max_depth - The maximum depth of the tree.

max_features - The number of features to consider when looking for the best split

min_samples_leaf - The minimum number of samples required to be at a leaf node.

min_samples_split - The minimum number of samples required to split an internal node

n_estimators - The number of trees in the forest.

- Train Set

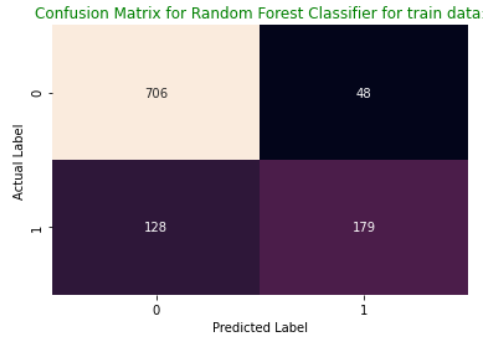
```
Accuracy for Random Forest Classifier model for train data is: 0.8341187558906692


Classification report for Random Forest Classifier model for train data is:
              precision    recall  f1-score   support

           0       0.85      0.94      0.89       754
           1       0.79      0.58      0.67       307

    accuracy                           0.83      1061
   macro avg       0.82      0.76      0.78      1061
weighted avg       0.83      0.83      0.83      1061


Confusion Matrix for Random Forest Classifier model for train data is:
```



AUC Score of Train set is – 0.87

- Test Set

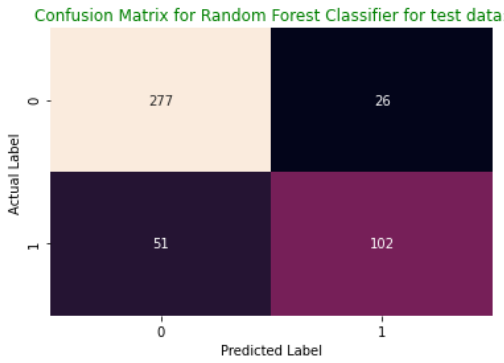
```
Accuracy for Random Forest Classifier model for test data is: 0.831140350877193


Classification report for Random Forest Classifier model for test data is:
              precision    recall  f1-score   support

           0       0.84      0.91      0.88       303
           1       0.80      0.67      0.73       153

    accuracy                           0.83       456
   macro avg       0.82      0.79      0.80       456
weighted avg       0.83      0.83      0.83       456


Confusion Matrix for Random Forest Classifier model for test data is:
```



AUC Score of Test set is – 0.89

- ROC Curve for Train and Test set – Random Forest Classifier Model

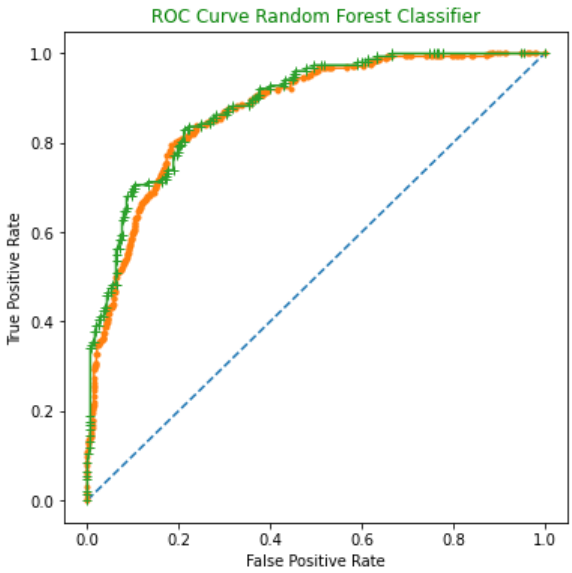


Fig.19

Inferences:

- Random Forest Classifier model is performing well in both train and test set and there is no problem of over-fitting or under-fitting, hence it's a valid model.
  AUC score for train is 0.87
  AUC score for test is 0.89
- For test set for both 0 and 1 value's:
  This is model is performing well in test data and is able to correctly identify the labels for test set with acceptable values of recall and precision.

Boosting Model – Using Ada Boost Classifier

Here we are building a Ada Boost Classifier Model without hyper parameter tuning, as using Grid Search CV did not improved the accuracy score significantly for both train and test set.

Base estimator for an Ada Boost Classifier Model is decisiontreeclassifier by default.

- Train Set

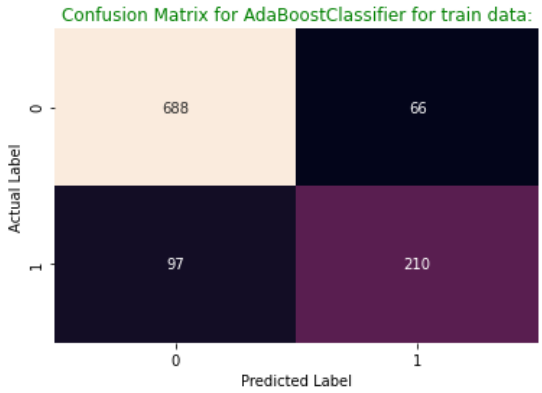
```
Accuracy for AdaBoostClassifier model for train data is: 0.8463713477851084


Classification report for AdaBoostClassifier model for train data is:
              precision    recall  f1-score   support

           0       0.88      0.91      0.89       754
           1       0.76      0.68      0.72       307

    accuracy                           0.85      1061
   macro avg       0.82      0.80      0.81      1061
weighted avg       0.84      0.85      0.84      1061


Confusion Matrix for AdaBoostClassifier model for train data is:
```



AUC Score of Train set is – 0.86

- Test Set

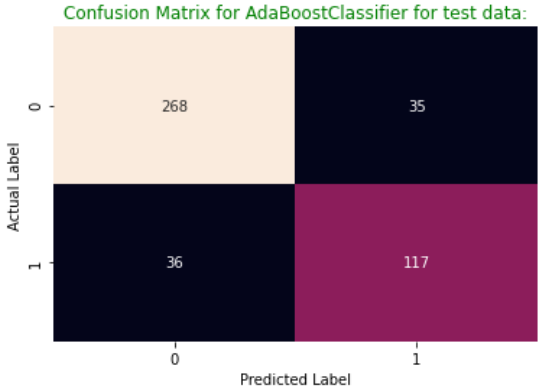
```
Accuracy for AdaBoostClassifier model for test data is: 0.8442982456140351


Classification report for AdaBoostClassifier model for test data is:
              precision    recall  f1-score   support

           0       0.88      0.88      0.88       303
           1       0.77      0.76      0.77       153

    accuracy                           0.84       456
   macro avg       0.83      0.82      0.83       456
weighted avg       0.84      0.84      0.84       456


Confusion Matrix for AdaBoostClassifier model for test data is:
```



AUC Score of Test set is – 0.93

- ROC Curve for Train and Test set – Ada Boost Classifier Model

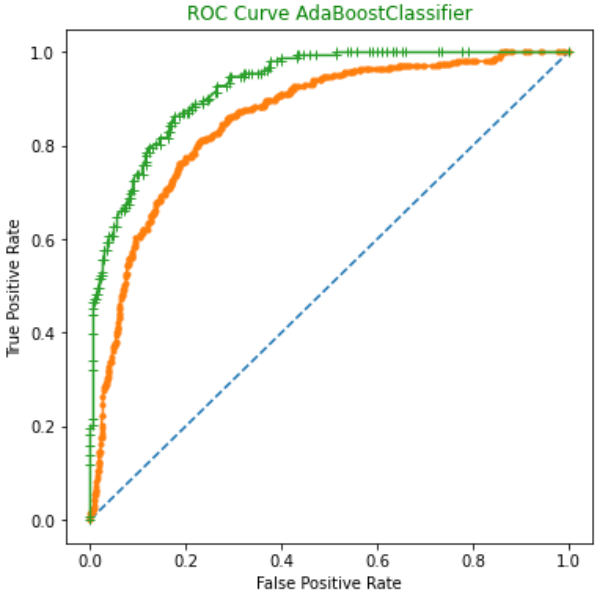


Fig.20

Inferences:

- Ada Boost Classifier model is performing well in both train and test set and there is no problem of over-fitting or under-fitting, hence it's a valid model.
  AUC score for train is 0.86
  AUC score for test is 0.93
- For test set for both 0 and 1 value's:
  This is model is performing well in test data and is able to correctly identify the labels for test set with acceptable values of recall and precision.

Boosting Model – Using Gradient Boosting Classifier

Here we are building a Gradient Boosting Classifier Model without hyper parameter tuning, as using Grid Search CV did not improved the accuracy score significantly for both train and test set.

Using Grid Search CV using learning_rate and n_estimators as parameters to tune resulted in decrease in both train and test accuracy score.

- Train Set

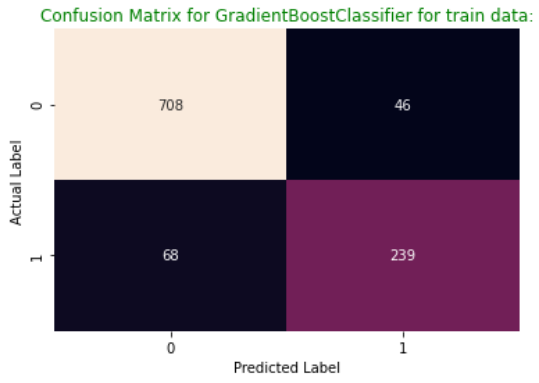
```
Accuracy for GradientBoostClassifier model for train data is: 0.8925541941564562

Classification report for GradientBoostClassifier model for train data is:
              precision    recall  f1-score   support

           0       0.91      0.94      0.93       754
           1       0.84      0.78      0.81       307

    accuracy                           0.89      1061
   macro avg       0.88      0.86      0.87      1061
weighted avg       0.89      0.89      0.89      1061
```

```
Confusion Matrix for GradientBoostClassifier model for train data is:
```



AUC Score of Train set is – 0.89

- Test Set

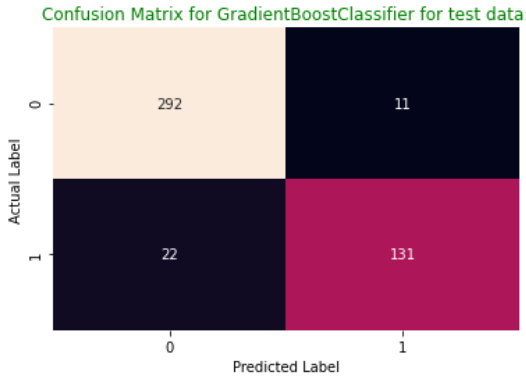
```
Accuracy for GradientBoostClassifier model for test data is: 0.9276315789473685


Classification report for GradientBoostClassifier model for test data is:
              precision    recall  f1-score   support

           0       0.93      0.96      0.95       303
           1       0.92      0.86      0.89       153

    accuracy                           0.93       456
   macro avg       0.93      0.91      0.92       456
weighted avg       0.93      0.93      0.93       456


Confusion Matrix for GradientBoostClassifier model for test data is:
```



AUC Score of Test set is – 0.98

- ROC Curve for Train and Test set – Gradient Boosting Classifier Model

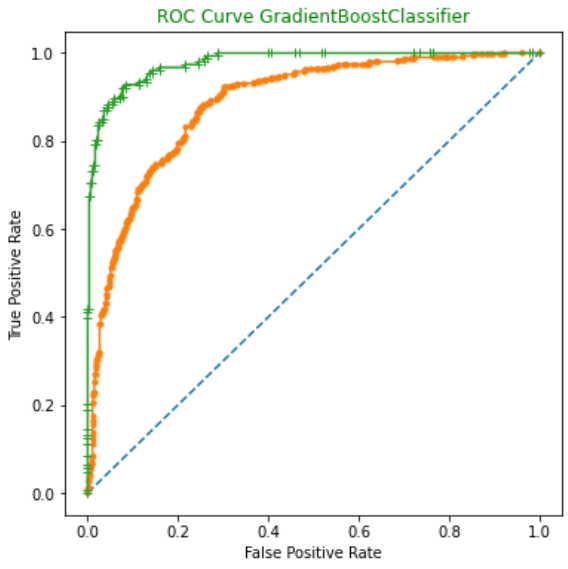


Fig.21

Inferences:

- Gradient Boosting Classifier model is performing well in both train and test set and there is no problem of over-fitting or under-fitting, hence it's a valid model.
  AUC score for train is 0.89
  AUC score for test is 0.98
- For test set for both 0 and 1 value's:
  This is model is performing well in test data and is able to correctly identify the labels for test set with acceptable values of recall and precision.

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. Final Model: Compare the models and write inference which model is best/optimized.

List of Models performed:

1) Logistic Regression performed well with Grid Search CV

2) LDA performed well with Grid Search CV

3) KNN performed well with Grid Search CV (on scaled dataset)

4) Gaussian Naïve Bayes performed without Grid Search CV

5) Bagging – Random Forest Classifier performed well with Grid Search CV

6) Boosting – Ada Boosting performed without Grid Search CV

7) Boosting – Gradient Boosting performed without Grid Search CV

- Reason for not performing Grid Search CV on some model is because the model performance decreased by using it.

MODEL PERFORMACE																										
																			Boosting							
	Logistice Regression				LDA				KNN				Gaussian Naïve Bayes				Bagging - Random Forest				Ada Boosting				Gradient Boosting			
	Train		Test		Train		Test		Train		Test		Train		Test		Train		Test		Train		Test		Train		Test	
	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
AUC ROC SCORE	0.88		0.9		0.88		0.9		0.87		0.92		0.89		0.88		0.87		0.89		0.86		0.93		0.89		0.98	
Accuracy	0.83		0.85		0.83		0.84		0.84		0.84		0.83		0.82		0.83		0.83		0.84		0.84		0.89		0.92	
Precision	0.86	0.76	0.87	0.81	0.86	0.74	0.88	0.77	0.88	0.77	0.87	0.81	0.88	0.73	0.87	0.74	0.85	0.79	0.84	0.8	0.88	0.76	0.88	0.77	0.91	0.84	0.93	0.92
Recall	0.92	0.64	0.91	0.74	0.91	0.65	0.89	0.76	0.92	0.68	0.91	0.72	0.9	0.69	0.87	0.75	0.94	0.58	0.91	0.67	0.91	0.68	0.88	0.76	0.94	0.78	0.96	0.86
F1 Score	0.89	0.69	0.89	0.77	0.89	0.69	0.88	0.77	0.9	0.72	0.89	0.76	0.89	0.71	0.87	0.74	0.89	0.67	0.88	0.73	0.89	0.72	0.88	0.77	0.93	0.81	0.95	0.89

Table.17

The dataset give to us is very small having only 1525 records, so as expected most of the models are performing very similar in both train and test. There is no problem of overfitting or under fitting in any the models that I have built and accuracy of both train and test set is with 10 % of each other.

Best Model:

Out of all the models only Gradient Boosting Classifier is performing best in term of accuracy of both train and test set.

Train Accuracy: 0.89
Test Accuracy: 0.92

AUC_ROC score of both train and test are also highest among all the other models.

AUC_ROC score train: 0.89
AUC_ROC score test: 0.98

Rest all models have move or less same accuracy at 83-84%.

1.8 Based on these predictions, what are the insights?

Prediction Purpose

As discussed earlier in model comparison section, Gradient Boosting Classifier model is performing best out of all the models. Hence we can take Gradient Boosting Classifier as our final model for making prediction for real word unknown data.

Recommendation

- As per our best model we can find most important features in our dataset.

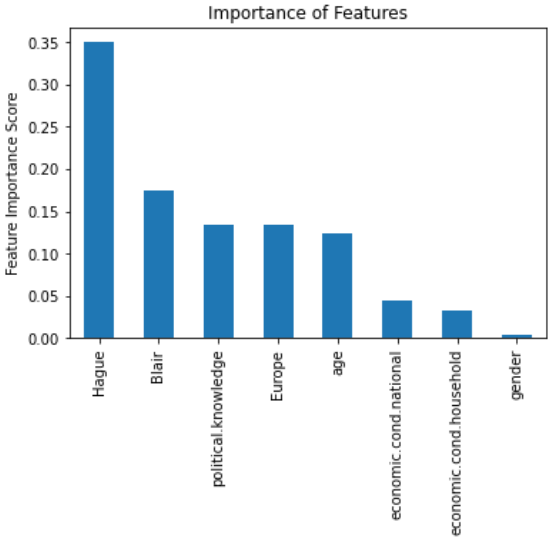


Fig.22

- As also observed in coefficient's of Logistic Regression model earlier we can say none of the coefficients were strong predictors of our target variable. If we have to choose from these weak predictors we can say Variable Hague is good predictor among these weak predictors.
- The dataset accuracy for different models can be increase if we have more number of records. The given dataset is very small due to which all models are performing more or less the same.
- Organisation should consider using more features and more number of records to improve the model performance as even when using Grid Search in most of the models is not showing significant improvements.

Appendix

- A model is said to perform well when it runs well in train as well as test data both
- The auc_ruc_score of both train and test should not differ more than 10% for the model to be valid.
- Higher the auc_ruc_score the better the model.
- If the difference between the auc_ruc_score for train and test set is greater than 10% then problem for overfitting and under fitting may arise.
  'Overfitting' is when model perform well in train but not in test set.
  'Under-fitting' is when model does not perform well in train but do well in test set.
  'Best Fit' is when model perform well in train as well as in test to a similar level.

- Confusion matrix and classification report is made for all the models.
- In classification report - '1' is our value of importance for model i.e. people who have claimed insurance.
  - Recall indicates how many of the actual data points are identified as True data points by the model.
  - Precision indicates the points that are identified as positive by the model, how many are really positive.
  - The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

  We will focus on recall and precision value in Classification Report for each model.

- An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:
  True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

  $$TPR=TP/TP+FN$$

  False Positive Rate (FPR) is defined as follows:

  $$FPR=FP/FP+TN$$
- An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

Problem 2:

Executive Summary

In this particular project, we are going to work on the inaugural corpora from the nltk (Natural Language Tool Kit) library 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

Data Set Used

Here we are using inaugural speech dataset from nltk (Natural Language Tool Kit) library. This dataset contains speeches of '58' Presidents of the United States of America.

```
['1789-Washington.txt',
 '1793-Washington.txt',
 '1797-Adams.txt',
 '1801-Jefferson.txt',
 '1805-Jefferson.txt',
 '1809-Madison.txt',
 '1813-Madison.txt',
 '1817-Monroe.txt',
 '1821-Monroe.txt',
 '1825-Adams.txt',
 '1829-Jackson.txt',
 '1833-Jackson.txt',
 '1837-VanBuren.txt',
 '1841-Harrison.txt',
 '1845-Polk.txt',
 '1849-Taylor.txt',
 '1853-Pierce.txt',
 '1857-Buchanan.txt',
 '1861-Lincoln.txt',
 '1865-Lincoln.txt',
 '1869-Grant.txt',
 '1873-Grant.txt',
 '1877-Hayes.txt',
 '1881-Garfield.txt',
 '1885-Cleveland.txt',
 '1889-Harrison.txt',
 '1893-Cleveland.txt',
 '1897-McKinley.txt',
 '1901-McKinley.txt',
 '1905-Roosevelt.txt',
 '1909-Taft.txt',
 '1913-Wilson.txt',
 '1917-Wilson.txt',
 '1921-Harding.txt',
 '1925-Coolidge.txt',
 '1929-Hoover.txt',
 '1933-Roosevelt.txt',
 '1937-Roosevelt.txt',
 '1941-Roosevelt.txt',
 '1945-Roosevelt.txt',
 '1949-Truman.txt',
 '1953-Eisenhower.txt',
 '1957-Eisenhower.txt',
 '1961-Kennedy.txt',
 '1965-Johnson.txt',
 '1969-Nixon.txt',
 '1973-Nixon.txt',
 '1977-Carter.txt',
 '1981-Reagan.txt',
 '1985-Reagan.txt',
 '1989-Bush.txt',
 '1993-Clinton.txt',
 '1997-Clinton.txt',
 '2001-Bush.txt',
 '2005-Bush.txt',
 '2009-Obama.txt',
 '2013-Obama.txt',
 '2017-Trump.txt']
```

Table.18

We are only going to use the speeches as mentioned above and working on them individually to find numbers of characters, words and sentences in them.

2.1 Find the number of characters, words, and sentences for the mentioned documents.

To identify the number of characters in the speech I am using the 'len' function. For number of words and sentences in the speech/text I am using the sentence tokenizer and word tokenizer from the nltk library.

1st Speech - President Franklin D. Roosevelt in 1941

On each national day of inauguration since 1789, the people have renewed their sense of dedication to the United States.\n\nIn Washington \'s day the task of the people was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the people was to preserve that Nati on from disruption from within.\n\nIn this day the task of the people is to save that Nation and its institutions from disruption from without. \n\nTo us there has come a time, in the midst of swift happenings, to pause for a moment and take stock -- to recall what our place in history has been, and to rediscover what we are and what we may be. If we do not, we risk the real peril of inaction.\n\nLives of nations are determi ned not by the count of years, but by the lifetime of the human spirit. The life of a man is three-score years and ten: a little more, a little less. The life of a nation is the fullness of the measure of its will to live.\n\nThere are men who doubt this. There are men who believe that democ racy, as a form of Government and a frame of life, is limited or measured by a kind of mystical and artificial fate that, for some unexplained r eason, tyranny and slavery have become the surging wave of the future -- and that freedom is an ebbing tide.\n\nBut we Americans know th at this is not true.\n\nEight years ago, when the life of this Republic seemed frozen by a fatalistic terror, we proved that this is not true. We were in the midst of shock -- but we acted. We acted quickly, boldly, decisively.\n\nThese later years have been living years -- fruitful years fo r the people of this democracy. For they have brought to us greater security and, I hope, a better understanding that life\'s ideals are to be m easured in other than material things.\n\nMost vital to our present and our future is this experience of a democracy which successfully survi ved crisis at home; put away many evil things; built new structures on enduring lines; and, through it all, maintained the fact of its democracy .\n\nFor action has been taken within the three-way framework of the Constitution of the United States. The coordinate branches of the Gov ernment continue freely to function. The Bill of Rights remains inviolate. The freedom of elections is wholly maintained. Prophets of the dow nfall of American democracy have seen their dire predictions come to naught.\n\nDemocracy is not dying.\n\nWe know it because we have s een it revive--and grow.\n\nWe know it cannot die -- because it is built on the unhampered initiative of individual men and women joined to gether in a common enterprise -- an enterprise undertaken and carried through by the free expression of a free majority.\n\nWe know it bec ause democracy alone, of all forms of government, enlists the full force of men\'s enlightened will.\n\nWe know it because democracy alone has constructed an unlimited civilization capable of infinite progress in the improvement of human life.\n\nWe know it because, if we look b elow the surface, we sense it still spreading on every continent -- for it is the most humane, the most advanced, and in the end the most unc onquerable of all forms of human society.\n\nA nation, like a person, has a body--a body that must be fed and clothed and housed, invigorat ed and rested, in a manner that measures up to the objectives of our time.\n\nA nation, like a person, has a mind -- a mind that must be kept informed and alert, that must know itself, that understands the hopes and the needs of its neighbors -- all the other nations that live within t he narrowing circle of the world.\n\nAnd a nation, like a person, has something deeper, something more permanent, something larger than t he sum of all its parts. It is that something which matters most to its future -- which calls forth the most sacred guarding of its present.\n\nIt i s a thing for which we find it difficult -- even impossible -- to hit upon a single, simple word.\n\nAnd yet we all understand what it is -- the spi rit -- the faith of America. It is the product of centuries. It was born in the multitudes of those who came from many lands -- some of high deg ree, but mostly plain people, who sought here, early and late, to find freedom more freely.\n\nThe democratic aspiration is no mere recent p hase in human history. It is human history. It permeated the ancient life of early peoples. It blazed anew in the middle ages. It was written in Magna Charta.\n\nIn the Americas its impact has been irresistible. America has been the New World in all tongues, to all peoples, not becaus e this continent was a new-found land, but because all those who came here believed they could create upon this continent a new life -- a lif e that should be new in freedom.\n\nIts vitality was written into our own Mayflower Compact, into the Declaration of Independence, into th e Constitution of the United States, into the Gettysburg Address.\n\nThose who first came here to carry out the longings of their spirit, and t he millions who followed, and the stock that sprang from them -- all have moved forward constantly and consistently toward an ideal which i n itself has gained stature and clarity with each generation.\n\nThe hopes of the Republic cannot forever tolerate either undeserved poverty or self-serving wealth.\n\nWe know that we still have far to go; that we must more greatly build the security and the opportunity and the kn owledge of every citizen, in the measure justified by the resources and the capacity of the land.\n\nBut it is not enough to achieve these pur poses alone. It is not enough to clothe and feed the body of this Nation, and instruct and inform its mind. For there is also the spirit. And of t he three, the greatest is the spirit.\n\nWithout the body and the mind, as all men know, the Nation could not live.\n\nBut if the spirit of Ame rica were killed, even though the Nation\'s body and mind, constricted in an alien world, lived on, the America we know would have perished .\n\nThat spirit -- that faith -- speaks to us in our daily lives in ways often unnoticed, because they seem so obvious. It speaks to us here in th e Capital of the Nation. It speaks to us through the processes of governing in the sovereignties of 48 States. It speaks to us in our counties, in our cities, in our towns, and in our villages. It speaks to us from the other nations of the hemisphere, and from those across the seas -- the en slaved, as well as the free. Sometimes we fail to hear or heed these voices of freedom because to us the privilege of our freedom is such an o ld, old story.\n\nThe destiny of America was proclaimed in words of prophecy spoken by our first President in his first inaugural in 1789 -- wo rds almost directed, it would seem, to this year of 1941: "The preservation of the sacred fire of liberty and the destiny of the republican mod el of government are justly considered deeply, finally, staked on the experiment intrusted to the hands of the American people."\n\nIf we los e that sacred fire--if we let it be smothered with doubt and fear -- then we shall reject the destiny which Washington strove so valiantly and s o triumphantly to establish. The preservation of the spirit and faith of the Nation does, and will, furnish the highest justification for every sacr ifice that we may make in the cause of national defense.\n\nIn the face of great perils never before encountered, our strong purpose is to pr otect and to perpetuate the integrity of democracy.\n\nFor this we muster the spirit of America, and the faith of America.\n\nWe do not retr eat. We are not content to stand still. As Americans, we go forward, in the service of our country, by the will of God.\n

Number of characters in the speech – 7571

Number of sentences in the speech – 68

Number of words in the speech – 1526

2nd Speech - President John F. Kennedy in 1961

Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fello w citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well as a beginning -- signifying ren ewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our forebears l prescribed nearly a century an d three quarters ago.\n\nThe world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for which our forebears fought are still at issue around the globe -- the be lief that the rights of man come not from the generosity of the state, but from the hand of God.\n\nWe dare not forget today that we are the heirs of that first revolution. Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new g eneration of Americans -- born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage -- and unwilling to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are committed today at home and around the world.\n\nLet every nation know, whether it wishes us well or ill, that we shall pay any price, bear any burden, meet any hardship, support any friend, oppose any foe, in order to assure the survival and the success of liberty.\n\nThis much we pledge -- and more.\n\nTo those old allies whose cultural and spiritual origins we share, we pledge the loyalty of faithful friends. United, t here is little we cannot do in a host of cooperative ventures. Divided, there is little we can do -- for we dare not meet a powerful challenge at odds and split asunder.\n\nTo those new States whom we welcome to the ranks of the free, we pledge our word that one form of colonial co ntrol shall not have passed away merely to be replaced by a far more iron tyranny. We shall not always expect to find them supporting our vi ew. But we shall always hope to find them strongly supporting their own freedom -- and to remember that, in the past, those who foolishly s ought power by riding the back of the tiger ended up inside.\n\nTo those peoples in the huts and villages across the globe struggling to break the bonds of mass misery, we pledge our best efforts to help them help themselves, for whatever period is required -- not because the Com munists may be doing it, not because we seek their votes, but because it is right. If a free society cannot help the many who are poor, it cann ot save the few who are rich.\n\nTo our sister republics south of our border, we offer a special pledge -- to convert our good words into good deeds -- in a new alliance for progress -- to assist free men and free governments in casting off the chains of poverty. But this peaceful revolu tion of hope cannot become the prey of hostile powers. Let all our neighbors know that we shall join with them to oppose aggression or subv ersion anywhere in the Americas. And let every other power know that this Hemisphere intends to remain the master of its own house.\n\nT o that world assembly of sovereign states, the United Nations, our last best hope in an age where the instruments of war have far outpaced t he instruments of peace, we renew our pledge of support--to prevent it from becoming merely a forum for invective -- to strengthen its shiel d of the new and the weak -- and to enlarge the area in which its writ may run.\n\nFinally, to those nations who would make themselves our adversary, we offer not a pledge but a request: that both sides begin anew the quest for peace, before the dark powers of destruction unleas hed by science engulf all humanity in planned or accidental self-destruction.\n\nWe dare not tempt them with weakness. For only when our arms are sufficient beyond doubt can we be certain beyond doubt that they will never be employed.\n\nBut neither can two great and powe rful groups of nations take comfort from our present course -- both sides overburdened by the cost of modern weapons, both rightly alarme d by the steady spread of the deadly atom, yet both racing to alter that uncertain balance of terror that stays the hand of mankind\'s final wa r.\n\nSo let us begin anew -- remembering on both sides that civility is not a sign of weakness, and sincerity is always subject to proof. Let us never negotiate out of fear. But let us never fear to negotiate.\n\nLet both sides explore what problems unite us instead of belaboring those problems which divide us.\n\nLet both sides, for the first time, formulate serious and precise proposals for the inspection and control of arm s -- and bring the absolute power to destroy other nations under the absolute control of all nations.\n\nLet both sides seek to invoke the wo nders of science instead of its terrors. Together let us explore the stars, conquer the deserts, eradicate disease, tap the ocean depths, and en courage the arts and commerce.\n\nLet both sides unite to heed in all corners of the earth the command of Isaiah -- to "undo the heavy burd ens ... and to let the oppressed go free."\n\nAnd if a beachhead of cooperation may push back the jungle of suspicion, let both sides join in cr eating a new endeavor, not a new balance of power, but a new world of law, where the strong are just and the weak secure and the peace pr eserved.\n\nAll this will not be finished in the first 100 days. Nor will it be finished in the first 1,000 days, nor in the life of this Administration , nor even perhaps in our lifetime on this planet. But let us begin.\n\nIn your hands, my fellow citizens, more than in mine, will rest the final s uccess or failure of our course. Since this country was founded, each generation of Americans has been summoned to give testimony to its n ational loyalty. The graves of young Americans who answered the call to service surround the globe.\n\nNow the trumpet summons us again -- not as a call to bear arms, though arms we need; not as a call to battle, though embattled we are -- but a call to bear the burden of a long t wilight struggle, year in and year out, "rejoicing in hope, patient in tribulation" -- a struggle against the common enemies of man: tyranny, po verty, disease, and war itself.\n\nCan we forge against these enemies a grand and global alliance, North and South, East and West, that can a ssure a more fruitful life for all mankind? Will you join in that historic effort?\n\nIn the long history of the world, only a few generations have been granted the role of defending freedom in its hour of maximum danger. I do not shrink from this responsibility -- I welcome it. I do not b elieve that any of us would exchange places with any other people or any other generation. The energy, the faith, the devotion which we bri ng to this endeavor will light our country and all who serve it -- and the glow from that fire can truly light the world.\n\nAnd so, my fellow A mericans: ask not what your country can do for you -- ask what you can do for your country.\n\nMy fellow citizens of the world: ask not what America will do for you, but what together we can do for the freedom of man.\n\nFinally, whether you are citizens of America or citizens of t he world, ask of us the same high standards of strength and sacrifice which we ask of you. With a good conscience our only sure reward, with history the final judge of our deeds, let us go forth to lead the land we love, asking His blessing and His help, but knowing that here on earth God\'s work must truly be our own.\n

Number of characters in the speech – 7618

Number of sentences in the speech – 52

Number of words in the speech – 1543

3rd Speech - President Richard Nixon in 1973

Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and good country we s hare together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroa d and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe cent ral question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not be what other postwar peri ods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad.\n\nLet us resolve th at this will be what it can become: a time of great responsibilities greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis past year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendships, and by our missions to Peking and to Moscow, we were able to establish the base for a new and more durable pat tern of relationships among the nations of the world. Because of America\'s bold initiatives, 1972 will be long remembered as the year of the greatest progress since the end of World War II toward a lasting peace in the world.\n\nThe peace we seek in the world is not the flimsy peac e which is merely an interlude between wars, but a peace which can endure for generations to come.\n\nIt is important that we understand both the necessity and the limitations of America\'s role in maintaining that peace.\n\nUnless we in America work to preserve the peace, the re will be no peace.\n\nUnless we in America work to preserve freedom, there will be no freedom.\n\nBut let us clearly understand the new nature of America\'s role, as a result of the new policies we have adopted over these past four years.\n\nWe shall respect our treaty commit ments.\n\nWe shall support vigorously the principle that no country has the right to impose its will or rule on another by force.\n\nWe shall continue, in this era of negotiation, to work for the limitation of nuclear arms, and to reduce the danger of confrontation between the great powers.\n\nWe shall do our share in defending peace and freedom in the world. But we shall expect others to do their share.\n\nThe time h as passed when America will make every other nation\'s conflict our own, or make every other nation\'s future our responsibility, or presum e to tell the people of other nations how to manage their own affairs.\n\nJust as we respect the right of each nation to determine its own fut ure, we also recognize the responsibility of each nation to secure its own future.\n\nJust as America\'s role is indispensable in preserving the world\'s peace, so is each nation\'s role indispensable in preserving its own peace.\n\nTogether with the rest of the world, let us resolve to move forward from the beginnings we have made. Let us continue to bring down the walls of hostility which have divided the world for too l ong, and to build in their place bridges of understanding -- so that despite profound differences between systems of government, the people of the world can be friends.\n\nLet us build a structure of peace in the world in which the weak are as safe as the strong -- in which each resp ects the right of the other to live by a different system -- in which those who would influence others will do so by the strength of their ideas, and not by the force of their arms.\n\nLet us accept that high responsibility not as a burden, but gladly -- gladly because the chance to build s uch a peace is the noblest endeavor in which a nation can engage; gladly, also, because only if we act greatly in meeting our responsibilities a broad will we remain a great Nation, and only if we remain a great Nation will we act greatly in meeting our challenges at home.\n\nWe have the chance today to do more than ever before in our history to make life better in America -- to ensure better education, better health, bette r housing, better transportation, a cleaner environment -- to restore respect for law, to make our communities more livable -- and to insure t he God-given right of every American to full and equal opportunity.\n\nBecause the range of our needs is so great -- because the reach of ou r opportunities is so great -- let us be bold in our determination to meet those needs in new ways.\n\nJust as building a structure of peace ab road has required turning away from old policies that failed, so building a new era of progress at home requires turning away from old policie s that have failed.\n\nAbroad, the shift from old policies to new has not been a retreat from our responsibilities, but a better way to peace.\ n\nAnd at home, the shift from old policies to new will not be a retreat from our responsibilities, but a better way to progress.\n\nAbroad an d at home, the key to those new responsibilities lies in the placing and the division of responsibility. We have lived too long with the consequ ences of attempting to gather all power and responsibility in Washington.\n\nAbroad and at home, the time has come to turn away from the condescending policies of paternalism -- of "Washington knows best."\n\nA person can be expected to act responsibly only if he has responsi bility. This is human nature. So let us encourage individuals at home and nations abroad to do more for themselves, to decide more for them selves. Let us locate responsibility in more places. Let us measure what we will do for others by what they will do for themselves.\n\nThat is why today I offer no promise of a purely governmental solution for every problem. We have lived too long with that false promise. In trusting too much in government, we have asked of it more than it can deliver. This leads only to inflated expectations, to reduced individual effort, a nd to a disappointment and frustration that erode confidence both in what government can do and in what people can do.\n\nGovernment must learn to take less from people so that people an do more for themselves.\n\nLet us remember that America was built not by governme nt, but by people -- not by welfare, but by work -- not by shirking responsibility, but by seeking responsibility.\n\nIn our own lives, let each of us ask -- not just what will government do for me, but what can I do for myself?\n\nIn the challenges we face together, let each of us ask -- n ot just how can government help, but how can I help?\n\nYour National Government has a great and vital role to play. And I pledge to you th at where this Government should act, we will act boldly and we will lead boldly. But just as important is the role that each and every one of u s must play, as an individual and as a member of his own community.\n\nFrom this day forward, let each of us make a solemn commitment i n his own heart: to bear his responsibility, to do his part, to live his ideals -- so that together, we can see the dawn of a new age of progress f or America, and together, as we celebrate our 200th anniversary as a nation, we can do so proud in the fulfillment of our promise to ourselve s and to the world.\n\nAs America\'s longest and most difficult war comes to an end, let us again learn to debate our differences with civility and decency. And let each of us reach out for that one precious quality government cannot provide -- a new level of respect for the rights an d feelings of one another, a new level of respect for the individual human dignity which is the cherished birthright of every American.\n\nAb ove all else, the time has come for us to renew our faith in ourselves and in America.\n\nIn recent years, that faith has been challenged.\n\n Our children have been taught to be ashamed of their country, ashamed of their parents, ashamed of America\'s record at home and of its ro le in the world.\n\nAt every turn, we have been beset by those who find everything wrong with America and little that is right. But I am confi dent that this will not be the judgment of history on these remarkable times in which we are privileged to live.\n\nAmerica\'s record in this c entury has been unparalleled in the world\'s history for its responsibility, for its generosity, for its creativity and for its progress.\n\nLet us be proud that our system has produced and provided more freedom and more abundance, more widely shared, than any other system in the hi story of the world.\n\nLet us be proud that in each of the four wars in which we have been engaged in this century, including the one we are now bringing to an end, we have fought not for our selfish advantage, but to help others resist aggression.\n\nLet us be proud that by our bo ld, new initiatives, and by our steadfastness for peace with honor, we have made a break-through toward creating in the world what the worl d has not known before -- a structure of peace that can last, not merely for our time, but for generations to come.\n\nWe are embarking her e today on an era that presents challenges great as those any nation, or any generation, has ever faced.\n\nWe shall answer to God, to histor y, and to our conscience for the way in which we use these years.\n\nAs I stand in this place, so hallowed by history, I think of others who ha ve stood here before me. I think of the dreams they had for America, and I think of how each recognized that he needed help far beyond him self in order to make those dreams come true.\n\nToday, I ask your prayers that in the years ahead I may have God\'s help in making decisio ns that are right for America, and I pray for your help so that together we may be worthy of our challenge.\n\nLet us pledge together to mak e these next four years the best four years in America\'s history, so that on its 200th birthday America will be as young and as vital as when it

began, and as bright a beacon of hope for all the world.\n\nLet us go forward from here confident in hope, strong in our faith in one another, sustained by our faith in God who created us, and striving always to serve His purpose.\n

Number of characters in the speech – 9991

Number of sentences in the speech – 68

Number of words in the speech – 2006

2.2 Remove all the stopwords from all three speeches.

Here we are removing stop-words, punctuations and special characters by separately appending them to the stop word list.

1st Speech - President Franklin D. Roosevelt in 1941

Number of words in speech – 1526

After removal numbers of stop-words in speech – 625

2nd Speech - President John F. Kennedy in 1961

Number of words in speech – 1543

After removal numbers of stop-words in speech – 689

3rd Speech - President Richard Nixon in 1973

Number of words in speech – 2006

After removal numbers of stop-words in speech – 832

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)

The top three words that occur most number of times as per the frequency count of each words in that speech are as mentioned below:

1st Speech - President Franklin D. Roosevelt in 1941

Words	Occurrence Times
nation	12
know	10
spirit	9

2nd Speech - President John F. Kennedy in 1961

Words	Occurrence Times
let	16
us	12
world	8

3rd Speech - President Richard Nixon in 1973

Words	Occurrence Times
us	26
let	22
america	21

2.4 Plot the word cloud of each of the speeches of the variable

Word cloud of all words used in the three speeches after removing all the stop words are as below:

1st Speech - President Franklin D. Roosevelt in 1941

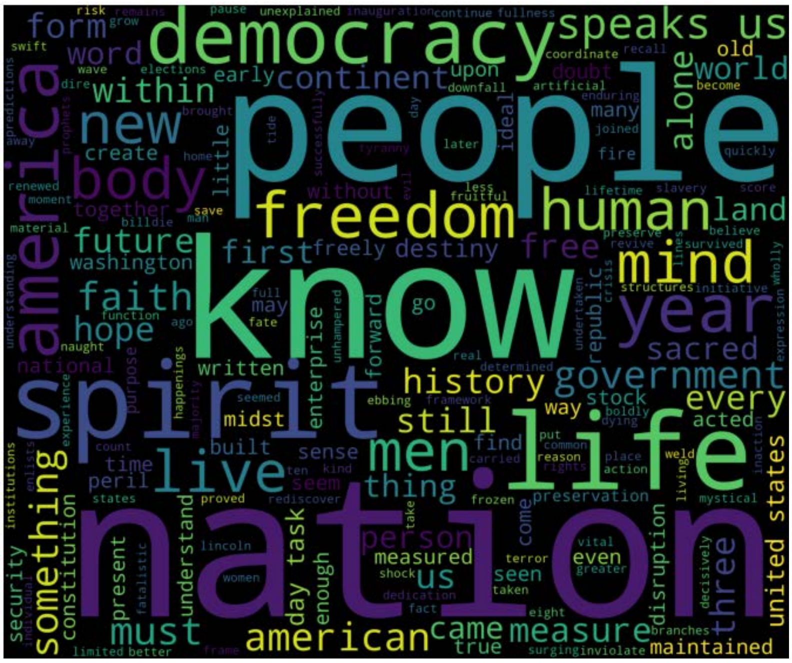


Fig.23

2nd Speech - President John F. Kennedy in 1961

Fig.24

3rd Speech - President Richard Nixon in 1973

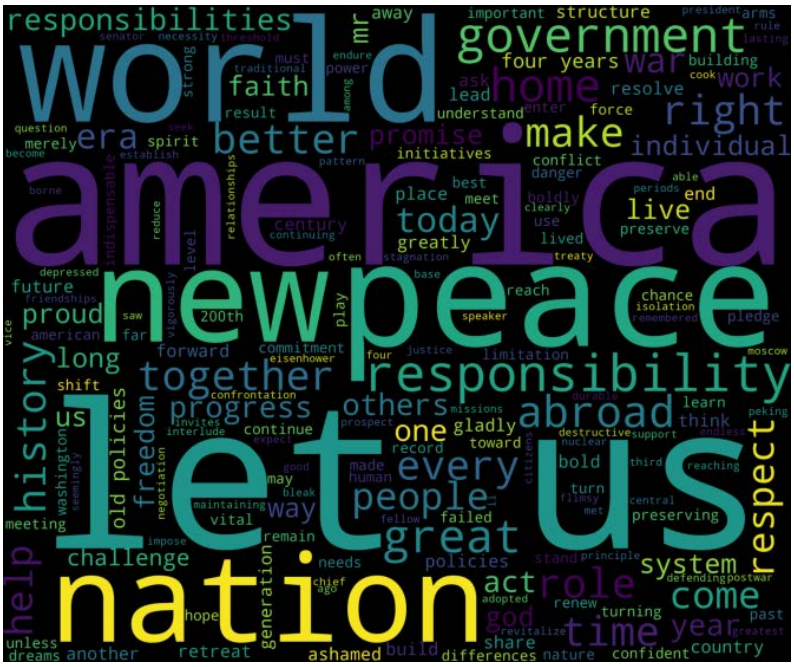


Fig.25

THE END