BUSINESS REPORT

ON

MACHINE LEARNING

By Kshitij Nishant

Table of Contents

Contents

Problem 1:

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.

Answer:

After reading the data,

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

Table 1

Variable Name Description

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.

Description of data:

	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 2

Info About the data

<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)

As we can see, there are two object datatypes, gender and vote. Rest all are integer types.

Shape of data: (1517, 9)

We even checked for null values and found none in any of the columns:

vote	0
age	0
economic.cond.national	0
economic.cond.household	0
Blair	0
Hague	0
Europe	0
political.knowledge	0
gender	0

I checked for duplicate values and found 8 of them and dropped them:

```
Total no of duplicate values = 8
```

	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 3

After dropping, shape of data: (1517, 9), i.e, 1517 rows and 9 columns remain.

Skewness of data is as follow:

Hague	0.146191
age	0.139800
Europe	-0.141891
economic.cond.household	-0.144148
economic.cond.national	-0.238474
political.knowledge	-0.422928
Blair	-0.539514

Inference:

1. "Unnamed: 0" was avariable I dropped that simply represented the index in the data.

2. Number of rows in dataset is 1517 and number of columns is 9.

3. There are a total of 10 variables present from which we dropped "Unnamed: 0".
2 categorical variables: vote, gender
7 numeric variables: age, economic.cond.household, economic.cond.national, Blair, Hague, Europe, political.knowledge

4. I did a descriptive analysis of the data in table 2 above which gives us information like mean, standard deviation, min-max etc of the data.

5. From null value check I was sure there is no null data present.

6. There were 8 duplicate data which I dropped.

7. Skewness is a measure of asymmetry of probability distribution of real-valued variable about its mean. Here, only 2 variables were positively skewed and rest are all negatively skewed with Blair having max skewness.

1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

Answer:

EDA

Univariate Anaysis

1. Age

Minimum age: 24
Maximum age: 93

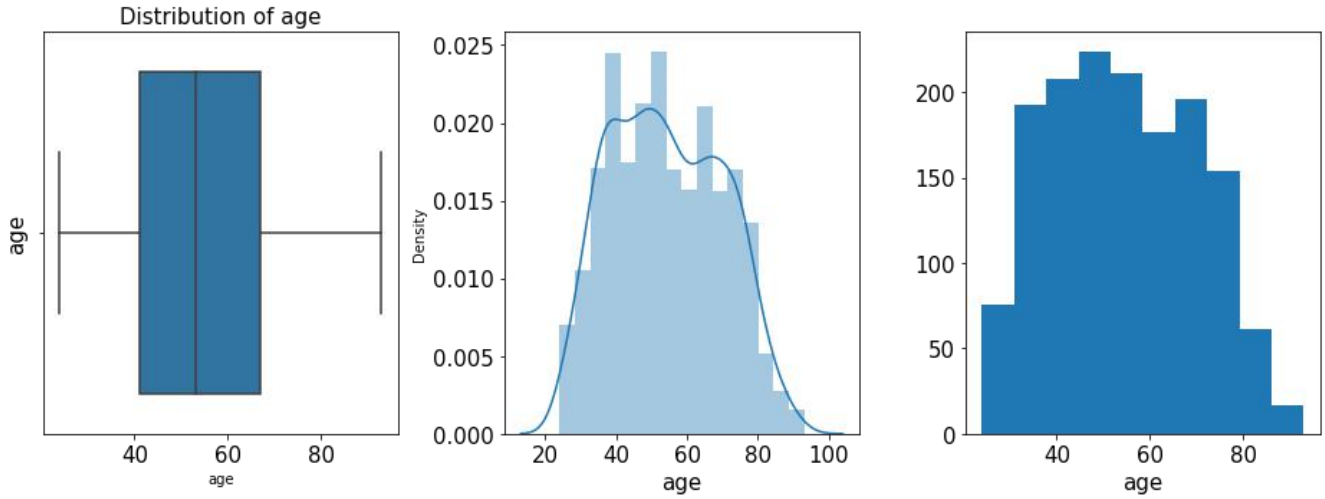

Figure 1

2. Economic conditions of national

Minimum economic.cond.national: 1
Maximum economic.cond.national: 5
Mean value: 3.245220830586684
Median value: 3.0
Standard deviation: 0.8817924638047195
Null values: False

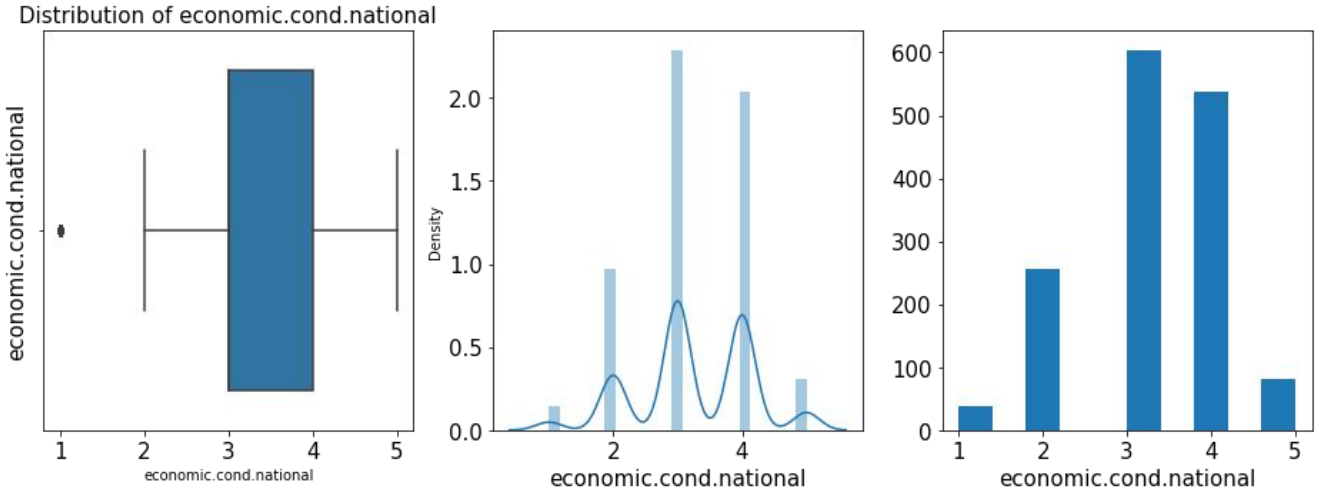

Figure 2

3. Economic conditions of household

Minimum economic.cond.household: 1
Maximum economic.cond.household: 5
Mean value: 3.1377719182597232
Median value: 3.0

Standard deviation: 0.9310694297616856
Null values: False

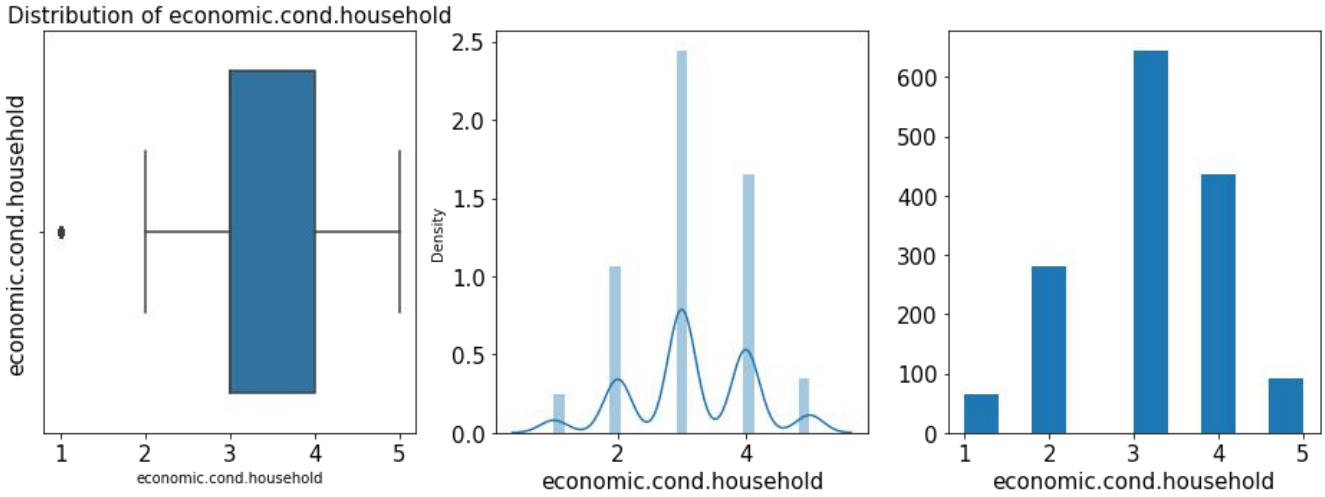

Figure 3

4. Blair

Minimum Blair: 1
Maximum Blair: 5
Mean value: 3.3355306526038233
Median value: 4.0
Standard deviation: 1.1747718854032745
Null values: False

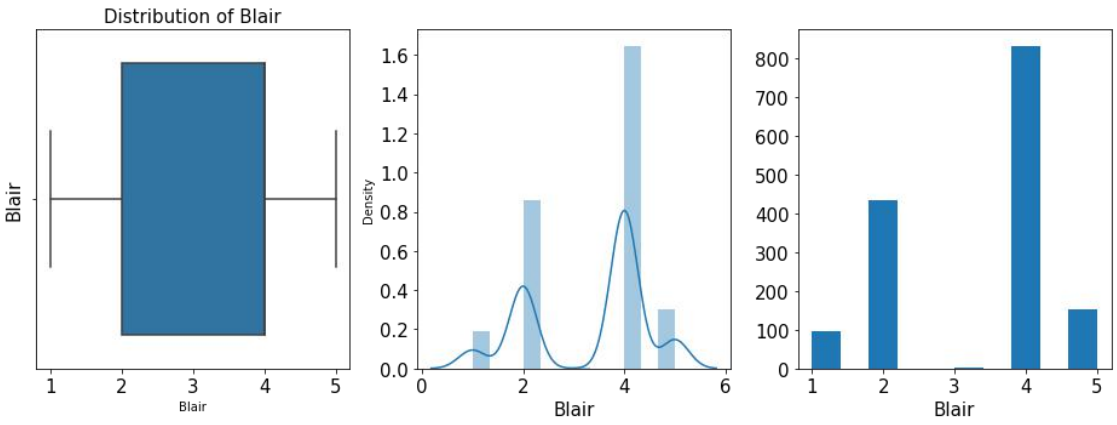

Figure 4

5. Hague

Range of values: 4
Minimum Hague: 1
Maximum Hague: 5
Mean value: 2.7495056031641396
Median value: 2.0
Standard deviation: 1.2324793557178417
Null values: False

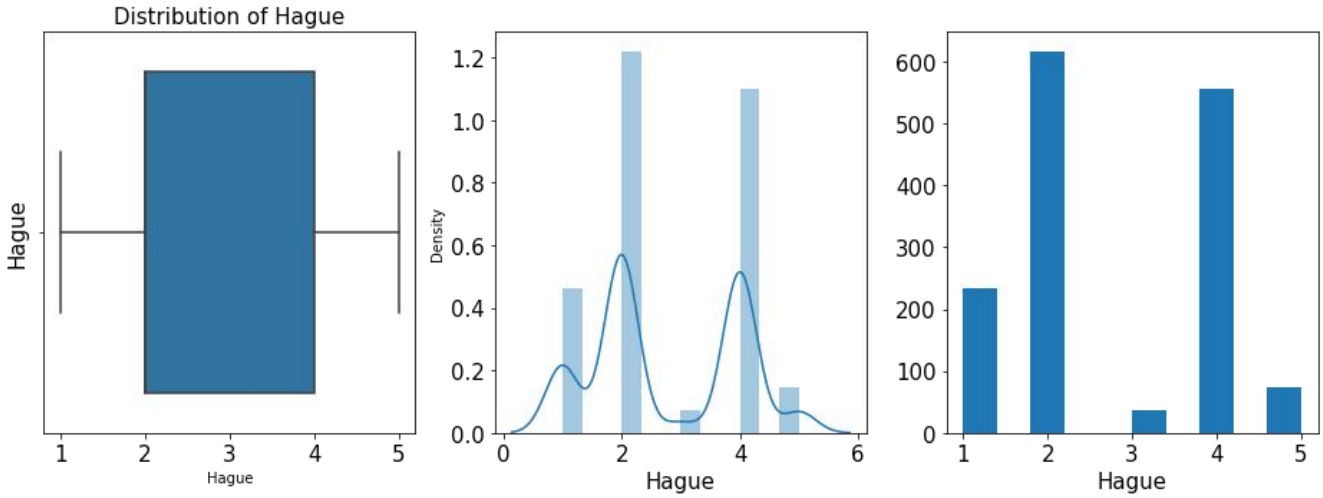

Figure 5

6. Europe

Range of values: 10
Minimum Europe: 1
Maximum Europe: 11
Mean value: 6.7402768622280815
Median value: 6.0
Standard deviation: 3.299043305366668
Null values: False

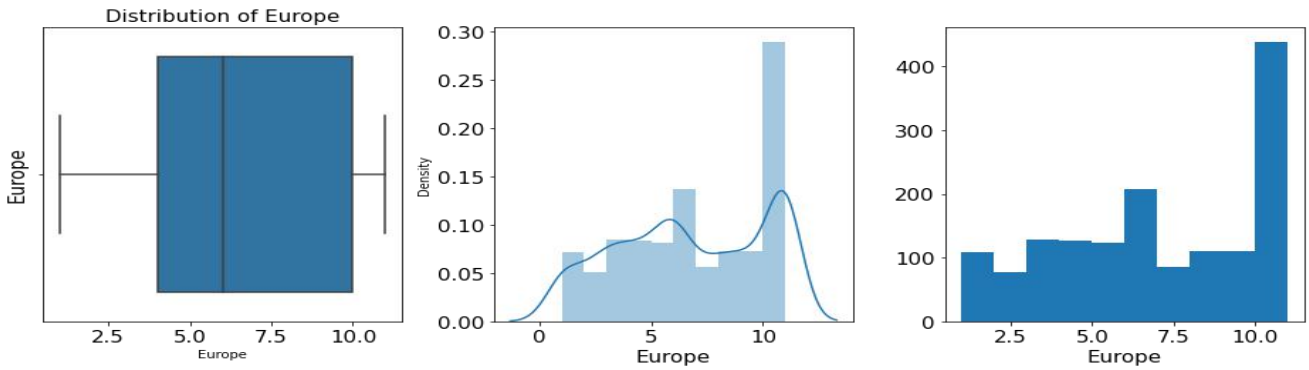

Figure 6

7. Political Knowledge

Range of values: 3
Minimum political.knowledge: 0
Maximum political.knowledge: 3
Mean value: 1.5405405405405406
Median value: 2.0
Standard deviation: 1.0844173188138866
Null values: False

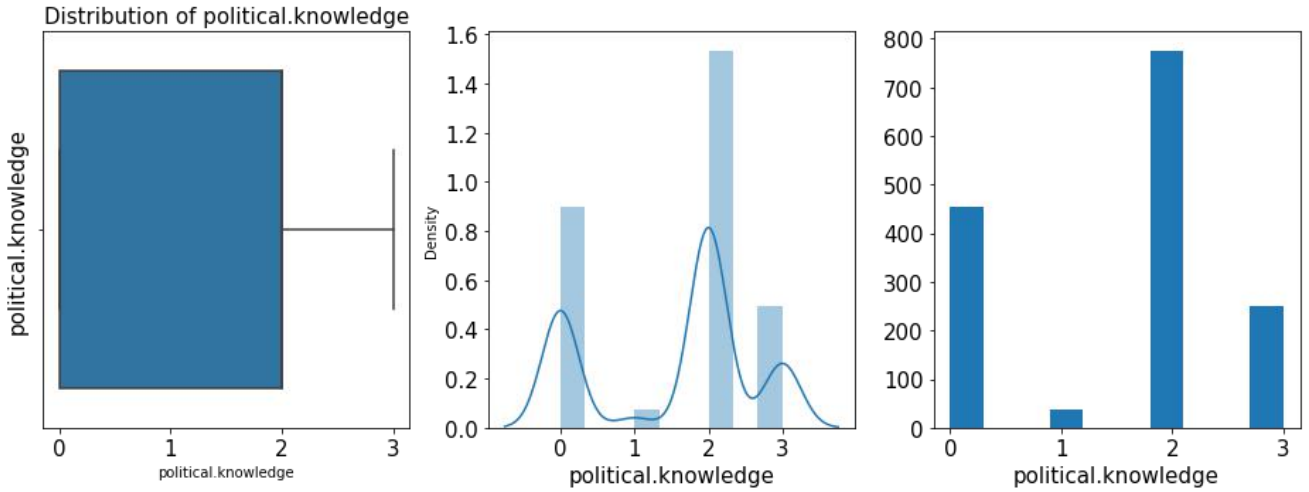

Figure 7

We can see all numerical variables are normally distributed but are multi modal in some instances.
We compare each variable's boxplot:

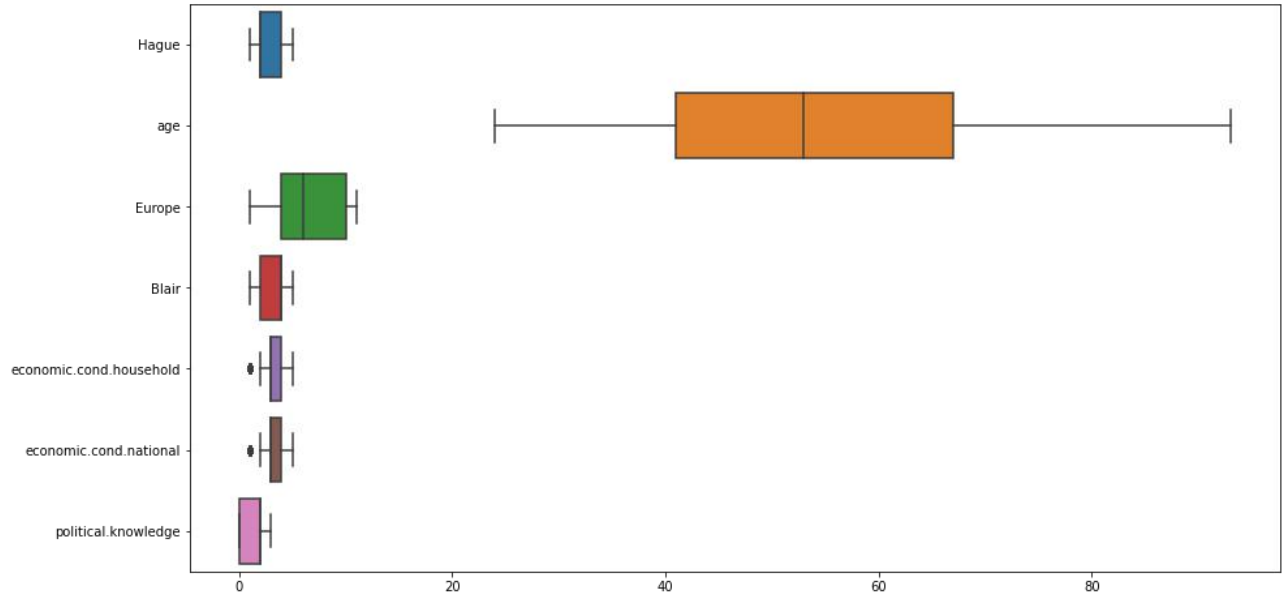

Figure 8

And saw there are outliers in economic household and economic national and will treated them.

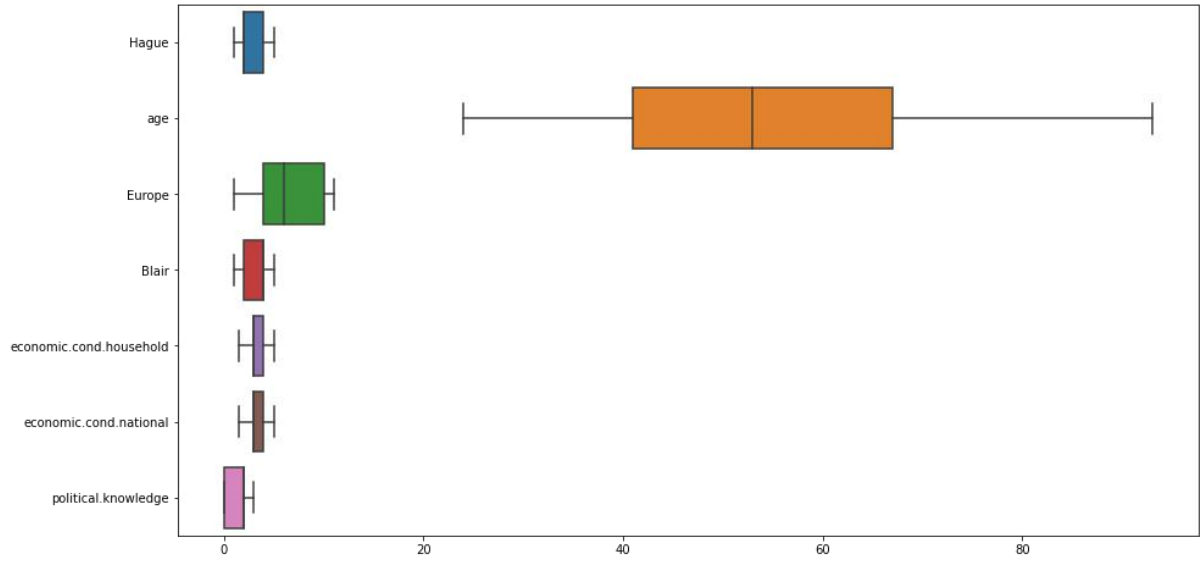

Figure 9

Bivariate Analysis

PairPlot:

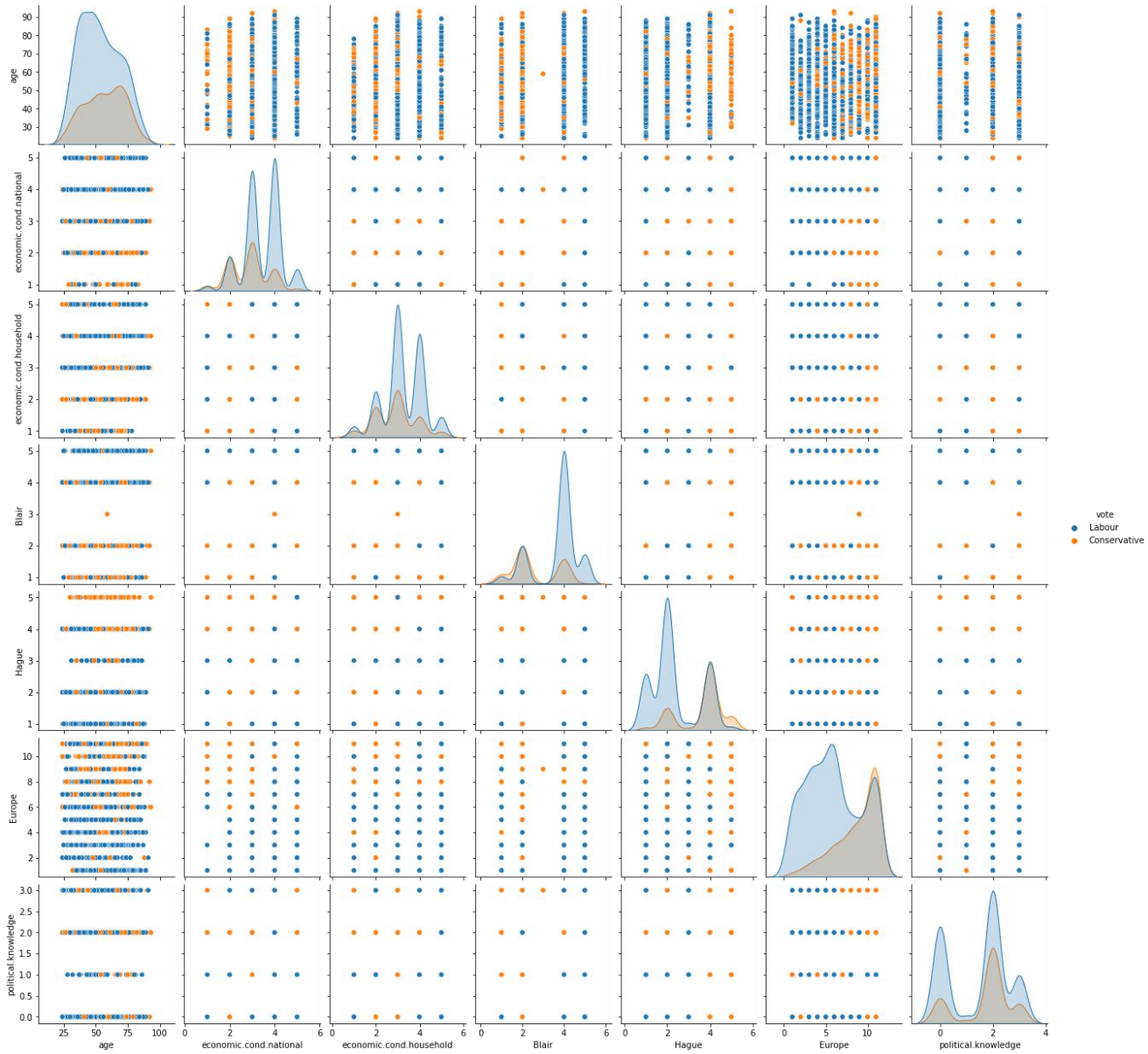

Figure 10

As such there is no strong relation present between variables. We also check with the heatmap and did not find much multicollinearity among the variables.

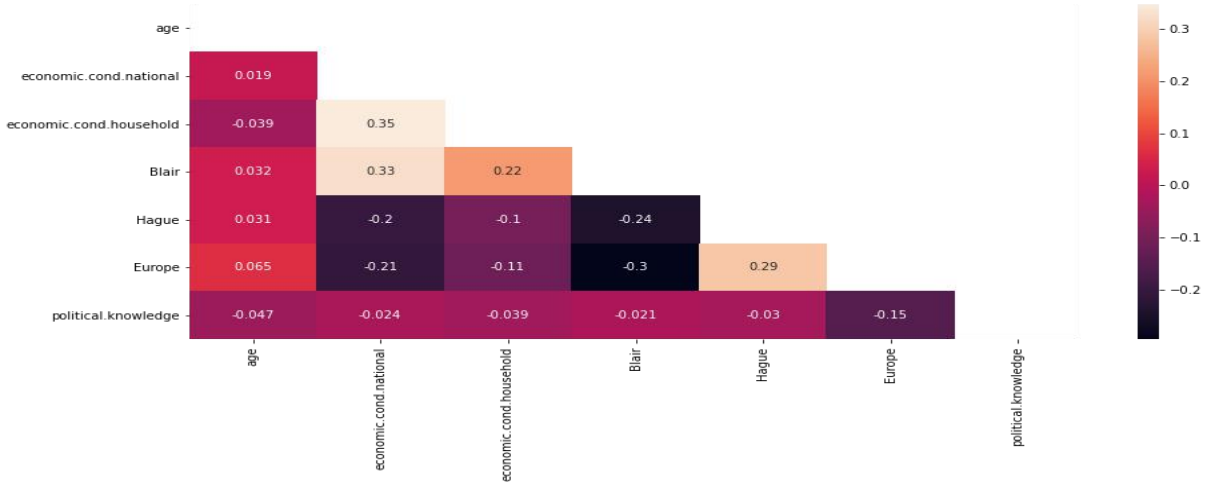

Figure 11

We do analysis of Blair and Age, Hague and Age,Economic.cond.household and age & vote and age:

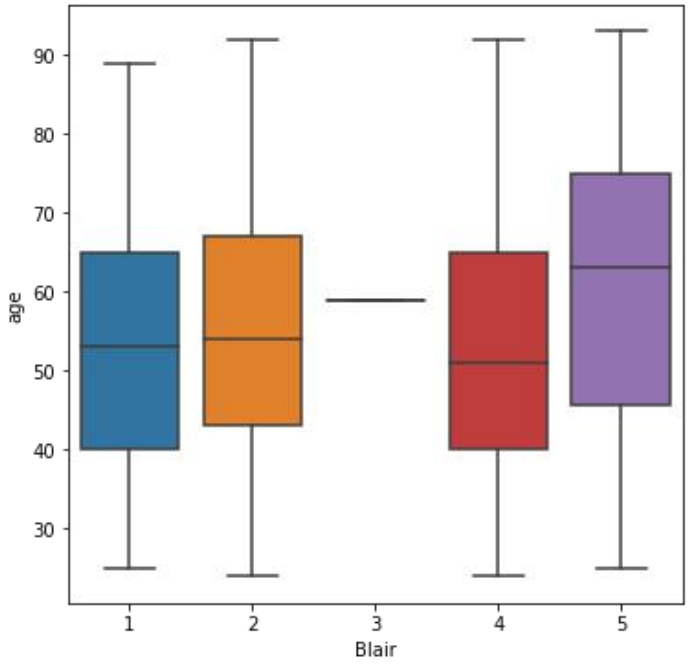

Figure 12

People above age 45 years think Blair doing good job.

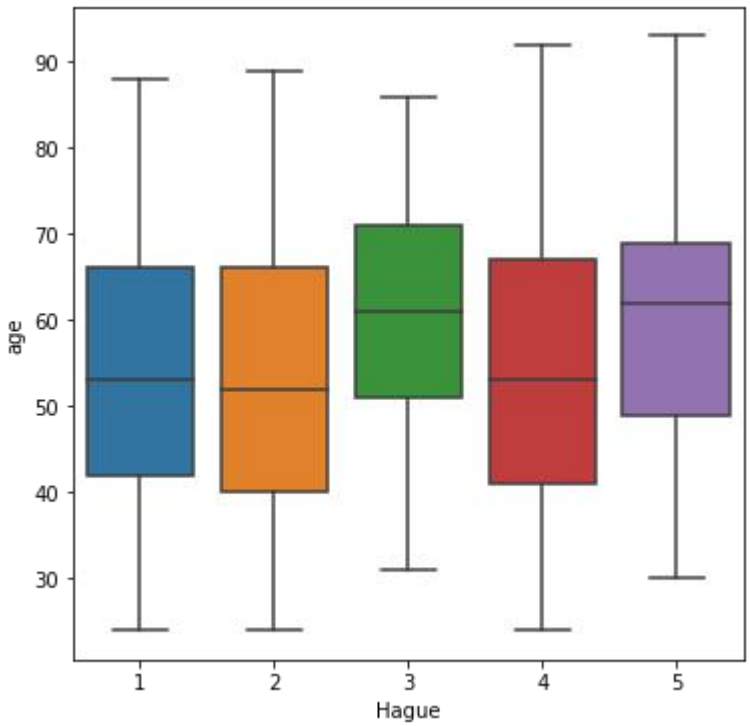

Figure 13

Hague has a slightly more concentration that of Blair for people above 50 years.

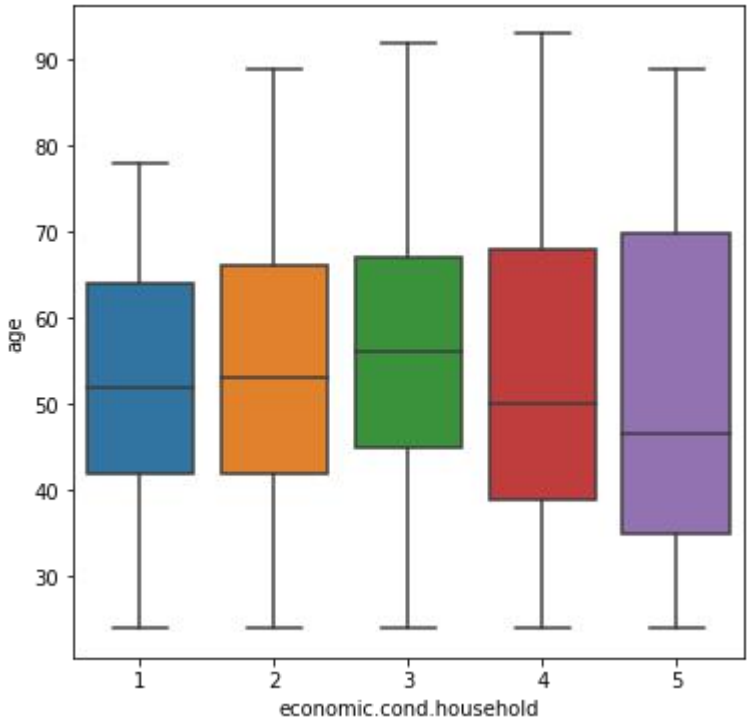

Figure 14

More number of people aged 60 and above voted for Conservative:

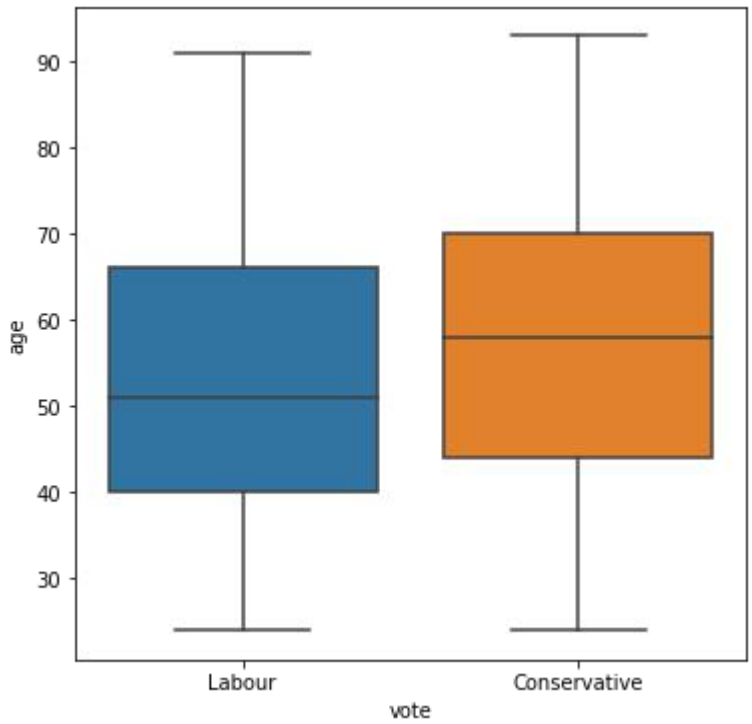

Figure 15

I even made different countplot to see how many of Blair and Hague think of the current election status:

1. Blair count on economic.cond.household

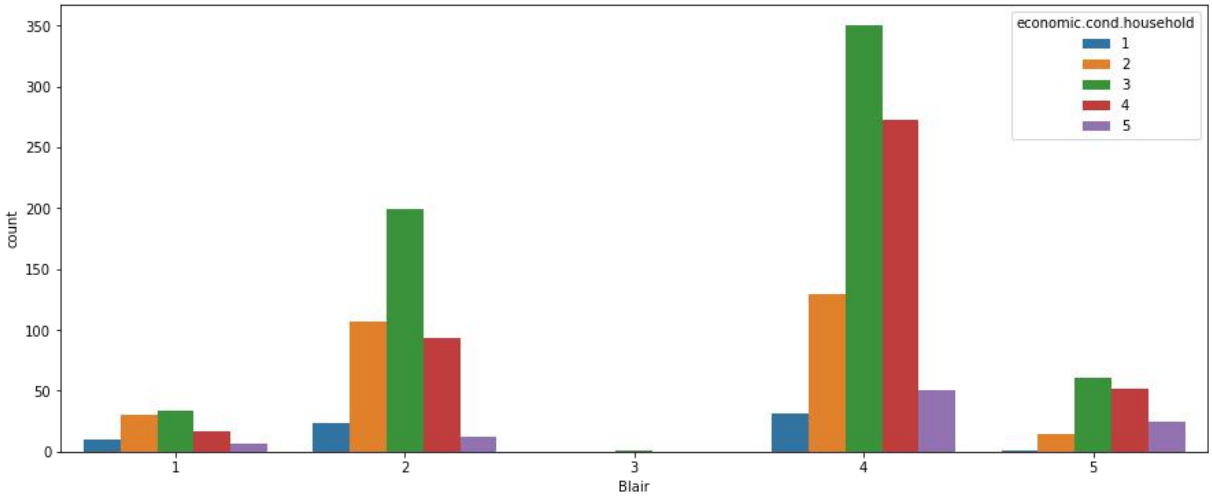

Figure 16

2. Hague count on economic.cond.household

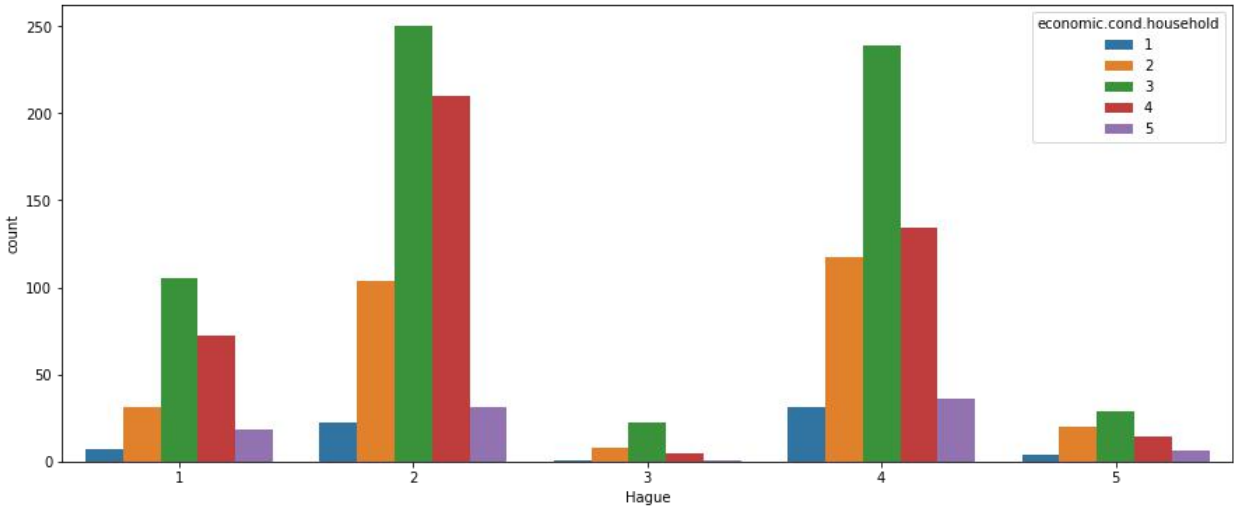

Figure 17

3. Blair count on economic.cond.national

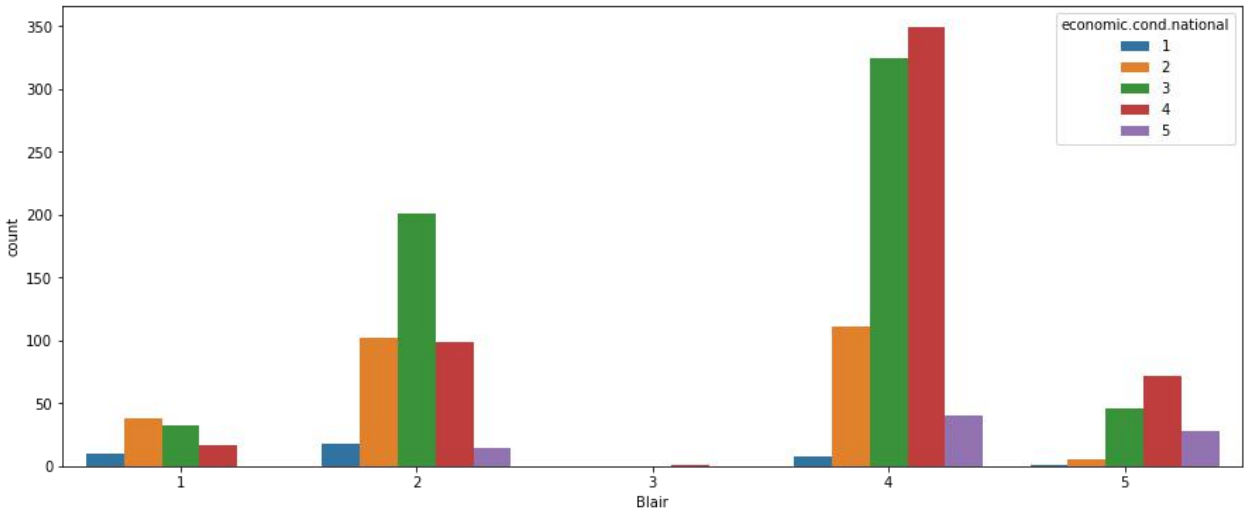

Figure 18

4. Hague count on economic.cond.national

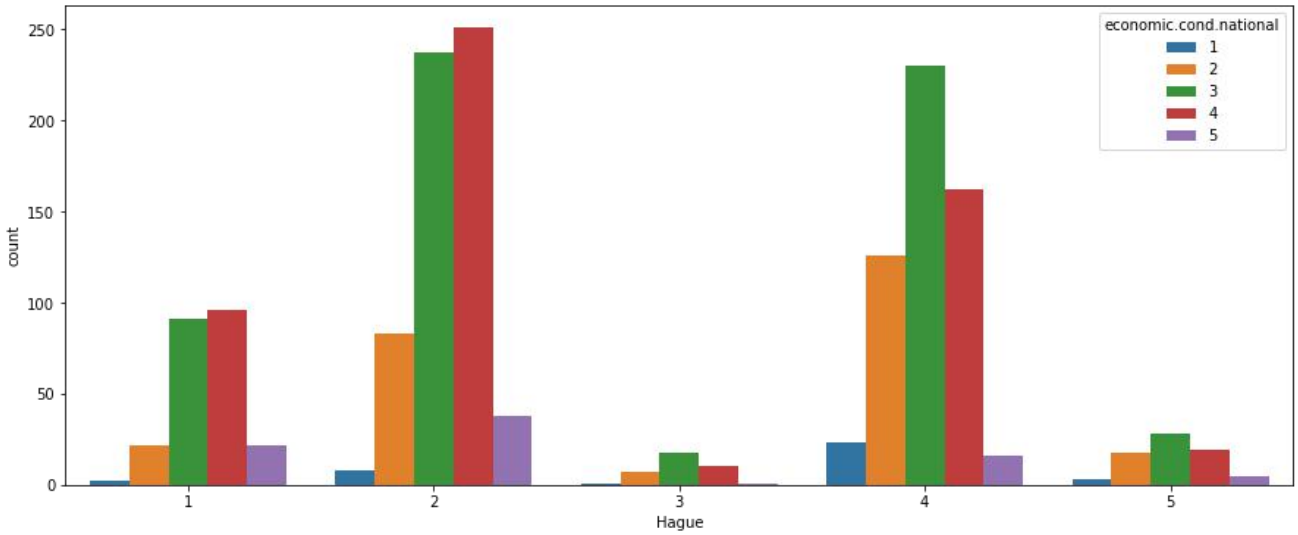

Figure 19

5. Blair count on political knowledge

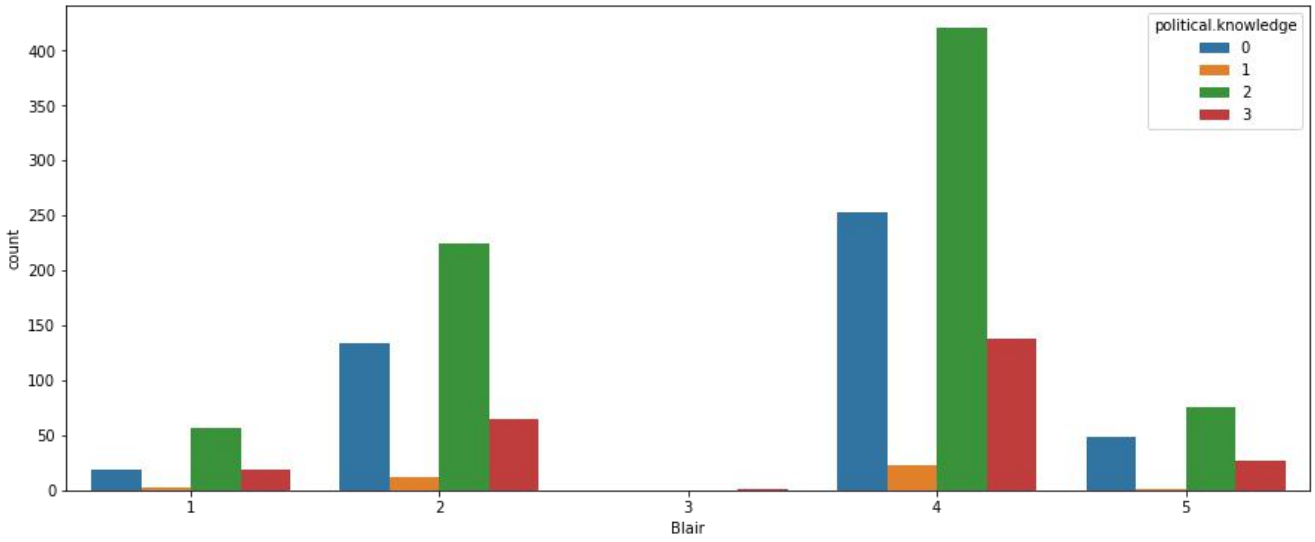

Figure 20

6. Hague count on political.knowledge

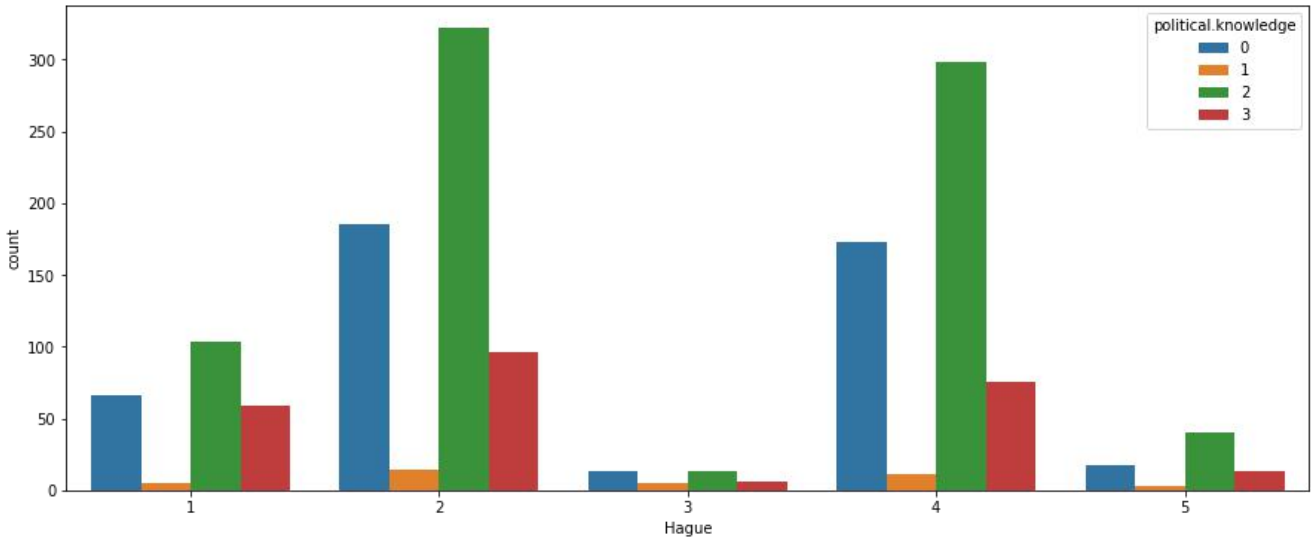

Figure 21

7. Blair count on Europe

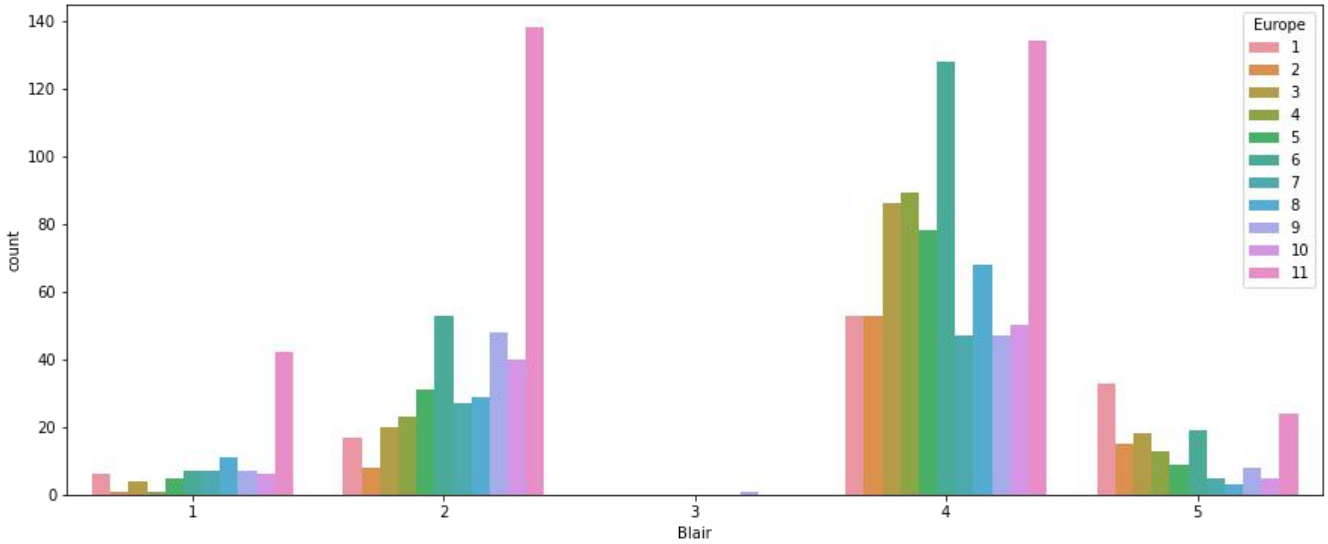

Figure 22

8. Hague count on Europe

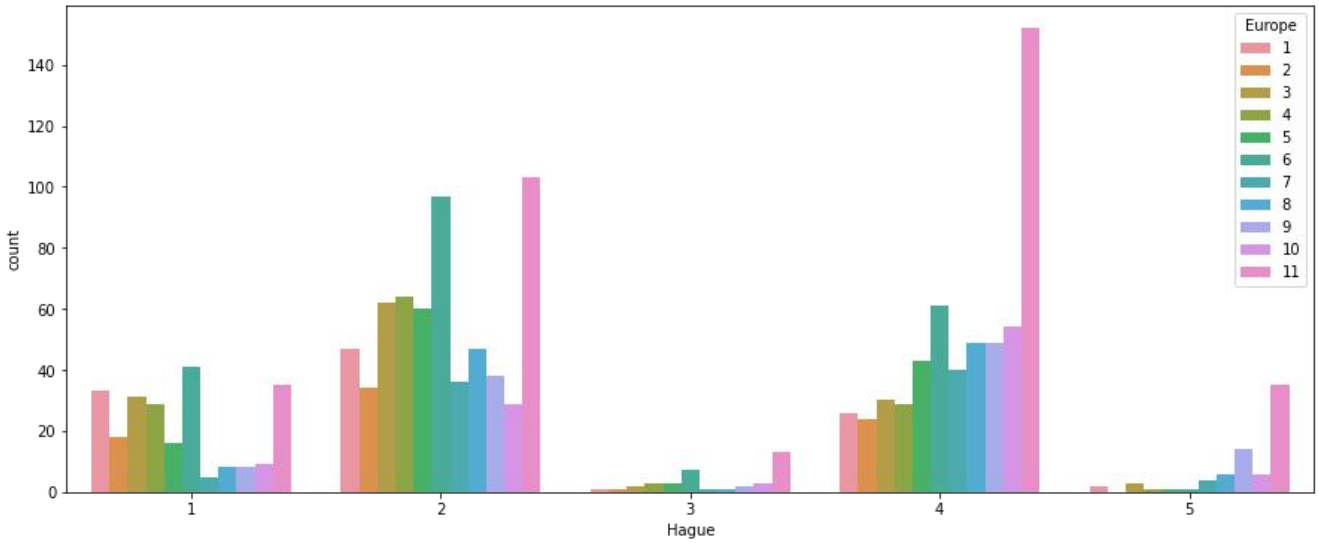

Figure 23

And I checked for the gender vs vote countplot as well:

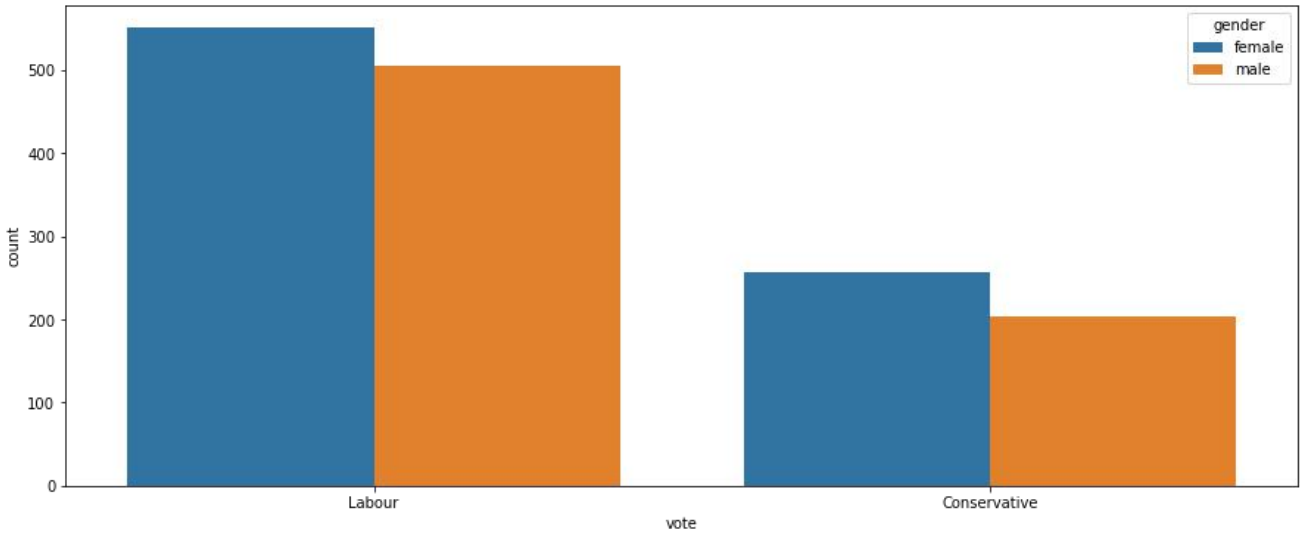

Figure 24

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.

Answer:

I encoded the object datatypes and converted them to integer:

Column Name: vote
['Labour', 'Conservative']
Categories (2, object): ['Conservative', 'Labour']
[1 0]

Column Name: gender
['female', 'male']
Categories (2, object): ['female', 'male']
[0 1]

Scaling is done so that data which belongs to wide range can be brought together in similar relative range and thus bring out the best performance of a model.

We preform scaling while dealing with Linear and Logistic Regression as these are very sensitive to range of datapoints. In addition it's useful in reducing multi-collinearity.
So, it depends on the model weather it requires scaling or not. Usually, the distance-based methods like KNN require scaling as unscaled data can cause a bias.

This is how data looks like before scaling:

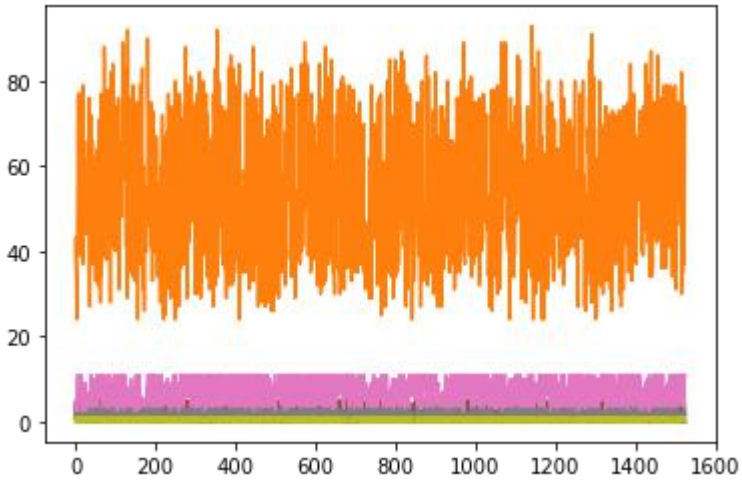

Figure 25

I used Z-score to scale the data and here how it looks after:

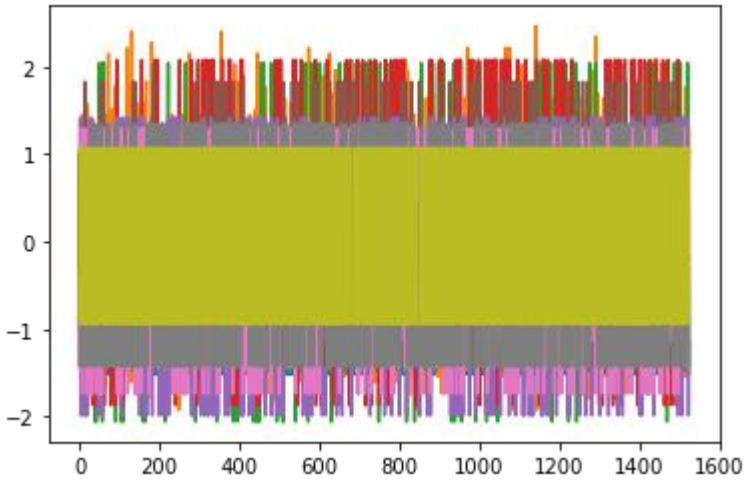

Figure 26

Data Splitting into train and test sets:

Here, the target variable is 'vote':

Vote data distribution

1 0.69677
0 0.30323
Name: vote, dtype: float64

Now we split data in ratio of 70:30 with 30% being test data and 70% being train data:

training set for independent variables is (1061, 8)
training set for dependent variables is (1061,)
test set for independent variables is (456, 8)
test set for independent variables is (456,)

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 behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

Answer:

Logistic Regression Model

I built a simple model at first with the data.
Parameters involved: penalty, solver, max_iter, tol etc

TRAIN SET Results: Accuracy - 83%

TEST SET Results: Accuracy - 83%

Inference:
Model preforms well with no presence of overfitting or under-fitting

Linear Discriminant Analysis(LDA)

Parameters involved: solver,shrinkage, etc

TRAIN SET Results: Accuracy - 83%

TEST SET Results: Accuracy - 83%

Inference:
This model also preformed well with same accuracy of train and test data.
No overfitting or underfitting of data.

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

Answer:

Naive Bayes

Train Accuracy - 83%
Test Accuracy - 82%

Inference:
The model preformed well without under-fitting and over-fitting.

KNN Model

Them main disadvantage of this model is that it's very slow when large volume of data is present.
Parameters involved: n_neighbours, weights, algorithm, metric etc

Train Accuracy - 86%
Test Accuracy - 82%

Inference:
There is a slight overfitting in this model.
I chose nearest neighbours to be 5 because it was giving the best performance.

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

Answer:

Model Tuning is the process of maximizing a model's performance without overfitting or creating too high of a variance. This is achieved by selecting the right 'hyper-parameters'.

Grid Search is one of the most common methods of optimizing the parameters. Here, a set of parameters are taken and the best combination is picked to evaluate the dataset, using cross-validation.

Overfitting means the model works well on Train set but relatively poor on Test set. Under-fitting is the exact opposite of over-fitting.

Bagging Model(Using Random Forest):

This is an ensemble technique. These kinds of techniques are used to combine several base models to get an optimal model. It also improves the performance of existing ML algorithms used in classification or regression. Most commonlys used with tree-based algorithm. It is a parallel method.

TRAIN SET:

Accuracy : 96%

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.90	0.94	307
1	0.96	0.99	0.98	754
accuracy			0.97	1061
macro avg	0.97	0.95	0.96	1061
weighted avg	0.97	0.97	0.97	1061

Confusion Matrix:

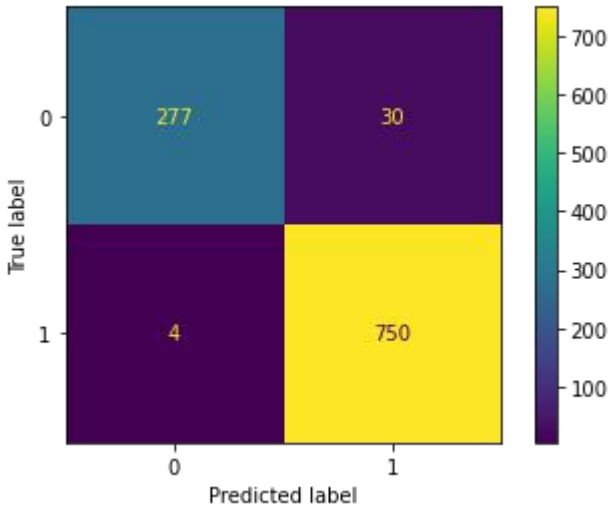

Figure 27

AUC - 99%

ROC Curve:

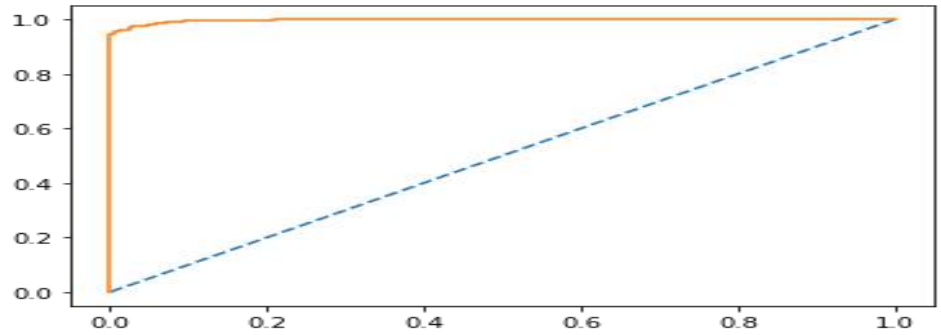

Figure 28

TEST SET:

Accuracy - 82%

Classification Report-

	precision	recall	f1-score	support
0	0.78	0.68	0.73	153
1	0.85	0.90	0.88	303
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.83	0.83	0.83	456

Confusion Matrix :

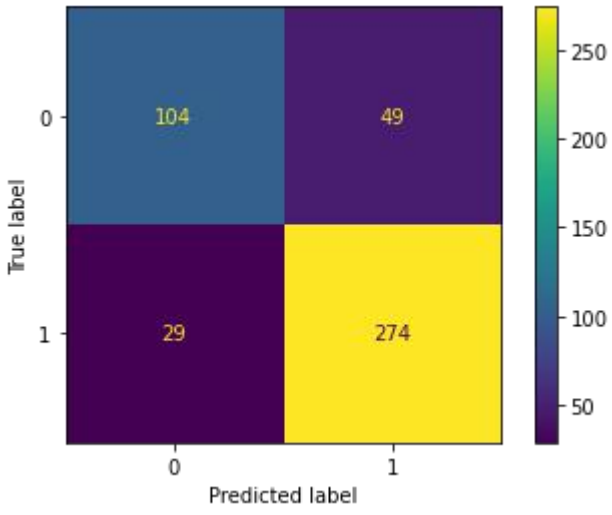

Figure 29

AUC - 89%

ROC Curve:

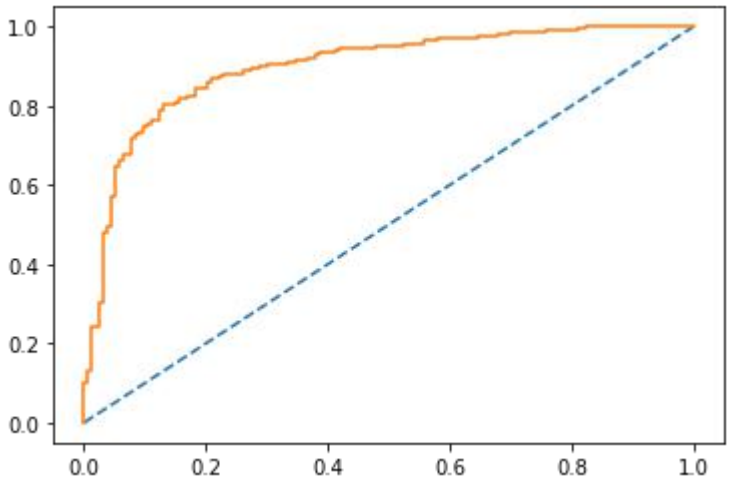

Figure 30

As we can see this model is highly over-fitted.

BOOSTING

Also an ensemble method to convert weak learners to strong learners. Unlike bagging it's a sequential method where results from one weak learner becomes input for another and so on, thus improving the performance of the model.This is an iterative process.

Misclassified data gain a higher weight and examples that are classified correctly will lose weight.Thus, future weak learner focues more on misclassified input data. They are also tree-based methods.

1. ADA BOOSTING

This is used to increase efficiency of binary classifiers, but now used to improve multiclass classifiers as well.

TRAIN SET

Accuracy - 85%

Classification report:

	precision	recall	f1-score	support
0	0.76	0.70	0.73	307
1	0.88	0.91	0.90	754
accuracy			0.85	1061
macro avg	0.82	0.80	0.81	1061
weighted avg	0.85	0.85	0.85	1061

Confusion Matrix:

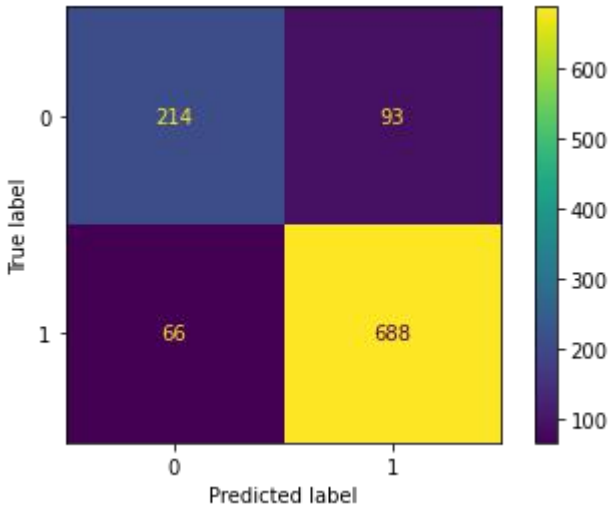

Figure 31

AUC - 91%

ROC Curve:

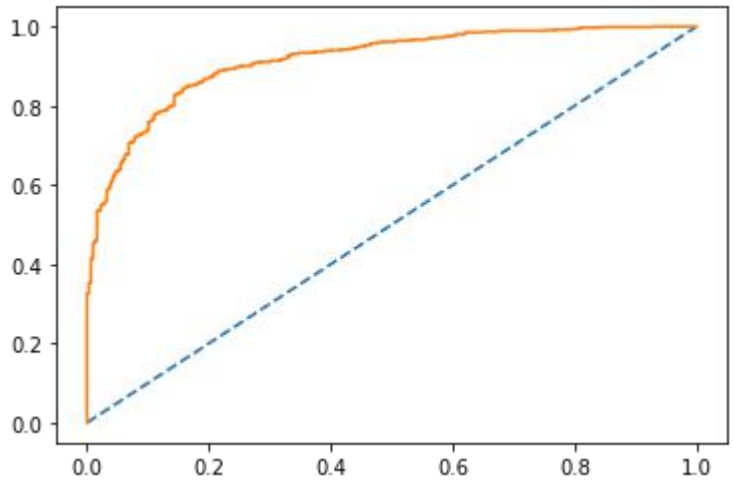

Figure 32

TEST SET

Accuracy - 81%

Classification report:

	precision	recall	f1-score	support
0	0.75	0.67	0.71	153
1	0.84	0.88	0.86	303
accuracy			0.81	456
macro avg	0.79	0.78	0.79	456
weighted avg	0.81	0.81	0.81	456

Confusion Matrix:

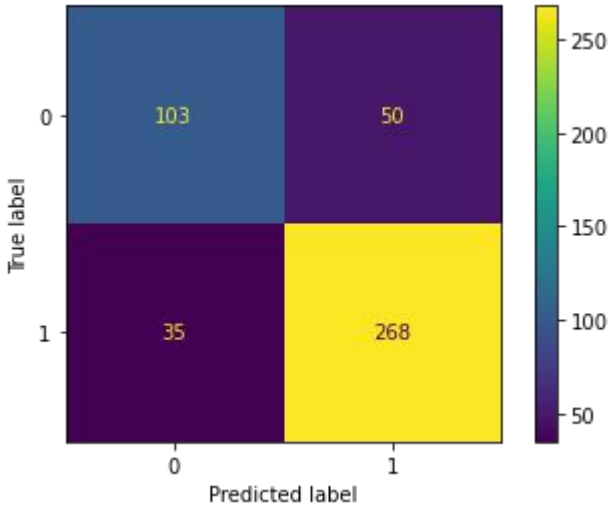

Figure 33

AUC - 87%

ROC Curve -

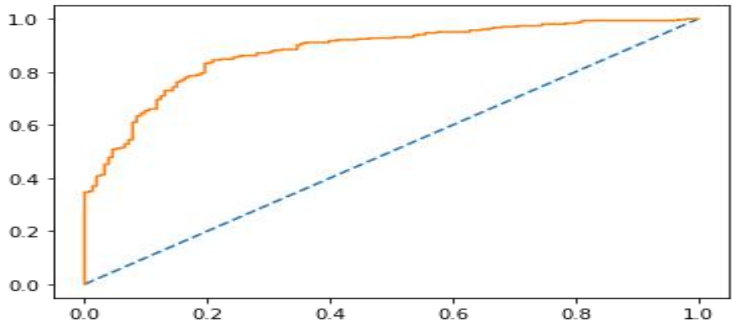

Figure 34

Even here we can see sight overfitting

2. Gradient Boosting

It works by sequentially adding misidentified predictors and under-fitted predictions to the ensemble, ensuring the errors identified previously are corrected. It tries to fit the new predictor to the residual errors made by previous one.

TRAIN SET

Accuracy- 89%

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.78	0.81	307
1	0.91	0.94	0.93	754
accuracy			0.89	1061
macro avg	0.88	0.86	0.87	1061
weighted avg	0.89	0.89	0.89	1061

Confusion Matrix:

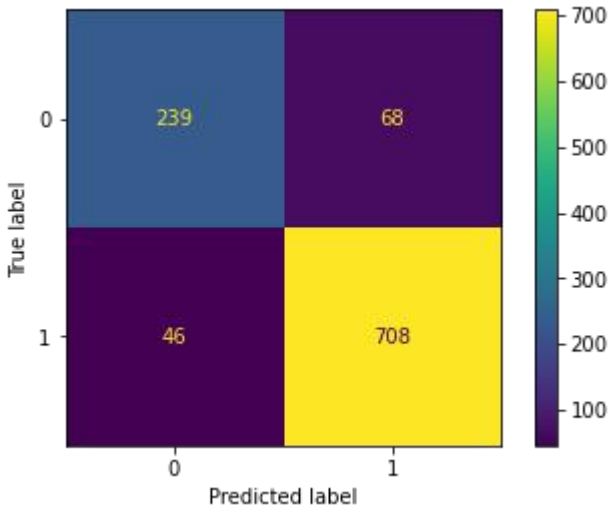

Figure 35

AUC- 95%

ROC Curve-

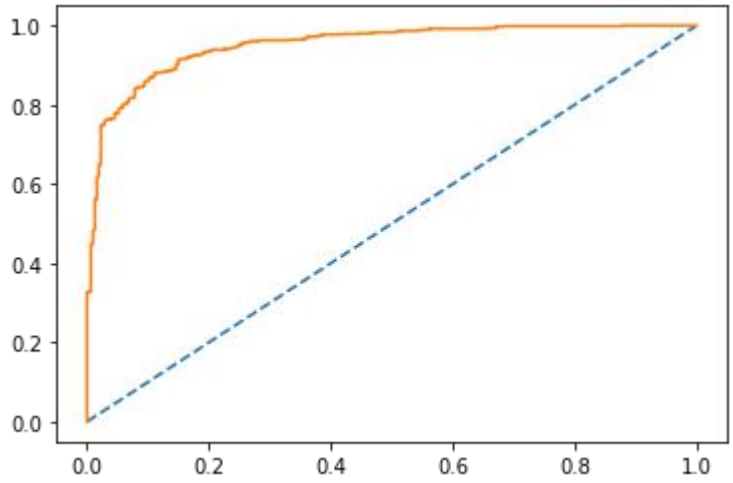

Figure 36

TEST SET

Accuracy - 83%

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.69	0.74	153
1	0.85	0.91	0.88	303

accuracy			0.84	456
macro avg	0.82	0.80	0.81	456
weighted avg	0.83	0.84	0.83	456

Confusion Matrix:

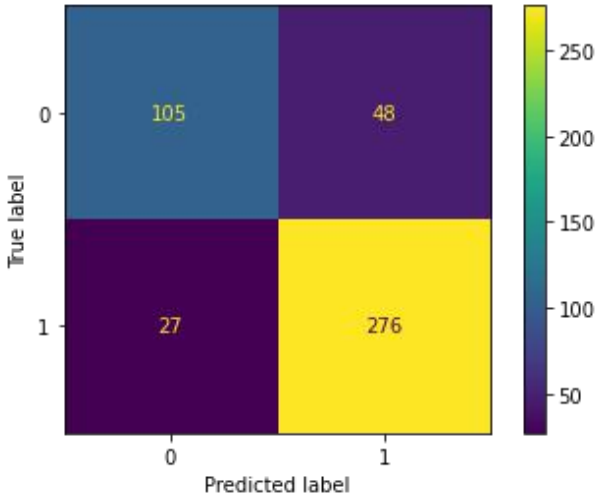

Figure 37

AUC- 89%

ROC Curve:

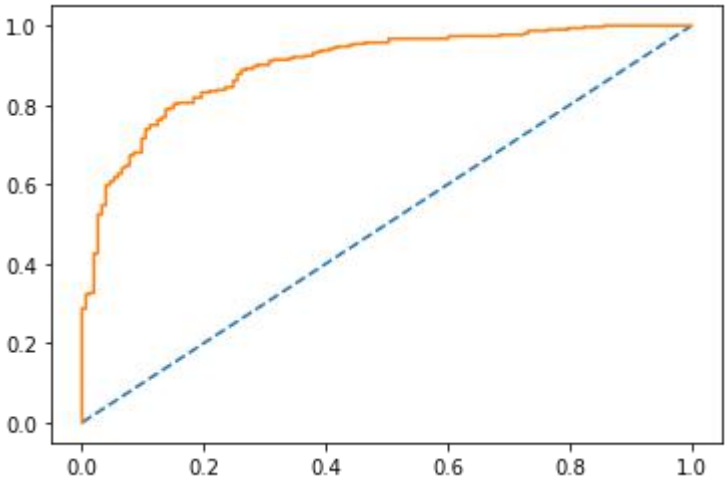

Figure 38

Here also we see slight over-fitting.

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)

Answer:

Logistic Regression Model

(Before Tuning)

TRAIN SET

Classification Report

	precision	recall	f1-score	support
0	0.75	0.64	0.69	307
1	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.81	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Confusion matrix:

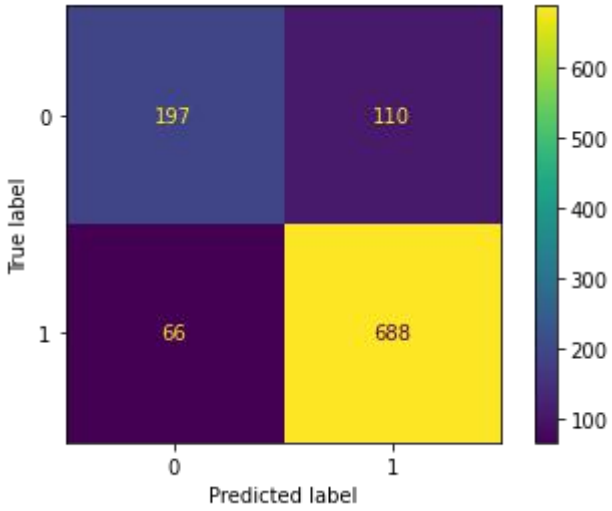

Figure 39

AUC- 89%

ROC Curve:

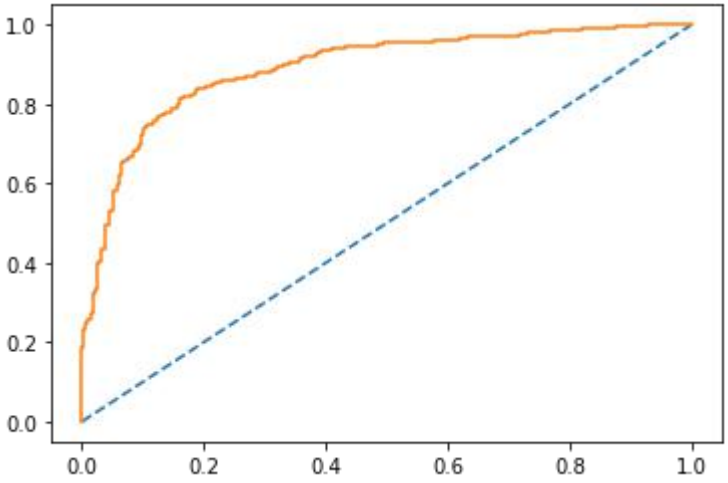

Figure 40

TEST SET

Classification Report:

	precision	recall	f1-score	support
0	0.76	0.73	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456

macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

Confusion Matrix:

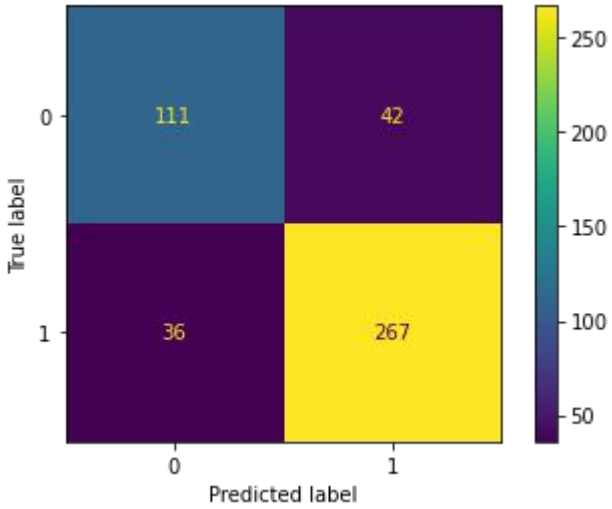

Figure 41

AUC-88%

ROC Curve-

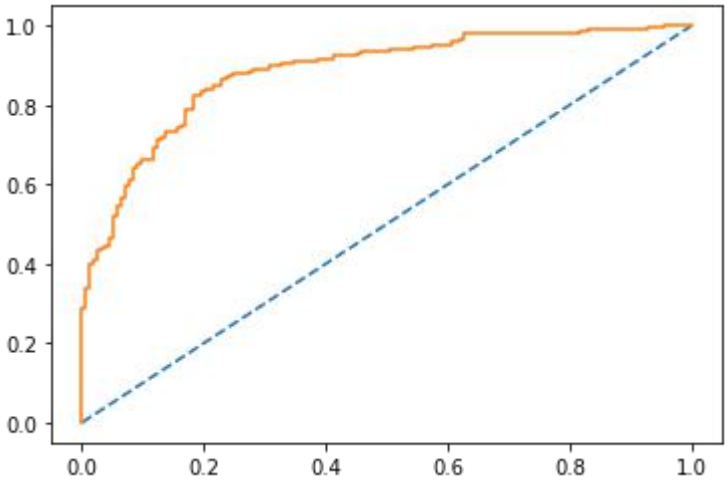

Figure 42

LDA

TRAIN SET

Classification Matrix:

	precision	recall	f1-score	support
0	0.74	0.65	0.69	307
1	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Confusion Matrix:

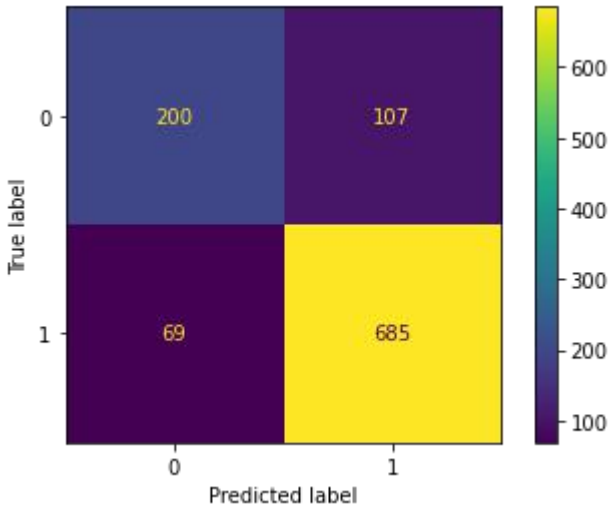

Figure 43

AUC- 89%

ROC Curve-

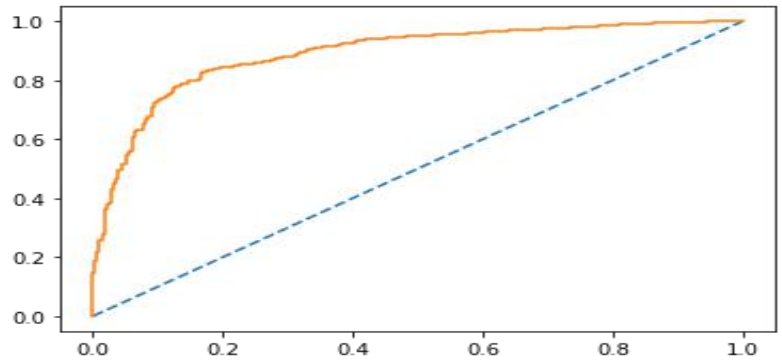

Figure 44

TEST SET

Classification Report:

	precision	recall	f1-score	support
0	0.76	0.73	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

Confusion Matrix

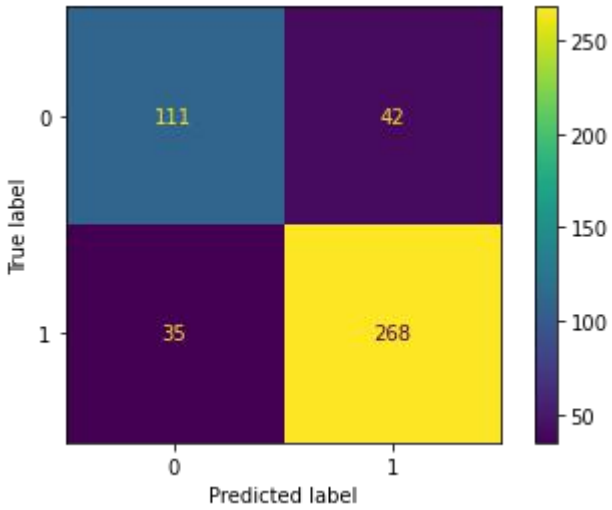

Figure 45

AUC- 88 %

ROC Curve:

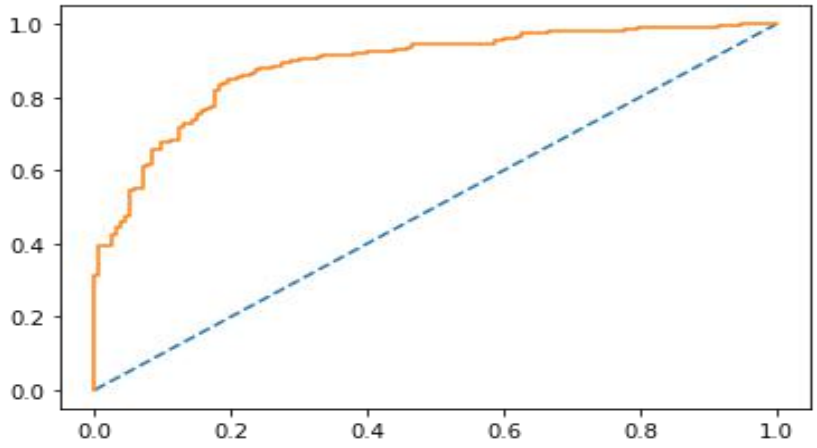

Figure 46

Naive Bayes Model

TRAIN SET

Classification Report

	precision	recall	f1-score	support
0	0.72	0.69	0.71	307
1	0.88	0.89	0.88	754
accuracy			0.83	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.83	0.83	1061

Confusion Matrix:

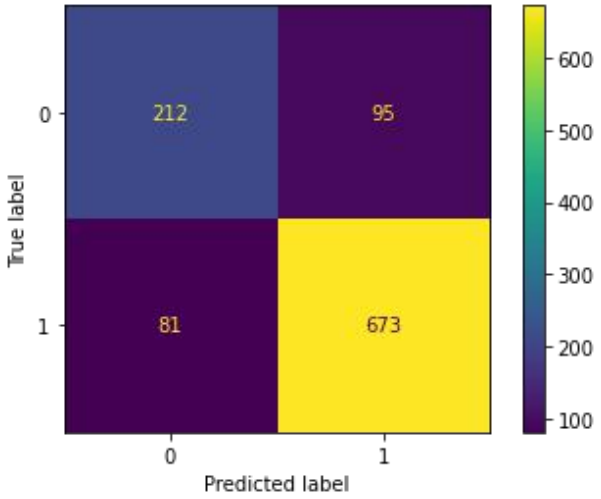

Figure 47

AUC- 88%

ROC Curve:

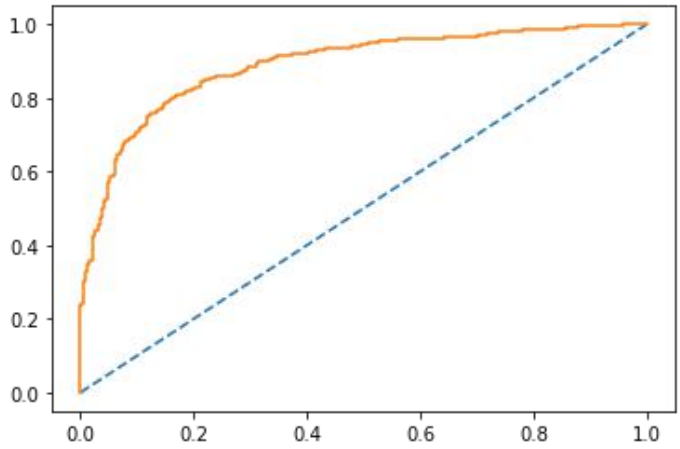

Figure 48

TEST SET

Classification Report:

	precision	recall	f1-score	support
0	0.74	0.73	0.73	153
1	0.87	0.87	0.87	303
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456

weighted avg	0.82	0.82	0.82	456

Confusion Matrix:

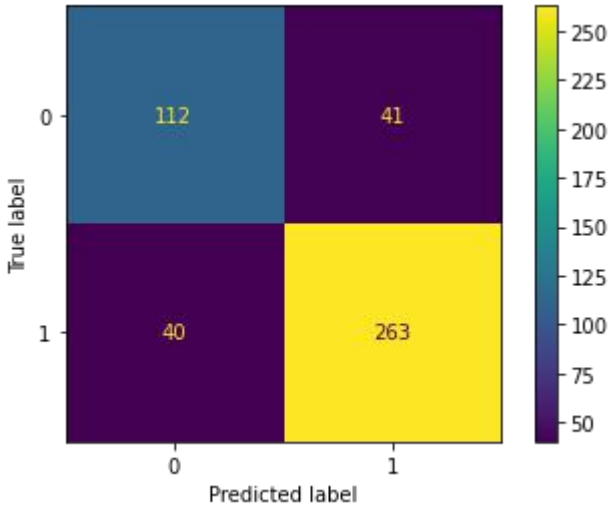

Figure 49

AUC- 87%

ROC CURVE:

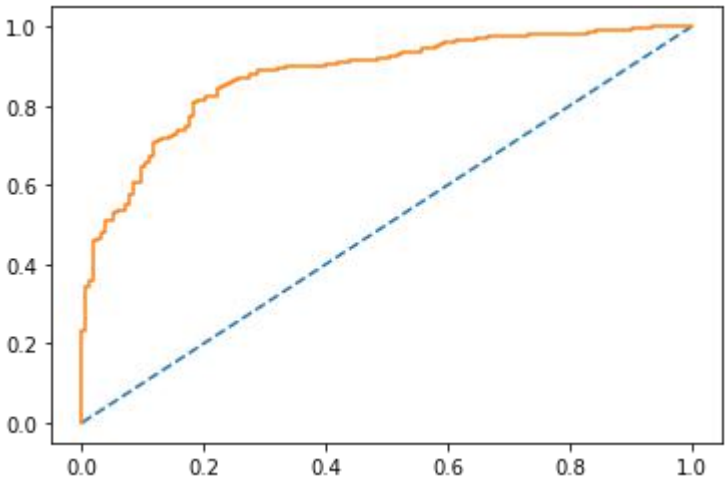

Figure 50

KNN MODEL

TRAIN SET

Classification Report

	precision	recall	f1-score	support
0	0.80	0.68	0.74	307
1	0.88	0.93	0.90	754
accuracy			0.86	1061
macro avg	0.84	0.81	0.82	1061
weighted avg	0.85	0.86	0.85	1061

Confusion Matrix

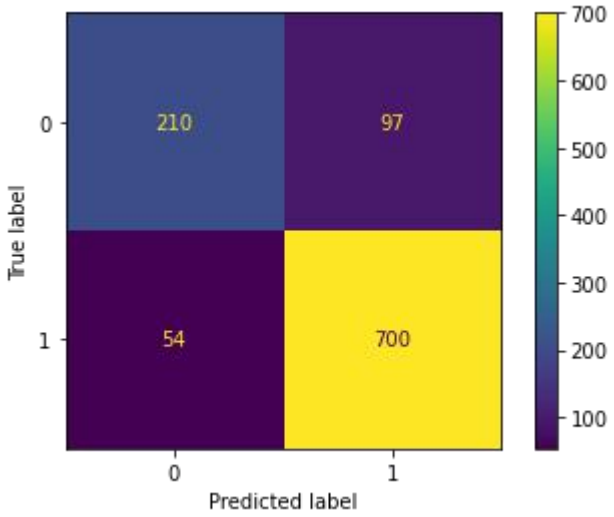

Figure 51

AUC - 92%

ROC Curve:

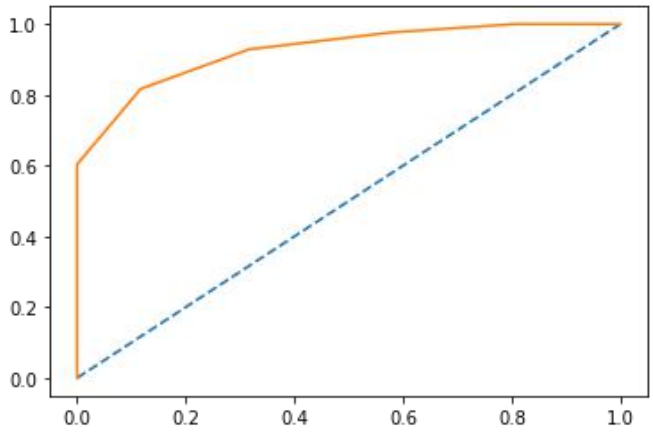

Figure 52

TEST SET

Classification Report

	precision	recall	f1-score	support
0	0.78	0.65	0.71	153
1	0.84	0.91	0.87	303
accuracy			0.82	456
macro avg	0.81	0.78	0.79	456
weighted avg	0.82	0.82	0.82	456

Confusion Matrix

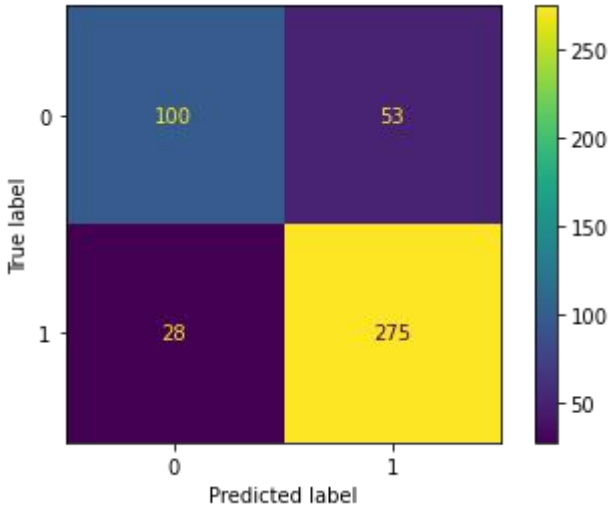

Figure 53

AUC - 86%

ROC Curve-

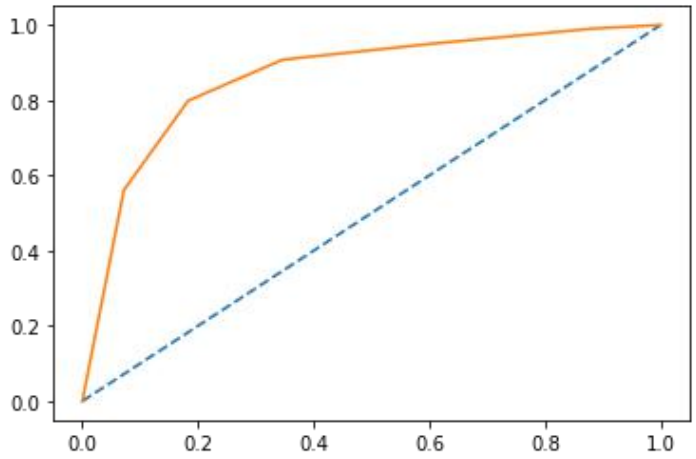

Figure 54

Models after tuning are also checked:

Tuned Logistic Regression Model

TRAIN SET

Accuracy - 83%

Classification report:

	precision	recall	f1-score	support
0	0.75	0.64	0.69	307
1	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.81	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Confusion Matrix:

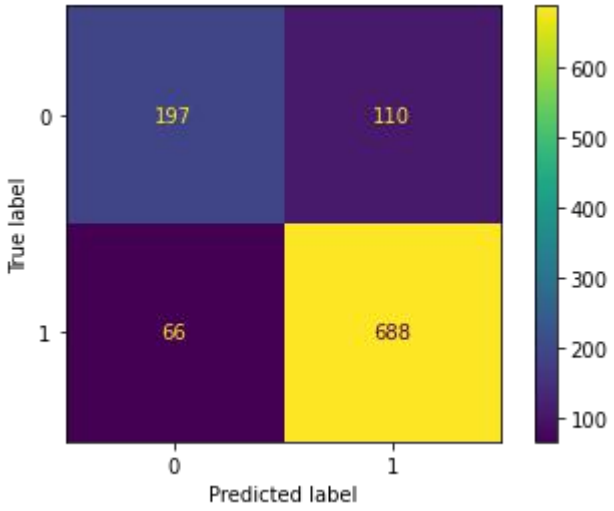

Figure 55

AUC: 89%

ROC Curve-

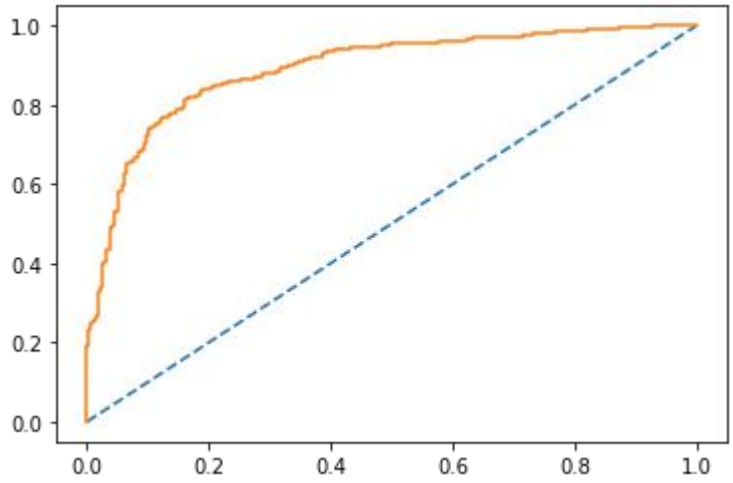

Figure 56

TEST SET

Accuracy: 83%

Classification Report:

	precision	recall	f1-score	support
0	0.76	0.73	0.74	153
1	0.86	0.88	0.87	303

accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

Confusion Matrix:

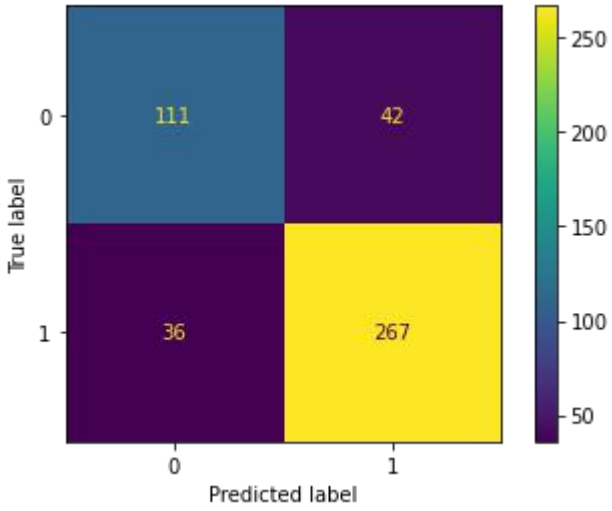

Figure 57

AUC- 88%

ROC Curve-

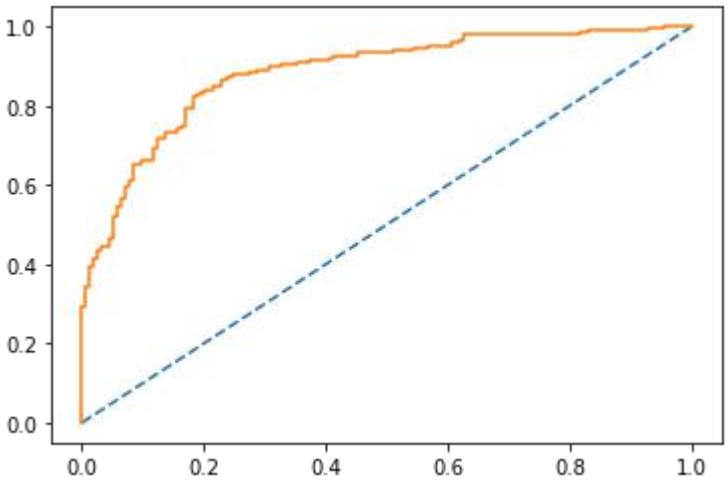

Figure 58

TUNED KNN MODEL

TRAIN SET

Accuracy- 85%

Classification Report:

	precision	recall	f1-score	support
0	0.78	0.66	0.72	307
1	0.87	0.93	0.90	754
accuracy			0.85	1061
macro avg	0.83	0.80	0.81	1061
weighted avg	0.85	0.85	0.85	1061

AUC- 89%

ROC Curve-

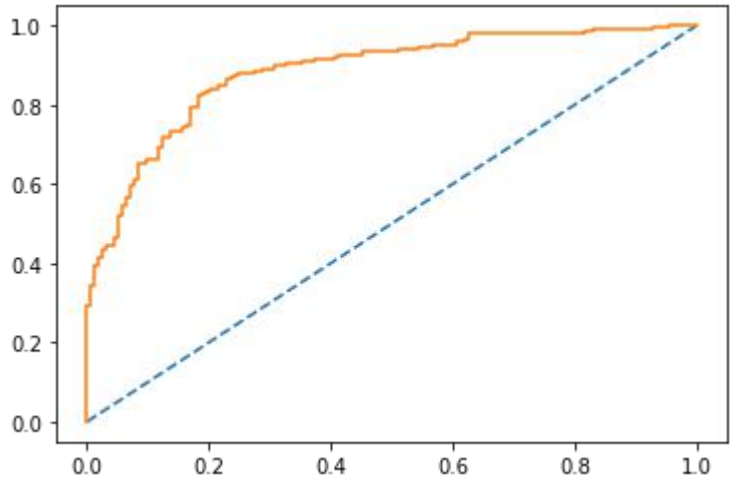

Figure 59

TEST SET

Accuracy - 82%

Classification Report:

	precision	recall	f1-score	support

0	0.78	0.65	0.71	153
1	0.84	0.91	0.87	303
accuracy			0.82	456
macro avg	0.81	0.78	0.79	456
weighted avg	0.82	0.82	0.82	456

AUC-87%

ROC Curve-

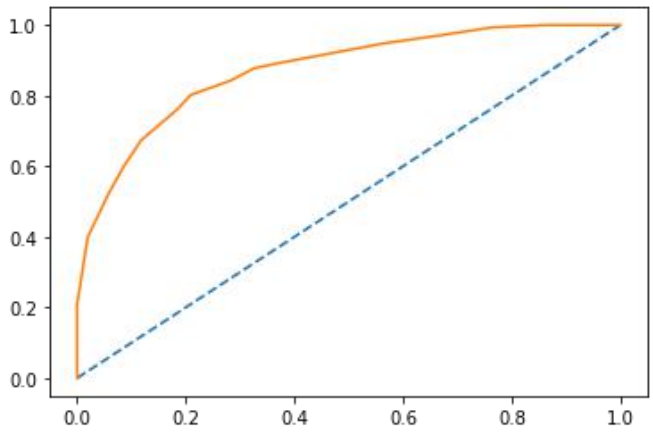

Figure 60

For KNN I selected n_neighbours = 5 to be the best parameter.

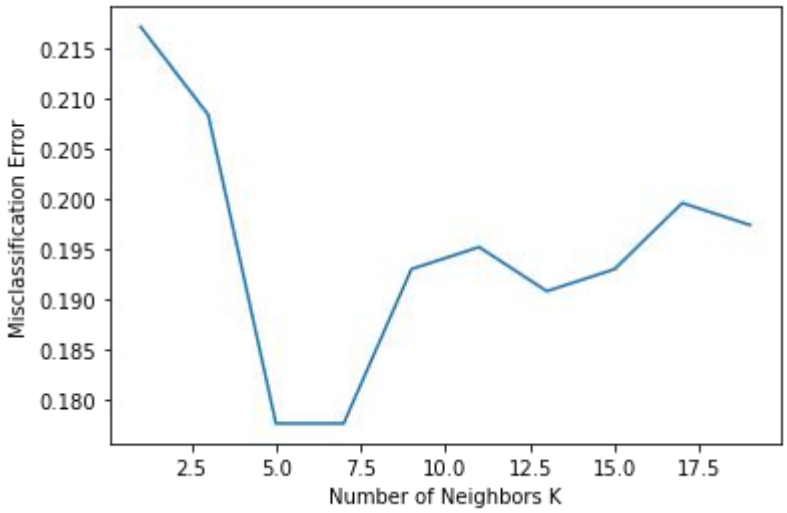

Figure 61

TUNED ADA BOOSTING

TRAIN SET

Accuracy - 77%

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.41	0.56	153
1	0.76	0.96	0.85	303
accuracy			0.78	456
macro avg	0.81	0.69	0.70	456
weighted avg	0.79	0.78	0.75	456

Confusion Matrix:

[[63 90]
[11 292]]

TEST SET

Accuracy- 79%

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.41	0.56	153
1	0.76	0.96	0.85	303
accuracy			0.78	456
macro avg	0.81	0.69	0.70	456
weighted avg	0.79	0.78	0.75	456

Confusion Matrix:
[[63 90]
 [11 292]]

TUNED GRADIENT BOOSTING

TRAIN SET

Accuracy - 88%

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.41	0.56	153
1	0.76	0.96	0.85	303
accuracy			0.78	456
macro avg	0.81	0.69	0.70	456
weighted avg	0.79	0.78	0.75	456

Confusion Matrix:
[[63 90]
 [11 292]]
TEST SET

Accuracy- 84%

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.69	0.74	153
1	0.85	0.91	0.88	303
accuracy			0.84	456
macro avg	0.83	0.80	0.81	456
weighted avg	0.84	0.84	0.83	456

Confusion Matrix:
[[106 47]
 [27 276]]

Model Comparison

We built many different models and even tuned them to check which of them gives the best results. After going through all the parameters I decide to consider Logistic Regression model as it was the most stable model of all with a good train and test accuracy of 83%. It didn't have any under-fitting or over-fitting as well. It had the most optimal recall, precision and f1-score as well.

So, Logistic Regression model was the best optimised model for the given dataset.

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

Answer:

Insights and Recommendations:

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

1. *Using Logistic Regresssion model without scaling for predicting the outcome as it has the best optimised performance.*
2. *Hyper-parameters tuning is an important aspect in building a model. There are limitations to this as to process these combinations, it is consumes a lot of power. But it does provide better results.*
3. *Gathering more data will help in improving the results.*
4. *We can also create a function in which all models predict the outcome in sequence. This will help in better understanding and probability of what the outcome will be.*
5. *Blair has more count points in terms of economic household than Hague.*
6. *Blair has more count points in terms of economic national than Hague.*
7. *Even if we see the graphs and data they suggest that in the whole Europe, Blair is leading.*
8. *In terms of political knowledge also Blair is considered better.*

9. So according to the model and overall data analysis I suggest Blair has a higher chance of winning and covering maximum seats after the election ends.

End of First Question

Problem 2:

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

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)

Answer:

Characters Used:

1. Number of characters in Roosevelt speech: 7571
2. Number of characters in Nixon speech: 9991
3. Number of characters in Kennedy speech: 7618

Words Used:

1. Number of words in Roosevelt speech: 1536
2. Number of words in Nixon speech: 2028
3. Number of words in Kennedy speech: 1546

Sentences used:

1. Number of sentences in Roosevelt speech: 68
2. Number of sentences in Nixon speech: 69
3. Number of sentences in Kennedy speech: 52

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.

Answer:

To remove stopwords, there is package called “stopwords” in nltk.corpus library.

The stopwords library contain all stop words like ‘and’, ‘a’, ‘is’, ‘.’, ‘of’ etc, that usually don’t have any importance in understanding the sentiments in machine learning algorithm. These stopwords present in the package are universally accepted stopwords and we can add more using the .extend() function or remove them as per our requirement.
We also need to specify the language we are working on with before defining the functions, as there are many language packages. Stemming helps the processor understand the words that have similar meaning.

Word count before removal of stopwords is given in the above question and after removal of stopwords the wrod count is as follows:

1. Word count after cleaning Roosevelt speech: 632
2. Word count after cleaning Kennedy speech: 697
3. Word count after cleaning Nixon speech: 836

1. *Sample Sentence for Roosevelt’s speech after removal of stopwords*:

“national day inauguration since 1789 people renewed sense dedication united states washington day task people create weld together nation lincoln day task people preserve nation disruption within day task people save nation institutions disruption without us come time midst swift happenings pause moment take stock recall

place history rediscover may risk real peril inaction lives nations determined count years lifetime human spirit life man three score years ten little little less life nation fullness measure live men doubt men believe democracy form government frame life limited measured kind mystical artificial fate 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 hope better understanding life ideals measured material things vital present future experience democracy successfully survived crisis home put away many evil things built new structures enduring lines 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 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 enlightened know democracy alone constructed unlimited civilization capable infinite progress improvement human life know look surface sense still spreading every continent humane advanced end unconquerable forms human society nation like person body body must fed clothed housed invigorated rested manner measures objectives time nation like person mind mind must kept informed alert must know 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 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 body mind constricted alien world lived 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 preservation sacred fire liberty destiny republican model government justly considered deeply finally staked experiment intrusted hands american people ." lose sacred fire let smothered doubt fear shall reject destiny washington strove valiantly triumphantly establish preservation spirit faith nation 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 ”

2. *Sample Sentence for Nixon’s speech after removal of stopwords*:

“ mr vice president mr speaker mr chief justice senator cook mrs eisenhower fellow citizens great good country share together met four years ago america bleak spirit depressed prospect seemingly endless war abroad destructive conflict home meet today stand threshold new era peace world central question us shall use peace let us resolve era enter postwar periods often time retreat isolation leads stagnation home invites new danger abroad let us resolve become time great responsibilities greatly borne renew spirit promise america enter third century nation past year saw far reaching results new policies peace continuing revitalize traditional friendships missions peking moscow able establish base new durable pattern

relationships among nations world america bold initiatives 1972 long remembered year greatest progress since end world war ii toward lasting peace world peace seek world flimsy peace merely interlude wars peace endure generations come important understand necessity limitations america role maintaining peace unless america work preserve peace peace unless america work preserve freedom freedom let us clearly understand new nature america role result new policies adopted past four years shall respect treaty commitments shall support vigorously principle country right impose rule another force shall continue era negotiation work limitation nuclear arms reduce danger confrontation great powers shall share defending peace freedom world shall expect others share time passed america make every nation conflict make every nation future responsibility presume tell people nations manage affairs respect right nation determine future also recognize responsibility nation secure future america role indispensable preserving world peace nation role indispensable preserving peace together rest world let us resolve move forward beginnings made let us continue bring walls hostility divided world long build place bridges understanding despite profound differences systems government people world friends let us build structure peace world weak safe strong respects right live different system would influence others strength ideas force arms let us accept high responsibility burden gladly gladly chance build peace noblest endeavor nation engage gladly also act greatly meeting responsibilities abroad remain great nation remain great nation act greatly meeting challenges home chance today ever history make life better america ensure better education better health better housing better transportation cleaner environment restore respect law make communities livable insure god given right every american full equal opportunity range needs great reach opportunities great let us bold determination meet needs new ways building structure peace abroad required turning away old policies failed building new era progress home requires turning away old policies failed abroad shift old policies new retreat responsibilities better way peace home shift old policies new retreat responsibilities better way progress abroad home key new responsibilities lies placing division responsibility lived long consequences attempting gather power responsibility washington abroad home time come turn away condescending policies paternalism washington knows best ." person expected act responsibly responsibility human nature let us encourage individuals

home nations abroad decide let us locate responsibility places let us measure others today offer promise purely governmental solution every problem lived long false promise trusting much government asked deliver leads inflated expectations reduced individual effort disappointment frustration erode confidence government people government must learn take less people people let us remember america built government people welfare work shirking responsibility seeking responsibility lives let us ask government challenges face together let us ask government help help national government great vital role play pledge government act act boldly lead boldly important role every one us must play individual member community day forward let us make solemn commitment heart bear responsibility part live ideals together see dawn new age progress america together celebrate 200th anniversary nation proud fulfillment promise world america longest difficult war comes end let us learn debate differences civility decency let us reach one precious quality government cannot provide new level respect rights feelings one another new level respect individual human dignity cherished birthright every american else time come us renew faith america recent years faith challenged children taught ashamed country ashamed parents ashamed america record home role world every turn beset find everything wrong america little right confident judgment history remarkable times privileged live america record century unparalleled world history responsibility generosity creativity progress let us proud system produced provided freedom abundance widely shared system history world let us proud four wars engaged century including one bringing end fought selfish advantage help others resist aggression let us proud bold new initiatives steadfastness peace honor made break toward creating world world known structure peace last merely time generations come embarking today era presents challenges great nation generation ever faced shall answer god history conscience way use years stand place hallowed history think others stood think dreams america think recognized needed help far beyond order make dreams come true today ask prayers years ahead may god help making decisions right america pray help together may worthy challenge let us pledge together make next four years best four years america history 200th birthday america young vital began bright beacon hope world let us go forward confident hope strong faith one another sustained faith god created us striving always serve purpose”

Sample Sentence for Kennedy's speech after removal of stopwords:

"vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change sworn almighty god solemn oath forebears l prescribed nearly century three quarters ago world different man holds mortal hands power abolish forms human poverty forms human life yet revolutionary beliefs forebears fought still issue around globe belief rights man come generosity state hand god dare forget today heirs first revolution let word go forth time place friend foe alike torch passed new generation americans born century tempered war disciplined hard bitter peace proud ancient heritage unwilling witness permit slow undoing human rights nation always committed committed today home around world let every nation know whether wishes us well ill shall pay price bear burden meet hardship support friend oppose foe order assure survival success liberty much pledge old allies whose cultural spiritual origins share pledge loyalty faithful friends united little cannot host cooperative ventures divided little dare meet powerful challenge odds split asunder new states welcome ranks free pledge word one form colonial control shall passed away merely replaced far iron tyranny shall always expect find supporting view shall always hope find strongly supporting freedom remember past foolishly sought power riding back tiger ended inside peoples huts villages across globe struggling break bonds mass misery pledge best efforts help help whatever period required communists may seek votes right free society cannot help many poor cannot save rich sister republics south border offer special pledge convert good words good deeds new alliance progress assist free men free governments casting chains poverty peaceful revolution hope cannot become prey hostile powers let neighbors know shall join oppose aggression subversion anywhere americas let every power know hemisphere intends remain master house world assembly sovereign states united nations last best hope age instruments war far outpaced instruments peace renew pledge support prevent becoming merely forum invective strengthen shield new weak enlarge area writ may run finally nations would make adversary offer pledge request sides begin anew quest peace dark powers destruction unleashed science engulf humanity planned

accidental self destruction dare tempt weakness arms sufficient beyond doubt certain beyond doubt never employed neither two great powerful groups nations take comfort present course sides overburdened cost modern weapons rightly alarmed steady spread deadly atom yet racing alter uncertain balance terror stays hand mankind final war let us begin anew remembering sides civility sign weakness sincerity always subject proof let us never negotiate fear let us never fear negotiate let sides explore problems unite us instead belaboring problems divide us let sides first time formulate serious precise proposals inspection control arms bring absolute power destroy nations absolute control nations let sides seek invoke wonders science instead terrors together let us explore stars conquer deserts eradicate disease tap ocean depths encourage arts commerce let sides unite heed corners earth command isaiah undo heavy burdens ... let oppressed go free ." beachhead cooperation may push back jungle suspicion let sides join creating new endeavor new balance power new world law strong weak secure peace preserved finished first 100 days finished first 1 000 days life administration even perhaps lifetime planet let us begin hands fellow citizens mine rest final success failure course since country founded generation americans summoned give testimony national loyalty graves young americans answered call service surround globe trumpet summons us call bear arms though arms need call battle though embattled call bear burden long twilight struggle year year rejoicing hope patient tribulation struggle common enemies man tyranny poverty disease war forge enemies grand global alliance north south east west assure fruitful life mankind join historic effort long history world generations granted role defending freedom hour maximum danger shrink responsibility welcome believe us would exchange places people generation energy faith devotion bring endeavor light country serve glow fire truly light world fellow americans ask country ask country fellow citizens world ask america together freedom man finally whether citizens america citizens world ask us high standards strength sacrifice ask good conscience sure reward history final judge deeds let us go forth lead land love asking blessing help knowing earth god work must truly ”

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)

Answer:

Top 3 words for Roosevelt's speech:
['nation', 'know', 'spirit']

From these {'nation': 12, 'know': 10, 'spirit': 9, 'life': 9, 'democracy': 9, 'us': 8, 'people': 7, 'america': 7, 'years': 6, 'freedom': 6, ...}

Top 3 words for Nixon's speech:
['us', 'let', 'america']

From these {'us': 26, 'let': 22, 'america': 21, 'peace': 19, 'world': 18, 'new': 15, 'nation': 11, 'responsibility': 11, 'government': 10, 'great': 9, ...}

Top 3 words for Kennedy's speech:
['let', 'us', 'world']

From these {'let': 16, 'us': 12, 'world': 8, 'sides': 8, 'new': 7, 'pledge': 7, 'citizens': 5, 'power': 5, 'shall': 5, 'free': 5, ...}

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

Answer:

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.
To create a word cloud we will again remove the stopwords but won't perform stemming. As removing stopwords would remove the filter the unwanted words that possibly have no sentiment analysis.

Word Cloud for Roosevelt's Speech

Figure 62

Word Cloud for Nixon's speech

Figure 63

Word cloud for Kennedy's speech

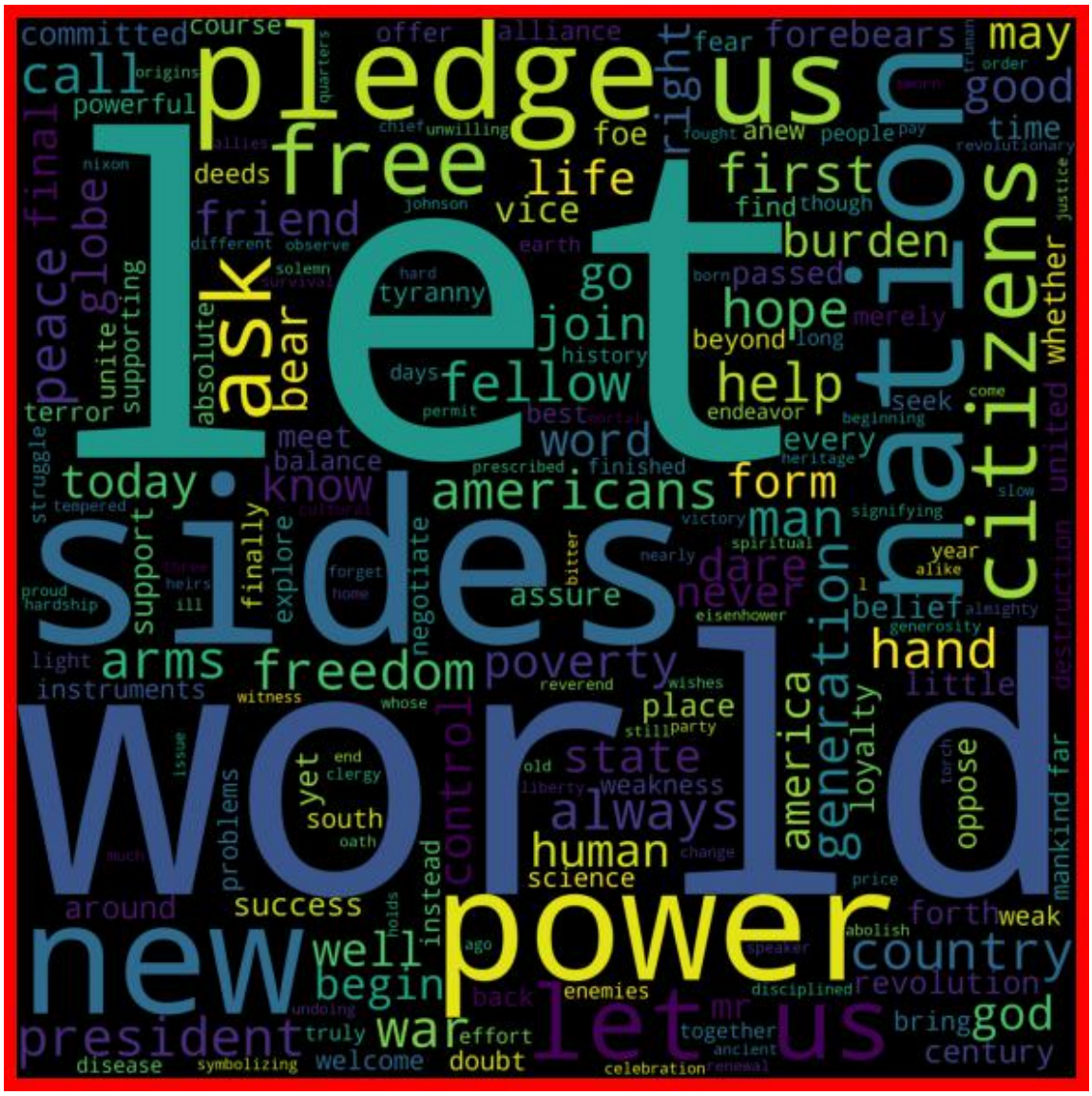

Figure 64

Insights:

1. Based on outputs we can see there are some similar words that are present in all the speeches.
2. These words may be the point which inspired many people and also got them the seat of president of USA.
3. Among the speeches, "nation" is a word highlighted in all three.

End of Problem 2
