Business Report

By – SAHIL SAXENA

MACHINE LEARNING

Machine Learning is the study of computer algorithms that can improve automatically through experience and by the use of data.

In this business problem we will look and machine learning algorithms at work

1. To predict the Exit Poll of an upcoming elections for a news channel

2. Text analysis of US president speeches

CONTENT

CONTENT OF TABLE

CONTENT OF FIGURE

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.

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.

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

SOLUTION:

TOP 5 ROWS OF DATAFRAME

	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.1. 1 – DATA FRAME

Unnamed:0 column is a redundant column, so there is no need to keep it for further analysis and it will also help in reducing storage memory usage.

ROWS – 1525
COLUMNS - 10

DESCRIPTIVE ANALYSIS AND NULL VALUE CHECK

DATA INFORMATION				
RangeIndex: 1525 entries, 0 to 1524				
Data columns (total 9 columns):				
#	Column	Count	Non-Null	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.1. 2 – DATA INFORMATION

There are 8 duplicated rows in the data frame which are just redundant rows, so we are these rows from the data frame for further analyses

DATA SUMMARY NUMERICAL VARIABLES	count	mean	std	min	25%	50%	75%	max
age	1517	54.24127	15.701741	24	41	53	67	93
economic.cond.national	1517	3.245221	0.881792	1	3	3	4	5
economic.cond.household	1517	3.137772	0.931069	1	3	3	4	5
Blair	1517	3.335531	1.174772	1	2	4	4	5
Hague	1517	2.749506	1.232479	1	2	2	4	5
Europe	1517	6.740277	3.299043	1	4	6	10	11
political.knowledge	1517	1.540541	1.084417	0	0	2	2	3

Table.1. 3 – DATA SUMMARY NUMERICAL VARIABLES

DATA SUMMARY CATEGORICAL VARIABLES	count	unique	top	freq
vote	1517	2	Labour	1057
gender	1517	2	female	808

Table.1. 4- DATA SUMMRY CATEGORICAL VARIABLE

CATEGORICAL VARIABLES COUNT AND DETAILS

VOTE: 2
Conservative 460
Labour 1057

GENDER: 2
male 709
female 808

INFERENCE

1. There are 1525 rows and 9 columns in the data frame.
2. There are no null values present in the data frame.
3. There are 07 int and 02 object type data present in the data.
4. All the data type are found correct as per their values in the data frame, So there are no anomalies present in the data.
5. There is 107.4+ KB memory usage in the data.

INFERENCE

1. The minimum age for voting is 24 years in Europe.
2. The Economic condition for national is good as the mean is on the higher side.
3. Same for economic condition of the household is also good as its mean is also on the higher side.
4. The Assessment of the Labour leader is good as its mean is 3.3 which is on higher side.
5. The assessment of the Conservative leader is not good its mean is on the lower side.
6. Labour has the most vote.
7. There are more female voters than male voters.

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

SOLUTION:

UNIVARIATE ANALYSIS

DISTRIBUTION OF VARIABLES

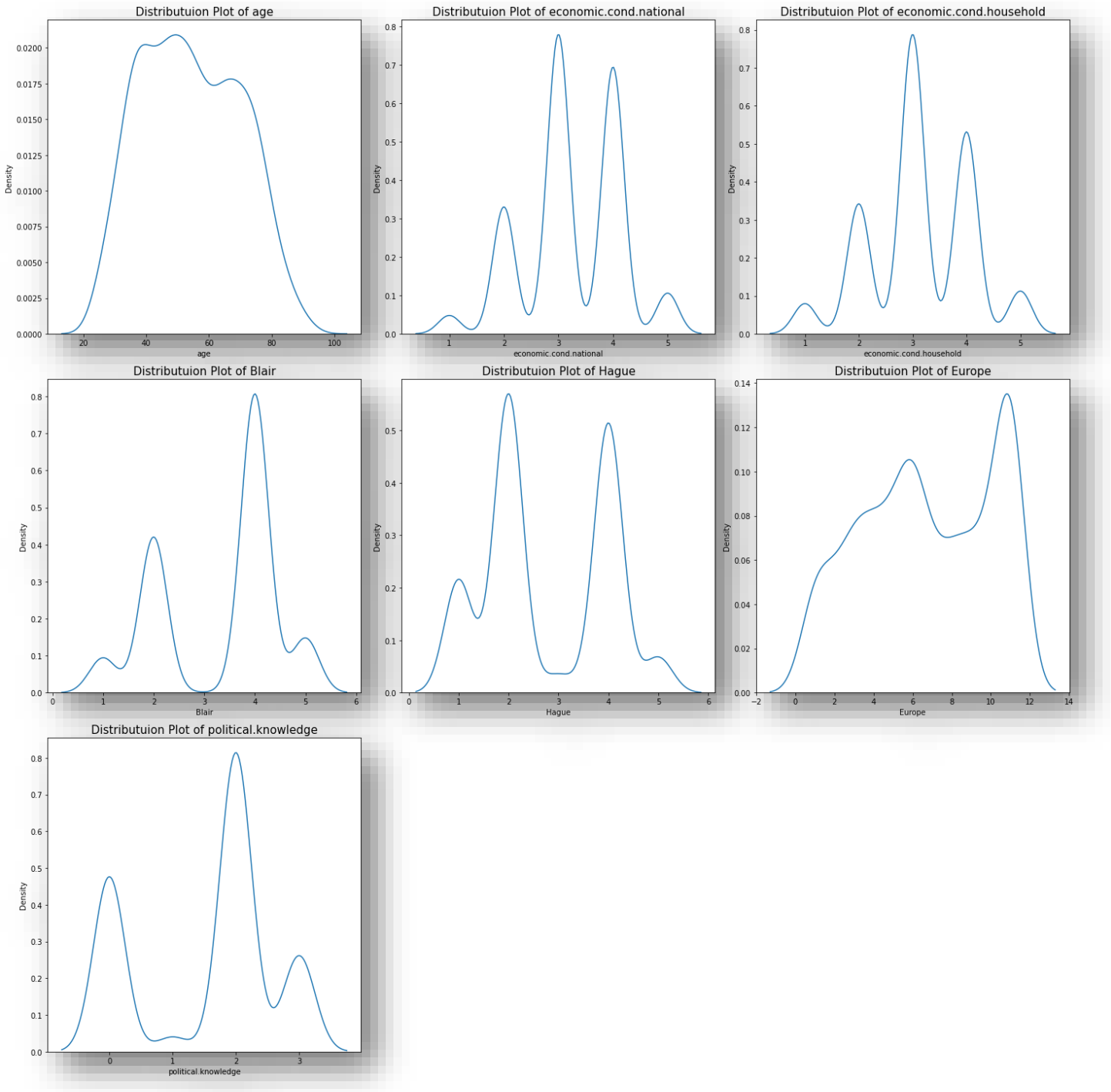


FIGURE.1. 1 – DISTRBUTION PLOT

As we can Check from the above Plot and stat

1. Age and Hague has very slight skewness towards right side and the value is almost near to zero which can help us assume normal distribution.
2. economic.cond.national, economic.cond.household, Europe all these variables are slightly left skewed.
3. Blair and Political.knowledge are the two predictors which have high negative stats value i.e., -0.539 and -0.423 resp, which shows moderately left skewed distribution.

```
Skewness of age
0.13979987012068112

Skewness of economic.cond.national
0.23847421478161793

Skewness of economic.cond.household
0.14414766882077137

Skewness of Blair
0.5395141989831328

Skewness of Hague
0.1461913444629453

Skewness of Europe
0.14189094981032258
```

OUTLIER CHECK FOR VARIABLES

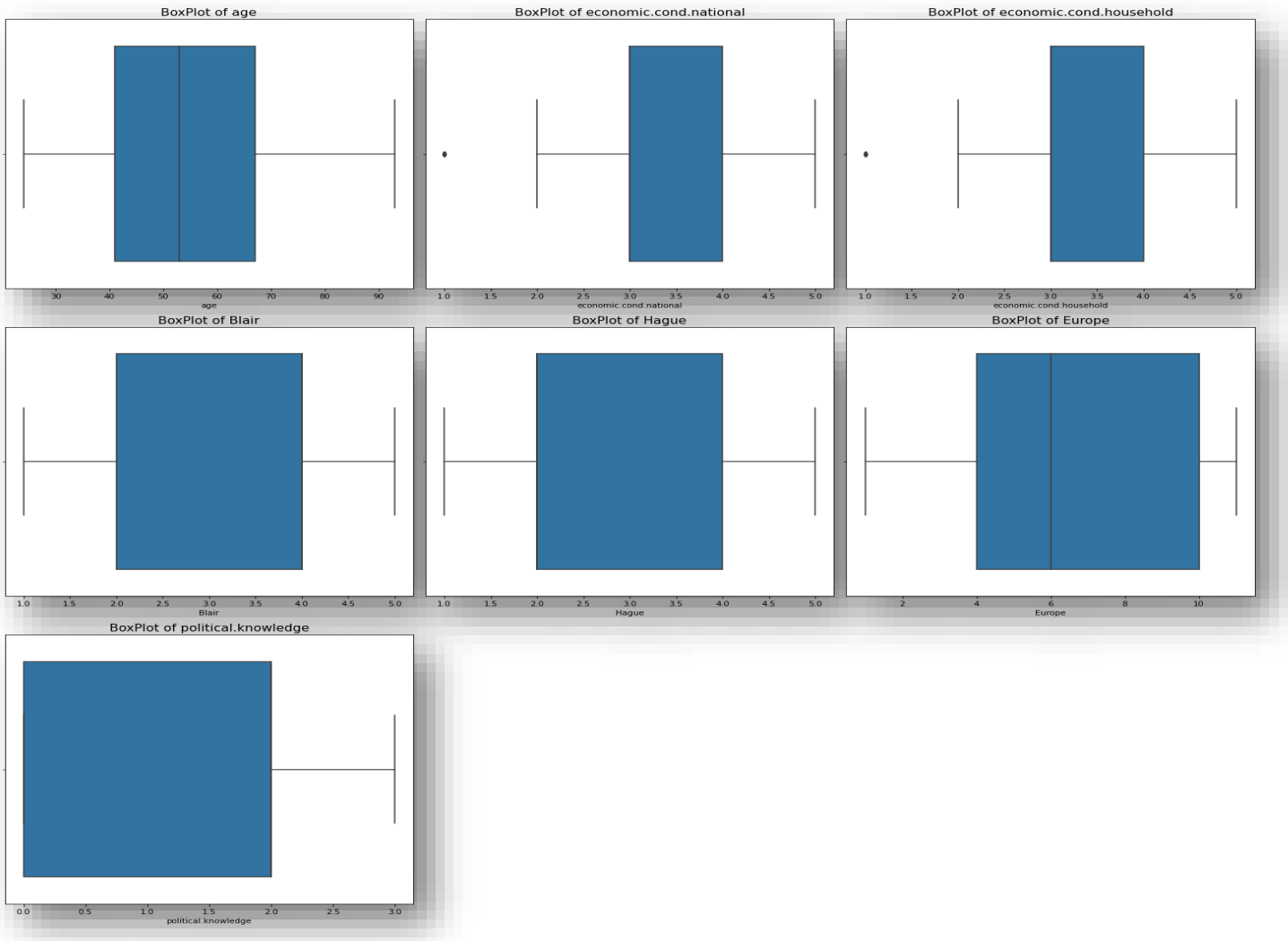


FIGURE.1. 2 – BOX PLOT

As We can Check from the Boxplot

1. economic.cond.national, economic.cond.household are the only two predictors which have only one outlier present in the data.

CATEGORICAL VARIABLE COUNT

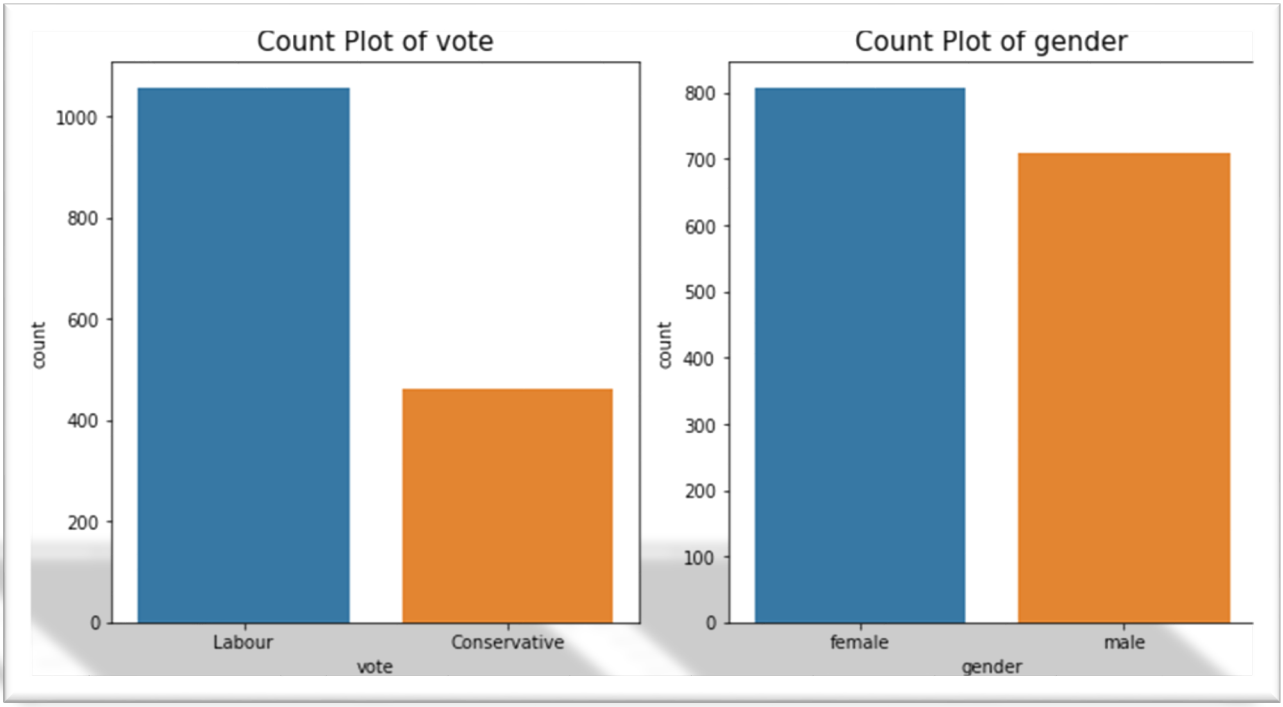
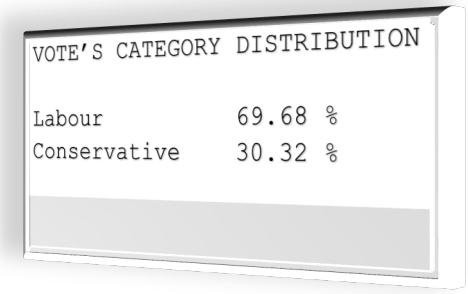


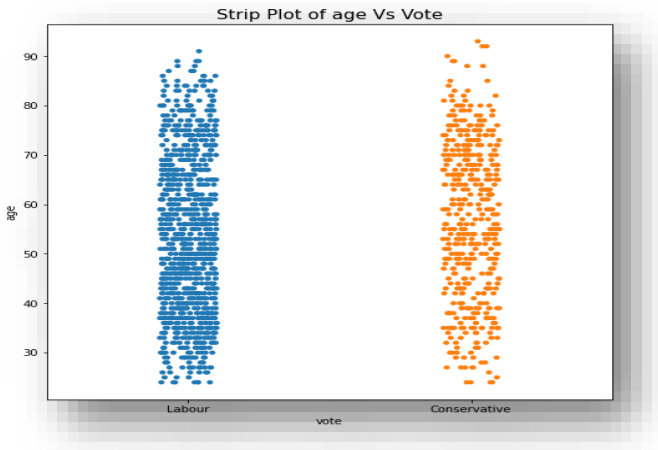
FIGURE.1. 3 – COUNT PLOT

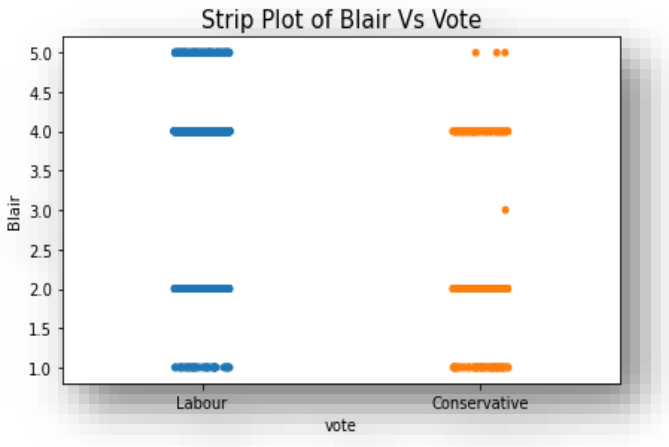
INFERENCE

1. As per the given data labour party has higher chance of winning the elections as they have 70 % vote in their favor.
2. There are more female voters than male voters.



BI-VARIATE ANALYSIS





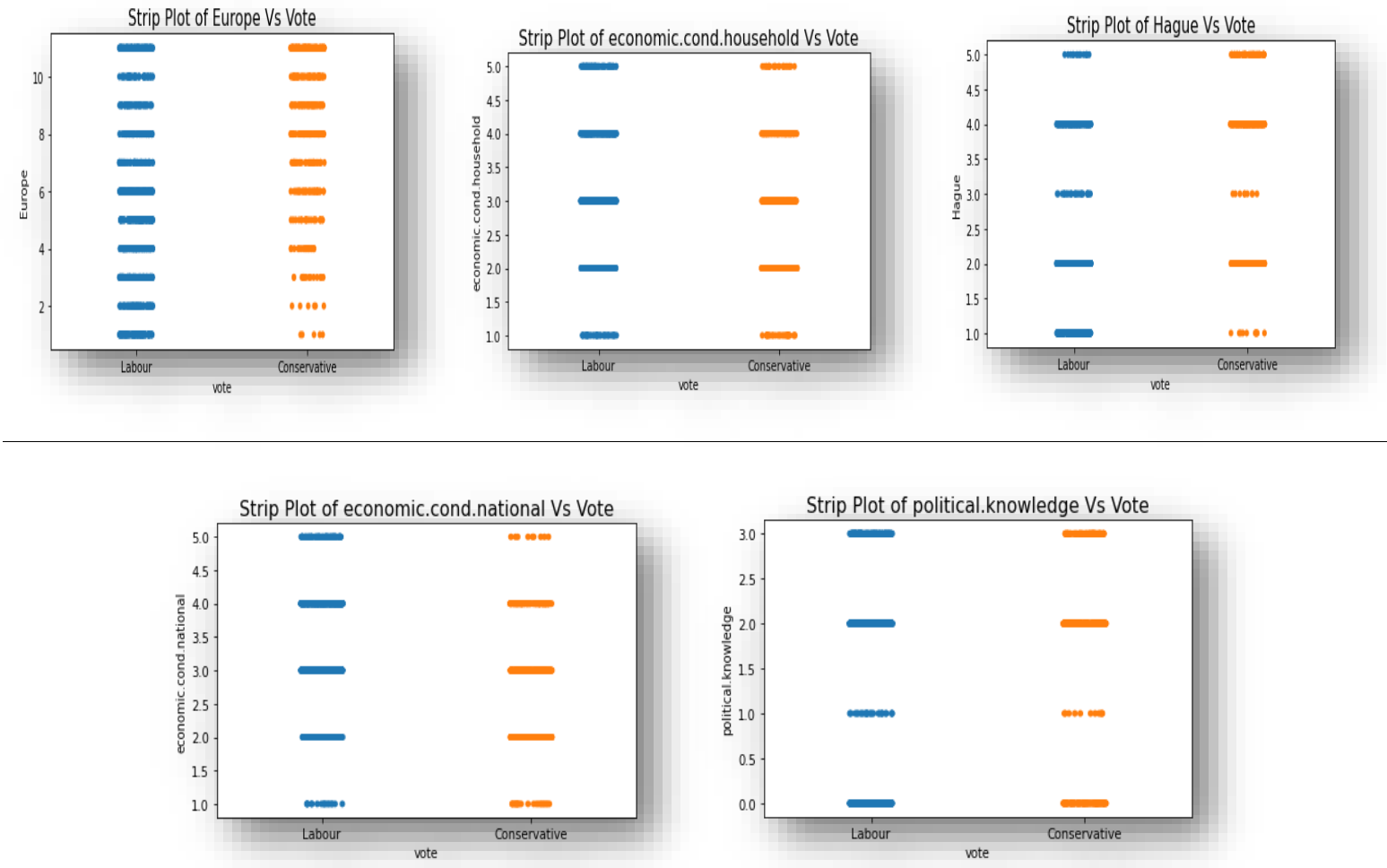


FIGURE.1. 4 – BI-VARIATE PLOT

As We can check and infer from the above plots

1. All the age groups Vote for both parties.
2. There are a greater number of People who vote for labour than conservative party thinks that the national economic condition is at the best.
3. There are a smaller number of People who vote for conservative than labour party thinks that the household economic condition is not at the best.
4. Voters of both parties have high Eurosceptic sentiment.

CHECKING FOR VARIABLES CORRELATION

FIGURE.1. 5 – PAIR PLOT

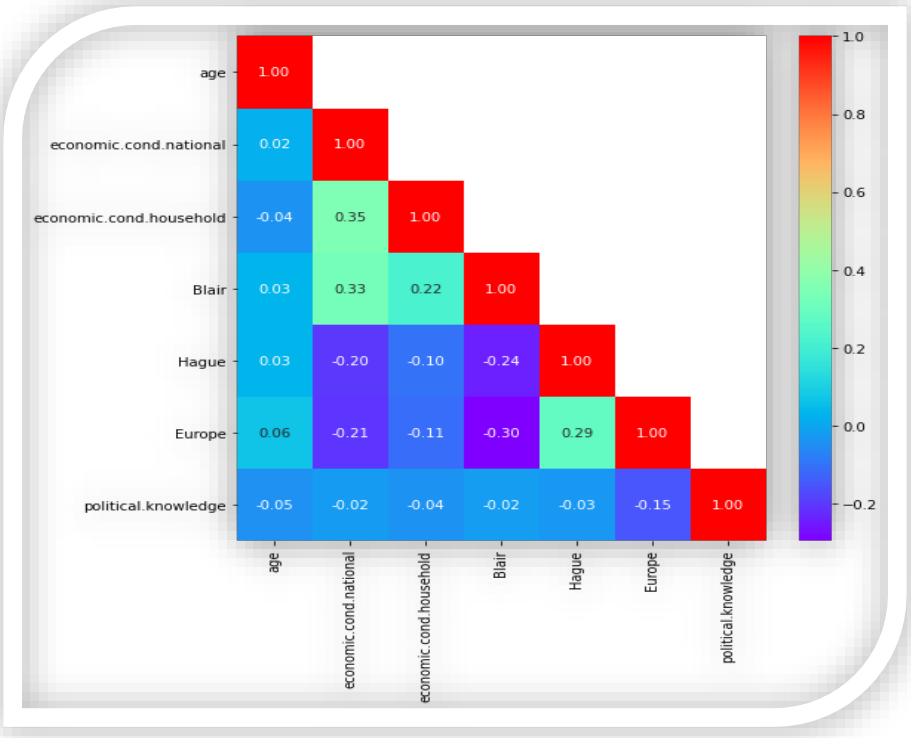


FIGURE.1. 6- CORRELATION PLOT

INFERENCE

As we can check from both the pair plot and correlation plots of variables that there is not much correlation present in the data,

Only few variables like economic.cond.national – economic.cond.household – blair

Europe – Hague

Are moderate positively correlated.

Q-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).

SOLUTION:

ENCODING THE DATA

AFTER ENCODING

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

Table.1. 5- CODING DATA

VOTE - ENCODING

0 - Conservative party - 460

1 - Labour party - 1057

GENDER - ENCODING

0 – Female - 858

1 – Male -709

IS SCALING NECESSARY OR NOT?

DATA SUMMARY NUMERICAL VARIABLES

	count	mean	std	min	25%	50%	75%	max
age	1517	54.24127	15.701741	24	41	53	67	93
economic.cond. national	1517	3.245221	0.881792	1	3	3	4	5
economic.cond. household	1517	3.137772	0.931069	1	3	3	4	5
Blair	1517	3.335531	1.174772	1	2	4	4	5
Hague	1517	2.749506	1.232479	1	2	2	4	5
Europe	1517	6.740277	3.299043	1	4	6	10	11
political.knowledge	1517	1.540541	1.084417	0	0	2	2	3

Table.1. 6 – SCALING DATA SUMMARY

ENCODING TECHINIQUES

LABEL ENCODING

Label encoding has been used on vote (target variable). As it a target variable we can encode it we can leave it as well.

ONE HOT ENCODING.

One hot encoding has been used on Gender (Predictor Variable) because to encode and convert it into numerical variable one hot encoding provides the best result.

COMMENT:

As per the data in the data frame, out of 9 columns Vote is our target variable and gender are a categorical variable then apart from age column all other columns are ordinal columns, so these are rating values so by scaling them will not be an affective idea and it will be more useful in model performance without scaling. The models which we are going to build are Logistic regression, Linear discriminant analysis, Naive Bayes, K nearest neighbors, in which Except KNN all other models are not much affected by scaling.

KNN is a distance-based model in all the independent variable should be on same scale so that distance of each data point can be measured correctly. So, for building KNN we need to scale the data.

DATA SPLITTING 70:30
TRAIN AND TEST SETS.

X_train – (1061, 8)

X_test – (456, 8)

y_ train – (1061,)

y_ test - *(456)*

TECHNIQUE USED

TRAIN AND TEST SPLIT

We used train and test slip technique from sklearn, it divides the data frame into four parts

X_train – Contains 70 % of predictor variable data for training the model.

X_test – Contains 30 % of predictor variable data for testing the model.

Y_train – Contains 70% of target variable for training the model.

Y_test - Contains 30% of target variable for testing the model.

Q-1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis). Interpret the inferences of both models.

SOLUTION:

LOGISTIC REGRESSION MODEL

This model is trained without any hyper parameters just with default parameters.

- The Coefficient of age is -0.014958728279182067
- The Coefficient of economic.cond.national is 0.6284920003187814
- The Coefficient of economic.cond.household is 0.06305081800974509
- The Coefficient of Blair is 0.6008794573583591
- The Coefficient of Hague is -0.8233082296459409
- The Coefficient of Europe is -0.21162032502062234
- The Coefficient of political.knowledge is -0.32183229467434155
- The Coefficient of gender_male is 0.19191024895010797

IMPORTANT FEATURES

FOR PREDICTING 1

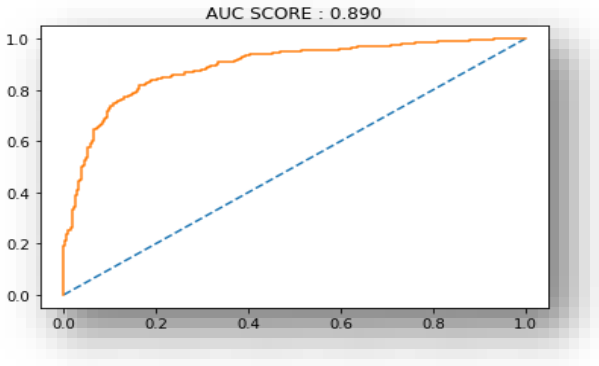
- Economic.cond.national
- Blair
- Gender
- Economic.cond.household

FOR PREDICTING 0

- Hague
- Political.knowledge
- Europe
- Age

TRAIN DATA

CLASSIFICATION REPORT				
	precision	recall f1	f1 score	support
0	0.74	0.64	0.69	307
1	0.86	0.91	0.88	754
accuracy			0.83	1061
macro avg	0.8	0.77	0.79	1061
weighted avg	0.83	0.83	0.83	1061



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (74%) – 74% of prediction that people who will vote for conservative is correct.

2. RECALL (64%) – 64% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

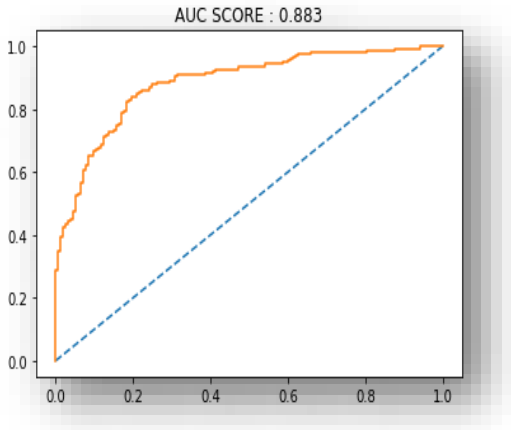
1. PRECISION (86%) – 86 % of prediction that people will vote for Labour.

2. RECALL (91%) – 91 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 83 % predictions are correct.

TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.76	0.74	0.75	153
1	0.87	0.88	0.88	303
accuracy			0.84	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.84	0.83	456



INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (76%) – 76% of prediction that people who will vote for conservative is correct.

2. RECALL (74%) – 74% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (87%) – 87 % of prediction that people will vote for Labour.

2. RECALL (88%) – 88 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 84 % predictions are correct.

MODEL VALIDNESS

As we can check Recall percentage for 0 train and test has a difference of 10 % but it is in permissible limits and all other values are within their limits, so we can say it is a robust model and it neither an overfit nor an under fit model.

LINEAR DISCRIMINANT ANALYSIS

This model is trained without any hyper parameters just with default parameters.

- The Coefficient of age is -0.020037048856610347
- The Coefficient of economic.cond.national is 0.6049204499917704
- The Coefficient of economic.cond.household is 0.05006904695697722
- The Coefficient of Blair is 0.7424003897819801
- The Coefficient of Hague is -0.9266343785776759
- The Coefficient of Europe is -0.22361192469849597
- The Coefficient of political.knowledge is -0.4303348424332059
- The Coefficient of gender_male is 0.14907997566596054

IMPORTANT FEATURES

FOR PREDICTING 1

- Blair
- Economic.cond.national
- Gender
- Economic.cond.household

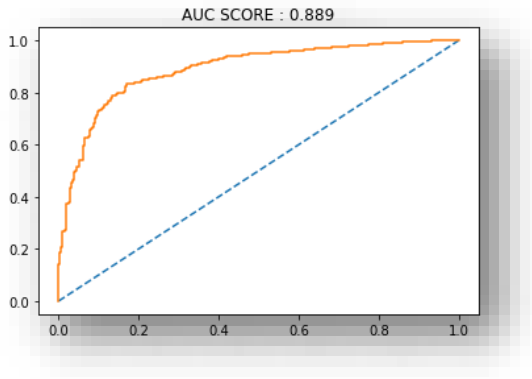
FOR PREDICTING 0

- Hague
- Political.knowledge
- Europe
- Age

TRAIN DATA

CLASSIFICATION REPORT

	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	0.8	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (74%) – 74% of prediction that people who will vote for conservative is correct.

2. RECALL (65%) – 65% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

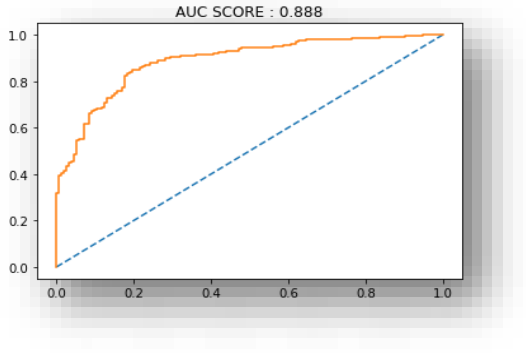
1. PRECISION (86%) – 86 % of prediction that people will vote for Labour.

2. RECALL (91%) – 91 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 83 % predictions are correct.

TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.77	0.73	0.74	153
1	0.86	0.89	0.88	303
accuracy			0.83	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456



INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (77%) – 77% of prediction that people who will vote for conservative is correct.

2. RECALL (73%) – 73% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (86%) – 86 % of prediction that people will vote for Labour.

2. RECALL (89%) – 89 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 83 % predictions are correct.

MODEL VALIDNESS

As we can check from train and test reports that all the all the parameters have almost similar values, so we can say it is a robust model and it neither an overfit nor an under fit model.

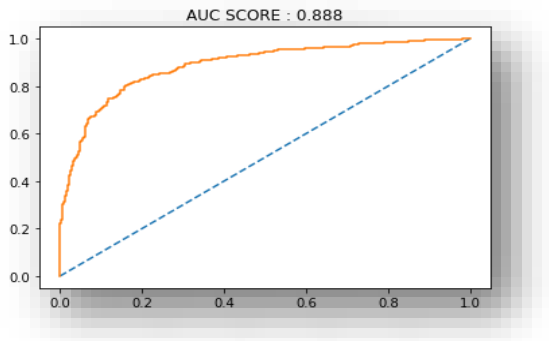
Q-1.5) Apply KNN Model and Naïve Bayes Model. Interpret the inferences of each model.

NAÏVE BAYES MODEL.

This model is trained without any hyper parameters just with default parameters.

TRAIN DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.73	0.69	0.71	307
1	0.88	0.9	0.89	754
accuracy			0.84	1061
macro avg	0.8	0.79	0.8	1061
weighted avg	0.83	0.84	0.83	1061



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (73%) – 73% of prediction that people who will vote for conservative is correct.

2. RECALL (69%) – 69% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

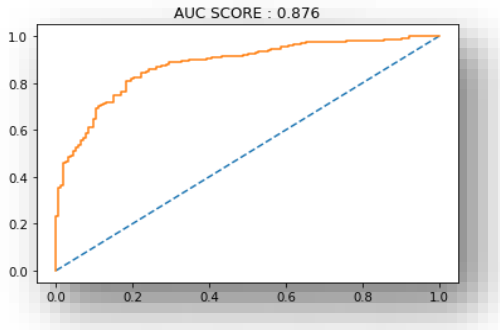
1. PRECISION (88%) – 88 % of prediction that people will vote for Labour.

2. RECALL (90%) – 90 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 84 % predictions are correct.

TEST DATA.

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	0.8	0.8	0.8	456
weighted avg	0.82	0.82	0.82	456



INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (74%) – 74% of prediction that people who will vote for conservative is correct.

2. RECALL (73%) – 73% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (87%) – 87 % of prediction that people will vote for Labour.

2. RECALL (87%) – 87 % of people vote for labour are predicted correctly.

MODEL VALIDNESS

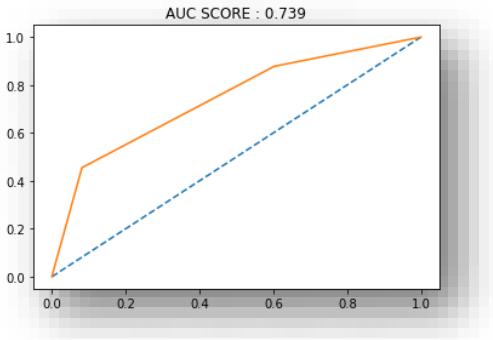
As we can check from train and test reports that all the all the parameters have almost similar values, so we can say it is a robust model and it neither an overfit nor an under fit model.

K-NEAREST NEIGHBOR MODEL.

This model is trained without any hyper parameters just with default parameters.

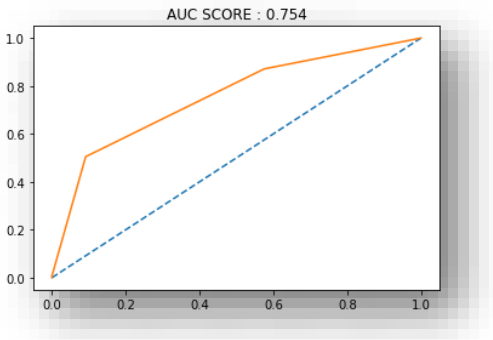
TRAIN DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.77	0.71	0.74	307
1	0.88	0.91	0.9	754
accuracy			0.85	1061
macro avg	0.83	0.81	0.82	1061
weighted avg	0.85	0.85	0.85	1061



TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.76	0.71	0.73	153
1	0.86	0.88	0.87	303
accuracy			0.82	456
macro avg	0.81	0.8	0.8	456
weighted avg	0.82	0.82	0.82	456



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (73%) – 73% of prediction that people who will vote for conservative is correct.

2. RECALL (69%) – 69% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (88%) – 88 % of prediction that people will vote for Labour.

2. RECALL (90%) – 90 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 85 % predictions are correct.

INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (76%) – 76 % of prediction that people who will vote for conservative is correct.

2. RECALL (71%) – 71% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (86%) – 86 % of prediction that people will vote for Labour.

2. RECALL (88%) – 88 % of people vote for labour are predicted correctly.

MODEL VALIDNESS

As we can check from train and test reports that all the all the parameters have almost similar values, so we can say it is a robust model and it neither an overfit nor an under fit model.

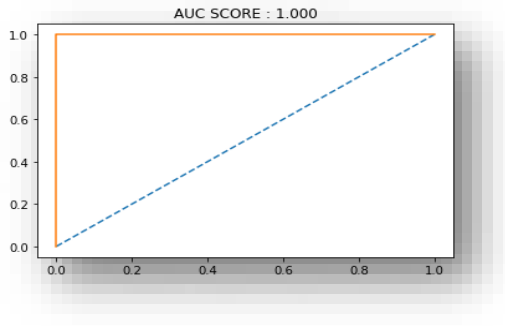
Q-1.6) Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.

SOLUTION:

BAGGING

This model is trained without any hyper parameters just with default parameters.

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	1	1	1	307
1	1	1	1	754
accuracy			1	1061
macro avg	1	1	1	1061
weighted avg	1	1	1	1061



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (100%) – 100% of prediction that people who will vote for conservative is correct.

2. RECALL (100%) – 100% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

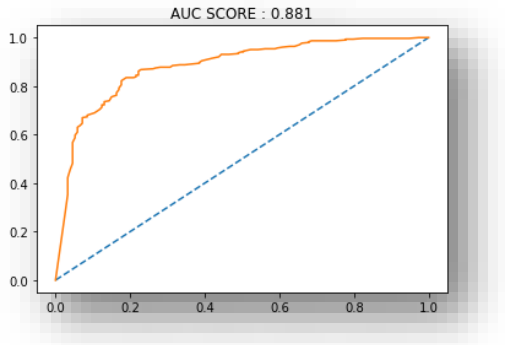
1. PRECISION (100%) – 100 % of prediction that people will vote for Labour.

2. RECALL (100%) – 100 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 100 % predictions are correct.

TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.74	0.71	0.72	153
1	0.86	0.88	0.87	303
accuracy			0.82	456
macro avg	0.8	0.79	0.8	456
weighted avg	0.82	0.82	0.82	456



INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (74%) – 74% of prediction that people who will vote for conservative is correct.

2. RECALL (71%) – 71% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (86%) – 86 % of prediction that people will vote for Labour.

2. RECALL (88%) – 88 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 82 % predictions are correct.

MODEL VALIDNESS

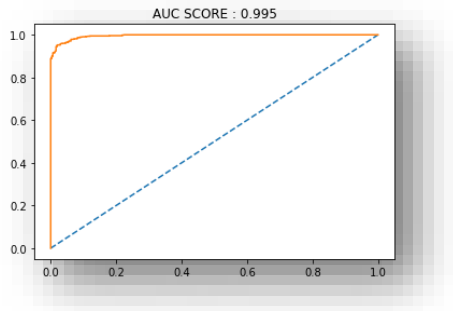
As we can check from train and test reports that all the all the parameters have a significance difference (over the permissible limits), so we can say it is an overfit model.

BAGGING WITH RANDOM FOREST

This model is trained without any hyper parameters just with default parameters.

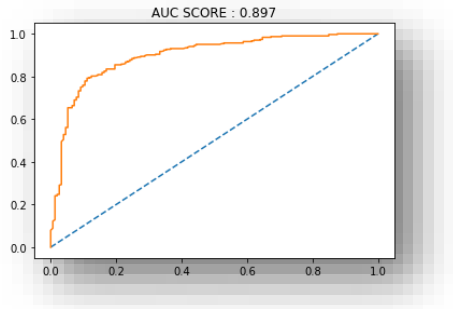
TRAIN DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.96	0.92	0.94	307
1	0.97	0.98	0.98	754
accuracy			0.97	1061
macro avg	0.96	0.95	0.96	1061
weighted avg	0.97	0.97	0.96	1061



TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.78	0.67	0.72	153
1	0.84	0.9	0.87	303
accuracy			0.82	456
macro avg	0.81	0.79	0.8	456
weighted avg	0.82	0.82	0.82	456



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (96%) – 96% of prediction that people who will vote for conservative is correct.

2. RECALL (92%) – 92% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (97%) – 97 % of prediction that people will vote for Labour.

2. RECALL (98%) – 98 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 97 % predictions are correct.

INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (78%) – 78% of prediction that people who will vote for conservative is correct.

2. RECALL (67%) – 67% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (84%) – 84 % of prediction that people will vote for Labour.

2. RECALL (90%) – 90% of people vote for labour are predicted correctly.

OVERALL ACCURACY – 82 % predictions are correct.

MODEL VALIDNESS

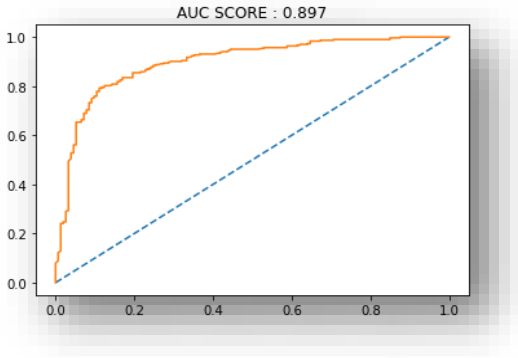
As we can check from train and test reports that all the all the parameters have a significance difference (over the permissible limits), so we can say it is an overfit model.

BOOSTING (ADA BOOSTING)

This model is trained without any hyper parameters just with default parameters.

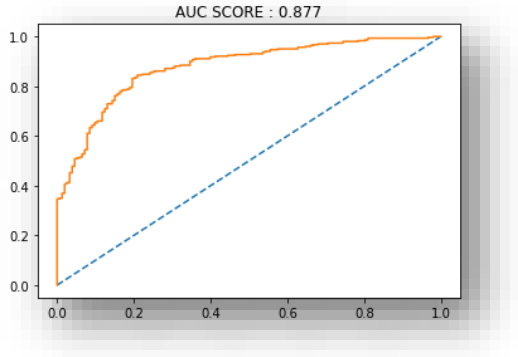
TRAIN DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.76	0.7	0.73	307
1	0.88	0.91	0.9	754
accuracy			0.85	1061
macro avg	0.82	0.8	0.81	1061
weighted avg	0.85	0.85	0.85	1061



TEST DATA

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



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (96%) – 96% of prediction that people who will vote for conservative is correct.

2. RECALL (92%) – 92% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (97%) – 97 % of prediction that people will vote for Labour.

2. RECALL (98%) – 98 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 97 % predictions are correct.

INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (78%) – 78% of prediction that people who will vote for conservative is correct.

2. RECALL (67%) – 67% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (84%) – 84 % of prediction that people will vote for Labour.

2. RECALL (90%) – 90% of people vote for labour are predicted correctly.

OVERALL ACCURACY – 82 % predictions are correct.

MODEL VALIDNESS

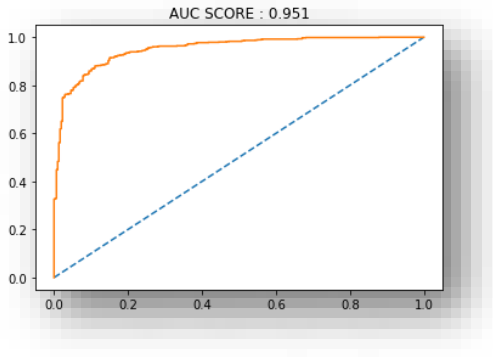
As we can check from train and test reports that all the all the parameters have almost similar values, so we can say it is a robust model and it neither an overfit nor an under fit model.

BOOSTING (GRADIENT BOOSTING)

This model is trained without any hyper parameters just with default parameters.

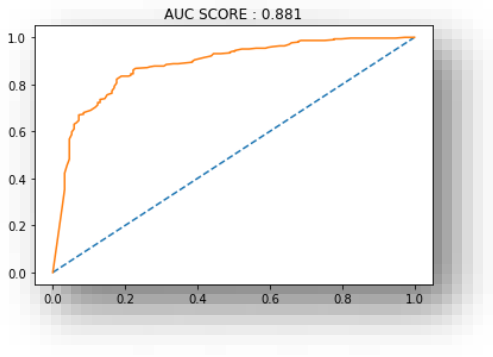
TRAIN DATA

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



TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.8	0.69	0.74	153
1	0.85	0.91	0.88	303
accuracy			0.84	456
macro avg	0.83	0.8	0.81	456
weighted avg	0.84	0.84	0.83	456



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (84%) – 84% of prediction that people who will vote for conservative is correct.

2. RECALL (78%) – 78% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (91%) – 91 % of prediction that people will vote for Labour.

2. RECALL (94%) – 94 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 89 % predictions are correct.

INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (80%) – 80% of prediction that people who will vote for conservative is correct.

2. RECALL (69%) – 69% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (85%) – 85% of prediction that people will vote for Labour.

2. RECALL (91%) – 91% of people vote for labour are predicted correctly.

OVERALL ACCURACY – 84 % predictions are correct.

MODEL VALIDNESS

As we can check Recall percentage for 0 train and test has a difference of 9 % but it is in permissible limits and all other values are within their limits, so we can say it is a robust model and it neither an overfit nor an under fit model.

MODEL TUNING

LOGISTIC REGRESSION WITH GRID SEARCH

This model is trained with Grid-Search best parameters.

HYPER PARAMETERS

Penalty: L2

Solver: Newton-Cg

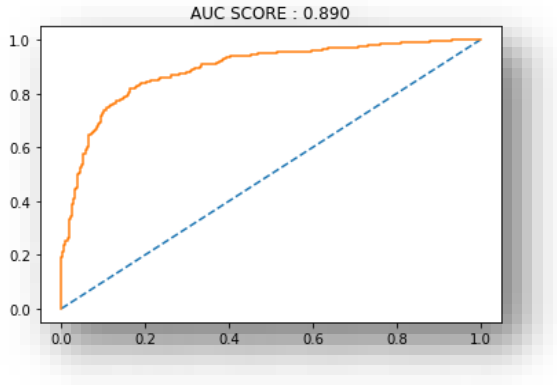
Tol: 0.0001

*** NOTE***

For hyper parameters different combinations of parameters have been tried and after each evaluation, it is found that Logistic Regression Model is performing best with these parameters.

TRAIN DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	**0.74**	**0.64**	**0.69**	**307**
1	**0.86**	**0.91**	**0.88**	**754**
accuracy			**0.83**	**1061**
macro avg	**0.8**	**0.77**	**0.79**	**1061**
weighted avg	**0.83**	**0.83**	**0.83**	**1061**



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (74%) – 74% of prediction that people who will vote for conservative is correct.

2. RECALL (64%) – 64 % of people vote for conservative are predicted correctly.

People Who will vote for Labour.

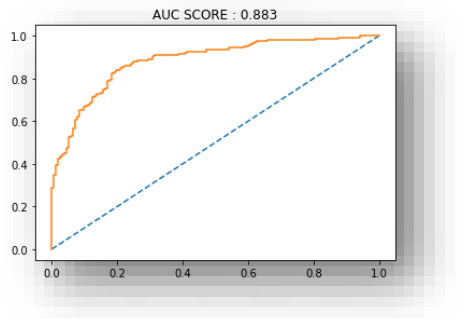
1. PRECISION (86%) – 86 % of prediction that people will vote for Labour.

2. RECALL (91%) – 91 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 83 % predictions are correct.

TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	**0.76**	**0.74**	**0.75**	**153**
1	**0.87**	**0.88**	**0.88**	**303**
accuracy			**0.84**	**456**
macro avg	**0.82**	**0.81**	**0.81**	**456**
weighted avg	**0.83**	**0.84**	**0.83**	**456**



INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (76%) – 76% of prediction that people who will vote for conservative is correct.

2. RECALL (74%) – 74% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (87%) – 87% of prediction that people will vote for Labour.

2. RECALL (88%) – 88% of people vote for labour are predicted correctly.

OVERALL ACCURACY – 84 % predictions are correct.

MODEL VALIDNESS

As we can check Recall percentage for 0 train and test has a difference of 10 % but it is in permissible limits and all other values are within their limits, so we can say it is a robust model and it neither an overfit nor an under fit model.

LINEAR DISCRIMINANT ANALYSIS WITH GRID SEARCH

This model is trained with Grid-Search best parameters.

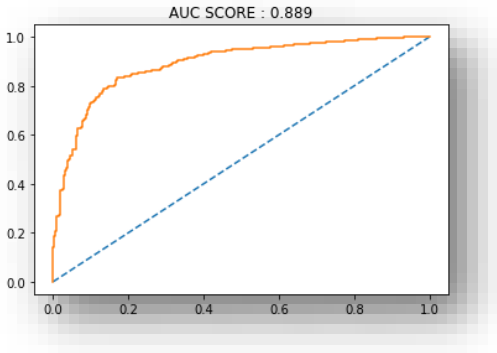
Solver - SVD

*** NOTE***

For hyper parameters different combinations of hyper parameters have been tried and after each evaluation, it is found that Linear discriminant analysis Model is performing best with these parameters.

TRAIN DATA

CLASSIFICATION REPORT				
	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	0.8	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (74%) – 74% of prediction that people who will vote for conservative is correct.

2. RECALL (65%) – 65% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

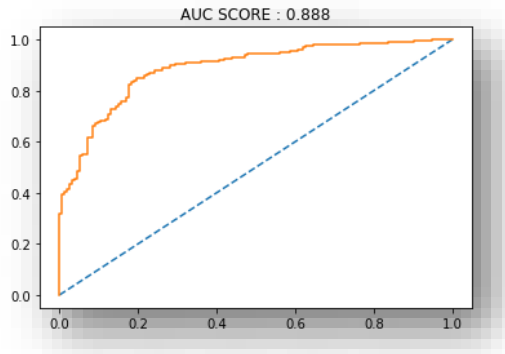
1. PRECISION (86%) – 86 % of prediction that people will vote for Labour.

2. RECALL (91%) – 91 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 83 % predictions are correct.

TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.77	0.73	0.74	153
1	0.86	0.89	0.88	303
accuracy			0.83	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456



INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (77%) – 77% of prediction that people who will vote for conservative is correct.

2. RECALL (73%) – 73% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (86%) – 86% of prediction that people will vote for Labour.

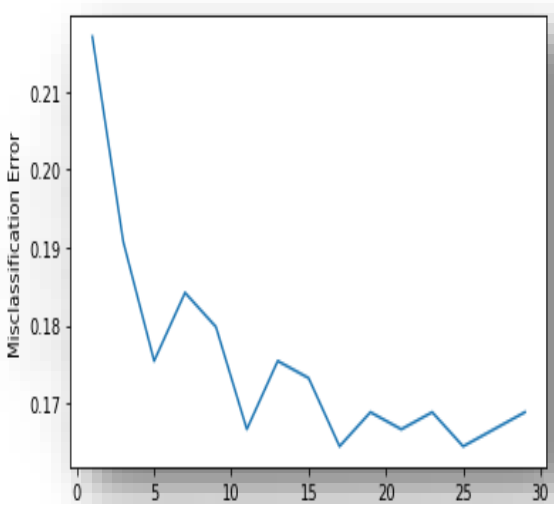
2. RECALL (89%) – 89% of people vote for labour are predicted correctly.

OVERALL ACCURACY – 83 % predictions are correct.

MODEL VALIDNESS

As we can check Recall percentage for 0 train and test has a difference of 8 % but it is in permissible limits and all other values are within their limits, so we can say it is a robust model and it neither an overfit nor an under fit

K-NEAREST NEIGHBORS WITH GRID SEARCH



This model is trained with Grid-Search best parameters.

HYPER PARAMETER

K VALUE – 25

MISCLASSIFICATION ERROR															
K_Value	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29
MCE	0.22	0.19	0.18	0.18	0.18	0.17	0.18	0.17	0.16	0.17	0.17	0.17	0.16	0.17	0.17

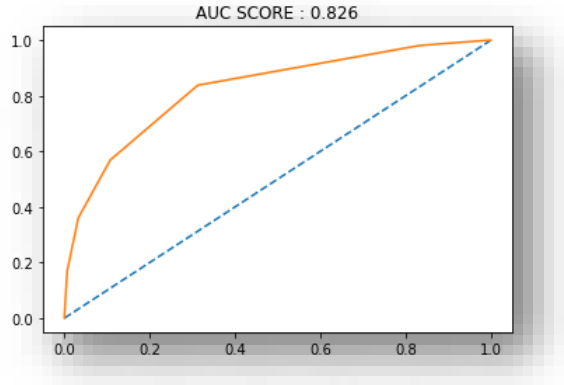
MCE – Mis Classification error

MCE = 1- Accuracy

As we can check at k – 25 misclassification error value is the least i.e. .16, so we decided to tune the model fot this value of k.

TRAIN DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	**0.76**	**0.64**	**0.69**	**307**
1	**0.86**	**0.92**	**0.89**	**754**
accuracy			**0.84**	**1061**
macro avg	**0.81**	**0.78**	**0.79**	**1061**
weighted avg	**0.83**	**0.84**	**0.83**	**1061**



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (76%) – 76% of prediction that people who will vote for conservative is correct.

2. RECALL (64%) – 64% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

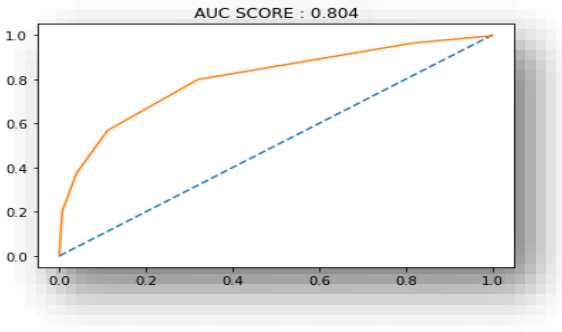
1. PRECISION (86%) – 86 % of prediction that people will vote for Labour.

2. RECALL (92%) – 92 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 84 % predictions are correct.

TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	**0.8**	**0.67**	**0.73**	**153**
1	**0.85**	**0.92**	**0.88**	**303**
accuracy			**0.84**	**456**
macro avg	**0.83**	**0.8**	**0.81**	**456**
weighted avg	**0.83**	**0.84**	**0.83**	**456**



INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (80%) – 80% of prediction that people who will vote for conservative is correct.

2. RECALL (67%) – 67% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (85%) – 85% of prediction that people will vote for Labour.

2. RECALL (92%) – 92% of people vote for labour are predicted correctly.

OVERALL ACCURACY – 84 % predictions are correct.

MODEL VALIDNESS

As we can check from train and test reports that all the all the parameters have almost similar values, so we can say it is a robust model and it neither an overfit nor an under fit model.

BAGGING CLASSIFIER WITH RANDOM FOREST AND GRID SEARCH

This model is trained with Grid-Search best parameters.

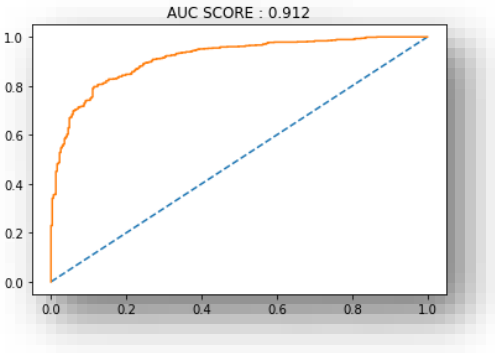
1. Max depth: 5
2. Max features: 4
3. Min samples leaf: 10
4. Min samples split: 50
5. N estimators: 300

*** NOTE***

For hyper parameters different combinations of hyper parameters have been tried and after each evaluation, it is found that Bagging Classifier And Random Forest with grid search Model is performing best with these parameters.

TRAIN DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.79	0.68	0.73	307
1	0.88	0.93	0.9	754
accuracy			0.85	1061
macro avg	0.83	0.8	0.82	1061
weighted avg	0.85	0.85	0.85	1061



INFERENCE FOR TRAIN DATA

People who will vote for Conservative party.

1. PRECISION (79%) – 79% of prediction that people who will vote for conservative is correct.

2. RECALL (68%) – 68% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

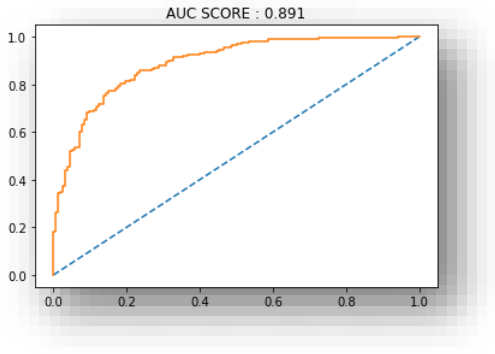
1. PRECISION (88%) – 88 % of prediction that people will vote for Labour.

2. RECALL (93%) – 93 % of people vote for labour are predicted correctly.

OVERALL ACCURACY – 85 % predictions are correct.

TEST DATA

CLASSIFICATION REPORT				
	precision	recall	f1 score	support
0	0.79	0.67	0.73	153
1	0.85	0.91	0.88	303
accuracy			0.83	456
macro avg	0.82	0.79	0.8	456
weighted avg	0.83	0.83	0.83	456



INFERENCE FOR TEST DATA

People who will vote for Conservative party.

1. PRECISION (79%) – 79% of prediction that people who will vote for conservative is correct.

2. RECALL (67%) – 67% of people vote for conservative are predicted correctly.

People Who will vote for Labour.

1. PRECISION (85%) – 85% of prediction that people will vote for Labour.

2. RECALL (91%) – 91% of people vote for labour are predicted correctly.

OVERALL ACCURACY – 83% predictions are correct.

MODEL VALIDNESS

As we can check from train and test reports that all the all the parameters have almost similar values, so we can say it is a robust model and it neither an overfit nor an under fit model.

COMMENT:

All the model performance are good and robust, some models like Bagging classifier, Bagging classifier with grid search and bagging classifier with random forest are over fitted model which cannot be included for model comparison. KNN with grid search has shown significant improvement than KNN with default parameters.

As the data is a balance data with 69 % and 31 % division in the target variable, so accuracy will be preferred above other parameters. Still other parameters will be taken into account but accuracy will be given more weightage.

Hyper-Parameters have been chosen on the basis of data to and chose the best parameters using the grid search technique.

Q-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.

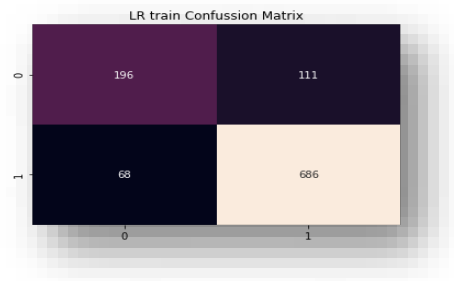
SOLUTION:

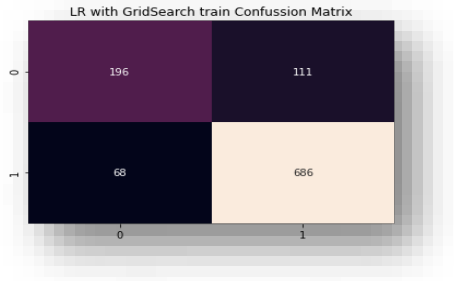
MODEL COMPARISON

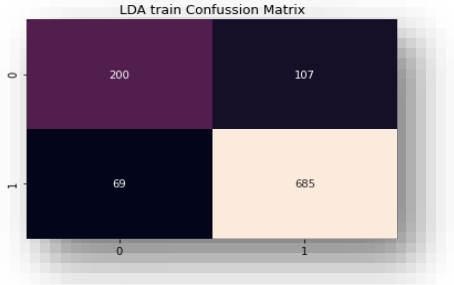
MODELS

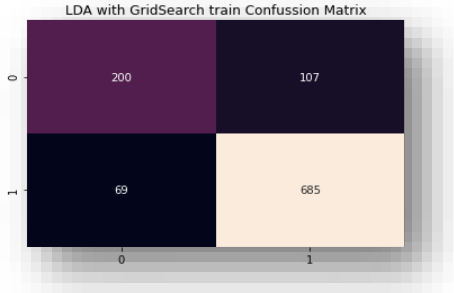
1. LR – LOGISTIC REGRESSION
2. LDA – LINEAR DISCRIMINANT ANALYSIS
3. GNB – GAUSSIAN NAÏVE BAYES
4. KNN – K NEAREST NEIGHBOUR
5. BC – BAGGING CLASSIFIER
6. BC AND RF – BAGGING CLASSIFIER WITH BASE ESTIMATOR RANDOM FOREST.
7. ABC – ADA BOOST CLASSIFIER
8. GB – GRADIENT BOOSTING
9. LR WITH GS – LOGISTIC REGRESSION WITH GRID SEARCH
10. LDA WITH GS – LINEAR DISCRIMINANT ANALYSIS WITH GRID SEARCH
11. KNN WITH GS – K – NEAREST NEIGHBOUR WITH GRID SEARCH.
12. BC & RF WITH GS – BAGGING CLASSIFIER AND RANDOM FOREST WITH GRID SEARCH

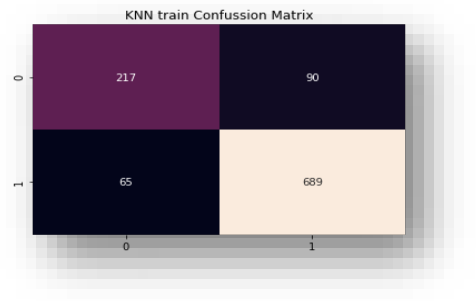
MODEL'S CONFUSION MATRIX FOR TRAIN DATA

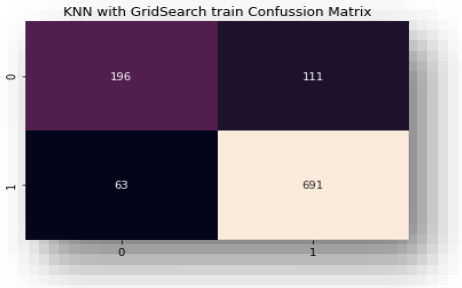


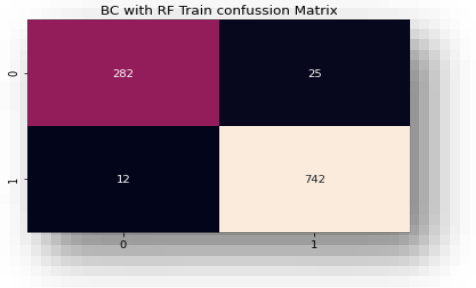


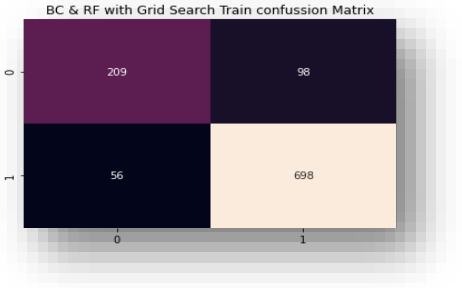


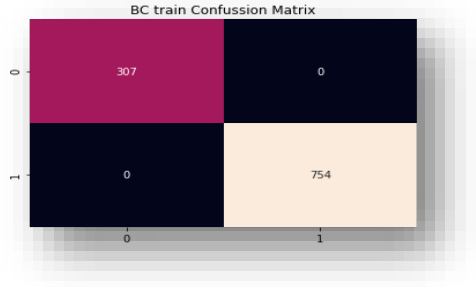


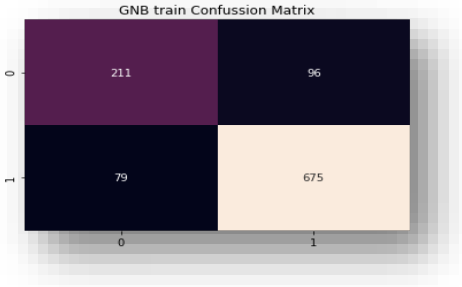


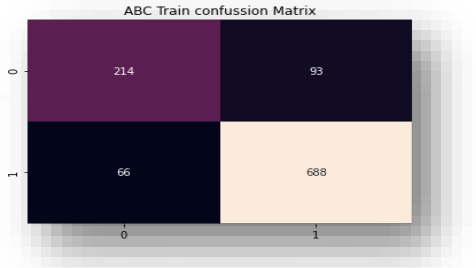


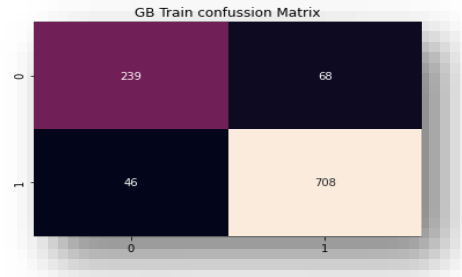




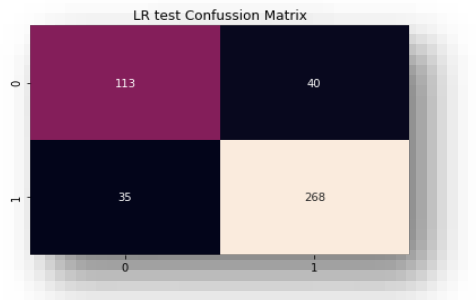


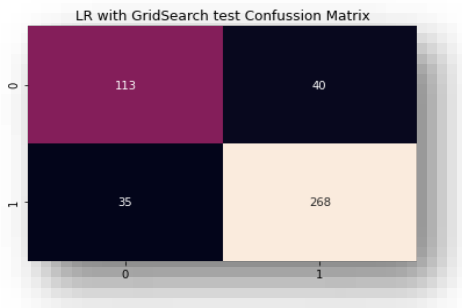


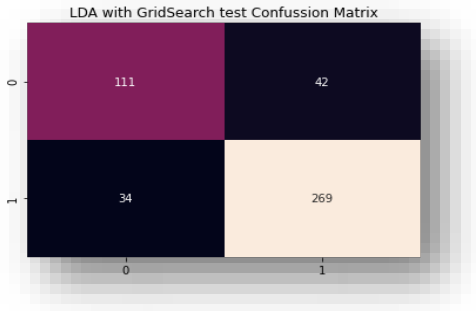


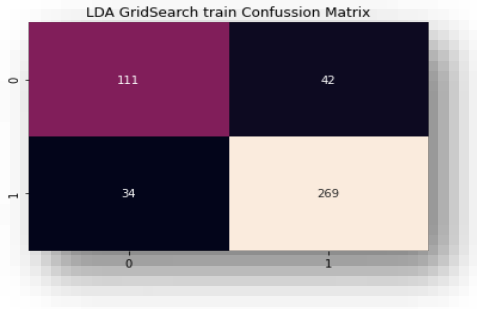


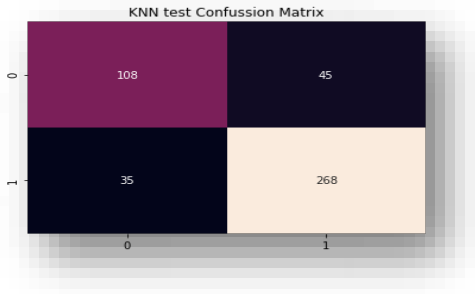
MODEL'S CONFUSION MATRIX FOR TEST DATA

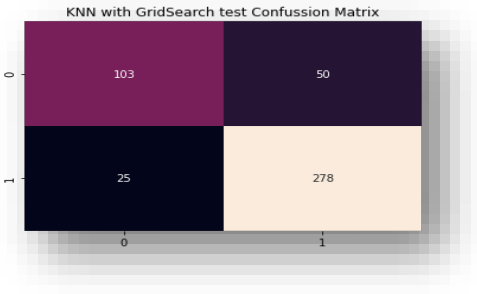


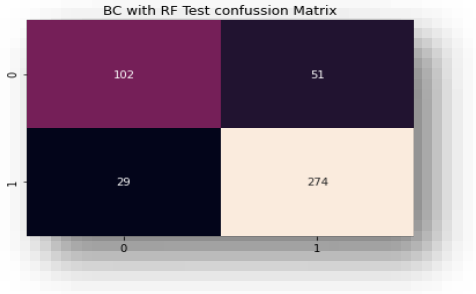


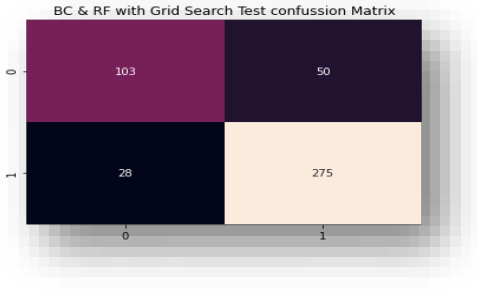


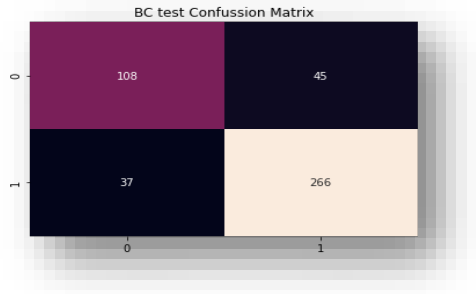


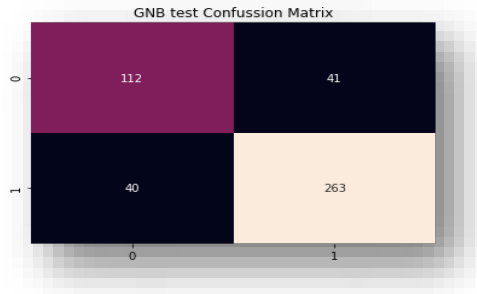


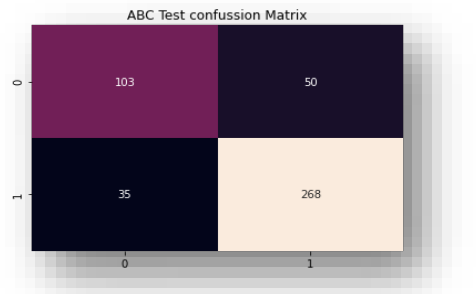


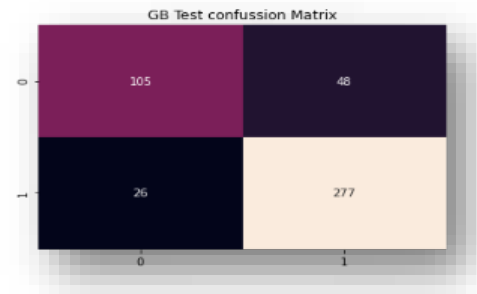












As for the model comparison we can check from the above model performance reports that Bagging Classifier and Bagging and Random Forest Classifier are both overfitted model and these models cannot be a good model for this business problem. So we are not including these models in model comaparison table below.

MODELS PERFORMANCE TABLE

MODEL COMPARISON												
SR NO.	MODEL		PRESICION		RECALL		F1 SCORE		ACCURACY		AUC SCORE	
			TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
1	LOGISTIC REGRESSION	0	0.74	0.76	0.64	0.74	0.69	0.75	0.83	0.84	0.89	0.883
		1	0.86	0.87	0.91	0.88	0.88	0.88				
2	LOGISTIC REGRESSION WITH GRID SEARCH	0	0.74	0.76	0.64	0.74	0.69	0.75	0.83	0.84	0.89	0.883
		1	0.86	0.87	0.91	0.88	0.88	0.88				
3	LINEAR DISCRIMINANT ANALYSIS	0	0.74	0.77	0.65	0.73	0.69	0.74	0.83	0.83	0.889	0.888
		1	0.86	0.86	0.91	0.89	0.89	0.88				
4	LINEAR DISCRIMINANT ANALYSIS WITH GRID SEARCH	0	0.74	0.77	0.65	0.73	0.69	0.74	0.83	0.83	0.889	0.888
		1	0.86	0.86	0.91	0.89	0.89	0.88				
5	GAUSSIAN NAÏVE BAYES	0	0.73	0.74	0.69	0.73	0.71	0.73	0.84	0.82	0.888	0.876
		1	0.88	0.87	0.9	0.87	0.89	0.87				
6	K-NEAREST NEIGHBOR	0	0.77	0.76	0.71	0.71	0.74	0.73	0.85	0.82	0.739	0.754
		1	0.88	0.86	0.91	0.88	0.9	0.87				
7	K-NEAREST NEIGHBOR WITH GRID SEARCH	0	0.76	0.8	0.64	0.67	0.69	0.73	0.84	0.84	0.826	0.804
		1	0.86	0.85	0.92	0.92	0.89	0.88				
8	BAGGING CLASSIFIER AND RANDOM FOREST WITH GRID SEARCH	0	0.79	0.79	0.68	0.67	0.73	0.73	0.85	0.83	0.912	0.891
		1	0.88	0.85	0.93	0.91	0.9	0.88				
9	ADA BOOST CLASSIFIER	0	0.76	0.75	0.7	0.67	0.73	0.71	0.85	0.81	0.915	0.877
		1	0.88	0.84	0.91	0.88	0.9	0.86				
10	GRADIENT BOOSTING	0	0.84	0.8	0.78	0.69	0.81	0.74	0.89	0.84	0.951	0.882
		1	0.91	0.85	0.94	0.91	0.93	0.88				

Table.1. 7 – MODEL COMPARISON

COMMENT:

After comparing all the models, If the company wants to get the result more specific towards 0 (Conservative) they can opt for the model which has highest score for recall. Similarly if the company needs to get the result more specific prediction about the 1 (Labour) than they need to chose the model which has highest recall value for 1.

But as this is news channel and company is more interested in overall prediction of election regarding which party has more chances of winning and covering how many seats we will prefer the model which has the highest accuracy amoint all models.

So keeping all aspects in mind we will choose Bagging & Random Forest with Grid search model as it has one of the highest accuracy and other parameters for both train and test data have almost similar scores. So this model will be the best for this business problem.

Q1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.

SOLUTION:

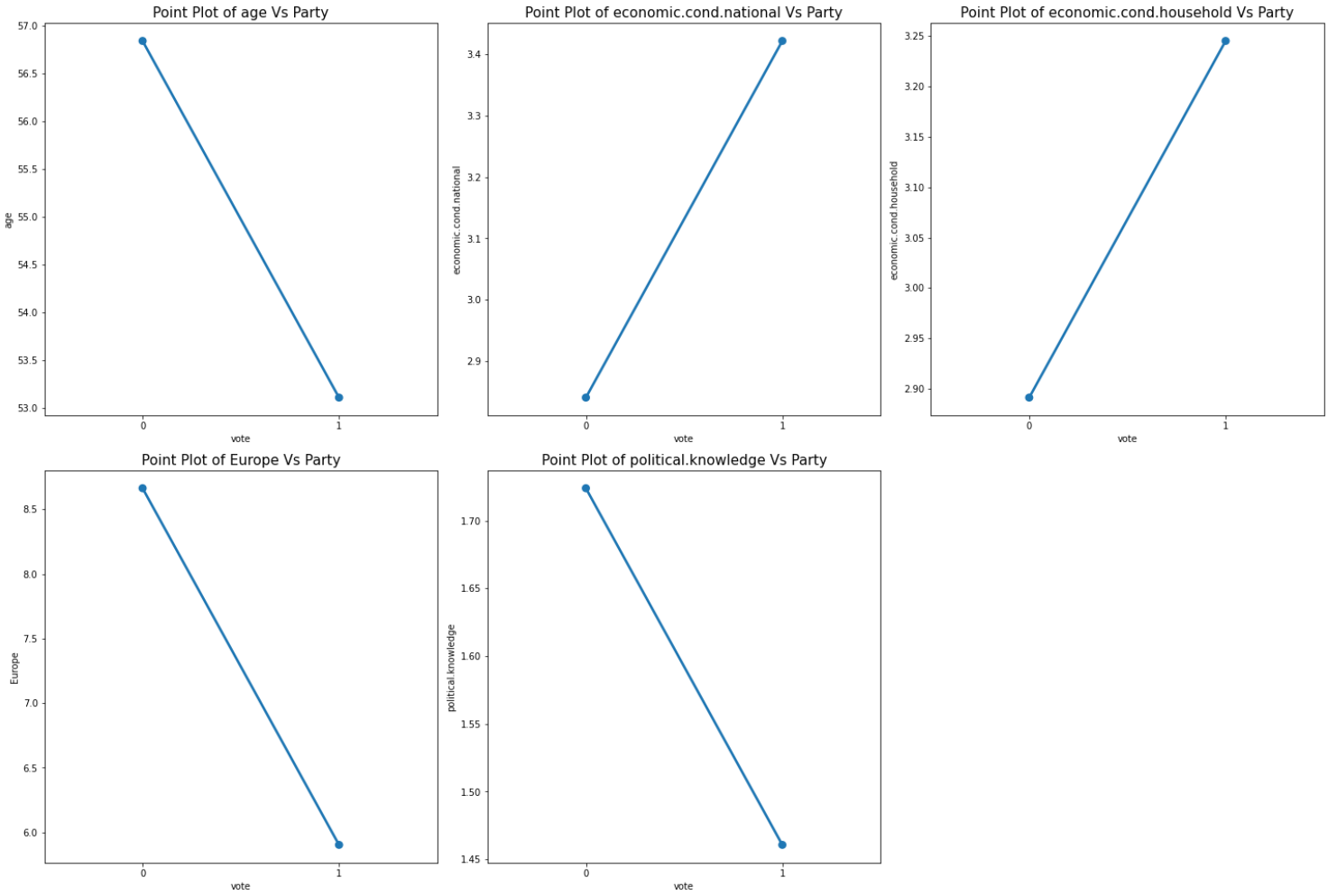


FIGURE.1. 7- BUSINESS INSIGHTS

COMMENT

Company can gather more data regarding conservative party there are less data, so that model can give better accuracy for predicting the Exit poll.

BUSINESS INSIGHTS

1. People with higher age are more leaned towards voting Conservative party, so company can gather more data with elder people.
2. People who believe that National Economic condition and Household Economic condition is higher are more likely to vote for Labour Party. Company can use this to differ between people who support either party and reason out why.
3. People with high Euro Sceptic thinking are more leaned to vote for Conservative Party.
4. People think that conservative party has more political knowledge than Labour, so this can be topic of debate.

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:

President Franklin D. Roosevelt in 1941

President John F. Kennedy in 1961

President Richard Nixon in 1973

After downloading the speech of all three presidents we have stored all these speeches in a dataset and converted it into a data frame to perform some analysis on the data.

DATA SUMMARY

1ST COLUMN IS THE INDEX

2ND COLUMN IS THE NAME OF THE PRESIDENT.

3RD COLUMN IS THE SPEECH GIVEN BY THE RESPECTICVE PRESIDENT

PRESIDENT SPEECH DATA		
	President	Speech
0	Roosevelt	On each national day of inauguration since 178...
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...

TABLE.2. 1- PRESIDENT SPEECH

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

SOLUTION:

Number of Characters, Word and Sentences

COUNTS					
	President	Speech	Character Count	Word Count	Sentence Count
0	Roosevelt	On each national day of inauguration since 178...	7571	1360	68
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	7618	1390	52
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9991	1819	68

TABLE.2. 2 - COUNTS

As we can check from the above table, there are different columns which gives the number of character, word and sentence present in each president's speech.

Q2.2) 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.

SOLUTION:

TEXT CLEANING

1. Converting all text into lower case.

Before converting to lower case.

PRESIDENT SPEECH DATA		
	President	Speech
0	Roosevelt	On each national day of inauguration since 178...
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...

After converting to lower case

LOWER CASE.		
	President	Speech
0	Roosevelt	on each national day of inauguration since 178...
1	Kennedy	vice president johnson, mr. speaker mr. chief…
2	Nixon	mr. vice president, mr. speaker, mr. chief jus...

TABLE.2. 3- LOWER CASE SPEECH

As we can from before and after table than all the words have been converted to lower.

This is Necessary because python is case sensitive and any letter in capital in any word it will consider as different word, so to perform any analysis on text data we need to convert all the text in to lower case.

2. Punctuation Removal.

PUNCTUATION

Punctuation is the use of spacing, conventional signs, and certain typographical devices as aids to the understanding and correct reading of written text, whether read silently or aloud.

Punctuations have been removed from all the speeches of each president.

Let's check President Roosevelt Speech first sentence.

Before Stop Words removal:

"On each National Day of Inauguration since 1789, the people have re newed their sense of dedication to the United States."

After Removal of Stop Words:

"National day inauguration since 1789 people renewed sense dedicatio n united states"

As we can check "," after 1789 and "." After united states has been removed from the sentence.

3. Stop Words Removal.

Stop words have been removed from all the speeches of each president.

Let's check President Roosevelt Speech first sentence.

Before Stop Words removal:

"On each National Day of Inauguration since 1789, the people have renewed thei r sense of dedication to
the United States."

After Removal of Stop Words:

"National day inauguration since 1789 people renewed sense dedication united st ates"

As we can check "On", "each", "of", "the", "have", "their", "to" has been removed from the sentence.

STOP WORDS

Stop words are any word in a stop list which are filtered out before or after processing of natural language data. There is no single universal list of stop words used by all natural language processing tools, nor any agreed upon rules for identifying stop words, and indeed not all tools even use such a list.

COUNTS				
	President	Speech	Word Count	Word Count After Cleaning
0	Roosevelt	national day inauguration since 1789 people re...	1360	627
1	Kennedy	vice president johnson mr speaker mr chief jus...	1390	693
2	Nixon	mr vice president mr speaker mr chief justice ...	1819	833

TABLE.2. 4- STOPWORD BEFORE AND AFTER COUNT

Q-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).

SOLUTION:

MOST OCCURING WORDS

1. For President Roosevelt Speech.

MOST FREQUENT WORDS		
	Most Common Word	Frequency
0	nation	11
1	know	10
2	spirit	9

TABLE.2. 5 – ROOSEVELT'S MOST FREQUENT WORDS

Above table consist of the words which occurred most of the time in president Roosevelt's speech and their frequency. "nation", "know", "spirit" is most frequently used words by President Roosevelt in his speech.

2. For President Kennedy

MOST FREQUENT WORDS		
	Most Common Word	Frequency
0	let	16
1	us	12
2	world	8

TABLE.2. 6 – KENNEDY'S MOST FREQUENT WORDS

Above table consist of the words which occurred most of the time in president Kennedy's speech and their frequency. "let", "us", "world" is most frequently used words by President Roosevelt in his speech.

3. For President Nixon

MOST FREQUENT WORDS		
	Most Common Word	Frequency
0	us	26
1	let	22
2	peace	19

TABLE.2. 7 - NIXON'S MOST FREQUENT WORDS

Above table consist of the words which occurred most of the time in president Nixon's speech and their frequency. "us", "let", "peace" is most frequently used words by President Roosevelt in his speech.

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

SOLUTION:

WORD CLOUD

A tag cloud is a visual representation of text data, which is often used to depict keyword metadata on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

President Roosevelt's Speech Word cloud.

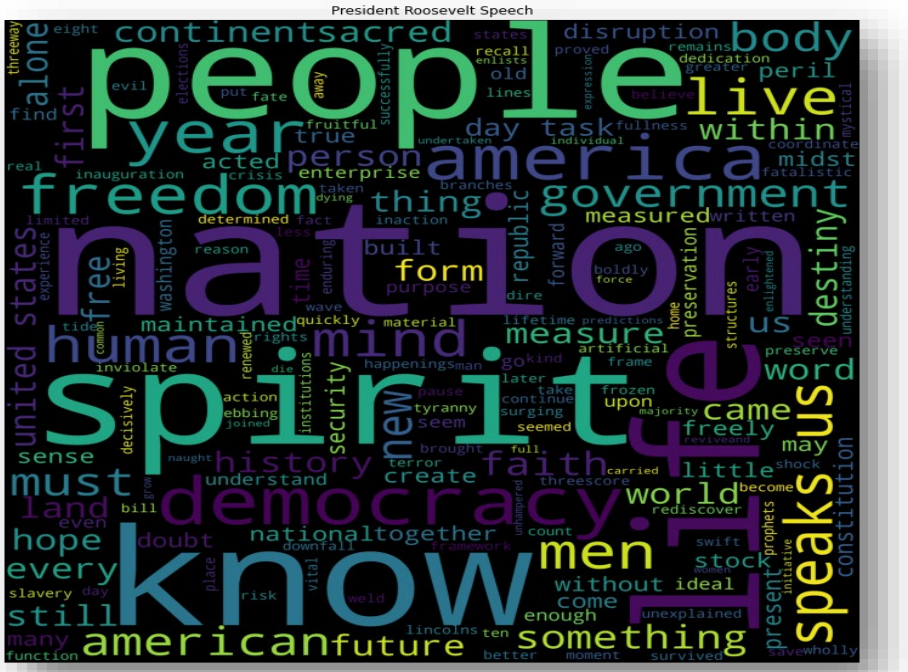


FIGURE.2. 1 – ROOSEVELT'S SPEECH WORD CLOUD

This is the world cloud for president Roosevelt's Speech and we can see that "nation", "spirit", "know" are the most frequent occurring word in his speech which we have already checked in the above question.

President Kennedy's speech world cloud.

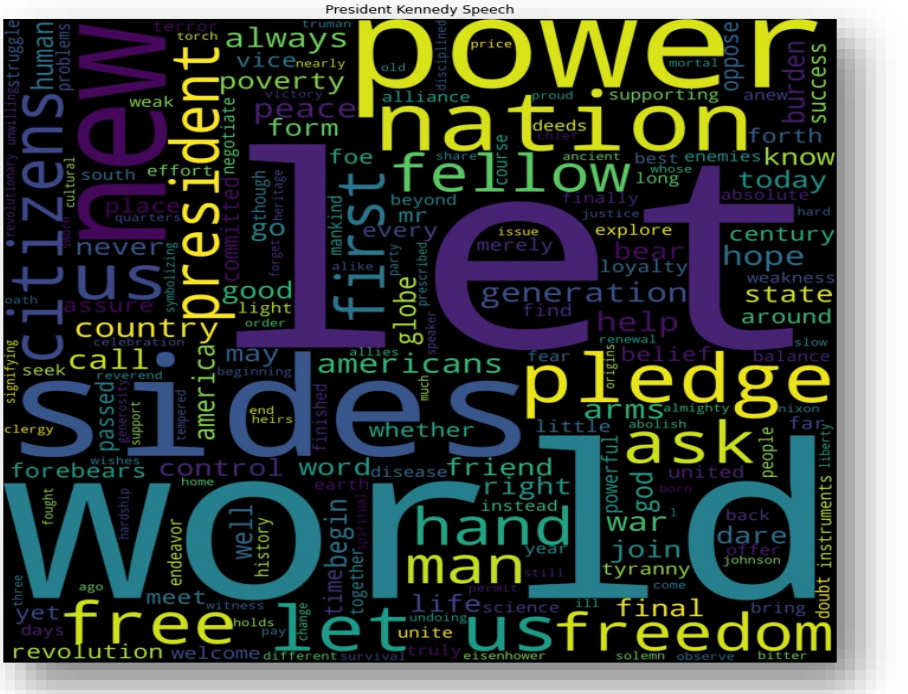


FIGURE.2. 2 - KENNEDY'S SPEECH WORD CLOUD

This is the world cloud for president Roosevelt's Speech and we can see that "let", "us", "world" is the most frequent occurring word in his speech which we have already checked in the above question.

President Nixon Speech Word Cloud.

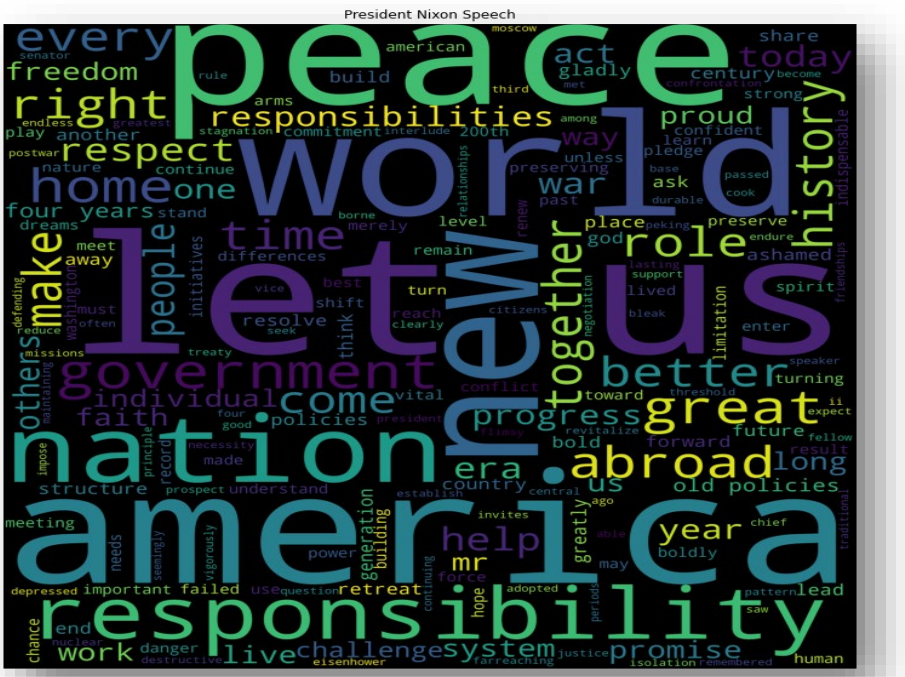


FIGURE.2. 3 - NIXON'S SPEECH WORD CLOUD

This is the world cloud for president Roosevelt's Speech and we can see that "let", "us", "peace" is the most frequent occurring word in his speech which we have already checked in the above question.