greatlearning

Learning for Life

BUSINESS REPORT

PREPARED BY - RUPESH KUMAR

SUBMISSION DATE: 14/11/2021

Table of Contents

List of Figures

List of Figures

List of Figures

List of Tables

List of Tables

List of Tables

List of Tables

Problem Statement 1: Machine Learning

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.

EXECUTIVE SUMMARY

A Leading News Channel **CNBE**. wants to analyze the recent elections. The dataset consists of 1525 voters with 9 variables. Based on the different attributes of the dataset we 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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis , visualization & apply various supervised learning algorithms i.e. **Logistics Regression , LDA ,KNN & Navie Bayes** 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. Explore the dataset using central tendency and other parameters. The data consists of 1525 different voters with 9 unique activities . Analyse the different attributes of the voters which can help in predicting which party a voter will vote for on the basis of the given information.This assignment should help the news channel to create an exit poll that will help in predicting overall win and seats covered by a particular party.

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

EDA - Data Description -

Checking the Records of the Dataset :

Head of the Dataset - First 10 Records of the Dataset.
Tail of the Dataset - Last 10 Records of the Dataset.

	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
5	6	Labour	47	3	4	4	4	4	2	male
6	7	Labour	57	2	2	4	4	11	2	male
7	8	Labour	77	3	4	4	1	1	0	male
8	9	Labour	39	3	3	4	4	11	0	female
9	10	Labour	70	3	2	5	1	11	2	male

	Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1515	1516	Conservative	82	2	2	2	1	11	2	female
1516	1517	Labour	30	3	4	4	2	4	2	male
1517	1518	Labour	76	4	3	2	2	11	2	male
1518	1519	Labour	50	3	4	4	2	5	2	male
1519	1520	Conservative	35	3	4	4	2	8	2	male
1520	1521	Conservative	67	5	3	2	4	11	3	male
1521	1522	Conservative	73	2	2	4	4	8	2	male
1522	1523	Labour	37	3	3	5	4	2	2	male
1523	1524	Conservative	61	3	3	1	4	11	2	male
1524	1525	Conservative	74	2	3	2	4	11	0	female

Tab:1 Records of the Dataset Head & Tail

Note: Dropping the Unnamed: 0 Column & Changing the name of columns. We are going to drop the column unnamed:0 as it is useless for the model.checking the dataset again.

	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
5	Labour	47	3	4	4	4	4	2	male
6	Labour	57	2	2	4	4	11	2	male
7	Labour	77	3	4	4	1	1	0	male
8	Labour	39	3	3	4	4	11	0	female
9	Labour	70	3	2	5	1	11	2	male

Tab:2 Records of the Dataset With Finalised Columns

Observation: Now we have all the columns which are useful for the model.

Data Dictionary for Problem Statement 1 :

Features/variables	Description
vote :	Party choice: Conservative or Labour
age :	in years
economic_cond_national :	Assessment of current national economic conditions, 1 to 5.
economic_cond_household :	Assessment of current household economic conditions, 1 to 5.
Blair :	Assessment of the Labour leader, 1 to 5.
Hague :	Assessment of the Conservative leader, 1 to 5.
Europe :	an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
political_knowledge :	Knowledge of parties' positions on European integration, 0 to 3.
gender :	female or male.

Tab:3 Data Dictionary of the Dataset

Checking the Summary of the Dataset :

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

Tab:4 Summary of the Dataset

Insights

- From the above table we can infer the count,mean, std , 25% , 50% ,75% and min & max values of the all numeric variables present in the dataset.
- From the above table we can infer the count, unique ,top ,freq of all the categorical variables present in the dataset.
- Their is no bad value present in the dataset.

Skewness of the Dataset :

In statistics, skewness is a measure of asymmetry of the probability distribution about its mean and helps describe the shape of the probability distribution. Basically it measures the level of how much a given distribution is different from a normal distribution (which is symmetric).

Skewness is a well-established statistical concept for continuous and to a lesser extent for discrete quantitative statistical variables.Here we are going to check the skewness of the features which are present in our dataset.

Features / Columns	Skewness	Observation
age	0.14462077228942483	slightly right skewed
economic_cond_national	-0.2404528899412957	slightly left skewed
economic_cond_household	-0.14955204997804528	slightly left skewed
Blair	-0.5354186518673825	slightly left skewed
Hague	0.1520996272526911	slightly right skewed
Europe	-0.13594670991422228	slightly left skewed
political_knowledge	-0.42683782344871657	slightly left skewed

Tab:5 Skewness of the Dataset

Insights :

From the above skewness table we infer that most of the features are slightly left skewed.

Checking the Shape of the Dataframe :

Shape attribute tells us number of observations and variables we have in the data set. It is used to check the dimension of data. The Election_Data.xlsx data set has 1525 observations (rows) and 9 variables (columns) in the dataset.

No. of Rows	No. of Columns
1525	9

Tab:6 Shape of the Dataframe

Checking the Appropriateness of Datatypes & Information of the Dataframe :

The info() function is used to print a concise summary of a DataFrame. This method prints information about a DataFrame including the index d-type and column d-types, non-null values and memory usage.

S.No.	Features / Columns	Non-Null Count	Dtype
1	vote	1525 non-null	object
2	age	1525 non-null	int64
3	economic_cond_national	1525non-null	int64
4	economic_cond_household	1525 non-null	int64
5	Blair	1525 non-null	int64
6	Hague	1525 non-null	int64
7	Europe	1525 non-null	int64
8	political_knowledge	1525 non-null	int64
9	gender	1525 non-null	object

Tab:7 Appropriateness of Datatypes & Information of the Dataframe

Insights :

From the above results we can see that there is no null values present in the dataset.Their are total 1525 rows & 9 columns in this dataset,indexed from 0 to 1524. Out of 9 variables 7 are int64 , 2 variables are object. Memory used by the dataset: 107.4+ KB.

Checking for Null Values :

S.No.	Features / Columns	Null Count
1	vote	0
2	age	0
3	economic_cond_national	0
4	economic_cond_household	0
5	Blair	0
6	Hague	0
7	Europe	0
8	political_knowledge	0
9	gender	0

Tab:8 Checking Null Values.

Insights :

From the above output we infer that their are no null values present in the dataset.

Checking for Anomalies in the Dataset : .

vote :

array(['Labour', 'Conservative'], dtype=object)

age :

array([43, 36, 35, 24, 41, 47, 57, 77, 39, 70, 66, 59, 51, 79, 37, 38, 53,
44, 60, 56, 61, 55, 62, 76, 27, 52, 48, 72, 42, 54, 50, 46, 33, 58,
64, 32, 71, 28, 34, 68, 67, 88, 40, 78, 65, 74, 82, 49, 84, 81, 45,
69, 31, 63, 89, 83, 29, 92, 73, 75, 26, 90, 25, 80, 30, 86, 85, 87,93, 91])

economic_cond_national

array([3, 4, 2, 1, 5])

economic_cond_household

array([3, 4, 2, 1, 5])

Blair

array([4, 5, 2, 1, 3])

Hague

array([1, 4, 2, 5, 3])

Europe

array([2, 5, 3, 4, 6, 11, 1, 7, 9, 10, 8])

political_knowledge

array([2, 0, 3, 1])

gender

array(['female', 'male'], dtype=object)

Tab:9 Checking for Anomalies for variables in the Dataset

Observations

No Anomalies found in the Dataset.

Checking the Value counts on all the Categorical Column :

S.No.	Vote	Count
1	Labour	1063
2	Conservative	462

Tab:10 Checking the Value counts on the Categorical Column - Vote

Insights:

- As per the given Data Dictionary there are 2 parties for voting Conservative and Labour.
- 1063 people voted for Labour party.
- 462 people voted for Conservative party.

S.No.	Gender	Count
1	female	812
2	male	713

Tab:11 Checking the Value counts on the Categorical Column - Gender

Insights :

- 812 female voters present in the dataset.
- 713 male voters present in the dataset

Checking Duplicate Values :

Here we found the no of duplicated rows in data set i.e. 8 , as we know that duplicated rows are not useful we decided drop them by using .drop( ) function.

After using the .drop function ( ) , we drop all the duplicate rows of the data ,check the shape of the data once again.

Number of Rows :1517
Number of Columns :9

Observation :
Number of duplicate rows = 0

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

Univariate Analysis of continuous Numerical Variables.

A histogram takes as input a numeric variable only. The variable is cut into several bins, and the number of observation per bin is represented by the height of the bar. It is possible to represent the distribution of several variable on the same axis using this technique.

A box-plot gives a nice summary of one or several numeric variables. The line that divides the box into 2 parts represents the median of the data. The end of the box shows the upper and lower quartiles. The extreme lines show the highest and lowest value excluding outliers.

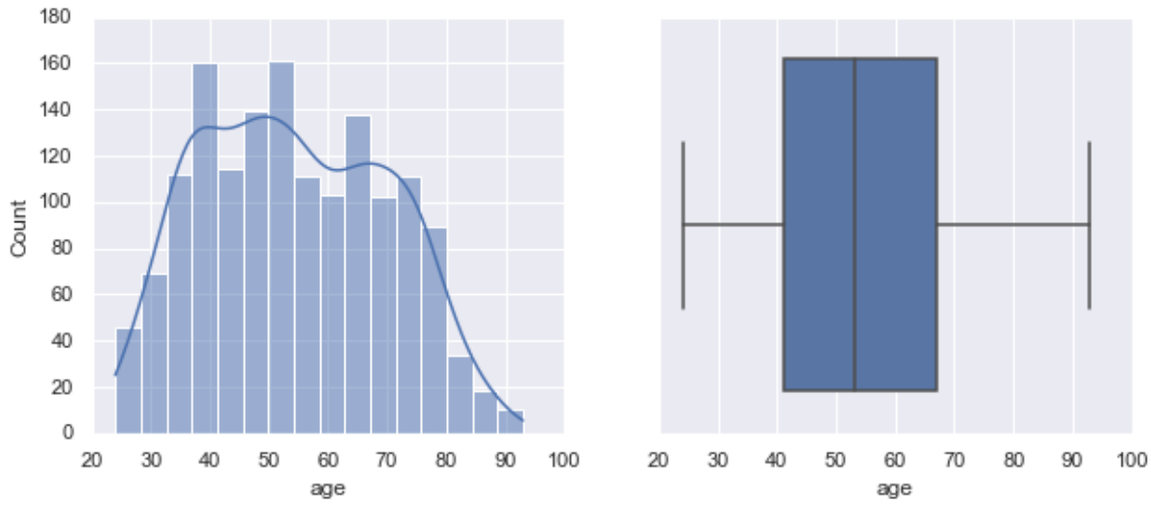

FIG: 1 Histogram & Box-Plot of Age

Statistical Summary	Values
count	1517.0
mean	54.241266
std	15.701741
min	24.0
25%	41.0
50%	53.0
75%	67.0
max	93.0
skewness	0.1396615989084527

Tab:12 Statistical Description of Age

Insights :

- Age : age of voters ranges from a minimum of 24 to maximum of 93.
- The average Age : age of voters is around 54.241266.
- The standard deviation of Age : age of voters is 15.701741.
- 25% , 50% (median) and 75 % of Age : age of voters are 41 , 53 and 67.
- Skewness indicating that the distribution is slightly right skewed.
- Age: age of voters don't have outliers.

Note :

As we have only one continuous variable with us and we plot the above the histogram & box-plot for the age variable. For the remaining discrete variables we are going to plot count-plot.

Univariate Analysis of Discrete Variables :

PieChart :

A pie chart is a circle divided into sectors that each represent a proportion of the whole. It is often used to show proportion, where the sum of the sectors equal 100%.

economic_cond_national -
Assessment of current national economic conditions, 1 to 5.The participants have assessed several parameters on a scale of 1 to 5. Here, is that 5 stands for a high score, 1 stands for a poor score (for e.g., a participant who gives a score of 5 for the current national economic translates to him/her perceiving the economic conditions to be very good or a participant who gives a score of 1 for the current national economic translates to him/her perceiving the economic conditions to be very poor).

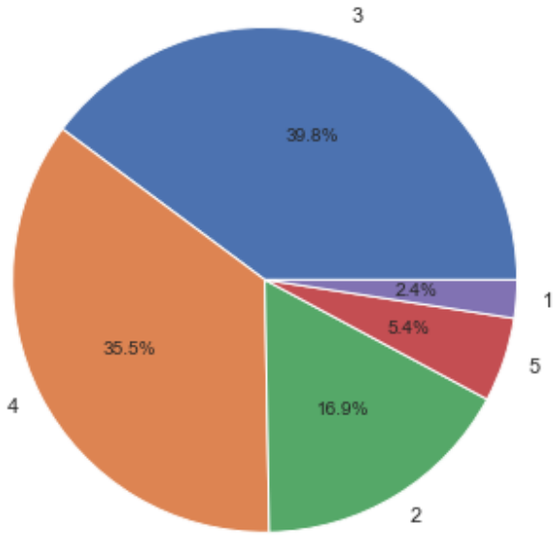

FIG: 2 Pie-Plot of economic_cond_national

Insights :

- Most of the participants/voters around (39.8%) gives a score of 3 for the current national economic conditions.
- 36.5% participants/voters gives a score of 4 for the current national economic conditions.
- 16.9% participants/voters gives a score of 2 for the current national economic conditions.
- 5.4% participants/voters gives a score of 5 for the current national economic conditions.
- Only 2.4% participants/voters gives a score of 1 for the current national economic conditions.

economic_cond_household :
Assessment of current household economic conditions, 1 to 5.The participants/voters have assessed several parameters on a scale of 1 to 5. Here, is that 5 stands for a high score and 1 stands for a poor score (for e.g., a participant/voter who gives a score of 5 for the current household economic translates to him/her perceiving the household economic conditions to be very good or a participant/voter who gives a score of 1 for the current household economic translates to him/her perceiving the household economic conditions to be very poor).

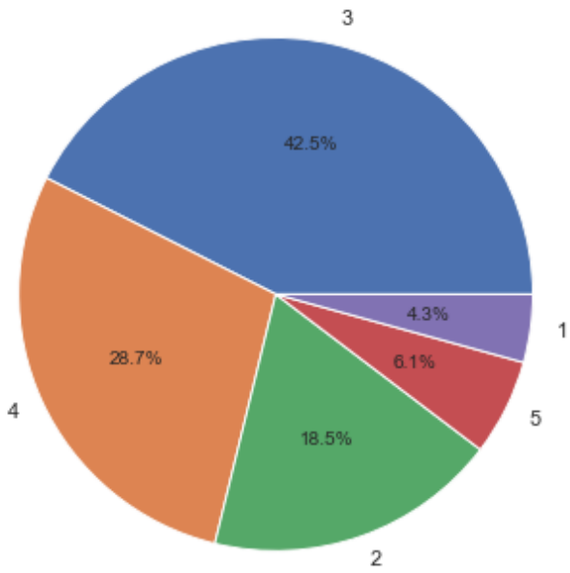

FIG: 3 Pie-Plot of economic_cond_household

Insights :

- Most of the participants/voters around (42.5%) gives a score of 3 for the current household economic conditions.
- 28.7% participants/voters gives a score of 4 for the current household economic conditions.
- 18.5% participants/voters gives a score of 2 for the current household economic conditions.
- 6.1% participants/voters gives a score of 5 for the current household economic conditions.
- Only 4.3% participants/voters gives a score of 1 for the current household economic conditions.

Blair

Assessment of the Labour leader (Tony Blair - Name for the person contesting for the labour party), 1 to 5.
The participants/voters have assessed several parameters on a scale of 1 to 5. Here, is that 5 stands for a highest score and 1 stands for a least score.

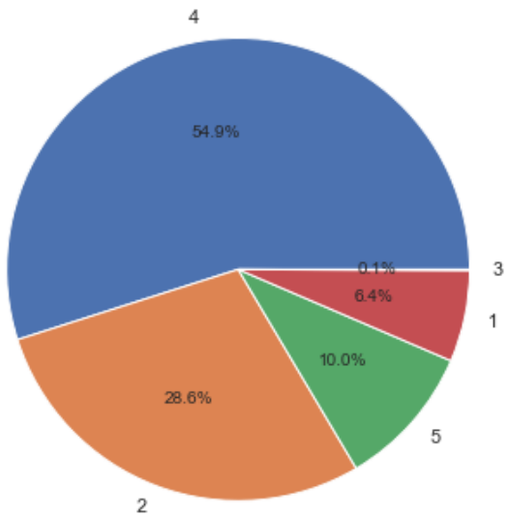

FIG: 4 Pie-Plot of Blair

Insights :

- Most of the participants/voters around (54.9%) gives a score of 4 to Blair.
- 28.6% of the participants/voters gives a score of 2 to Blair.
- 10% of the participants/voters gives a score of 5 to Blair.
- 6.4% of the participants/voters gives a score of 1 to Blair.
- Only 0.1% of the participants/voters gives a score of 3 to Blair.

Hague

Assessment of the Conservative leader (Hague- Name for the person contesting for the Conservative party), 1 to 5.The participants/voters have assessed several parameters on a scale of 1 to 5. Here, is that 5 stands for a highest score and 1 stands for a least score.

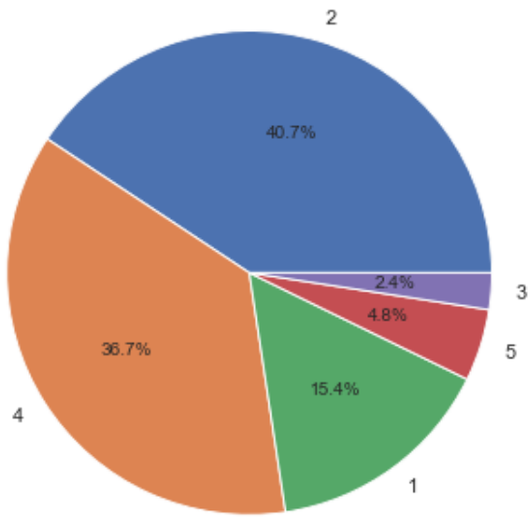

FIG: 5 Pie-Plot of Hague

Insights :

- Most of the participants/voters around (40.7%) gives a score of 2 to Hague.
- 36.7% of the participants/voters gives a score of 4 to Hague.
- 15.4% of the participants/voters gives a score of 1 to Hague.
- 4.8% of the participants/voters gives a score of 5 to Hague. of the
- Only 2.8% of the participants/voters gives a score of 1 to Hague.

Europe :

An 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.As per the data dictionary we know that a 11-point scale that measures Eurosceptic means criticism of the European Union (EU) and European integration and 1 point scale being the opposite of Eurosceptism which is also known as European Unionism.In the europe column 5-6 is considered to be neutral, 1-4 being european unionism and 7-11 being eurosceptic.

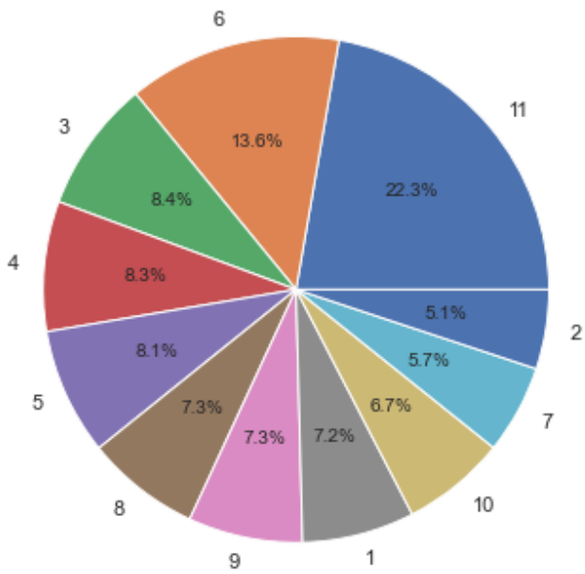

FIG: 6 Pie-Plot of Europe

Insights :

- 29% of the voters have represent European Unionism sentiments.
- Around 21.7% of the voters are considered to be neutral.
- 49.3% of the voters have represent Eurosceptic sentiment.

political_knowledge :

knowledge of the voters regarding the party's position on European integration,0 to 3.
level 0 stands the least, level 3 stands for the highest.

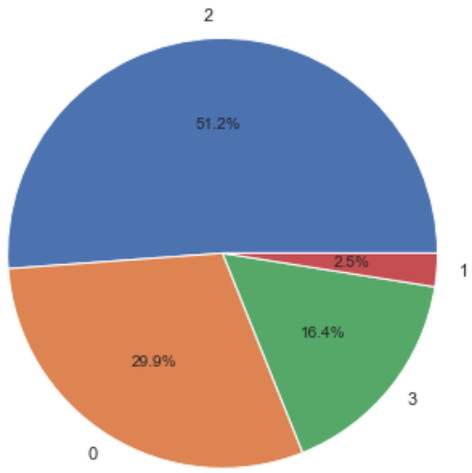

FIG: 7 Pie-Plot of political_knowledge

Insights :

- 29.9% of the voters have least i.e. (0) political knowledge - (knowledge of the voters regarding the party's position on European integration).
- 16.4% of voters have high i.e(3) political knowledge - (knowledge of the voters regarding the party's position on European integration).
- 51.2% of voters have fair i.e.(2) political knowledge - (knowledge of the voters regarding the party's position on European integration).
- 2.5% of voters have level(1) political knowledge - (knowledge of the voters regarding the party's position on European integration).

Univariate Analysis of Categorical Variables :

vote

Party choice: Conservative or Labour.

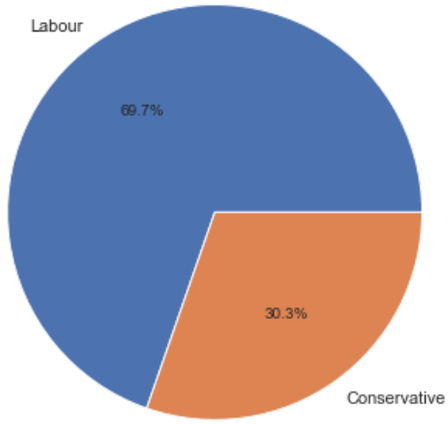

FIG:8 Pie-Plot of Vote

Insights :

- 69.7% of the participants/voters vote for Labour Party.
- 30.3% of the participants/voters vote for Conservative Party.

gender :
female or male.

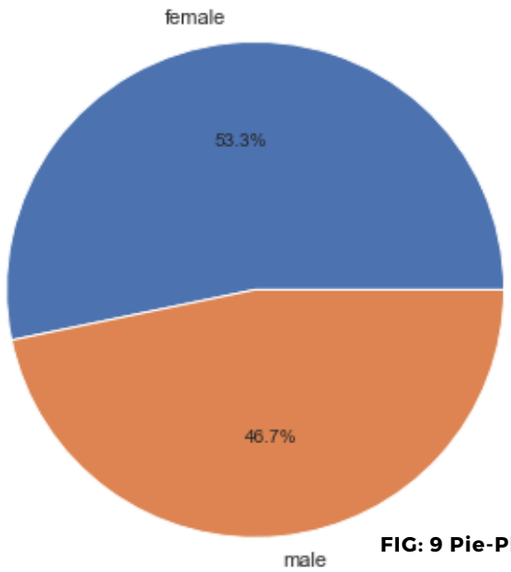

FIG: 9 Pie-Plot of Gender

Insights :

- **53.3% of the voters are female.**
- **46.7% of the voters are males.**

Bivariate Analysis :

Scatter Plot :

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

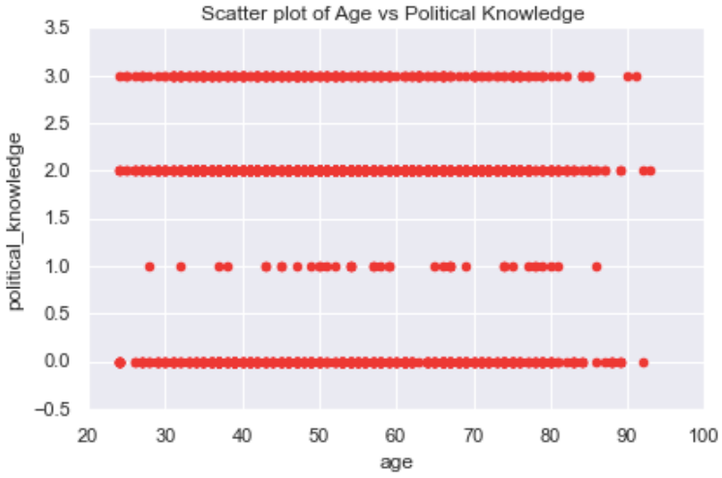

FIG: 10 Scatter Plot of Age VS Political_Knowledge

Insights :

- Their is no such relationship between age & political_knowledge.

FIG: 11 Scatter Plot of Age VS Economic Cond. National

Insights :

- Their is no such relationship between age & economic_cond_national.

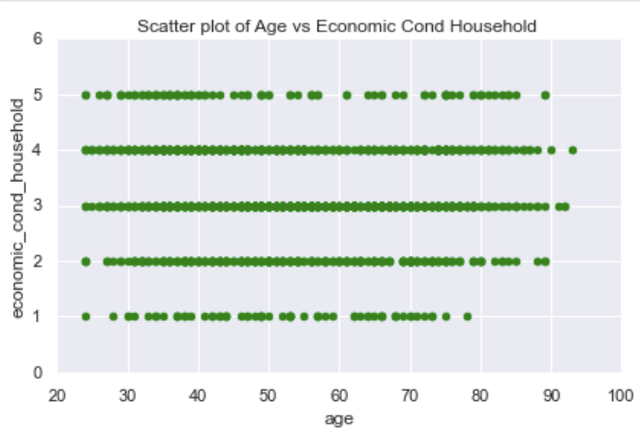

FIG: 12 Scatter Plot of Age VS Economic Cond. Household

Insights :

- Their is no such relationship between age & economic_cond_household.

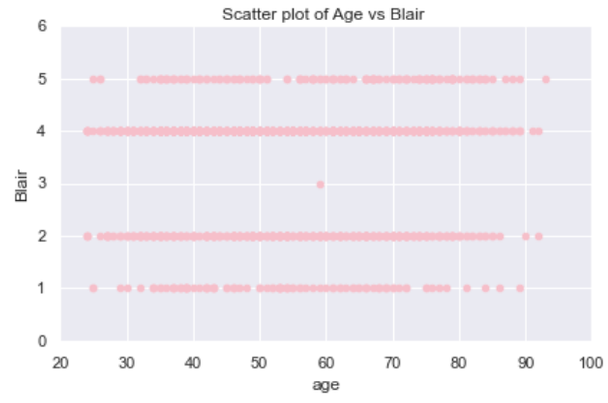

FIG: 13 Scatter Plot of Age VS Blair

Insights :

- Their is no such relationship between age & Blair.

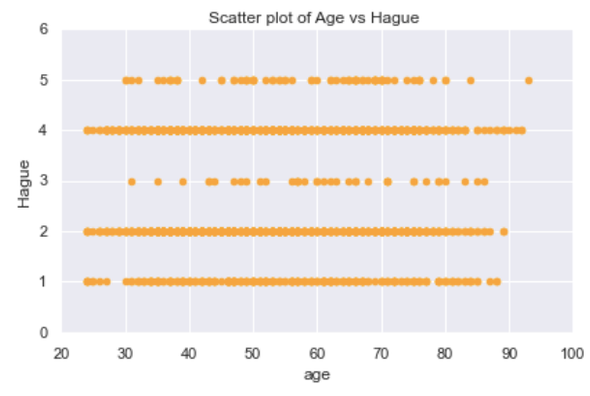

FIG: 14 Scatter Plot of Age VS Hague

Insights :

- Their is no such relationship between age & Hague.

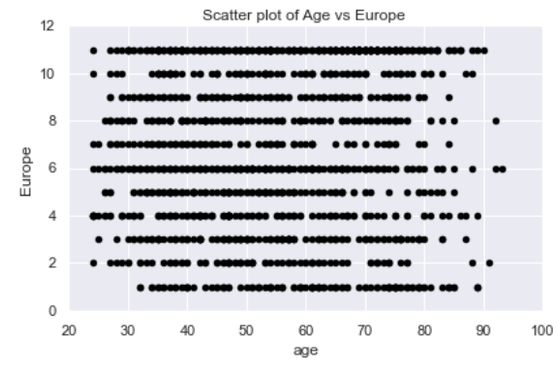

FIG: 15 Scatter Plot of Age VS Europe

Insights :

- Their is no such relationship between age & Europe.

Count-plot with Hue :

A count-plot is kind of like a histogram or a bar graph for categorical variables.
Hue :This parameter take column name for colour encoding.

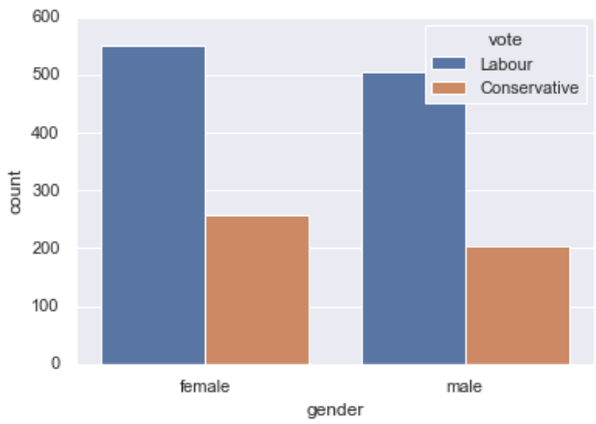

FIG: 16 Count-Plot with Hue Gender VS Vote

Insights :

- Majority of voters female or male voted for labour party.
- Female voters are more than as compared to male voters.
- Among female voters majority of voters voted for Labour party & similarly among the male voters majority of voters voted for Labour party.
- Their is less number of voters female or male who voted for conservative party as compared to the labour party.

Box-plot :

A box-plot gives a nice summary of one or several numeric variables. The line that divides the box into 2 parts represents the median of the data. The end of the box shows the upper and lower quartiles. The extreme lines show the highest and lowest value excluding outliers.

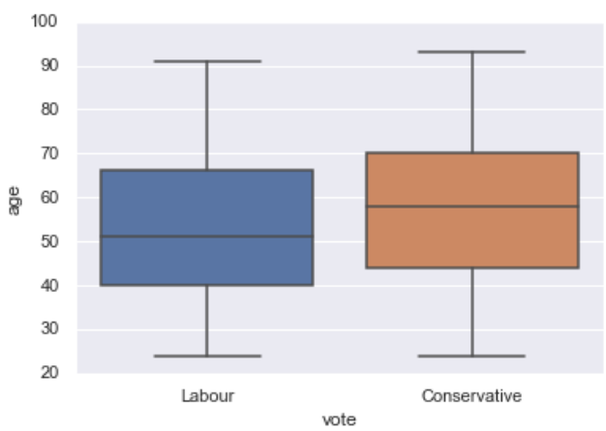

FIG: 17 Box-Plot of Vote VS Age

Insights :

- 50% voters who voted for labour party are in the age level of 48-52.
- 50% voters who voted for conservative party are in the age level of 55-58.
- Median of age of voters who voted for conservative party is more than the labour party.
- It can be easily observed that relatively younger people have voted for “Labour” party in comparison to that of older people who voted for “Conservative” party.

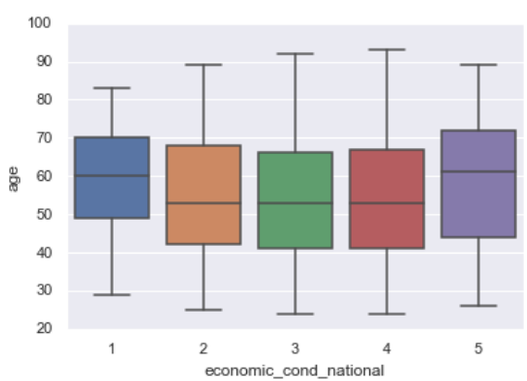

FIG: 18 Box-Plot of Economic Cond. National VS Age

Insights :

- A large proportion of people assess the national economic condition to be average to good (rated 3, 4 and 5), it is interesting to note that the older people in the age range of 50-70 feel that the economic condition is poor.

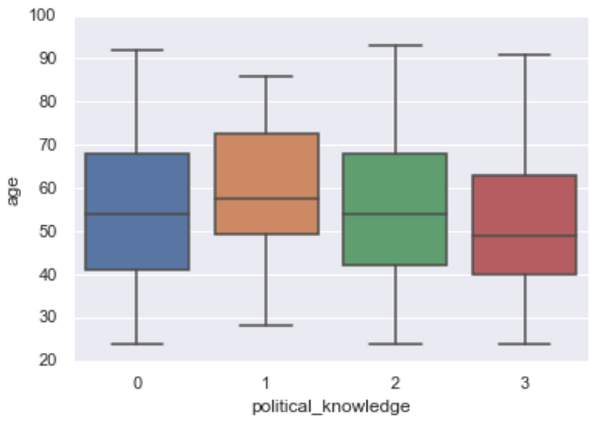

FIG: 19 Box-Plot of Political Knowledge VS Age

Insights :

- There is an evenly distributed number of people when it comes to thier political knowledge about their party's position on European integration.

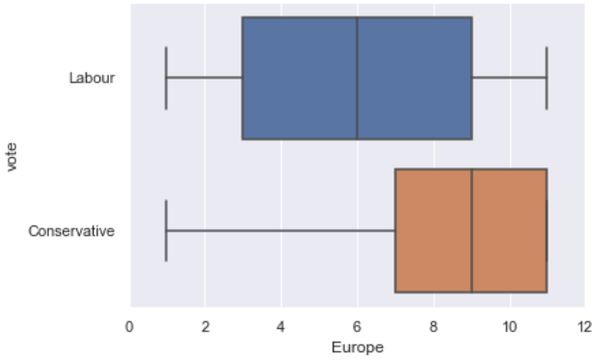

FIG: 20 Box-Plot of Europe VS Vote

Insights :

- Distribution is larger for the Labour party as compared the Conservative party.
- Majority of people with 'Eurosceptic' sentiment voted for "Labour" party.

Multivariate Analysis :

Heat-map :

A correlation heat-map uses coloured cells, typically in a monochromatic scale, to show a 2D correlation matrix (table) between two discrete dimensions or event types.Correlation heat-maps are ideal for comparing the measurement for each pair of dimension values.Darker Shades have higher Correlation , while lighter shades have smaller values of Correlation as compared to darker shades values.Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

Checking for Correlations :

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge
age	1.000000	0.018687	-0.038868	0.032084	0.031144	0.064562	-0.046598
economic_cond_national	0.018687	1.000000	0.347687	0.326141	-0.200790	-0.209150	-0.023510
economic_cond_household	-0.038868	0.347687	1.000000	0.215822	-0.100392	-0.112897	-0.038528
Blair	0.032084	0.326141	0.215822	1.000000	-0.243508	-0.295944	-0.021299
Hague	0.031144	-0.200790	-0.100392	-0.243508	1.000000	0.285738	-0.029906
Europe	0.064562	-0.209150	-0.112897	-0.295944	0.285738	1.000000	-0.151197
political_knowledge	-0.046598	-0.023510	-0.038528	-0.021299	-0.029906	-0.151197	1.000000

Tab:13 Correlation Table

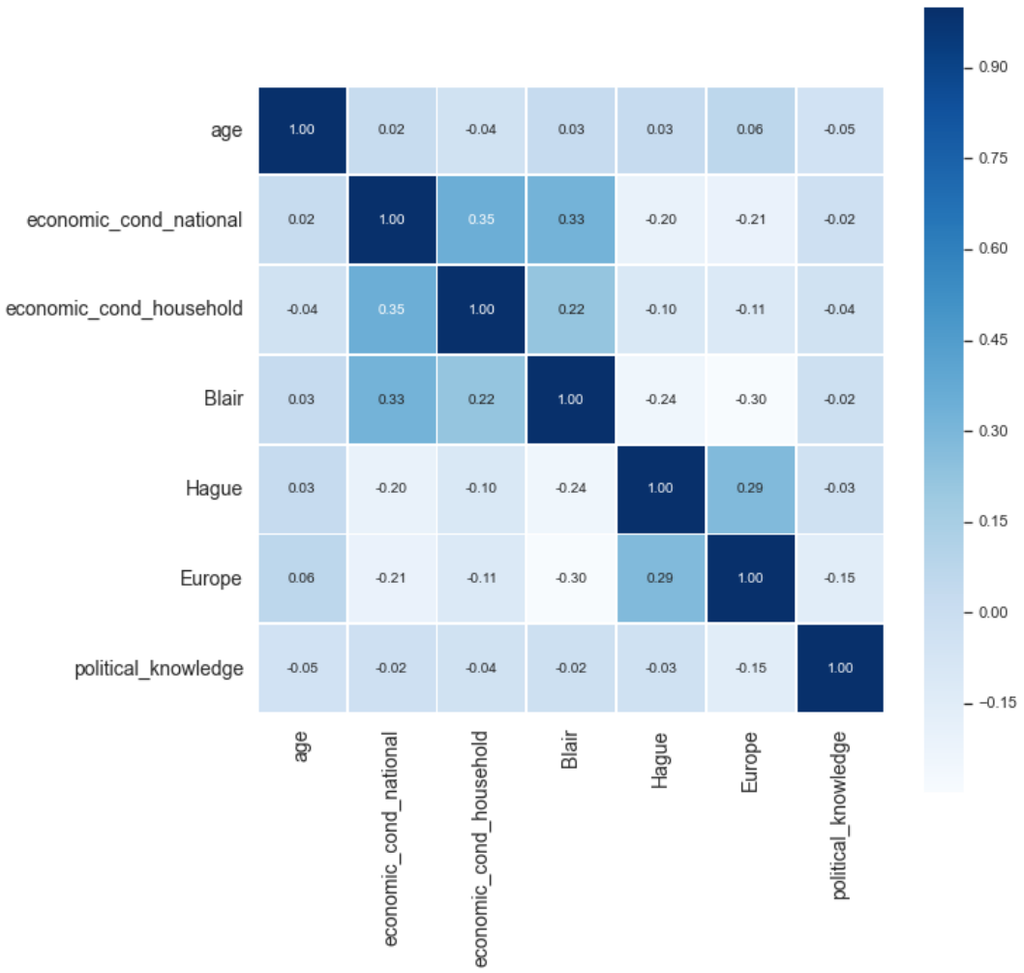

FIG: 21 Heat-Map of Problem 1

Insights :

- economic_cond_national with economic_cond_household shows max correlation i.e.0.35 , but the correlation is weak.
- economic_cond_national with Blair shows weak correlation i.e.0.33.
- economic_cond_household with Blair shows weak correlation i.e.0.22.
- economic_cond_national with Hague & Europe shows poorest correlation i.e.-0.20,-0.21.

- **Note:** Their is no such relationship between the variables , hence their is no problem of multicollinearity in the dataset.

Pairplot :

Pairplot shows the relationship between the variables in the form of scatter-plot and the distribution of the variable in the form of histogram.

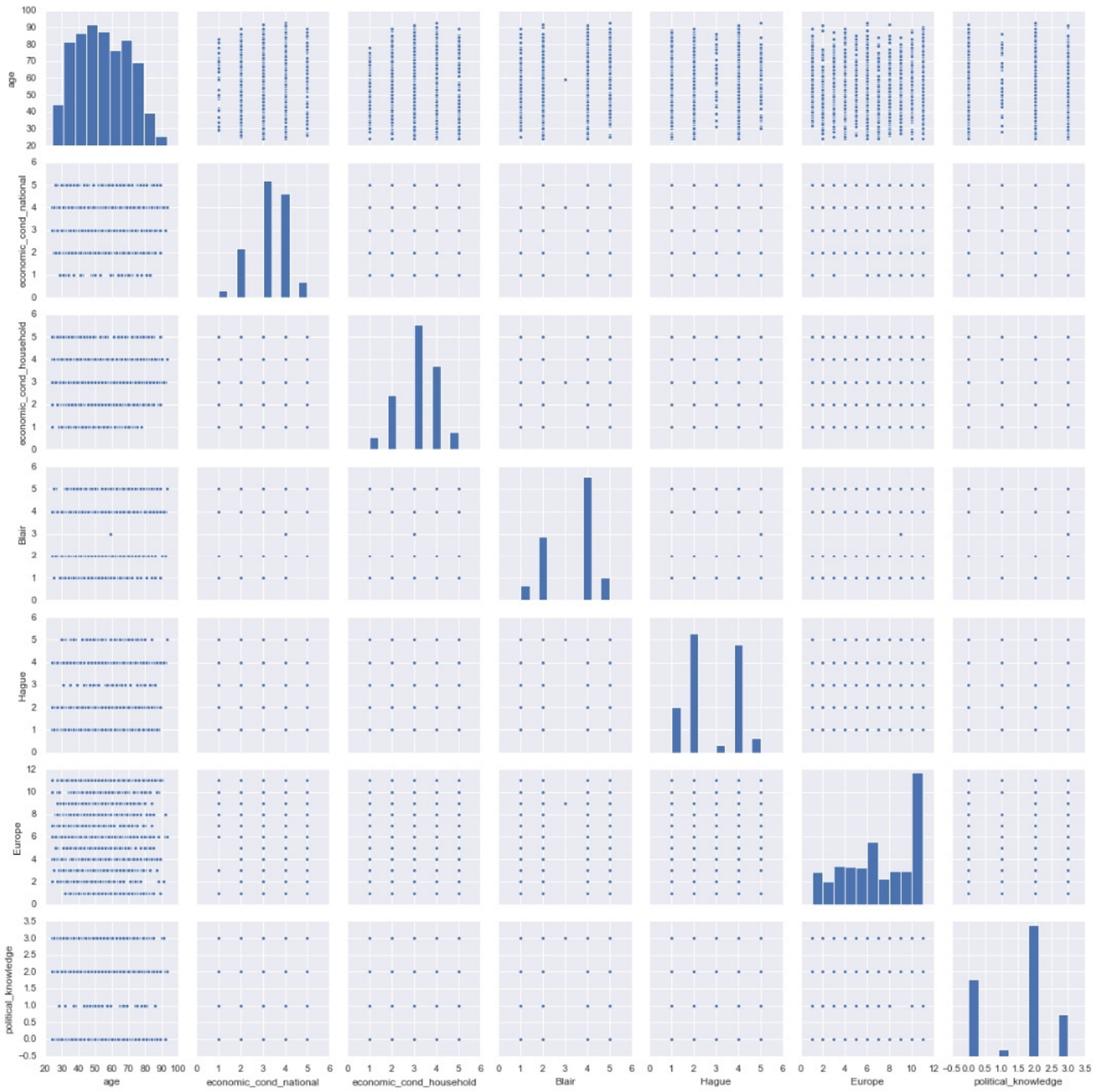

FIG: 22 Pair Plot of Problem 1

Insights :

- Their is no such relationship between the given features in the dataset.

Checking for Outliers in the dataset -To check for outliers, we will be plotting the box plots..

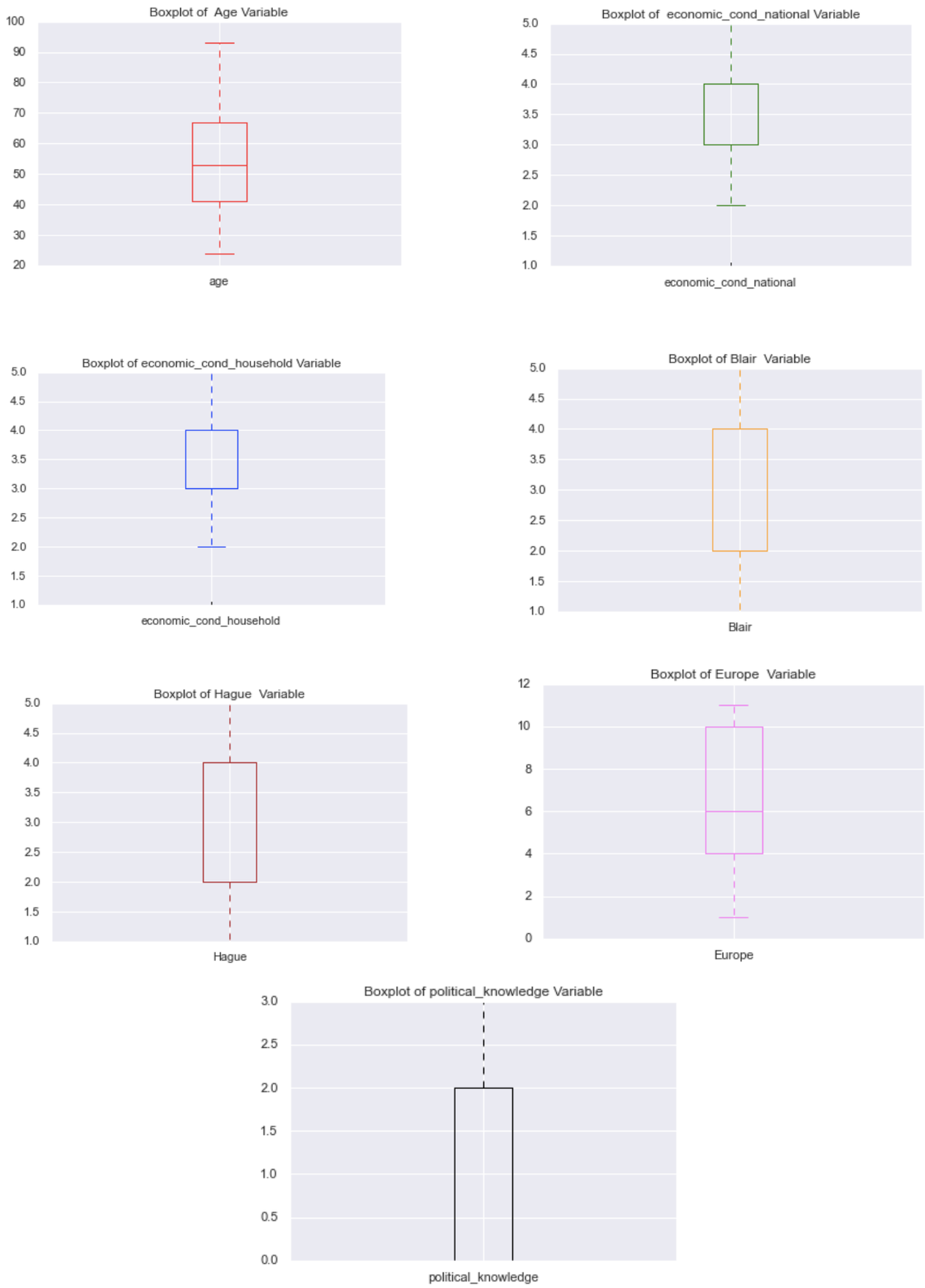

FIG: 23 Outlier Detection of Problem 1

Insights:

- From the above plotted boxplots we clearly infer that their is no outliers present in the dataset.

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).

In this problem we are going to use **Label Encoding** refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.In the given problem we are doing label encoding of the vote and gender variable for machine learning model as rest all other variables are already in numeric form.

S.No.	vote	Encode
1	Labour	1
2	Conservative	0

Tab:14 Encode Table of Categorical Column - Vote

Sr.No.	Gender	Encode
1	Female	0
2	Male	1

Tab:15 Encode Table of Categorical Column - Gender

Checking the Head of the Dataset after Encoding :

	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender
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

Tab:16 Records of Dataset after Encoding

Scaling :

- As we know that scaling means that we're transforming our data so that it fits within a specific scale, like 0-100 or 0-1.
- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data preprocessing to handle highly varying magnitudes or values or units. By looking at the dataset
- Here in the given dataset we have only one continuous variable i.e. age rest variables are discrete ordinal in nature having a specific range.As we know that we perform scaling when we have more than 1 continuous variable to handle varying magnitudes or values or units.Moreover their is no outliers present in the given dataset too. So scaling in the given dataset isn't so much effective.

Proportion of 1s and 0s :

Proportion of 1 and 0	Ratio
1	0.69677
0	0.30323

Tab:17 Proportion of 1s & 0s

Extracting the target column into separate vectors for Training set and Test set :

```
X = df_1.drop("vote", axis=1)

y = df_1.pop("vote")
```

Train-Test Split for Machine Learning Model -

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets.

In the given problem, we are advised to split the training and the testing data in the ratio of (70: 30).Here we are split the data into train and test part , like x_train , x_test , train_labels & test_labels ,by using train_test_split func() from sk-learn library here , we are taking 70 % data for training and 30 % data for testing.

Dimensions of the Training and Test Data :

Train & Test Data	Rows & Columns
X_train	(1061,8)
train_labels	(1061,)
X_test	(456,8)
test_labels	(456,)

Tab:18 Checking the Dimension of the Training & Test Data

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

Logistic Regression- is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.The hypothesis of logistic regression tends it to limit the cost function between 0 and 1 .

In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function' or also known as the 'logistic function'** - In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

Building Logistic Regression Model -

```
model = LogisticRegression()
model.fit(X_train, train_labels)
```

We can create Logistic Regression model by the help of sklearn lib by import import Logistic Regression. Now fit train data into the model, check predictions on train and test data.

Predicting on Training and Test Dataset :

lr_ytrain_predict

```
array([0, 1, 1, ..., 1, 1, 1], dtype=int8)
```

lr_ytest_predict

```
array([1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0,
       0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1,
       1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
       0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1,
       1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0,
       0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0,
       1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1,
       0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1], dtype=int8)
```

Tab:19 Logistic Regression Prediction on the Train & Test Data

Getting the Predicted Probability For Train & Test Data :

Train Data -

	0	1
0	0.929415	0.070585
1	0.096428	0.903572
2	0.296155	0.703845
3	0.111361	0.888639
4	0.016040	0.983960

Test Data -

	0	1
0	0.427092	0.572908
1	0.144547	0.855453
2	0.005856	0.994144
3	0.845964	0.154036
4	0.059527	0.940473

Tab:20 Logistic Regression Predicted Probability on the Train & Test Data

Model Evaluation-Logistic Regression :

AUC and ROC for the Training Data :

```
AUC: 0.890
Text(0, 0.5, 'True Positive Rate')
```

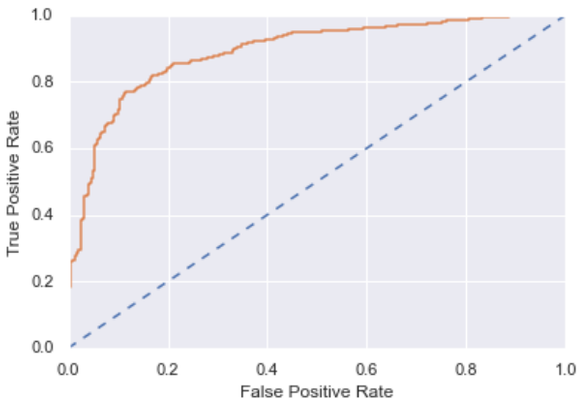

FIG: 24 AUC-ROC Curve Logistic Regression Model (Train Data)

Confusion Matrix for the Training Data :

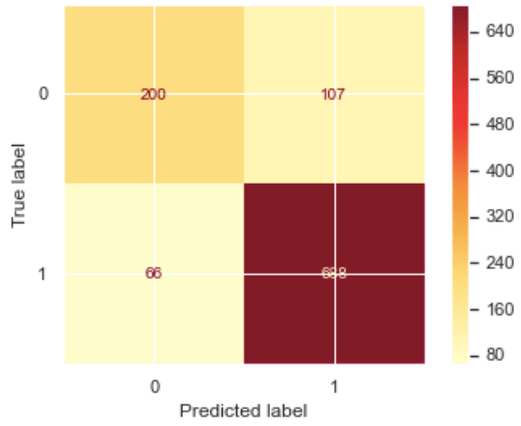

FIG: 25 Confusion Matrix Plot Logistic Regression Model (Train Data)

Train Data Accuracy----------0.8369462770970783

Classification Report of Training Data :

	precision	recall	f1-score	support
0	0.75	0.65	0.70	307
1	0.87	0.91	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

Tab:21 Logistic Regression Classification Report Train Data

AUC and ROC for the Test Data :

```
AUC: 0.877
[<matplotlib.lines.Line2D at 0x7fd6052cc520>]
```

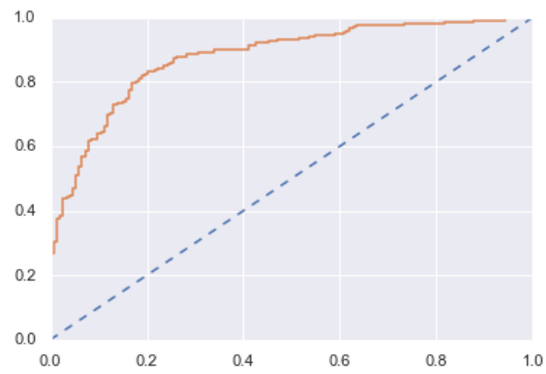

FIG: 26 AUC-ROC Curve Logistic Regression Model (Test Data)

Confusion Matrix for Test Data

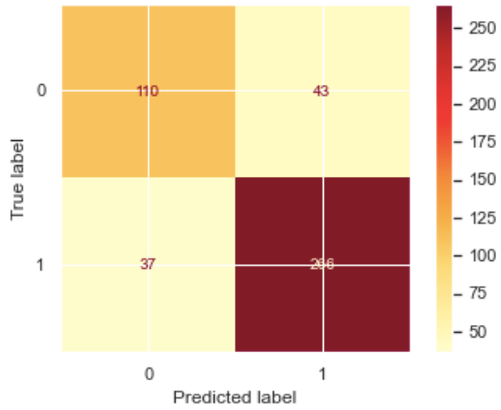

FIG: 27 Confusion Matrix Plot Logistic Regression Model (Test Data)

Test Data Accuracy----------0.8245614035087719

Classifiaction Report of Test Data :

	precision	recall	f1-score	support
0	0.75	0.72	0.73	153
1	0.86	0.88	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

Tab:22 Logistic Regression Classification Report Test Data

Conclusion Logistic Regression Model :

Train Data Class 0:

- AUC: 89%
- Accuracy: 84%
- Precision: 75%
- Recall: 65%
- f1-Score: 70%

Test Data Class 0:

- AUC: 87.7%
- Accuracy: 82%
- Precision: 75%
- Recall: 72%
- f1-Score: 73%

Train Data Class 1:

- AUC: 89%
- Accuracy: 84%
- Precision: 87%
- Recall: 91%
- f1-Score: 89%

Test Data Class 1:

- AUC: 87.7%
- Accuracy: 82%
- Precision: 86%
- Recall: 88%
- f1-Score: 87%

- On comparing the Train & Test results of the Logistics Regression Model , we conclude their is no problem of **underfitting or overfitting** of the model.
- As **accuracy , precision ,recall & f1 score** are quite similar for train & test. Hence model is good to predict the results.

Linear Discriminant Analysis, or LDA for short, is a predictive modeling algorithm for multi-class classification. It can also be used as a dimensionality reduction technique, providing a projection of a training dataset that best separates the examples by their assigned class.

Building LDA Model -

```
clf = LinearDiscriminantAnalysis()
model=clf.fit(X_train,train_labels)
```

We can create LDA model by the help of sklearn lib by import import Linear Discriminant Analysis. Now fit train data into the model, check predictions on train and test data.

Predicting on Training and Test Dataset :

lda_ytrain_predict

```
array([0, 1, 1, ..., 1, 1, 1], dtype=int8)
```

lda_ytest_predict

```
array([1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0,
       0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1,
       1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,
       0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1,
       1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0,
       0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0,
       1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1,
       0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1], dtype=int8)
```

Tab:23 LDA Prediction on the Train & Test Data

Getting the Predicted Probability :

Train Data -

	0	1
0	0.949216	0.050784
1	0.078241	0.921759
2	0.307389	0.692611
3	0.078963	0.921037
4	0.012161	0.987839

Test Data -

	0	1
0	0.462093	0.537907
1	0.133955	0.866045
2	0.006414	0.993586
3	0.861210	0.138790
4	0.056545	0.943455

Tab:24 LDA Predicted Probability on the Train & Test Data

Model Evaluation - LDA Model :

AUC and ROC for the Training Data :

```
AUC: 0.889
Text(0, 0.5, 'True Positive Rate')
```

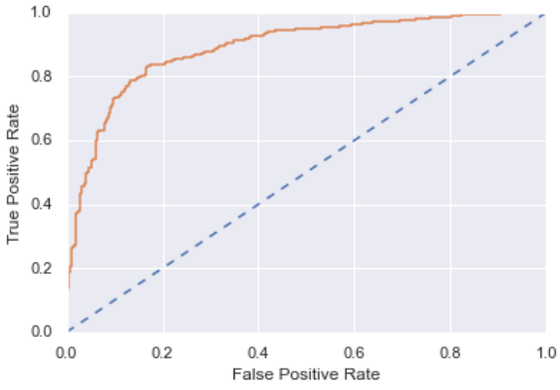

FIG: 28 AUC-ROC Curve LDA Model (Train Data)

Confusion Matrix for the Training Data :

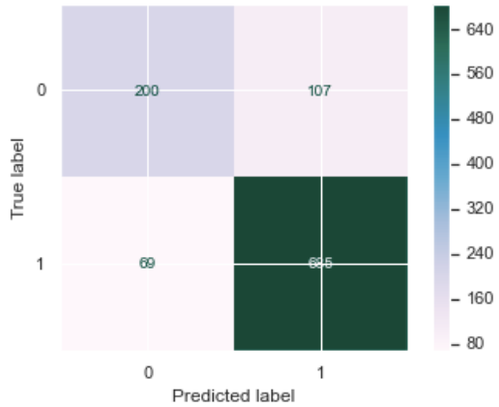

FIG: 29 Confusion Matrix Plot LDA Model (Train Data)

Train Data Accuracy----------0.8341187558906692

Classification Report of Train Data :

	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

Tab:25LDA Classification Report Train Data

AUC and ROC for the Test Data :

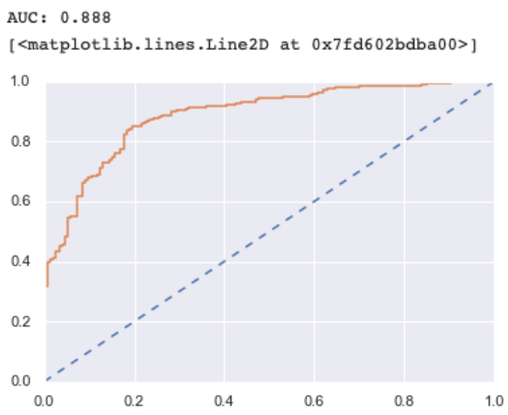

FIG: 30 AUC-ROC Curve LDA Model (Test Data)

Confusion Matrix for Test Data :

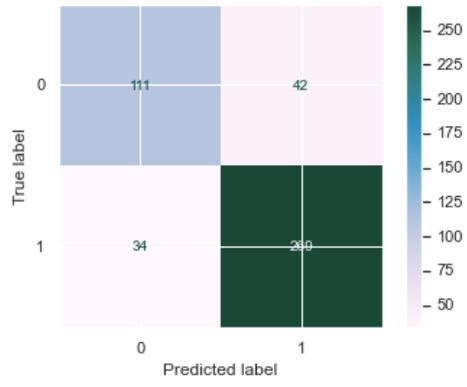

FIG: 31 Confusion Matrix Plot LDA Model (TestData)

Test Data Accuracy----------0.8333333333333334

Classifiaction Report of Test Data :

	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

Tab:26 LDA Classification Report Test Data

Conclusion LDA Model :

Train Data Class 0:

- AUC: 88.9%
- Accuracy: 83%
- Precision: 74%
- Recall: 65%
- f1-Score: 69%

Train Data Class 1:

- AUC: 88.9%
- Accuracy: 83%
- Precision: 86%
- Recall: 91%
- f1-Score: 89%

Test Data Class 0:

- AUC: 88.8%
- Accuracy: 83%
- Precision: 77%
- Recall: 73%
- f1-Score: 74%

Test Data Class 1:

- AUC: 88.8%
- Accuracy: 83%
- Precision: 86%
- Recall: 89%
- f1-Score: 88%

- On comparing the Train & Test results of the LDA Model , we conclude their is no problem of **underfitting or overfitting** of the model.
- As **accuracy , precision ,recall & f1 score** are quite similar for train & test. Hence model is good to predict the results.

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

The abbreviation KNN stands for **"K-Nearest Neighbour"**. It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'**.By default KNN Classifier( ) takes k =5.**

Building KNN Model -

```
KNN_model=KNeighborsClassifier()
KNN_model.fit(X_train,train_labels)
```

We can create KNN model by the help of sklearn lib by import import KNeighborsClassifier. Now fit train data into the model, check predictions on train and test data.

Predicting the Training and Testing data:

knn_ytrain_predict

```
array([0, 1, 1, ..., 1, 1, 1], dtype=int8)
```

knn_ytest_predict

```
array([0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1,
       0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1,
       1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0,
       0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0,
       1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1,
       1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int8)
```

Tab:27 KNN Prediction on the Train & Test Data

Getting the Predicted Probability :

Train Data -

	0	1
0	0.8	0.2
1	0.0	1.0
2	0.4	0.6
3	0.2	0.8
4	0.0	1.0

Test Data -

	0	1
0	0.6	0.4
1	0.4	0.6
2	0.2	0.8
3	0.4	0.6
4	0.0	1.0

Tab:28 KNN Predicted Probability on the Train & Test Data

Model Evaluation - KNN Model :

AUC and ROC for the Training Data :

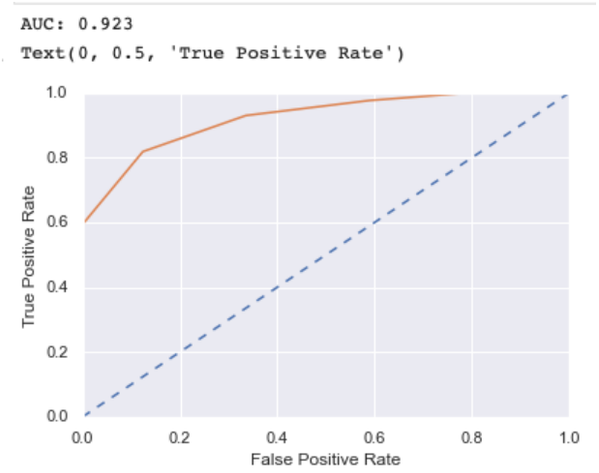

FIG: 32 AUC-ROC Curve KNN Model (Train Data)

Confusion Matrix for the Training Data :

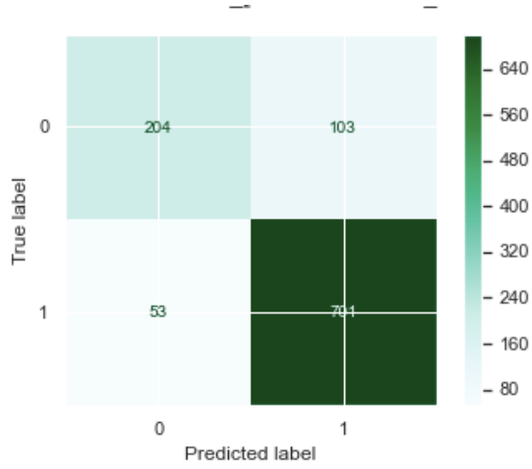

FIG: 33 Confusion Matrix Plot KNN Model (Train Data)

Train Data Accuracy----------0.8529688972667295

Classification Report of Train Data :

	precision	recall	f1-score	support
0	0.79	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

Tab:29 KNN Classification Report Train Data

AUC and ROC for the Test Data :

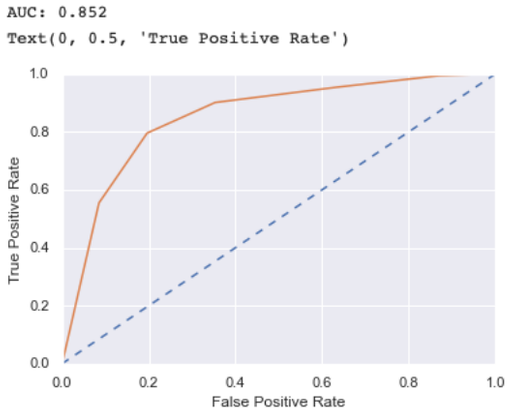

FIG: 34 AUC-ROC Curve KNN Model (Test Data)

Confusion Matrix for Test Data :

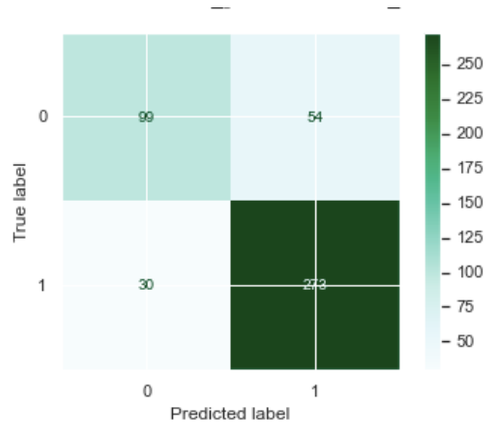

FIG: 35 Confusion Matrix Plot KNN Model (Test Data)

Test Data Accuracy----------0.8157894736842105

Classifiaction Report of Test Data :

	precision	recall	f1-score	support
0	0.77	0.65	0.70	153
1	0.83	0.90	0.87	303
accuracy			0.82	456
macro avg	0.80	0.77	0.78	456
weighted avg	0.81	0.82	0.81	456

Tab:30 KNN Classification Report Test Data

Conclusion KNN Model :

Train Data Class 0:

- AUC: 92.3%
- Accuracy: 85%
- Precision: 79%
- Recall: 66%
- f1-Score: 72%

Test Data Class 0:

- AUC: 85.2%
- Accuracy: 82%
- Precision: 77%
- Recall: 65%
- f1-Score: 70%

Train Data Class 1:

- AUC: 92.3%
- Accuracy: 85%
- Precision: 87%
- Recall: 93%
- f1-Score: 90%

Test Data Class 1:

- AUC: 85.2%
- Accuracy: 82%
- Precision: 83%
- Recall: 90%
- f1-Score: 87%

- On comparing the Train & Test results of the KNN Model , we conclude their is no problem of **underfitting or overfitting** of the model.
- As **accuracy , precision ,recall & f1 score** are quite similar for train & test. Hence model is good to predict the results.

- **Naïve Bayes algorithm** is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. ... Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class.
- Naive Bayes is a classification algorithm that is suitable for binary and multiclass classification. It is a supervised classification technique used to classify future objects by assigning class labels to instances/records using conditional probability.

Building Navie Bayes Model -

```
NB_model = GaussianNB()
NB_model.fit(X_train,train_labels)
```

We can create Navie Bayes model by the help of sklearn lib by import import GaussianNB.Now fit train data into the model, check predictions on train and test data.

Predicting the Training and Testing Data :

nb_ytrain_predict

```
array([0, 1, 1, ..., 1, 1, 1], dtype=int8)
```

nb_ytest_predict

```
array([0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0,
       0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0,
       0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0,
       1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,
       0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
       0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1,
       1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0,
       0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0,
       1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1,
       0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1], dtype=int8)
```

Tab:31 Gaussian Navie Bayes Prediction on the Train & Test Data

Getting the Predicted Probability :

Train Data -

	0	1
0	0.984678	0.015322
1	0.065437	0.934563
2	0.271735	0.728265
3	0.080026	0.919974
4	0.007648	0.992352

Test Data -

	0	1
0	0.536792	0.463208
1	0.120285	0.879715
2	0.000332	0.999668
3	0.945240	0.054760
4	0.039267	0.960733

Tab:32 Gaussian Navie Bayes Predicted Probability on the Train & Test Data

Model Evaluation - Gaussian Naive Bayes Model :

AUC and ROC for the Training Data :

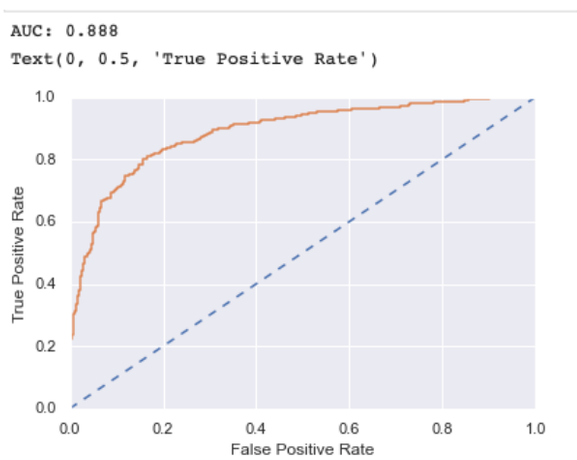

FIG: 36 AUC-ROC Curve Gaussian Naive Bayes Model (Train Data)

Confusion Matrix for the Training Data

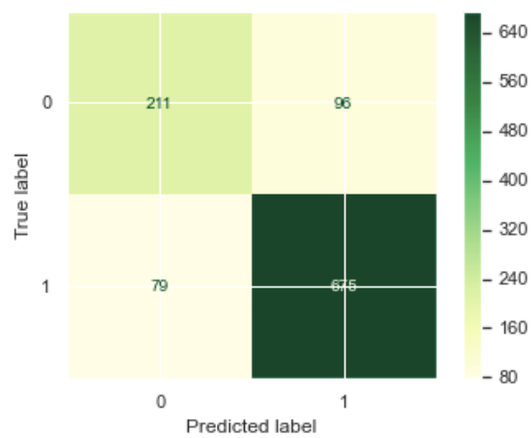

FIG: 37 Confusion Matrix Plot Gaussian Naive Bayes Model (Train Data)

Train Data Accuracy----------0.8350612629594723

Classification Report of Train Data :

	precision	recall	f1-score	support
0	0.73	0.69	0.71	307
1	0.88	0.90	0.89	754
accuracy			0.84	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.84	0.83	1061

Tab:33 Gaussian Navie Bayes Classification Report Train Data

AUC and ROC for the Test Data :

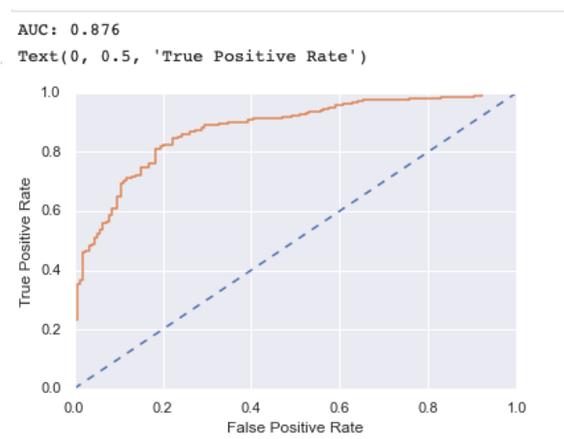

FIG: 38 AUC-ROC Curve Gaussian Naive Bayes Model (Test Data)

Confusion Matrix for the Test Data :

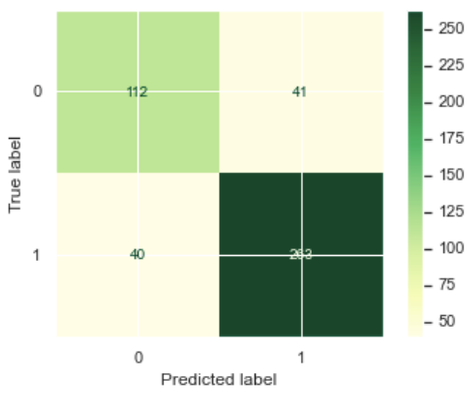

FIG: 39 Confusion Matrix Plot Gaussian Naive Bayes Model (Test Data)

Test Data Accuracy----------0.8223684210526315

Classification Report of Test Data :

	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

Tab:34 Gaussian Navie Bayes Classification Report Test Data

Conclusion Gaussian Naive Bayes Model :

Train Data Class 0:

- AUC: 88.8%
- Accuracy: 84%
- Precision: 73%
- Recall: 69%
- f1-Score: 71%

Train Data Class 1:

- AUC: 88.8%
- Accuracy: 84%
- Precision: 88%
- Recall: 90%
- f1-Score: 89%

Test Data Class 0:

- AUC: 87.6%
- Accuracy: 82%
- Precision: 74%
- Recall: 73%
- f1-Score: 73%

Test Data Class 1:

- AUC: 87.6%
- Accuracy: 82%
- Precision: 87%
- Recall: 87%
- f1-Score: 87%

- On comparing the Train & Test results of the Navie Bayes Model , we conclude their is no problem of **underfitting or overfitting** of the model.
- As **accuracy , precision ,recall & f1 score** are quite similar for train & test. Hence model is good to predict the results.

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

In this question we are building various models like Random Forest ,Logistic Regression ,Random Forest with Bagging , Boosting Models like Ada Boost & Gradient Boost and moreover here we also tunned the model which we built above. 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 & will compare in 1.7 down the line.

- A **Random forest** is a **machine learning technique** that's used to solve regression and classification problems. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees.
- It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Building Random Forest Base Model -

```
rfcl = RandomForestClassifier()
rfcl = rfcl.fit(X_train, train_labels)
```

We can create Random Forest Base Model by the help of sklearn lib by import RandomForestClassifier. Now fit train data into the model, check predictions on train and test data.

Predicting on Training and Test dataset

rf_ytrain_predict

```
array([0, 1, 0, ..., 1, 1, 1], dtype=int8)
```

rf_ytest_predict

```
array([0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1,
       0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
       1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0,
       0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0,
       1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1], dtype=int8)
```

Tab:35 Random Forest Base Model Prediction on the Train & Test Data

Getting the Predicted Probability :

Train Data -

	0	1
0	0.72	0.28
1	0.31	0.69
2	0.01	0.99
3	0.79	0.21
4	0.09	0.91

Test Data -

	0	1
0	0.72	0.28
1	0.31	0.69
2	0.01	0.99
3	0.79	0.21
4	0.09	0.91

Tab:36 Random Forest Base Model Predicted Probability on the Train & Test Data

Model Evaluation - Random Forest Base Model

AUC and ROC for the Training Data

```
AUC: 1.000
Text(0, 0.5, 'True Positive Rate')
```

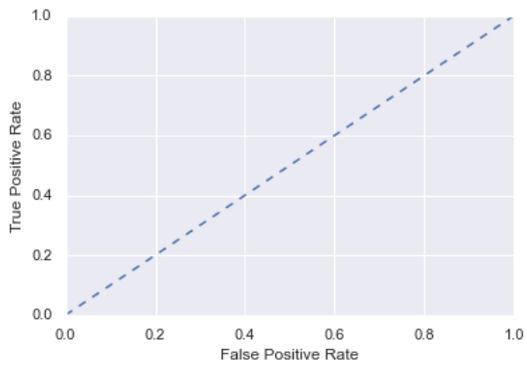

FIG: 40 AUC-ROC Curve Random Forest Base Model (Train Data)

Confusion Matrix for the Training Data

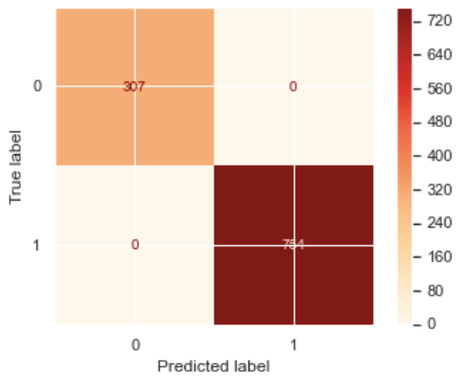

FIG: 41 Confusion Matrix Plot Random Forest Base Model (Train Data)

Train Data Accuracy----------1.0

Classification Report of Training Data

	precision	recall	f1-score	support
0	1.00	1.00	1.00	307
1	1.00	1.00	1.00	754
accuracy			1.00	1061
macro avg	1.00	1.00	1.00	1061
weighted avg	1.00	1.00	1.00	1061

Tab:37 Random Forest Base Model Classification Report Train Data

AUC and ROC for the Test Data

```
AUC: 0.892
Text(0, 0.5, 'True Positive Rate')
```

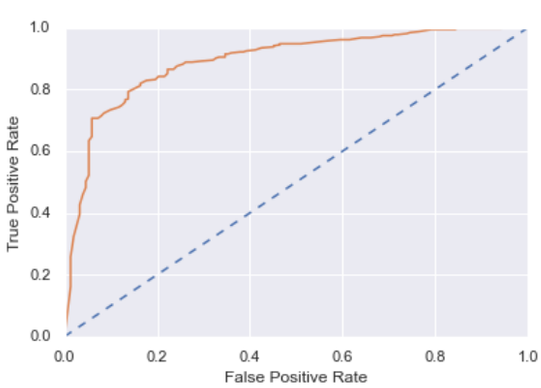

FIG: 42 AUC-ROC Curve Random Forest Base Model (Test Data)

Confusion Matrix for Test Data

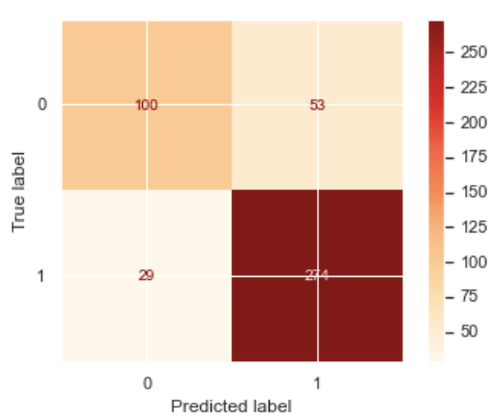

FIG: 43 Confusion Matrix Plot Random Forest Base Model (Test Data)

Test Data Accuracy----------0.8201754385964912

Classifiaction Report of Test Data :

	precision	recall	f1-score	support
0	0.78	0.65	0.71	153
1	0.84	0.90	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

Tab:38 Random Forest Base Model Classification Report Test Data

Conclusion Random Forest Base Model :

Train Data Class 0:

- AUC: 100%
- Accuracy: 100%
- Precision: 100%
- Recall: 100%
- f1-Score: 100%

Train Data Class 1:

- AUC: 100%
- Accuracy: 100%
- Precision: 100%
- Recall: 100%
- f1-Score: 100%

Test Data Class 0:

- AUC: 89.2%
- Accuracy: 82%
- Precision: 78%
- Recall: 65%
- f1-Score: 71%

Test Data Class 0:

- AUC: 89.2%
- Accuracy: 82%
- Precision: 84%
- Recall: 90%
- f1-Score: 87%

Variable Importance

	Imp
age	0.213364
Europe	0.191816
Hague	0.174736
Blair	0.138352
economic_cond_national	0.092993
economic_cond_household	0.080528
political_knowledge	0.073670
gender	0.034541

Insights :

age , Europe ,Hague Blair are the important variables for predictions.

- On comparing the Train & Test results of the Random Forest Model , we conclude their is problem of **overfitting** of the model.As the difference between train and test accuracies is more than 10%, it is a overfit model.
- As **accuracy , precision ,recall & f1 score** for train data 100 % and peform less on the test data as we also know that no model have 100 % accurate in the practical world. Hence we reject this model as it is not a good model to predict results.
- Moreover as saw here in the model we have model overfit issue we can resolve this issue by applying bagging technique.So in the next model we apply bagging the same random forest and check its perfomance on train & test data.

Bagging of Random Forest - Bagging is an ensemble algorithm that fits multiple models on different subsets of a training dataset, then combines the predictions from all models. Random forest is an extension of bagging that also randomly selects subsets of features used in each data sample.

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.Lets build Random Forest with Bagging technique.

Building Bagged Random Forest Model -

```
rf_bag = RandomForestClassifier()
Bagging_model=BaggingClassifier(base_estimator=rf_bag,n_estimators=100,random_state=1)
Bagging_model.fit(X_train,train_labels)
```

We can create Bagged Random Forest Model by the help of sklearn lib by import import RandomForestClassifier & import BaggingClassifier Now fit train data into the bagging classifier model & in base_estimator insert RandomForestClassifier().Now check predictions on train and test data.

Predicting the Training and Testing data:

rf_bag_ytrain_predict

```
array([0, 1, 0, ..., 1, 1, 1], dtype=int8)
```

rf_bag_ytest_predict

```
array([0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0,
       0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
       1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0,
       0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0,
       1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
       0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1], dtype=int8)
```

Tab:39 Random Forest Bagged Model Prediction on the Train & Test Data

Getting the Predicted Probability

Train Data -

	0	1
0	0.9047	0.0953
1	0.1279	0.8721
2	0.5896	0.4104
3	0.0425	0.9575
4	0.0338	0.9662

Test Data -

	0	1
0	0.6813	0.3187
1	0.2965	0.7035
2	0.0410	0.9590
3	0.7510	0.2490
4	0.1168	0.8832

Tab:40 Random Forest Bagged Model Predicted Probability on the Train & Test Data

Model Evaluation - Bagging of Random Forest

AUC and ROC for the Training Data

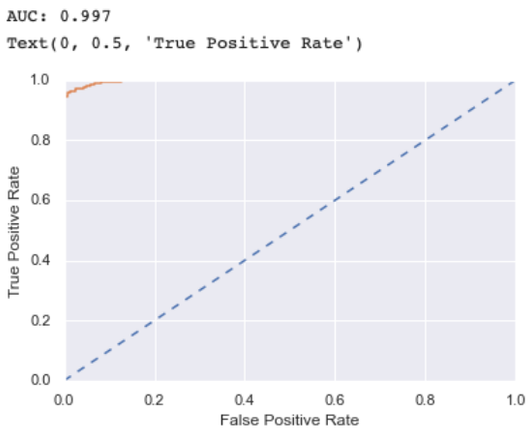

FIG: 44 AUC-ROC Curve Random Forest Bagged Model (Train Data)

Confusion Matrix for the Training Data

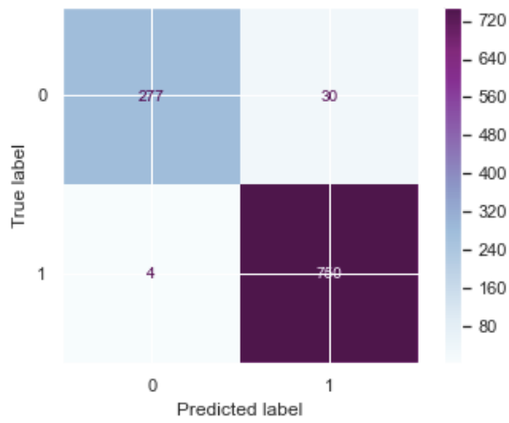

FIG: 45 Confusion Matrix Plot Random Forest Bagged Model (Train Data)

Train Data Accuracy ----------0.9679547596606974

Classification Report of Train Data

	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

Tab:41 Random Forest Bagged Model Classification Report Train Data

AUC and ROC for the Test Data

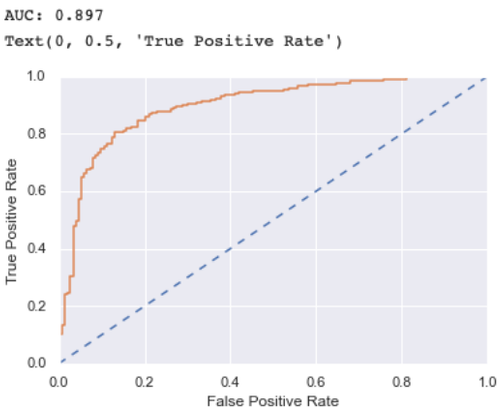

FIG: 46 AUC-ROC Curve Random Forest Bagged Model (Test Data)

Confusion Matrix for Test Data

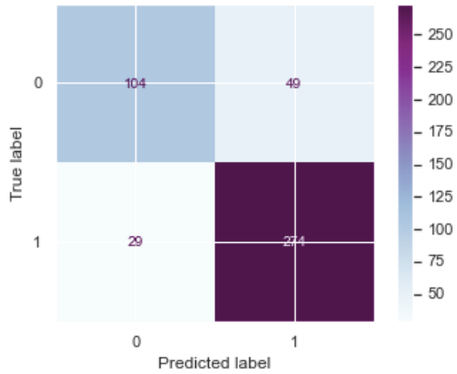

FIG: 47 Confusion Matrix Plot Random Forest Bagged Model (Test Data)

Test Data Accuracy----------0.8289473684210527

Classifiaction Report of Test Data

	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

Tab:42 Random Forest Bagged Model Classification Report Test Data

Conclusion Bagging of Random Forest

Train Data Class 0:

- AUC: 99.7%
- Accuracy: 97%
- Precision: 99%
- Recall: 90%
- f1-Score: 94%

Test Data Class 0:

- AUC: 89.7%
- Accuracy: 83%
- Precision: 78%
- Recall: 68%
- f1-Score: 73%

Train Data Class 1:

- AUC: 99.7%
- Accuracy: 97%
- Precision: 96%
- Recall: 99%
- f1-Score: 98%

Test Data Class 1:

- AUC: 89.7%
- Accuracy: 83%
- Precision: 85%
- Recall: 90%
- f1-Score: 88%

Variable Importance

	Imp
age	0.213364
Europe	0.191816
Hague	0.174736
Blair	0.138352
economic_cond_national	0.092993
economic_cond_household	0.080528
political_knowledge	0.073670
gender	0.034541

Insights :
age , Europe ,Hague Blair are the important variables for predictions.

- On comparing the Train & Test results of the Bagged Random Forest Model , we conclude their is also problem of **overfitting** of the model.As the difference between train and test accuracies is more than 10% model perform poor on test data as compared to the train data , it is a overfit model too.
- Hence we reject this model also as it is not a good model to predict results.

- **AdaBoost algorithm,** short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

- AdaBoost (Adaptive Boosting) : It works on similar method as discussed above. It fits a sequence of weak learners on different weighted training data. If prediction is incorrect using the first learner, then it gives higher weight to observation which have been predicted incorrectly.

Building Ada Boost Model :

```
ADB_model = AdaBoostClassifier(n_estimators=100,random_state=1)
ADB_model.fit(X_train,train_labels)
```

We can create Ada Boost Model by the help of sklearn lib by import AdaBoostClassifier.Now fit train data into the model, check predictions on train and test data.

Predicting the Training and Testing data

ADB_ytrain_predict

```
array([0, 1, 1, ..., 1, 1, 1], dtype=int8)
```

ADB_ytest_predict

```
array([0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0,
       0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
       0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1,
       1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0,
       0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int8)
```

Tab:43 Ada Boost Model Prediction on the Train & Test Data

Getting the Predicted Probability

Train Data -

	0	1
0	0.501152	0.498848
1	0.493238	0.506762
2	0.497106	0.502894
3	0.492405	0.507595
4	0.491090	0.508910

Test Data -

	0	1
0	0.502173	0.497827
1	0.496755	0.503245
2	0.480057	0.519943
3	0.505525	0.494475
4	0.493868	0.506132

Tab:44 Ada Boost Model Predicted Probability on the Train & Test Data

Model Evaluation - ADA Boosting Model

AUC and ROC for the Training Data

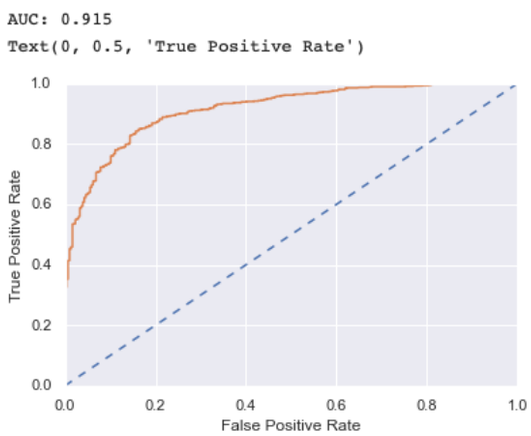

FIG: 48 AUC-ROC Curve Ada Boost Model (Train Data)

Confusion Matrix for the Training Data

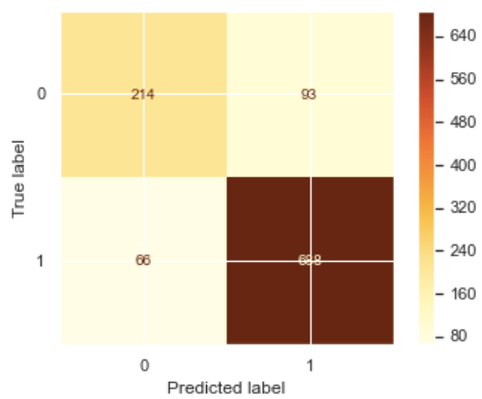

FIG: 49 Confusion Matrix Plot Ada Boost Model (Train Data)

Train Data Accuracy ----------0.8501413760603205

Classification Report of Train Data

	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

Tab:45 Ada Boost Model Classification Report Train Data

AUC and ROC for the Test Data

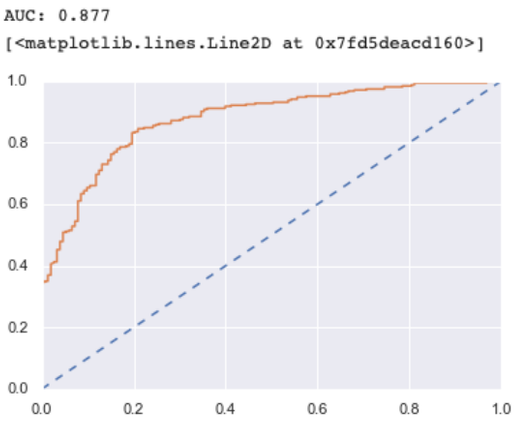

FIG: 50 AUC-ROC Curve Ada Boost Model (Test Data)

Confusion Matrix for Test Data

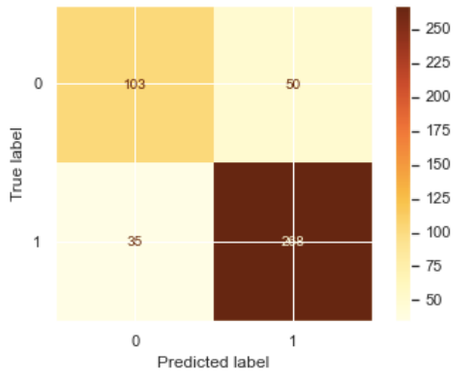

FIG: 51 Confusion Matrix Plot Ada Boost Model (Test Data)

Test Data Accuracy----------0.8135964912280702

Classifiaction Report of Test Data

	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

Tab:46 Ada Boost Model Classification Report Test Data

Conclusion Ada Boost Model

Train Data Class 0:

- AUC: 91.5%
- Accuracy: 85%
- Precision: 76%
- Recall: 70%
- f1-Score: 73%

Test Data Class 0:

- AUC: 87.7%
- Accuracy: 81%
- Precision: 75%
- Recall: 67%
- f1-Score: 71%

Train Data Class 1:

- AUC: 91.5%
- Accuracy: 85%
- Precision: 88%
- Recall: 91%
- f1-Score: 90%

Test Data Class 1:

- AUC: 87.7%
- Accuracy: 81%
- Precision: 84%
- Recall: 88%
- f1-Score: 86%

- On comparing the Train & Test results of the Ada Boost Model , we conclude their is no problem of **underfitting or overfitting** of the model.
- As **accuracy , precision ,recall & f1 score** are quite similar for train & test. Hence model is good to predict the results.

- **Gradient Boosting Model-** Gradient boosting is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

- Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model inorder to minimize the error.

- Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

Building Gradient Boosting Model :

```
gbcl = GradientBoostingClassifier(random_state=1)
gbcl = gbcl.fit(X_train,train_labels)
```

We can create Gradient Boosting Model by the help of sklearn lib by import GradientBoostingClassifier.Now fit train data into the model, check predictions on train and test data.

Predicting the Training and Testing data

gbcl_ytrain_predict

```
array([0, 1, 1, ..., 1, 1, 1], dtype=int8)
```

gbcl_ytest_predict

```
array([0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0,
       0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
       0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0,
       0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1,
       0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1], dtype=int8)
```

Tab:47 Gradient Boosting Model Prediction on the Train & Test Data

Getting the Predicted Probability

Train Data -

	0	1
0	0.833670	0.166330
1	0.071905	0.928095
2	0.335574	0.664426
3	0.049894	0.950106
4	0.031041	0.968959

Test Data -

	0	1
0	0.690657	0.309343
1	0.236942	0.763058
2	0.001102	0.998898
3	0.840247	0.159753
4	0.111644	0.888356

Tab:48 Gradient Boosting Model Predicted Probability on the Train & Test Data

Model Evaluation - Gradient Boosting Model

AUC and ROC for the Training Data

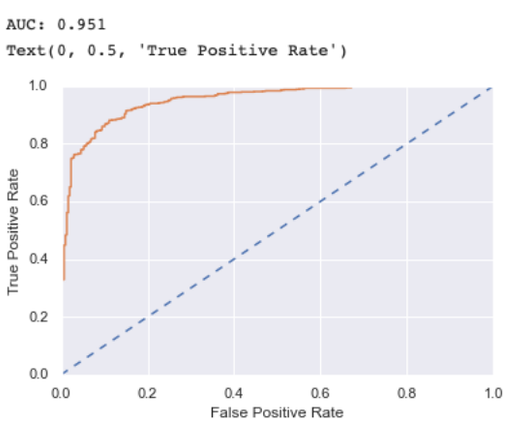

FIG: 52 AUC-ROC Curve Gradient Boosting Model (Train Data)

Confusion Matrix for the Training Data

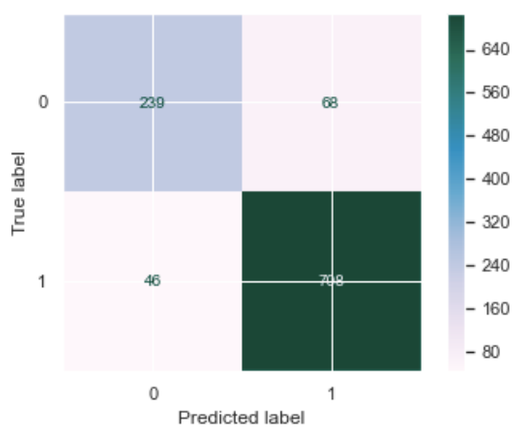

FIG: 53 Confusion Matrix Plot Gradient Boosting Model (Train Data)

Train Data Accuracy----------0.8925541941564562

Classification Report of Train Data

	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

Tab:49 Gradient Boosting Model Classification Report Train Data

AUC and ROC for the Test Data

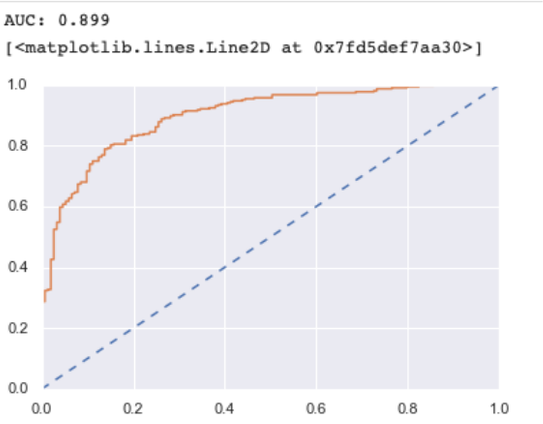

FIG: 54 AUC-ROC Curve Gradient Boosting Model (Test Data)

Confusion Matrix for Test Data

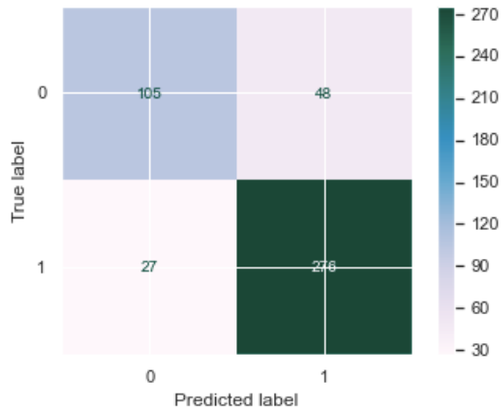

FIG: 55 Confusion Matrix Plot Gradient Boosting Model (Test Data)

Test Data Accuracy----------0.8355263157894737

Classifiaction Report of Test Data

	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

Tab:50 Gradient Boosting Model Classification Report Test Data

Conclusion Gradient Boost Model

Train Data Class 0:

- AUC: 95.1%
- Accuracy: 89%
- Precision: 84%
- Recall: 78%
- f1-Score: 81%

Test Data Class 0:

- AUC: 89.9%
- Accuracy: 84%
- Precision: 80%
- Recall: 69%
- f1-Score: 74%

Train Data Class 1:

- AUC: 95.1%
- Accuracy: 89%
- Precision: 91%
- Recall: 94%
- f1-Score: 93%

Test Data Class 1:

- AUC: 89.9%
- Accuracy: 84%
- Precision: 85%
- Recall: 91%
- f1-Score: 88%

- On comparing the Train & Test results of the Gradient Boosting Model , we conclude their is no problem of **underfitting or overfitting** of the model.
- As **accuracy , precision ,recall & f1 score** are quite good for train & test and balanced. Hence model is good to predict the results.

Tunned Logistic Regression Model

In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function' or also known as the 'logistic function' -** In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

Grid Search for finding out the optimal values for the hyper parameters-

Grid search builds a model for every combination of hyperparameters specified and evaluates each model. A more efficient technique for hyperparameter tuning is the Randomized search — where random combinations of the hyperparameters are used to find the best solution.
As per the industries standards we are taking various hyper parameters to build our logistic regression ,hyperparametrs are listed below :

penalty :['l2','l1','none']
solver':['sag','lbfgs','newton-cg','saga']
tol':[0.0001,0.00001]
n_jobs=-1,

The reason behind using the hyper parameters (Machine learning algorithms have hyperparameters that allow you to tailor the behavior of the algorithm to your specific dataset. Hyperparameters are different from parameters, which are the internal coefficients or weights for a model found by the learning algorithm) is to increase the performance our model as **penalty ,solver & tol** are the hyper parameters for the model tuning.we are taking penalty l2,l1 & none as these are most general penalty for a logistic model.

- **L1 regularization** adds an L1 penalty equal to the absolute value of the magnitude of coefficients. In other words, it limits the size of the coefficients. L1 can yield sparse models (i.e. models with few coefficients); Some coefficients can become zero and eliminated. Lasso regression uses this method.
- **L2 regularization** adds an L2 penalty equal to the square of the magnitude of coefficients. L2 will not yield sparse models and all coefficients are shrunk by the same factor (none are eliminated). Ridge regression and SVMs use this method.

we are taking various solver here like 'sag','lbfgs','newton-cg','saga' ,reson behind using these solver as solver the solver uses a Coordinate Descent (CD) algorithm that solves optimization problems by successively performing approximate minimization along coordinate directions or coordinate hyperplanes.

- newton-cg: Solver which calculates Hessian explicitly which can be computationally expensive in high dimensions.
- sag: Stands for Stochastic Average Gradient Descent. More efficient solver with large datasets.
- The SAGA solver is a variant of SAG that also supports the non-smooth penalty L1 option (i.e. L1 Regularization). This is therefore the solver of choice for sparse multinomial logistic regression and it's also suitable for very Large dataset.
- lbfgs — Stands for Limited-memory Broyden–Fletcher–Goldfarb–Shanno. It approximates the second derivative matrix updates with gradient evaluations. It stores only the last few updates, so it saves memory. It isn't super fast with large data sets. It will be the default solver as of Scikit-learn version 0.22.
- n_jobs=-1, we use this to use all the cores of the system for faster execution.

The best estimator for building our model are obtained by using the grid search cv function() are tabluate below:

penalty :['none']
solver':['sag']
tol':[0.0001]
n_jobs=-1,

Result :

From grid search we get **penalty = 'none'** , **solver = 'sag' , tol = 0.0001** , we are going to use these parameter to build our logistic regression model. We can create logistic regression model by the help of sklearn lib by import LogisticRegression. Now fit these parameters into the model with train data , check predictions on train and test data.

Predicting on Training and Test dataset

lr_tunned_ytrain_predict

```
array([0, 1, 1, ..., 1, 1, 1], dtype=int8)
```

lr_tunned_ytest_predict

```
array([1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0,
       0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1,
       1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1,
       0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1,
       1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1,
       1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0,
       0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0,
       1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1,
       0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1], dtype=int8)
```

Tab:51 Tuned Logistic Regression Model Prediction on the Train & Test Data

Getting the Predicted Probability

Train Data -

	0	1
0	0.938636	0.061364
1	0.098090	0.901910
2	0.272524	0.727476
3	0.115593	0.884407
4	0.016852	0.983148

Test Data -

	0	1
0	0.431510	0.568490
1	0.158065	0.841935
2	0.005284	0.994716
3	0.852118	0.147882
4	0.061924	0.938076

Tab:52 Tuned Logistic Regression Model Predicted Probability on the Train & Test Data

Model Evaluation - Tunned Logistic Regression Model

AUC and ROC for the Training Data

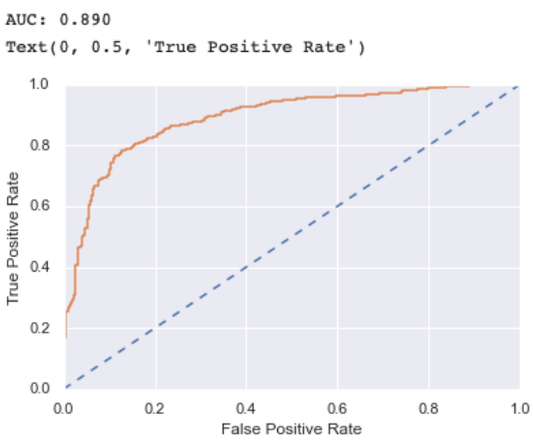

FIG: 56 AUC-ROC Curve Tuned Logistic Regression Model (Train Data)

Confusion Matrix for the Training Data

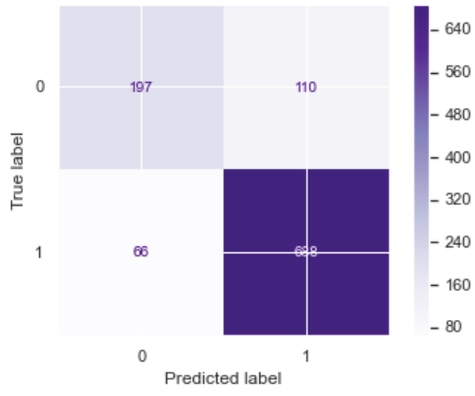

FIG: 57 Confusion Matrix Plot Tuned Logistic Regression Model (Train Data)

Train Data Accuracy----------0.8341187558906692

Classification Report of Training Data

	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

Tab:53 Tuned Logistic Regression Model Classification Report Train Data

AUC and ROC for the Test Data

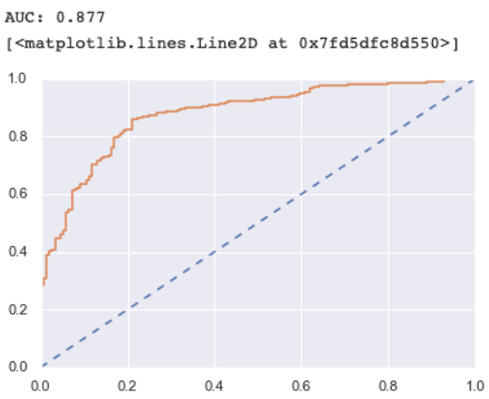

FIG: 58 AUC-ROC Curve Tuned Logistic Regression Model (Test Data)

Confusion Matrix for Test Data

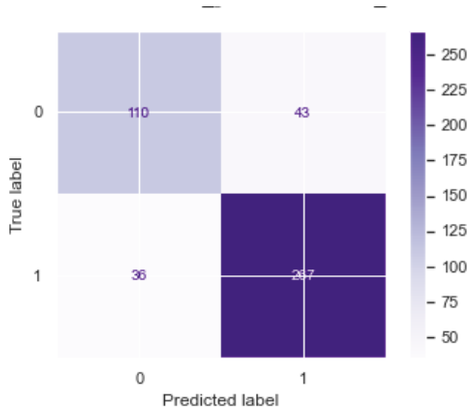

FIG: 59 Confusion Matrix Plot Tuned Logistic Regression Model (Test Data)

Test Data Accuracy----------0.8267543859649122

Classifiaction Report of Test Data

	precision	recall	f1-score	support
0	0.75	0.72	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.80	456
weighted avg	0.83	0.83	0.83	456

Tab:54 Tuned Logistic Regression Model Classification Report Test Data

Conclusion Tunned Logistic Regression Model

Train Data Class 0:

- AUC: 89%
- Accuracy: 83%
- Precision: 75%
- Recall: 64%
- f1-Score: 69%

Test Data Class 0:

- AUC: 87.7%
- Accuracy: 83%
- Precision: 75%
- Recall: 72%
- f1-Score: 74%

Train Data Class 1:

- AUC: 89%
- Accuracy: 83%
- Precision: 86%
- Recall: 91%
- f1-Score: 89%

Test Data Class 1:

- AUC: 87.7%
- Accuracy: 83%
- Precision: 86%
- Recall: 88%
- f1-Score: 87%

- On comparing the Train & Test results of the Tunned Logistic Regression Model , we conclude their is no problem of **underfitting or overfitting** of the model.
- As **accuracy , precision ,recall & f1 score** are quite similar for train & test and balanced. Hence model is good to predict the results.

LDA Model Tuning with Custom Cutoff Tunning

Change the cut-off values for maximum accuracy for LDA Model :

We will do this exercise only on the training data.By doing this we are changing the default cutoff of 0.5 and try to check the model on various cutoff from 0.1 to 1, and find on which cutoff we get the best result of accuracy , precision , recall and f1 score and will that cut-off and make predictions on that cutoff to improve our performance.

Train Data :

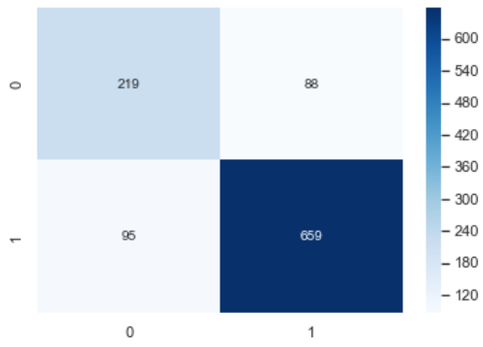

FIG: 60 Confusion Matrix Plot Tuned LDA Custom Cutoff Model(Train Data)

Classification Report of the custom cut-off train data:

	precision	recall	f1-score	support
0	0.70	0.71	0.71	307
1	0.88	0.87	0.88	754
accuracy			0.83	1061
macro avg	0.79	0.79	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Tab:55 LDA Model Custom Cutoff Model Classification Report Train Data

At cutoff 0.6 we get the better value for accuracy score , F1 score , recall & Precision.

Test Data :

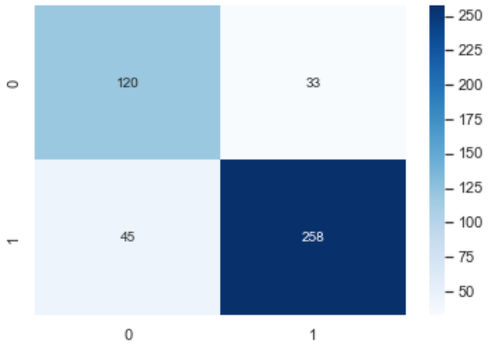

FIG: 61 Confusion Matrix Plot Tuned LDA Custom Cutoff Model(Test Data)

Classification Report of the custom cut-off test data:

	precision	recall	f1-score	support
0	0.73	0.78	0.75	153
1	0.89	0.85	0.87	303
accuracy			0.83	456
macro avg	0.81	0.82	0.81	456
weighted avg	0.83	0.83	0.83	456

Tab:56 LDA Model Custom Cutoff Model Classification Report Test Data

Conclusion :

We can take 0.6 cutoff for slightly better performance of the LDA model.Now we have balanced & more reliable scores, will get better results.

Tunned KNN Model

Run the KNN with no of neighbours to be 1,3,5..19 and *Find the optimal number of neighbours from K=1,3,5,7....19 using the Mis classification error Hint: Misclassification error (MCE) = 1 - Test accuracy score. Calculated MCE for each model with neighbours = 1,3,5...19 and find the model with lowest MCE.

Plot misclassification error vs k (with k value on X-axis) using matplotli

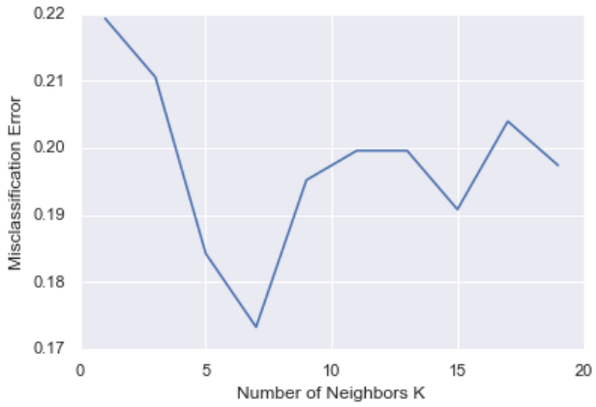

FIG: 62 Misclassification Error VS K-Value Plot

For K = 7 it is giving the least misclassification error lets build the model for k=7 & check the evaluation metrics.As per industry standard we always take value in odd number because each instance in essence votes for their class and the class with the most votes is taken as the prediction. If you are using K and you have an even number of classes (e.g. 2) it is a good idea to choose a K value with an odd number to avoid a tie.

Building KNN Model -

```
KNN_model=KNeighborsClassifier(n_neighbors=7)
KNN_model.fit(X_train,train_labels)
```

We can create KNN model by the help of sklearn lib by import import KNeighborsClassifier. Now fit train data into the model & set n_neighbors =7. check predictions on train and test data.

Predicting the Training and Testing data

KNN_model_tunned_ytrain_predict

```
array([0, 1, 1, ..., 1, 1, 1], dtype=int8)
```

KNN_model_tunned_ytest_predict

```
array([0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
       0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1,
       1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1,
       1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0,
       0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0,
       1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1,
       1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int8)
```

Tab:57 Tuned KNN Model Prediction on the Train & Test Data

Getting the Predicted Probability

Train Data -

	0	1
0	0.714286	0.285714
1	0.142857	0.857143
2	0.285714	0.714286
3	0.285714	0.714286
4	0.000000	1.000000

Test Data -

	0	1
0	0.571429	0.428571
1	0.285714	0.714286
2	0.142857	0.857143
3	0.285714	0.714286
4	0.142857	0.857143

Tab:58 Tuned KNN Model Predicted Probability on the Train & Test Data

Model Evaluation - Tunned KNN Model

AUC and ROC for the Training Data

```
AUC: 0.917
Text(0, 0.5, 'True Positive Rate')
```

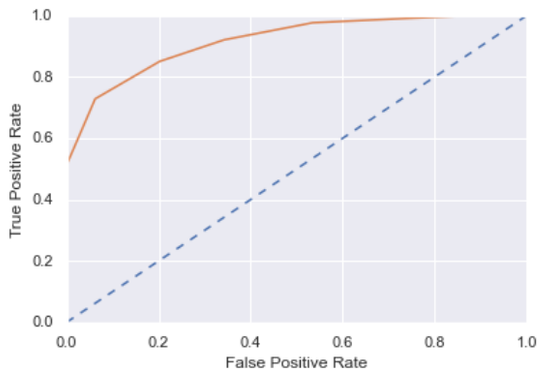

FIG: 63 AUC-ROC Curve Tuned KNN Model (Train Data)

Confusion Matrix for the Training Data

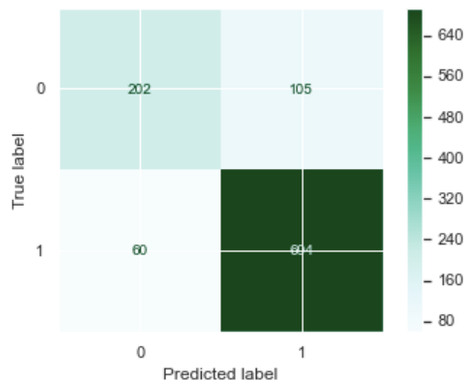

FIG: 64 Confusion Matrix Plot Tuned KNN Model (Train Data)

Train Data Accuracy----------0.8444863336475024

Classification Report of Train Data

	precision	recall	f1-score	support
0	0.77	0.66	0.71	307
1	0.87	0.92	0.89	754
accuracy			0.84	1061
macro avg	0.82	0.79	0.80	1061
weighted avg	0.84	0.84	0.84	1061

Tab:59 Tuned KNN Model Classification Report Train Data

AUC and ROC for the Test Data

```
AUC: 0.863
Text(0, 0.5, 'True Positive Rate')
```

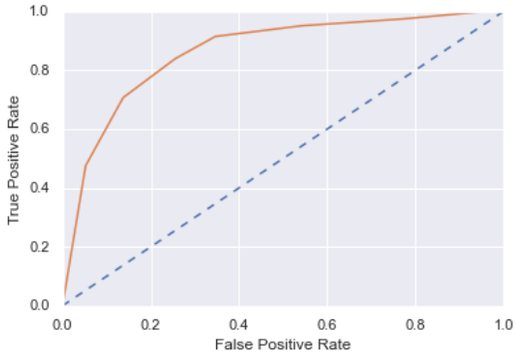

FIG: 65 AUC-ROC Curve Tuned KNN Model (Test Data)

Confusion Matrix for the Test Data

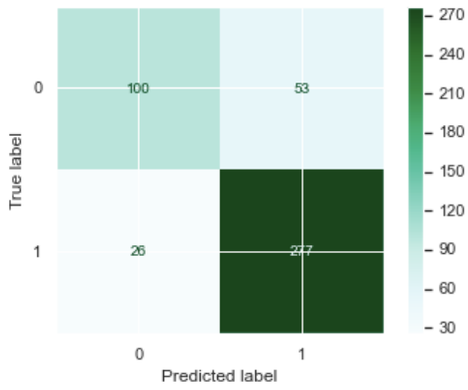

FIG: 66 Confusion Matrix Plot Tuned KNN Model (Test Data)

Test Data Accuracy----------0.8267543859649122

Classification Report of Test Data

	precision	recall	f1-score	support
0	0.79	0.65	0.72	153
1	0.84	0.91	0.88	303
accuracy			0.83	456
macro avg	0.82	0.78	0.80	456
weighted avg	0.82	0.83	0.82	456

Tab:60 Tuned KNN Model Classification Report Test Data

Conclusion Tunned KNN Model

Train Data Class 0:

- AUC: 91.7%
- Accuracy: 84%
- Precision: 77%
- Recall: 66%
- f1-Score: 71%

Test Data Class 0:

- AUC: 86.3%
- Accuracy: 83%
- Precision: 79%
- Recall: 65%
- f1-Score: 72%

Train Data Class 1:

- AUC: 91.7%
- Accuracy: 84%
- Precision: 87%
- Recall: 92%
- f1-Score: 89%

Test Data Class 1:

- AUC: 86.3%
- Accuracy: 83%
- Precision: 84%
- Recall: 91%
- f1-Score: 88%

- On comparing the Train & Test results of the KNN Model , we conclude their is no problem of **underfitting or overfitting** of the model.
- As **accuracy , precision ,recall & f1 score** are quite similar for train & test and balanced. Hence model is good to predict the results.

Cross Validation on Naive Bayes Model

Cross Validation on Train Data

```
array([0.80373832, 0.78301887, 0.8490566 , 0.83962264, 0.90566038,
       0.8490566 , 0.78301887, 0.83962264, 0.81132075, 0.82075472])
```

Cross Validation on Test Data

```
array([0.7826087 , 0.80434783, 0.86956522, 0.80434783, 0.86956522,
       0.86956522, 0.88888889, 0.82222222, 0.75555556, 0.82222222])
```

Tab:61 Cross Validation Score of Navie Bayes Model for Train & Test Data for Model Validation.

Conclusion Cross Validation on Naive Bayes Model:

After 10 fold cross validation, scores both on train and test data set respectively for all 10 folds are almost same.Hence our model is valid.

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

All the models are built above with their train & test accuracy , AUC AND ROC curve and score and Confusion Matrix and classification report. Here we will compare all the models on basis of their performance,AUC/ROC Score , Accuracy , Precision ,Recall for train & test data with tabular representation below & decide the final model for the deployment.In the given problem our is work to built a model 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.According the context of the of the business problem here in the given problem ,precision and recall & f1 play vital role so we will compare precision & recall f1 values for each model and select our final model for deployment.

Comparison of the Performance Metrics of All the Models on Train Data

Model Name (Train Data)	Precision		Recall		f1-score		Model Accuracy	AUC/ROC
	0(Cons)	1(Labour)	0(Cons)	1(Labour)	0(Cons)	1(Labour)		
Logistic Regression	75	87	65	91	70	89	84	89
LDA	74	86	65	91	69	89	83	88.9
KNN	79	87	66	93	72	90	85	92.3
Naïve Baise	73	88	69	90	71	89	84	88.8
Random Forest Base Model	100	100	100	100	100	100	100	100
Bagging with Random Forest	99	96	90	99	94	98	97	99.7
ADA Boosting Model	76	88	70	91	73	90	85	91.5
Gradient Boosting Model	84	91	78	94	81	93	89	95.1
Tunned Logistic Regression Model	75	86	64	91	69	89	83	89
LDA Model with Custom cutoff	70	88	71	87	71	88	83	90
Tunned KNN Model	77	87	66	92	71	89	84	91.7

Tab:62 Comparison of the Performance Metrics of All the Models (Train Data)

Comparison of the Performance Metrics of All the Models on Test Data

Model Name (Test Data)	Precision		Recall		f1-score		Model Accuracy	AUC/ROC
	0(Cons)	1(Labour)	0(Cons)	1(Labour)	0(Cons)	1(Labour)		
Logistic Regression	75	86	72	88	73	87	82	87.7
LDA	77	86	73	89	74	88	83	88.8
KNN	77	83	65	90	70	90	82	85.2
Naïve Baise	74	87	73	87	73	87	82	87.6
Random Forest Base Model	78	84	65	90	71	87	82	89.2
Bagging with Random Forest	78	85	68	90	71	88	83	89.7
ADA Boosting Model	75	84	67	88	71	86	81	87.7
Gradient Boosting Model	80	85	69	91	74	88	84	89.9
Tunned Logistic Regression Model	75	86	72	88	74	87	83	87.7
LDA Model with Custom cutoff	73	89	78	85	75	87	83	89
Tunned KNN Model	79	84	65	91	72	88	83	86.3

Tab:63 Comparison of the Performance Metrics of All the Models (Test Data)

Conclusion of Comparison :

Study over the classification reports of all models and deployment of performance improvement techniques/methods and fine-tuning the model, we find that among the acceptable models are:

- Logistic Regression
- LDA
- KNN
- Navie Bayes
- Ada Boosting
- Gradient Boosting Model
- Tunned Logistic Regression Model
- LDA Model with Custom Cutoff
- Tunned KNN Model

This is because the above models are balanced and show some degree of performance improvement when run on test data. Also, they do not have an overfitting problem on the Training Data. However, we will choose the best model among the above.

The models that are being rejected are:

- Random Forest Base Model
- Bagged Random Forest Model

Random Forest Base Model & Bagged Random Forest suffer from the problem of overfitting and therefore these models are unreliable.

Final Model Selection :

As per the context of business problem and the given dataset we finalised KNN algorithm as it is a good choice because we have a small dataset and the data is noise free and labeled plus KNN works well with smaller dataset because it is a lazy learner. It needs to store all the data and then makes decision only at run time.Moreover , KNN model have good balanced precision , recall and f1 score.KNN model is reliable to predict which party a voter will vote for on the basis of the given information.

1.8 Based on these predictions, what are the insights?

Insights :

- The Election_Data.xlsx data set has 1525 observations (rows) and 9 variables (columns) in the dataset.
- Their are total 1525 rows & 9 columns in this dataset,indexed from 0 to 1524. Out of 9 variables 7 are int64 , 2 variables are object. Memory used by the dataset: 107.4+ KB.
- There are no null values present in the dataset.
- No Anomalies found in the Dataset.
- Out of 1524 voters 1063 people voted for Labour party and 462 people voted for Conservative party.
- Out of 1524 voters 812 female voters present in the dataset and 713 male voters present in the dataset.
- Most of the participants/voters around (39.8%) gives a score of 3 for the current national economic conditions.
- 36.5% participants/voters gives a score of 4 for the current national economic conditions.
- 16.9% participants/voters gives a score of 2 for the current national economic conditions.
- 5.4% participants/voters gives a score of 5 for the current national economic conditions.
- Only 2.4% participants/voters gives a score of 1 for the current national economic conditions.

- Most of the participants/voters around (42.5%) gives a score of 3 for the current household economic conditions.
- 28.7% participants/voters gives a score of 4 for the current household economic conditions.
- 18.5% participants/voters gives a score of 2 for the current household economic conditions.
- 6.1% participants/voters gives a score of 5 for the current household economic conditions.
- Only 4.3% participants/voters gives a score of 1 for the current household economic conditions.

- Most of the participants/voters around (54.9%) gives a score of 4 to Blair.
- 28.6% of the participants/voters gives a score of 2 to Blair.
- 10% of the participants/voters gives a score of 5 to Blair.
- 6.4% of the participants/voters gives a score of 1 to Blair.
- Only 0.1% of the participants/voters gives a score of 3 to Blair.

- Most of the participants/voters around (40.7%) gives a score of 2 to Hague.
- 36.7% of the participants/voters gives a score of 4 to Hague.
- 15.4% of the participants/voters gives a score of 1 to Hague.
- 4.8% of the participants/voters gives a score of 5 to Hague. of the
- Only 2.8% of the participants/voters gives a score of 1 to Hague.

- 29% of the voters have represent European Unionism sentiments.
- Around 21.7% of the voters are considered to be neutral.
- 49.3% of the voters have represent Eurosceptic sentiment.

- 29.9% of the voters have least i.e. (0) political knowledge - (knowledge of the voters regarding the party's position on European integration).
- 16.4% of voters have high i.e(3) political knowledge - (knowledge of the voters regarding the party's position on European integration).
- 51.2% of voters have fair i.e.(2) political knowledge - (knowledge of the voters regarding the party's position on European integration).
- 2.5% of voters have level(1) political knowledge - (knowledge of the voters regarding the party's position on European integration).

- 53.3% of the voters are female.
- 46.7% of the voters are males.
- 69.7% of the participants/voters vote for Labour Party.
- 30.3% of the participants/voters vote for Conservative Party.
- Their is no such relationship between the variables , hence their is no problem of multicollinearity in the dataset.
- Their is no outliers present in the dataset.
- age , Europe ,Hague Blair are the important variables for predictions.
- As per the analysis we come to come know that chances of Labour party to win the election is more as compared to the Conservative party.
- As per the context of business problem and the given dataset we finalised KNN algorithm as it is a good choice because we have a small dataset and the data is noise free and labeled plus KNN works well with smaller dataset because it is a lazy learner. It needs to store all the data and then makes decision only at run time.Moreover , KNN model have good balanced precision , recall and f1 score.KNN model is reliable to predict which party a voter will vote for on the basis of the given information.

Recommendations :

- News Channel can take large sample of the dataset for better predictions as more is the amount of sample data we can do better prediction.So we apply the other machine learning algorithms like Random Forest , ANN, etc which works far better on the large dataset.
- As Sample size is an important consideration for research. Larger sample sizes provide more accurate mean values, identify outliers that could skew the data in a smaller sample and provide a smaller margin of error that helps news channel in predicting overall win and seats covered by a particular party.

Problem Statement 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

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

Loading the Speeches of the Presidents of the United States of America.

Speech 1 :

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

Checking the Number of Characters Present in the Speech 1 -

The Number of the Characters Present in the Speech 1(1941-Roosevelt.txt) is 7571.

Checking the Number of Words Present in the Speech 1 -

The Number of the Words Present in the Speech 1(1941-Roosevelt.txt) is 1360.

Checking the Number of Sentences Present in the Speech 1 -

The Number of Sentences Present in the Speech 1 (1941-Roosevelt.txt) is 68

Speech 2 :

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

Checking the Number of Characters Present in the Speech 2 -

The number of the characters present in the speech 2 ('1961-Kennedy.txt') is 7618.

Checking the Number of Words Present in the Speech 2 -

The Number of the Words Present in the Speech 2 ('1961-Kennedy.txt') is 1390.

Checking the Number of Sentences Present in the Speech 2 -

The Number of Sentences Present in the Speech 2 ('1961-Kennedy.txt') is 52.

Speech 3

'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and good country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilities greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis past year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendships, and by our missions to Peking and to Moscow, we were able to establish the base for a new and more durable pattern of relationships among the nations of the world. Because of America\'s bold initiatives, 1972 will be long remembered as the year of the greatest progress since the end of World War II toward a lasting peace in the world.\n\nThe peace we seek in the world is not the flimsy peace which is merely an interlude between wars, but a peace which can endure for generations to come.\n\nIt is important that we understand both the necessity and the limitations of America\'s role in maintaining that peace.\n\nUnless we in America work to preserve the peace, there will be no peace.\n\nUnless we in America work to preserve freedom, there will be no freedom.\n\nBut let us clearly understand the new nature of America\'s role, as a result of the new policies we have adopted over these past four years.\n\nWe shall respect our treaty commitments.\n\nWe shall support vigorously the principle that no country has the right to impose its will or rule on another by force.\n\nWe shall continue, in this era of negotiation, to work for the limitation of nuclear arms, and to reduce the danger of confrontation between the great powers.\n\nWe shall do our share in defending peace and freedom in the world. But we shall expect others to do their share.\n\nThe time has passed when America will make every other nation\'s conflict our own, or make every other nation\'s future our responsibility, or presume to tell the people of other nations how to manage their own affairs.\n\nJust as we respect the right of each nation to determine its own future, we also recognize the responsibility of each nation to secure its own future.\n\nJust as America\'s role is indispensable in preserving the world\'s peace, so is each nation\'s role indispensable in preserving its own peace.\n\nTogether with the rest of the world, let us resolve to move forward from the beginnings we have made. Let us continue to bring down the walls of hostility which have divided the world for too long, and to build in their place bridges of understanding -- so that despite profound differences between systems of government, the people of the world can be friends.\n\nLet us build a structure of peace in the world in which the weak are as safe as the strong -- in which each respects the right of the other to live by a different system -- in which those who would influence others will do so by the strength of their ideas, and not by the force of their arms.\n\nLet us accept that high responsibility not as a burden, but gladly -- gladly because the chance to build such a peace is the noblest endeavor in which a nation can engage; gladly, also, because only if we act greatly in meeting our responsibilities abroad will we remain a great Nation, and only if we remain a great Nation will we act greatly in meeting our challenges at home.\n\nWe have the chance today to do more than ever before in our history to make life better in America -- to ensure better education, better health, better housing, better transportation, a cleaner environment -- to restore respect for law, to make our communities more livable -- and to insure the God-given right of every American to full and equal opportunity.\n\nBecause the range of our needs is so great -- because the reach of our opportunities is so great -- let us be bold in our determination to meet those needs in new ways.\n\nJust as building a structure of peace abroad has required turning away from old policies that failed, so building a new era of progress at home requires turning away from old policies that have failed.\n\nAbroad, the shift from old policies to new has not been a retreat from our responsibilities, but a better way to peace.\n\nAnd at home, the shift from old policies to new will not be a retreat from our responsibilities, but a better way to progress.\n\nAbroad and at home, the key to those new responsibilities lies in the placing and the division of responsibility.

We have lived too long with the consequences of attempting to gather all power and responsibility in Washington.\n\nAbroad and at home, the time has come to turn away from the condescending policies of paternalism -- of "Washington knows best."\n\nA person can be expected to act responsibly only if he has responsibility. This is human nature. So let us encourage individuals at home and nations abroad to do more for themselves, to decide more for themselves. Let us locate responsibility in more places. Let us measure what we will do for others by what they will do for themselves.\n\nThat is why today I offer no promise of a purely governmental solution for every problem. We have lived too long with that false promise. In trusting too much in government, we have asked of it more than it can deliver. This leads only to inflated expectations, to reduced individual effort, and to a disappointment and frustration that erode confidence both in what government can do and in what people can do.\n\nGovernment must learn to take less from people so that people an do more for themselves.\n\nLet us remember that America was built not by government, but by people -- not by welfare, but by work -- not by shirking responsibility, but by seeking responsibility.\n\nIn our own lives, let each of us ask -- not just what will government do for me, but what can I do for myself?\n\nIn the challenges we face together, let each of us ask -- not just how can government help, but how can I help?\n\nYour National Government has a great and vital role to play. And I pledge to you that where this Government should act, we will act boldly and we will lead boldly. But just as important is the role that each and every one of us must play, as an individual and as a member of his own community.\n\nFrom this day forward, let each of us make a solemn commitment in his own heart: to bear his responsibility, to do his part, to live his ideals -- so that together, we can see the dawn of a new age of progress for America, and together, as we celebrate our 200th anniversary as a nation, we can do so proud in the fulfillment of our promise to ourselves and to the world.\n\nAs America\'s longest and most difficult war comes to an end, let us again learn to debate our differences with civility and decency. And let each of us reach out for that one precious quality government cannot provide -- a new level of respect for the rights and feelings of one another, a new level of respect for the individual human dignity which is the cherished birthright of every American.\n\nAbove all else, the time has come for us to renew our faith in ourselves and in America.\n\nIn recent years, that faith has been challenged.\n\nOur children have been taught to be ashamed of their country, ashamed of their parents, ashamed of America\'s record at home and of its role in the world.\n\nAt every turn, we have been beset by those who find everything wrong with America and little that is right. But I am confident that this will not be the judgment of history on these remarkable times in which we are privileged to live.\n\nAmerica\'s record in this century has been unparalleled in the world\'s history for its responsibility, for its generosity, for its creativity and for its progress.\n\nLet us be proud that our system has produced and provided more freedom and more abundance, more widely shared, than any other system in the history of the world.\n\nLet us be proud that in each of the four wars in which we have been engaged in this century, including the one we are now bringing to an end, we have fought not for our selfish advantage, but to help others resist aggression.\n\nLet us be proud that by our bold, new initiatives, and by our steadfastness for peace with honor, we have made a break-through toward creating in the world what the world has not known before -- a structure of peace that can last, not merely for our time, but for generations to come.\n\nWe are embarking here today on an era that presents challenges great as those any nation, or any generation, has ever faced.\n\nWe shall answer to God, to history, and to our conscience for the way in which we use these years.\n\nAs I stand in this place, so hallowed by history, I think of others who have stood here before me. I think of the dreams they had for America, and I think of how each recognized that he needed help far beyond himself in order to make those dreams come true.\n\nToday, I ask your prayers that in the years ahead I may have God\'s help in making decisions that are right for America, and I pray for your help so that together we may be worthy of our challenge.\n\nLet us pledge together to make these next four years the best four years in America\'s history, so that on its 200th birthday America will be as young and as vital as when it began, and as bright a beacon of hope for all the world.\n\nLet us go forward from here confident in hope, strong in our faith in one another, sustained by our faith in God who created us, and striving always to serve His purpose.\n'

Checking the Number of Characters Present in the Speech 3 -

The Number of the Characters Present in the Speech 3 ('1973-Nixon.txt') is 9991.

Checking the Number of Words Present in the Speech 3 -

The Number of the Words Present in the Speech 3 ('1973-Nixon.txt') is 1819.

Checking the Number of Sentences Present in the Speech 3 -

The Number of sentences Present in the Speech 3 ('1973-Nixon.txt') is 68.

Table of characters,words and sentences of Speeches Problem 2.1 :

S.No	Speech_Name	Characters_count	words_count	sentence_count
1	1941-Roosevelt.txt	7571	1360	68
2	1961-Kennedy.txt	7618	1390	52
3	1973-Nixon.txt	9991	1819	68

Tab:64 Table of Characters , Words, and Sentence of Speeches Problem 2.1

2.2 Remove all the stopwords from all three speeches.

Create DataFrame(df_s1,df_s2,df_s3) respectively for Speech1,Speech2 and Speech3.

DataFrame of Speech 1

	speech1
0	On each national day of inauguration since 178...
1	In Washington's day the task of the people was...
2	In Lincoln's day the task of the people was to...
3	In this day the task of the people is to save ...
4	To us there has come a time, in the midst of s...
...	...
63	In the face of great perils never before encou...
64	For this we muster the spirit of America, and ...
65	We do not retreat.
66	We are not content to stand still.
67	As Americans, we go forward, in the service of...

68 rows × 1 columns

Tab:65 Dataframe of Speech 1

DataFrame of Speech 2

	speech2
0	Vice President Johnson, Mr. Speaker, Mr. Chief...
1	For I have sworn I before you and Almighty God...
2	The world is very different now.
3	For man holds in his mortal hands the power to...
4	And yet the same revolutionary beliefs for whi...
5	We dare not forget today that we are the heirs...
6	Let the word go forth from this time and place...
7	Let every nation know, whether it wishes us we...
8	This much we pledge -- and more.
9	To those old allies whose cultural and spiritu...

Tab:66 Dataframe of Speech2

DataFrame of Speech 3

	speech3
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...
1	As we meet here today, we stand on the thresho...
2	The central question before us is: How shall w...
3	Let us resolve that this era we are about to e...
4	Let us resolve that this will be what it can b...
...	...
63	As I stand in this place, so hallowed by histo...
64	I think of the dreams they had for America, an...
65	Today, I ask your prayers that in the years ah...
66	Let us pledge together to make these next four...
67	Let us go forward from here confident in hope,...

68 rows × 1 columns

Tab:67 Dataframe of Speech 3

Stopwords in English

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'h
is', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'who
m', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing'
, 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'd
uring', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there
', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than
', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'could
n', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mu
stn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

EDA Steps for Speech 1-('1941-Roosevelt.txt')

Remove punctuation and special character like(--)

	speech1
0	On each national day of inauguration since 178...
1	In Washingtons day the task of the people was ...
2	In Lincolns day the task of the people was to ...
3	In this day the task of the people is to save ...
4	To us there has come a time in the midst of sw...

Tab:68 Remove punctuation and special character Speech1

Convert the Speech into Lower Case Conversion and check the Data frame of Speech 1

	speech1
0	on each national day of inauguration since 178...
1	in washingtons day the task of the people was ...
2	in lincolns day the task of the people was to ...
3	in this day the task of the people is to save ...
4	to us there has come a time in the midst of sw...
...	...
63	in the face of great perils never before encou...
64	for this we muster the spirit of america and t...
65	we do not retreat
66	we are not content to stand still
67	as americans we go forward in the service of o...

68 rows × 1 columns

Tab:69 DataFrame with Lower Case of Speech 1

Checking of Stopwords in Each Sentence of Speech1

	speech1	stopwords
0	on each national day of inauguration since 178...	9
1	in washingtons day the task of the people was ...	8
2	in lincolns day the task of the people was to ...	9
3	in this day the task of the people is to save ...	12
4	to us there has come a time in the midst of sw...	26
5	if we do not we risk the real peril of inaction	7
6	lives of nations are determined not by the cou...	11
7	the life of a man is threescore years and ten ...	8
8	the life of a nation is the fullness of the me...	11
9	there are men who doubt this	4

Tab:70 Stopwords in Each Sentence of Speech1

Sum of All Stopwords in Speech1----711

Sentences of Speech 1 After Removal of Stopwords

	speech1	stopwords	speech_1_after_remove_stopwords
0	on each national day of inauguration since 178...	9	national day inauguration since 1789 people re...
1	in washingtons day the task of the people was ...	8	washingtons day task people create weld togeth...
2	in lincolns day the task of the people was to ...	9	lincolns day task people preserve nation disru...
3	in this day the task of the people is to save ...	12	day task people save nation institutions disru...
4	to us there has come a time in the midst of sw...	26	us come time midst swift happenings pause mome...

Tab:71 Sentences of Speech 1 After Removal of Stopwords

Stopwords Count Comparison Before and After Speech 1

	speech1	stopwords	speech_1_after_remove_stopwords	after_remove_stopwords
0	on each national day of inauguration since 178...	9	national day inauguration since 1789 people re...	0
1	in washingtons day the task of the people was ...	8	washingtons day task people create weld togeth...	0
2	in lincolns day the task of the people was to ...	9	lincolns day task people preserve nation disru...	0
3	in this day the task of the people is to save ...	12	day task people save nation institutions disru...	0
4	to us there has come a time in the midst of sw...	26	us come time midst swift happenings pause mome...	0
5	if we do not we risk the real peril of inaction	7	risk real peril inaction	0
6	lives of nations are determined not by the cou...	11	lives nations determined count years lifetime ...	0
7	the life of a man is threescore years and ten ...	8	life man threescore years ten little little less	0
8	the life of a nation is the fullness of the me...	11	life nation fullness measure live	0
9	there are men who doubt this	4	men doubt	0

Tab:72 Stopwords Count Comparison Before and After Speech 1

Sample Sentence of Speech 1 Before Removal of Stopwords -

'on each national day of inauguration since 1789 the people have renewed their sense of dedication to the united states'

Sample Sentence of Speech 1 after Removal of Stopwords -

'national day inauguration since 1789 people renewed sense dedication united states'

Word Count With Stopwords Speech 1

	speech1	word_count_with_stopwords
0	on each national day of inauguration since 178...	20
1	in washingtons day the task of the people was ...	16
2	in lincolns day the task of the people was to ...	17
3	in this day the task of the people is to save ...	20
4	to us there has come a time in the midst of sw...	41

Tab:73 Word Count With Stopwords Speech 1

Total Word Count Sum With Stopwords Speech 1----------1338

Word Count Without Stopwords Speech 1

	speech1	word_count_without_stopwords
0	on each national day of inauguration since 178...	11
1	in washingtons day the task of the people was ...	8
2	in lincolns day the task of the people was to ...	8
3	in this day the task of the people is to save ...	8
4	to us there has come a time in the midst of sw...	15

Tab:74 Word Count Without StopwordsSpeech 1

Total Word Count Sum Without Stopwords Speech 1----------627

EDA Steps for Speech 2-('1961-Kennedy.txt')

Remove Punctuation and Special Character like(--)

	speech2
0	Vice President Johnson Mr Speaker Mr Chief Jus...
1	For I have sworn I before you and Almighty God...
2	The world is very different now
3	For man holds in his mortal hands the power to...
4	And yet the same revolutionary beliefs for whi...

Tab:75 Remove punctuation and special character Speech2

Convert the Speech into Lower Case Conversion and check the Data frame of Speech 2

	speech2
0	vice president johnson mr speaker mr chief jus...
1	for i have sworn i before you and almighty god...
2	the world is very different now
3	for man holds in his mortal hands the power to...
4	and yet the same revolutionary beliefs for whi...
5	we dare not forget today that we are the heirs...
6	let the word go forth from this time and place...
7	let every nation know whether it wishes us wel...
8	this much we pledge and more
9	to those old allies whose cultural and spiritu...

Tab:76 DataFrame with Lower Case of Speech 2

Check Stopwords in Each Sentence of Speech 2

	speech2	stopwords
0	vice president johnson mr speaker mr chief jus...	13
1	for i have sworn i before you and almighty god...	12
2	the world is very different now	4
3	for man holds in his mortal hands the power to...	10
4	and yet the same revolutionary beliefs for whi...	22

Tab:77 Stopwords in Each Sentence of Speech 2

Sum of All Stopwords in Speech 2----------672

Speech 2 After Removing Stopwords

	speech2	stopwords	speech_2_after_remove_stopwords
0	vice president johnson mr speaker mr chief jus...	13	vice president johnson mr speaker mr chief jus...
1	for i have sworn i before you and almighty god...	12	sworn almighty god solemn oath forebears I pre...
2	the world is very different now	4	world different
3	for man holds in his mortal hands the power to...	10	man holds mortal hands power abolish forms hum...
4	and yet the same revolutionary beliefs for whi...	22	yet revolutionary beliefs forebears fought sti...

Tab:78 Sentences of Speech 2 After Removal of Stopwords

Stopwords Count Comparison Before and After of Speech 2

	speech2	stopwords	speech_2_after_remove_stopwords	after_remove_stopwords
0	vice president johnson mr speaker mr chief jus...	13	vice president johnson mr speaker mr chief jus...	0
1	for i have sworn i before you and almighty god...	12	sworn almighty god solemn oath forebears I pre...	0
2	the world is very different now	4	world different	0
3	for man holds in his mortal hands the power to...	10	man holds mortal hands power abolish forms hum...	0
4	and yet the same revolutionary beliefs for whi...	22	yet revolutionary beliefs forebears fought sti...	0
5	we dare not forget today that we are the heirs...	8	dare forget today heirs first revolution	0
6	let the word go forth from this time and place...	40	let word go forth time place friend foe alike ...	0
7	let every nation know whether it wishes us wel...	15	let every nation know whether wishes us well i...	0

Tab:79 Stopwords Count Comparison Before and After Speech 2

Sample Sentence of Speech 2 Before Removal of Stopwords -

'vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens we observe today not a victory of party but a celebration of freedom symbolizing an end as well as a beginning signifying renewal as well as change'

Sample Sentence 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'

Word Count With Stopwords of Speech 2

	speech2	word_count_with_stopwords
0	vice president johnson mr speaker mr chief jus...	46
1	for i have sworn i before you and almighty god...	25
2	the world is very different now	6
3	for man holds in his mortal hands the power to...	22
4	and yet the same revolutionary beliefs for whi...	39

Tab:80 Word Count With Stopwords Speech 2

Total Word Count Sum with Stopwords of Speech 2----------1365

Word Count Without Stopwords of Speech 2

	speech2	word_count_without_stopwords
0	vice president johnson mr speaker mr chief jus...	33
1	for i have sworn i before you and almighty god...	13
2	the world is very different now	2
3	for man holds in his mortal hands the power to...	12
4	and yet the same revolutionary beliefs for whi...	17

Tab:81 Word Count Without Stopwords Speech 2

Total Word Count Sum Without Stopwords of Speech 2 ----------693

EDA Steps for Speech 3-('1973-Nixon.txt')

Remove Punctuation and Special Character like(--)

	speech3
0	Mr Vice President Mr Speaker Mr Chief Justice ...
1	As we meet here today we stand on the threshol...
2	The central question before us is How shall we...
3	Let us resolve that this era we are about to e...
4	Let us resolve that this will be what it can b...

Tab:82 Remove punctuation and special character Speech 3

Convert the Speech into Lower Case Conversion and check the Data frame of Speech 3

	speech3
0	mr vice president mr speaker mr chief justice ...
1	as we meet here today we stand on the threshol...
2	the central question before us is how shall we...
3	let us resolve that this era we are about to e...
4	let us resolve that this will be what it can b...
...	...
63	as i stand in this place so hallowed by histor...
64	i think of the dreams they had for america and...
65	today i ask your prayers that in the years ahe...
66	let us pledge together to make these next four...
67	let us go forward from here confident in hope ...

68 rows × 1 columns

Tab:83 DataFrame with Lower Case of Speech 3

Check Stopwords in Each Sentence of Speech 3

	speech3	stopwords
0	mr vice president mr speaker mr chief justice ...	17
1	as we meet here today we stand on the threshol...	11
2	the central question before us is how shall we...	6
3	let us resolve that this era we are about to e...	21
4	let us resolve that this will be what it can b...	21

Tab:84 Stopwords in Each Sentence of Speech 3

Sum of All Stopwords in Speech 3----------969

Speech 3 After Removing Stopwords

	speech3	stopwords	speech_3_after_remove_stopwords
0	mr vice president mr speaker mr chief justice ...	17	mr vice president mr speaker mr chief justice ...
1	as we meet here today we stand on the threshol...	11	meet today stand threshold new era peace world
2	the central question before us is how shall we...	6	central question us shall use peace
3	let us resolve that this era we are about to e...	21	let us resolve era enter postwar periods often...
4	let us resolve that this will be what it can b...	21	let us resolve become time great responsibilit...

Tab: 85 Sentences of Speech 3 After Removal of Stopwords

Stopwords Count Comparison Before and After of Speech 3

	speech3	stopwords	speech_3_after_remove_stopwords	after_remove_stopwords
0	mr vice president mr speaker mr chief justice ...	17	mr vice president mr speaker mr chief justice ...	0
1	as we meet here today we stand on the threshol...	11	meet today stand threshold new era peace world	0
2	the central question before us is how shall we...	6	central question us shall use peace	0
3	let us resolve that this era we are about to e...	21	let us resolve era enter postwar periods often...	0
4	let us resolve that this will be what it can b...	21	let us resolve become time great responsibilit...	0

Tab:86 Stopwords Count Comparison Before and After Speech 3

Sample Sentence of Speech 3 Before Removal of Stopwords -

mr vice president mr speaker mr chief justice senator cook mrs eisenhower and my fellow citizens of this great and good country we share together when we met here four years ago america was bleak in spirit depressed by the prospect of seemingly endless war abroad and of destructive conflict at home'

Sample Sentence of Speech 3 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'

Word Count With Stopwords of Speech3

	speech3	word_count_with_stopwords
0	mr vice president mr speaker mr chief justice ...	52
1	as we meet here today we stand on the threshol...	19
2	the central question before us is how shall we...	12
3	let us resolve that this era we are about to e...	39
4	let us resolve that this will be what it can b...	38

Tab: 87 Word Count With Stopwords Speech 3

Total Word Count Sum With Stopwords of Speech 3 ----------1802

Word Count Without Stopwords of Speech 3

	speech3	word_count_without_stopwords
0	mr vice president mr speaker mr chief justice ...	35
1	as we meet here today we stand on the threshol...	8
2	the central question before us is how shall we...	6
3	let us resolve that this era we are about to e...	18
4	let us resolve that this will be what it can b...	17

Tab: 88 Word Count Without Stopwords Speech 3

Total Word Count Sum Without Stopwords of Speech 3 ----------833

Table of All Speech1,Speech2 and Speech3 Word Count Sum Before Removing Stopwords and After Removing Stopwords

Speech_Name	words_count_with_stopwords	words_count_without_stopwords
1941-Roosevelt.txt	1338	627
1961-Kennedy.txt	1365	693
1973-Nixon.txt	1802	833

Tab: 89 Table of All Speech1,Speech2 and Speech3 Word Count Sum Before Removing Stop-words and After Removing Stop-words

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)

Speech 1 - Top Three Words. (After Removing the Stopwords) in ('1941-Roosevelt.txt') with frequency count ----[('nation', 11), ('know', 10), ('spirit', 9)].

Speech 2 - Top Three Words. (After Removing the Stopwords) in ('1961-Kennedy.txt') with frequency count ----[('let', 16), ('us', 12), ('world', 8)].

Speech 3 - Top Three Words. (After Removing the Stopwords) in ('1973-Nixon.txt') with frequency count ----[('us', 26), ('let', 22), ('peace', 19)].

Note-As we know that we can include "us" and "let" and "know" in stopwords but in FAQ this is mention that we do not remove these words,so we are not removing these words from the list else we can remove according to the context and sentiments of the business problem.

Speech_Name	1st Most Occured Word	2nd Most Occured word	3rd Most Occured Word
1941-Roosevelt.txt	nation	know	spirit
1961-Kennedy.txt	let	us	world
1973-Nixon.txt	us	let	peace

Tab: 90 Table of Top 3 Words of All the Speeches

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

Word Cloud Speech 1

FIG: 67 Word Cloud Speech 1

Word Cloud Speech 2

FIG: 68 Word Cloud Speech 2

Word Cloud Speech 3

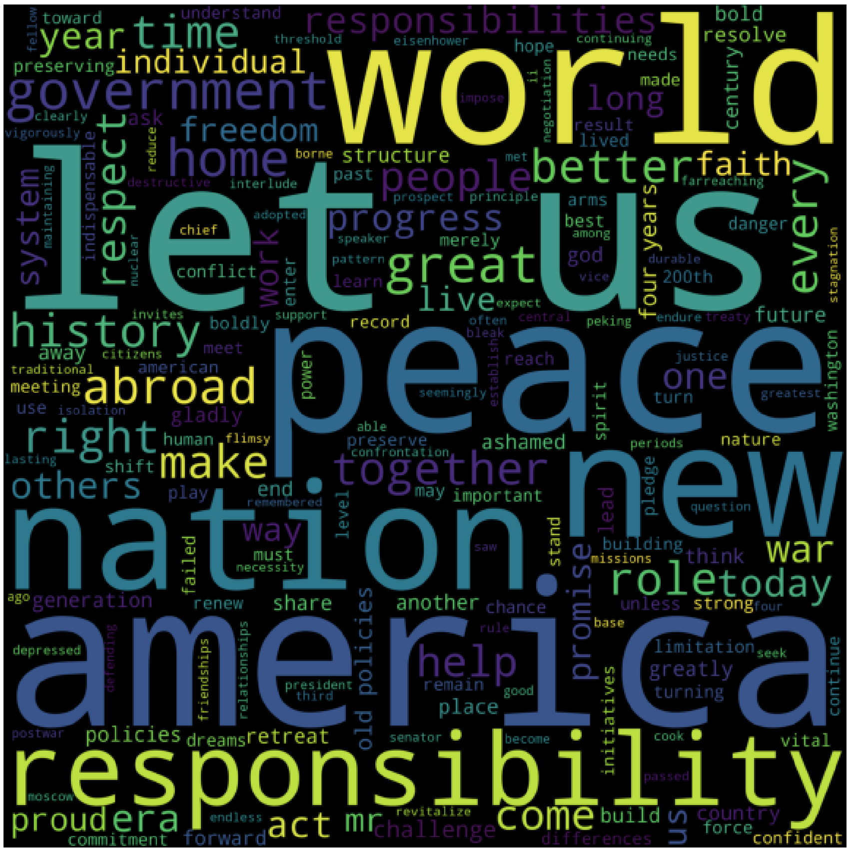

FIG: 69 Word Cloud Speech 3

Thank You