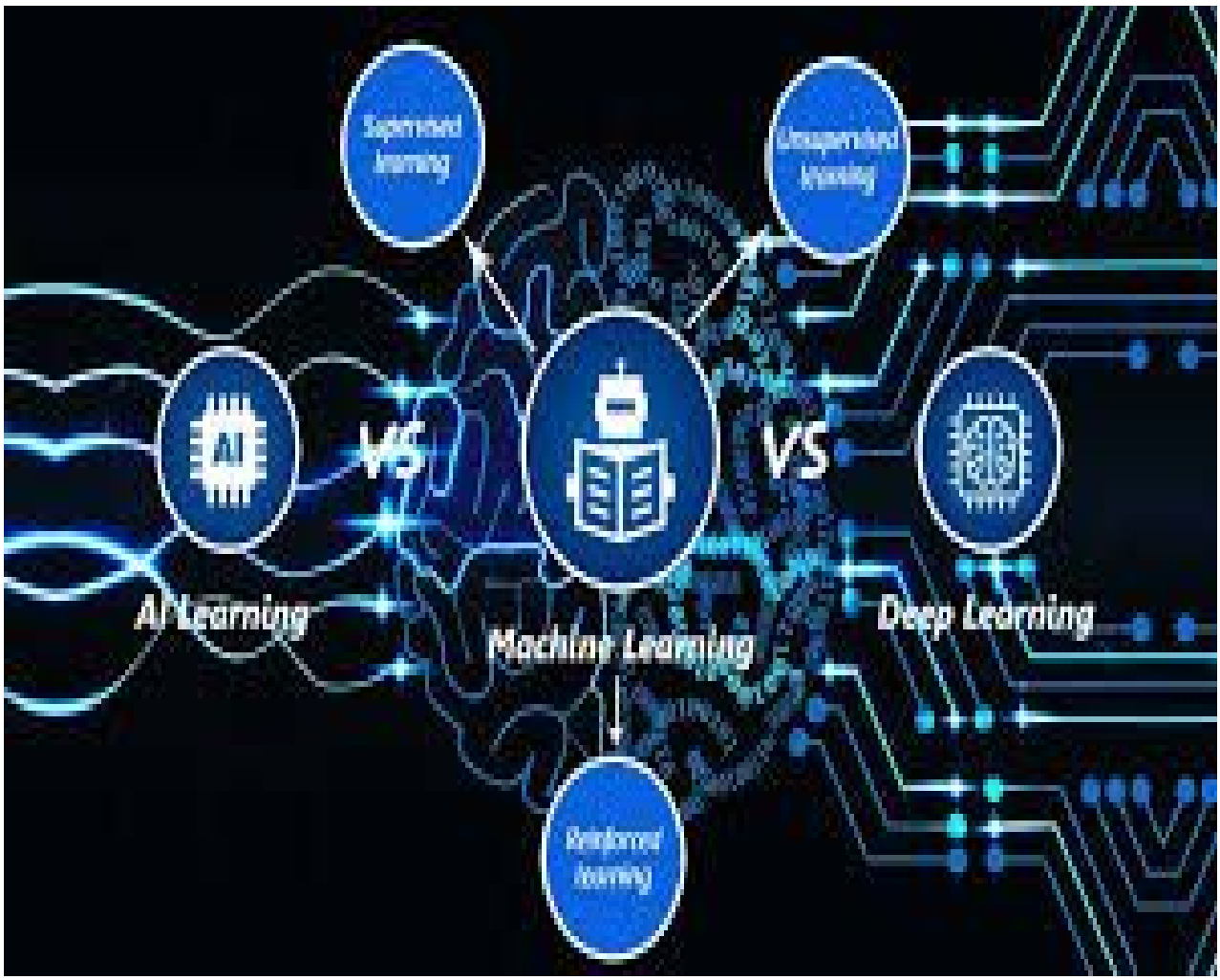


BUSINESS REPORT

19.07.2020

SHOBANADEVI R

shobanadevir2096@gmail.com

7395985368

Ramco Systems

TABLE OF CONTENTS

Case Snippet1:You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

DATA SECTION: The dataset of Election data has 1525 rows and 9 variables

Descriptive Statistics:

Import the data:

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

Shape of Data: Checking the duplicates

Before (1517, 9)
After deleting duplicate (1517, 9)

Info of the dataset

```
Data columns (total 9 columns):
vote                    1517 non-null object
age                     1517 non-null int64
economic.cond.national  1517 non-null int64
economic.cond.household 1517 non-null int64
Blair                   1517 non-null int64
Hague                   1517 non-null int64
Europe                  1517 non-null int64
political.knowledge     1517 non-null int64
gender                  1517 non-null object
dtypes: int64(7), object(2)
memory usage: 106.7+ KB
```

Null Value condition:The dataset has no null values.

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

Vote - To predict the value.

Conservative -460

Labour -1057

INFERENCE:

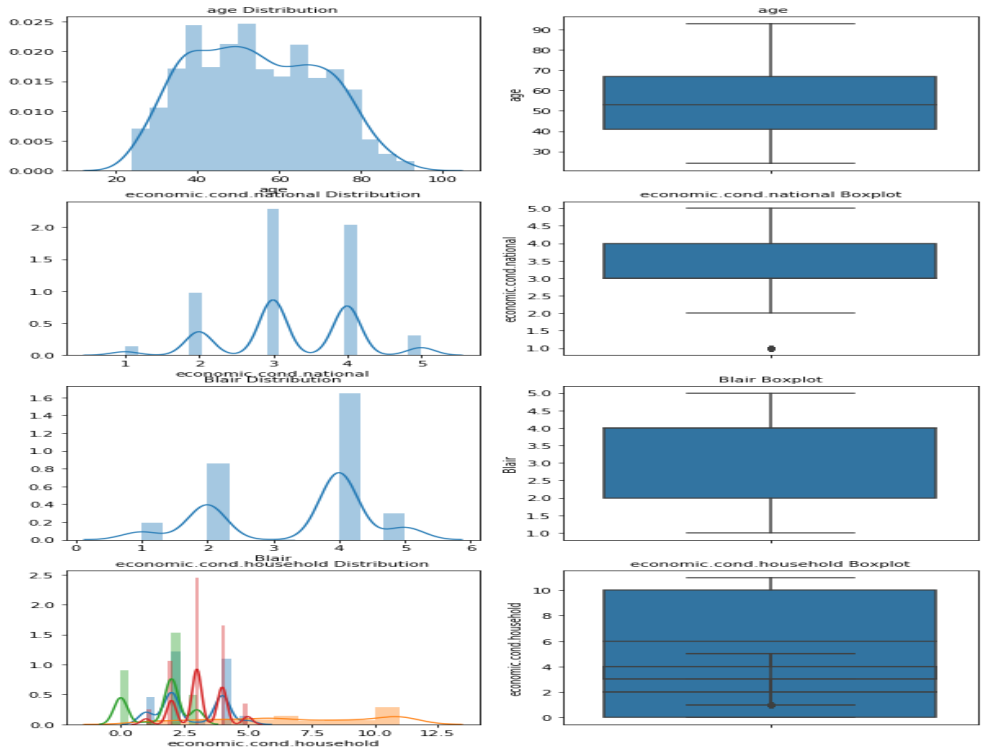
- Here all columns in the dataset have mostly number as datatype, which will doesn't have much variation in the dataset.

- Only two columns has object type, encoding that into numerical will make it easy for building models.
- Age column has data values which have continuous distribution values in the required ranges.
- Vote is the variable which we need to predict which has two value counts.

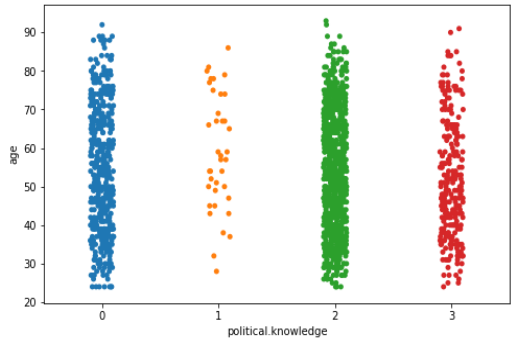
ANALYSIS SECTION

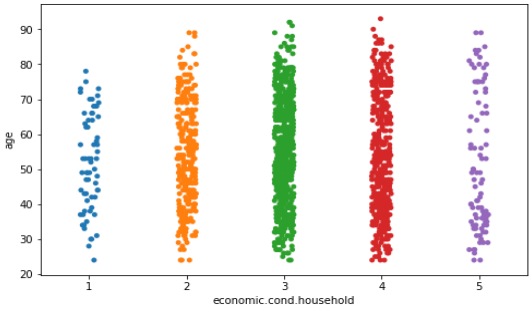
Univariate Analysis: Inference

- From the Univariate analysis of Age it shows wide distribution along 20-100.People started showing interest to vote after 30 age, and mostly people between age of 30-60 showed more high interest in voting.
- Comparing both economic conditional national and household has higher ratings.
- Thus by seeing the analysis the age, blair,hogue will also help to decide who will win the election

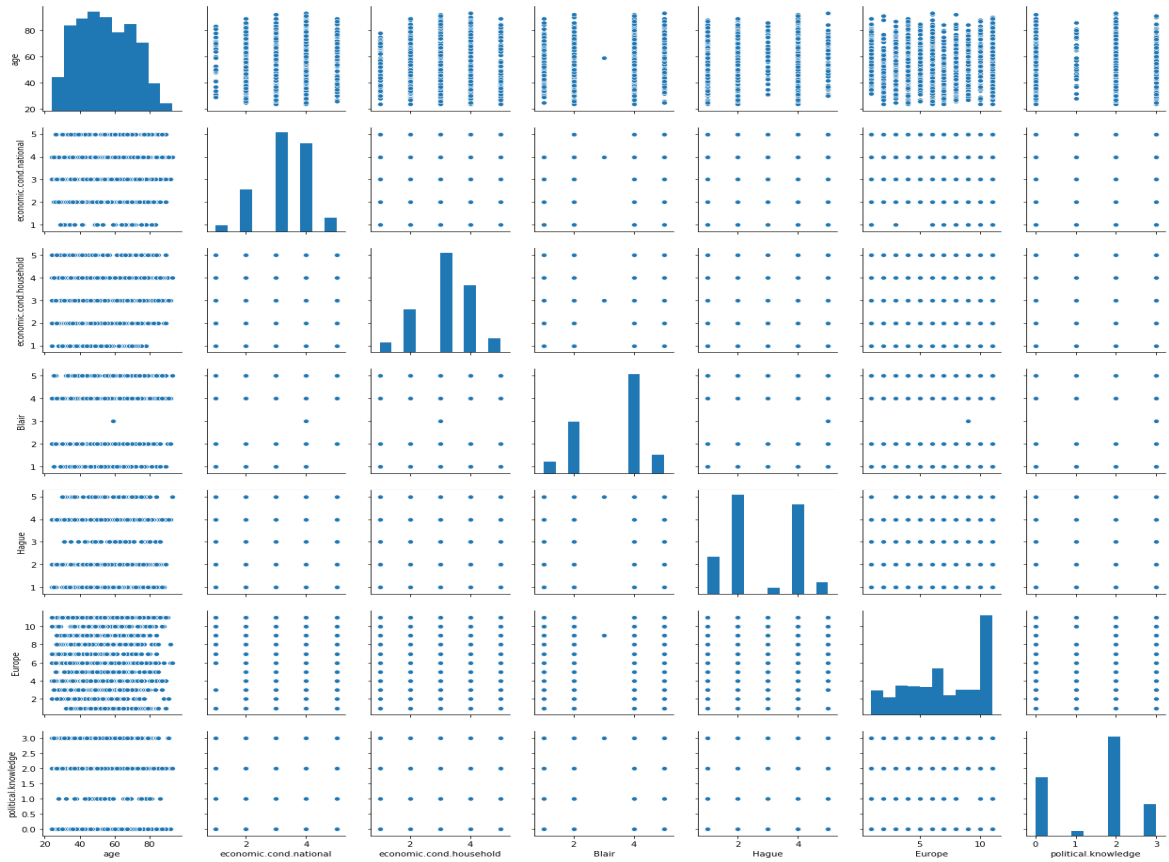


Bivariate Analysis: Having a look between age and political knowledge, the most common thing between all ages of their political knowledge is average.

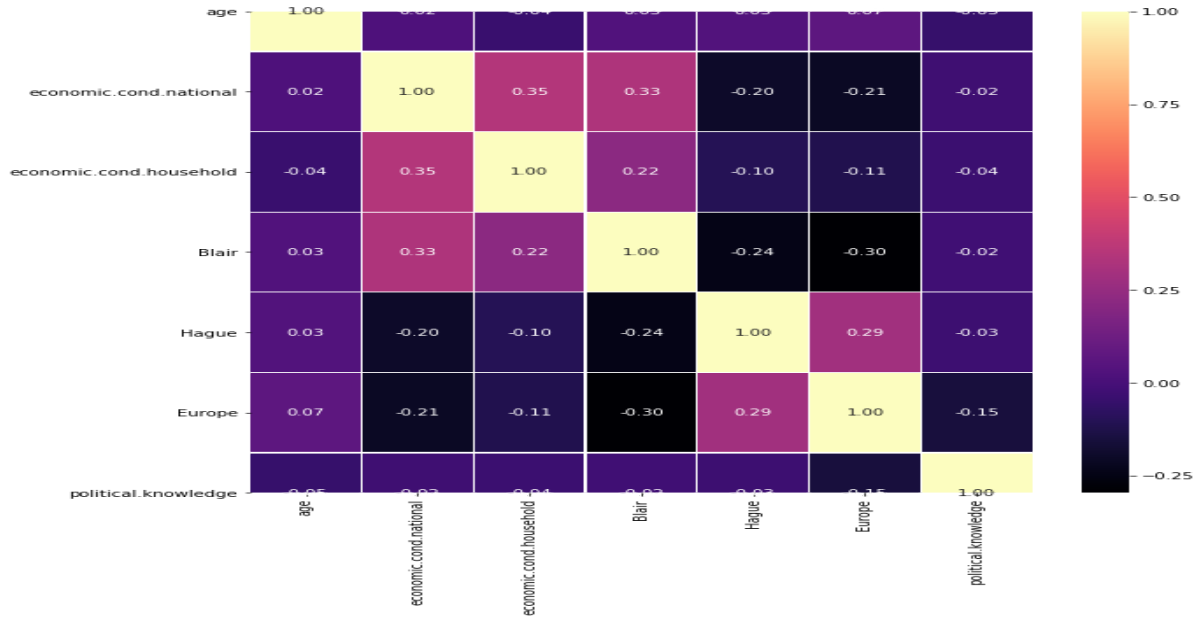




Pair plot:Only age has normal continuous ranges with the data, other variables have data as integers which are in very small ranges.



From the heatmap we can easily understand that the variables is not highly correlated, the maximum of correlation is 0.33

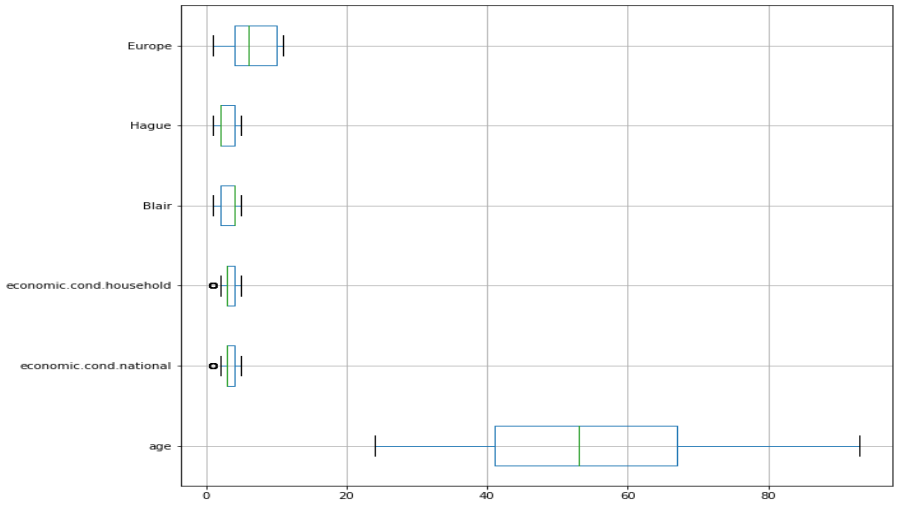


Exploratory Data Analysis: From the statistics we can understand the IOR ranges in the data.

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000
mean	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295
std	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315
min	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000
50%	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000
75%	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000
max	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000

Outlier Check:

The given dataset has no outliers as all the variables are integer and values ranges are in common.



ENCODING: encoding the string values as numbers will help us in an easy way for predicting.

```
feature: vote
[Labour, Conservative]
Categories (2, object): [Conservative, Labour]
[1 0]

feature: gender
[female, male]
Categories (2, object): [female, male]
[0 1]
```

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

SCALING: Here the values ranges are common and not having variation in the dataset, so scaling is not necessary in this dataset.

Train & Test Split Data: Splitting the data into two variables as X and Y in ratio of (70:30)

X trainset -(1061, 8) X testset - (1061, 1) Y trainset - (456, 8) Y Testset - (456, 1)

MODELLING:

Applying various models on the train and test set to predict the best model.

Logistic Regression:

Train Data score:

```
0.8369462770970783
[[198 109]
 [ 64 690]]
              precision    recall  f1-score   support

           0       0.76      0.64      0.70       307
           1       0.86      0.92      0.89       754

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

Test Data Score:

```
0.8289473684210527
[[110  43]
 [ 35 268]]
              precision    recall  f1-score   support

           0       0.76      0.72      0.74       153
           1       0.86      0.88      0.87       303

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

LDA Model:

Train Data Score :

```
0.8341187558906692
[[200 107]
 [ 69 685]]
              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
```

Test Data Score:

```
0.8333333333333334
[[111  42]
 [ 34 269]]
              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
```

KNN Model:

Train Data Score:

```
0.8529688972667295
[[204 103]
 [ 53 701]]
              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
```

Test Data Score:

```
0.8157894736842105
[[ 99  54]
 [ 30 273]]
              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
```

Naïve Bayes Model:

Train data Score

```
0.8350612629594723
[[211  96]
 [ 79 675]]
              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
```

Test Data Score

```
0.8223684210526315
[[112  41]
 [ 40 263]]
              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
```

SVM:

Train data Score:

```
0.9198868991517436
[[243  64]
 [ 21 733]]
              precision    recall  f1-score   support

           0       0.92      0.79      0.85       307
           1       0.92      0.97      0.95       754

    accuracy                           0.92      1061
   macro avg       0.92      0.88      0.90      1061
weighted avg       0.92      0.92      0.92      1061
```

Test data Score:

```
0.7960526315789473
[[ 88  65]
 [ 28 275]]
              precision    recall  f1-score   support

           0       0.76      0.58      0.65       153
           1       0.81      0.91      0.86       303

    accuracy                           0.80       456
   macro avg       0.78      0.74      0.75       456
weighted avg       0.79      0.80      0.79       456
```

BAGGING:

Train Data Set:

```
0.9670122525918945
[[278  29]
 [  6 748]]
              precision    recall  f1-score   support

           0       0.98      0.91      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
```

Test Data Set:

```
0.8289473684210527
[[104  49]
 [ 29 274]]
              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
```

BOOSTING:

Train Data Set:

```
0.8501413760603205
[[214  93]
 [ 66 688]]
              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
```

Test Data Set:

```
0.8135964912280702
[[103  50]
 [ 35 268]]
              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
```

Performance Metrics:

Model	Train Accuracy	Test Accuracy	Confusion matrix Train data	Confusion matrix Test data	ROC_AUC Score	ROC_AUC curve
Logistic Regression	0.8369	0.8289	[[198 109] [64 690]]	[[110 43] [35 268]]	0.833	0.832
LDA	0.834	0.833	[[200 107] [69 685]]	[[111 42] [34 269]]	0.833	0.8331
KNN	0.852	0,825	[[204 103] [53 701]]	[[99 54] [30 273]]	0.82	0.85
Naïve Bayes Model	0.835	0.822	[[211 96] [79 675]]	[[112 41] [40 263]]	0.84	0.82
SVM	0.919	0.796	[[243 64] [21 733]]	[[88 65] [28 275]]	0.92	0.80
Bagging(RF)	0.967	0.828	[[278 29] [6 748]]	[[104 49] [29 274]]	0.97	0.83
Boosting	0.8501	0.8135	[[214 93] [66 688]]	[[103 50] [35 268]]	0.85	0.81

FINAL MODEL:

BAGGING RANDOM FOREST

- ***Model Score for Train Data - 0.9670122525918945***
- ***Model Score for Test Data - 0.8289473684210527***
- ***Accuracy Score for Train Data - 0.83***
- ***Accuracy Score for Test Data - 0.97***
- ***Compared to other models, Bagging Recall value is greater and hence performs as the best model.***

Inference:

Interest is 1 i.e Labour - 1

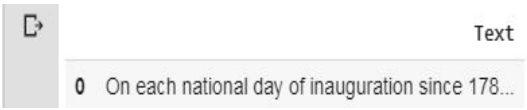
- *Let's look at the performance of all the models on the Train Data set, here recall percentage shows the total relevant results correctly classified by the algorithm hence we will compare Recall of class "1" for all models.*
- *Train data Recall is 0.99, Test data recall - 0.90 has good performance in Bagging with Random forest.*
- *Here the train data and test data fulfil has proved that the model is well fit for the prediction.*
- *As our prediction we can predict that **Labour** will win the election and will cover 90% of seats in the election.*

Case Snippet2:In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

1. President Franklin D. Roosevelt in 1941

DATA SECTION:

Here we are going to analyse the text file for this case to understand the basics of text mining.The Text file extracted from the give code and defined in Dataframe`



	Text
0	On each national day of inauguration since 178...

ANALYSIS SECTION: Basic Exploration in Text Mining

1. Number of Characters:6331

2. Number of Words

	Text	word_count
0	On each national day of inauguration since 178...	1323

2. Number of Sentences :68

Removal of StopWords:

Before - The counts of stopwords present in speech.

	Text	word_count	stopwords
	On each national day of inauguration since 178...	1360	632

After - Removal of Stopword

	Text	stopwords
	On national day inauguration since 1789 people...	0

Frequency of Words: The most used words in the speech and its top three are represented here.

nation 11

know 10

spirit 9

Wordcloud:

2.1973-Nixon.txt

DATA SECTION: Text file extracted from the given code and defined in a dataframe.

	Text
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...

ANALYSIS SECTION:Basic Exploration in Text Mining.

1. Number of words

	Text	word_count
)	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769

2. Number of characters

	Text	char_count
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9991

3. Number of sentences:68

Removal of Stop words:

Before - Removal of Stopwords

	Text	char_count	word_count	stopwords
0	mr. vice president, mr. speaker, mr. chief jus...	9991	1769	958

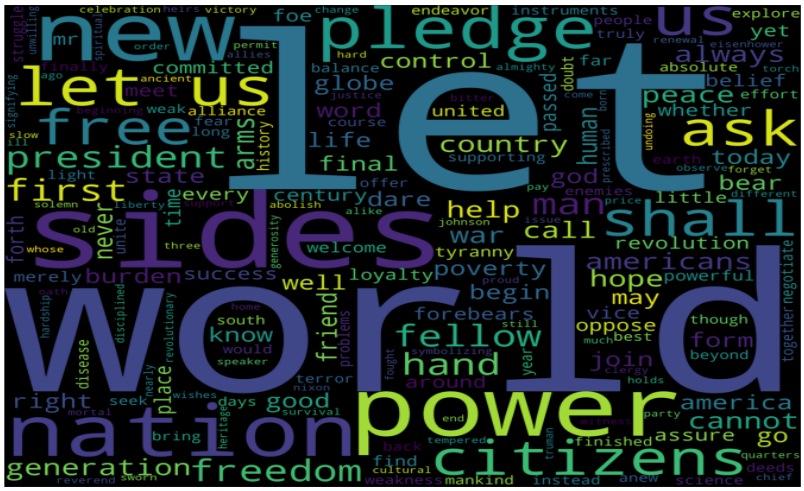
After - Removal of Stopwords

	Text	char_count	word_count	stopwords
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9991	1769	0

Frequency of words: The most used words in the speech with its top three are represented here.

```
us        26
let       22
peace     19
world     16
```

Wordcloud :



3.President John F. Kennedy in 1961

DATA SECTION: Text file extracted from the given code and defined in a dataframe.

	Text
0	Vice President Johnson, Mr. Speaker, Mr. Chief...

ANALYSIS SECTION: Basic Exploration of Text Mining

1. Number of Words

	Text	word_count
0	vice president johnson mr speaker mr chief jus...	1365

2. Number of Characters

	Text	char_count
0	vice president johnson mr speaker mr chief jus...	7362

3. Number of Sentences :52

Stop words

Before - the count of stopwords present in speech

	Text	word_count	char_count	stopwords
0	vice president johnson mr speaker mr chief jus...	1365	7362	672

After Removal of stopwords

	Text	word_count	char_count	stopwords
0	vice president johnson mr speaker mr chief jus...	1365	7362	0

Frequency of words: the most used words in speech

let	16
us	12
world	8
sides	8
pledge	7

Wordcloud: