Business Report
Machine Learning

`

Arnab Ghosal

28 July 2021

Problem 1:

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

Dataset for Problem: Election_Data.xlsx

Data Ingestion: 11 marks

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks)

Reading 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

Descriptive Statistics

	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

Null value Check

```
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
```

We can see that there is no null values in this dataset.

Duplicate value Check

We have found there were **8** duplicate value were existing , we had to drop them to clean the data .

Inference

1. There are total **1525 rows** & **10 columns** are present in the dataset.
2. Out of 10 columns there are 8 **numeric** columns and 2 **categoric** columns are there .
3. There is no null value existing the dataset.
4. Initially we have found 8 duplicate records in the dataset but we have dropped them eventually in-terms of cleaning the data .
5. Out of two categoric columns “voters” are distributed b/w two categoric columns , **“labour”** & **“Conservative”** where distribution ratio is not uniform on the other “Gender “ column is distributed b/w **“Male”** & **“Female”.**

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

Shape of the Data

Dataset is having 1525 rows & 10 columns

Info of the Data

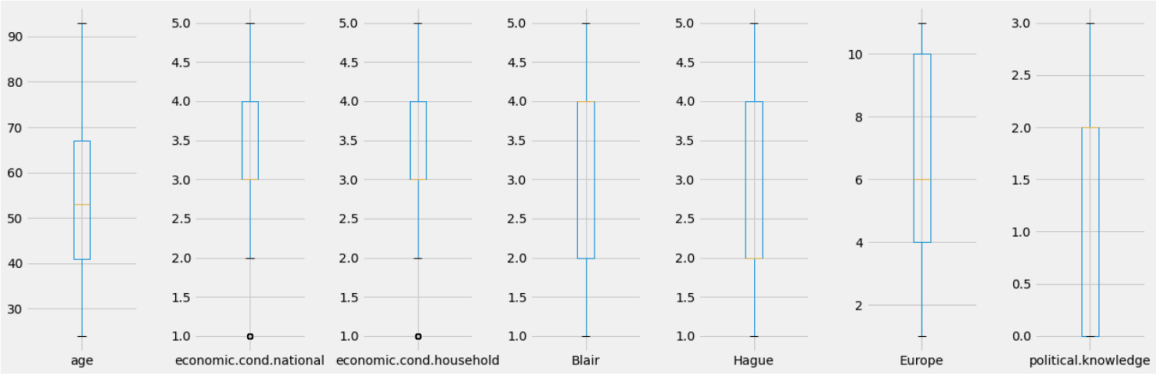
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   Unnamed: 0              1525 non-null   int64 
 1   vote                    1525 non-null   object
 2   age                     1525 non-null   int64 
 3   economic.cond.national  1525 non-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
dtypes: int64(8), object(2)
memory usage: 119.3+ KB
```

Printing Unique value count for categoric variables

```
vote   No of Levels: 2
Labour          1063
Conservative     462
Name: vote, dtype: int64


gender   No of Levels: 2
female    812
male      713
Name: gender, dtype: int64
```

Checking for outliers



We can see from the above box-plots that age columns doesn't have any outlier & other columns like "**economic.cond.national**" & "**economic.cond.household**" are ordinal variable so ideally those columns can not have outliers .

Now before going ahead for Univariate analysis we wanted to convert "**age**" column **.** As we can see that Age variable is having discrete values so we will try to convert this to ordinal values **.**

After the conversion dataset were as below :

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender	age_group
0	Labour	43	3	3	4	1	2	2	female	40s
1	Labour	36	4	4	4	4	5	2	male	30s
2	Labour	35	4	4	5	2	3	2	male	30s
3	Labour	24	4	2	2	1	4	0	female	20s
4	Labour	41	2	2	1	1	6	2	male	40s
5	Labour	47	3	4	4	4	4	2	male	40s
6	Labour	57	2	2	4	4	11	2	male	50s
7	Labour	77	3	4	4	1	1	0	male	70s
8	Labour	39	3	3	4	4	11	0	female	30s
9	Labour	70	3	2	5	1	11	2	male	70s

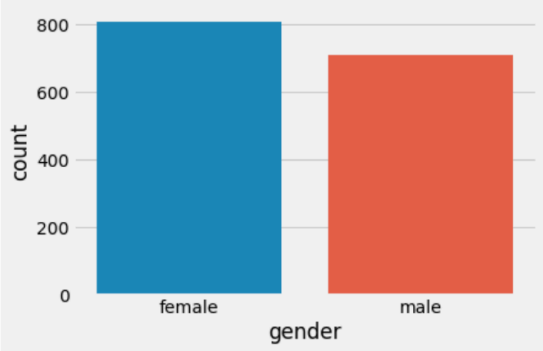
Univariate Analysis

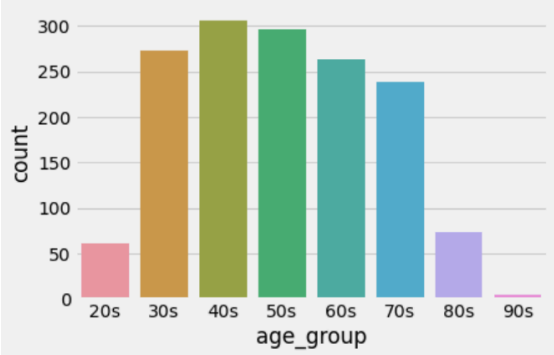
Univariate analysis is the simplest form of analysing data. “Uni” means “one”, so in other words our data takes only one variable at a time. It doesn't deal with causes or relationships (unlike regression) and it's major purpose is to describe; It takes data (individual column) separately , summarises that data and finds patterns in the data._

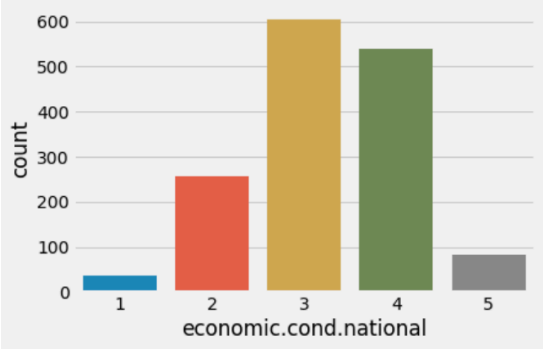
To evaluate univariate analysis we have segregated numeric & categorical columns.

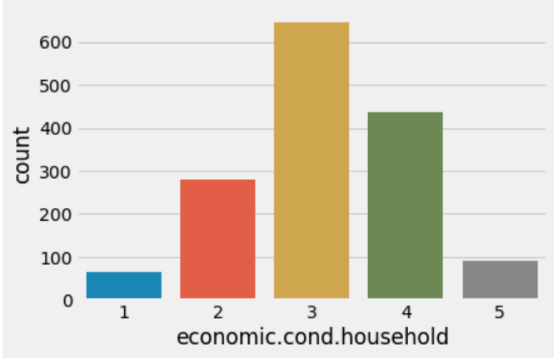
Numeric columns are explained by box plot /histogram or count plot .

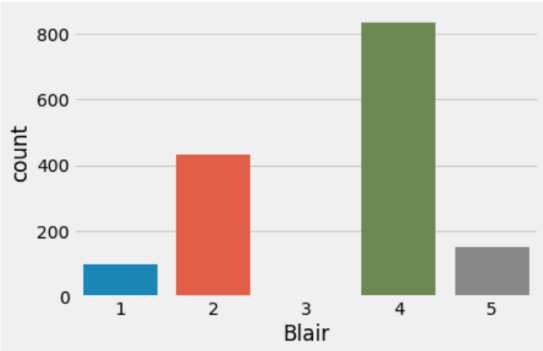
Categoric variables are explained by Strip-plot.

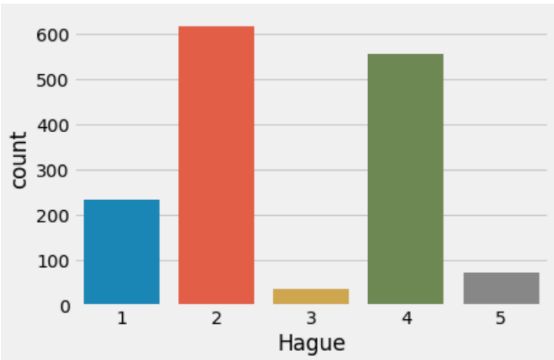


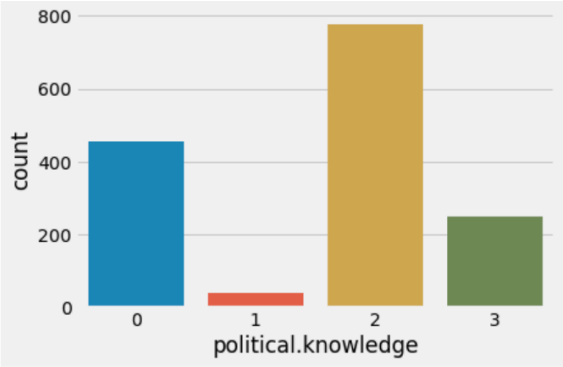






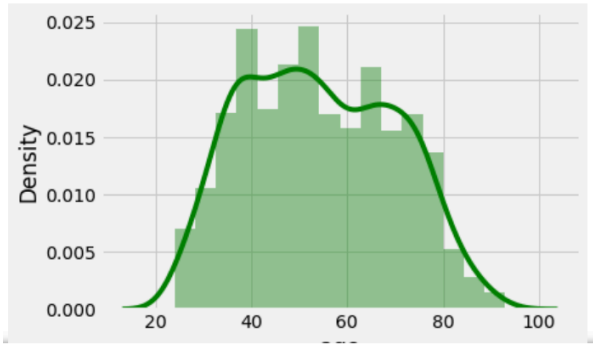




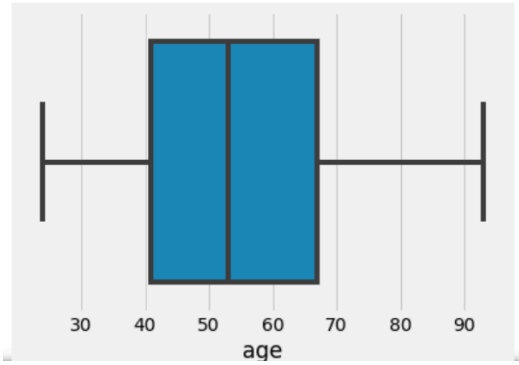


Numeric variables are also explained through histogram & Box-plot .

```
Description of age
--------------------------------------------------------------------------------
count    1517.000000
mean       54.241266
std        15.701741
min        24.000000
25%        41.000000
50%        53.000000
75%        67.000000
max        93.000000
Name: age, dtype: float64
--------------------------------------------------------------------------------
Distribution of age
--------------------------------------------------------------------------------
```




```
BoxPlot of age
--------------------------------------------------------------------------------
```



The output displays, total 8*3 = 24 distinct charts/columns & descriptions. Hence I have put the screenshot of only one variable which is **age** .

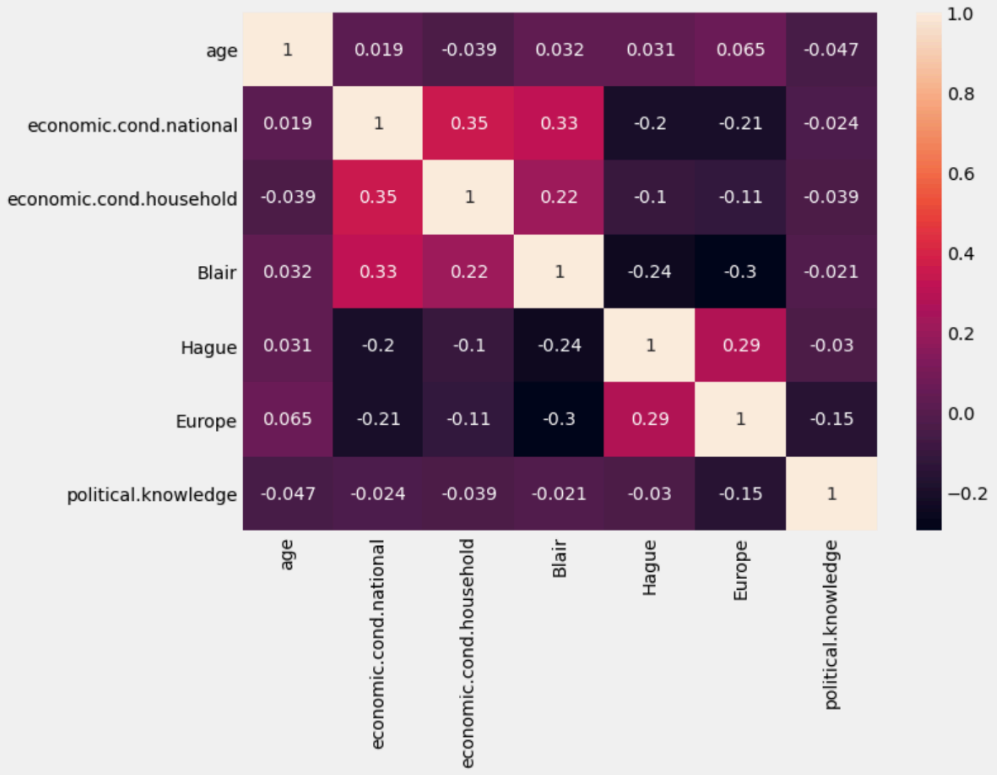
(Please refer the python notebook for better clarity)

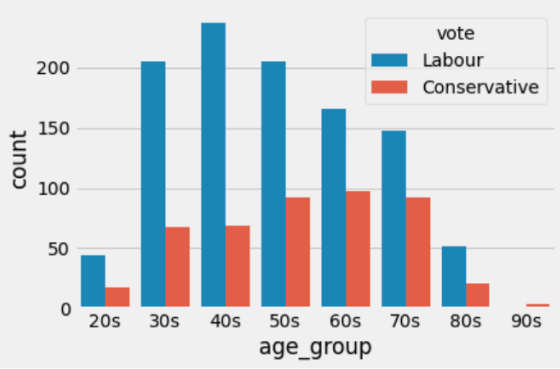
Bi-Variate Analysis

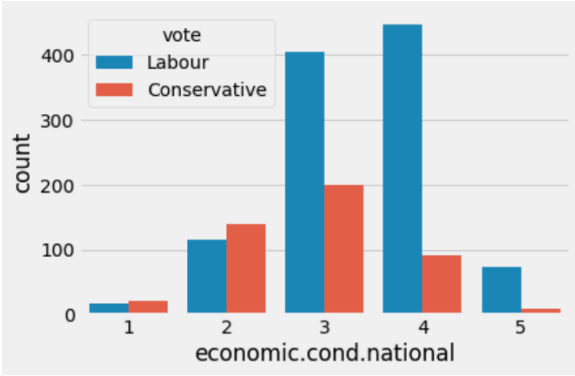
As the name suggest it gives analysis of two specific variable at a time and we can derive correlation b/w them. We can use Heat-map for the same and further to highlight each correlated pair we can use scatter-plot / bar plot etc.

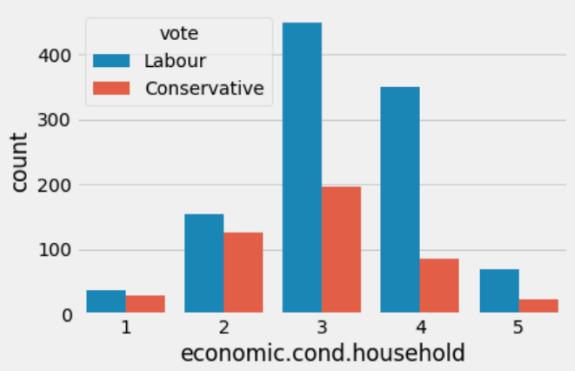
Similar to Univariate , we have done the Bi-variate analysis as well separately for categorical as well as numeric columns.

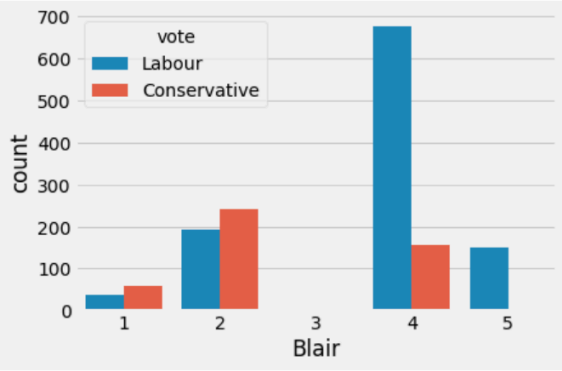
To evaluate numeric columns we have implemented **Heat-Map** and tried analysing further through **Count Plot & Strip Plot**

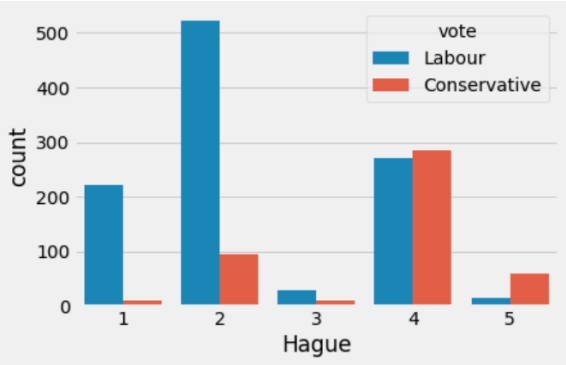


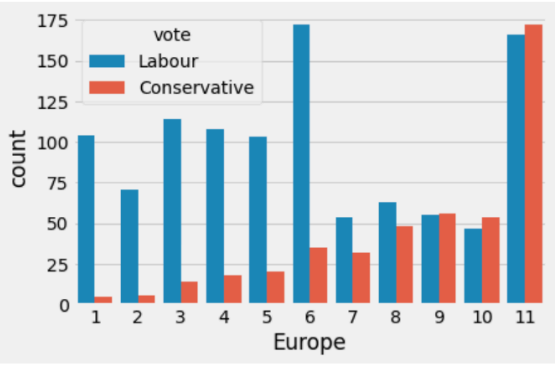


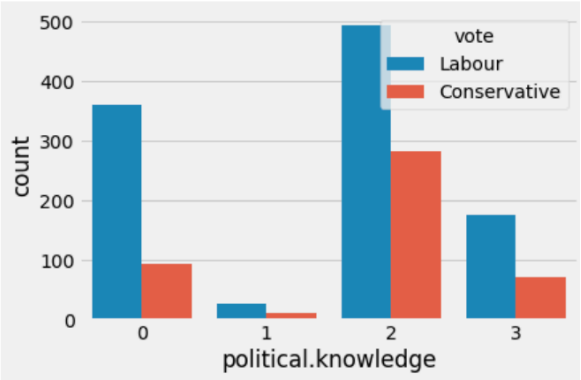


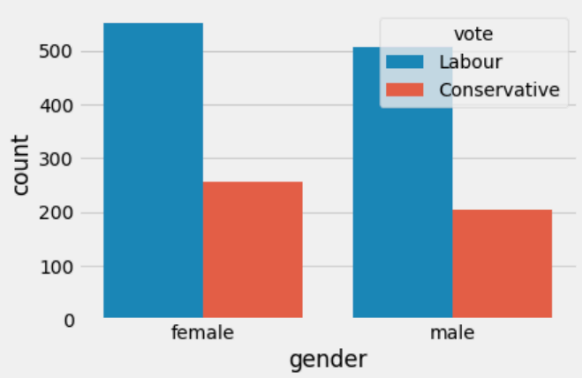

vote
Labour
Conservative
count
economic.cond.household

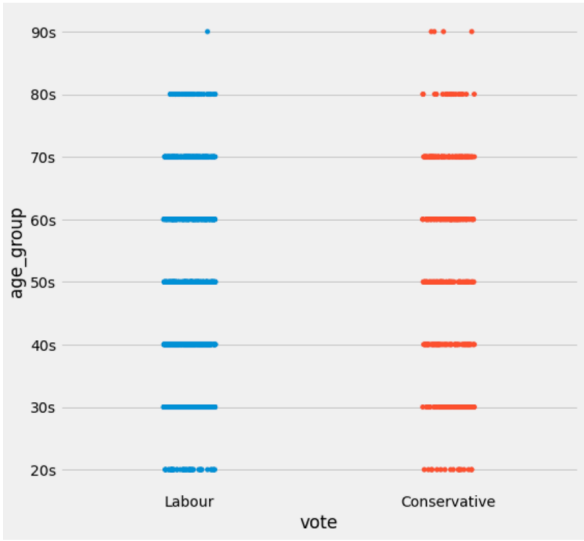

vote
Labour
Conservative
count
Blair

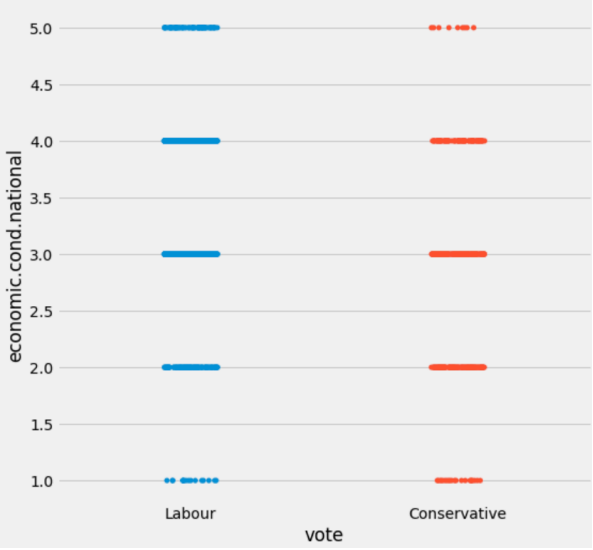

vote
Labour
Conservative
count
Hague

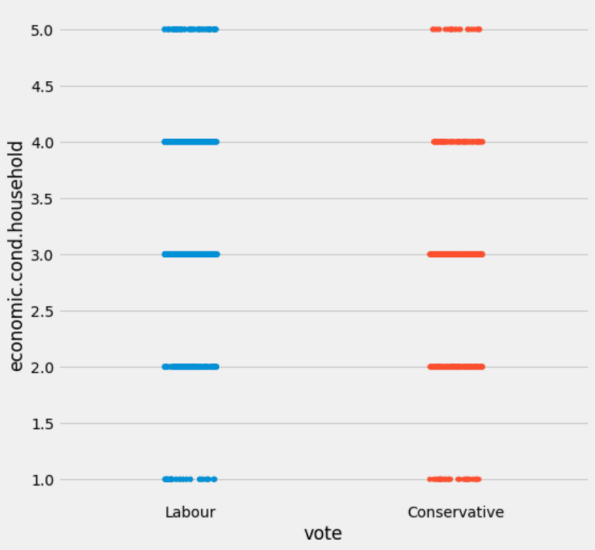

vote
Labour
Conservative
count
Europe

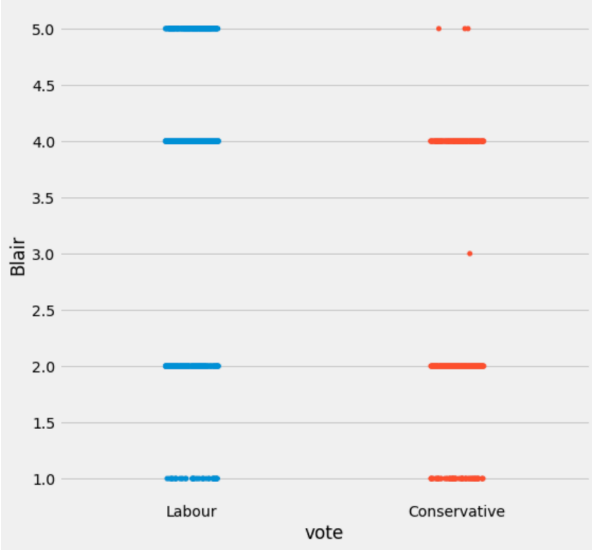

vote
Labour
Conservative
count
political.knowledge

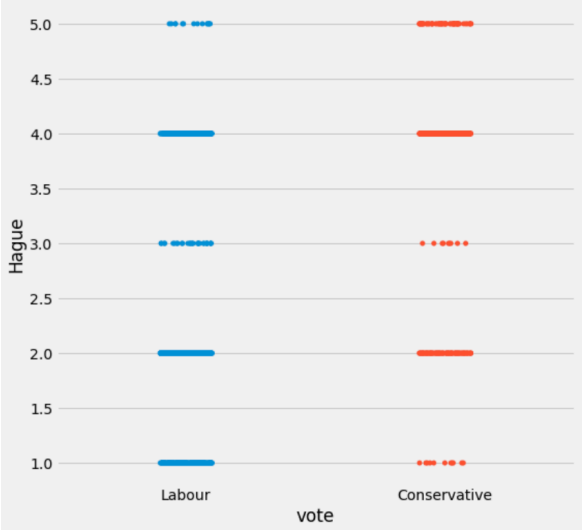

vote
Labour
Conservative
count
female
male
gender

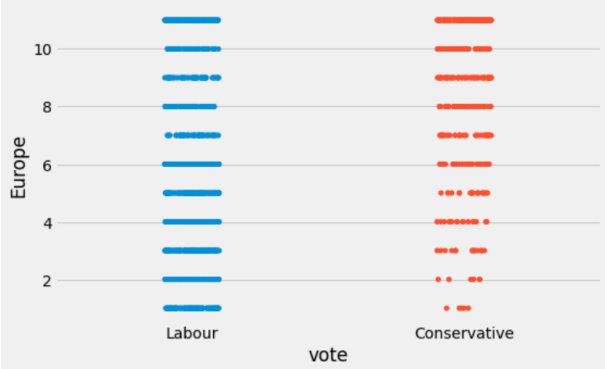

age_group
Labour
Conservative
vote

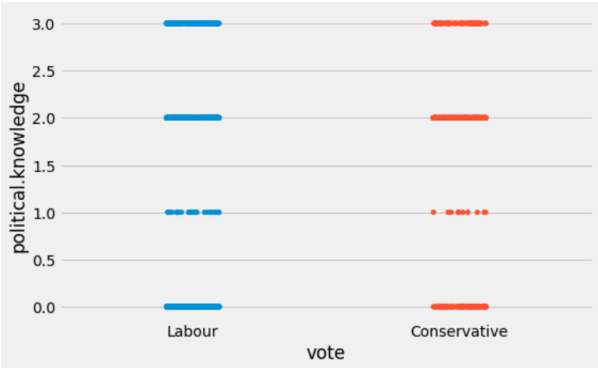

economic.cond.national
Labour
Conservative
vote











In the above graph & charts we tried putting multiple columns corresponding to the target or dependent variable and tried the influence or impact over it .

Checking Pair-Plot corresponding to target variable

In the **Pair-Plot** we wanted to compare all variable's distribution corresponding to the target variable **"Vote"**

Data Preparation: 4 marks

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). (4 Marks)

(Please refer the python notebook for better clarity)

As we know that we have here 2 Categorical Variables which are **'Gender'** & **"Vote"** hence we have done one hot encoding with dropping of first column to avoid multicollinearity
(Please refer the python notebook for better clarity)

Here we can see that Age Group ranges from **20** to **100** and all other variables most of them are Ordinal Variables like rating **['vote', 'economic.cond.national', ' economic.cond.household,'Blair', 'Hague', 'Europe', 'political.knowledge', 'gender']**, So for better understanding and interpretation of the Model we are Doing Binning of Age Column as below .

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender	age_group
0	Labour	43	3	3	4	1	2	2	female	40s
1	Labour	36	4	4	4	4	5	2	male	30s
2	Labour	35	4	4	5	2	3	2	male	30s
3	Labour	24	4	2	2	1	4	0	female	20s
4	Labour	41	2	2	1	1	6	2	male	40s

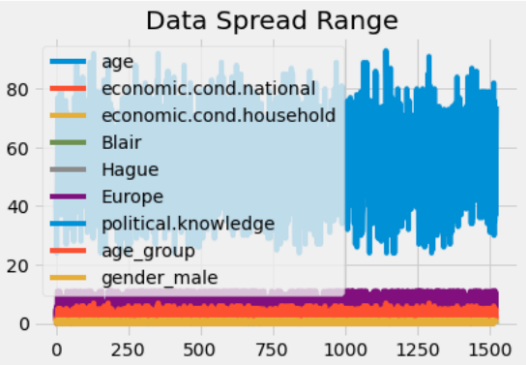
Now we encode Both the columns **"Gender"** & **"Age_group"**. We have done one hot encoding for "Gender" & label encoding for "Age_group".

After encoding the dataset become as below :

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

Is Scaling necessary here or not?

As we can see below Data range graph and can derive that data ranges are being spread between 0 to 100 & most of them are ordinal so it has no meaning to scale the ordinal variables , So we are not doing scaling in this case .



(Please refer the python notebook for better clarity)

Here we have separated our Dependent variable (Vote) from the dataset & we had split the data into train & test of (70:30) ratio by using python train_test_split function by passing below parameters .

For each model we have first trained the model and then test that model so we have split the data into train and test set by passing below parameters into train test split function.

After splitting :

```
X_train:  (1061, 9)
X_test:  (456, 9)
y_train:  (1061,)
y_test:  (456,)
```

Modeling: 22 marks

1.4 Apply Logistic Regression and LDA (linear discriminant analysis). (4 marks)

(Please refer the python notebook for better clarity)

Logistic Regression

We have implemented Logistic Regression by using the following parameter

max_iter=100000
n_jobs=2
random_state=1

We wanted to achieve the best possible outcome from the model hence we applied **GridSearchCV** with following parameters to provide multiple combination and get the best parameter picked among the same.

Best Parameter :

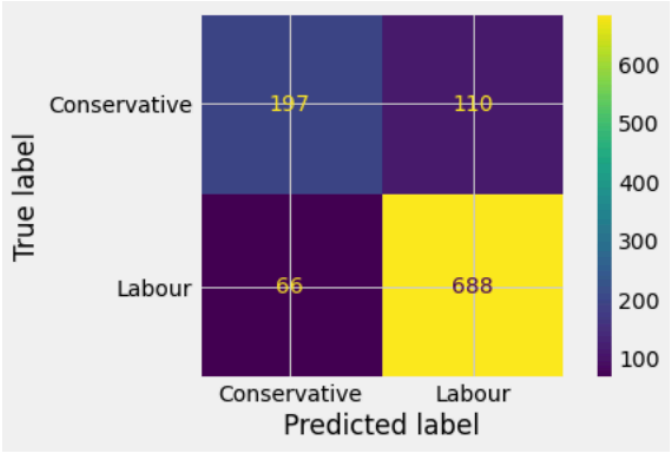
```
'penalty': 'none', 'solver': 'lbfgs', 'tol': 0.0001}
```

Best Estimator

```
LogisticRegression(max_iter=100000, n_jobs=2, penalty='none',
random_state=1)
```

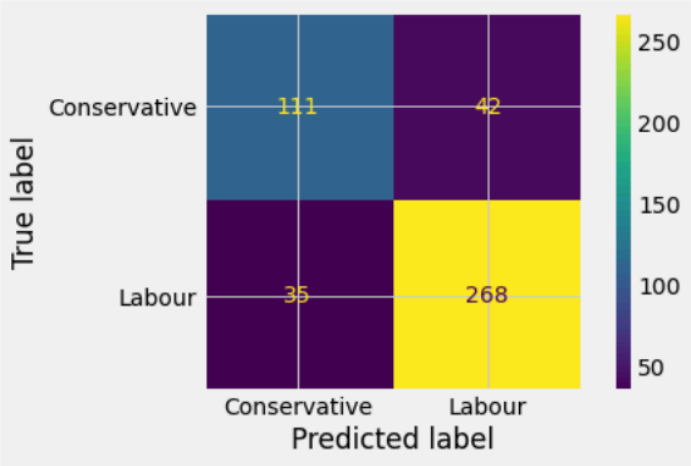
Confusion matrix & Classification Report on train data

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



Confusion matrix & Classification Report on test data

	precision	recall	f1-score	support
Conservative	0.76	0.73	0.74	153
Labour	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



Train Accuracy , Train_recall_labour, Train_recall_conservative

0.8331762488218661 , 0.6416938110749185 , 0.9124668435013262

Test Accuracy , Test_recall_labour, Test_recall_conservative

0.831140350877193 0.7254901960784313 0.8844884488448845

Linear Discriminant Analysis

We have implemented **LDA** by using the following parameter

params={'solver':['lsqr','eigen'],
'n_components':[1,7,2]}

We wanted to achieve the best possible outcome from the model hence we applied **GridSearchCV** with following parameters to provide multiple combination and get the best parameter picked among the same.

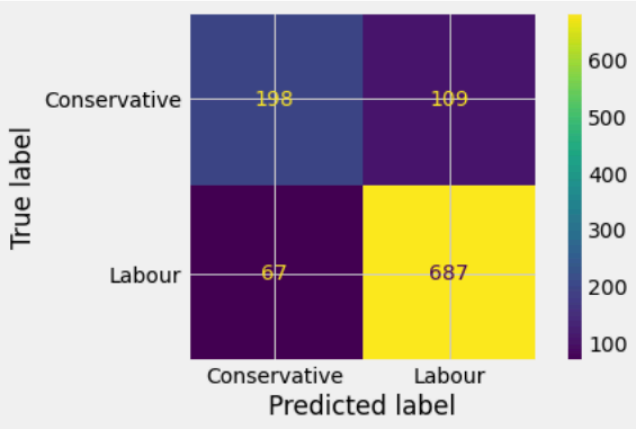
Best Parameter :
`{'n_components': 1, 'solver': 'lsqr'}`

Best Estimator

LinearDiscriminantAnalysis(n_components=1, solver='lsqr')

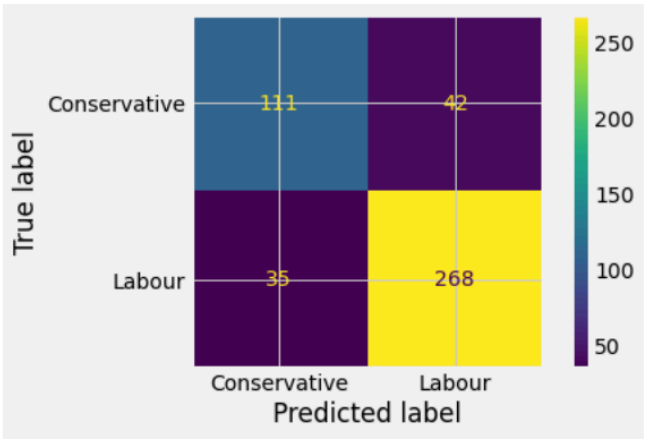
Confusion matrix & Classification Report on train data

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



Confusion matrix & Classification Report on test data

	precision	recall	f1-score	support
Conservative	0.76	0.73	0.74	153
Labour	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



In terms of deriving the model's result & comparing them we see in **Logistic Regression & Linear Discriminant Analysis** performed quite well.
In terms of comparing the train & test data we can feel that both looks like a good fit model but still corresponding to the classification report on the test data Logistic Regression slightly performed better . Hence I am choosing Logistic Regression is my best model b/w these two.

Train Accuracy , Train_recall_labour, Train_recall_conservative

0.8341187558906692 0.6449511400651465 0.9111405835543767

Test Accuracy , Test_recall_labour, Test_recall_conservative

0.831140350877193 0.7254901960784313 0.8844884488448845

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results. (4 marks)

(Please refer the python notebook for better clarity)

KNN Model

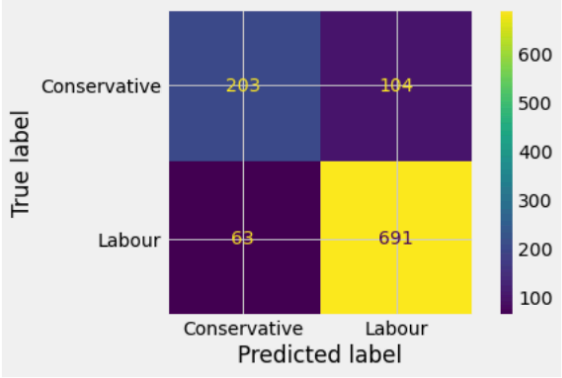
We have implemented **KNN** by using the following parameter

n_neighbors=7

As we know in KNN model initially we have to assume a number which is optimal nearest neighbour and later we can fine tune the same by finding minimum **MCE** or **Misclassification Error**

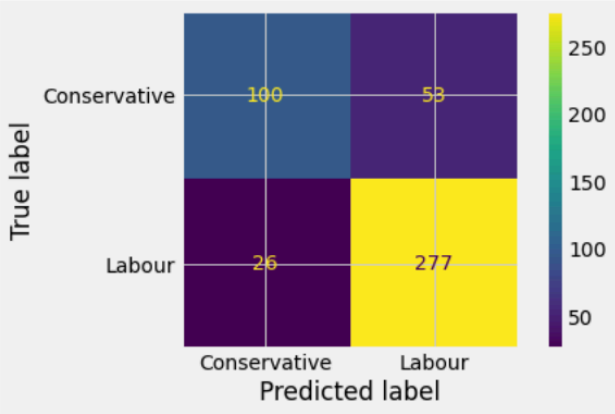
Confusion matrix & Classification Report on train data

	precision	recall	f1-score	support
Conservative	0.76	0.66	0.71	307
Labour	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



Confusion matrix & Classification Report on test data

	precision	recall	f1-score	support
Conservative	0.79	0.65	0.72	153
Labour	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



Train Accuracy , Train_recall_labour, Train_recall_conservative

0.8426013195098964 0.6612377850162866 0.916445623342175

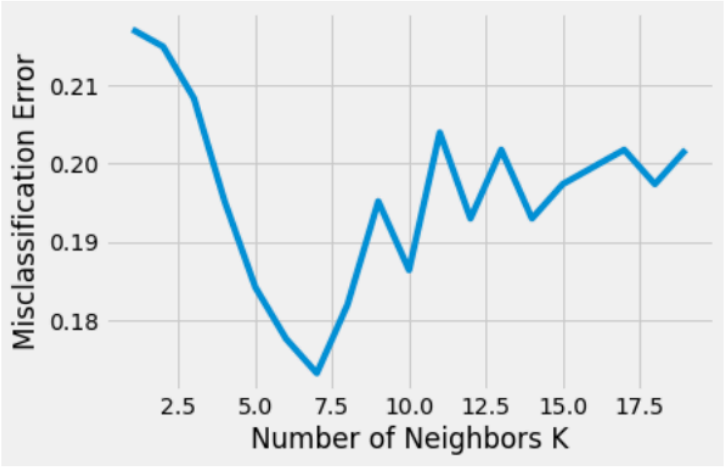
Test Accuracy , Test_recall_labour, Test_recall_conservative

0.8267543859649122 0.6535947712418301 0.9141914191419142

Tuning KNN Model with MCE Evaluation (k values from 1-20 with 2 interval & last number exclusive)

```
[0.2171052631578947,
 0.2149122807017544,
 0.20833333333333337,
 0.19517543859649122,
 0.1842105263157895,
 0.17763157894736847,
 0.17324561403508776,
 0.18201754385964908,
 0.19517543859649122,
 0.1864035087719298,
 0.20394736842105265,
 0.19298245614035092,
 0.20175438596491224,
 0.19298245614035092,
 0.19736842105263153,
 0.19956140350877194,
 0.20175438596491224,
 0.19736842105263153,
 0.20175438596491224]
```

Plotting misclassification error (MCE) vs k



We can see from the above graph that when k value 7 , then only the misclassification error is minimal, hence we can conclude saying that K=7 is the optimal number of nearest neighbour.

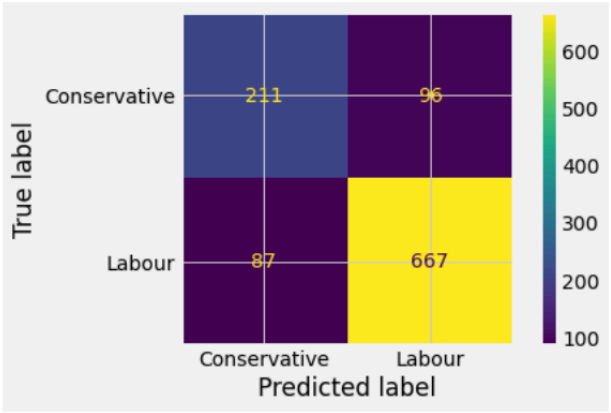
NAIVE BAYES

We have implemented **Naive Bayes Model** by using the following parameter

```
GNB_model = GaussianNB()
GNB_model.fit(X_train, y_train)
```

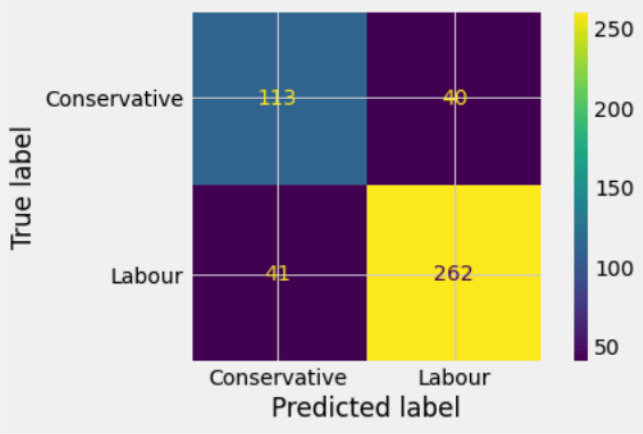
Confusion matrix & Classification Report on train data

	precision	recall	f1-score	support
Conservative	0.71	0.69	0.70	307
Labour	0.87	0.88	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



Confusion matrix & Classification Report on test data

	precision	recall	f1-score	support
Conservative	0.73	0.74	0.74	153
Labour	0.87	0.86	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



Train Accuracy , Train_recall_labour, Train_recall_conservative

0.827521206409048 0.6872964169381107 0.8846153846153846

Test Accuracy , Test_recall_labour, Test_recall_conservative

0.8223684210526315 0.738562091503268 0.8646864686468647

In terms of deriving the model's result & comparing them we see in **KNN & Naive Bayes** performed quite well.

In terms of comparing the train & test data we can see that Naive Bayes model's classification report based on test data is better than KNN , as recall labour is pretty low than Naive Bayes , But in Naive Bayes including accuracy both recall value (Labour & Conservative) returns a better value.

1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. (7 marks)

(Please refer the python notebook for better clarity)

In the above implementation we tried using **GridSearchCv** & **CrossValidation** to get the best possible outcome in one shot, hence I can consider above model are implemented in a tuned way .

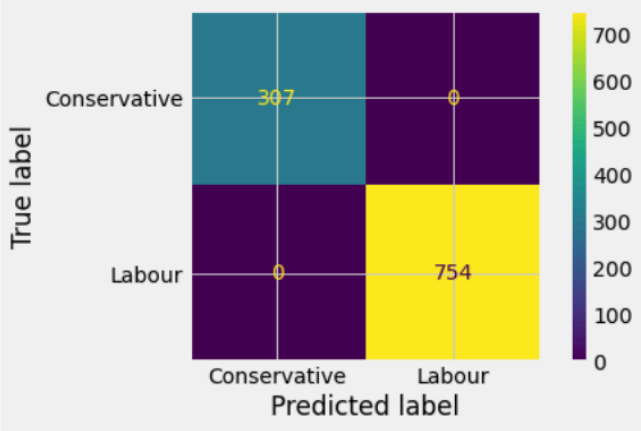
Random Forest

We have implemented **RF** by using the following parameter

```
RF_model=RandomForestClassifier(n_estimators=100,random_state=1)
RF_model.fit(X_train, y_train)
```

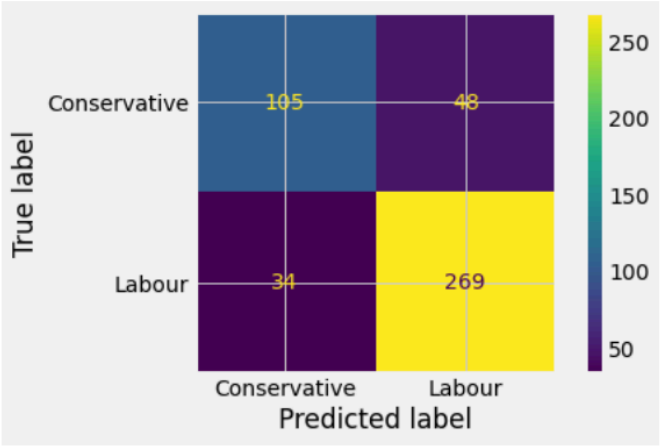
Confusion matrix & Classification Report on train data

	precision	recall	f1-score	support
Conservative	1.00	1.00	1.00	307
Labour	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



Confusion matrix & Classification Report on train data

	precision	recall	f1-score	support
Conservative	0.76	0.69	0.72	153
Labour	0.85	0.89	0.87	303
accuracy			0.82	456
macro avg	0.80	0.79	0.79	456
weighted avg	0.82	0.82	0.82	456



Train Accuracy , Train_recall_labour, Train_recall_conservative

```
1.0 1.0 1.0
```

Test Accuracy , Test_recall_labour, Test_recall_conservative

```
0.8201754385964912 0.6862745098039216 0.8877887788778878
```

Looking at the **RF** model we can see that this is an overfitted model hence instead of rely on this model we decided to implement **Bagging on RF**

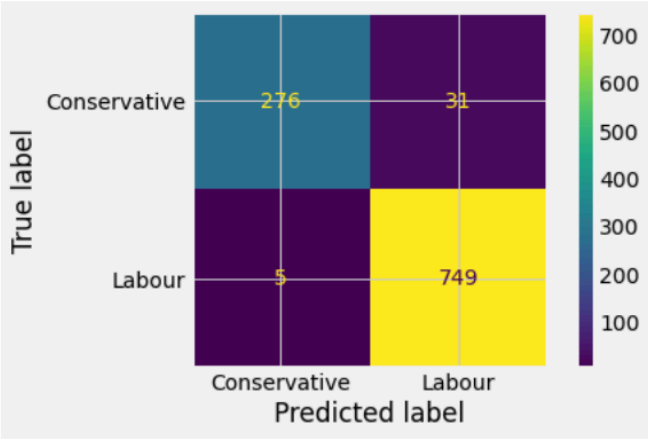
Bagging with RF

We have implemented **Bagging** by using the following parameter

```
Bagging_model=BaggingClassifier(base_estimator=RF_model,n_es
timators=100,random_state=1)
Bagging_model.fit(X_train, y_train)
```

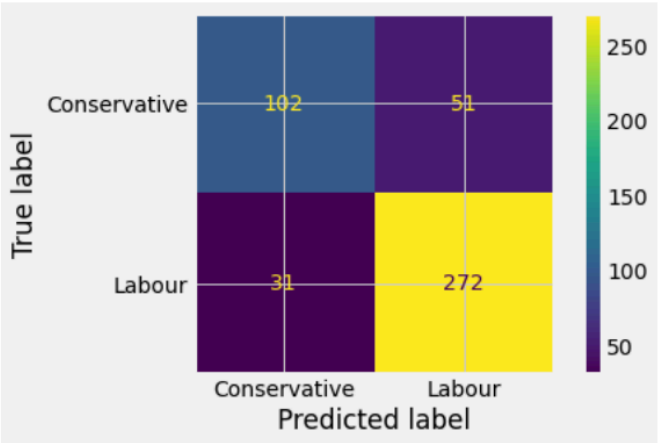
Confusion matrix & Classification Report on train data

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



Confusion matrix & Classification Report on test data

	precision	recall	f1-score	support
Conservative	0.77	0.67	0.71	153
Labour	0.84	0.90	0.87	303
accuracy			0.82	456
macro avg	0.80	0.78	0.79	456
weighted avg	0.82	0.82	0.82	456



Train Accuracy , Train_recall_labour, Train_recall_conservative

```
0.9660697455230914 0.8990228013029316 0.993368700265252
```

Test Accuracy , Test_recall_labour, Test_recall_conservative

```
0.8201754385964912 0.6666666666666666 0.8976897689768977
```

From the test recall for labour we can see a drastic drop , hence this will also be considered as overfitted model.

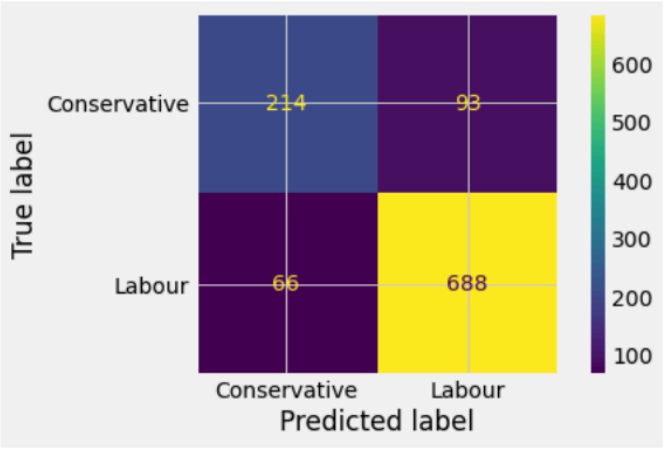
Ada-Boost Model

We have implemented **Ada-Boost** by using the following parameter

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

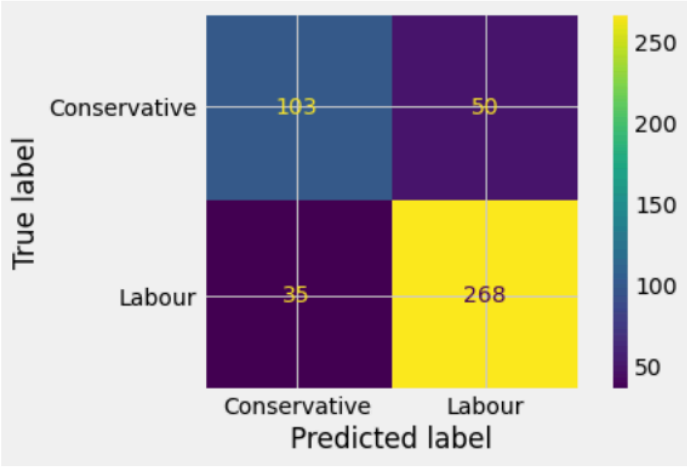
Confusion matrix & Classification Report on train data

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



Confusion matrix & Classification Report on test data

	precision	recall	f1-score	support
Conservative	0.75	0.67	0.71	153
Labour	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



Train Accuracy , Train_recall_labour, Train_recall_conservative

0.8501413760603205 0.6970684039087948 0.9124668435013262

Test Accuracy , Test_recall_labour, Test_recall_conservative

0.8135964912280702 0.673202614379085 0.8844884488448845

This looks like a good fit model but still we should look for a better model as recall for labour is very low in train & test both report .

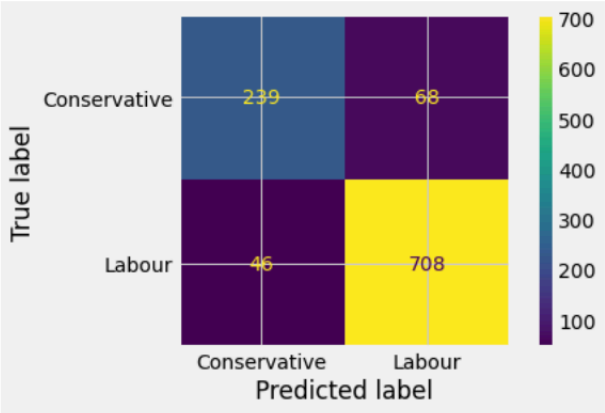
Gradient Boost Model

We have implemented **Gradient Boost** by using the following parameter

```
GBC_model = GradientBoostingClassifier(random_state=1)
GBC_model = GBC_model.fit(X_train, y_train)
```

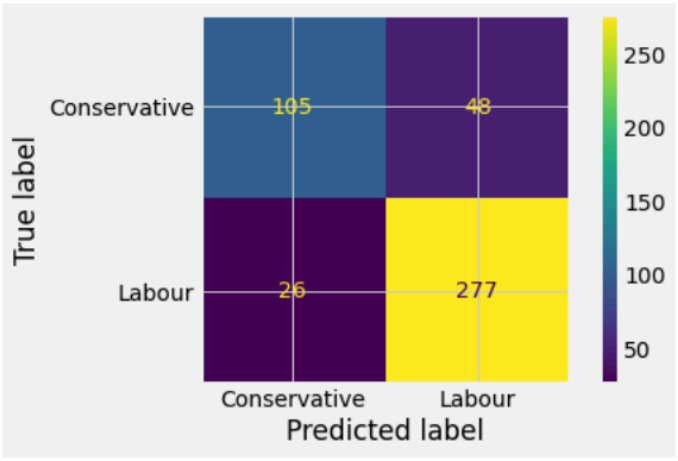
Confusion matrix & Classification Report on train data

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



Confusion matrix & Classification Report on test data

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



Train Accuracy , Train_recall_labour, Train_recall_conservative

0.8925541941564562 0.7785016286644951 0.9389920424403183

Test Accuracy , Test_recall_labour, Test_recall_conservative

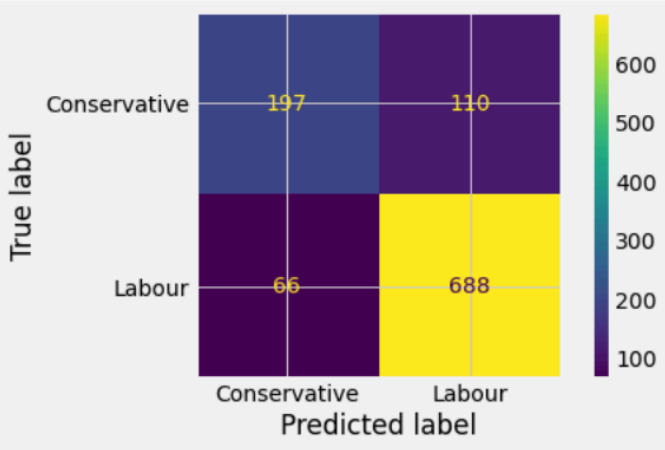
0.8377192982456141 0.6862745098039216 0.9141914191419142

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/optimised. (7 marks)

Logistic Regression

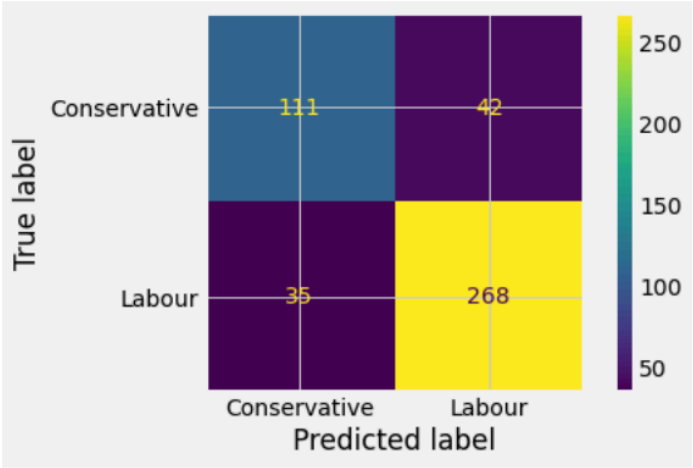
Confusion matrix & Classification Report on train data

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

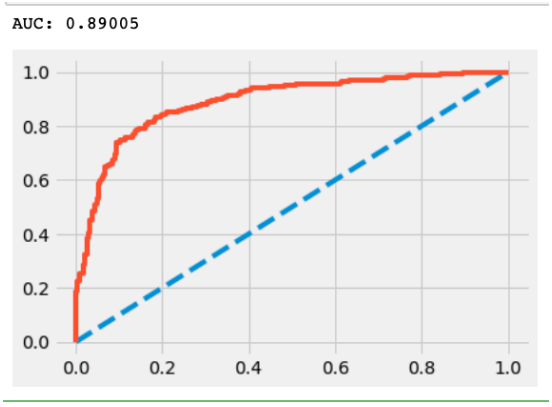


Confusion matrix & Classification Report on train data

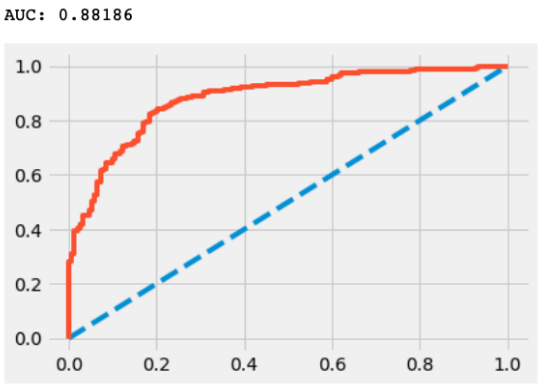
	precision	recall	f1-score	support
Conservative	0.76	0.73	0.74	153
Labour	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456



AUC and ROC of the train & test dataset



Train Data

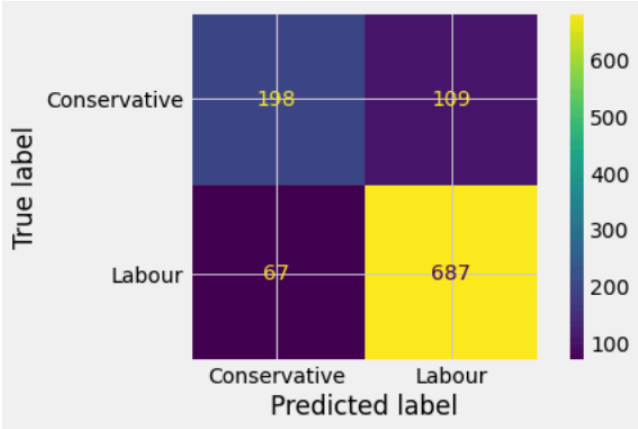


Test Data

Linear Discriminant Analysis

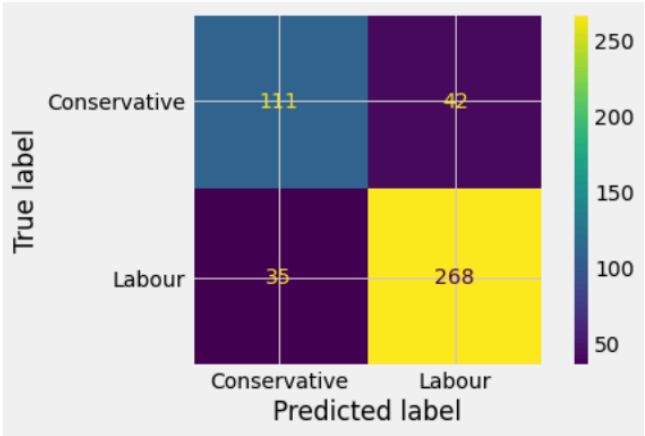
Confusion matrix & Classification Report on train data

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



Confusion matrix & Classification Report on test data

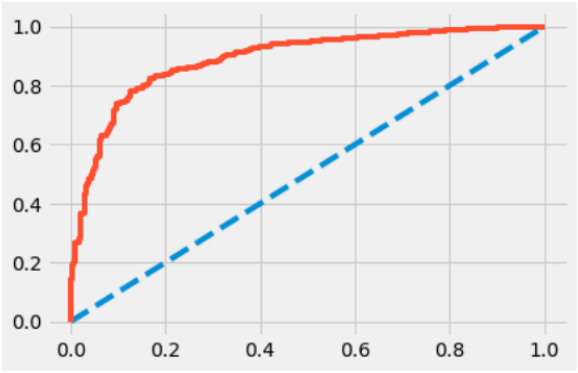
	precision	recall	f1-score	support
Conservative	0.76	0.73	0.74	153
Labour	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456



AUC and ROC of the train & test dataset

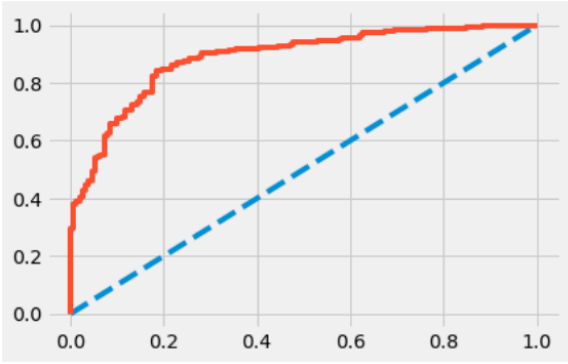
Train Data

AUC: 0.88961



Test Data

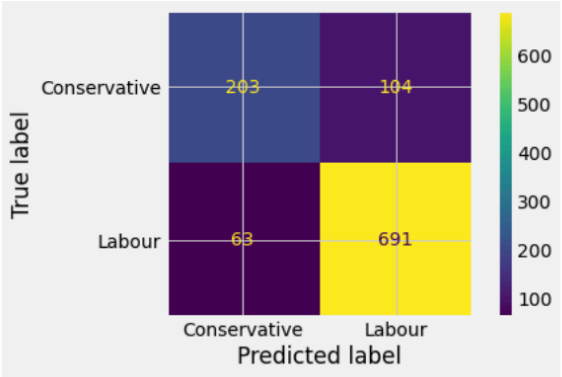
AUC: 0.88706



KNN

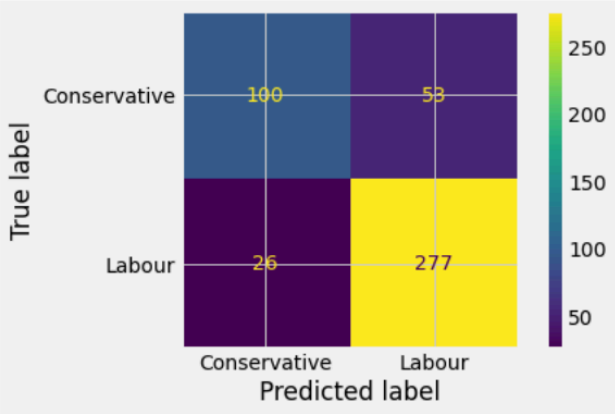
Confusion matrix & Classification Report on train data

	precision	recall	f1-score	support
Conservative	0.76	0.66	0.71	307
Labour	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



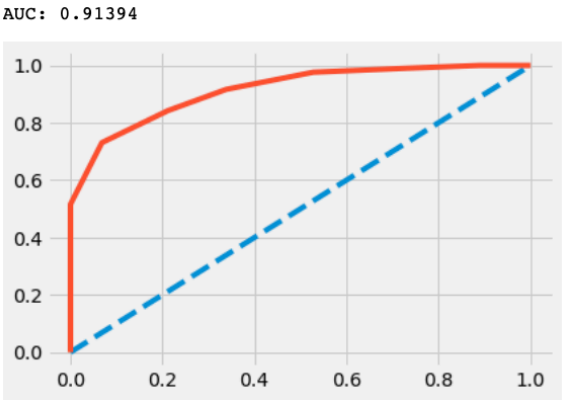
Confusion matrix & Classification Report on test data

	precision	recall	f1-score	support
Conservative	0.79	0.65	0.72	153
Labour	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

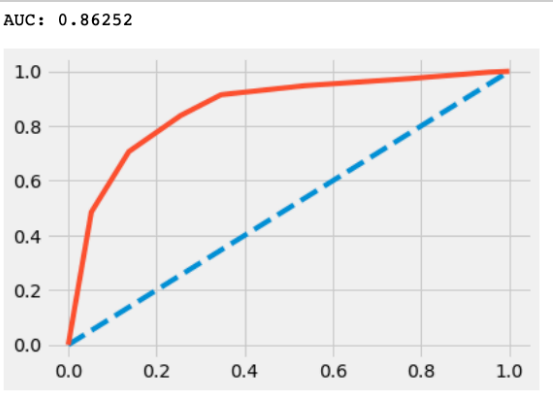


AUC and ROC of the train & test dataset

Train Data



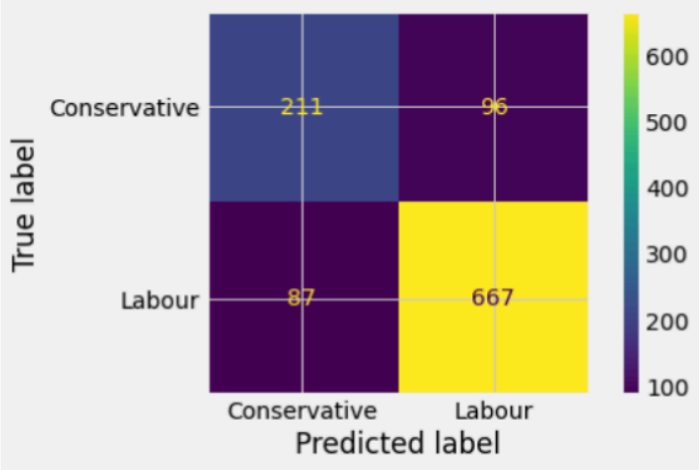
Test Data



NAIVE BAYES

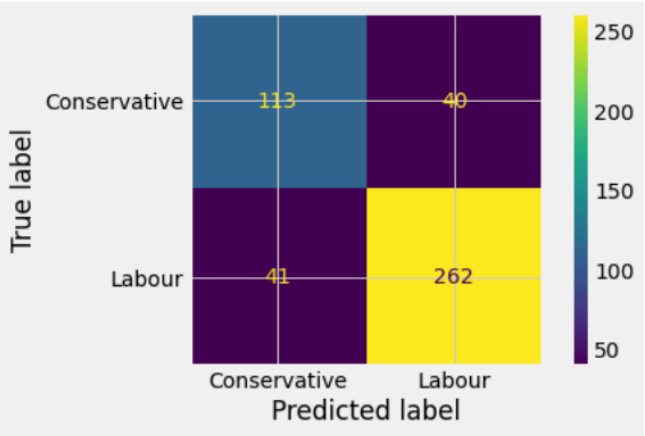
Confusion matrix & Classification Report on train data

	precision	recall	f1-score	support
Conservative	0.71	0.69	0.70	307
Labour	0.87	0.88	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



Confusion matrix & Classification Report on test data

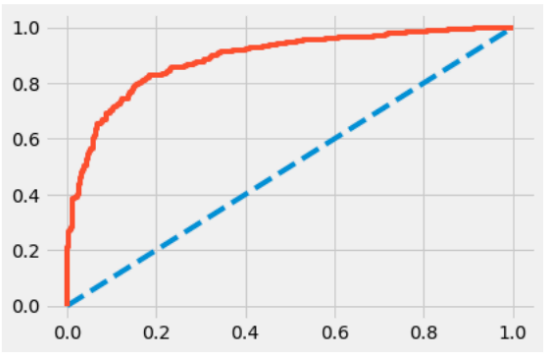
	precision	recall	f1-score	support
Conservative	0.73	0.74	0.74	153
Labour	0.87	0.86	0.87	303
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456



AUC and ROC of the train & test dataset

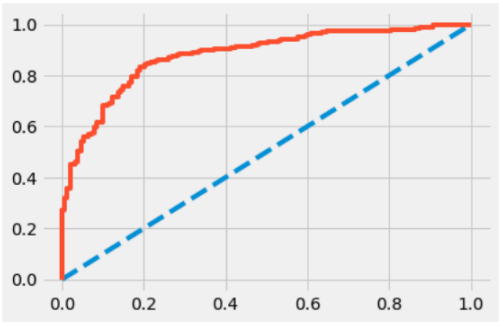
Train Data

AUC: 0.88807



Test Data

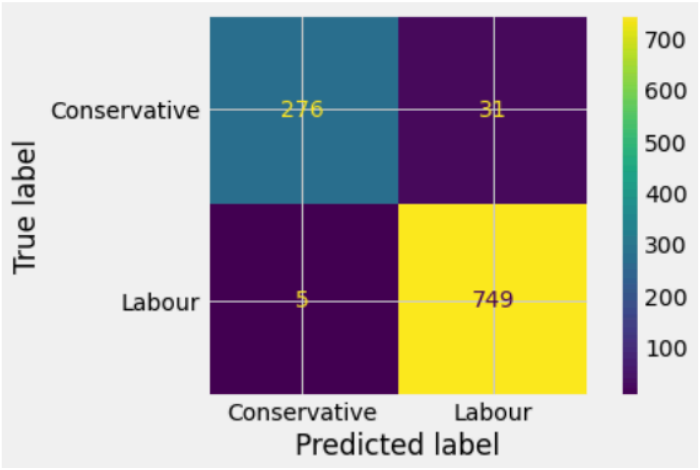
AUC: 0.87955



Bagging with RF

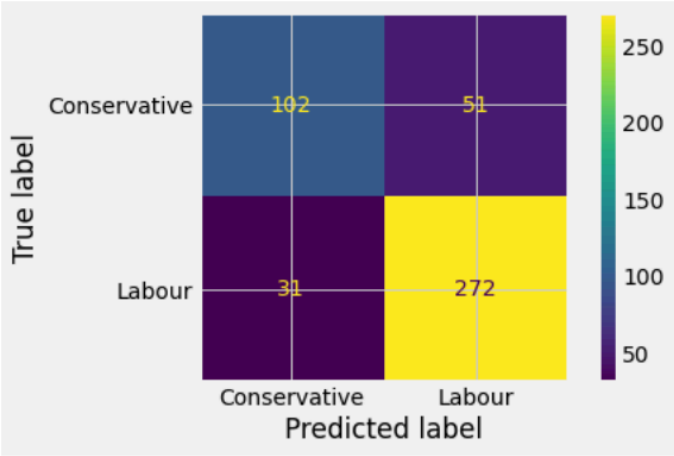
Confusion matrix & Classification Report on train data

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



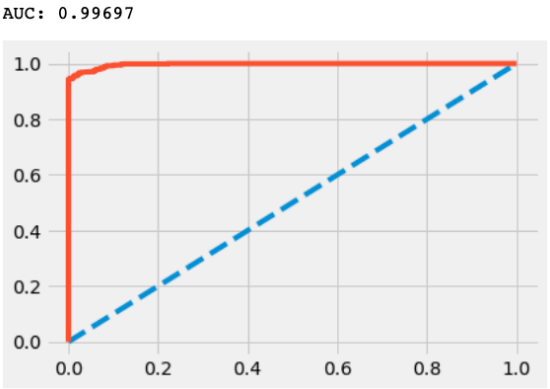
Confusion matrix & Classification Report on test data

	precision	recall	f1-score	support
Conservative	0.77	0.67	0.71	153
Labour	0.84	0.90	0.87	303
accuracy			0.82	456
macro avg	0.80	0.78	0.79	456
weighted avg	0.82	0.82	0.82	456

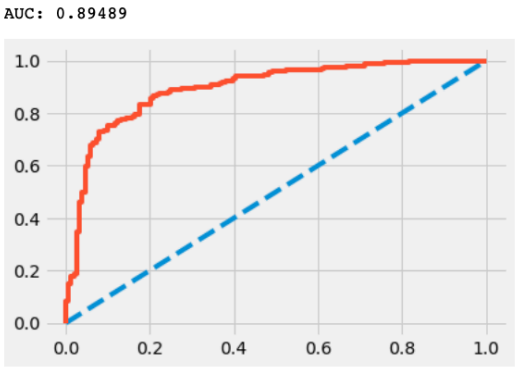


AUC and ROC of the train & test dataset

Train Data



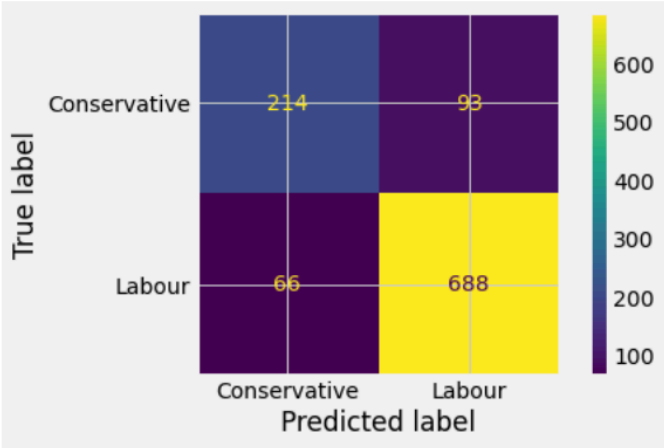
Test Data



Ada-Boost

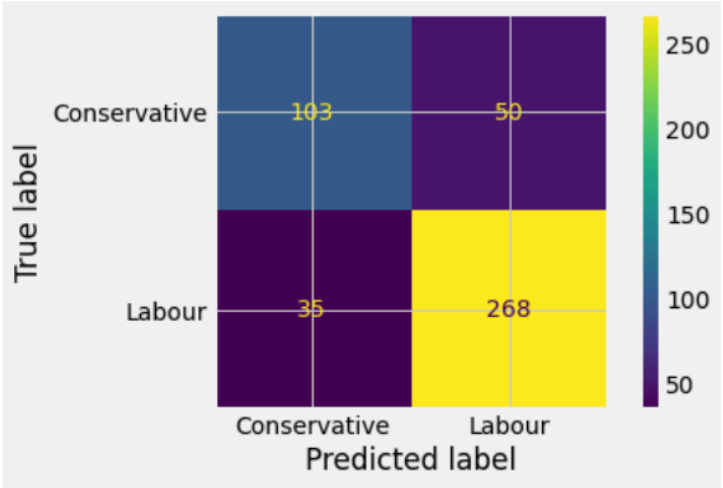
Confusion matrix & Classification Report on train data

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



Confusion matrix & Classification Report on test data

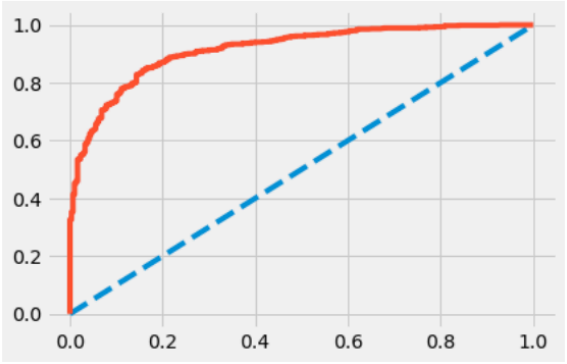
	precision	recall	f1-score	support
Conservative	0.75	0.67	0.71	153
Labour	0.84	0.88	0.86	303
accuracy			0.81	456
macro avg	0.79	0.78	0.79	456
weighted avg	0.81	0.81	0.81	456



AUC and ROC of the train & test dataset

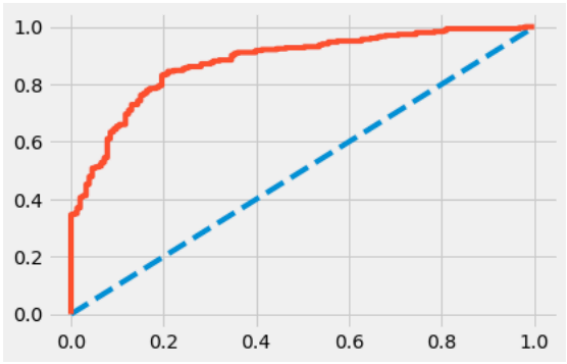
Train Data

AUC: 0.91481



Test Data

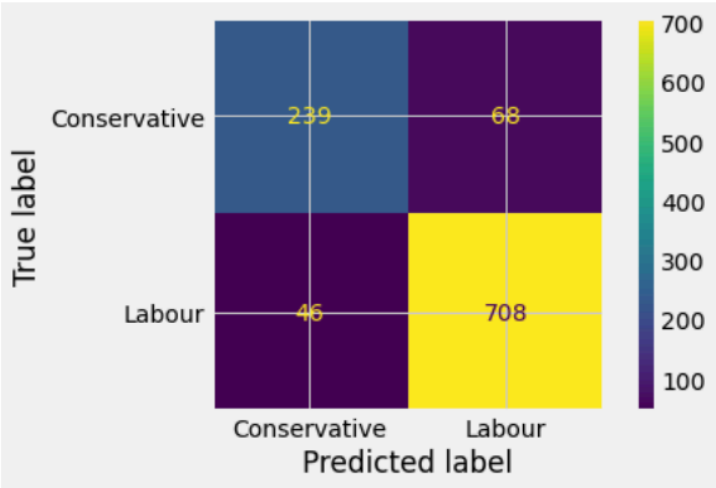
AUC: 0.87738



Gradient-Boost

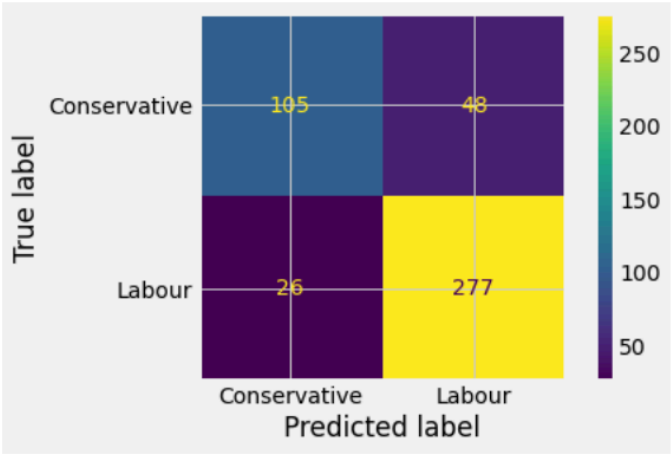
Confusion matrix & Classification Report on train data

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



Confusion matrix & Classification Report on test data

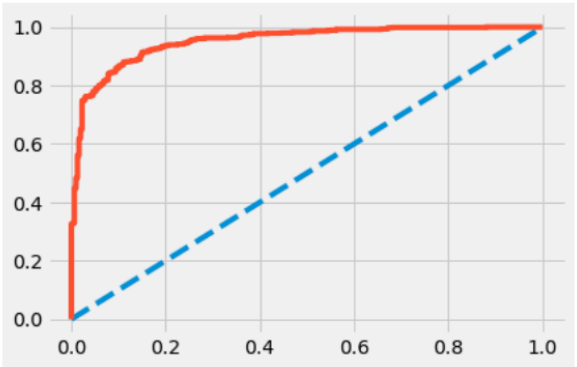
	precision	recall	f1-score	support
Conservative	0.80	0.69	0.74	153
Labour	0.85	0.91	0.88	303
accuracy			0.84	456
macro avg	0.83	0.80	0.81	456
weighted avg	0.84	0.84	0.83	456



AUC and ROC of the train & test dataset

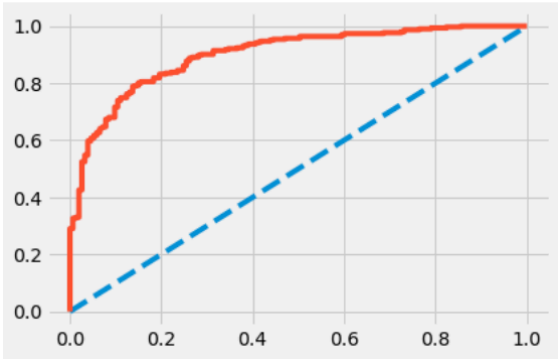
Train Data

AUC: 0.95116



Test Data

AUC: 0.89908



Comparison of Different Models

	Train Recall	Test Recall	Accuracy Train	Accuracy Test
Naive-Bayes	0.884615	0.864686	0.827521	0.822368
LDA	0.911141	0.884488	0.834119	0.831140
ADABoost	0.912467	0.884488	0.850141	0.813596
GradientBoost	0.938992	0.914191	0.892554	0.837719
KNN	0.916446	0.914191	0.842601	0.826754
LR	0.912467	0.884488	0.833176	0.831140
RF	1.000000	0.887789	1.000000	0.820175
Bagging	0.993369	0.897690	0.966070	0.820175

So as per the above test data, best performing model is - KNN . Though we can see that **GRADIENT BOOST** has better value still overall model wise **KNN** model's prediction is much more accurate .

Best Performing models are - **GRADIENT BOOST, KNN , LDA & LR**

We have observed that **KNN** & **Naive Bayes** train &test data are almost equal.

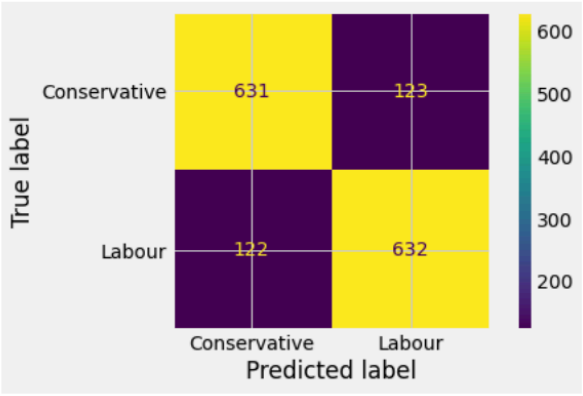
Hence we decided to evaluate the same after implementing **SMOTE** , as we found that target variable's ratio distribution is not uniform .

(Please refer notebook for better clarity)

Naive Bayes After SMOTE

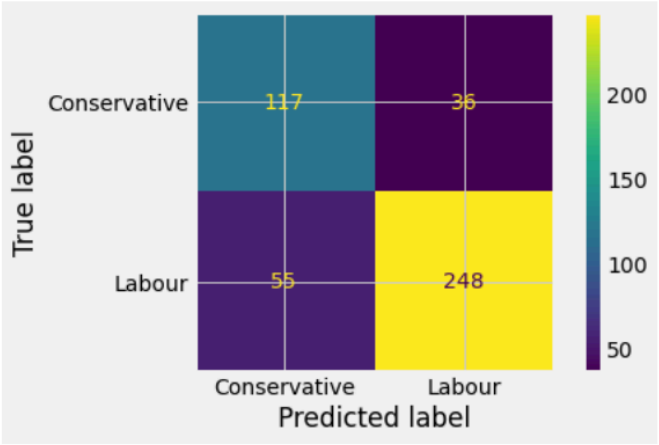
Confusion matrix & Classification Report on train data

	precision	recall	f1-score	support
Conservative	0.84	0.84	0.84	754
Labour	0.84	0.84	0.84	754
accuracy			0.84	1508
macro avg	0.84	0.84	0.84	1508
weighted avg	0.84	0.84	0.84	1508



Confusion matrix & Classification Report on test data

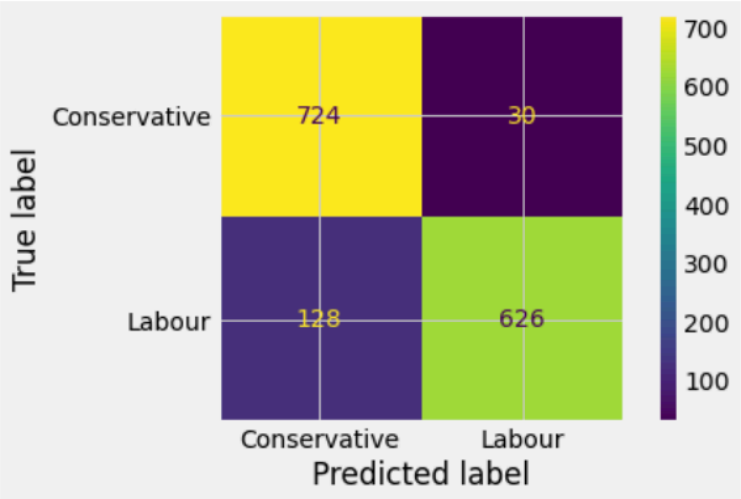
	precision	recall	f1-score	support
Conservative	0.68	0.76	0.72	153
Labour	0.87	0.82	0.84	303
accuracy			0.80	456
macro avg	0.78	0.79	0.78	456
weighted avg	0.81	0.80	0.80	456



KNN After SMOTE

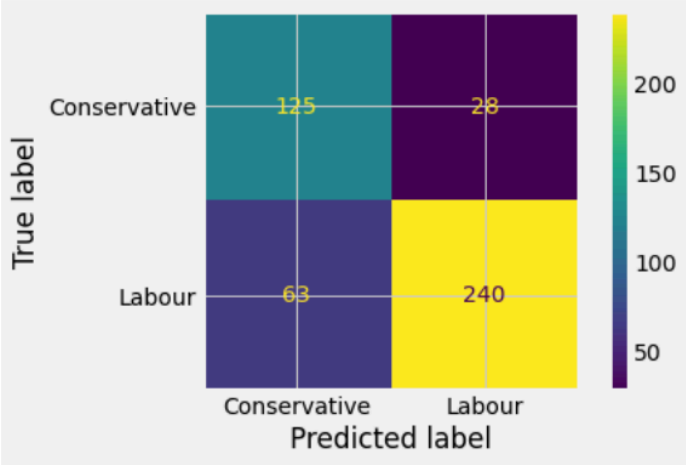
Confusion matrix & Classification Report on train data

	precision	recall	f1-score	support
Conservative	0.85	0.96	0.90	754
Labour	0.95	0.83	0.89	754
accuracy			0.90	1508
macro avg	0.90	0.90	0.89	1508
weighted avg	0.90	0.90	0.89	1508



Confusion matrix & Classification Report on test data

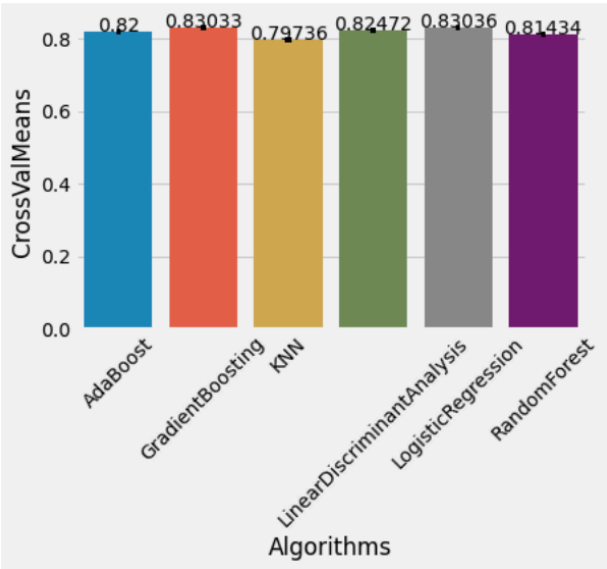
	precision	recall	f1-score	support
Conservative	0.66	0.82	0.73	153
Labour	0.90	0.79	0.84	303
accuracy			0.80	456
macro avg	0.78	0.80	0.79	456
weighted avg	0.82	0.80	0.80	456



Model building is an iterative process. Model performance both on the test and train dataset can be improved using feature engineering, feature extraction, hyper parameter tuning (including combination of various parameters).

Model has to match the business objective and hence various permutation and combinations can be tried on to refine the model.Hence we have tried multiple permutation combination to validate model's performance.

Here below we wanted to show a comparison graph based on the way above models are performed.



1.8 Based on these predictions, what are the insights? (5 marks)

As per the problem statement we have to predict which party a voter will vote on the basis of the given information. To solve the same we have made Logistic Regression , Linear Discriminant Analysis Model, k- Nearest neighbour & Naïve Bays Model to find out the best results Gradient Boost is performing slightly better than other models in terms of accuracy .

As we have seen from heat map that there is very less correlation between the variables which is good for the model.

Observations

- Voters between Age group of 30s -70s are voting more.
- Voters in their 20s & 80s, 90s are voting significantly less
- Significantly More no. of females have voted for Labour party
- Most of the people, who have Euro-sceptic attitudes given vote to the Labour Party.
- Labour party has more votes than Conservative party

Recommendations

- We have to gather more data like ratings on their previous leadership qualities (How they have performed previously), Religion of the respondent etc. would certainly help to gain more impactful insights .
- CNBE can conduct Online surveys by which they can reduce their actual cost on surveys in result they can gather more data in much more easier way.
- CNBE can also give free Magazines, ebooks etc. or online Coupons to the voters if they participate in surveys to engage more people .
- Company should collect the ratings based on the performance of the leader towards Current issues & liabilities.

Problem 2:

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

- **President Franklin D. Roosevelt in 1941**
- **President John F. Kennedy in 1961**
- **President Richard Nixon in 1973**

(Hint: use .words(), .raw(), .sent() for extracting counts)

2.1 Find the number of characters, words, and sentences for the mentioned documents. – 3 Marks

(Please refer the python notebook for better clarity)

Number of **characters** in each Speech are as below :

Number of character in Roosevelt file : **7571**
Number of character in Kennedy file : **7618**
Number of character in Nixon file : **9991**

Number of **words** in each Speech are as below :

Number of words in Roosevelt file : **1536**
Number of words in Kennedy file : **1546**
Number of words in Nixon file : **2028**

Number of **sentences** in each Speech are as below :

Number of Sentences in Roosevelt Speech : **68**
Number of Sentences in Kennedy Speech : **52**
Number of Sentences in Nixon Speech : **69**

2.2 Remove all the stop-words from all three speeches. – 3 Marks

(Please refer the python notebook for better clarity)

Removal of stop-words from **1941-Roosevelt Speech**

Initially we found **1360 words** in 1941-Rossevelt Speech , out of which **654** were stop-words.

	Text	word_count	stop_count	stop	upper
0	On each national day of inauguration since 178...	1360	654	[each, of, the, have, their, of, to, the, the,...	3

Below are the list of stop-words found from the above text :

```
['each', 'of', 'the', 'have', 'their', 'of', 'to', 'the', 'the', 'of', 'the', 'was', 'to', 'and', 'a', 'the', 'of',
'the', 'was', 'to', 'that', 'from', 'from', 'this', 'the', 'of', 'the', 'is', 'to', 'that', 'and', 'its', 'from',
'from', 'there', 'has', 'a', 'in', 'the', 'of', 'to', 'for', 'a', 'and', 'to', 'what', 'our', 'in', 'has', 'and', 'to',
'what', 'we', 'are', 'and', 'what', 'we', 'we', 'do', 'we', 'the', 'of', 'of', 'are', 'not', 'by', 'the', 'of', 'but',
'by', 'the', 'of', 'the', 'of', 'a', 'is', 'and', 'a', 'a', 'of', 'a', 'is', 'the', 'of', 'the', 'of', 'its', 'will',
'to', 'are', 'who', 'are', 'who', 'that', 'as', 'a', 'of', 'and', 'a', 'of', 'is', 'or', 'by', 'a', 'of', 'and', 'for',
'some', 'and', 'have', 'the', 'of', 'the', 'and', 'that', 'is', 'an', 'we', 'that', 'this', 'is', 'not', 'when', 'the',
'of', 'this', 'by', 'a', 'we', 'that', 'this', 'is', 'not', 'were', 'in', 'the', 'of', 'but', 'we', 'have', 'been',
'for', 'the', 'of', 'this', 'they', 'have', 'to', 'a', 'that', 'are', 'to', 'be', 'in', 'other', 'than', 'to', 'our',
'and', 'our', 'is', 'this', 'of', 'a', 'which', 'at', 'on', 'through', 'it', 'the', 'of', 'its', 'has', 'been', 'the',
'of', 'the', 'of', 'the', 'of', 'the', 'to', 'of', 'of', 'is', 'of', 'the', 'of', 'have', 'their', 'to', 'is', 'not',
'it', 'because', 'we', 'have', 'it', 'it', 'because', 'it', 'is', 'on', 'the', 'of', 'and', 'in', 'a', 'an', 'and',
'through', 'by', 'the', 'of', 'a', 'it', 'because', 'of', 'all', 'of', 'the', 'of', 'it', 'because', 'has', 'an', 'of',
'in', 'the', 'of', 'it', 'if', 'we', 'below', 'the', 'we', 'it', 'on', 'for', 'it', 'is', 'the', 'most', 'the', 'most',
'and', 'in', 'the', 'the', 'most', 'of', 'all', 'of', 'a', 'has', 'a', 'that', 'be', 'and', 'and', 'and', 'in', 'a',
'that', 'up', 'to', 'the', 'of', 'our', 'a', 'has', 'a', 'a', 'that', 'be', 'and', 'that', 'that', 'the', 'and', 'the',
'of', 'its', 'all', 'the', 'other', 'that', 'the', 'of', 'the', 'a', 'a', 'has', 'more', 'than', 'the', 'of', 'all',
'its', 'is', 'that', 'which', 'most', 'to', 'its', 'which', 'the', 'most', 'of', 'its', 'is', 'a', 'for', 'which', 'we',
'it', 'to', 'a', 'we', 'all', 'what', 'it', 'is', 'the', 'the', 'of', 'is', 'the', 'of', 'was', 'in', 'the', 'of',
'those', 'who', 'from', 'some', 'of', 'but', 'who', 'and', 'to', 'more', 'is', 'no', 'in', 'is', 'the', 'of', 'in',
'the', 'was', 'in', 'the', 'its', 'has', 'been', 'has', 'been', 'the', 'in', 'all', 'to', 'all', 'not', 'because',
'this', 'was', 'a', 'but', 'because', 'all', 'those', 'who', 'here', 'they', 'this', 'a', 'a', 'that', 'should', 'be',
'in', 'was', 'into', 'our', 'own', 'into', 'the', 'of', 'into', 'the', 'of', 'the', 'into', 'the', 'who', 'here', 'to',
'out', 'the', 'of', 'their', 'and', 'the', 'who', 'and', 'the', 'that', 'from', 'them', 'all', 'have', 'and', 'an',
'which', 'in', 'itself', 'has', 'and', 'with', 'each', 'of', 'the', 'or', 'that', 'we', 'have', 'to', 'that', 'we',
'more', 'the', 'and', 'the', 'and', 'the', 'of', 'in', 'the', 'by', 'the', 'and', 'the', 'of', 'the', 'it', 'is', 'not',
'to', 'these', 'is', 'not', 'to', 'and', 'the', 'of', 'this', 'and', 'and', 'its', 'there', 'is', 'the', 'of', 'the',
'the', 'is', 'the', 'the', 'and', 'the', 'as', 'all', 'the', 'not', 'if', 'the', 'of', 'were', 'the', 'and', 'in', 'an',
'the', 'we', 'have', 'that', 'to', 'in', 'our', 'in', 'because', 'they', 'so', 'to', 'here', 'in', 'the', 'of', 'the',
'to', 'through', 'the', 'of', 'in', 'the', 'of', 'to', 'in', 'our', 'in', 'our', 'in', 'our', 'and', 'in', 'our', 'to',
'from', 'the', 'other', 'of', 'the', 'and', 'from', 'those', 'the', 'the', 'as', 'as', 'the', 'we', 'to', 'or', 'these',
'of', 'because', 'to', 'the', 'of', 'our', 'is', 'such', 'an', 'of', 'was', 'in', 'of', 'by', 'our', 'in', 'his', 'in',
'it', 'to', 'this', 'of', 'of', 'the', 'of', 'and', 'the', 'of', 'the', 'of', 'are', 'on', 'the', 'to', 'the', 'of',
'the', 'we', 'that', 'we', 'it', 'be', 'with', 'and', 'then', 'we', 'the', 'which', 'so', 'and', 'so', 'to', 'of',
'the', 'and', 'of', 'the', 'and', 'the', 'for', 'that', 'we', 'in', 'the', 'of', 'the', 'of', 'before', 'our', 'is',
'to', 'and', 'to', 'the', 'of', 'this', 'we', 'the', 'of', 'and', 'the', 'of', 'do', 'not', 'are', 'not', 'to', 'we',
'in', 'the', 'of', 'our', 'by', 'the', 'will', 'of']
```

After removing these stop-words processed text are as below :

```
'nation day inaugur sinc peopl renew sens dedic unit state washington day task peopl creat weld togeth nation lincoln
day task peopl preserv nation disrupt within day task peopl save nation institut disrupt without us come time midst
swift happen paus moment take stock recal place histori rediscov may risk real peril inact live nation determin count
year lifetim human spirit life man threescor year ten littl littl less life nation full measur live men doubt men believ
democraci form govern frame life limit measur kind mystic artifici fate unexplain reason tyranni slaveri becom surg wave
futur freedom eb tide american know true eight year ago life republ seem frozen fatalist terror prove true midst shock
act act quickli boldli decis later year live year fruit year peopl democraci brought us greater secur hope better
understand life ideal measur materi thing vital present futur experi democraci success surviv crisi home put away mani
evil thing built new structur endur line maintain fact democraci action taken within threeway framework constitut unit
state coordin branch govern continu freeli function bill right remain inviol freedom elect wholli maintain prophet
downfal american democraci seen dire predict come naught democraci die know seen reviveand grow know cannot die built
unhamp initi individu men women join togeth common enterpris enterpris undertaken carri free express free major know
democraci alon form govern enlist full forc men enlighten know democraci alon construct unlimit civil capabl infinit
progress improv human life know look surfac sens still spread everi contin human advanc end unconquer form human societi
nation like person bodya bodi must fed cloth hous invigor rest manner measur object time nation like person mind mind
must kept inform alert must know understand hope need neighbor nation live within narrow circl world nation like person
someth deeper someth perman someth larger sum part someth matter futur call forth sacr guard present thing find
difficult even imposs hit upon singl simpl word yet understand spirit faith america product centuri born multitud came
mani land high degre mostli plain peopl sought earli late find freedom freeli democrat aspir mere recent phase human
histori human histori permeat ancient life earli peopl blaze anew middl age written magna charta america impact irresist
america new world tongu peopl contin newfound land came believ could creat upon contin new life life new freedom vital
written mayflow compact declar independ constitut unit state gettysburg address first came carri long spirit million
follow stock sprang move forward constantli consist toward ideal gain statur clariti gener hope republ cannot forev
toler either undeserv poverti selfserv wealth know still far go must greatli build secur opportun knowledg everi citizen
measur justifi resourc capac land enough achiev purpos alon enough cloth feed bodi nation instruct inform mind also
spirit three greatest spirit without bodi mind men know nation could live spirit america kill even though nation bodi
mind constrict alien world live america know would perish spirit faith speak us daili live way often unnot seem obviou
speak us capit nation speak us process govern sovereignti state speak us counti citi town villag speak us nation
hemispher across sea enslav well free sometim fail hear heed voic freedom us privileg freedom old old stori destini
america proclaim word propheci spoken first presid first inaugur word almost direct would seem year preserv sacr fire
liberti destini republican model govern justli consid deepli final stake experi intrust hand american peopl lose sacr
fireif let smother doubt fear shall reject destini washington strove valiantli triumphantli establish preserv spirit
faith nation furnish highest justif everi sacrific may make caus nation defens face great peril never encount strong
purpos protect perpetu integr democraci muster spirit america faith america retreat content stand still american go
forward servic countri god'
```

Removal of stop-words from **1961-Kennedy Speech**

Initially we found **1390 words** in 1941-Rossevelt Speech , out of which **642** were stop-words.

	Text	word_count	stop_count	stop	upper
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	1390	642	[we, not, a, of, but, a, of, --, an, as, as, a...	5

Below are the list of stop-words found from the above text :


```
['we', 'not', 'a', 'of', 'but', 'a', 'of', 'an', 'as', 'as', 'a', 'as', 'as', 'have', 'before', 'you', 'and', 'the',
'same', 'our', 'a', 'and', 'is', 'very', 'in', 'his', 'the', 'to', 'all', 'of', 'and', 'all', 'of', 'the', 'same',
'for', 'which', 'our', 'are', 'at', 'the', 'the', 'that', 'the', 'of', 'not', 'from', 'the', 'of', 'the', 'but', 'from',
'the', 'of', 'not', 'that', 'we', 'are', 'the', 'of', 'that', 'the', 'from', 'this', 'and', 'to', 'and', 'that', 'the',
'has', 'been', 'to', 'a', 'of', 'in', 'this', 'by', 'by', 'a', 'and', 'of', 'our', 'and', 'to', 'or', 'the', 'of',
'those', 'to', 'which', 'this', 'has', 'been', 'and', 'to', 'which', 'we', 'are', 'at', 'and', 'the', 'it', 'or',
'that', 'we', 'any', 'any', 'any', 'any', 'any', 'in', 'to', 'the', 'and', 'the', 'of', 'we', 'and', 'those', 'and',
'we', 'we', 'the', 'of', 'there', 'is', 'we', 'do', 'in', 'a', 'of', 'there', 'is', 'we', 'can', 'do', 'for', 'we',
'not', 'a', 'at', 'and', 'those', 'whom', 'we', 'to', 'the', 'of', 'the', 'we', 'our', 'that', 'of', 'not', 'have',
'to', 'be', 'by', 'a', 'more', 'not', 'to', 'them', 'our', 'we', 'to', 'them', 'their', 'own', 'and', 'to', 'in', 'the',
'those', 'who', 'by', 'the', 'of', 'the', 'up', 'those', 'in', 'the', 'and', 'the', 'to', 'the', 'of', 'we', 'our',
'to', 'them', 'for', 'is', 'not', 'because', 'the', 'be', 'doing', 'not', 'because', 'we', 'their', 'but', 'because',
'it', 'is', 'a', 'the', 'who', 'are', 'it', 'the', 'few', 'who', 'are', 'our', 'of', 'our', 'we', 'a', 'to', 'our',
'into', 'in', 'a', 'for', 'to', 'and', 'in', 'off', 'the', 'of', 'this', 'of', 'the', 'of', 'all', 'our', 'that', 'we',
'with', 'them', 'to', 'or', 'in', 'the', 'other', 'that', 'this', 'to', 'the', 'of', 'its', 'own', 'that', 'of', 'the',
'our', 'in', 'an', 'where', 'the', 'of', 'have', 'the', 'of', 'we', 'our', 'of', 'it', 'from', 'a', 'for', 'to', 'its',
'of', 'the', 'and', 'the', 'and', 'to', 'the', 'in', 'which', 'its', 'to', 'those', 'who', 'themselves', 'our', 'we',
'not', 'a', 'but', 'a', 'that', 'both', 'the', 'for', 'before', 'the', 'of', 'by', 'all', 'in', 'or', 'not', 'them',
'with', 'only', 'when', 'our', 'are', 'can', 'we', 'be', 'that', 'they', 'will', 'be', 'can', 'and', 'of', 'from',
'our', 'both', 'by', 'the', 'of', 'both', 'by', 'the', 'of', 'the', 'both', 'to', 'that', 'of', 'that', 'the', 'of',
'on', 'both', 'that', 'is', 'not', 'a', 'of', 'and', 'is', 'to', 'out', 'of', 'to', 'both', 'what', 'of', 'those',
'which', 'both', 'for', 'the', 'and', 'for', 'the', 'and', 'of', 'and', 'the', 'to', 'other', 'under', 'the', 'of',
'all', 'both', 'to', 'the', 'of', 'of', 'its', 'the', 'the', 'the', 'and', 'the', 'and', 'both', 'to', 'in', 'all',
'of', 'the', 'the', 'of', 'to', 'the', 'and', 'to', 'the', 'if', 'a', 'of', 'the', 'of', 'both', 'in', 'a', 'not', 'a',
'of', 'but', 'a', 'of', 'where', 'the', 'are', 'just', 'and', 'the', 'and', 'the', 'this', 'will', 'not', 'be', 'in',
'the', 'will', 'it', 'be', 'in', 'the', 'nor', 'in', 'the', 'of', 'this', 'nor', 'in', 'our', 'on', 'this', 'your',
'my', 'more', 'than', 'in', 'will', 'the', 'or', 'of', 'our', 'this', 'was', 'each', 'of', 'has', 'been', 'to', 'to',
'its', 'of', 'who', 'the', 'to', 'the', 'the', 'again', 'not', 'as', 'a', 'to', 'we', 'not', 'as', 'a', 'to', 'we',
'are', 'but', 'a', 'to', 'the', 'of', 'a', 'in', 'and', 'in', 'in', 'a', 'against', 'the', 'of', 'and', 'we', 'against',
'these', 'a', 'and', 'and', 'and', 'that', 'can', 'a', 'more', 'for', 'all', 'you', 'in', 'that', 'the', 'of', 'the',
'only', 'a', 'few', 'have', 'been', 'the', 'of', 'in', 'its', 'of', 'do', 'not', 'from', 'this', 'do', 'not', 'that',
'any', 'of', 'with', 'any', 'other', 'or', 'any', 'other', 'the', 'the', 'which', 'we', 'to', 'this', 'will', 'our',
'and', 'all', 'who', 'it', 'and', 'the', 'from', 'that', 'can', 'the', 'my', 'not', 'what', 'your', 'can', 'do', 'for',
'you', 'what', 'you', 'can', 'do', 'for', 'your', 'of', 'the', 'not', 'what', 'will', 'do', 'for', 'but', 'what', 'we',
'can', 'do', 'for', 'the', 'of', 'you', 'are', 'of', 'or', 'of', 'the', 'of', 'the', 'same', 'of', 'and', 'which', 'we',
'of', 'a', 'our', 'only', 'with', 'the', 'of', 'our', 'to', 'the', 'we', 'and', 'but', 'that', 'here', 'on', 'be',
'our']
```


After removing these stop-words processed text are as below :


```
'vice presid johnson mr speaker mr chief justic presid eisenhow vice presid nixon presid truman reverend clergi fellow
citizen observ today victori parti celebr freedom symbol end well begin signifi renew well chang sworn almighti god
solemn oath forebear l prescrib nearli centuri three quarter ago world differ man hold mortal hand power abolish form
human poverti form human life yet revolutionari belief forebear fought still issu around globe belief right man come
generos state hand god dare forget today heir first revolut let word go forth time place friend foe alik torch pass new
gener american born centuri temper war disciplin hard bitter peac proud ancient heritag unwil wit permit slow undo human
right nation alway commit commit today home around world let everi nation know whether wish us well ill shall pay price
bear burden meet hardship support friend oppos foe order assur surviv success liberti much pledg old alli whose cultur
spiritu origin share pledg loyalti faith friend unit littl cannot host cooper ventur divid littl dare meet power
challeng odd split asund new state welcom rank free pledg word one form coloni control shall pass away mere replac far
iron tyranni shall alway expect find support view shall alway hope find strongli support freedom rememb past foolishli
sought power ride back tiger end insid peopl hut villag across globe struggl break bond mass miseri pledg best effort
help help whatev period requir communist may seek vote right free societi cannot help mani poor cannot save rich sister
republ south border offer special pledg convert good word good deed new allianc progress assist free men free govern
cast chain poverti peac revolut hope cannot becom prey hostil power let neighbor know shall join oppos aggress subvers
anywher america let everi power know hemispher intend remain master hous world assembl sovereign state unit nation last
best hope age instrument war far outpac instrument peac renew pledg supportto prevent becom mere forum invect strengthen
shield new weak enlarg area writ may run final nation would make adversari offer pledg request side begin anew quest
peac dark power destruct unleash scienc engulf human plan accident selfdestruct dare tempt weak arm suffici beyond doubt
certain beyond doubt never employ neither two great power group nation take comfort present cours side overburden cost
modern weapon rightli alarm steadi spread deadli atom yet race alter uncertain balanc terror stay hand mankind final war
let us begin anew rememb side civil sign weak sincer alway subject proof let us never negoti fear let us never fear
negoti let side explor problem unit us instead belabor problem divid us let side first time formul seriou precis propos
inspect control arm bring absolut power destroy nation absolut control nation let side seek invok wonder scienc instead
terror togeth let us explor star conquer desert erad diseas tap ocean depth encourag art commerc let side unit heed
corner earth command isaiah undo heavi burden let oppress go free beachhead cooper may push back jungl suspicion let
side join creat new endeavor new balanc power new world law strong weak secur peac preserv finish first day finish first
day life administr even perhap lifetim planet let us begin hand fellow citizen mine rest final success failur cours sinc
countri found gener american summon give testimoni nation loyalti grave young american answer call servic surround globe
trumpet summon us call bear arm though arm need call battl though embattl call bear burden long twilight struggl year
year rejoic hope patient tribul struggl common enemi man tyranni poverti diseas war forg enemi grand global allianc
north south east west assur fruit life mankind join histor effort long histori world gener grant role defend freedom
```

hour maximum danger shrink respons welcom believ us would exchang place peopl gener energi faith devot bring endeavor light countri serv glow fire truli light world fellow american ask countri ask countri fellow citizen world ask america togeth freedom man final whether citizen america citizen world ask us high standard strength sacrific ask good conscienc sure reward histori final judg deed let us go forth lead land love ask bless help know earth god work must truli'

Removal of stop-words from **1973-Nixon Speech**

Initially we found **1819 words** in 1941-Rossevelt Speech , out of which **916** were stop-words.

	Text	word_count	stop_count	stop	upper
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1819	916	[and, my, of, this, and, we, we, here, was, in...	14

Below are the list of stop-words found from the above text :

['and', 'my', 'of', 'this', 'and', 'we', 'we', 'here', 'was', 'in', 'by', 'the', 'of', 'and', 'of', 'at', 'we', 'here', 'we', 'on', 'the', 'of', 'a', 'of', 'in', 'the', 'before', 'we', 'that', 'that', 'this', 'we', 'are', 'about', 'to', 'will', 'not', 'be', 'what', 'other', 'have', 'so', 'a', 'of', 'and', 'that', 'to', 'at', 'and', 'that', 'this', 'will', 'be', 'what', 'it', 'can', 'a', 'of', 'in', 'which', 'we', 'the', 'and', 'the', 'of', 'as', 'we', 'our', 'as', 'a', 'from', 'our', 'for', 'to', 'our', 'and', 'by', 'our', 'to', 'and', 'to', 'we', 'were', 'to', 'the', 'for', 'a', 'and', 'more', 'of', 'the', 'of', 'the', 'of', 'will', 'be', 'as', 'the', 'of', 'the', 'the', 'of', 'a', 'in', 'the', 'we', 'in', 'the', 'is', 'not', 'the', 'which', 'is', 'an', 'between', 'but', 'a', 'which', 'can', 'for', 'to', 'is', 'that', 'we', 'both', 'the', 'and', 'the', 'of', 'in', 'that', 'we', 'in', 'to', 'the', 'there', 'will', 'be', 'no', 'we', 'in', 'to', 'there', 'will', 'be', 'no', 'the', 'of', 'as', 'a', 'of', 'the', 'we', 'have', 'over', 'these', 'our', 'the', 'that', 'no', 'has', 'the', 'to', 'its', 'will', 'or', 'on', 'by', 'in', 'this', 'of', 'to', 'for', 'the', 'of', 'and', 'to', 'the', 'of', 'between', 'the', 'do', 'our', 'in', 'and', 'in', 'the', 'we', 'to', 'do', 'their', 'has', 'when', 'will', 'other', 'our', 'or', 'other', 'our', 'or', 'to', 'the', 'of', 'other', 'how', 'to', 'their', 'own', 'as', 'we', 'the', 'of', 'each', 'to', 'its', 'own', 'we', 'the', 'of', 'each', 'to', 'its', 'own', 'as', 'is', 'in', 'the', 'so', 'is', 'each', 'in', 'its', 'own', 'with', 'the', 'of', 'the', 'to', 'from', 'the', 'we', 'have', 'to', 'down', 'the', 'of', 'which', 'have', 'the', 'for', 'too', 'and', 'to', 'in', 'their', 'of', 'so', 'that', 'between', 'of', 'the', 'of', 'the', 'can', 'be', 'a', 'of', 'in', 'the', 'in', 'which', 'the', 'are', 'as', 'as', 'the', 'in', 'which', 'each', 'the', 'of', 'the', 'other', 'to', 'by', 'a', 'in', 'which', 'those', 'who', 'will', 'do', 'so', 'by', 'the', 'of', 'their', 'and', 'not', 'by', 'the', 'of', 'their', 'that', 'not', 'as', 'a', 'but', 'because', 'the', 'to', 'such', 'a', 'is', 'the', 'in', 'which', 'a', 'can', 'because', 'only', 'if', 'we', 'in', 'our', 'will', 'we', 'a', 'and', 'only', 'if', 'we', 'a', 'will', 'we', 'in', 'our', 'at', 'have', 'the', 'to', 'do', 'more', 'than', 'before', 'in', 'our', 'to', 'in', 'to', 'a', 'to', 'for', 'to', 'our', 'more', 'and', 'to', 'the', 'of', 'to', 'and', 'the', 'of', 'our', 'is', 'so', 'because', 'the', 'of', 'our', 'is', 'so', 'be', 'in', 'our', 'to', 'those', 'in', 'as', 'a', 'of', 'has', 'from', 'that', 'so', 'a', 'of', 'at', 'from', 'that', 'have', 'the', 'from', 'to', 'has', 'not', 'been', 'a', 'from', 'our', 'but', 'a', 'to', 'at', 'the', 'from', 'to', 'will', 'not', 'be', 'a', 'from', 'our', 'but', 'a', 'to', 'and', 'at', 'the', 'to', 'those', 'in', 'the', 'and', 'the', 'of', 'have', 'too', 'with', 'the', 'of', 'to', 'all', 'and', 'in', 'and', 'at', 'the', 'has', 'to', 'from', 'the', 'of', 'of', 'can', 'be', 'to', 'only', 'if', 'he', 'has', 'is', 'at', 'and', 'to', 'do', 'more', 'for', 'to', 'more', 'for', 'in', 'more', 'what', 'we', 'will', 'do', 'for', 'by', 'what', 'they', 'will', 'do', 'for', 'is', 'why', 'no', 'of', 'a', 'for', 'have', 'too', 'with', 'that', 'too', 'in', 'we', 'have', 'of', 'it', 'more', 'than', 'it', 'can', 'only', 'to', 'to', 'and', 'to', 'a', 'and', 'that', 'both', 'in', 'what', 'can', 'do', 'and', 'in', 'what', 'can', 'to', 'from', 'so', 'that', 'an', 'do', 'more', 'for', 'that', 'was', 'not', 'by', 'but', 'by', 'not', 'by', 'but', 'by', 'not', 'by', 'but', 'by', 'our', 'own', 'each', 'of', 'not', 'just', 'what', 'will', 'do', 'for', 'but', 'what', 'can', 'do', 'for', 'the', 'we', 'each', 'of', 'not', 'just', 'how', 'can', 'but', 'how', 'can', 'has', 'a', 'and', 'to', 'to', 'you', 'that', 'where', 'this', 'should', 'we', 'will', 'and', 'we', 'will', 'just', 'as', 'is', 'the', 'that', 'each', 'and', 'of', 'as', 'an', 'and', 'as', 'a', 'of', 'his', 'own', 'this', 'each', 'of', 'a', 'in', 'his', 'own', 'to', 'his', 'to', 'do', 'his', 'to', 'his', 'so', 'that', 'we', 'can', 'the', 'of', 'a', 'of', 'for', 'and', 'as', 'we', 'our', 'as', 'a', 'we', 'can', 'do', 'so', 'in', 'the', 'of', 'our', 'to', 'ourselves', 'and', 'to', 'the', 'and', 'most', 'to', 'an', 'again', 'to', 'our', 'with', 'and', 'each', 'of', 'out', 'for', 'that', 'a', 'of', 'for', 'the', 'and', 'of', 'a', 'of', 'for', 'the', 'which', 'is', 'the', 'of', 'all', 'the', 'has', 'for', 'to', 'our', 'in', 'ourselves', 'and', 'in', 'that', 'has', 'been', 'have', 'been', 'to', 'be', 'of', 'their', 'of', 'their', 'of', 'at', 'and', 'of', 'its', 'in', 'the', 'we', 'have', 'been', 'by', 'those', 'who', 'with', 'and', 'that', 'is', 'am', 'that', 'this', 'will', 'not', 'be', 'the', 'of', 'on', 'these', 'in', 'which', 'we', 'are', 'to', 'in', 'this', 'has', 'been', 'in', 'the', 'for', 'its', 'for', 'its', 'for', 'its', 'and', 'for', 'its', 'be', 'that', 'our', 'has', 'and', 'more', 'and', 'more', 'more', 'than', 'any', 'other', 'in', 'the', 'of', 'the', 'be', 'that', 'in', 'each', 'of', 'the', 'in', 'which', 'we', 'have', 'been', 'in', 'this', 'the', 'we', 'are', 'now', 'to', 'an', 'we', 'have', 'not', 'for', 'our', 'but', 'to', 'be', 'that', 'by', 'our', 'and', 'by', 'our', 'for', 'with', 'we', 'have', 'a', 'in', 'the', 'what', 'the', 'has', 'not', 'before', 'a', 'of', 'that', 'can', 'not', 'for', 'our', 'but', 'for', 'to', 'are', 'here', 'on', 'an', 'that', 'as', 'those', 'any', 'or', 'any', 'has', 'to', 'to', 'and', 'to', 'our', 'for', 'the', 'in', 'which', 'we', 'these', 'in', 'this', 'so', 'by', 'of', 'who', 'have', 'here', 'before', 'of', 'the', 'they', 'had', 'for', 'and', 'of', 'how', 'each', 'that', 'he', 'himself', 'in', 'to', 'those', 'your', 'that', 'in', 'the', 'have', 'in', 'that', 'are', 'for', 'and', 'for', 'your', 'so', 'that', 'we', 'be', 'of', 'our', 'to', 'these', 'the', 'in', 'so', 'that', 'on', 'its', 'will', 'be', 'as', 'and', 'as', 'as', 'when', 'it', 'and', 'as', 'a', 'of', 'for', 'all', 'the', 'from', 'here', 'in', 'in', 'our', 'in', 'by', 'our', 'in', 'who', 'and', 'to']

After removing these stop-words processed text are as below :

'mr vice presid mr speaker mr chief justic senat cook mr eisenhow fellow citizen great good countri share togeth met four year ago america bleak spirit depress prospect seemingli endless war abroad destruct conflict home meet today stand threshold new era peac world central question us shall use peac let us resolv era enter postwar period often time retreat isol lead stagnat home invit new danger abroad let us resolv becom time great respons greatli born renew spirit promis america enter third centuri nation past year saw farreach result new polici peac continu revit tradit friendship mission peke moscow abl establish base new durabl pattern relationship among nation world america bold initi long rememb year greatest progress sinc end world war ii toward last peac world peac seek world flimsi peac mere interlud war peac endur gener come import understand necess limit america role maintain peac unless america work preserv peac peac unless america work preserv freedom freedom let us clearli understand new natur america role result new polici adopt past four year shall respect treati commit shall support vigor principl countri right impos rule anoth forc shall continu era negoti work limit nuclear arm reduc danger confront great power shall share defend peac freedom world shall expect other

share time pass america make everi nation conflict make everi nation futur respons presum tell peopl nation manag affair respect right nation determin futur also recogn respons nation secur futur america role indispens preserv world peac nation role indispens preserv peac togeth rest world let us resolv move forward begin made let us continu bring wall hostil divid world long build place bridg understand despit profound differ system govern peopl world friend let us build structur peac world weak safe strong respect right live differ system would influenc other strength idea forc arm let us accept high respons burden gladli gladli chanc build peac noblest endeavor nation engag gladli also act greatli meet respons abroad remain great nation remain great nation act greatli meet challeng home chanc today ever histori make life better america ensur better educ better health better hous better transport cleaner environ restor respect law make commun livabl insur godgiven right everi american full equal opportun rang need great reach opportun great let us bold determin meet need new way build structur peac abroad requir turn away old polici fail build new era progress home requir turn away old polici fail abroad shift old polici new retreat respons better way peac home shift old polici new retreat respons better way progress abroad home key new respons lie place divis respons live long consequ attempt gather power respons washington abroad home time come turn away condescend polici patern washington know best person expect act respons respons human natur let us encourag individu home nation abroad decid let us locat respons place let us measur other today offer promis pure government solut everi problem live long fals promis trust much govern ask deliv lead inflat expect reduc individu effort disappoint frustrat erod confid govern peopl govern must learn take less peopl peopl let us rememb america built govern peopl welfar work shirk respons seek respons live let us ask govern challeng face togeth let us ask govern help help nation govern great vital role play pledg govern act act boldli lead boldli import role everi one us must play individu member commun day forward let us make solemn commit heart bear respons part live ideal togeth see dawn new age progress america togeth celebr th anniversari nation proud fulfil promis world america longest difficult war come end let us learn debat differ civil decenc let us reach one preciou qualiti govern cannot provid new level respect right feel one anoth new level respect individu human digniti cherish birthright everi american els time come us renew faith america recent year faith challeng children taught asham countri asham parent asham america record home role world everi turn beset find everyth wrong america littl right confid judgment histori remark time privileg live america record centuri unparallel world histori respons generos creativ progress let us proud system produc provid freedom abund wide share system histori world let us proud four war engag centuri includ one bring end fought selfish advantag help other resist aggress let us proud bold new initi steadfast peac honor made breakthrough toward creat world world known structur peac last mere time gener come embark today era present challeng great nation gener ever face shall answer god histori conscienc way use year stand place hallow histori think other stood think dream america think recogn need help far beyond order make dream come true today ask prayer year ahead may god help make decis right america pray help togeth may worthi challeng let us pledg togeth make next four year best four year america histori th birthday america young vital began bright beacon hope world let us go forward confid hope strong faith one anoth sustain faith god creat us strive alway serv purpos'

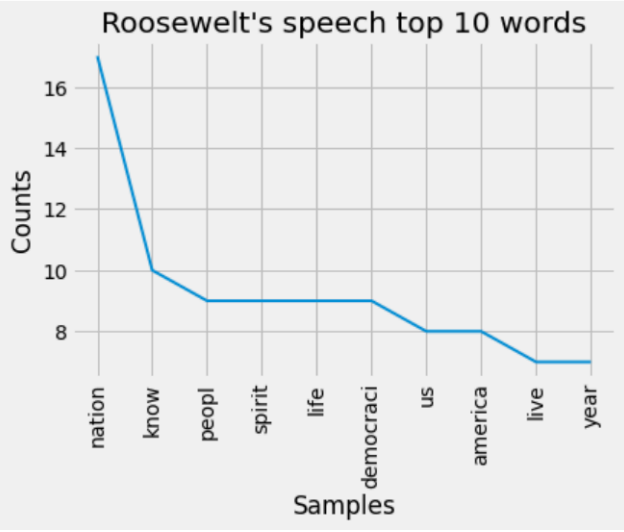
2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stop-words) – 3 Marks

(Please refer the python notebook for better clarity)

Roosevelt's Speech Top 3 words in terms of occurrence

('nation': 17 times)
('know' : 10 times)
('peopl': 9 times)

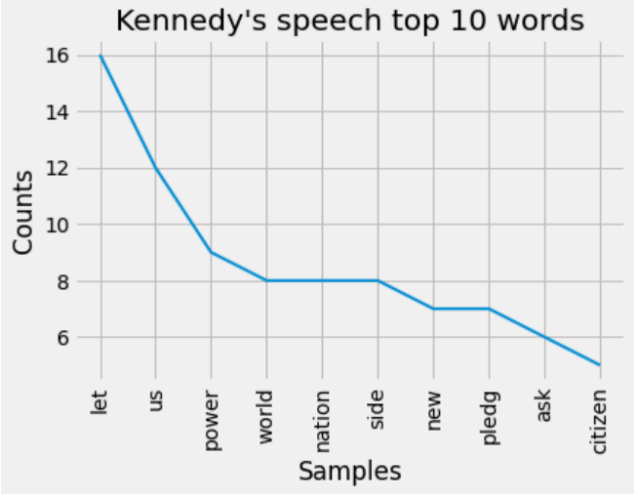
Let's visualise mostly used words



Kennedy's Speech Top 3 words in terms of occurrence

```
('let': 16 times)
('us' : 12 times)
('power': 9 times)
```

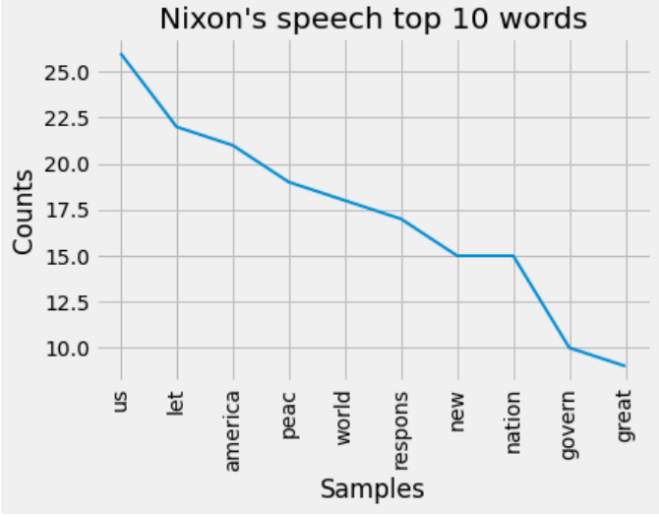
Let's visualise mostly used words



Nixon's Speech Top 3 words in terms of occurrence

```
('us': 26 times)
('let' : 22 times)
('america': 21 times)
```

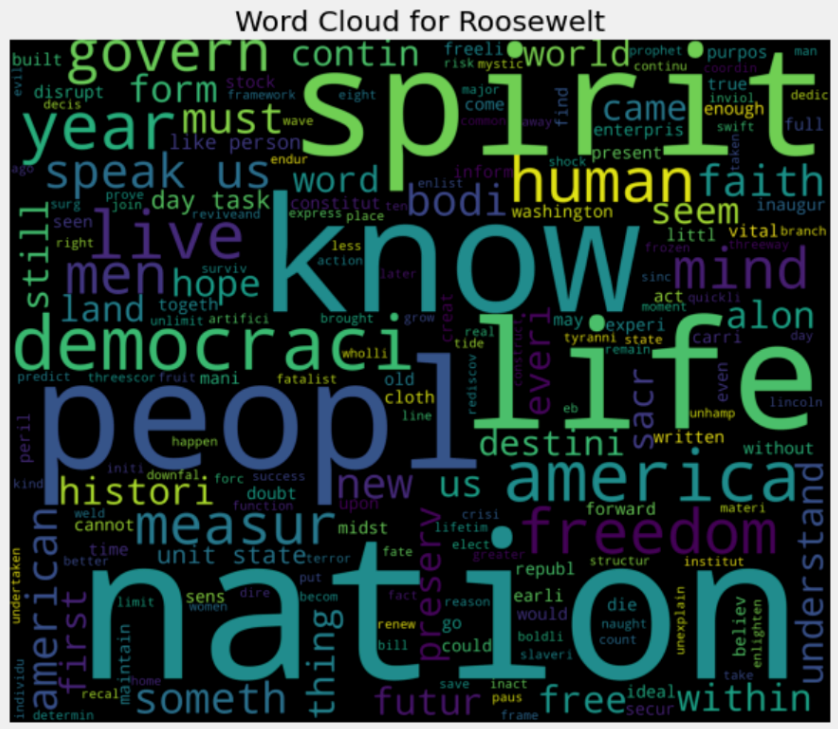
Let's visualise mostly used words



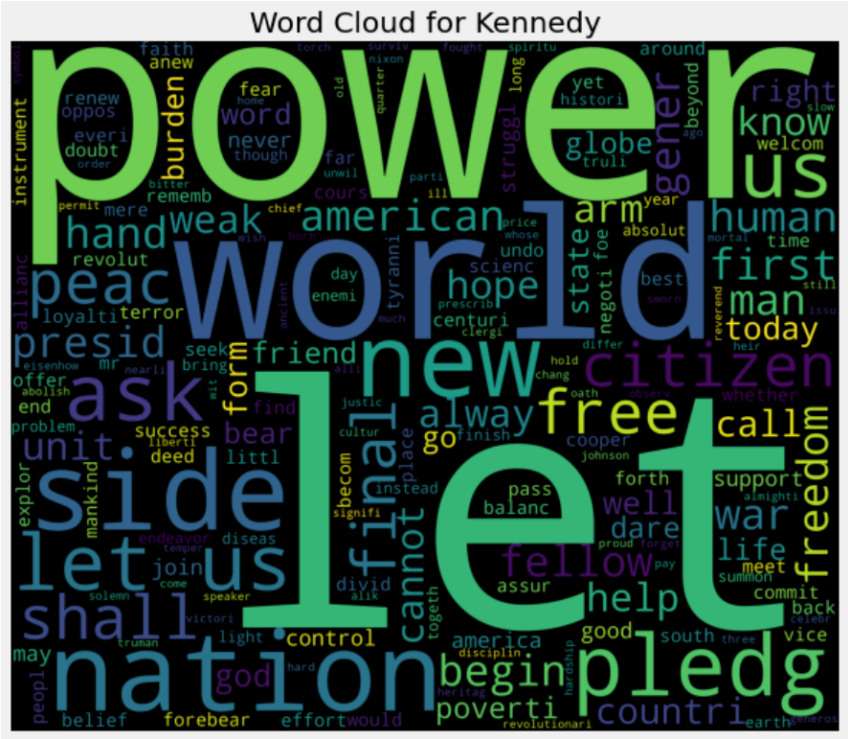
2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stop-words) – 3 Marks

(Please refer the python notebook for better clarity)

Roosevelt's word-cloud after cleaning the text



Kennedy's word-cloud after cleaning the text



Nixon's word-cloud after cleaning the text