MACHINE LEARNING PROJECT

Table of Contents

List of Tables

List of Figures

Problem 1:

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

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

Descriptive Statistics:

Table 1

#	Column	Non-Null Count	Dtype
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

Data Description:

- Vote: Party choice: Conservative or Labour
- Age: in years
- Economic.cond.national: Assessment of current national economic conditions, 1 to 5.
- Economic.cond.household: Assessment of current household economic conditions, 1 to 5.
- Blair: Assessment of the Labour leader, 1 to 5.
- Hague: Assessment of the Conservative leader, 1 to 5.
- Europe: an 11-point scale that measures respondents' attitudes toward European integration. High scoresrepresent 'Eurosceptic' sentiment.
- Political.knowledge: Knowledge of parties' positions on European integration, 0 to 3.
- Gender: female or male.

Table 2

	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

Table 3

	count	mean	std	min	25%	50%	75%	max
age	1525.0	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	2.746885	1.230703	1.0	2.0	2.0	4.0	5.0
Europe	1525.0	6.728525	3.297538	1.0	4.0	6.0	10.0	11.0
political.knowledge	1525.0	1.542295	1.083315	0.0	0.0	2.0	2.0	3.0

Table 4

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

Count of Labour party in the vote column : 1063
Count of Conservative party in the vote column   462

Table 5

vote	object
age	int64
economic.cond.national	int64
economic.cond.household	int64
Blair	int64
Hague	int64
Europe	int64
political.knowledge	int64
gender	object

Insights:

- Data consists of both categorical and numerical values
- There are total 1525 rows representing voters and 10 columns with 9 variables. Out of 10, 2 columns are of object type and 8 columns are of integer type.
- Data does not contain missing values.
- Minimum age of an individual voting is 24 years and maximum age is 93 years. Mean voting age is 54 years.
- Minimum assessment of current national economic conditions is 1 and a maximum assessment is 5 with an average assessment of 3.
- Minimum assessment of current household economic conditions 1 and a maximum assessment is 5 with an average assessment of 3.
- Minimum assessment of the Labour leader Tony Blair is 1 and maximum assessment is 5 with an average assessment of 4.
- Minimum assessment of the Conservative leader William Hague is 1 and maximum r assessment is 5 with an average assessment of 2.
- 75% of the voters on a 11-point scale that measures respondents attitudes toward European integration represent high ‘Eurosceptic’ sentiment with a maximum scale of 11 and a minimum scale of 1.

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

Univariate and Bivariate analysis:

- For variable "age" : Minimum voting age is 24 years and maximum voting age is 93 years. Mean voting age is54 years.
- For variable "economic.cond.national" : Minimum assessment of current national economic conditions is 1and a maximum assessment is 5 with an average assessment of 3.
- For variable "economic.cond.household" : Minimum assessment of current household economic conditions 1and a maximum assessment is 5 with an average assessment of 3.
- For variable "Blair" : Minimum assessment of the Labour leader Tony Blair is 1 and maximum assessment is 5 with an average assessment of 4.
- For variable "Hague" : Minimum assessment of the Conservative leader William Hague is 1 and maximum rassessment is 5 with an average assessment of 2.
- For variable "Europe" : 75% of the voters on a 11-point scale that measures respondents attitudes towardEuropean integration represent high 'Eurosceptic' sentiment with a maximum scale of 11 and a minimum scaleof 1.
- On an average knowledge of parties positions on European integration is 2.Approximately 25% of parties donot hold positions on European integration with a maximum holding of 3.
- The medians of variables "Blair" , "Hague","economic.cond.national" and "economic.cond.household" areidentical to the first quartile, which is why there is an overlap in the Boxplot (Figure: 2.2-2.5).This could bebecause data might have identical large proportion of low values.
- We can also confirm presence of outliers in variables "economic.cond.national" and"economic.cond.household".
- Since the lower quartile and middle quartile values are same (i.e. 0), variable "political.knowledge" does nothave a lower whisker and middle whisker

Age:

Figure 1

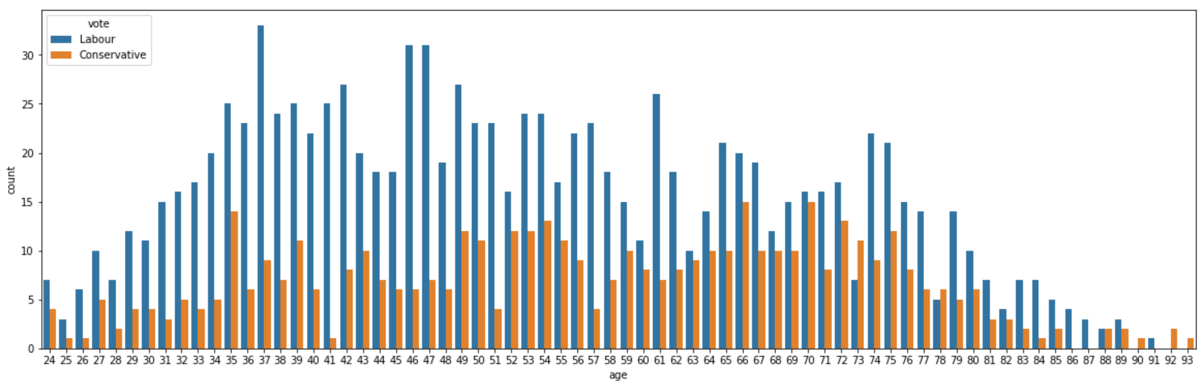
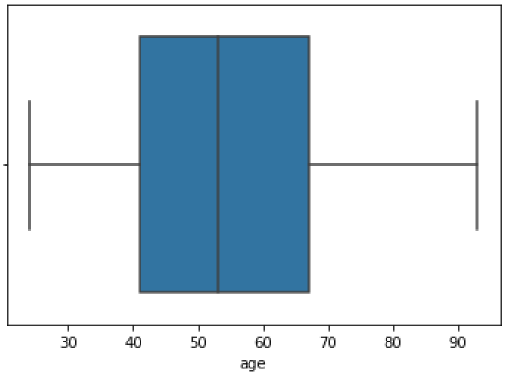


Figure 2



Economic. Cond. National

Figure 3

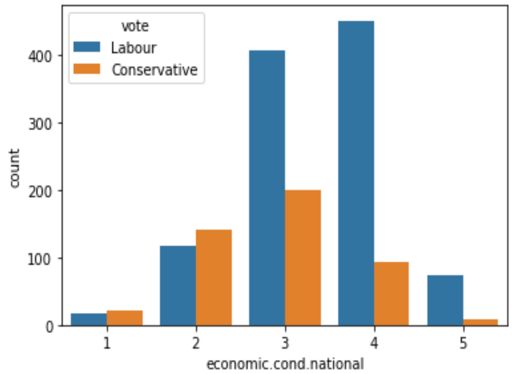
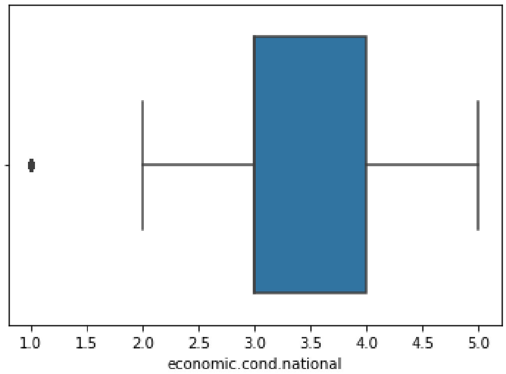


Figure 4



Economic. Cond. Household

Figure 5

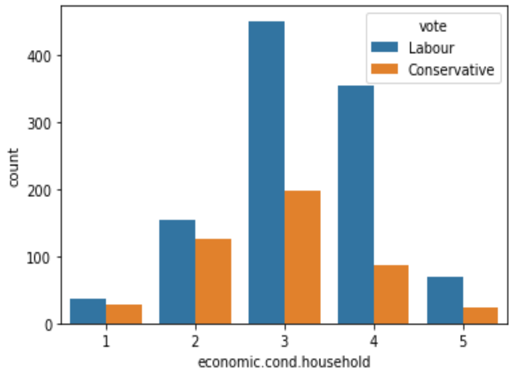
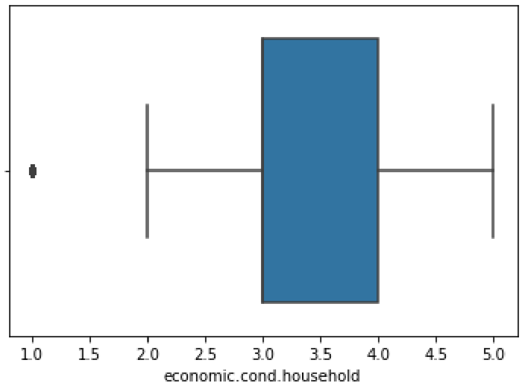


Figure 6



Blair

Figure 7

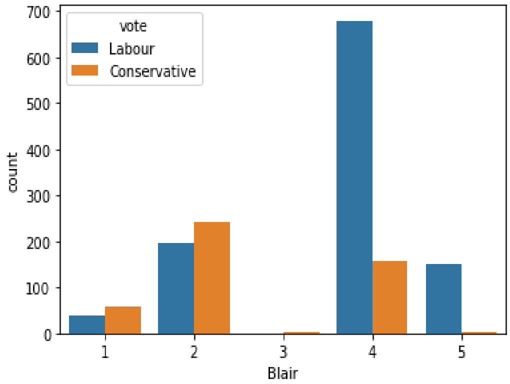
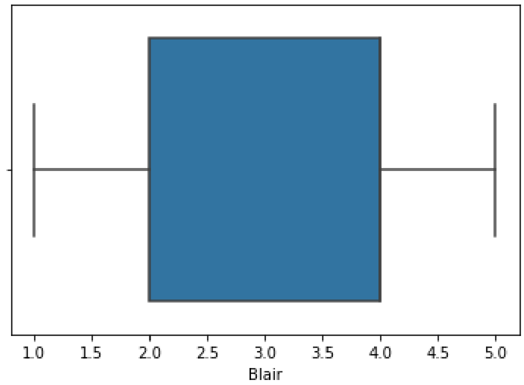


Figure 8



Hague

Figure 9

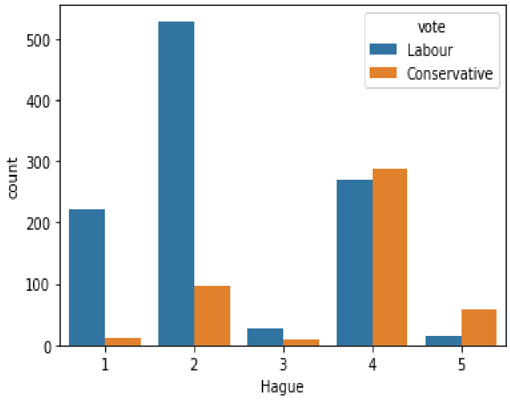
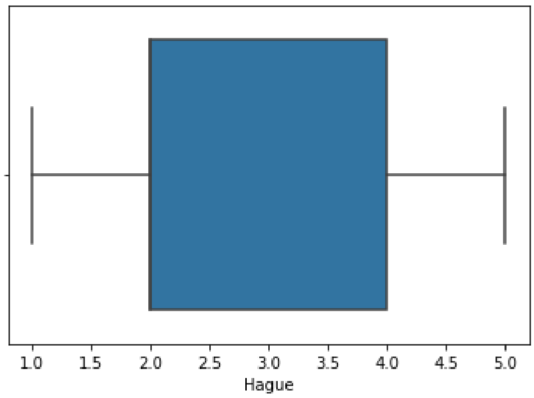


Figure 10



Europe

Figure 11

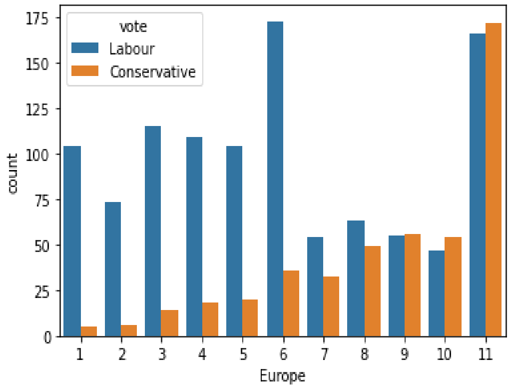
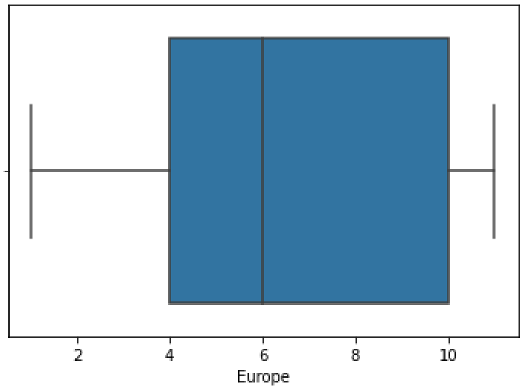


Figure 12



Political Knowledge

Figure 13

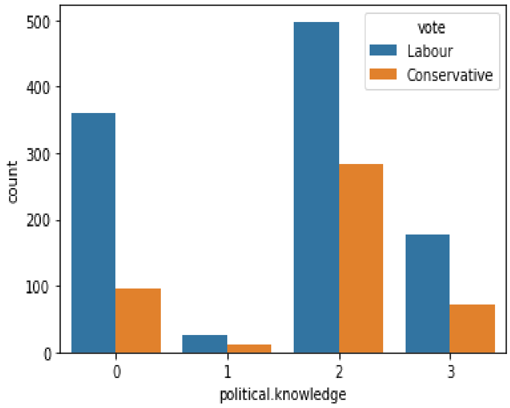
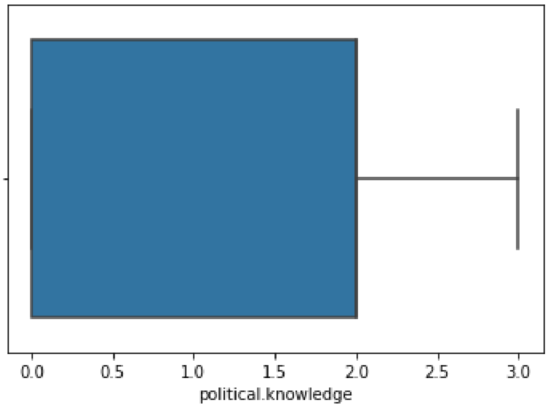


Figure 14



- Distribution is skewed to left tail for all the variable except for variables age and Hague, which has right tail.
- Also, since the skewness is ranging between -0.5 and 0.5 we can say that data is moderately skewed.
- Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to a longer or fatter tail on the right. The mean of positively skewed data will be greater than the median

Multivariate Analysis:

Figure 15

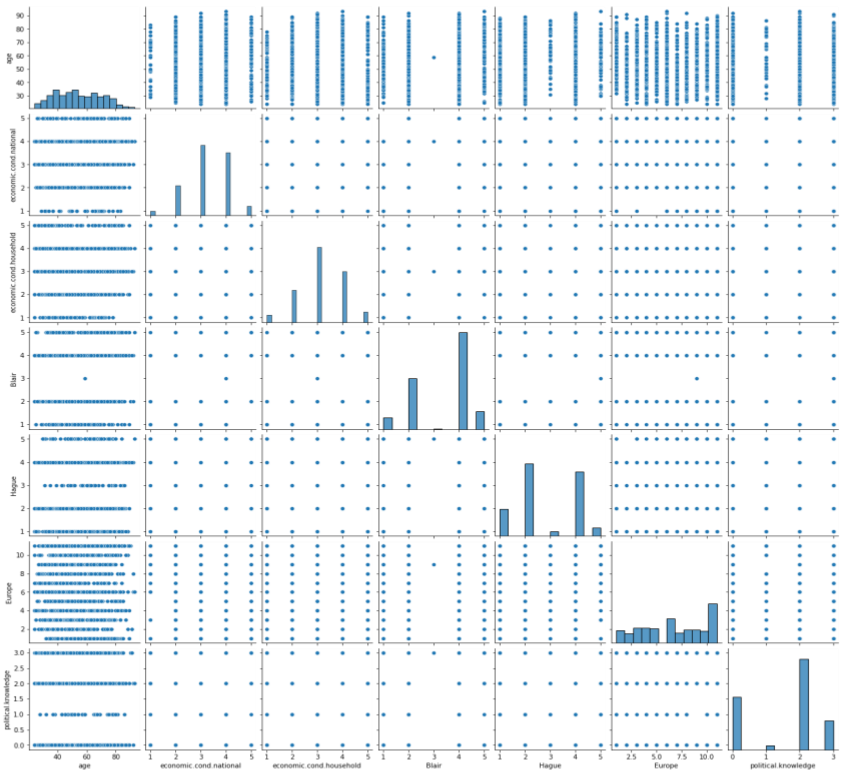


Figure 16

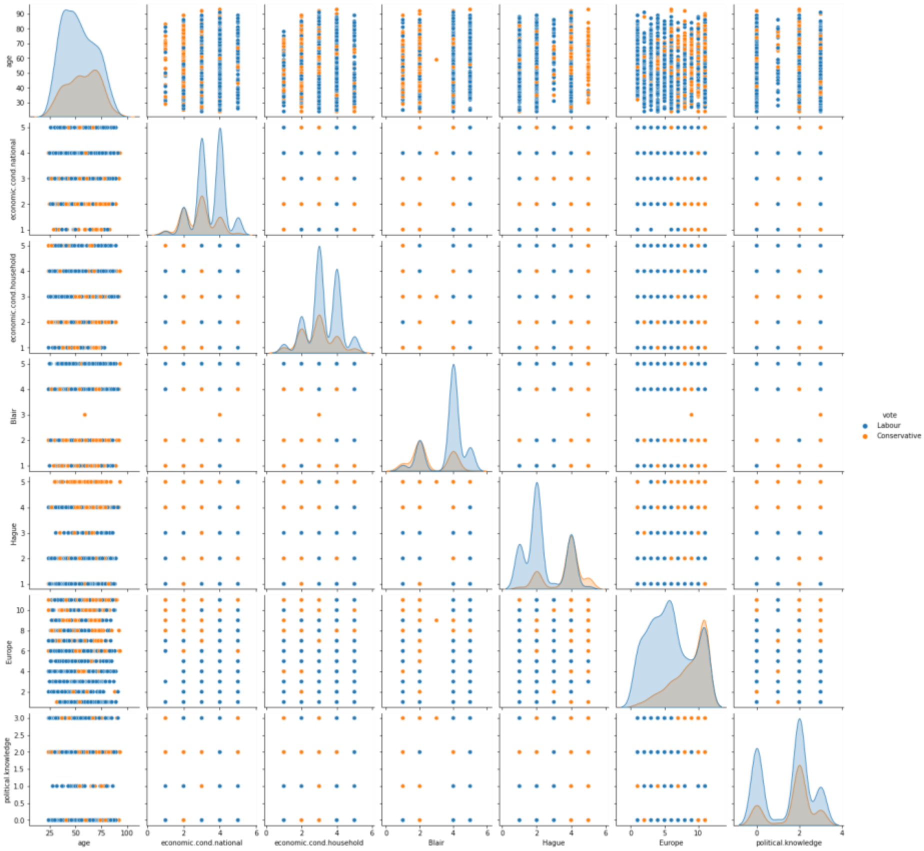
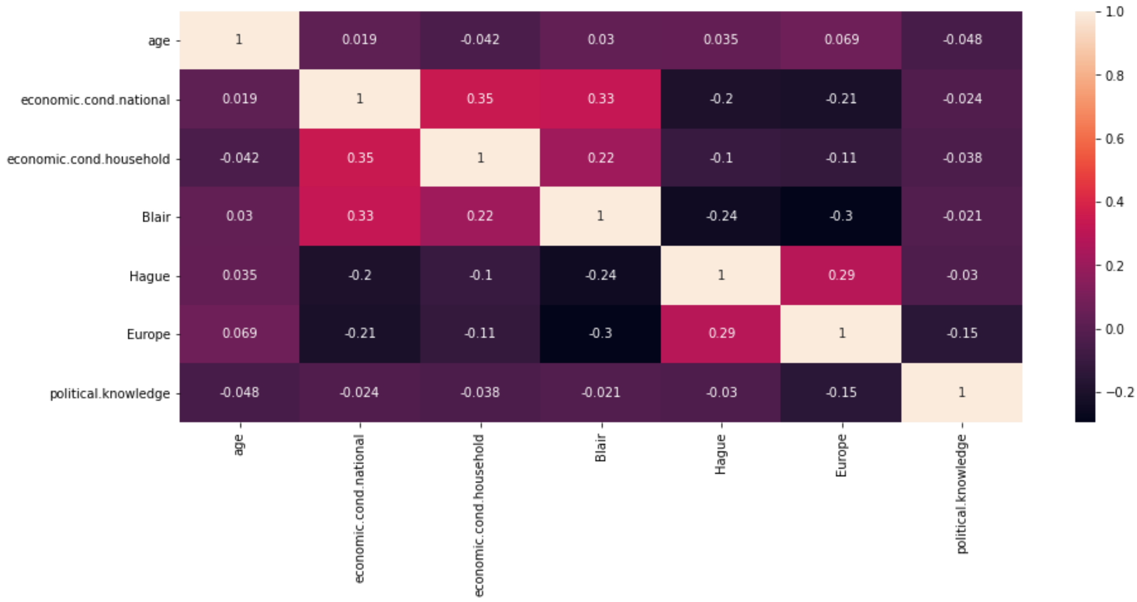
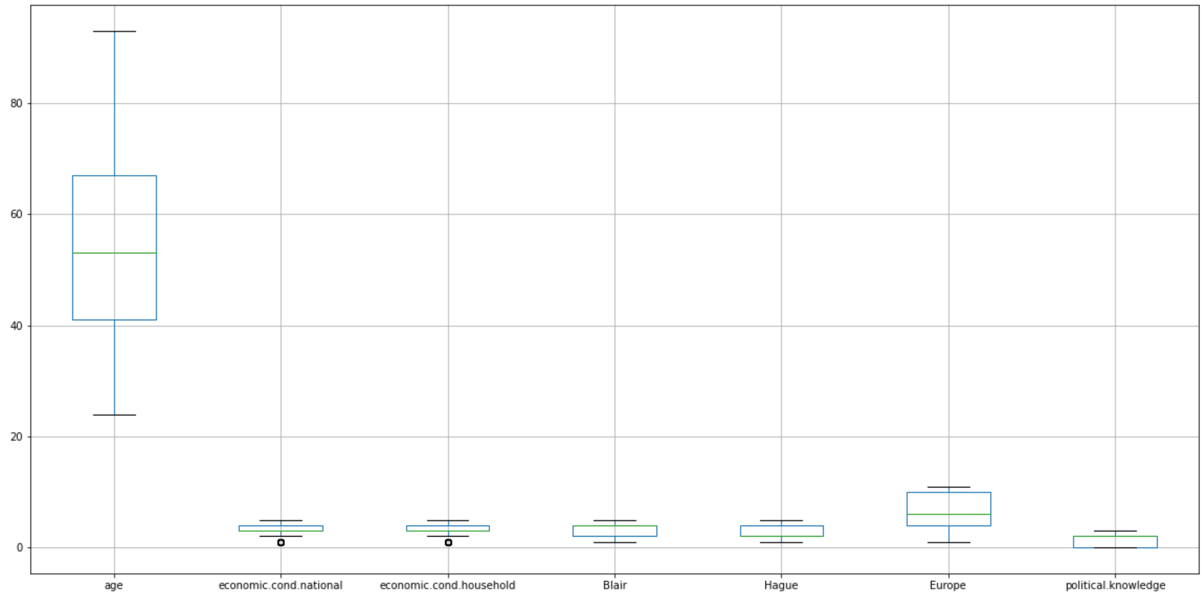


Figure 17



- Negative Correlation is an indication that mentioned variables move in the opposite direction whoever isvoting for Blair is obviously not voting for Hague.Hence there is a negative correlation between the twoindicating cause and effect relationship between the variables.
- In general,correlation values of -0.30 and + 0.30 represent weak correlation.Variables "Blair" and "Hague"both have weak correlation with national and household economic conditions but Blair has slightly better correlation with these parameters (not much of a difference).
- National economic conditions has very weak correlation with household economic condition

Figure 18



Clearly there is presence of outliers in variable economic.cond.household and economic.cond.national.

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).

Feature Scaling is performed when we are dealing with Gradient Descent Based algorithms (Linear and Logistic Regression, Neural Network) and Distance-based algorithms (KNN, K-means, SVM) as these are very sensitive to the range of the data points.

The Machine Learning algorithms that require the feature scaling are mostly KNN (K-Nearest Neighbours), Neural Networks, Linear Regression, and Logistic Regression.- The machine learning algorithms that do not require feature scaling is mostly non-linear ML algorithms such as Decision trees, Random Forest, AdaBoost, Naïve Bayes, etc.
Here, we are building a model, to predict which party a voter will vote for on the basis of the given information and to create an exit poll that will help in predicting overall win and seats covered by a particular party.

In order to do our analysis we are expected to build model using Logistic Regression, LDA, KNN Model and Naïve Bayes Model. For now we are not scaling the data and will do the scaling based on the models we will run ahead. Hence, as mentioned scaling might be necessary for two models and might not be necessary for the other two.

Table 6

age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	IsMale_or_not
-0.711973	-0.279218	-0.150948	0.566716	-1.419886	-1.434426	0.422643	-0.937059
-1.157661	0.856268	0.924730	0.566716	1.018544	-0.524358	0.422643	1.067169
-1.221331	0.856268	0.924730	1.418187	-0.607076	-1.131070	0.422643	1.067169
-1.921698	0.856268	-1.226625	-1.136225	-1.419886	-0.827714	-1.424148	-0.937059
-0.839313	-1.414704	-1.226625	-1.987695	-1.419886	-0.221002	0.422643	1.067169
-0.457295	-0.279218	0.924730	0.566716	1.018544	-0.827714	0.422643	1.067169
0.179402	-1.414704	-1.226625	0.566716	1.018544	1.295778	0.422643	1.067169
1.452797	-0.279218	0.924730	0.566716	-1.419886	-1.737782	-1.424148	1.067169
-0.966652	-0.279218	-0.150948	0.566716	1.018544	1.295778	-1.424148	-0.937059
1.007109	-0.279218	-1.226625	1.418187	-1.419886	1.295778	0.422643	1.067169

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

Linear Discriminant Analysis:

LDA model score for train data = 0.8369259606373008

Confusion Matrix for LDA Train data

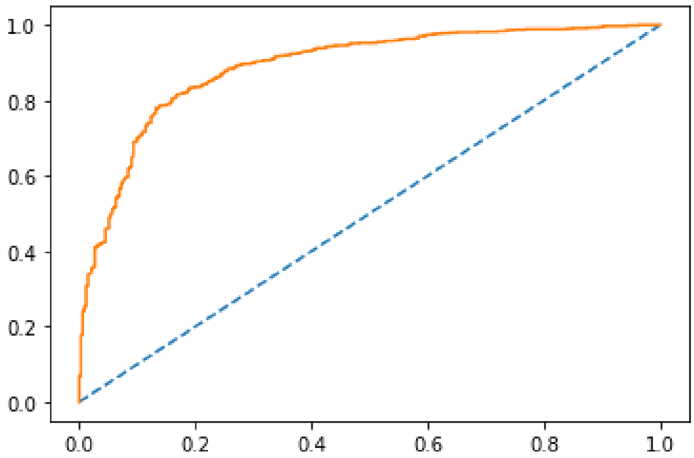
233	99
75	660

LDA Classification Report for LDA Train data

Table 7

	precision	recall	f1-score	support
0	0.76	0.70	0.73	332
1	0.87	0.90	0.88	735
accuracy			0.84	1067
macro avg	0.81	0.80	0.81	1067
weighted avg	0.83	0.84	0.84	1067

Figure 19



LDA model score for test data = 0.8187772925764192

Confusion Matrix for LDA Test data

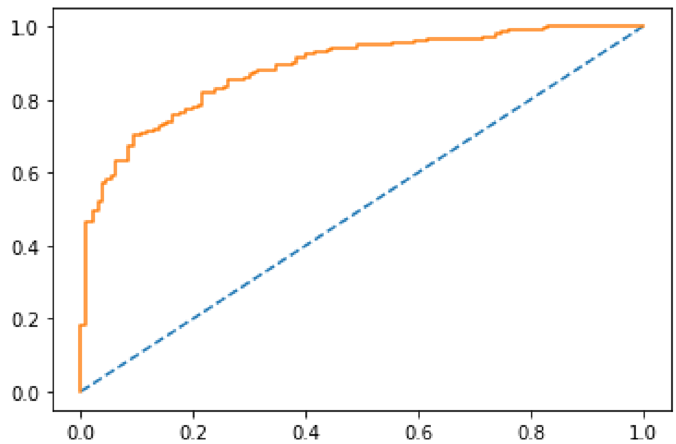
86	44
39	289

LDA Classification Report for LDA Test data

Table 8

	precision	recall	f1-score	support
0	0.69	0.66	0.67	130
1	0.87	0.88	0.87	328
accuracy			0.82	458
macro avg	0.78	0.77	0.77	458
weighted avg	0.82	0.82	0.82	458

Figure 20



Inference of LDA Model:

Using the confusion matrix, the True Positive, False Positive, False Negative, and True Negative values can be extracted which will aid in the calculation of the accuracy score, precision score, recall score, and f1score.Listing below model performance metrics before fine tuning the model:

Train Data:

True Positive: 233
False Positive: 75
False Negative: 99
True Negative: 660
AUC: 88%
Accuracy: 84%
Precision: 87%
f1-Score: 88%
Recall: 90%

Test Data:

True Positive: 86
False Positive: 39
False Negative: 44
True Negative: 289
AUC: 88.4%
Accuracy: 82%
Precision: 87%
f1-Score: 87%
Recall: 88%

We know that, FPR tells us what proportion of the negative class got incorrectly classified by the classifier. Here, we have higher TNR and a lower FPR which is desirable to classify the negative class. Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low indicating high Sensitivity/Recall, Precision, Specificity and F1 Score. Accuracy of the model is more than 70%, which can be considered as a good accuracy score. Train and Test data scores are mostly in line and the overall performance of model looks good. Hence, it can be inferred that overall this model can be considered as a good model.

Logistic Regression:

Logistic regression model score for train data = 0.8406747891283973

Confusion Matrix for Logistic regression model Train data

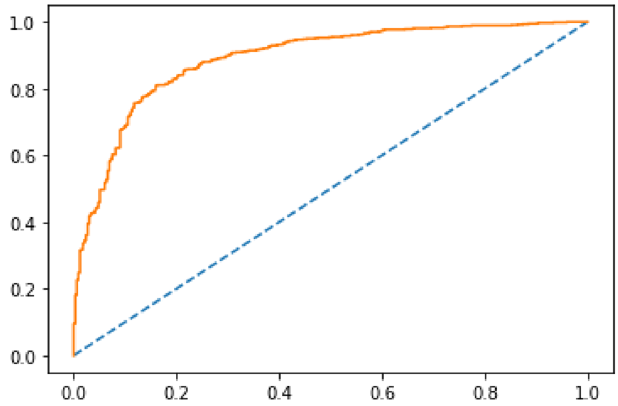
230	102
68	667

Logistic regression model Classification Report for Train data

Table 9

	precision	recall	f1-score	support
0	0.77	0.69	0.73	332
1	0.87	0.91	0.89	735
accuracy			0.84	1067
macro avg	0.82	0.80	0.81	1067
weighted avg	0.84	0.84	0.84	1067

Figure 21



Logistic regression model score for test data = 0.8406747891283973

Confusion Matrix for Logistic regression model Test data

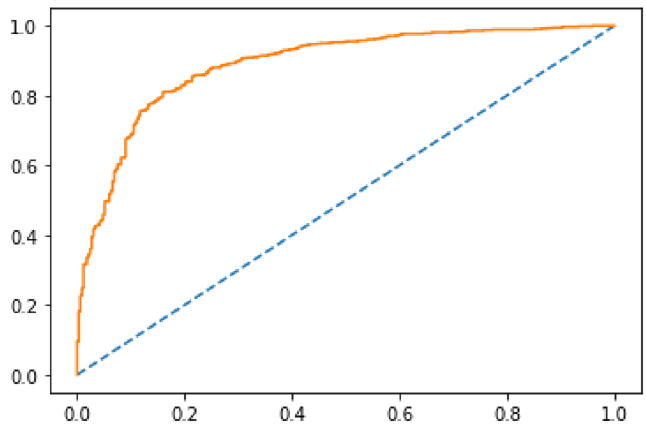
85	45
36	292

Logistic regression model Classification Report for Test data

Table 10

	precision	recall	f1-score	support
0	0.70	0.65	0.68	130
1	0.87	0.89	0.88	328
accuracy			0.82	458
macro avg	0.78	0.77	0.78	458
weighted avg	0.82	0.82	0.82	458

Figure 22



Inference of Logistic Regression Model

Train Data:

True Positive: 230
False Positive: 68
False Negative: 102
True Negative: 667
AUC: 88.9%
Accuracy: 84%
Precision: 87%
f1-Score: 89%
Recall:91%

Test Data:

True Positive: 85
False Positive: 36
False Negative: 45
True Negative: 292
AUC: 88.2 %
Accuracy: 82 %
Precision: 87 %
f1-Score: 88 %
Recall: 89 %

We know that, FPR tells us what proportion of the negative class got incorrectly classified by the classifier.Here, we have higher TNR and a lower FPR which is desirable to classify the negative class. Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low indicating high Sensitivity/Recall, Precision, Specificity and F1 Score.

Accuracy of the model is more than 70%, which can be considered as a good accuracy score. Train and Test data scores are mostly in line and the overall performance of model looks good. Hence, it can be inferred that overall this model can be considered as a good model.

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

KNN Model:

1) For K=5

KNN model score for train data = 0.8678915135608049

Confusion Matrix for KNN model Train data

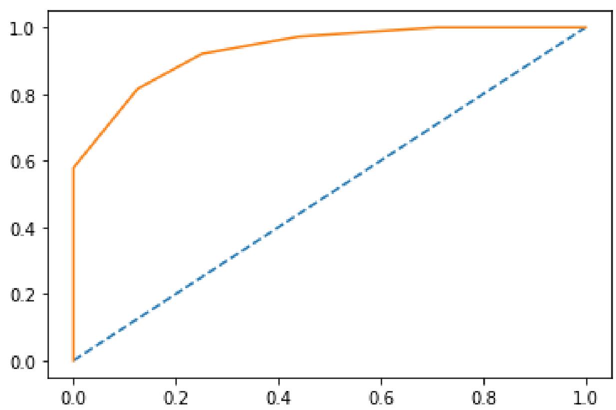
263	88
63	729

KNN model Classification Report for Train data

Table 11

	precision	recall	f1-score	support
0	0.81	0.75	0.78	351
1	0.89	0.92	0.91	792
accuracy			0.87	1143
macro avg	0.85	0.83	0.84	1143
weighted avg	0.87	0.87	0.87	1143

Figure 23



KNN model score for test data = 0.82460732984293

Confusion Matrix for KNN model Test data

81	30

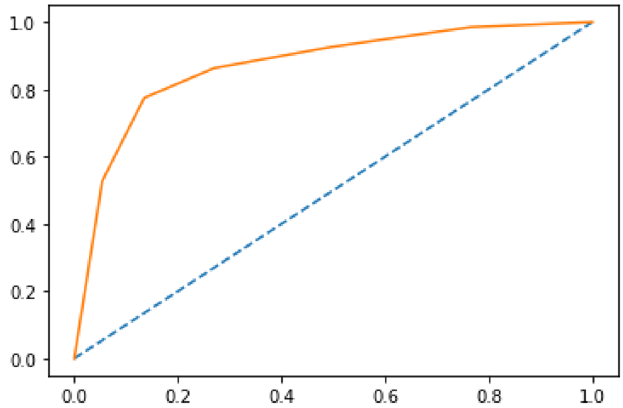
37	234

KNN model Classification Report for Test data

Table 12

	precision	recall	f1-score	support
0	0.69	0.73	0.71	111
1	0.89	0.86	0.87	271
accuracy			0.82	382
macro avg	0.79	0.80	0.79	382
weighted avg	0.83	0.82	0.83	382

Figure 24



Train Data:

True Positive: 263
False Positive: 63
False Negative: 88
True Negative: 729
AUC: 93.2 %
Accuracy: 87 %
Precision: 89 %
f1-Score: 91 %
Recall:92 %

Test Data:

True Positive: 81
False Positive: 37
False Negative: 30
True Negative: 234
AUC: 87 %
Accuracy: 82%

Precision: 89%
f1-Score: 87%
Recall: 86%

We can see a considerable difference in model AUC between Train and Test Data while the other parameters are mostly in line.

2) For K=7

KNN model score for train data = 0.8530183727034121

Confusion Matrix for KNN model Train data

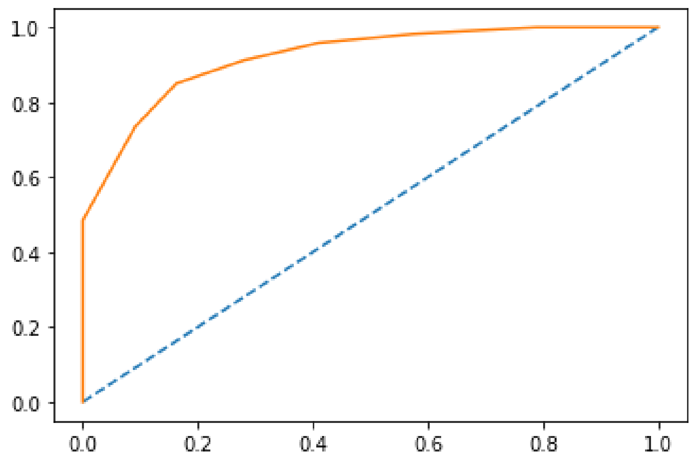
253	98
70	722

KNN model Classification Report for Train data

Table 13

	precision	recall	f1-score	support
0	0.78	0.72	0.75	351
1	0.88	0.91	0.90	792
accuracy			0.85	1143
macro avg	0.83	0.82	0.82	1143
weighted avg	0.85	0.85	0.85	1143

Figure 25



KNN model score for test data = 0.83

Confusion Matrix for KNN model Test data

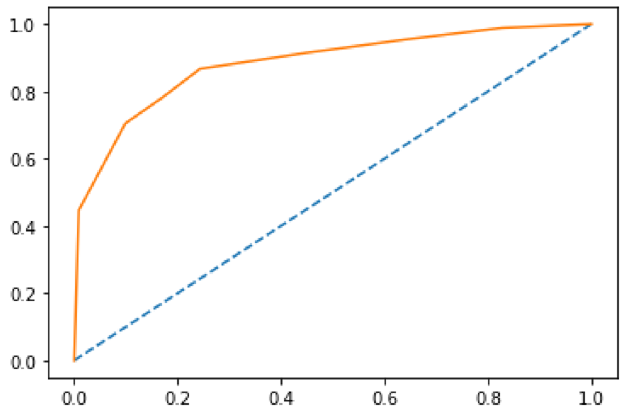
84	27
36	235

KNN model Classification Report for Test data

Table 14

	precision	recall	f1-score	support
0	0.70	0.76	0.73	111
1	0.90	0.87	0.88	271
accuracy			0.84	382
macro avg	0.80	0.81	0.80	382
weighted avg	0.84	0.84	0.84	382

Figure 26



Train Data:

True Positive: 253
False Positive: 70
False Negative: 98
True Negative: 722
AUC: 92 %
Accuracy: 85%
Precision: 88%
f1-Score: 90%
Recall: 91 %

Test Data:

True Positive: 84
False Positive: 36
False Negative: 27
True Negative: 235
AUC: 88 %

Accuracy: 84%
Precision: 90%
f1-Score: 88%
Recall: 87%

Inference:

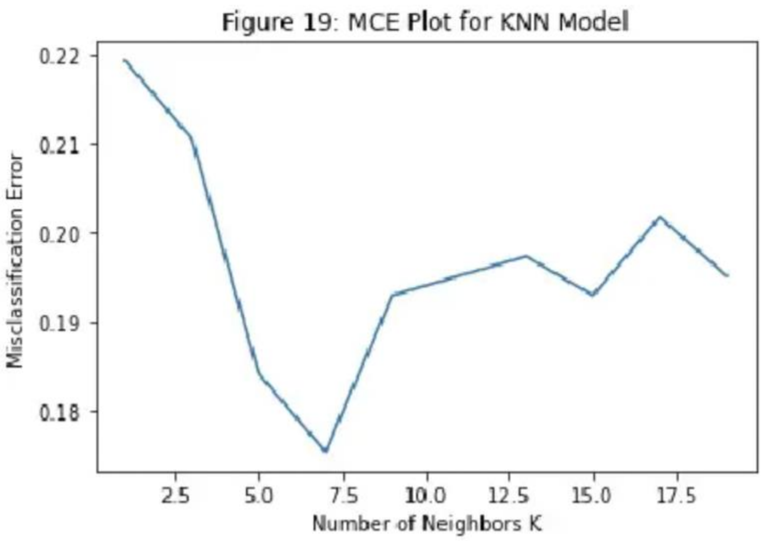
- KNN Model Score for Scaled Train Data for k=5 is 0.8539
- KNN Model Score for Scaled Test Data for k=5 is 0.8157
- KNN Model Score for Scaled Train Data with K=7 is 0.8482
- KNN Model Score for Scaled Test Data with K=7 is 0.8245

There is a slight improvement in Accuracy Score for Test data with K=7
Accuracy score of 85% is generally considered a good accuracy score.
Further, to find the optimal value of k we will look at the K=1,3,5,7....19 and store the train and test scores in a Data frame (ac_score) and using these scores, we will calculate the Misclassification error (MCE) and find the model with lowest Misclassification error (MCE) using the below mentioned formula: Misclassification error(MCE) = 1 - Test accuracy score

Ac_score:

0.7807017543859649
0.7894736842105263
0.8157894736842105
0.8245614035087719
0.8070175438596491
0.8048245614035088
0.8026315789473685
0.8070175438596491
0.7982456140350878
0.8048245614035088

Figure 27



Naïve Bayes Model:

Naïve model for train data

Naïve model score for train data = 0.8331771321462043

Confusion Matrix for Naïve model Train data

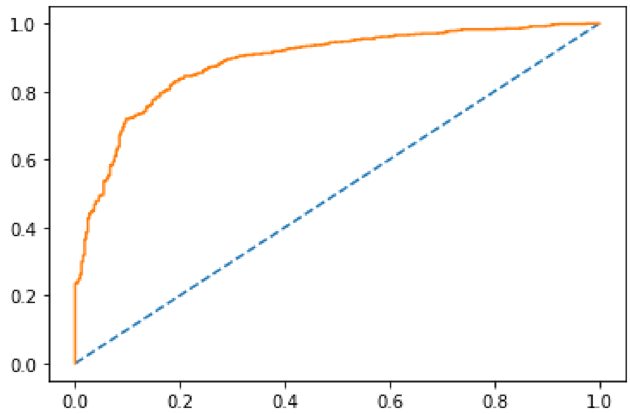
240	92
86	649

Naïve model Classification Report for Train data

Table 15

	precision	recall	f1-score	support
0	0.74	0.72	0.73	332
1	0.88	0.88	0.88	735
accuracy			0.83	1067
macro avg	0.81	0.80	0.80	1067
weighted avg	0.83	0.83	0.83	1067

Figure 28



Naïve model for test data

Naïve model score for test data = 0.8253275109170306

Confusion Matrix for Naïve model Test data

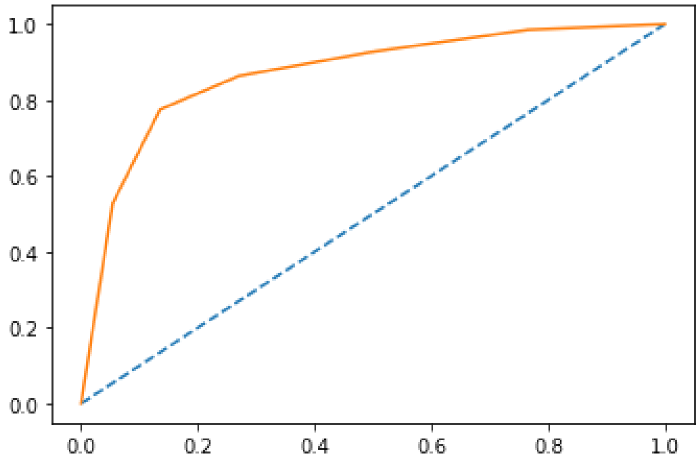
94	36
44	284

Naïve model Classification Report for Test data

Table 16

	precision	recall	f1-score	support
0	0.68	0.72	0.70	130
1	0.89	0.87	0.88	328
accuracy			0.83	458
macro avg	0.78	0.79	0.79	458
weighted avg	0.83	0.83	0.83	458

Figure 29



We know that, FPR tells us what proportion of the negative class got incorrectly classified by the classifier. Here, we have higher TNR and a lower FPR which is desirable to classify the negative class. Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low indicating high Sensitivity/Recall, Precision, Specificity and F1 Score. Accuracy of the model is more than 70%, which can be considered as a good accuracy score. Train and Test data scores are mostly in line and the overall performance of model looks good. Hence, it can be inferred that overall this model can be considered as a good model. After fine tuning the model we can see that model has given mostly the same performance with a very slight improvement in few parameters. Hence, we can say that fine tuning this particular model does not make much of a difference the model performance.

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

Bagging Model:

Bagging model for train data

Bagging model score for train data = 0.9990627928772259

Confusion Matrix for Bagging model Train data

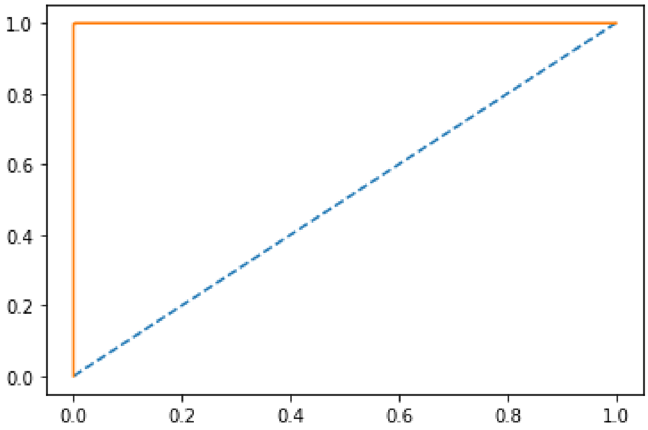
331	1
0	735

Bagging model Classification Report for Train data

Table 17

	precision	recall	f1-score	support
0	1.00	1.00	1.00	332
1	1.00	1.00	1.00	735
accuracy			1.00	1067
macro avg	1.00	1.00	1.00	1067
weighted avg	1.00	1.00	1.00	1067

Figure 30



Bagging model for test data

Bagging model score for test data = 0.7969432314410481

Confusion Matrix for Bagging model Test data

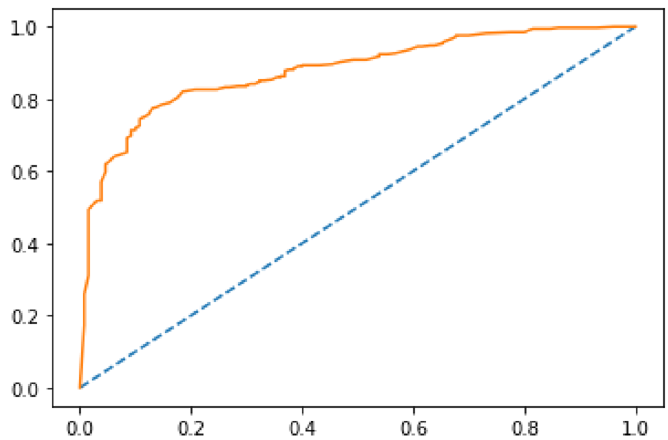
83	47
46	282

Bagging model Classification Report for Test data

Table 18

	precision	recall	f1-score	support
0	0.64	0.64	0.64	130
1	0.86	0.86	0.86	328
accuracy			0.80	458
macro avg	0.75	0.75	0.75	458
weighted avg	0.80	0.80	0.80	458

Figure 31



Inference:

Clearly, our model has better performance on the training set than on the test set, it is likely that model has overfitted. Hence, it might be a big red flag as our model has 100% accuracy on the training set but only82% accuracy on the test set.
Generally bagging is used to avoid problems of overfitting but in this model may be while sampling with replacements some observations got repeated in each subset. Hence, our model is over fitting.
We know that, FPR tells us what proportion of the negative class got incorrectly classified by the classifier. Here, we have higher TNR and a lower FPR which is desirable to classify the negative class. Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low for Test Data indicating high Sensitivity/Recall, Precision, Specificity and F1 Score.

AdaBoostClassifier Model:

AdaBoostClassifier model for train data

AdaBoostClassifier model score for train data = 0.8472352389878163

Confusion Matrix for AdaBoostClassifier model Train data

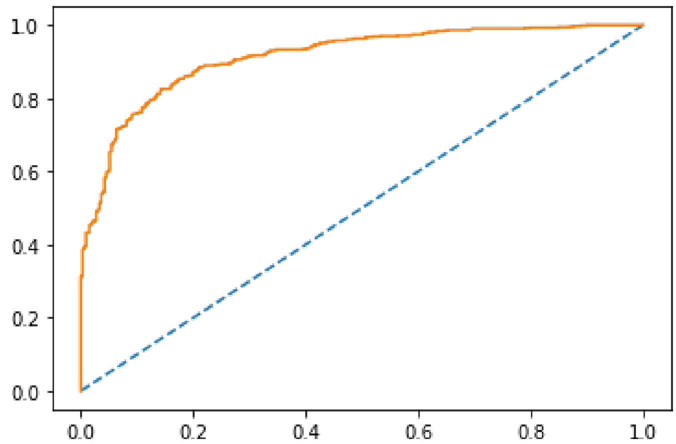
238	94
69	666

AdaBoostClassifier model Classification Report for Train data

Table 19

	precision	recall	f1-score	support
0	0.78	0.72	0.74	332
1	0.88	0.91	0.89	735
accuracy			0.85	1067
macro avg	0.83	0.81	0.82	1067
weighted avg	0.84	0.85	0.85	1067

Figure 32



AdaBoostClassifier model for test data

AdaBoostClassifier model score for test data = 0.8187772925764192

Confusion Matrix for AdaBoostClassifier model Test data

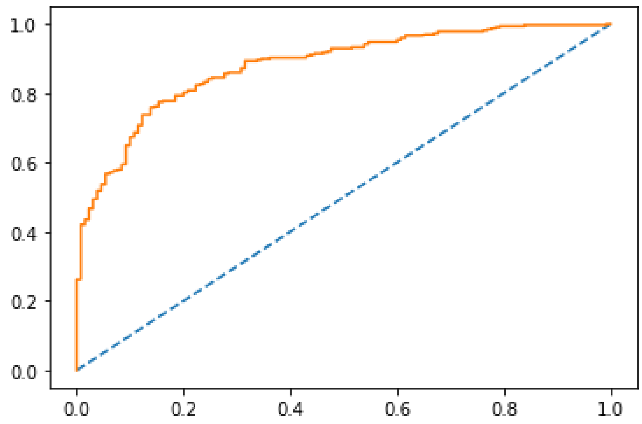
94	36
44	284

AdaBoostClassifier model Classification Report for Test data

Table 20

	precision	recall	f1-score	support
0	0.68	0.72	0.70	130
1	0.89	0.87	0.88	328
accuracy			0.83	458
macro avg	0.78	0.79	0.79	458
weighted avg	0.83	0.83	0.83	458

Figure 33



Inference:

Clearly,our model has better performance on the training set than on the test set.
We know that, FPR tells us what proportion of the negative class got incorrectly classified by theclassifier.
Here, we have higher TNR and a lower FPR which is desirable to classify the negative class.
Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low for indicating highSensitivity/Recall, Precision,Specificity and F1 Score.F1-score, Recall,Precision and AUC are better for train data.

GradientBoostingClassifier:

GradientBoostingClassifier model for train data

GradientBoostingClassifier model score for train data = 0.8865979381443299

Confusion Matrix for GradientBoostingClassifier model train data

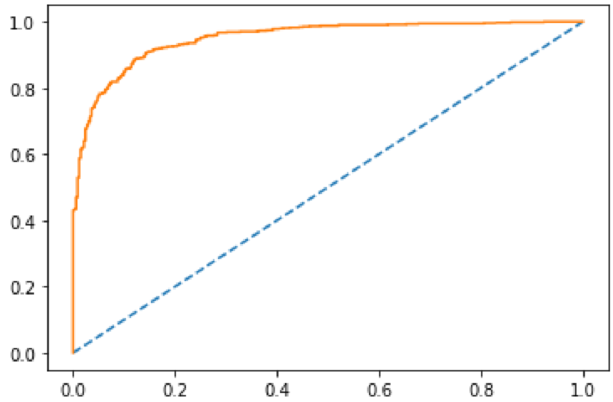
240	92
86	649

GradientBoostingClassifier model Classification Report for train data

Table 21

	precision	recall	f1-score	support
0	0.84	0.79	0.81	332
1	0.91	0.93	0.92	735
accuracy			0.89	1067
macro avg	0.87	0.86	0.87	1067
weighted avg	0.89	0.89	0.89	1067

Figure 34



GradientBoostingClassifier model for test data

GradientBoostingClassifier model score for test data = 0.8318777292576419

Confusion Matrix for GradientBoostingClassifier model test data

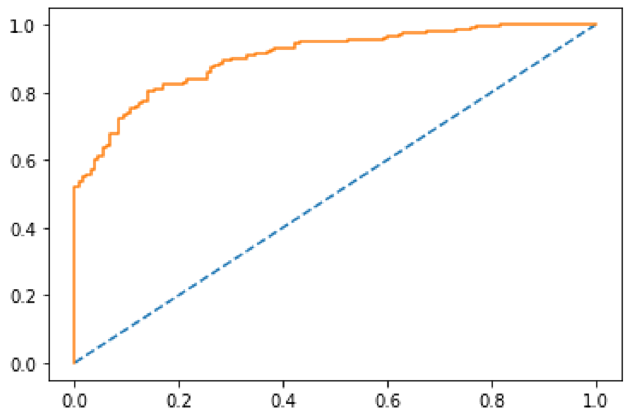
94	36
44	284

GradientBoostingClassifier model Classification Report for test data

Table 22

	precision	recall	f1-score	support
0	0.68	0.72	0.70	130
1	0.89	0.87	0.88	328
accuracy			0.83	458
macro avg	0.78	0.79	0.79	458
weighted avg	0.83	0.83	0.83	458

Figure 35



Inference:

Clearly, our model has better performance on the training set than on the test set. We know that, FPR tells us what proportion of the negative class got incorrectly classified by the classifier. Here, we have higher TNR and a lower FPR which is desirable to classify the negative class. Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low for indicating high Sensitivity/Recall, Precision, Specificity and F1 Score.F1-score, Recall, Precision and AUC are better for train data.

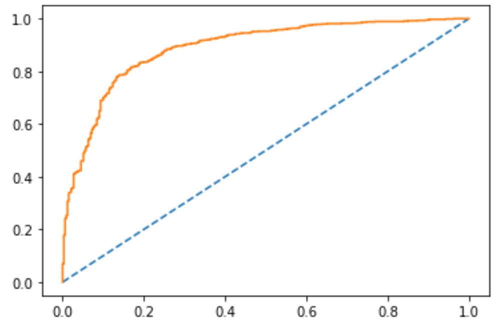
Model can be considered a good model.
The best technique to use between bagging and boosting depends on the data available, simulation, and any existing circumstances at the time. In this case, we might consider Boosting as a better technique since the model is overfitting for Train data with Boosting algorithm.

An estimate's variance is significantly reduced by boosting techniques during the combination procedure, thereby increasing the accuracy. Therefore, the results obtained demonstrate higher stability than the individual results.

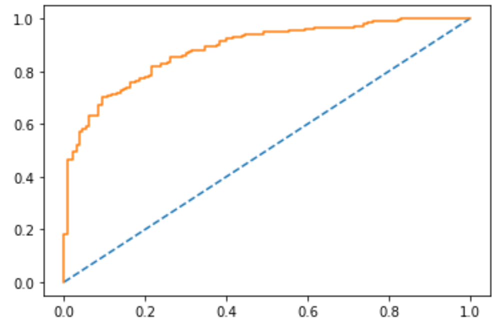
Boosting technique has generated a unified model with lower errors since it concentrates on optimizing the advantages and reducing shortcomings in a single model.

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

LDA model

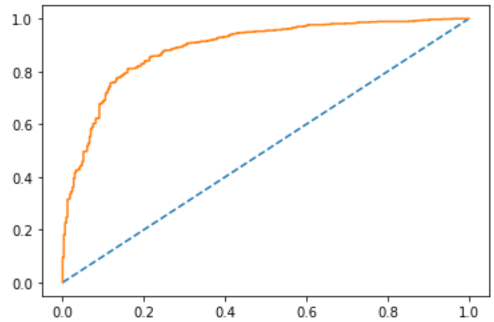


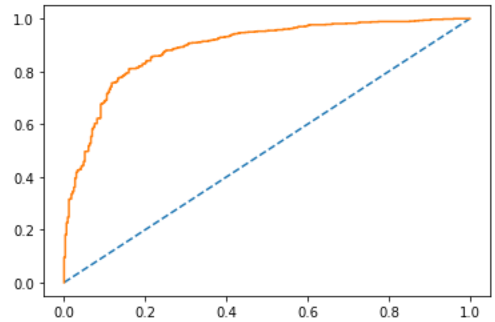
AUC curve : 0.889



AUC curve 0.884

Linear Regression Model

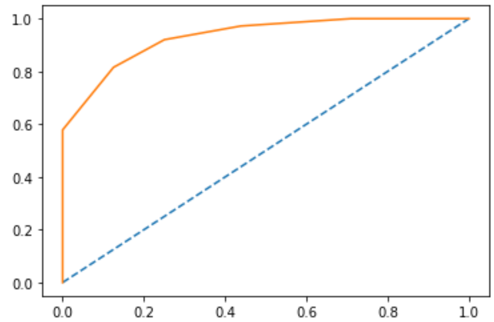


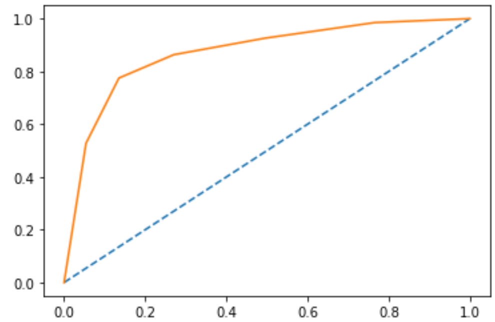


AUC curve : 0.889

AUC curve : 0.882

KNN Model

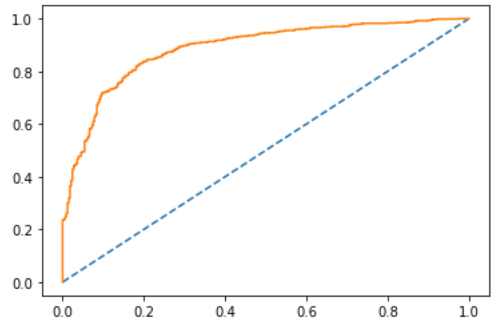


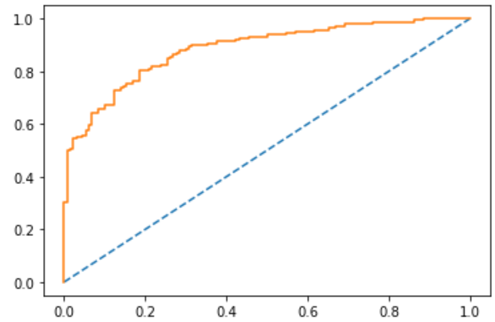


AUC curve: 0.932

AUC curve : 0.870

Naïve Bayes Model

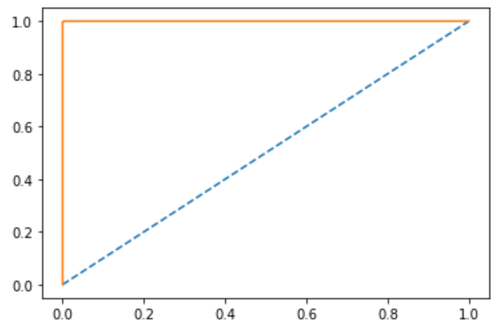


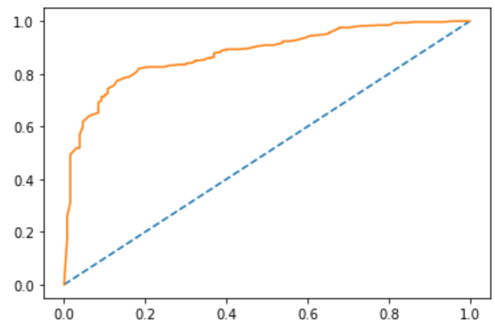


AUC curve: 0.886

AUC curve: 0.885

Bagging Model

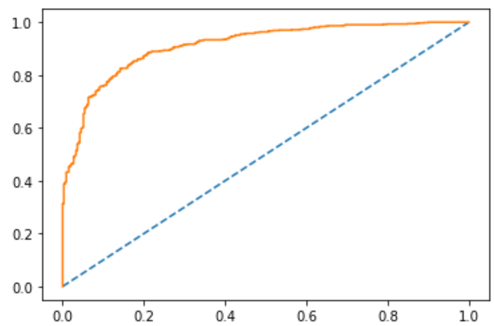


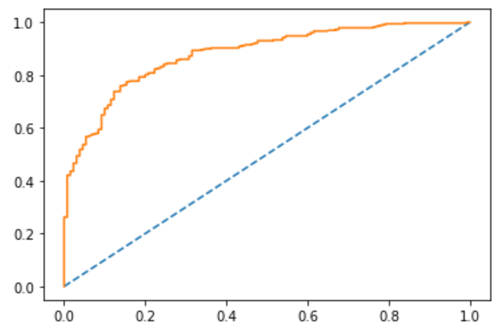


AUC curve: 1.0

AUC curve: 0.87

AdaBoostClassifier Model

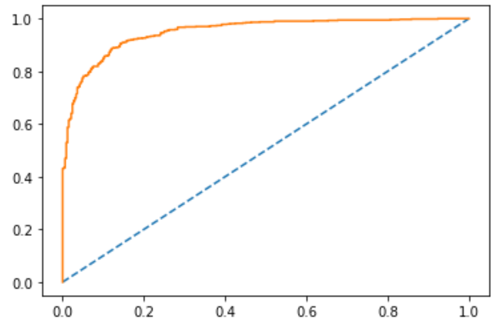


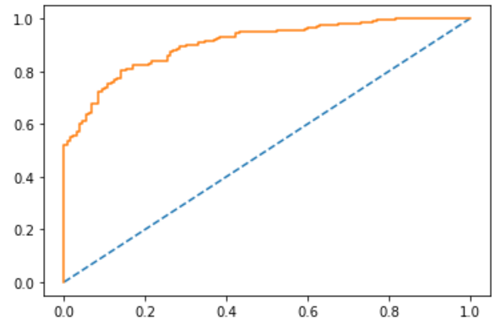


AUC curve: 0.91

AUC curve: 0.87

GradientBoostingClassifier Model





AUC curve:0.95 AUC curve: 0.90

Gradient Boosting Classifier is to be considered as the best model with AUC_ROC score of 95% for train data and 90% for test data and also with accuracy of 89%, when compared to the other models

1.8 Based on these predictions, what are the insights?

Gradient Boosting Classifier is to be considered as the best model with AUC_ROC score of 95% for train data and 90% for test data and also with accuracy of 89%, when compared to the other models.

Along with other parameters such as Recall value, AUC_SCORE and AUC_ROC_Curve, those results were pretty good is this model.

Labour party is performing better than Conservative from huge margin.

Female voters turn out is greater than the male voters.

Those who have better national economic conditions are preferring to vote for Labour party.

Persons having higher Eurosceptic sentiments conservative party are preferring to vote for Conservative party.

Those who have higher political knowledge have voted for Conservative party.

Looking at the assessment for both the leaders, Labour Leader is performing well as he has got better ratings in assessment.

Problem 2:

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

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

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

President Franklin D. Roosevelt in 1941

Length of all words in text 1941-Roosevelt is 1536

Word count in Roosevelt speech

Table 23

	speech	word_count
0	national day inauguration since 1789 people re...	11
1	washingtons day task people create weld together	7
2	lincolns day task people preserve disruption w...	7
3	day task people save institutions disruption w...	7
4	us come time midst swift happenings pause mome...	22

Character count in Roosevelt speech

Table 24

	speech	char_count
0	On each national day of inauguration since 178...	120
1	In Washington's day the task of the people was...	84
2	In Lincoln's day the task of the people was to...	96
3	In this day the task of the people is to save ...	108
4	To us there has come a time, in the midst of s...	248

Sentence count in Roosevelt speech

Table 25

	speech	sent_count
0	On each national day of inauguration since 178...	120
1	In Washington's day the task of the people was...	84
2	In Lincoln's day the task of the people was to...	96
3	In this day the task of the people is to save ...	108
4	To us there has come a time, in the midst of s...	248

President John F. Kennedy in 1961

Length of all words in text 1961-Kennedy is 1546

Word count in Kennedy speech

Table 26

	speech	word_count
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	73
1	The world is very different now. For man holds...	68

2	We dare not forget today that we are the heirs...	96
3	Let every nation know, whether it wishes us we...	40
4	This much we pledge -- and more.	7

Character count in Kennedy speech

Table 27

	speech	**char_count**
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	445
1	The world is very different now. For man holds...	355
2	We dare not forget today that we are the heirs...	512
3	Let every nation know, whether it wishes us we...	217
4	This much we pledge -- and more.	32

Sentence count in Kennedy speech

Table 28

	speech	**sent_count**
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	445
1	The world is very different now. For man holds...	355
2	We dare not forget today that we are the heirs...	512
3	Let every nation know, whether it wishes us we...	217

4	This much we pledge -- and more.	32

President Richard Nixon in 1973

Length of all words in text 1973-Nixon is 2028

Word count in Nixon speech

Table 29

	speech	**word_count**
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	25
1	When we met here four years ago, America was b...	27
2	As we meet here today, we stand on the thresho...	19
3	The central question before us is: How shall w...	51
4	Let us resolve that this will be what it can b...	38

Character count in Nixon speech

Table 30

	speech	**char_count**
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	155
1	When we met here four years ago, America was b...	156
2	As we meet here today, we stand on the thresho...	84

3	The central question before us is: How shall w...	269
4	Let us resolve that this will be what it can b...	199

2.2 Remove all the stopwords from all three speeches

Word count after removing stop words from Roosevelt Speech

Table 31

	speech	**word_count**
0	national day inauguration since 1789, people r...	11
1	washington's day task people create weld toget...	8
2	lincoln's day task people preserve nation disr...	8
3	day task people save nation institutions disru...	8
4	us come time, midst swift happenings, pause mo...	23

Word count after removing stop words from Kennedy Speech

Table 32

	speech	**word_count**
0	vice president johnson, mr. speaker, mr. chief...	48
1	world different now. man holds mortal hands po...	33
2	dare forget today heirs first revolution. let ...	48
3	let every nation know, whether wishes us well ...	25
4	much pledge -- more.	4

Word count after removing stop words from Nixon Speech

Table 33

	speech	word_count
0	mr. vice president, mr. speaker, mr. chief jus...	19
1	met four years ago, america bleak spirit, depr...	16
2	meet today, stand threshold new era peace world.	8
3	central question us is: shall use peace? let u...	26
4	let us resolve become: time great responsibili...	17

2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words.

Word occurred the most number of times in Roosevelt speech

nation	11
know	10
spirit	9

Word occurred the most number of times in Kennedy speech

let	16
us	12
sides	8

Word occurred the most number of times in Nixon speech

us	26
let	22
peace	19

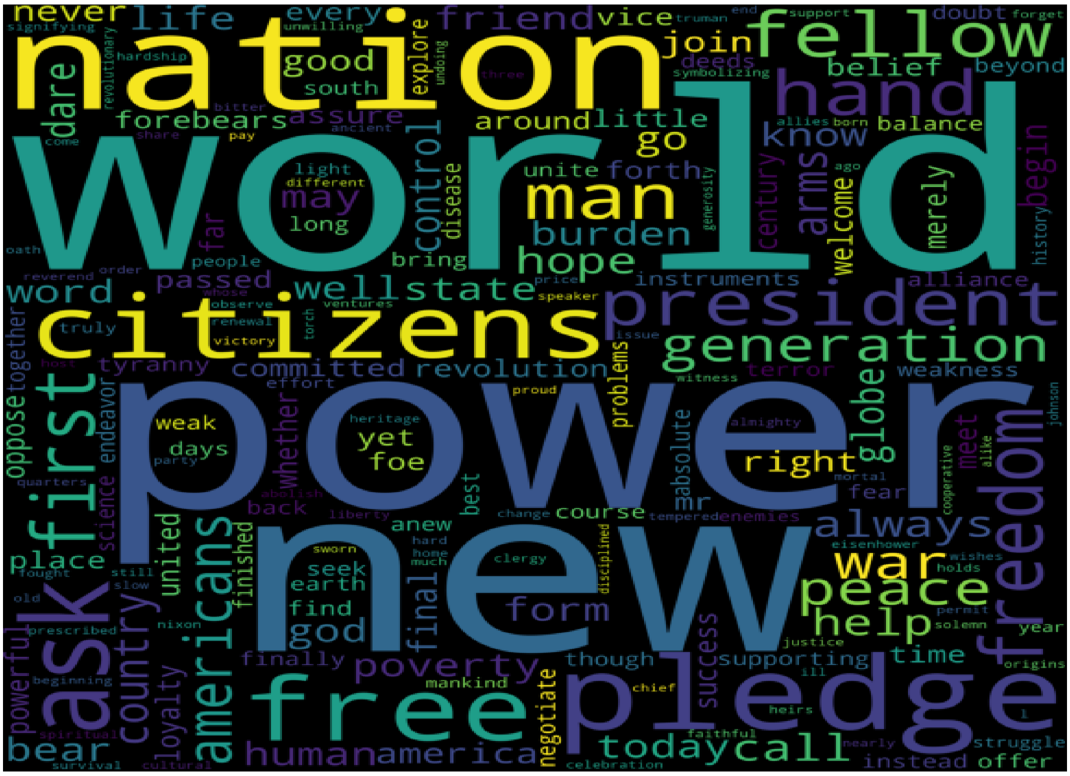
2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords)

Word cloud for Rossevelt

Figure 36

Word cloud for Kennedy

Figure 37



Word cloud for Nixon

Figure 38