MACHINE LEARNING BUSINESS REPORT

PGP-DSBA

JANUARY 17, 2021

Machine Learning Project
Srikanthpr.27@gmail.com

Table of Contents

1. Problem-1: Machine Learning Models

You are hired by one of the leading news channel 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: Election_Data.xlsx

Data Dictionary:

1	vote:	Party choice: Conservative or Labour
2	age:	in years
3	economic.cond.national:	Assessment of current national economic conditions, 1 to 5.
4	economic.cond.household:	Assessment of current household economic conditions, 1 to 5.
5	Blair:	Assessment of the Labour leader, 1 to 5.
6	Hague:	Assessment of the Conservative leader, 1 to 5.
7	Europe:	An 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
8	political.knowledge:	Knowledge of parties' positions on European integration, 0 to 3.
9	gender:	Female or male.

Data Ingestion:

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

Solution:

- Loaded the required packages and read the dataset.
- Remove the unwanted column from the original dataset has it is a serial number.
- Dataset has 1525 observations and 9 features including the target variable.

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

Figure-1: Head of the data

- To understand the data need to look into data structure, summary statistics, skewness and missing values of the data features:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

Figure-2: Data Structure of the data Features

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525.00000	1525.00000	1525.00000	1525.00000	1525.00000	1525.00000	1525.00000
mean	54.18230	3.24590	3.14033	3.33443	2.74689	6.72852	1.54230
std	15.71121	0.88097	0.92995	1.17482	1.23070	3.29754	1.08331
min	24.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.00000
25%	41.00000	3.00000	3.00000	2.00000	2.00000	4.00000	0.00000
50%	53.00000	3.00000	3.00000	4.00000	2.00000	6.00000	2.00000
75%	67.00000	4.00000	4.00000	4.00000	4.00000	10.00000	2.00000
max	93.00000	5.00000	5.00000	5.00000	5.00000	11.00000	3.00000

Figure-3a: Summary Statistics for Numerical data

	vote	gender
count	1525	1525
unique	2	2
top	Labour	female
freq	1063	812

Figure-3b: Summary Statistics for Categorical data

```
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
```

Figure-4: Missing Values

Key takeaways from the data structure and descriptive statistics:

→ There are no Missing values found for a given dataset and also we do not have a unique identifier, hence there is no necessity to check duplicate values.
→ Out of 9 features, we have 8 categorical features and one numerical features.
→ Age feature is the only continuous variable. Minimum age is 24; Maximum age is 93; Average age is 54.18; and has some amount of skewness present which is 0.15 lies between 0 and 0.5, hence the data is fairly symmetrical.
→ 2 variables: vote and gender is a Nominal variable.
  - **Vote:** Majority of the electors choose to vote Labour party compared to conservative party: 1063 out of 1525.
  - **Gender:** Female voters found to be more compared to male voters: 812 female voters out of 1525 voters.

→ 6 variables which are of integer data-type needs to be considered has ordinal variable data-type:
  - **'economic.cond.national'**: Assessment of current national economic conditions was given a rating on a scale of 1 to 5.
  - **'economic.cond.household'**: Assessment of current household economic conditions was given a rating on a scale of 1 to 5.
  - **'Blair'**: Assessment of the Labour leader was given a rating on a scale of 1 to 5.
  - **'Hague'**: Assessment of the conservative leader was given a rating on a scale of 1 to 5.
  - **'Europe'**: An 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
  - **'political.knowledge'**: Knowledge of parties' positions on European integration, was a given a rating on a scale of 0 to 3.

→ Count plot or bar plot will fetch more information for ordinal data type variables to gather more insights.

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

Solution:

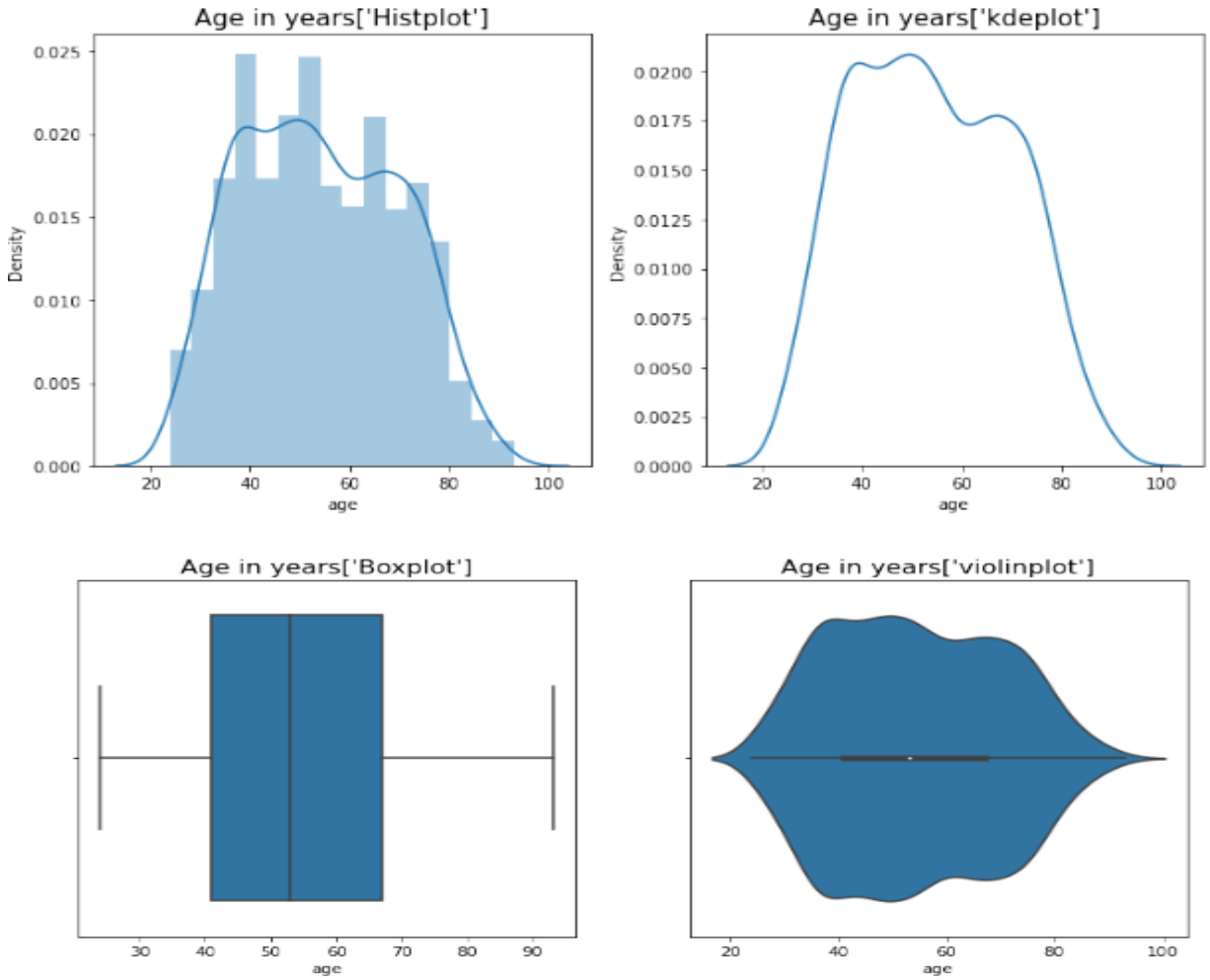


Figure-5a: Univariate Analysis on continuous [Age] variable

- Age column do not contain any outliers.
- Age column has some amount of skewness, hence the data is fairly symmetrical.

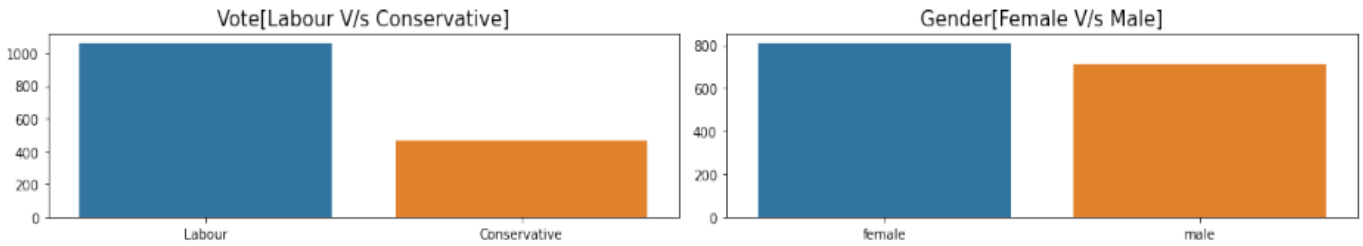
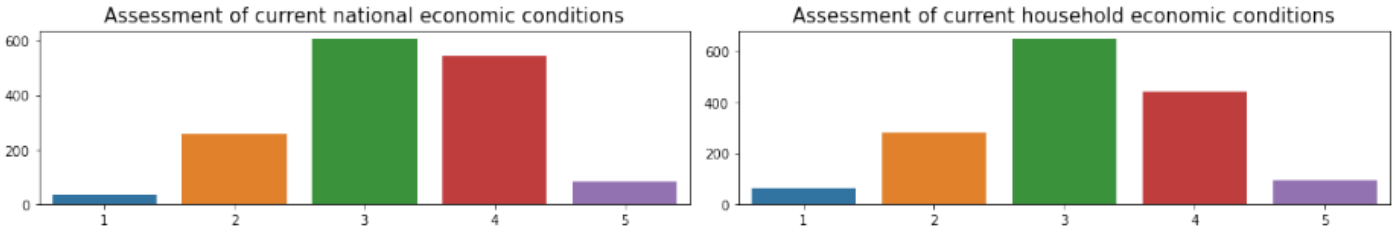


Figure-5b: Univariate Analysis on Nominal variable

- Frequency of labour party is very high compared to conservative party.
- Frequency of Female voters are slightly high compared to male voters.



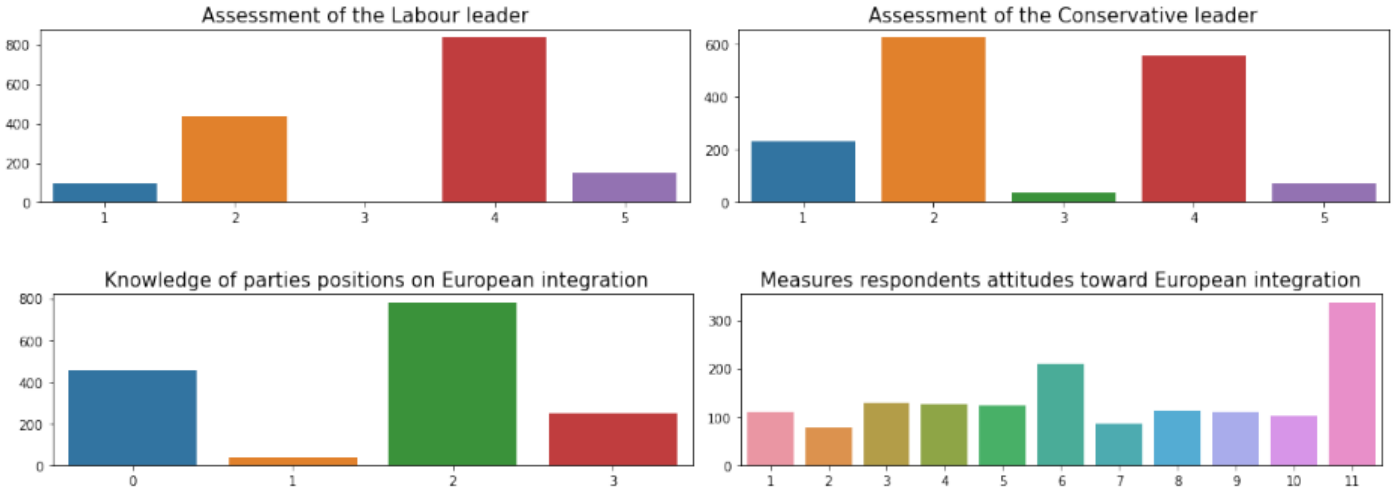


Figure-5c: Univariate Analysis on Ordinal variable

```
VOTE :  2
Conservative     462
Labour          1063
Name: vote, dtype: int64
```

```
GENDER :  2
female    812
male      713
Name: gender, dtype: int64
```

Figure-6a: Unique Count of Nominal variables

```
ECONOMIC.COND.NATIONAL :  5
1     37
2    257
3    607
4    542
5     82
Name: economic.cond.national, dtype: int64
```

```
ECONOMIC.COND.HOUSEHOLD :  5
1     65
2    280
3    648
4    440
5     92
Name: economic.cond.household, dtype: int64
```

```
BLAIR :  5
1     97
2    438
3      1
4    836
5    153
Name: Blair, dtype: int64
```

```
HAGUE :  5
1    233
2    624
3     37
4    558
5     73
Name: Hague, dtype: int64
```

```
EUROPE :  11
1     109
2      79
3     129
4     127
5     124
6     209
7      86
8     112
9     111
10    101
11    338
Name: Europe, dtype: int64
```

```
POLITICAL.KNOWLEDGE :  4
0    455
1     38
2    782
3    250
Name: political.knowledge, dtype: int64
```

Figure-6b: Unique Count of Ordinal variables

- Majority of the voters gave rating 3 and 4 for the current national and household economic conditions.
- Majority of the voters gave rating 4 for labour party leader Blair.
- For conservative party leader Hague, most of the voters choose to rate 2 and 4.
- Majority of the voters represents Eurosceptic sentiment.

Bi-variate Analysis and Multi-variate Analysis

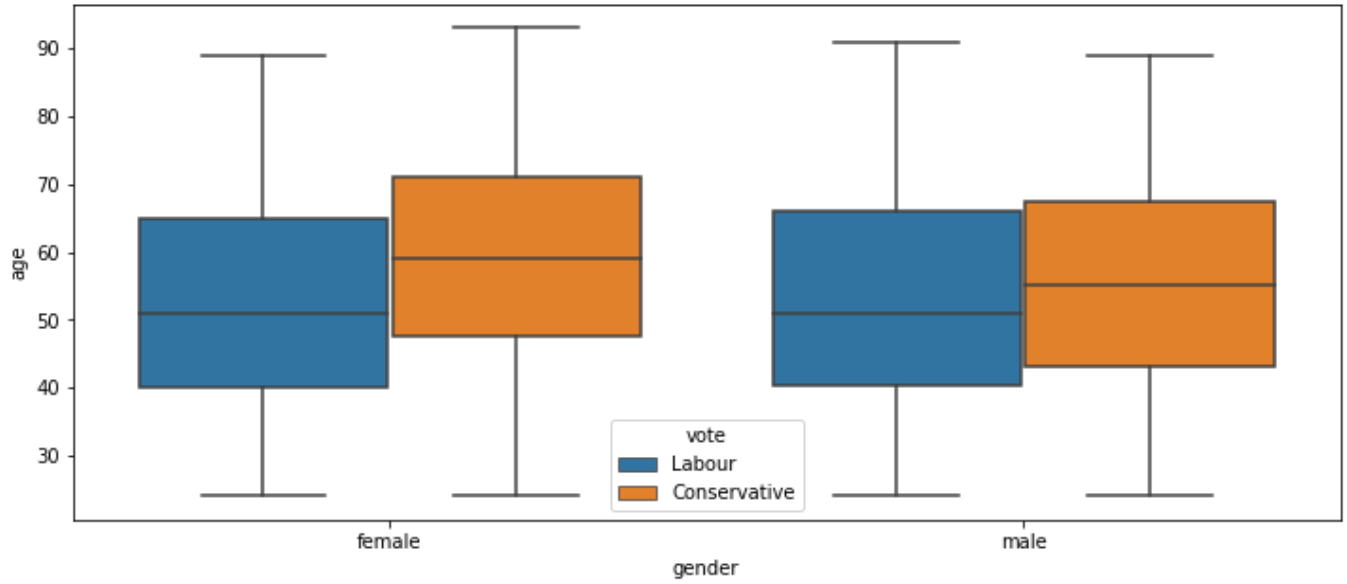
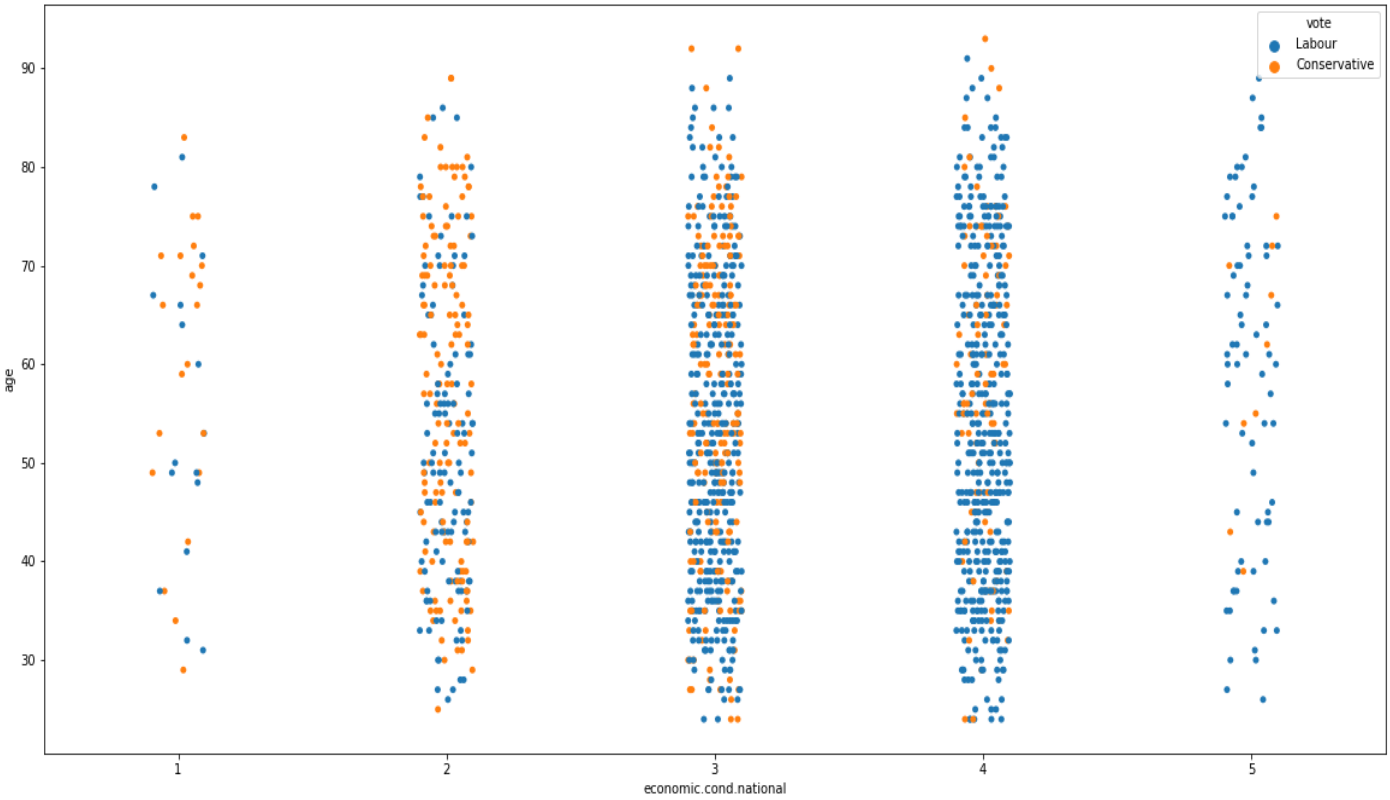


Figure-7a: One numeric v/s Nominal categorical variables

- Average age group near to 60 choose to elect conservative party whereas average age group near to 50 choose to elect Labour party.
- Middle 50% of the male and female people who fall under the age group within the range of 40 to 65 choose to elect Labour party.
- Middle 50% of the male and female people who fall under the age group within the range of 50 to 70 choose to elect conservative party.
- So we can see from the above boxplot that has the age increases people choose to elect conservative party.



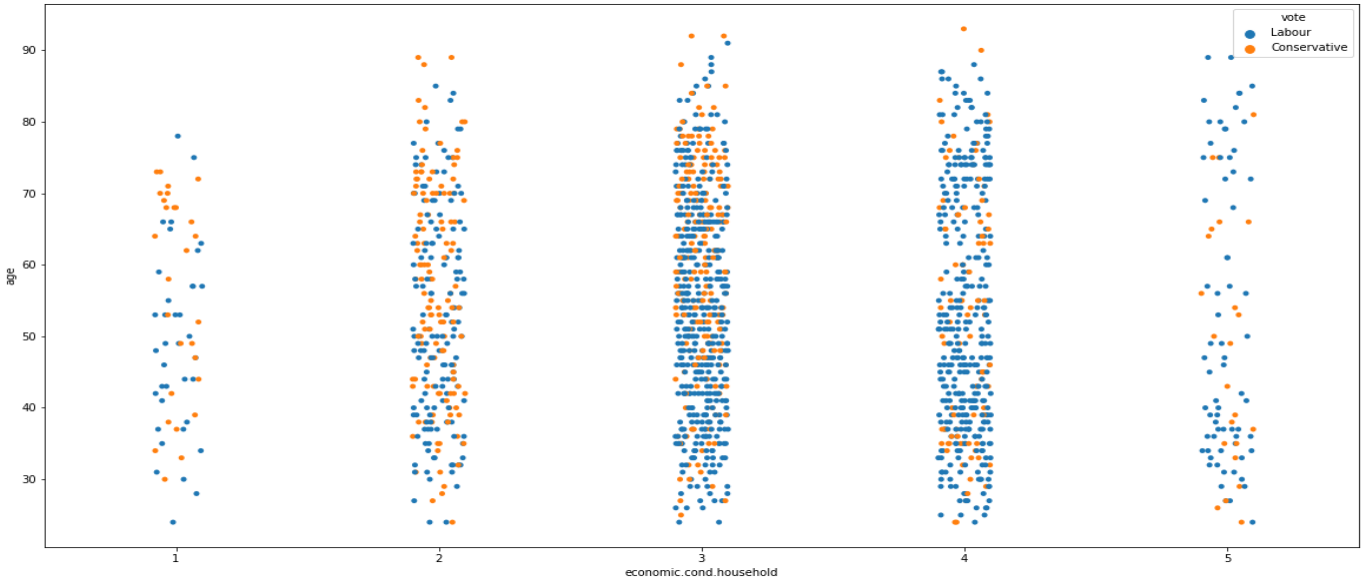
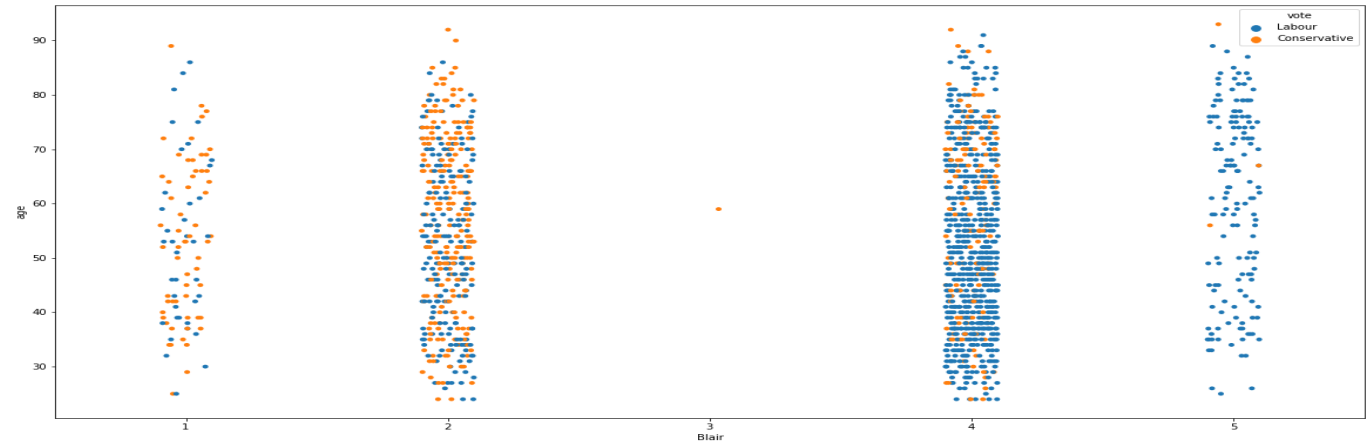


Figure-7b: One numeric v/s Ordinal categorical variables

- From the above strip plot, we could see that rating 3 and rating 4 were given by majority of the voters for the assessment of economic national as well as household conditions and most of them choose to elect labour party.
- Also we could say that voters who gave rating 1 or 2 choose select conservative party.



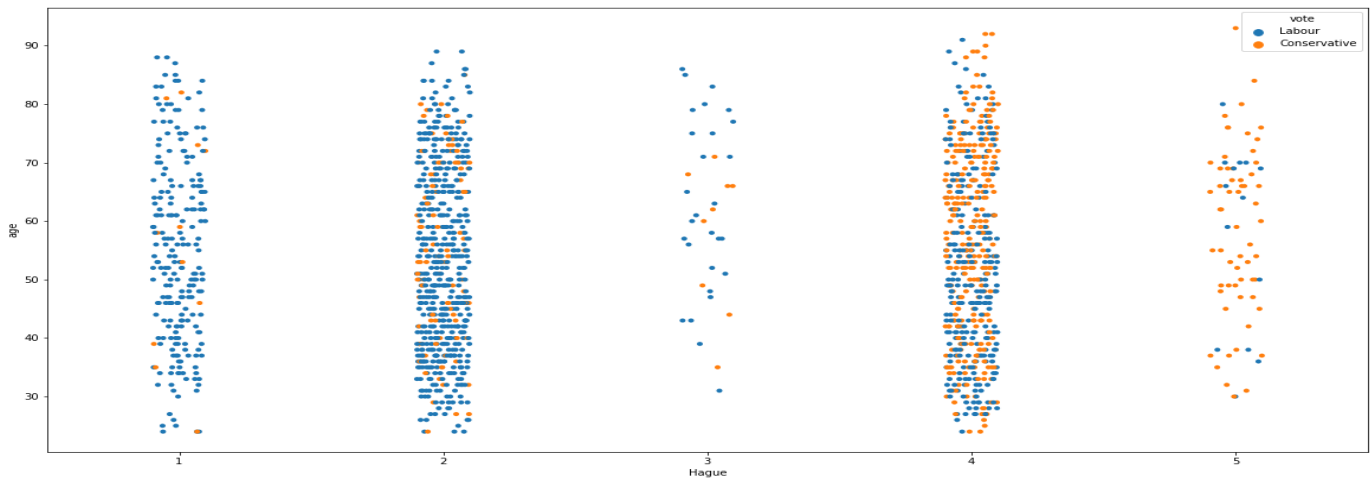


Figure-7c: One numeric v/s Ordinal categorical variables

- It's clear from the above strip plot, voters who gave rating 4 and 5 for the assessment of labour party leader Blair, choose to elect labour party.
- Also, it's very clear, voters who gave rating 3 and above for the assessment of conservative party leader Hague, choose to elect conservative party.

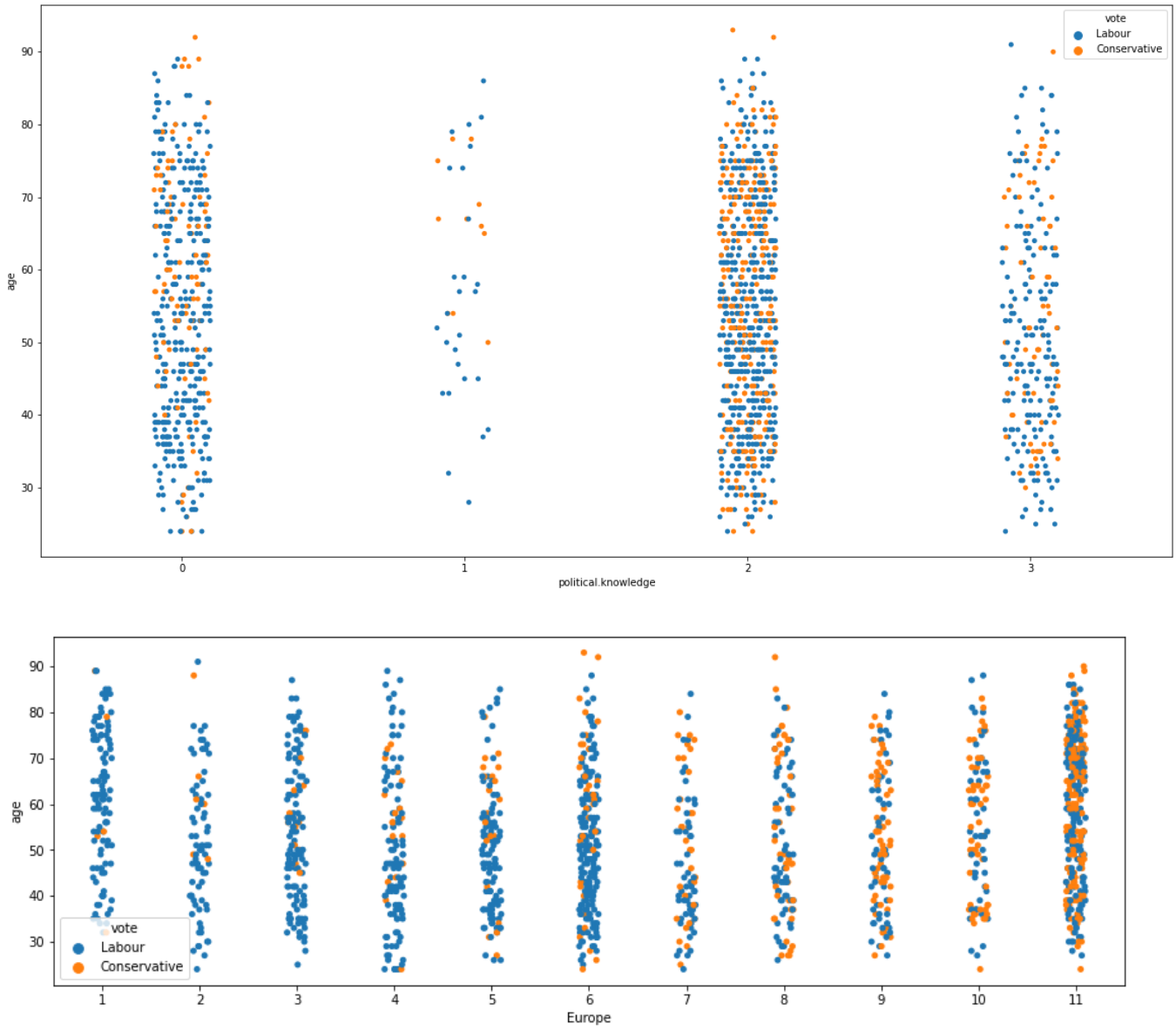


Figure-7d: One numeric v/s Ordinal categorical variables

- Voters who choose to vote for labour party do not possess a very high Eurosceptic sentiment, whereas voters who represent high Eurosceptic sentiment vote for conservative party.

Data Preparation:

1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).

Solution:

Data Balance:

vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
Conservative	462	462	462	462	462	462	462	462
Labour	1063	1063	1063	1063	1063	1063	1063	1063

Figure-8a: Data balance for target column

```
Labour          0.697049
Conservative    0.302951
Name: vote, dtype: float64
```

Figure-8b: Data balance for target column

- Distribution of Target variable is 70:30.
- Most of the voters elected to labour party. The ratio is almost 1:2
- The model's ability to predict class labour will be more compared to conservative.

Encode the Data:

- For a given dataset, we have 8 categorical variables, 6 variables which are of ordinal data type already integer data type, remaining 2 nominal variables gender and the target variable vote with the type as object, needs to be converted into categorical data type.

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


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

Figure-9: Encode the Data (having string values)

- After encoding the data, the target variable 'vote' was captured in to separate vector for training and test data set.
- Then, the data was split into train and test in the ratio of 70:30.

```
Number of rows and columns of the training set for the independent variables: (1067, 8)
Number of rows and columns of the training set for the dependent variable: (1067,)
Number of rows and columns of the test set for the independent variables: (458, 8)
Number of rows and columns of the test set for the dependent variable: (458,)
```

Figure-10: Data-Split [70:30]

Scaling:

- We have feature age in years with different unit weight and the remaining is of ratings ranging from 1 to 5; 1 to 11 and 0 to 3. Hence scaling is required for certain models to get accurate results.
- Scaled data can be done only for training set, for test set scaling is not required, because in the real world data will not be scaled, need to be passed has it is through the model with whatever measurements they come in.

Modelling:

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

Solution:

Apply Logistic Regression:

Feature scaling doesn't make a much effect for logistic regression test data accuracy, hence scaling is not performed for logistic regression.

Build a Logistic regression model using a grid search cross validation to get best parameter/ estimators for a given dataset.

```
{'max_iter': 10000, 'penalty': 'l1', 'solver': 'liblinear', 'tol': 0.0001}

LogisticRegression(max_iter=10000, n_jobs=4, penalty='l1', solver='liblinear',
                   verbose=True)
```

Figure-11: Best Estimators for logistic Regression

With the above parameter setting values were predicted for both training and test dataset.

Accuracy of the logistic regression model for train set is 83% and accuracy of the logistic regression model for test set is 85.4%. The model is a valid since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Apply Linear Discriminant Analysis:

Feature scaling doesn't make a much effect for LDA test data accuracy, hence scaling is not performed for LDA model.

Linear Discriminant analysis was built without any specific parameter setting and then values were predicted on both training and test dataset.

Accuracy of the LDA model for train set is 82.6% and accuracy of the LDA model for test set is 84.5%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

On comparing accuracy of both the models LR and LDA models, Test data accuracy of logistic regression is quite good compared to LDA model. Hence Logistic regression perform well for predicting Labour or conservative party.

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

Solution:

Apply Gaussian Naïve Bayes:

For Naive Bayes algorithm while calculating likelihoods of numerical features it assumes that the feature to be normally distributed and then we calculate probability using mean and variance of that feature only and also it assumes that all the predictors are independent to each other. Feature scaling doesn't matter. Performing a features scaling in this algorithms may not have much effect.

Now build a GaussianNB classifier. Then the classifier is trained using training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method with test set features as its parameters.

Accuracy of GaussianNB model for train set is 82.2% and accuracy of GaussianNB model for test set is 84.7%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Apply K-nearest neighbour [KNN-model]:

In order to have good KNN model, data requires pre-processing to make all independent variables similarly scaled and centred. Hence need to perform z-score on all numeric attributes in models that calculate distance and see the performance for KNN.

By default, value of n_neighbors=5, in order to get best KNN model need to try for different K values and find out for the corresponding k-value which is the least Misclassification error.

Misclassification error (MCE) = 1 - Test accuracy score

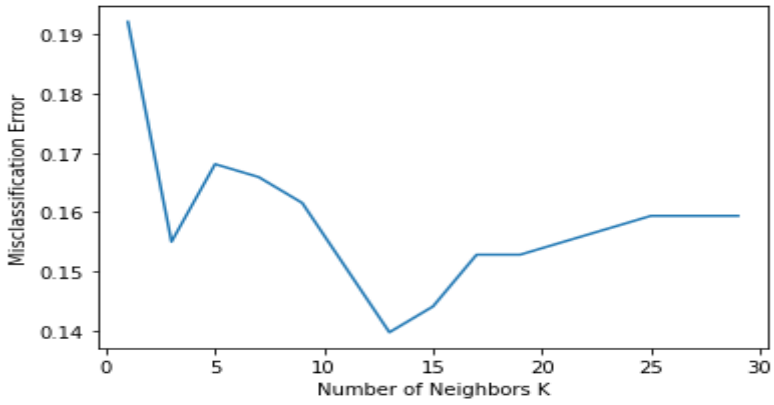


Figure-12: Plot misclassification error vs k (with k value on X-axis)

From the above plot the optimum number of neighbour or k-value is found to be 13.

Now build a KNeighborsClassifier using the value of n_neighbors=13 and metric= 'Euclidean'. Then the classifier is trained using scaled training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method with test set features as its parameters.

Accuracy of KNN model for train set is 84.25% and accuracy of KNN model for test set is 86%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

On comparing accuracy of both the models NB and KNN models, Test data accuracy of KNN model for n_neighbour=13 is quite good compared to NB model. Hence KNN model perform well for predicting Labour or conservative party.

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

Solution:

Apply Ensemble Random Forest model:

Feature scaling is not required for random forest model, hence scaling is not performed.

Build a Random Forest model using a grid search cross validation to get best parameter/ estimators for a given dataset.

```
{'max_depth': 7, 'max_features': 4, 'min_samples_leaf': 40, 'min_samples_split': 100, 'n_estimators': 501}

RandomForestClassifier(max_depth=7, max_features=4, min_samples_leaf=40,
                       min_samples_split=100, n_estimators=501,
                       random_state=27)
```

Figure-13: Best Estimators for Random-Forest

With the above parameter setting values were predicted for both training and test dataset.

Accuracy of the random forest model for train set is 82.28% and test set is 83.40%. The model is a valid since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Apply Bagging using base estimator has Random Forest:

Bagging was built using the above tuned random forest has a base estimator and n_estimators=100. Then the classifier is trained using training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method with test set features as its parameters.

```
BaggingClassifier(base_estimator=RandomForestClassifier(max_depth=7,
                                                        max_features=4,
                                                        min_samples_leaf=40,
                                                        min_samples_split=100,
                                                        n_estimators=501,
                                                        random_state=27),
                  n_estimators=100, random_state=27)
```

Figure-14: Best Estimators for Random-Forest

Accuracy of the Bagging model for train set is 81.17% and test set is 82.97%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

On comparing accuracy of Random Forest and Bagging models, Test data accuracy of Random Forest model quite good compared to Bagging model. Hence Random Forest model performs well for predicting Labour or conservative party.

Apply Gradient Boosting:

Gradient Boosting was built with n_estimators=100 and then classifier is trained using training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method.

Accuracy of the Gradient boosting model for train set is 88.8% and accuracy of the Gradient model for test set is 83.8%. The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Apply Ada Boosting:

AdaBoost was built with n_estimators=100 and then classifier is trained using training data for which we can use fit() method for training it. After building a classifier, model is ready to make predictions for which we can use predict() method.

Accuracy of the Adaboost model for train set is 84.4% and test set is 83.6%.The model is a valid model, since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

On comparing accuracy of Gradient Boosting and AdaBoost models, Test data accuracy of both the models are almost equal, since the difference between the train and test set is very low for Ada boost compared to Gradient Boost model, Ada Boost model performs well for predicting Labour or conservative party.

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.

Solution:

1. **Model Evaluation-Logistic Regression:**

- Accuracy on training set is 83% and on testing set is 85.4%.
- **Classification report:**

Classification report for Logistic Regression model on Training set is

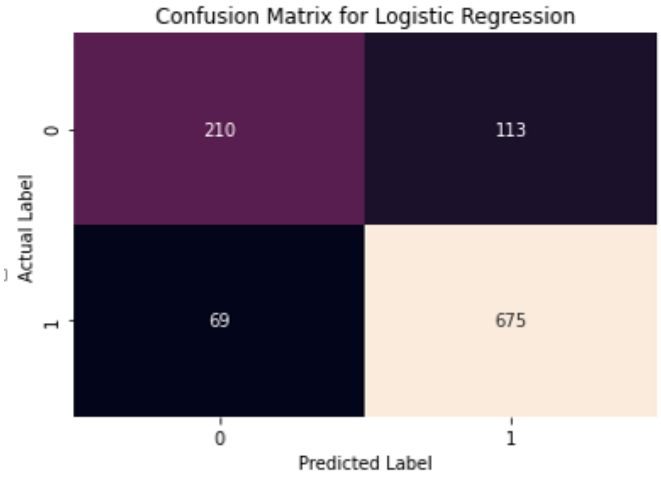
	precision	recall	f1-score	support
0	0.75	0.65	0.70	323
1	0.86	0.91	0.88	744
accuracy			0.83	1067
macro avg	0.80	0.78	0.79	1067
weighted avg	0.83	0.83	0.83	1067

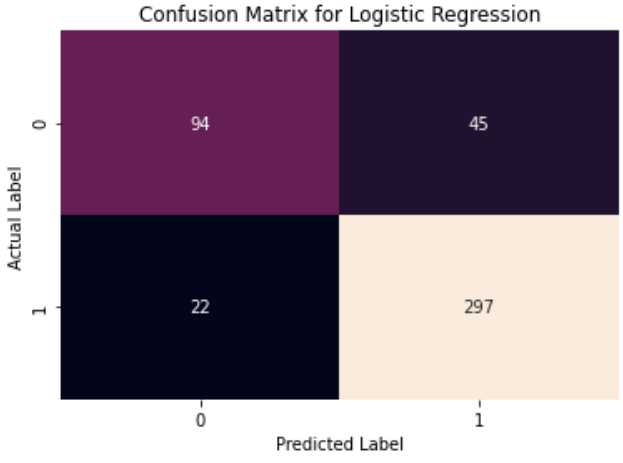
Classification report for Logistic Regression model on Testing set is

	precision	recall	f1-score	support
0	0.81	0.68	0.74	139
1	0.87	0.93	0.90	319
accuracy			0.85	458
macro avg	0.84	0.80	0.82	458
weighted avg	0.85	0.85	0.85	458

Figure-15: Classification Report on Train and Test set

- **Confusion Matrix:**





LR_train_precision 0.86
LR_train_recall 0.91
LR_train_f1 0.88

LR_test_precision 0.87
LR_test_recall 0.93
LR_test_f1 0.9

Figure-16: Confusion matrix on Train and Test set

- AUC score on the train dataset is 87.7% and on test dataset is 91.5%.
- **ROC Curve:**

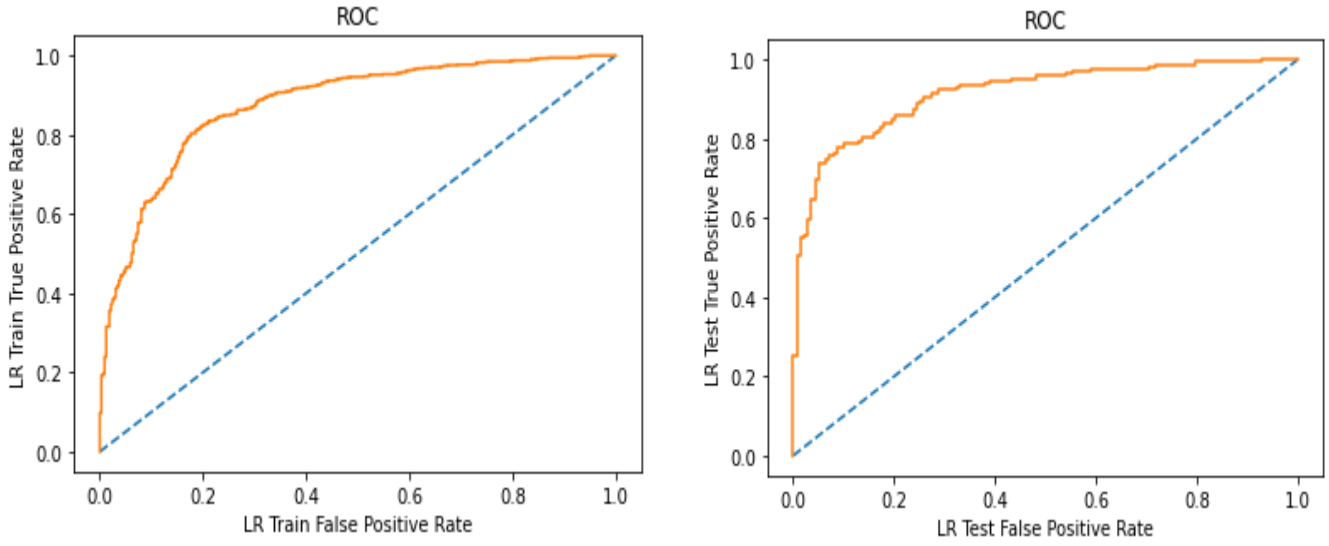


Figure-17: ROC Curve on Train and Test set

- **Logistic Regression Output:**

Model	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	83	87.7	86	91	88
Test set	85.4	91.6	87	93	90

Figure-18: Performance metrics Output

From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

2. Model Evaluation-Linear Discriminant Analysis:

- Accuracy on train set is 82.6% and on test set is 84.5%.
- **Classification report:**

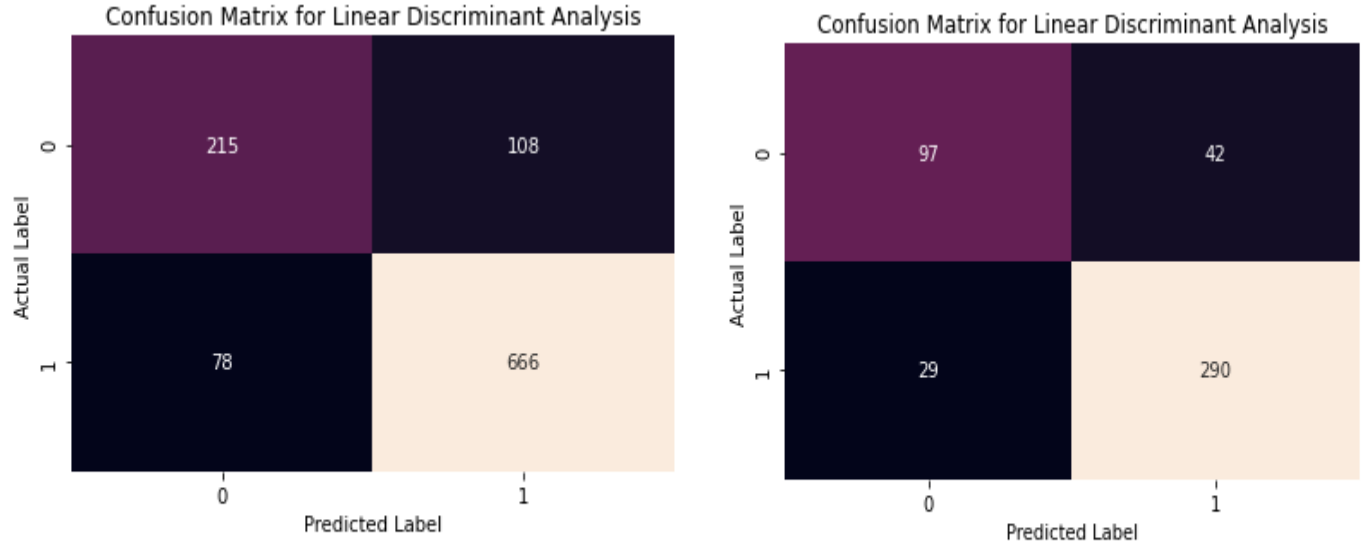
```
Classification report for Linear Discriminant Analysis model on Training set is
              precision    recall  f1-score   support

           0       0.73      0.67      0.70       323
           1       0.86      0.90      0.88       744

    accuracy                           0.83      1067
   macro avg       0.80      0.78      0.79      1067
weighted avg       0.82      0.83      0.82      1067
```

Classification report for Linear Discriminant Analysis model on Testing set is

	precision	recall	f1-score	support
0	0.77	0.70	0.73	139
1	0.87	0.91	0.89	319
accuracy			0.84	458
macro avg	0.82	0.80	0.81	458
weighted avg	0.84	0.84	0.84	458

Figure-19: Classification Report on Train and Test set

- **Confusion Matrix:**



LDA_train_precision 0.86
LDA_train_recall 0.9
LDA_train_f1 0.88

LDA_test_precision 0.87
LDA_test_recall 0.91
LDA_test_f1 0.89

Figure-20: Confusion matrix on Train and Test set

- AUC score on the train dataset is 87.6% and on test dataset is 91.5%.
- **ROC Curve:**

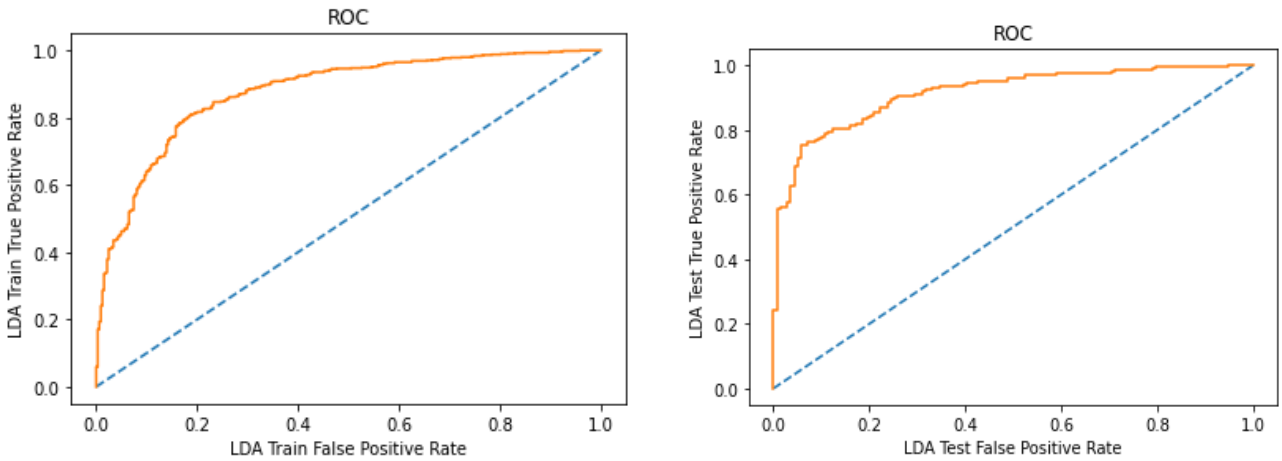
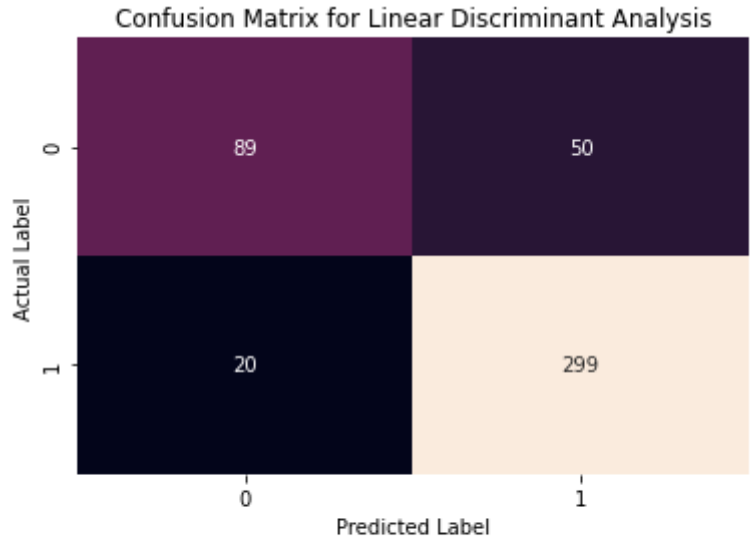


Figure-21: ROC Curve on Train and Test set

By changing the probability cut-off values; we see that 0.4 and 0.5 gives better accuracy than the rest of the custom cut-off values. But 0.4 cut-off gives us the best 'f1-score'. Hence, we will take the cut-off as 0.4 to get the optimum 'f1' score in order to improve the test set results.

- Accuracy for LDA model on testing set with cut-off value 0.4 is 84.7
- **Confusion matrix:**




```
LDA_test_precision_new  0.86
LDA_test_recall_new  0.94
LDA_test_f1_new  0.9
```

Figure-22: Confusion Matrix for LDA model on Testing set with cut-off value 0.4

- **LDA Output:**

Model	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	82.6	87.6	86	90	88
Test set	84.5	91.5	87	91	89
Improved Test set	84.7	Nil	86	94	90

Figure-23: Performance metrics Output

→ From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
→ By changing the probability cut-off value from default 0.5 to 0.4, we could see that the precision has improved from 91% to 94% on the test set and the model accuracy is of 84.7%
→ Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

3. Model Evaluation-Gaussian Naïve Bayes:

- Accuracy on train set is 82.2% and on test set is 84.7%.
- **Classification report:**

Classification report for Naive Bayes model on Training set is

	precision	recall	f1-score	support
0	0.71	0.69	0.70	323
1	0.87	0.88	0.87	744
accuracy			0.82	1067
macro avg	0.79	0.78	0.79	1067
weighted avg	0.82	0.82	0.82	1067

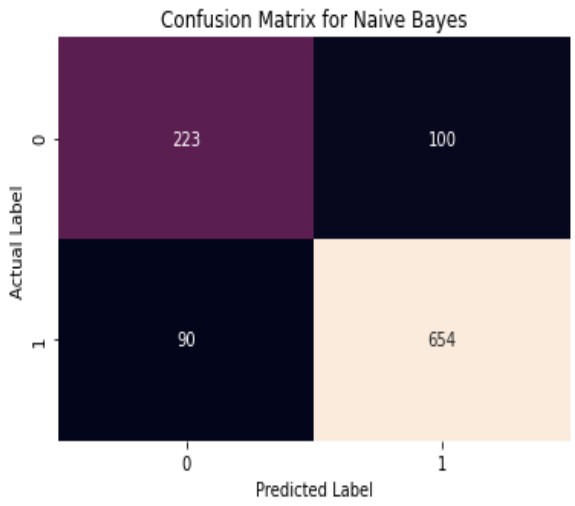
Classification report for Naive Bayes model on Testing set is

	precision	recall	f1-score	support
0	0.76	0.73	0.74	139
1	0.88	0.90	0.89	319
accuracy			0.85	458
macro avg	0.82	0.81	0.82	458
weighted avg	0.85	0.85	0.85	458

Figure-24: Classification Report on Train and Test set

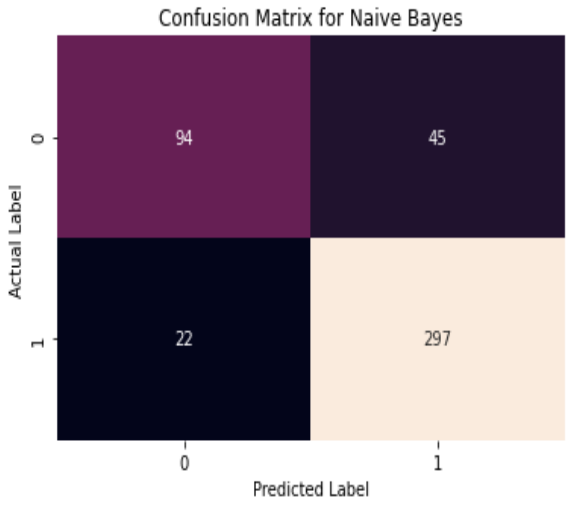
- **Confusion Matrix:**

Confusion Matrix for Naive Bayes model on Training set is



NB_train_precision 0.87
NB_train_recall 0.88
NB_train_f1 0.87

Confusion Matrix for Naive Bayes model on Testing set is



NB_test_precision 0.88
NB_test_recall 0.9
NB_test_f1 0.89

Figure-25: Confusion matrix on Train and Test set

- AUC score on the train dataset is 87.4% and on test dataset is 91%.
- **ROC Curve:**

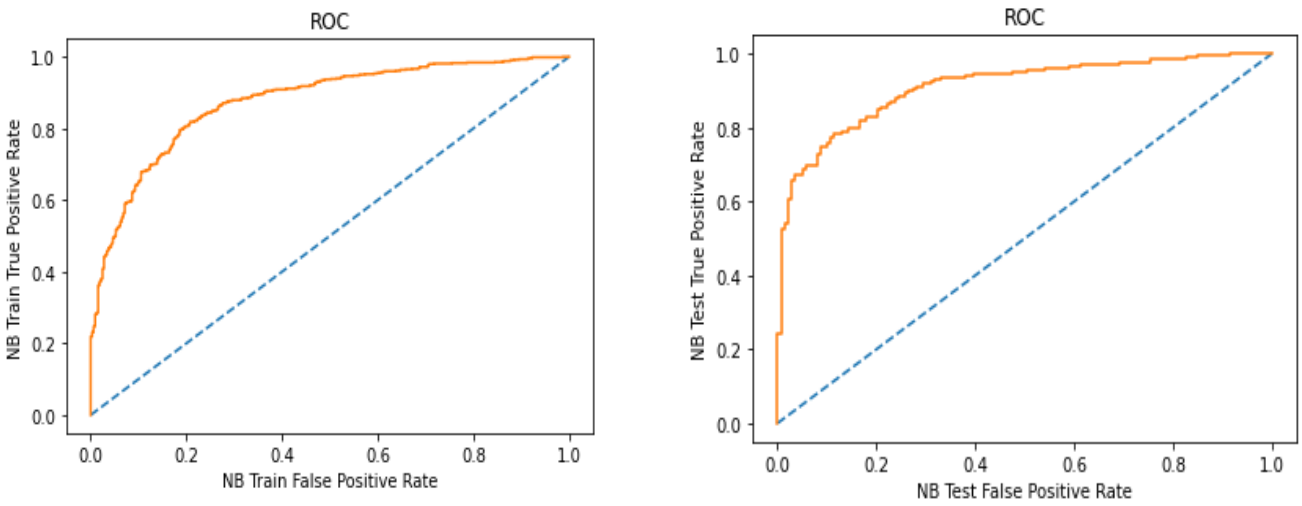
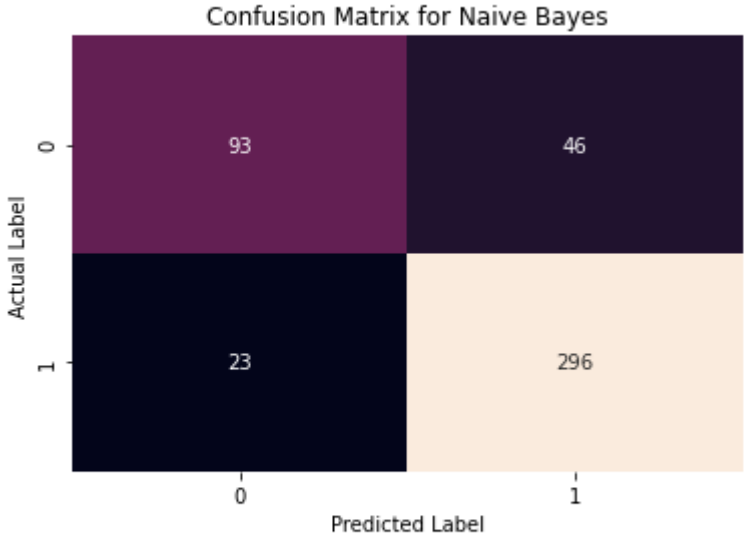


Figure-26: ROC Curve on Train and Test set

By changing the probability cut-off values; we see that 0.4 and 0.5 gives better accuracy than the rest of the custom cut-off values. But 0.4 cut-off gives us the best 'f1-score'. Hence, we will take the cut-off as 0.4 to get the optimum 'f1' score in order to improve the test set results.

- Accuracy for LDA model on testing set with cut-off value 0.4 is 84.9
- **Confusion matrix:**

```
Confusion Matrix for Naive Bayes model on Testing set with cut-off value 0.4 is
```




```
NB_test_precision_new  0.87
NB_test_recall_new  0.93
NB_test_f1_new  0.9
```

Figure-27: Confusion Matrix for Gaussian Naïve Bayes model on Testing set with cut-off value 0.4

- **Gaussian Naïve Bayes Output:**

Model	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	82.2	87.4	87	88	87
Test set	84.7	91	88	90	89
Improved Test set	84.9	-	87	93	90

Figure-28: Performance metrics Output

→ From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
→ By changing the probability cut-off value from default 0.5 to 0.4, we could see that the precision has improved from 90% to 93% on the test set and the model accuracy is of 84.9%
→ Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

4. Model Evaluation- KNN:

- Accuracy on train set is 84.3% and on test set is 86%.
- **Classification report:**

Classification report for K-nearest neighbour model on Training set is

	precision	recall	f1-score	support
0	0.77	0.69	0.73	323
1	0.87	0.91	0.89	744
accuracy			0.84	1067
macro avg	0.82	0.80	0.81	1067
weighted avg	0.84	0.84	0.84	1067

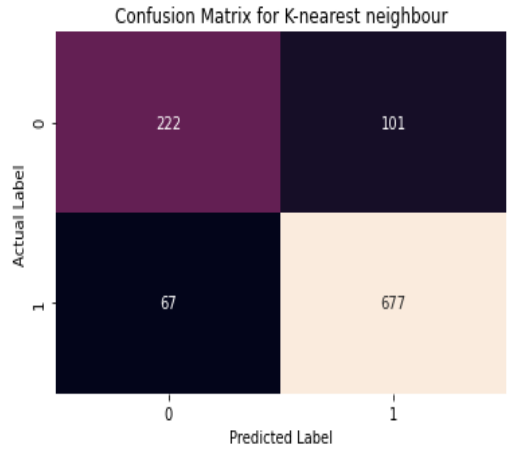
Classification report for K-nearest neighbour model on Test set is

	precision	recall	f1-score	support
0	0.80	0.73	0.76	139
1	0.89	0.92	0.90	319
accuracy			0.86	458
macro avg	0.84	0.82	0.83	458
weighted avg	0.86	0.86	0.86	458

Figure-29: Classification Report on Train and Test set

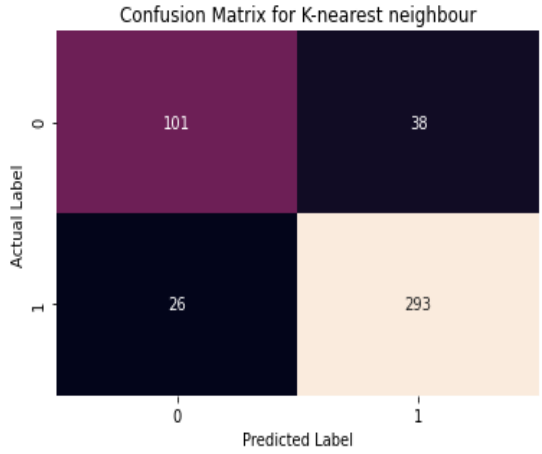
- **Confusion Matrix:**

Confusion Matrix for K-nearest neighbour model on Training set is



KNN_train_precision 0.87
KNN_train_recall 0.91
KNN_train_f1 0.89

Confusion Matrix for K-nearest neighbour model on Test set is



KNN_test_precision 0.89
KNN_test_recall 0.92
KNN_test_f1 0.9

Figure-30: Confusion matrix on Train and Test set

- AUC score on the train dataset is 91.1% and on test dataset is 89.4%.
- **ROC Curve:**

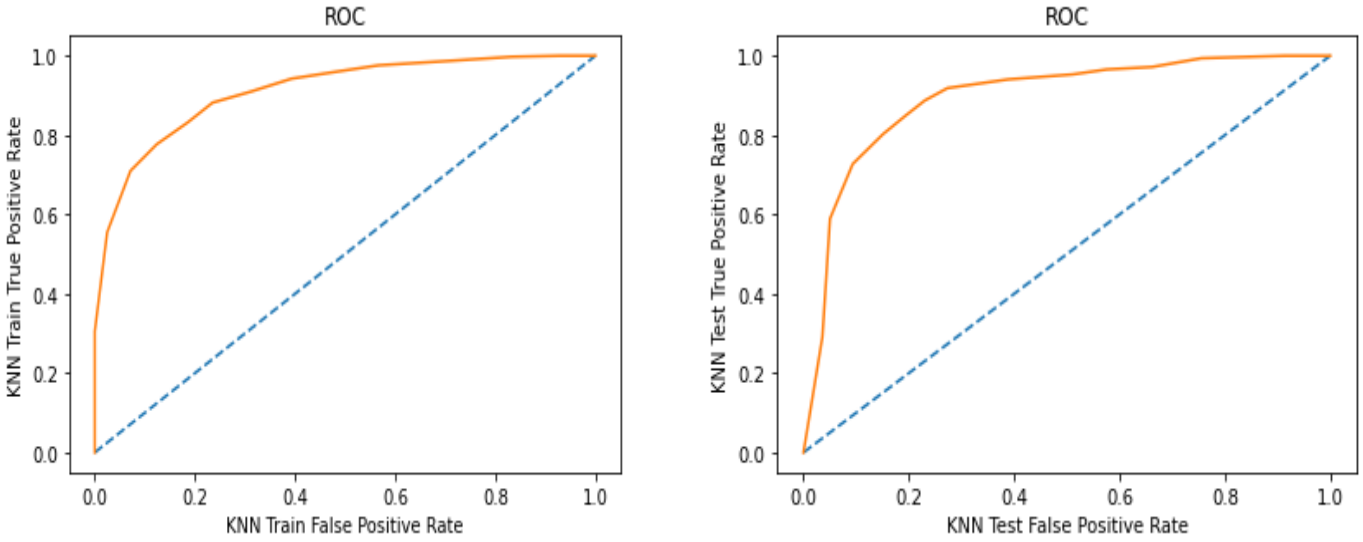


Figure-31: ROC Curve on Train and Test set

- **KNN Output:**

Model	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	84.3	91.1	87	91	89
Test set	86	89.4	89	92	90

Figure-32: Performance metrics Output

→ From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
→ Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

5. Model Evaluation- Random Forest:

- Accuracy on train set is 82.3% and on test set is 83.4%.
- **Classification report:**

Classification report for RandomForestClassifier model on Training set is

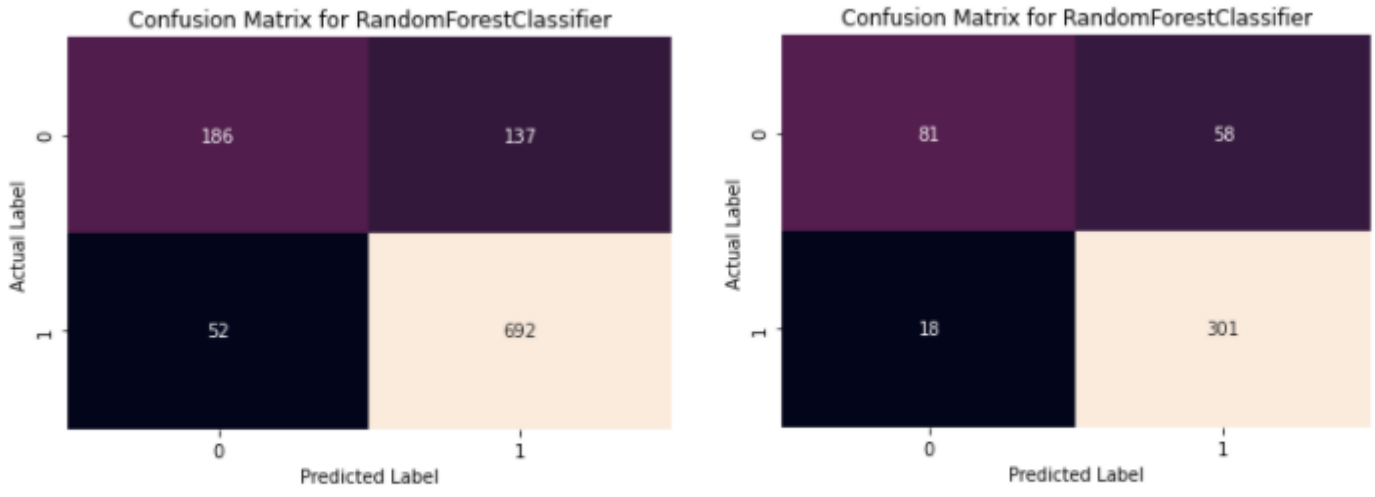
	precision	recall	f1-score	support
0	0.78	0.58	0.66	323
1	0.83	0.93	0.88	744
accuracy			0.82	1067
macro avg	0.81	0.75	0.77	1067
weighted avg	0.82	0.82	0.81	1067

Classification report for RandomForestClassifier model on Test set is

	precision	recall	f1-score	support
0	0.82	0.58	0.68	139
1	0.84	0.94	0.89	319
accuracy			0.83	458
macro avg	0.83	0.76	0.78	458
weighted avg	0.83	0.83	0.83	458

Figure-33: Classification Report on Train and Test set

- **Confusion Matrix:**



RF_train_precision 0.83
RF_train_recall 0.93
RF_train_f1 0.88

RF_test_precision 0.84
RF_test_recall 0.94
RF_test_f1 0.89

Figure-34: Confusion matrix on Train and Test set

- AUC score on the train dataset is 88.8% and on test dataset is 90%.
- **ROC Curve:**

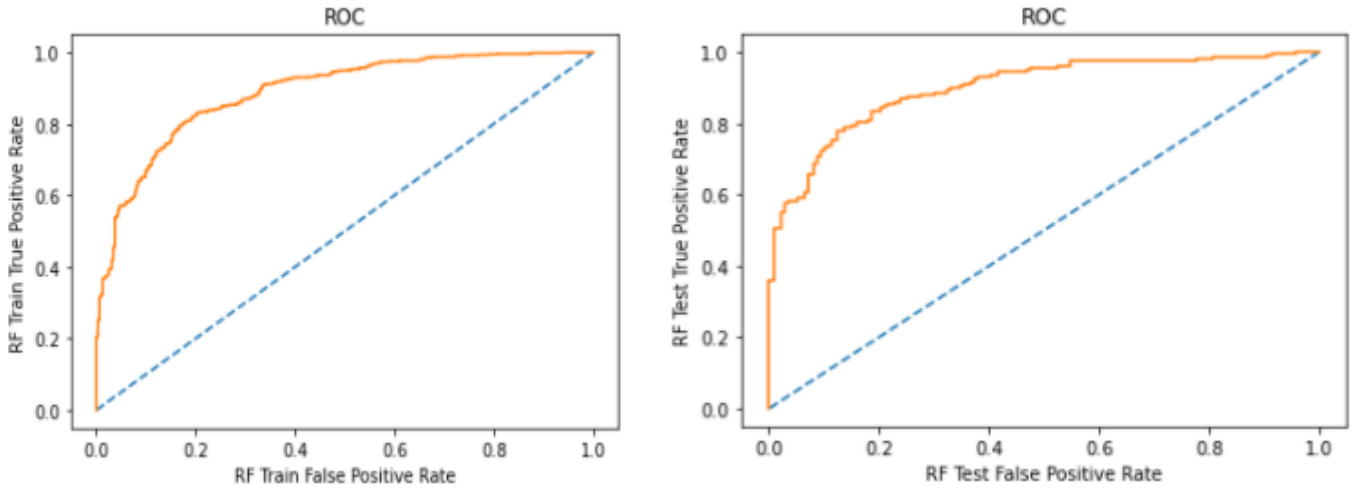


Figure-35: ROC Curve on Train and Test set

- **Random Forest Output:**

Model	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	82.3	88.8	83	93	88
Test set	83.4	90	84	94	89

Figure-36: Performance metrics Output

→ From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
→ Also, the model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

6. Model Evaluation- Bagging:

- Accuracy on train set is 81.2% and on test set is 83%.
- **Classification report:**

Classification report for BaggingClassifier model on Training set is

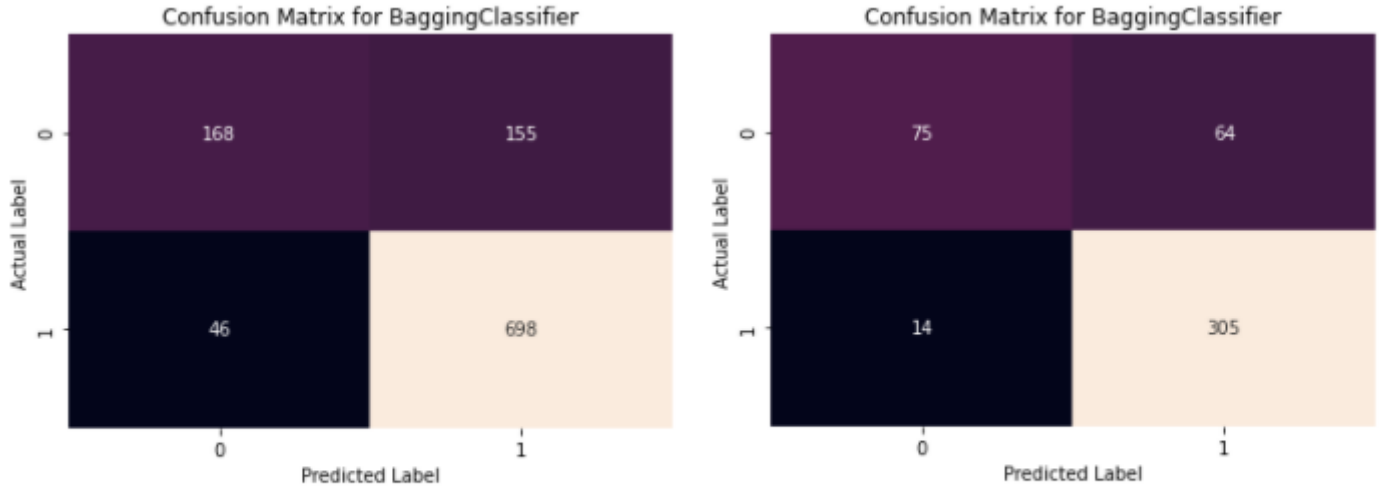
	precision	recall	f1-score	support
0	0.79	0.52	0.63	323
1	0.82	0.94	0.87	744
accuracy			0.81	1067
macro avg	0.80	0.73	0.75	1067
weighted avg	0.81	0.81	0.80	1067

Classification report for BaggingClassifier model on Test set is

	precision	recall	f1-score	support
0	0.84	0.54	0.66	139
1	0.83	0.96	0.89	319
accuracy			0.83	458
macro avg	0.83	0.75	0.77	458
weighted avg	0.83	0.83	0.82	458

Figure-37: Classification Report on Train and Test set

- **Confusion Matrix:**



BG_train_precision 0.82
BG_train_recall 0.94
BG_train_f1 0.87

BG_test_precision 0.83
BG_test_recall 0.96
BG_test_f1 0.89

Figure-38: Confusion matrix on Train and Test set

- AUC score on the train dataset is 88.1% and on test dataset is 89.7%.
- **ROC Curve:**

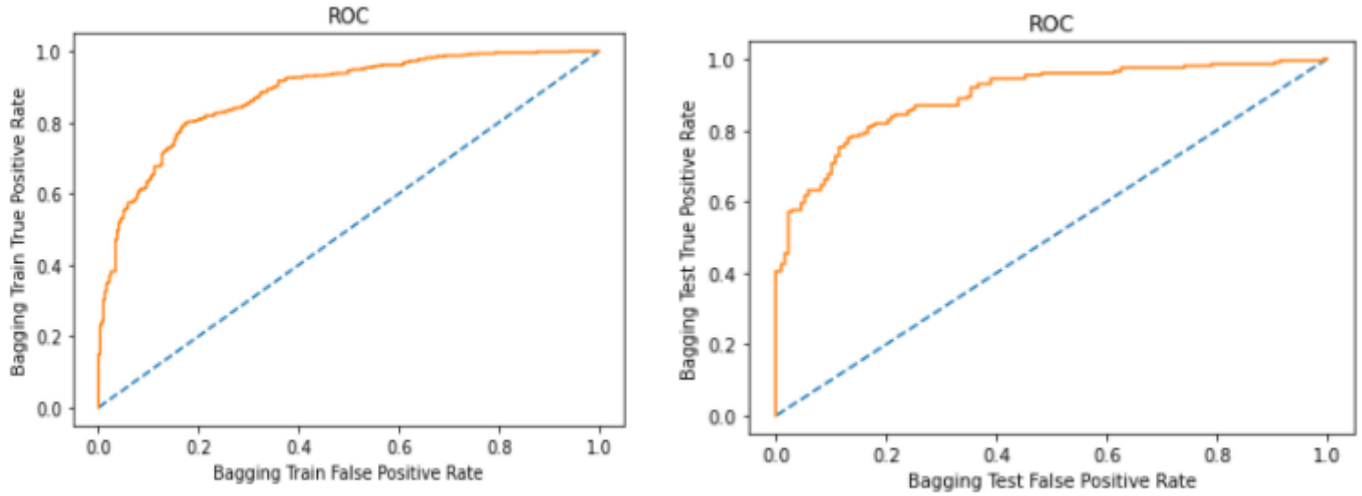


Figure-39: ROC Curve on Train and Test set

- **Bagging Output:**

Model	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	81.2	88.1	82	94	87
Test set	83	89.7	83	96	89

Figure-40: Performance metrics Output

→ From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
→ Also, model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

7. Model Evaluation- Gradient Boosting:

- Accuracy on train set is 88.8% and on test set is 83.8%.
- **Classification report:**

Classification report for GradientBoostingClassifier model on Training set is

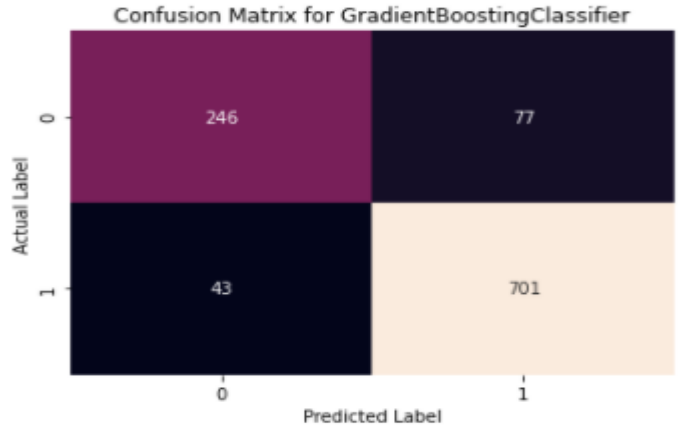
	precision	recall	f1-score	support
0	0.85	0.76	0.80	323
1	0.90	0.94	0.92	744
accuracy			0.89	1067
macro avg	0.88	0.85	0.86	1067
weighted avg	0.89	0.89	0.89	1067

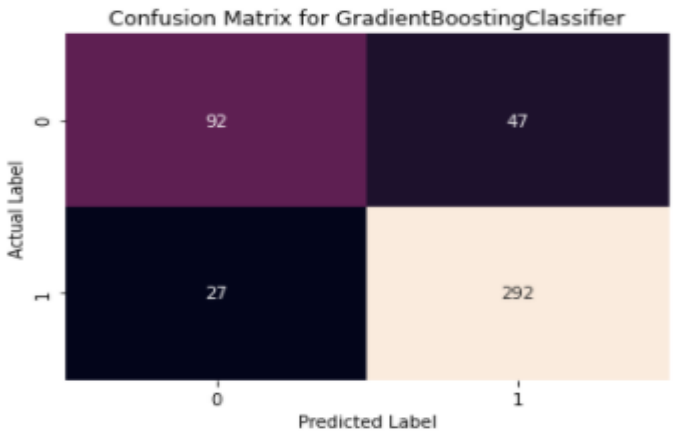
Classification report for GradientBoostingClassifier model on Test set is

	precision	recall	f1-score	support
0	0.77	0.66	0.71	139
1	0.86	0.92	0.89	319
accuracy			0.84	458
macro avg	0.82	0.79	0.80	458
weighted avg	0.83	0.84	0.83	458

Figure-41: Classification Report on Train and Test set

- **Confusion Matrix:**





GB_train_precision 0.9
GB_train_recall 0.94
GB_train_f1 0.92

GB_test_precision 0.86
GB_test_recall 0.92
GB_test_f1 0.89

Figure-42: Confusion matrix on Train and Test set

- AUC score on the train dataset is 94.8% and on test dataset is 90.8%.
- **ROC Curve:**

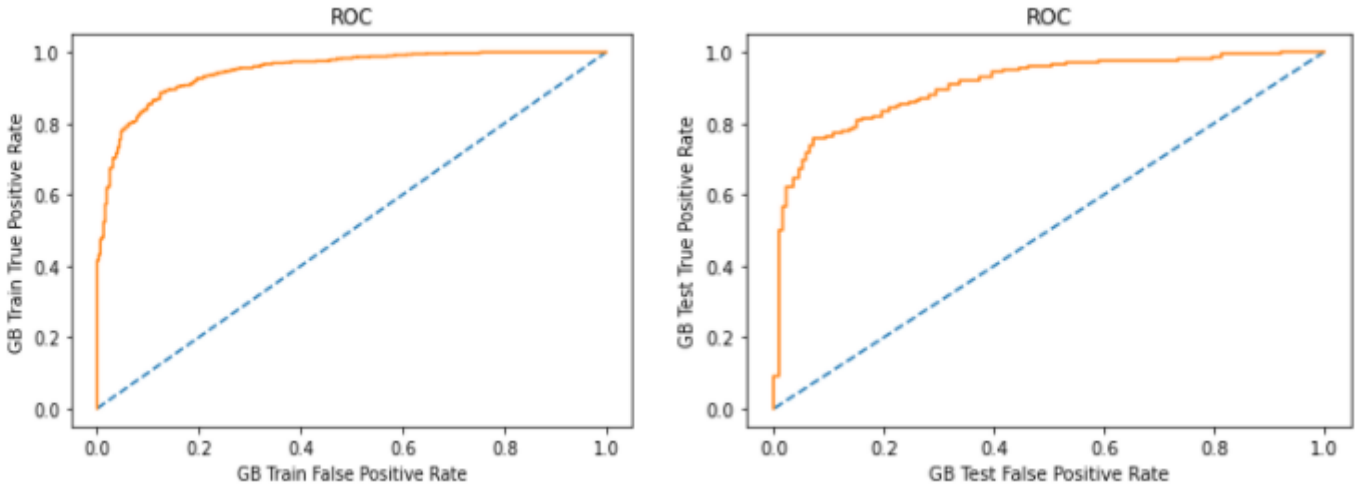


Figure-43: ROC Curve on Train and Test set

- **Gradient Boosting Output:**

Model	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	88.8	94.8	90	94	92
Test set	83.8	90.8	86	92	89

Figure-44: Performance metrics Output

→ From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.
→ Also, model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

8. Model Evaluation- Ada Boosting:

- Accuracy on train set is 84.4% and on test set is 83.6%.
- **Classification report:**

Classification report for AdaBoostClassifier model on Training set is

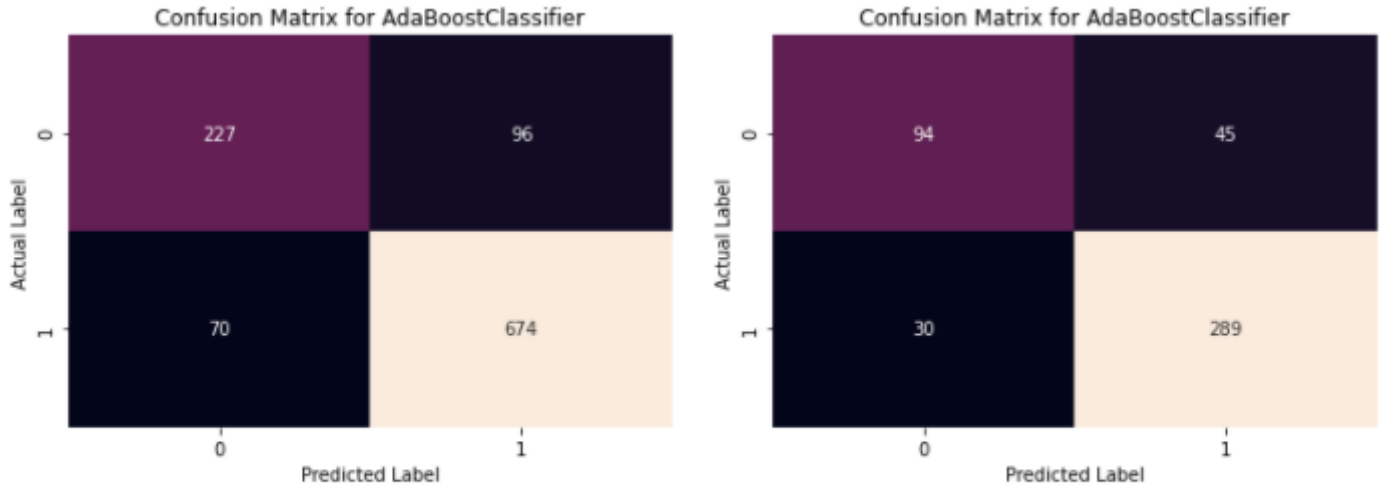
	precision	recall	f1-score	support
0	0.76	0.70	0.73	323
1	0.88	0.91	0.89	744
accuracy			0.84	1067
macro avg	0.82	0.80	0.81	1067
weighted avg	0.84	0.84	0.84	1067

Classification report for AdaBoostClassifier model on Test set is

	precision	recall	f1-score	support
0	0.76	0.68	0.71	139
1	0.87	0.91	0.89	319
accuracy			0.84	458
macro avg	0.81	0.79	0.80	458
weighted avg	0.83	0.84	0.83	458

Figure-45: Classification Report on Train and Test set

- **Confusion Matrix:**




```
AB_train_precision  0.88
AB_train_recall  0.91
AB_train_f1  0.89
```

```
AB_test_precision  0.87
AB_test_recall  0.91
AB_test_f1  0.89
```

Figure-46: Confusion matrix on Train and Test set

- AUC score on the train dataset is 90.2% and on test dataset is 90.6%.
- **ROC Curve:**

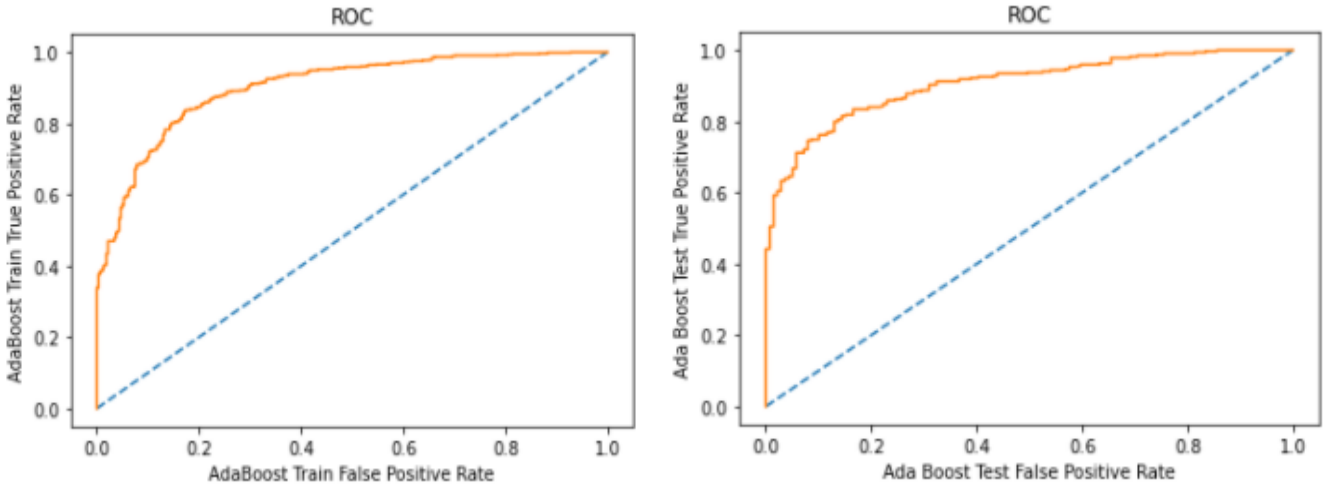


Figure-47: ROC Curve on Train and Test set

- **AdaBoost Output:**

Model	Accuracy in %	AUC in %	Precision in %	Recall in %	f1-Score
Train set	84.4	90.2	88	91	89
Test set	83.6	90.6	87	91	89

Figure-48: Performance metrics Output

→ From the above tabular output, we could clearly see that train and test set results are almost similar, and with the overall measures high, the model is a good model.

→ Also, model is a valid model since both the train and test set accuracy are almost similar and the difference between both the set lies within the range of 10%.

Inference:

1.8 Based on these predictions, what are the insights?

Solution:

	LR Train	LR Test	LDA Train	LDA Test[0.5]	LDA Test[0.4]	NB Train	NB Test[0.5]	NB Test[0.4]	KNN Train	KNN Test	RF Train	RF Test	Bagging Train	Bagging Test	Gradient-Boost Train	Gradient-Boost Test	AdaBoost Train	AdaBoost Test
Accuracy	0.829	0.854	0.826	0.845	0.847	0.822	0.847	0.849	0.843	0.860	0.823	0.834	0.812	0.830	0.888	0.838	0.844	0.836
AUC	0.877	0.915	0.876	0.915	NaN	0.874	0.910	NaN	0.911	0.894	0.888	0.900	0.881	0.897	0.948	0.908	0.902	0.906
Recall	0.910	0.930	0.900	0.910	0.940	0.880	0.900	0.930	0.910	0.920	0.930	0.940	0.940	0.960	0.940	0.920	0.910	0.910
Precision	0.860	0.870	0.860	0.870	0.860	0.870	0.880	0.870	0.870	0.890	0.830	0.840	0.820	0.830	0.900	0.860	0.880	0.870
F1 Score	0.880	0.900	0.880	0.890	0.900	0.870	0.890	0.900	0.890	0.900	0.880	0.890	0.870	0.890	0.920	0.890	0.890	0.890

Figure-49: Performance metrics Comparison of models

- By comparing recall score and precision score of all the models, Gaussian Naïve Bayes model performs well in predicting labour or conservative party.
- The test data recall of Naïve Bayes is 90% i.e only 10% of the people who is in favour of labour party, automatically he or she will be voting against the labour party.
- The test data precision of Naïve Bayes is 88% i.e only 12% of the people were made false predictions that the votes to be in favour were actually predicted against the labour party.

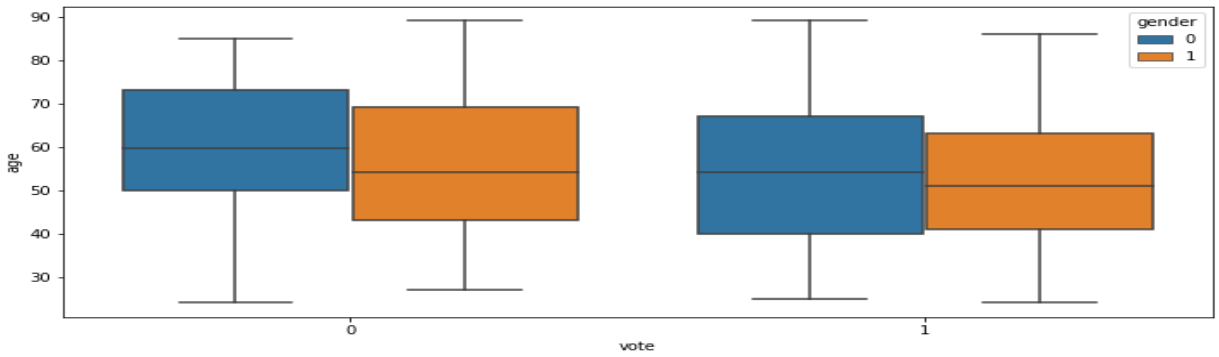


Figure-50: Plot Age vs Actual vote

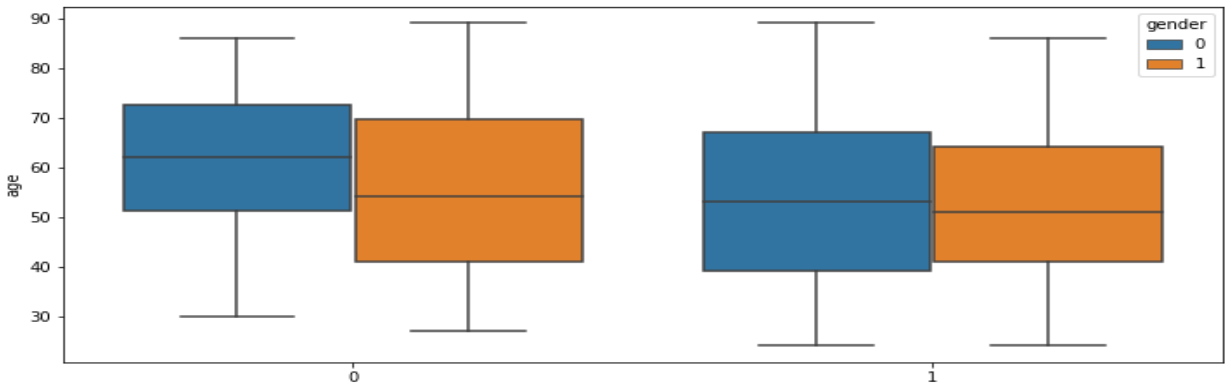


Figure-50: Plot Age vs Predicted vote

2. Problem-2: Text Analytics

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**

Code Snippet to extract the three speeches:

```
import nltk
nltk.download('inaugural')
from nltk.corpus import inaugural
inaugural.fileids()
inaugural.raw('1941-Roosevelt.txt')
inaugural.raw('1961-Kennedy.txt')
inaugural.raw('1973-Nixon.txt')
```

2.1 Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)

Solution:

Loaded the required packages and extract three speeches using the given code snippet.

Also import stopwords and punctuation (special characters) from nltk library for data-cleaning process.

→ Speech given by-President Franklin D. Roosevelt in 1941:

```
'On each national day of inauguration since 1789, the people have renewed their sen
se of dedication to the United States.\n\nIn Washington\'s day the task of the peop
le was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the p
eople was to preserve that Nation from disruption from within.\n\nIn this day the t
ask of the people is to save that Nation and its institutions from disruption from
without.\n\nTo us there has come a time, in the midst of swift happenings, to pause
for a moment and take stock -- to recall what our place in history has been, and to
rediscover what we are and what we may be. If we do not, we risk the real peril of
inaction.\n\nLives of nations are determined not by the count of years, but by the
lifetime of the human spirit. The life of a man is three-score years and ten: a lit
tle more, a little less. The life of a nation is the fullness of the measure of its
will to live.\n\nThere are men who doubt this. There are men who believe that democ
racy, as a form of Government and a frame of life, is limited or measured by a kind
of mystical and artificial fate that, for some unexplained reason, tyranny and slav
ery have become the surging wave of the future -- and that freedom is an ebbing tid
e.\n\nBut we Americans know that this is not true.\n\nEight years ago, when the lif
e of this Republic seemed frozen by a fatalistic terror, we proved that this is not
true. We were in the midst of shock -- but we acted. We acted quickly, boldly, deci
sively.\n\nThese later years have been living years -- fruitful years for the peopl
e of this democracy. For they have brought to us greater security and, I hope, a be
tter understanding that life\'s ideals are to be measured in other than material th
ings.\n\nMost vital to our present and our future is this experience of a democracy
which successfully survived crisis at home; put away many evil things; built new st
ructures on enduring lines; and, through it all, maintained the fact of its democra
cy.\n\nFor action has been taken within the three-way framework of the Constitution
of the United States. The coordinate branches of the Government continue freely to
function. The Bill of Rights remains inviolate. The freedom of elections is wholly
maintained. Prophets of the downfall of American democracy have seen their dire pre
```

Figure-1a: Sample speech by Franklin D. Roosevelt

- The total number of characters, words and sentences for the above mentioned speech given by president Franklin D. Roosevelt in 1941 are as below:

	1941-Roosevelt.txt
Character	7571
Words	1536
Sentences	68

→ **Speech given by-President John F. Kennedy in 1961:**

'Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well as a beginning -- signifying renewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our forebears l prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for which our forebears fought are still at issue around the globe -- the belief that the rights of man come not from the generosity of the state, but from the hand of God.\n\nWe dare not forget today that we are the heirs of that first revolution. Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans -- born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage -- and unwilling to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are committed today at home and around the world.\n\nLet every nation know, whether it wishes us well or ill, that we shall pay any price, bear any burden, meet any hardship, support any friend, oppose any foe, in order to assure the survival and the success of liberty.\n\nThis much we pledge -- and more.\n\nTo those old allies whose cultural and spiritual origins we share, we pledge the loyalty of faithful friends. United, there is little we cannot do in a host of cooperative ventures. Divided, there is little we can do -- for we dare not meet a powerful challenge at odds and split asunder.\n\nTo those new States whom we welcome to the ranks of the free, we pledge our word that one form of colonial control shall not have passed away merely to be replaced by a far more iron tyranny. We shall not always expect to find them supporting our view. But we shall always hope to find them strongly supporting their own freedom -- and to remember that, in the past, thos

Figure-1b: Sample speech by John F. Kennedy

- The total number of characters, words and sentences for the above mentioned speech given by president John F. Kennedy in 1961are as below:

	1961-Kennedy.txt
Character	7618
Words	1546
Sentences	52

→ Speech given by-President Richard Nixon in 1973:

'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and good country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilities greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis past year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendships, and by our missions to Peking and to Moscow, we were able to establish the base for a new and more durable pattern of relationships among the nations of the world. Because of America\'s bold initiatives, 1972 will be long remembered as the year of the greatest progress since the end of World War II toward a lasting peace in the world.\n\nThe peace we seek in the world is not the flimsy peace which is merely an interlude between wars, but a peace which can endure for generations to come.\n\nIt is important that we understand both the necessity and the limitations of America\'s role in maintaining that peace.\n\nUnless we in America work to preserve the peace, there will be no peace.\n\nUnless we in America work to preserve freedom, there will be no freedom.\n\nBut let us clearly understand the new nature of America\'s role, as a result of the new policies we have adopted over these past four years.\n\nWe shall respect our treaty commitments.\n\nWe shall support vigorously the principle that no country has the right to impose its will or rule on another by force.\n\nWe shall continue, in this era of negotiation, to work for the limitation of nuclear arms, and to reduce the danger of confrontation between the great

Figure-1c: Sample speech by Richard Nixon

- The total number of characters, words and sentences for the above mentioned speech given by president Richard Nixon in 1973 are as below:

	1973-Nixon.txt
Character	9991
Words	2028
Sentences	69

2.2 Find Remove all the stopwords from the three speeches.

Solution:

Data cleaning process done on all the three speeches by converting all the characters to lower case and then remove stopwords and special characters /punctuation's and assign it to new variable in the form lists.

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

Solution:

→ Most Frequent words in 1941-Roosevelt Speech:

	Count
nation	12
know	10
spirit	9
democracy	9
life	9

→ **Most Frequent words in 1961-Kennedy Speech:**

	Count
let	16
us	12
sides	8
world	8
new	7

→ **Most Frequent words in 1973-Nixon Speech:**

	Count
us	26
let	22
america	21
peace	19
world	18

2.4 Plot the word cloud of each of the three speeches. (After removing the stopwords).

Solution:

→ **Word Cloud for 1941-Roosevelt Speech:**

→ **Word Cloud for 1961-Kennedy Speech:**

→ **Word Cloud for 1973-Nixon Speech:**