PGPDSBA Online FEB A 2021

Pavan Kumar R Naik

PGP-DSBA Online

Feb A 2021

30/08/2021

Table of Contents

List of Figures

List of Tables

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.

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:

Q1.1. Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.

Solution:

Sample of Dataset:

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

Table 1.1. Dataset Sample (Election Data)

Summary of Dataset:

	Unnamed: 0	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000
mean	763.000000	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295
std	440.373894	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315
min	1.000000	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	382.000000	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000
50%	763.000000	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000
75%	1144.000000	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000
max	1525.000000	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000

Table 1.2. Dataset Summary (Election Data)

Remove the ‘Unnamed’ column from the dataset.

Type of Variables:

```
<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
```

Table 1.3. Type of Variables (Election Data)

Check for zero value in any numerical observation:

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

Table 1.4. Zero value in numerical observation (Election Data)

Check for null values in the dataset:

```
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
```

Table 1.5. Null value in dataset (Election Data)

Unique count of values in object variable:

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


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

Table 1.6. Count of values in object variable (Election Data)

Inference:

- Dataset has 10 columns out of which first column (Unnamed:0) is of no use for analysis and been removed
- Dataset has 1525 observations
- There are no missing values and also, we don't have unique identifier, hence there is no necessity in checking the duplicate values
- Age column is the only continuous variable with range between 24 and 93, avg. of 54.18
- 2 variables: vote and gender are nominal variables
    - Majority of electors choose to vote Labour party compared to Conservative party,1063:462
    - Female voters are more compared to male voters,812:713
- 6 variables of integer data type have ordinal data type
    - 'economic.cond.national': rating on scale 1 to 5
    - 'economic.cond.household': rating on scale 1 to 5
    - 'Blar': rating on scale 1 to 5
    - 'Hague': rating on scale 1 to 5
    - 'Europe': A 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
    - 'political.knowledge': Knowledge of parties' positions on European integration, 0 to 3.
- Age variable has some amount of skewness which is 0.15 lies between 0 to 0.5 , hence the data is fairly symmetrical

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

Solution:

Univariate Analysis:

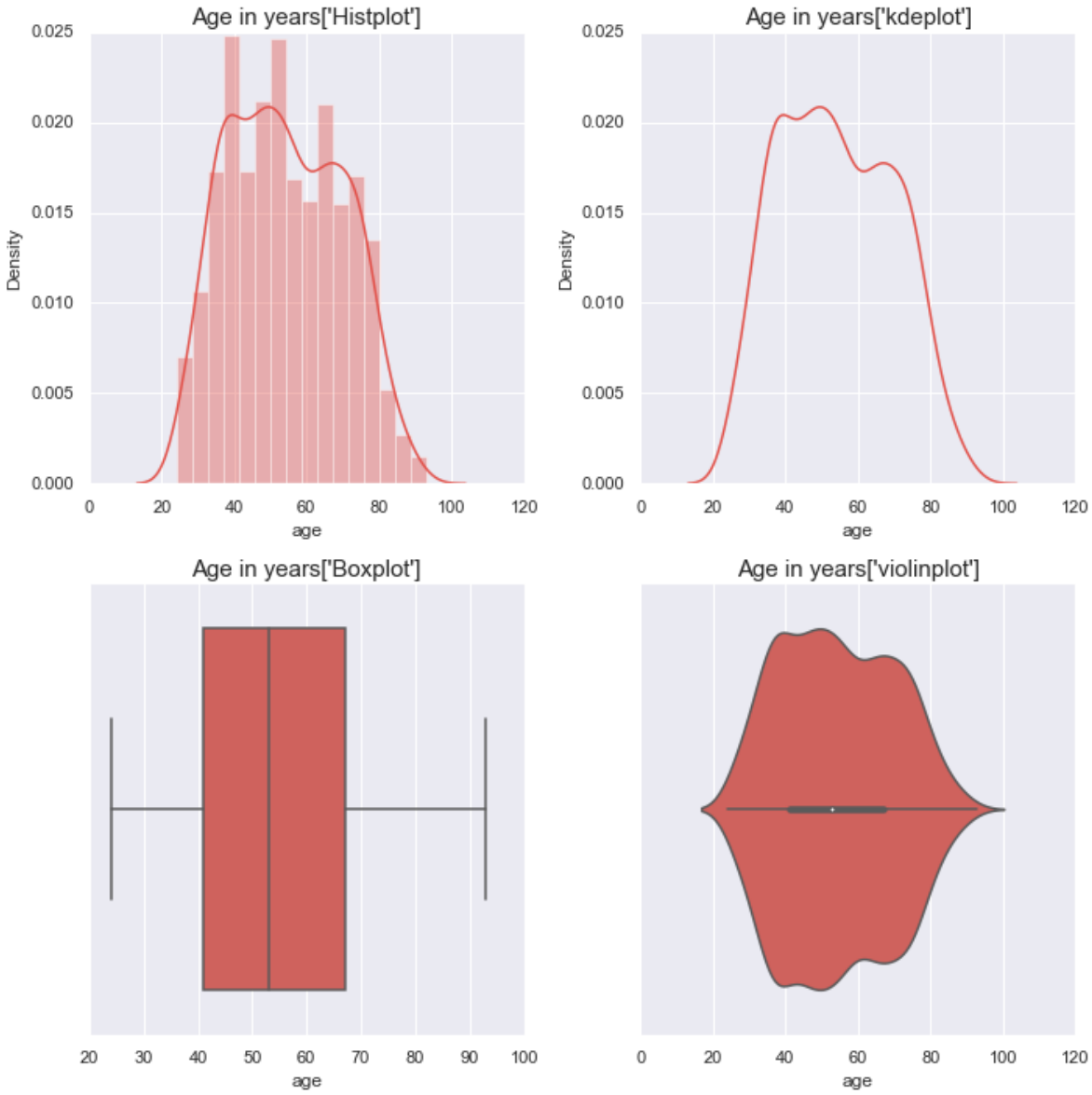


Fig 1.1. Univariate analysis on continuous variable (Age)

- Age variable do not contain any outliers
- Age variable has some amount of skewness 0.1446, data is fairly symmetrical

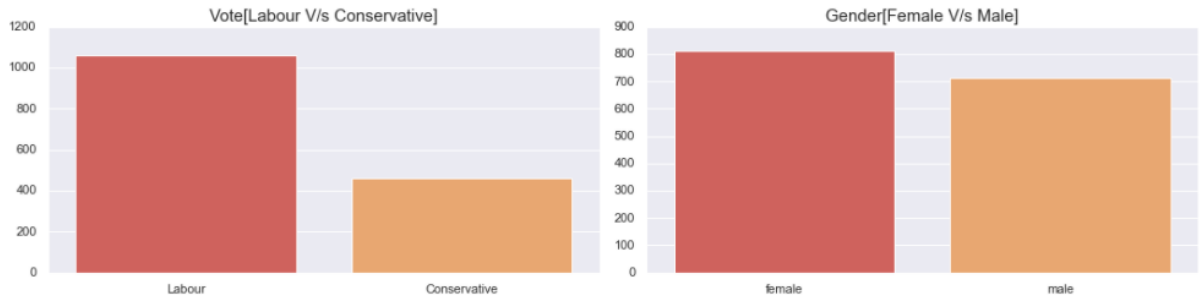


Fig 1.2. Univariate analysis on nominal variable

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

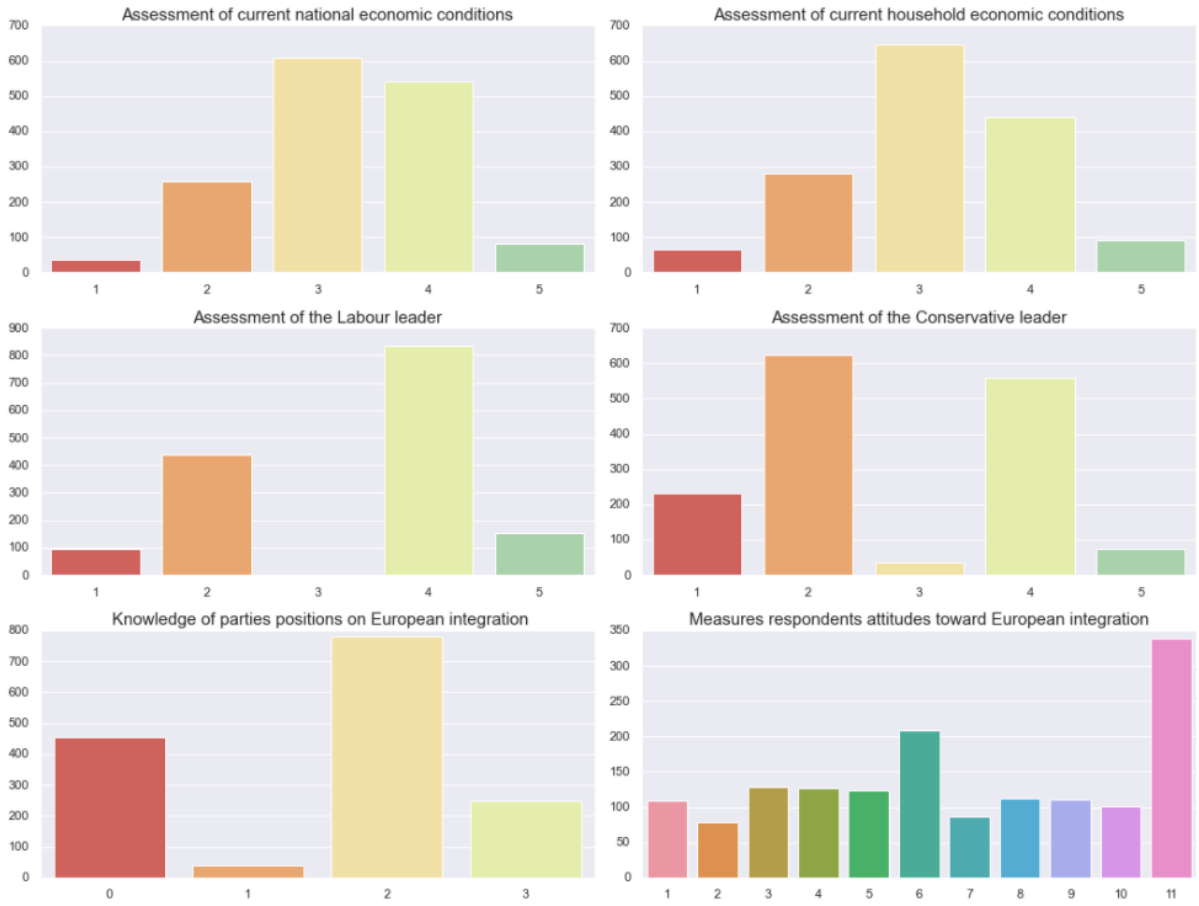


Fig 1.3. Univariate analysis on ordinal variable

```
VOTE :  2                    GENDER :  2
0     462                    0    812
1    1063                    1    713
Name: vote, dtype: int64     Name: gender, dtype: int64
```

Table 1.7. Unique count of nominal 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 Haque, most of the voters chose to rate 2 and 4
- Most of the voters represents Eurosceptic sentiment

```
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
```

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

Table 1.8. Unique count of ordinal variables

Bivariate and Multivariate analysis:

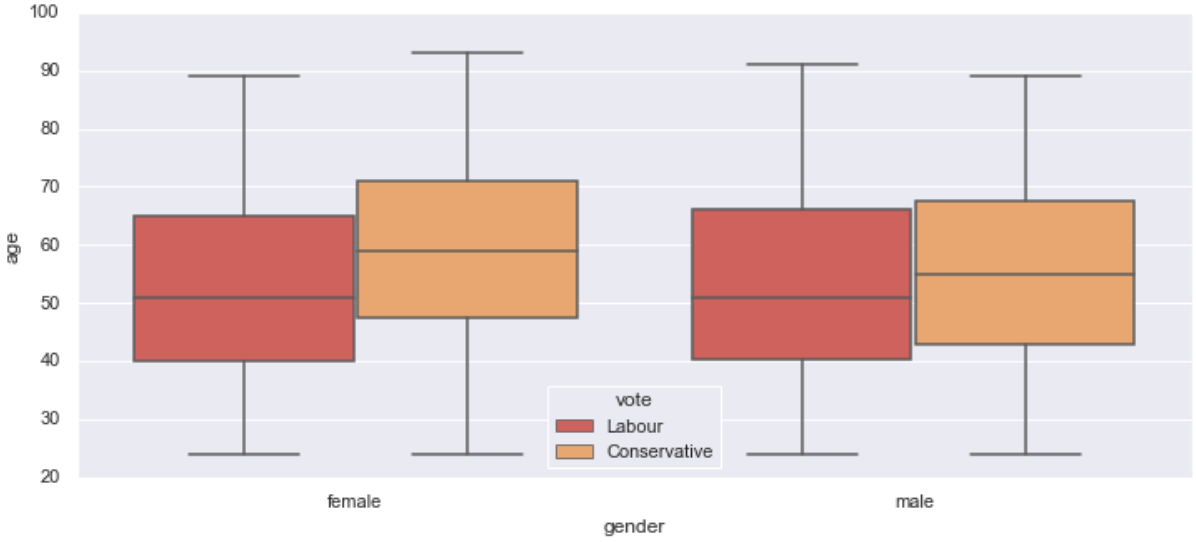


Fig 1.4. One Numeric vs Nominal categorical variables

- Avg. age near to 60 to elect conservative party whereas avg. age nears to 50 to elect labour party
- Middle 50% of the male and female people who fall under the age group within the age 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 age range of 50 to 70 choose to elect conservative party
- We can notice in the above box plot that as the age increases people choose to elect conservative party

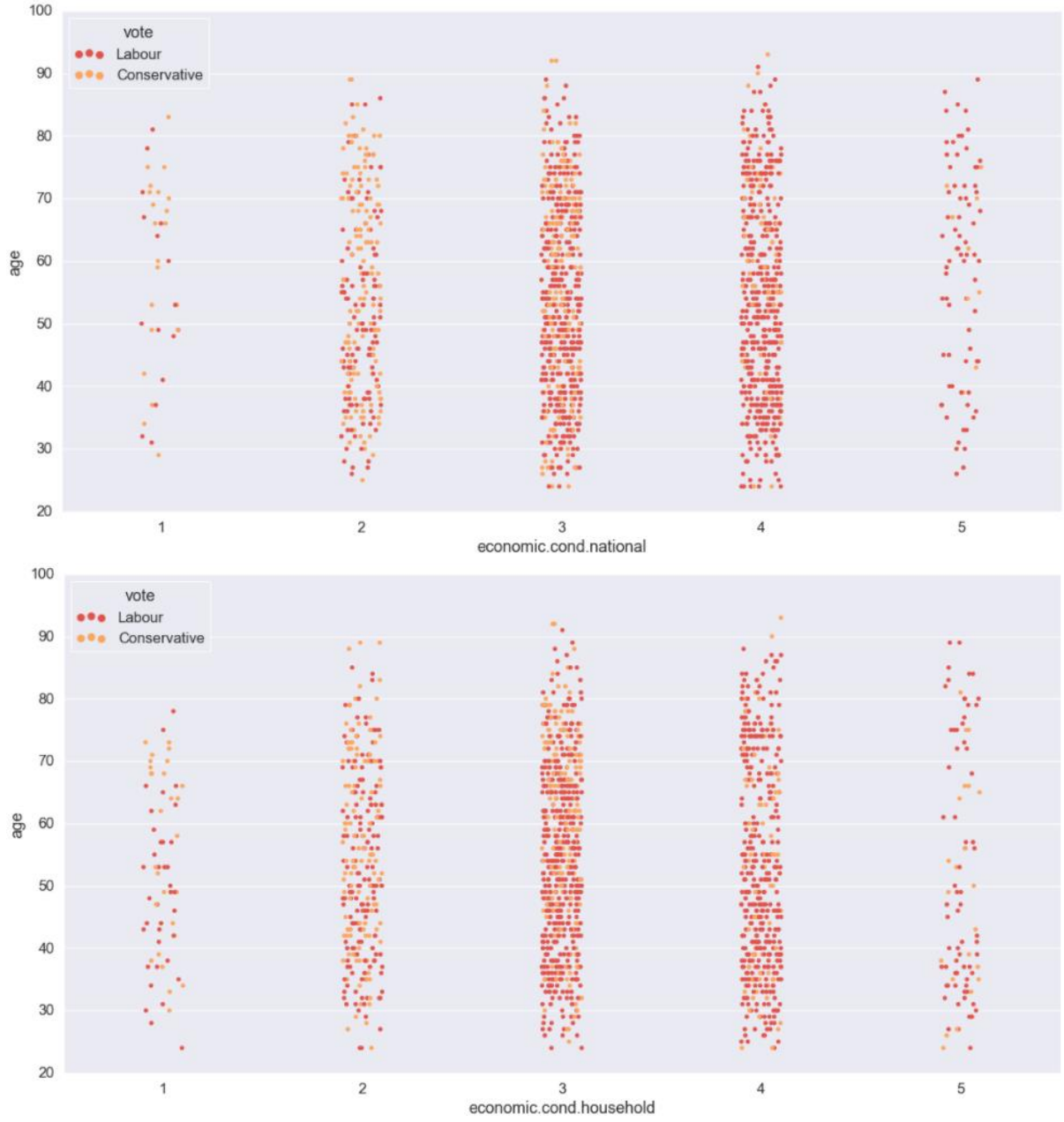


Fig 1.5. One Numeric(age) vs Ordinal categorical variables (ECN and ECH)

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

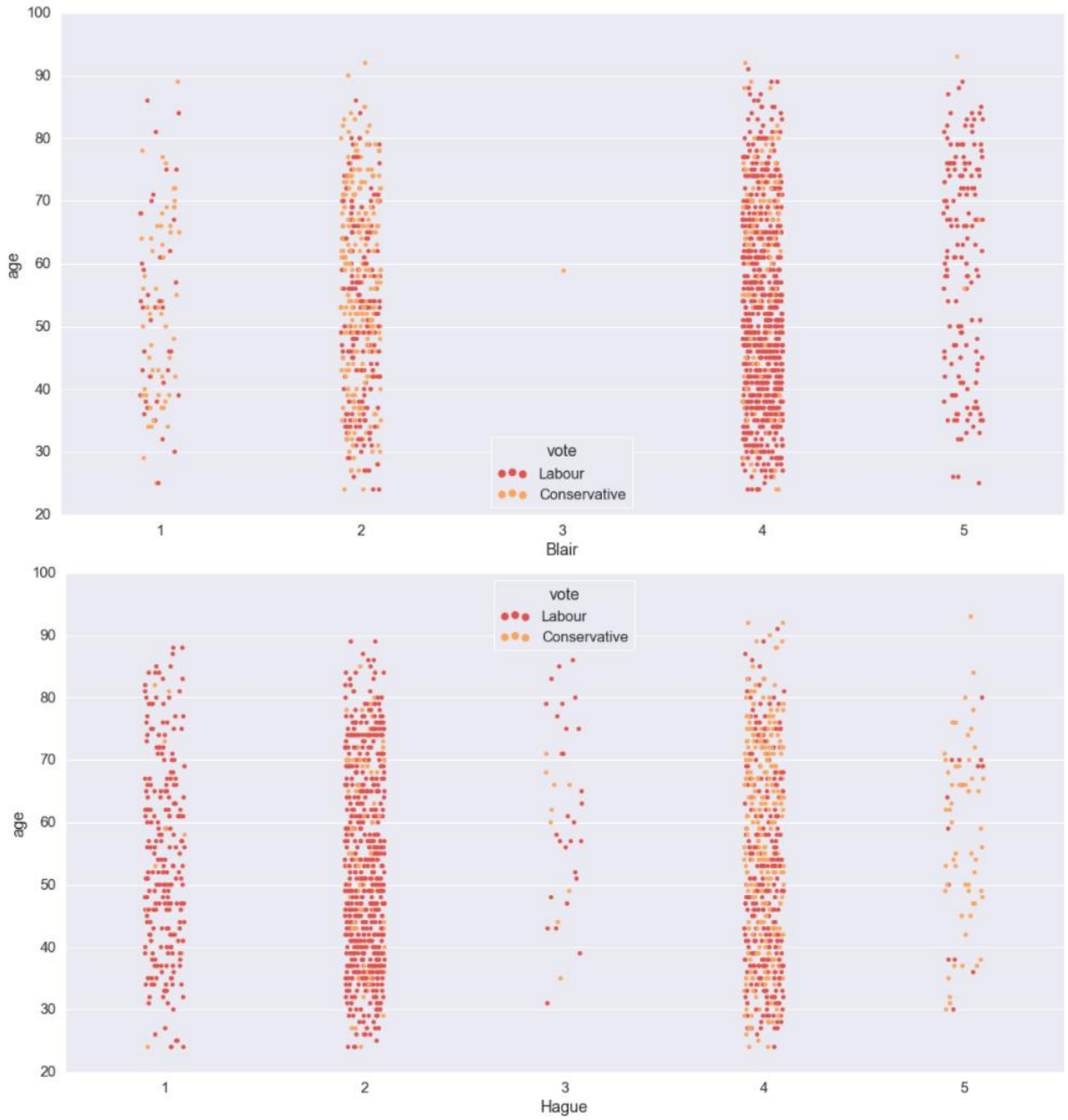


Fig 1.6. One Numeric(age) vs Ordinal categorical variables (Blair and Hague)

- Clear from the above plot that voters who gave rating 4 and 5 for the assessment of laboir party leader Blair, chose to elect labour party
- Voters who gave rating 3 and above for the assessment of conservative party leader Hague, chose to elect the conservative party

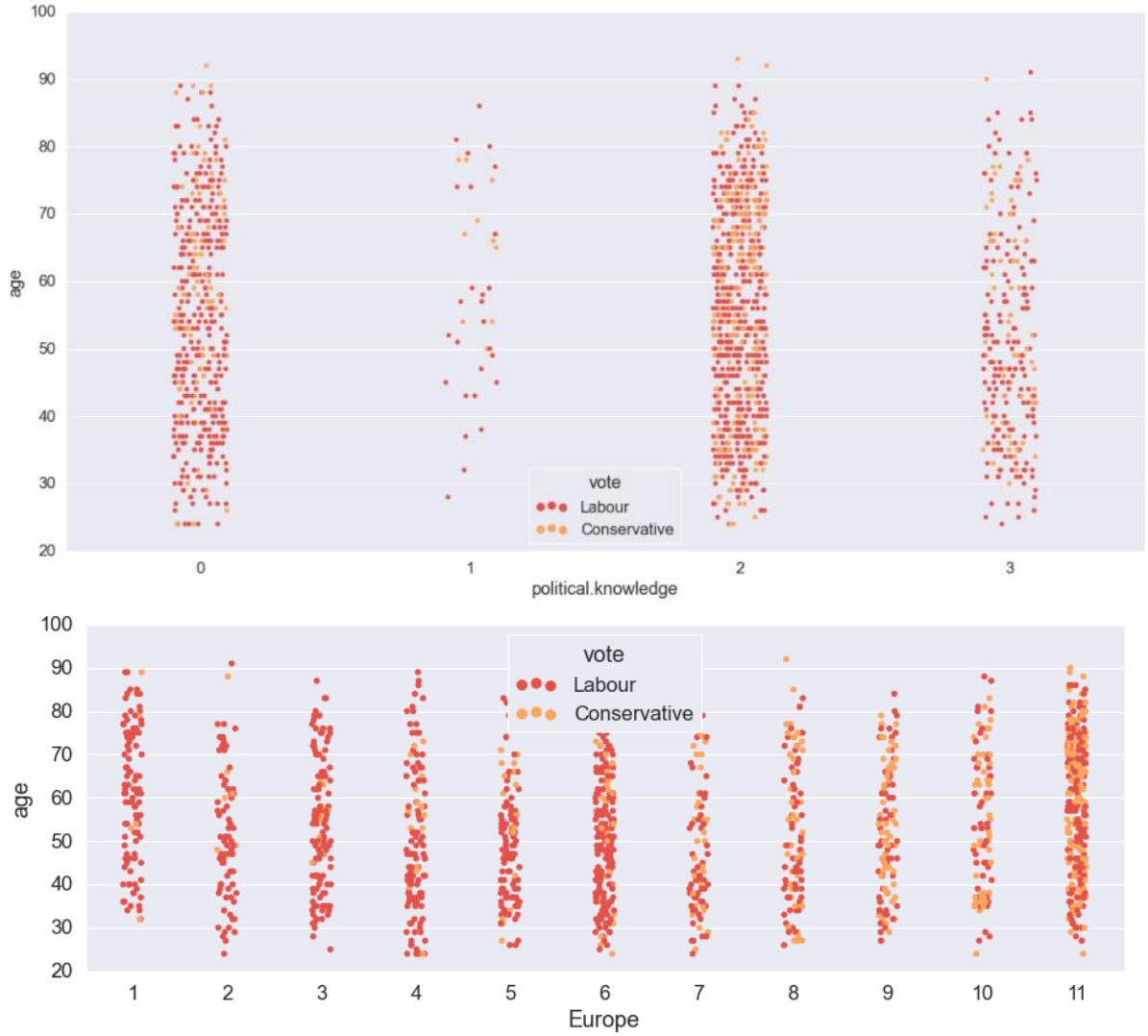


Fig 1.7. One Numeric(age) vs Ordinal categorical variables (Political Knowledge and Europe)

- Voters who choose to vote for the labour party do not posess a very high Eurosceptic sentiment, whereas voters who represent high Eurosceptic sentiment vote for conservative party
- Data has no outliers “the number of outliers is 0”

Data Visualization: EDA using sweet viz to visualize the summary for each variable as well to underrated data –

- Univariate

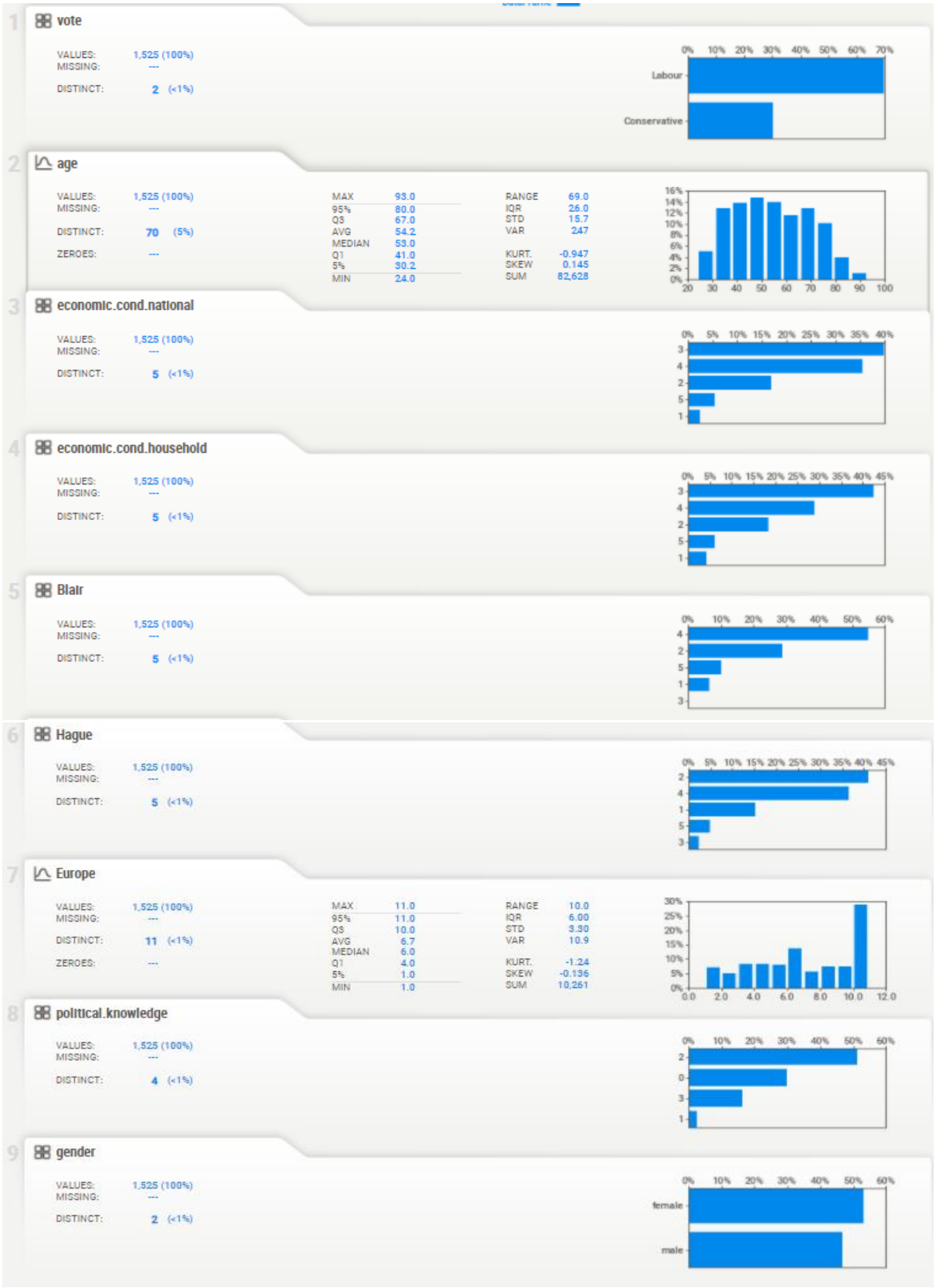


Fig 1.8. Sweet viz Univariate analysis

- Bivariate Analysis: Between target variable and others

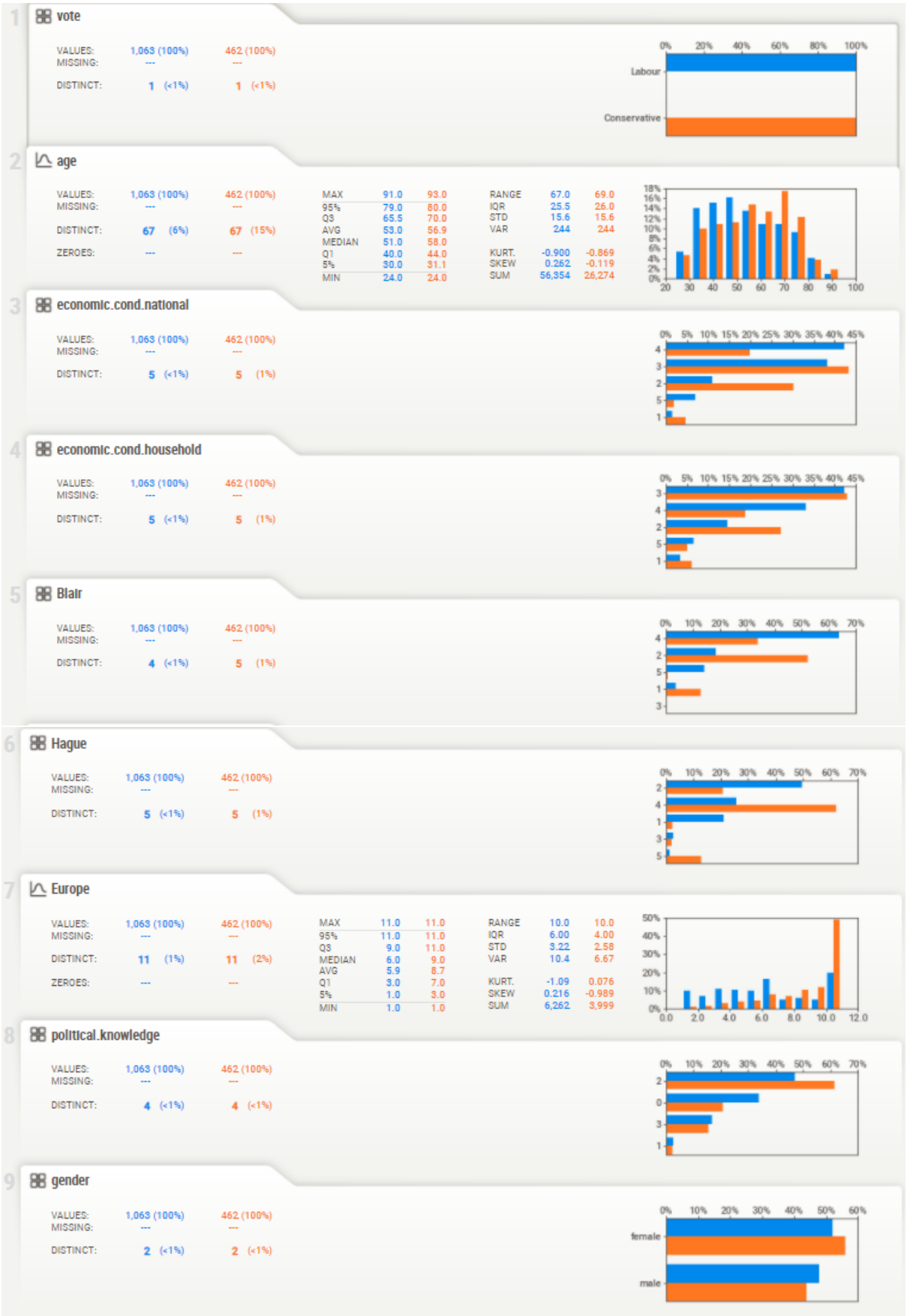


Fig 1.9 Sweet viz Bivariate analysis

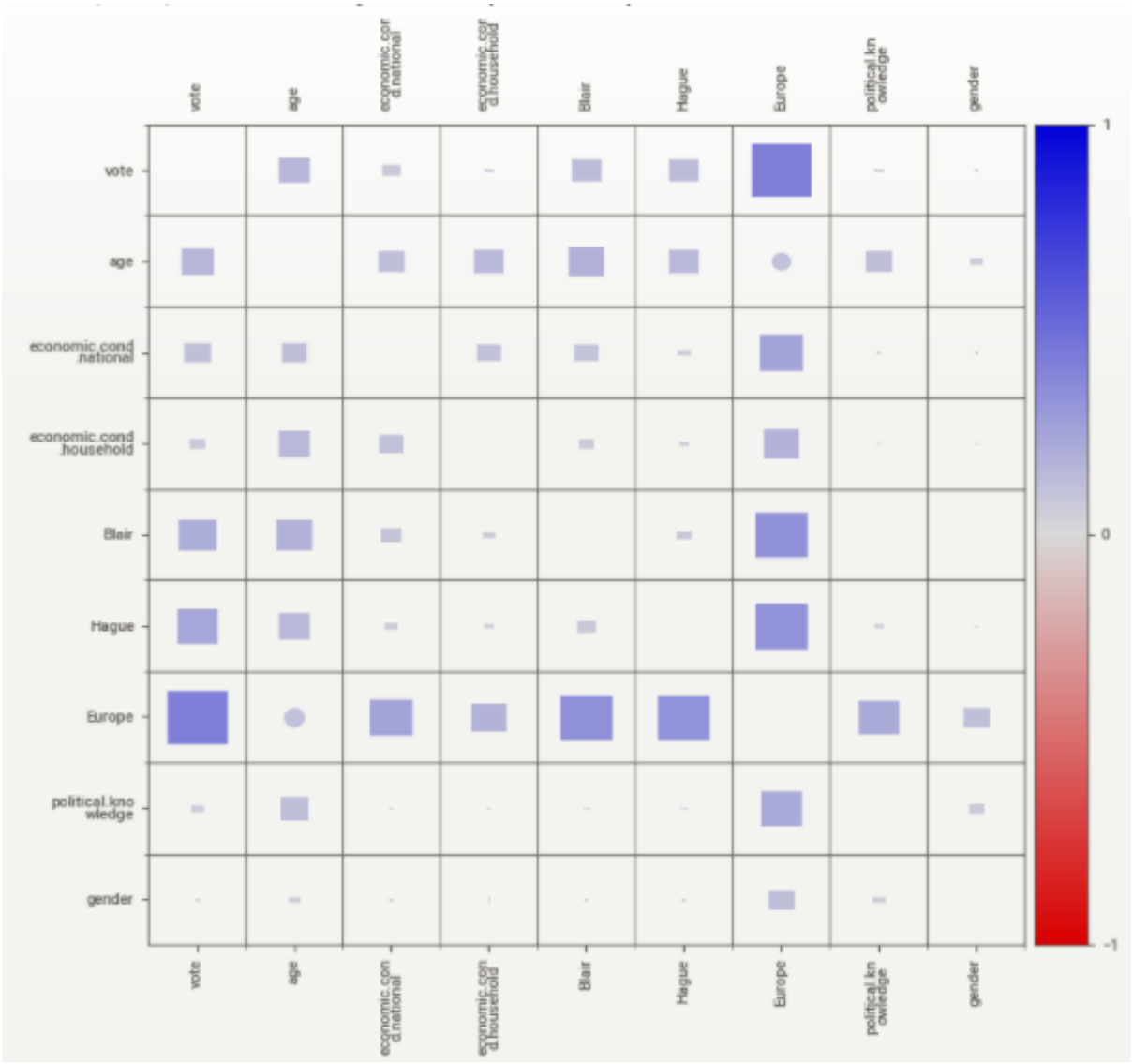


Fig 1.10. Sweet viz Multivariate analysis

Inference:

- Class proportion of target variable “vote” is more than 10 percent, so balanced for modelling
- Two third of the voters are from economic condition of 3 and 4
- 50% of the voters have assessed 4 and above for Blair,50% of the voters have assessed 2 and below for Hague
- 50% have political knowledge of below 2
- Majority of the voters belongs to economic conditions of 4 and 5 prefer Labour party and belonging to 2 and below prefer Conservative party
- Majority of female voters prefer Conservative party and male prefer Labour party
- Most number of voters between age 20 – 50 they prefer Labour, majority of voters above age of 50 prefer Conservative

Data Preparation:

Q1.3 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:

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

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

Table 1.9. Data balance for target variable

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

Data Encoding:

- For a given data set we have 8 categorical variables, 6 are ordinal which are of integer data type and 2 are nominal variable Gender and the target variable Vote as object which needs to be converted to 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]
```

Table 1.10. Data encoding

- Post encoding the data, the target variable vote was captured into a separate vector for training and testing dataset
- Data was split into 70:30 train and test ratio

```
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,)
```

Table 1.11. Data split 70:30

Scaling:

- We have feature age in years with different unit weight and the remaining variables are of ratings ranging from 1 to 5; 1 to 11 and 0 to 3. Hence scaling is required for certain models to get the accurate results
- Scaling will be done only on the training dataset

Data Modelling:

Q1.4 Apply Logistic Regression and LDA (linear discriminant analysis).

Solution:

Logistic Regression:

- Feature scaling is doesn't require for Logistic Regression, hence not scaled
- Build a Logistic Regression using a grid search cross validation to get the best parameter/estimators for a given dataset

```
{'max_iter': 5000, 'penalty': 'l1', 'solver': 'liblinear', 'tol': 0.01}

LogisticRegression(max_iter=5000, n_jobs=4, penalty='l1', solver='liblinear',
                   tol=0.01, verbose=True)
```

Table 1.12. Best Grid for LR

- Accuracy of Train set: 82.6%
- Accuracy of Test set: 85.8%

Linear Discriminant Analysis:

- Feature scaling is doesn't require for LDA, hence not scaled
- Model is built without specific parameter settings: LinearDiscriminantAnalysis()
- Accuracy of Train set: 82.6%
- Accuracy of Test set: 84.5%

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

Q1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results

Solution:

Gaussian Naïve Bayes:

- 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 feature scaling in this algorithm may not have much effect
- Build a model: GaussianNB()
- Accuracy of Train set: 82.2%
- Accuracy of Test set: 84.7

K-nearest neighbour [KNN-model]:

- To have good KNN model, data requires pre-processing to make all independent variables similarly scaled and centered. 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

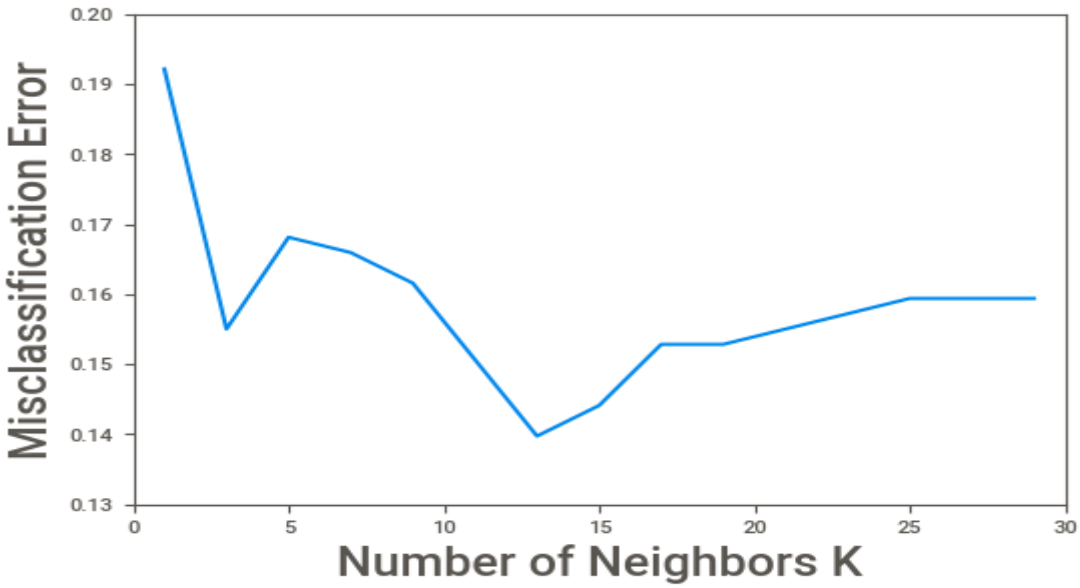


Fig 1.11. Misclassification error KNN

- Optimum number of neighbour or k-value is found to be 13
- Build a KNeighborsClassifier using the value of n_neighbors=13 and metric= 'Euclidean'
- Accuracy of Train set: 84.2%
- Accuracy of Test set: 86.0%

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

Q1.6 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

```
{'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=31)
```

Table 1.13. Best Grid for RF

- Accuracy of Train set: 82.0%
- Accuracy of Test set: 82.7%

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

```
BaggingClassifier(base_estimator=RandomForestClassifier(max_depth=7,
                                                        max_features=4,
                                                        min_samples_leaf=40,
                                                        min_samples_split=100,
                                                        n_estimators=501,
                                                        random_state=31),
                  n_estimators=100, random_state=31)
```

Table 1.14. Best Grid for Bagging

- Accuracy of Train set: 81.6%
- Accuracy of Test set: 82.7%

Comparing accuracy of Random Forest and Bagging models, Test data accuracy of Random Forest and Bagging are equal however the Train accuracy is better in 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
- Accuracy of Train set: 88.8%
- Accuracy of Test set: 83.8%

Apply Ada Boosting:

- AdaBoost was built with n_estimators=100
- Accuracy of Train set: 84.4%
- Accuracy of Test set: 83.6%

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

Q1.7 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:

Model Evaluation-Logistic Regression:

- Accuracy on training set is 82.7% and on testing set is 85.8%
- Classification report

Classification report for Logistic Regression model on Training set is

	precision	recall	f1-score	support
0	0.74	0.66	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

Table 1.15. Classification report (LR – Train)

Classification report for Logistic Regression model on Testing set is

	precision	recall	f1-score	support
0	0.81	0.69	0.75	139
1	0.87	0.93	0.90	319
accuracy			0.86	458
macro avg	0.84	0.81	0.82	458
weighted avg	0.86	0.86	0.85	458

Table 1.16. Classification report (LR – Test)

- Confusion matrix

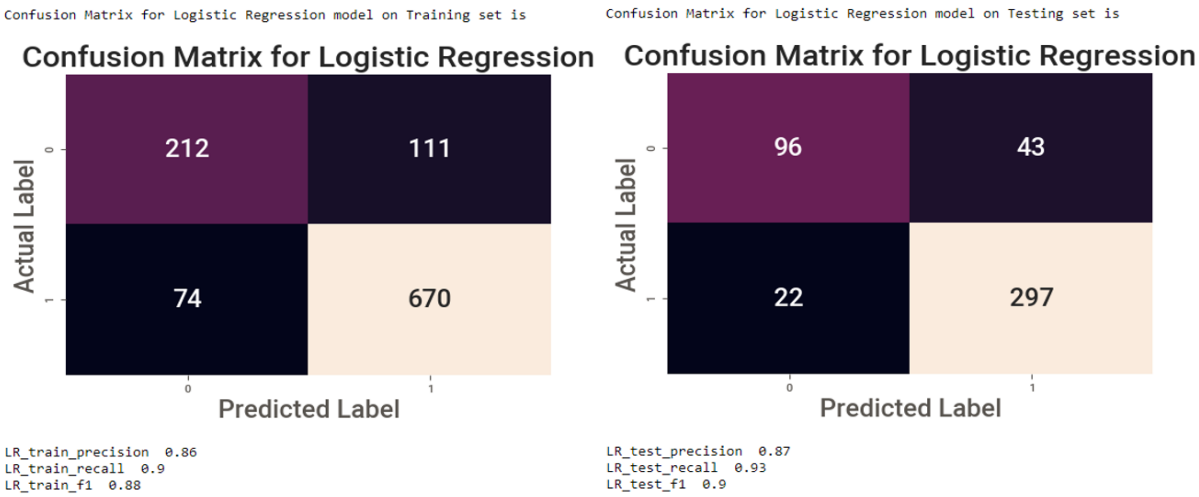


Fig 1.12. Confusion matrix LR

- AUC and ROC: AUC Train:87.6% and AUC Test:91.6%

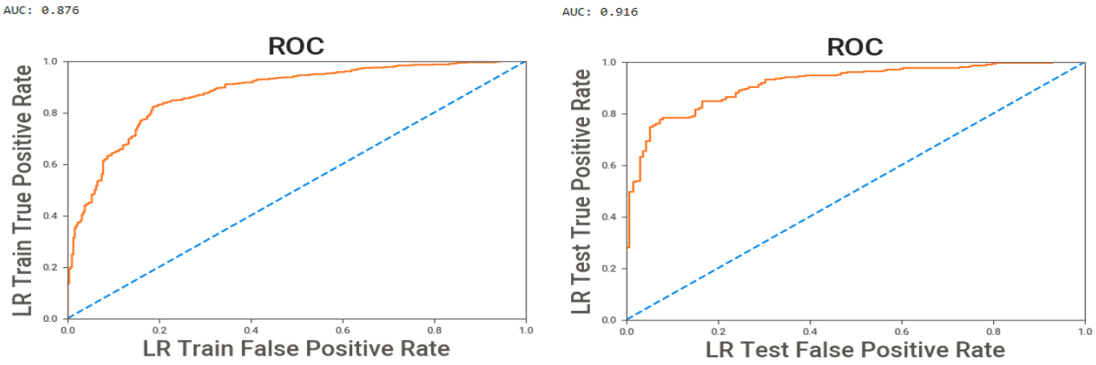


Fig 1.13. AUC and ROC – LR

Model Evaluation-Linear Discriminant Analysis:

- Accuracy on training set is 82.6% and on testing 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

Table 1.17. Classification report (LDA – Train)

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

Table 1.18. Classification report (LDA – Test)

- Confusion matrix

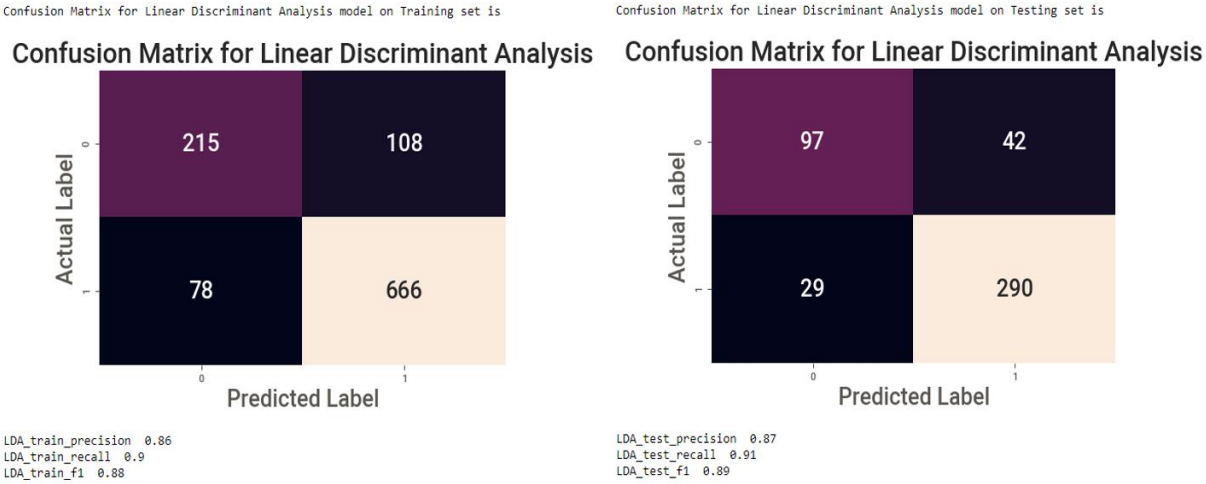


Fig 1.14. Confusion matrix LDA

- AUC and ROC: AUC Train:87.6% and AUC Test:91.5%

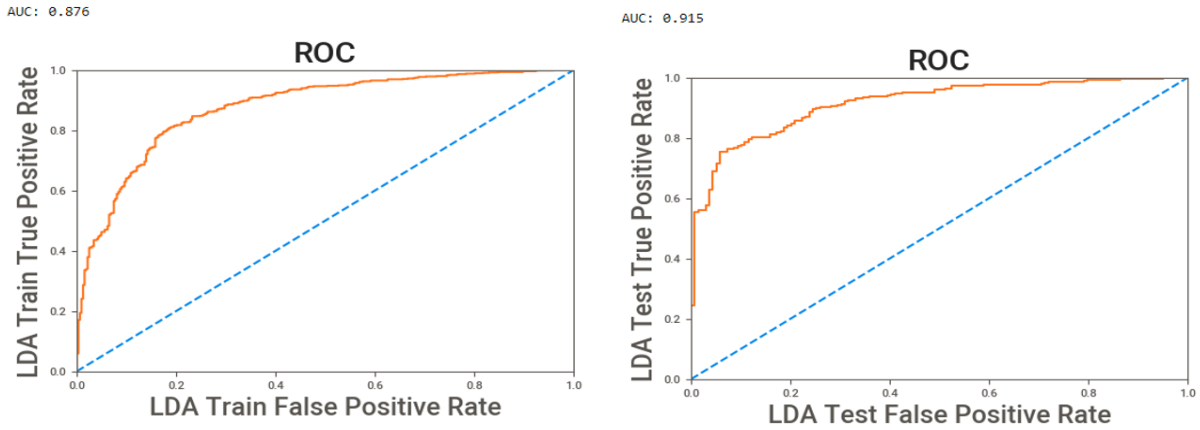
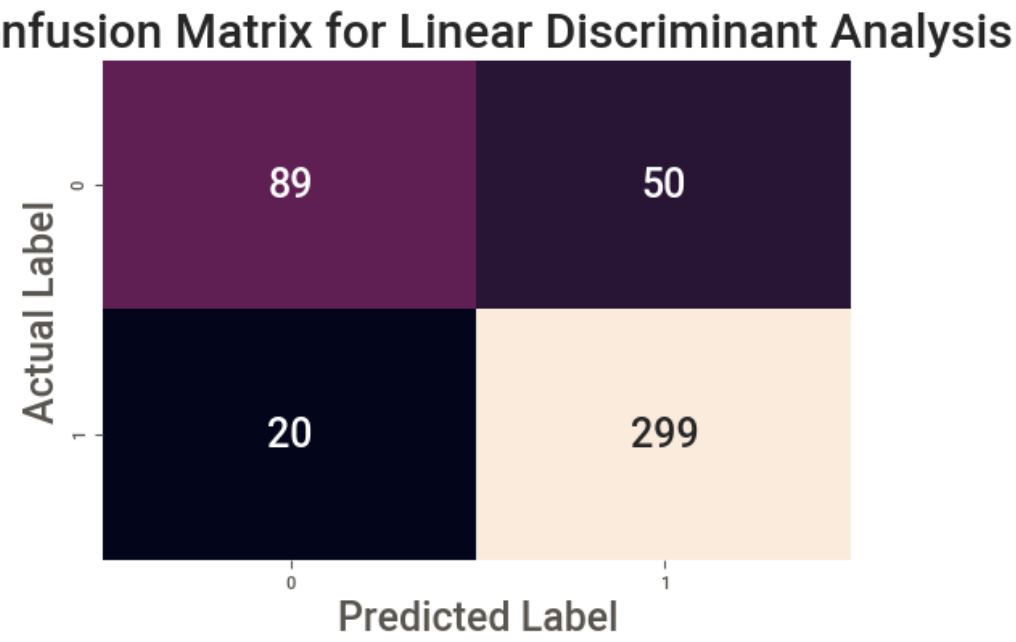


Fig 1.15. AUC and ROC – LDA

- 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:

```
Confusion Matrix for Linear Discriminant Analysis model on Testing set with cut-off value 0.4 is
```




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

Fig 1.16. Confusion matrix LDA – Test with cut off value 0.4

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%

Model Evaluation- Gaussian Naïve Bayes:

- Accuracy on training set is 82.2% and on testing 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
```

Table 1.19. Classification report (NB – Train)

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

Table 1.20. Classification report (NB – Test)

- Confusion matrix

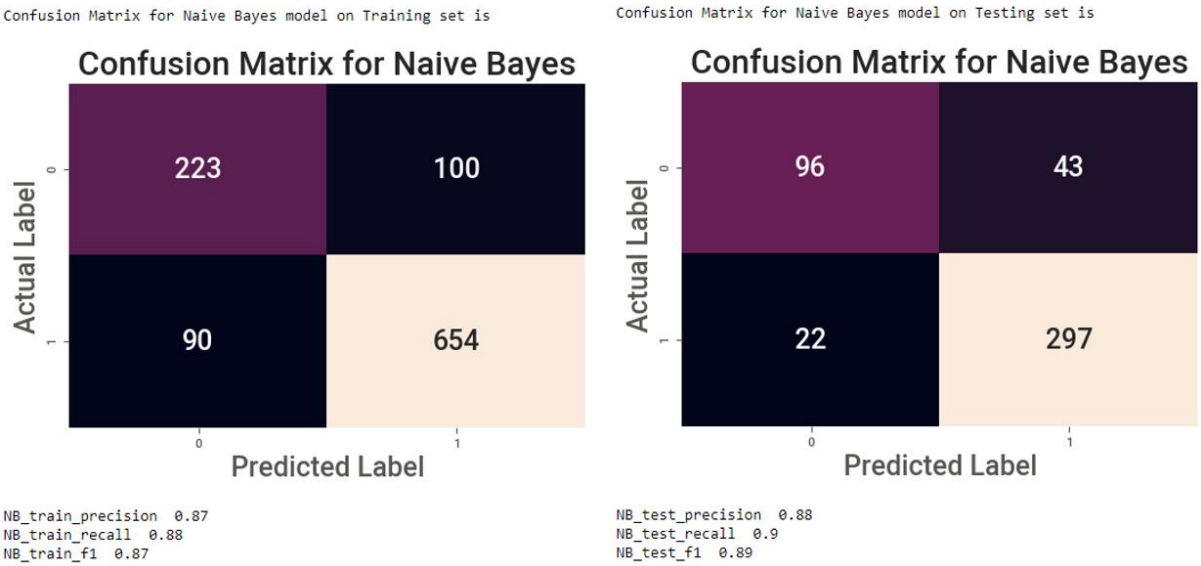


Fig 1.17. Confusion matrix NB

- AUC and ROC: AUC Train:87.4% and AUC Test:91.0%

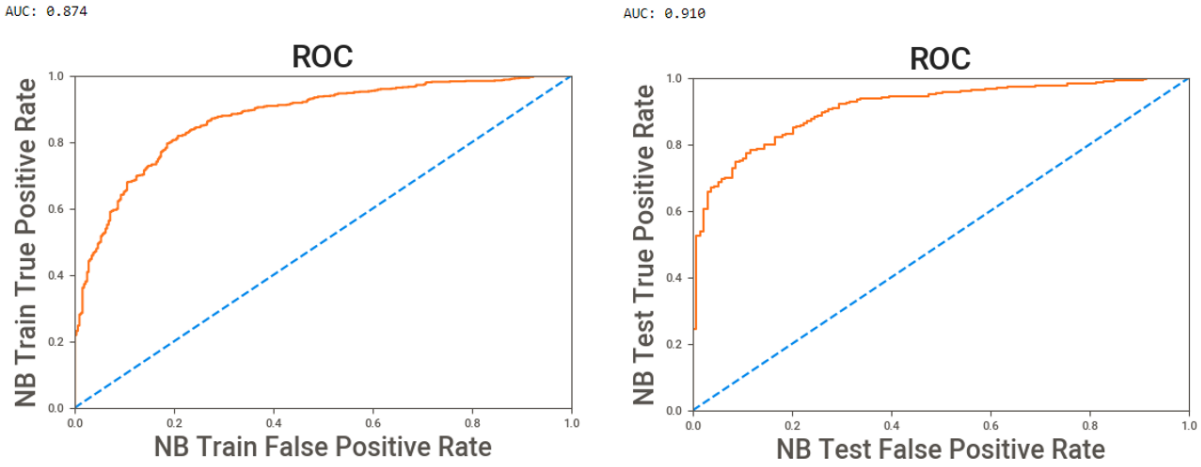
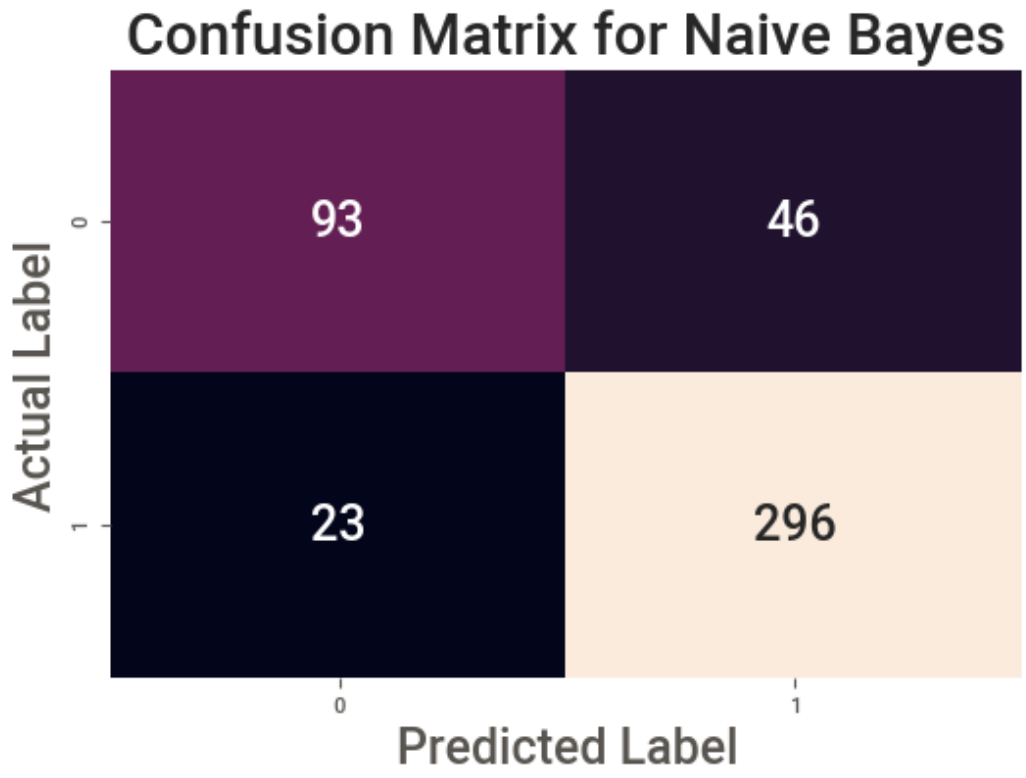


Fig 1.18. AUC and ROC – NB

- 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
```

Fig 1.19. Confusion matrix NB testing set with cut off 0.4

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%

Model Evaluation- KNN:

- Accuracy on training set is 84.3% and on testing set is 86.0%
- 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
```

Table 1.21. Classification report (KNN – Train)

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

Table 1.22. Classification report (KNN – Test)

- Confusion matrix

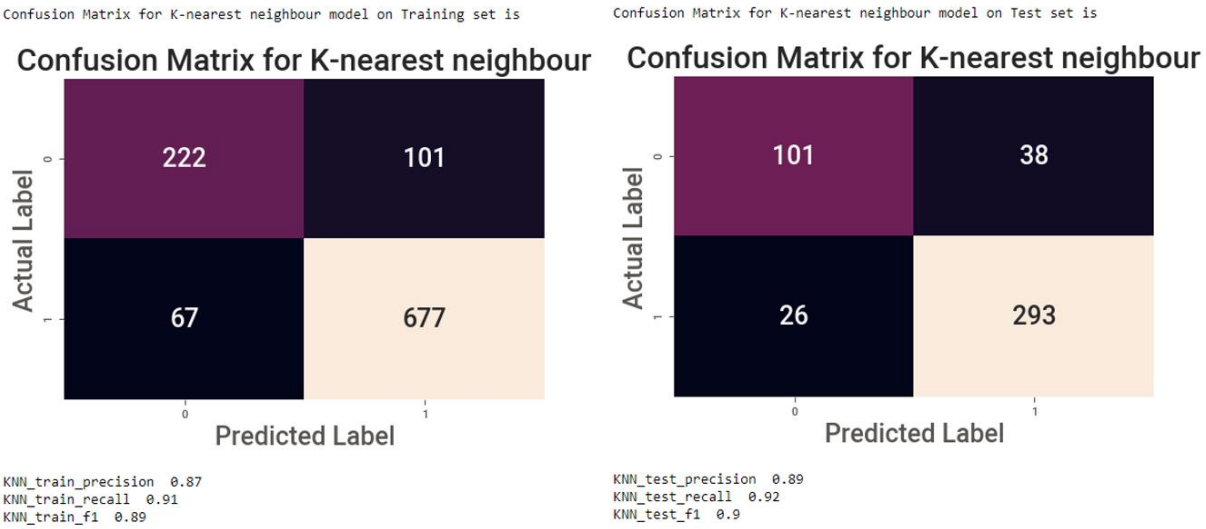


Fig 1.20. Confusion matrix KNN

- AUC and ROC: AUC Train:91.1% and AUC Test:89.4%

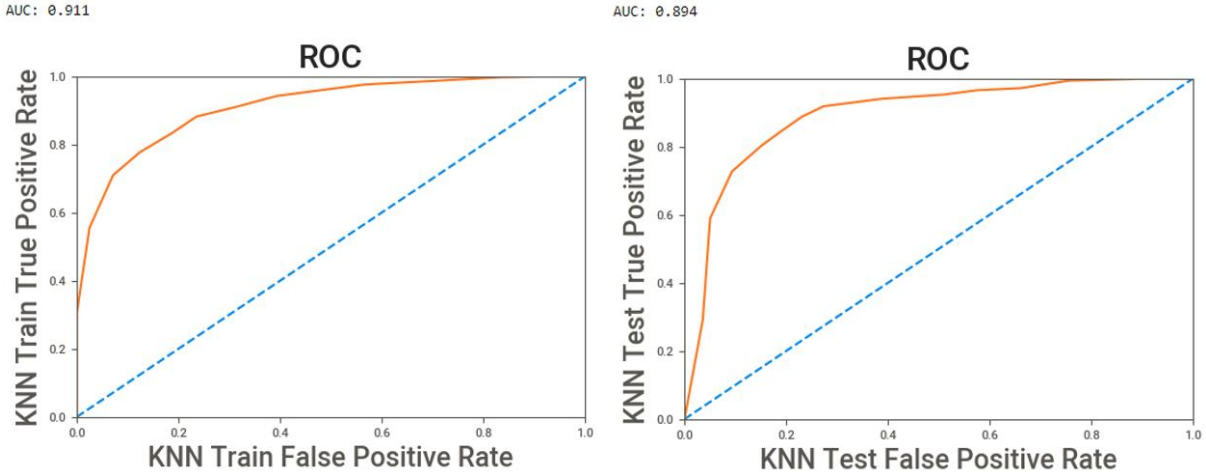


Fig 1.21. AUC and ROC – KNN

Model Evaluation- Random Forest:

- Accuracy on training set is 82.0% and on testing set is 82.8%
- Classification report

Classification report for RandomForestClassifier model on Training set is

	precision	recall	f1-score	support
0	0.78	0.57	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

Table 1.23. Classification report (RF – Train)

Classification report for RandomForestClassifier model on Test set is

	precision	recall	f1-score	support
0	0.82	0.55	0.66	139
1	0.83	0.95	0.88	319
accuracy			0.83	458
macro avg	0.82	0.75	0.77	458
weighted avg	0.83	0.83	0.82	458

Table 1.24. Classification report (RF – Test)

- Confusion matrix

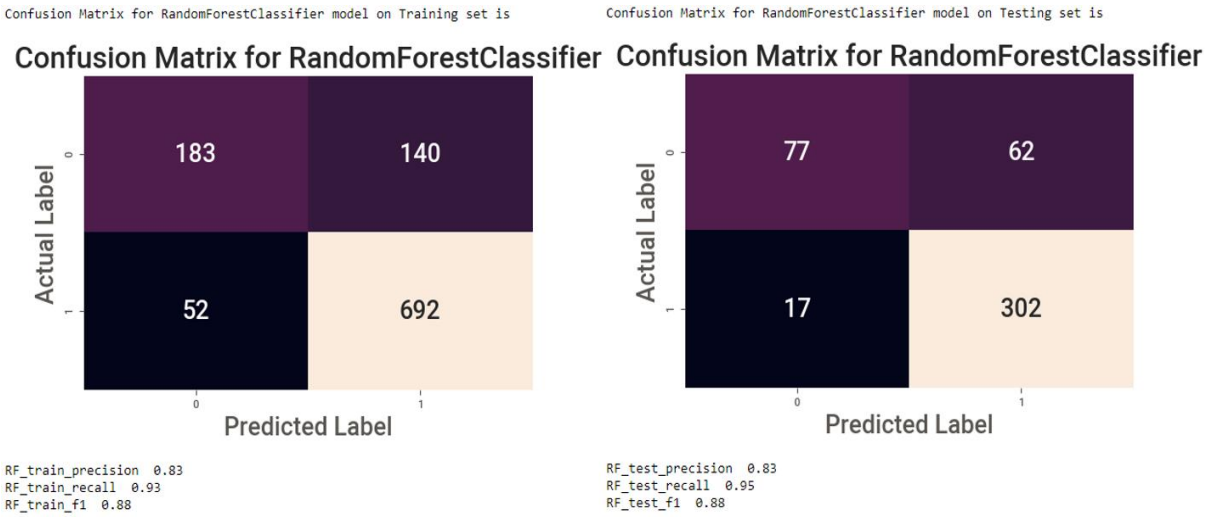


Fig 1.22. Confusion matrix RF

- AUC and ROC: AUC Train:88.8% and AUC Test:90.2%

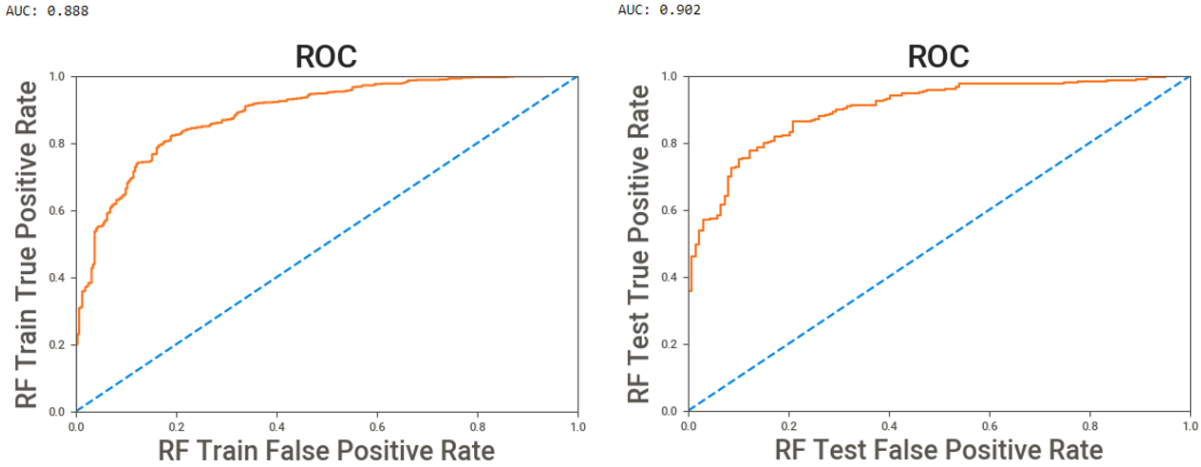


Fig 1.23. AUC and ROC – RF

Model Evaluation- Bagging:

- Accuracy on training set is 81.6% and on testing set is 82.8%
- Classification report

Classification report for BaggingClassifier model on Training set is

	precision	recall	f1-score	support
0	0.78	0.54	0.64	323
1	0.82	0.94	0.88	744
accuracy			0.82	1067
macro avg	0.80	0.74	0.76	1067
weighted avg	0.81	0.82	0.81	1067

Table 1.25. Classification report (Bag – Train)

Classification report for BaggingClassifier model on Test set is

	precision	recall	f1-score	support
0	0.83	0.54	0.66	139
1	0.83	0.95	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

Table 1.26. Classification report (Bag – Test)

- Confusion matrix

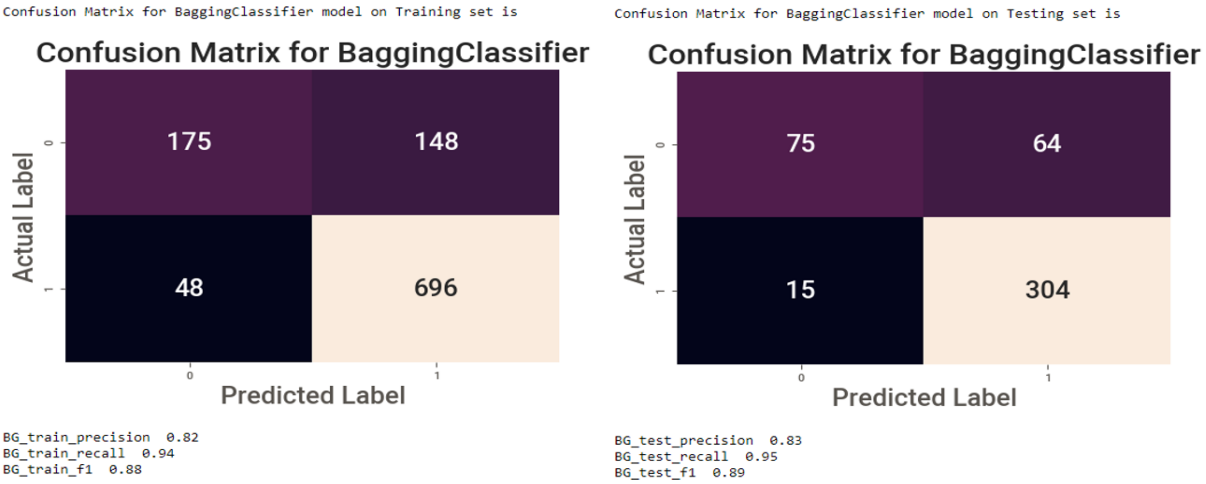


Fig 1.24. Confusion matrix Bag

- AUC and ROC: AUC Train:88.0% and AUC Test:89.5%

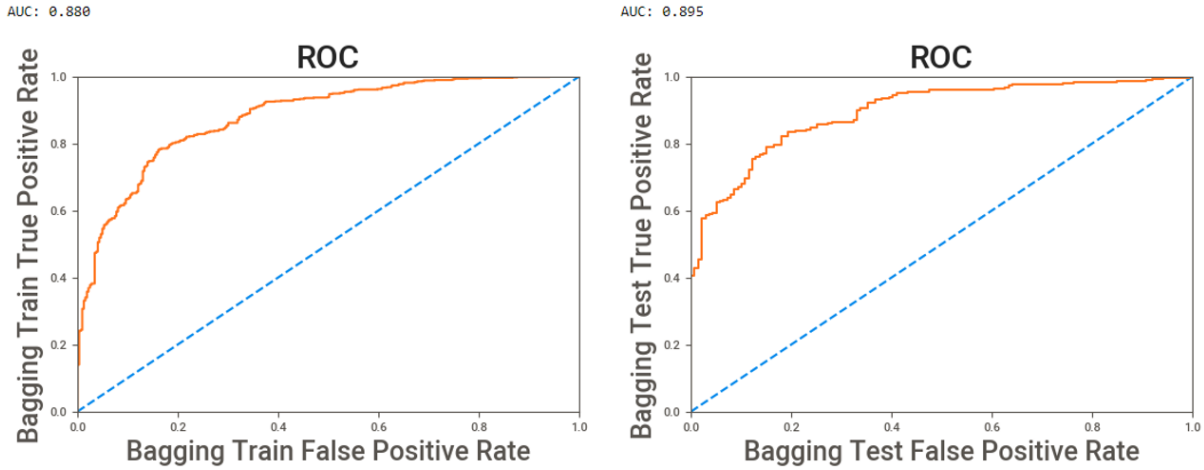


Fig 1.25. AUC and ROC – Bag

Model Evaluation- Gradient Boosting:

- Accuracy on training set is 88.8% and on testing 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

Table 1.27. Classification report (Gradient – Train)

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

Table 1.28. Classification report (Gradient – Test)

- Confusion matrix

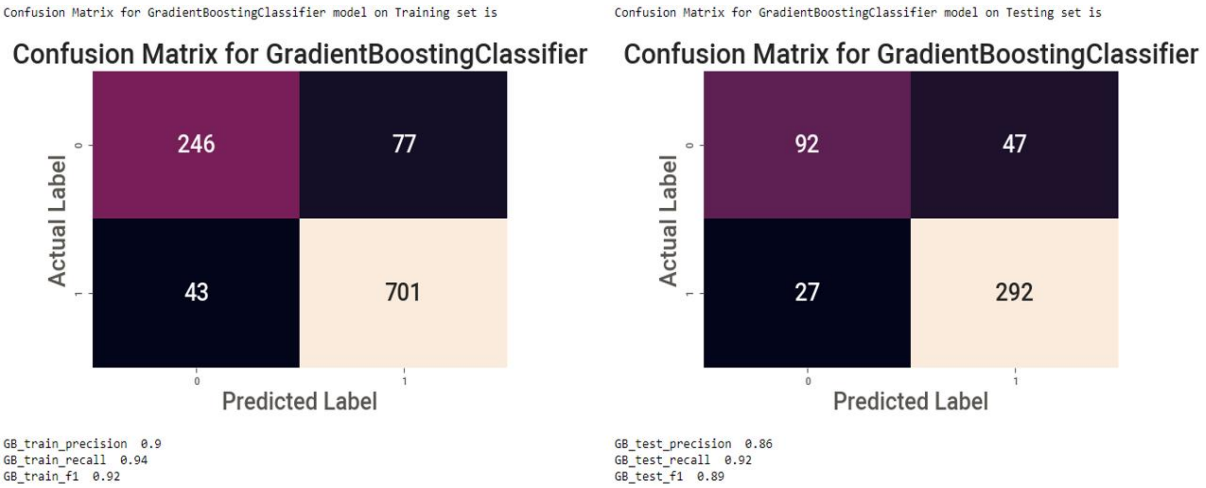


Fig 1.26 Confusion matrix Gradient

- AUC and ROC: AUC Train:94.8% and AUC Test:90.8%

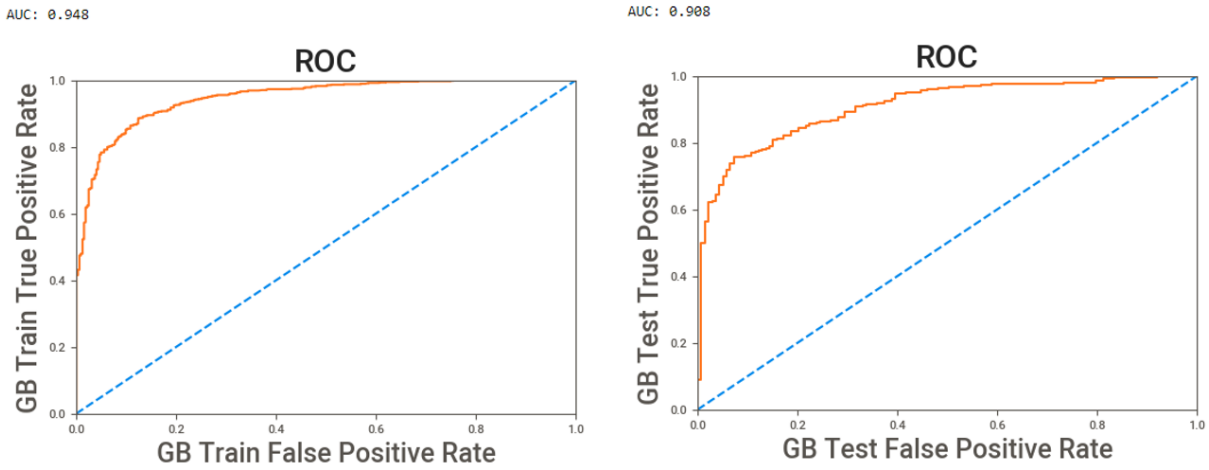


Fig 1.27 AUC and ROC – Gradient

Model Evaluation- Ada Boosting:

- Accuracy on training set is 84.4% and on testing 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

Table 1.29. Classification report (Ada – Train)

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

Table 1.30. Classification report (Ada – Test)

- Confusion matrix

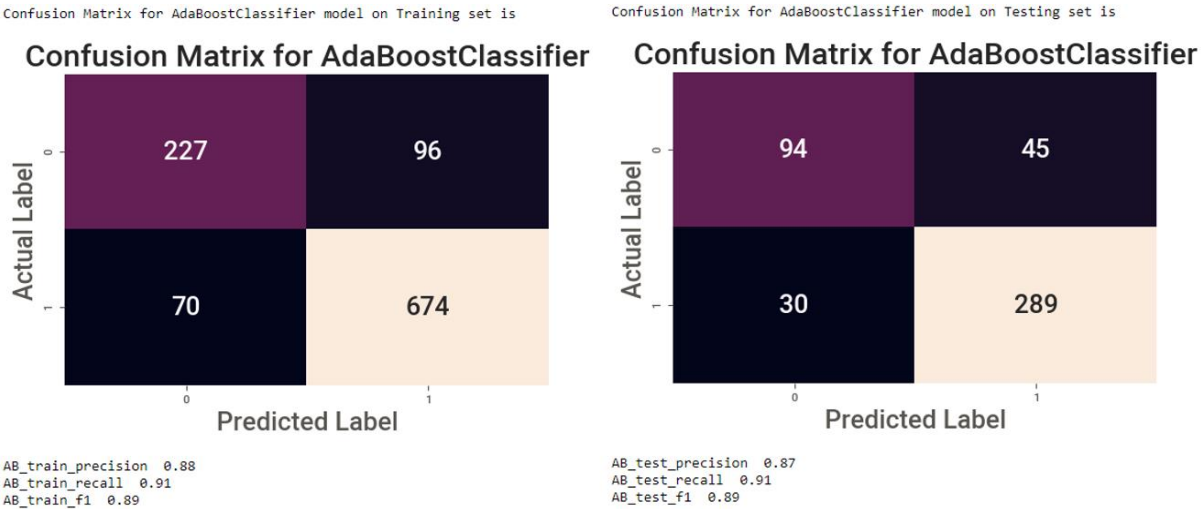


Fig 1.28 Confusion matrix Ada Boosting

- AUC and ROC: AUC Train:90.2% and AUC Test:90.6%

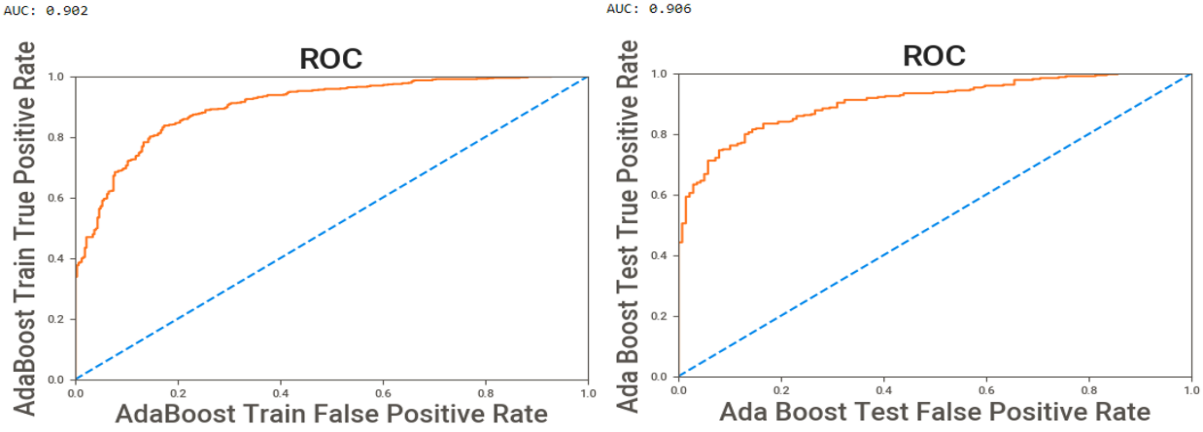


Fig 1.29 AUC and ROC – Ada Boosting

Q1.8 Based on these predictions, what are the insights?

Solution:

	Accuracy	AUC	Recall	Precision	F1 Score
LR Train	0.827	0.876	0.90	0.86	0.88
LR Test	0.858	0.916	0.93	0.87	0.90
LDA Train	0.826	0.876	0.90	0.86	0.88
LDA Test[0.5]	0.845	0.915	0.91	0.87	0.89
LDA Test[0.4]	0.847	NaN	0.94	0.86	0.90
NB Train	0.822	0.874	0.88	0.87	0.87
NB Test[0.5]	0.847	0.910	0.90	0.88	0.89
NB Test[0.4]	0.849	NaN	0.93	0.87	0.90
KNN Train	0.843	0.911	0.91	0.87	0.89
KNN Test	0.860	0.894	0.92	0.89	0.90
RF Train	0.820	0.888	0.93	0.83	0.88
RF Test	0.828	0.902	0.95	0.83	0.88
Bagging Train	0.816	0.880	0.94	0.82	0.88
Bagging Test	0.828	0.895	0.95	0.83	0.89
Gradient-Boost Train	0.888	0.948	0.94	0.90	0.92
Gradient-Boost Test	0.838	0.908	0.92	0.86	0.89
AdaBoost Train	0.844	0.902	0.91	0.88	0.89
AdaBoost Test	0.836	0.906	0.91	0.87	0.89

Table 1.31. Comparison of models

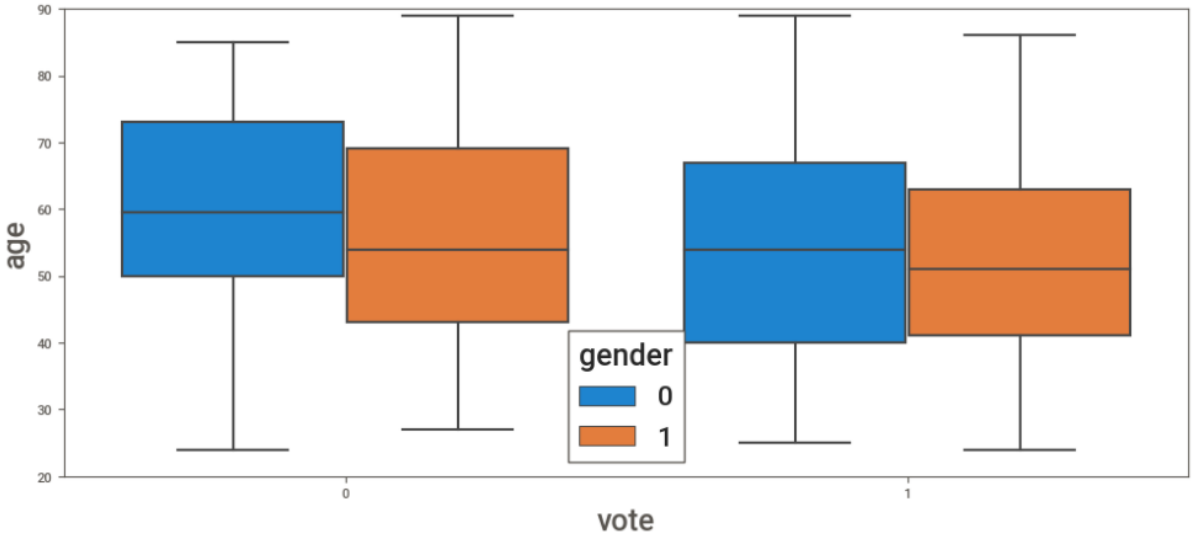


Fig 1.30 Age vs Actual Vote

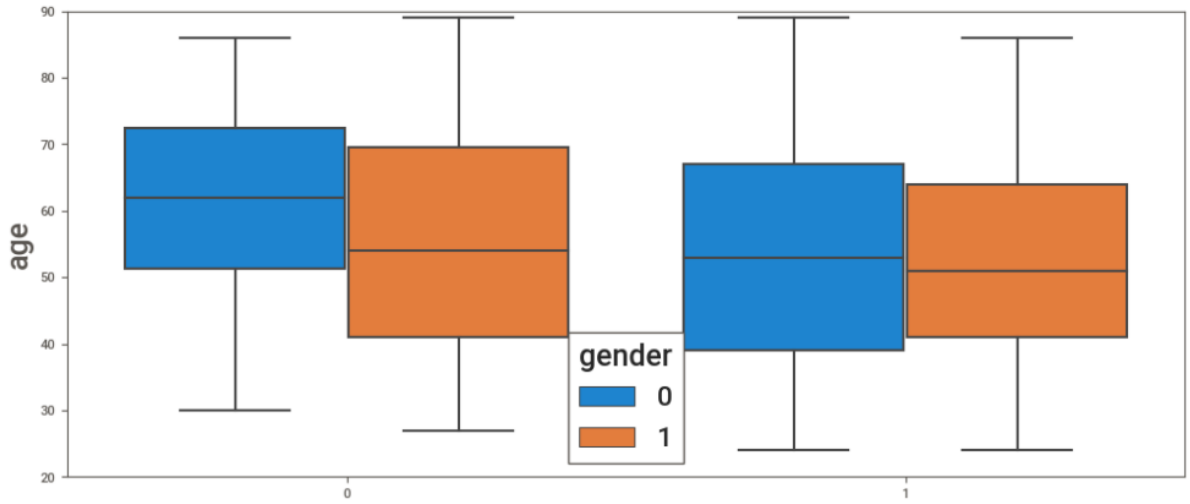


Fig 1.31 Age vs Predicted Vote

- 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 favor 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 favor were actually predicted against the Labour party.
- If people prefer Europe sentiments, then there is a high chance for them to vote for Conservative party
- 55% of voters have given assessment score of 4 for the Blair and Only 37 percentage of voters have given assessment score of 4 for the Hague.
  So, Based on Assessment labor party leader Blair has more assessment majority score of 4 than Conservative party Hague (Majority score for Hague is 2).
- If the voter has political knowledge of 2 then there is more chance for them to vote for conservative. If the voter has political knowledge of 0 and 3 then there is more chance for them to vote for Labor.
- More than 51 percentage of people belonging to age of 50 or less. The majority of voters of this age group less than 50 supports Labor party.

Problem 2:

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

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

Q2.1. Find the number of characters, words, and sentences for the mentioned documents.

Solution:

Below is the number of characters, words and sentences for the following speech President Franklin D. Roosevelt in 1941, President John F. Kennedy in 1961 and President Richard Nixon in 1973 respectively.

	1941-Roosevelt.txt
Character	7571
Words	1536
Sentences	68
	1961-Kennedy.txt
Character	7618
Words	1546
Sentences	52
	1973-Nixon.txt
Character	9991
Words	2028
Sentences	69

Table 2.1. Dataset characters, words and sentences

Q2.2 Remove all the stopwords from all 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.

1941-Roosevelt	1961-Kennedy	1973-Nixon
'national'	'vice'	'mr'
'day'	'president'	'vice'
'inauguration'	'johnson'	'president'
'since'	'mr'	'mr'
'1789'	'speaker'	'speaker'
'people'	'mr'	'mr'
'renewed'	'chief'	'chief'
'sense'	'justice'	'justice'
'dedication'	'president'	'senator'
'united'	'eisenhower'	'cook'
'states'	'vice'	'mrs'
'washington'	'president'	'eisenhower'
'day'	'nixon'	'fellow'
'task'	'president'	'citizens'
'people'	'truman'	'great'
'create'	'reverend'	'good'
'weld'	'clergy'	'country'
'together'	'fellow'	'share'
'nation'	'citizens'	'together'

Table 2.2. Cleaned words from 3 speeches

Q2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words.

Solution:

Most Frequent 3 words as below post cleaning.

1941-Roosevelt		1961-Kennedy		1973-Nixon	
	Count		Count		Count
nation	12	let	16	us	26
know	10	us	12	let	22
spirit	9	sides	8	america	21

Table 2.3. Top 3 words post cleaning

Q2.4 Plot the word cloud of each of the speeches of the variable

Solution:

Word Cloud for 1941-Roosevelt Speech (after cleaning):

Fig 2.1 Cloud – Speech 1941

Word Cloud for 1961-Kennedy Speech (after cleaning):

Fig 2.2 Cloud – Speech 1961

Word Cloud for 1973-Nixon Speech (after cleaning):

Fig 2.3 Cloud – Speech 1973

Thanks & regards,
Pavan Kumar R Naik