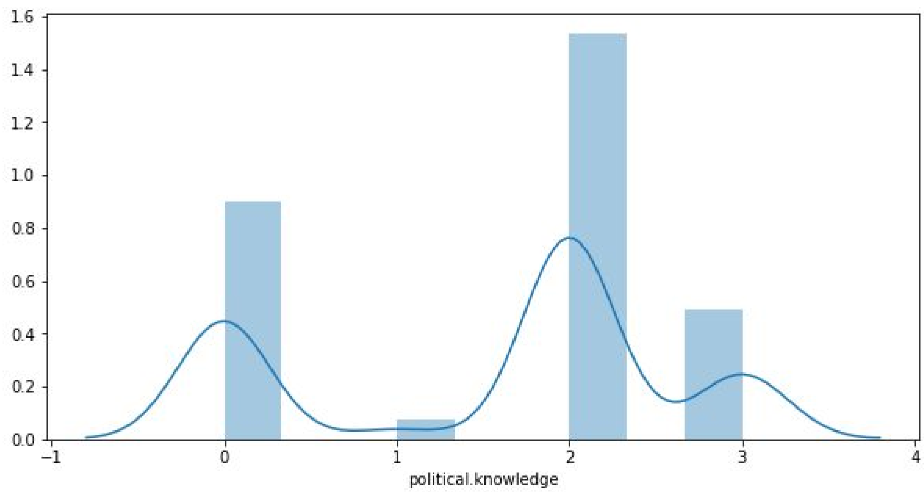
Machine Learning Project Report
Problem A

Data Ingestion

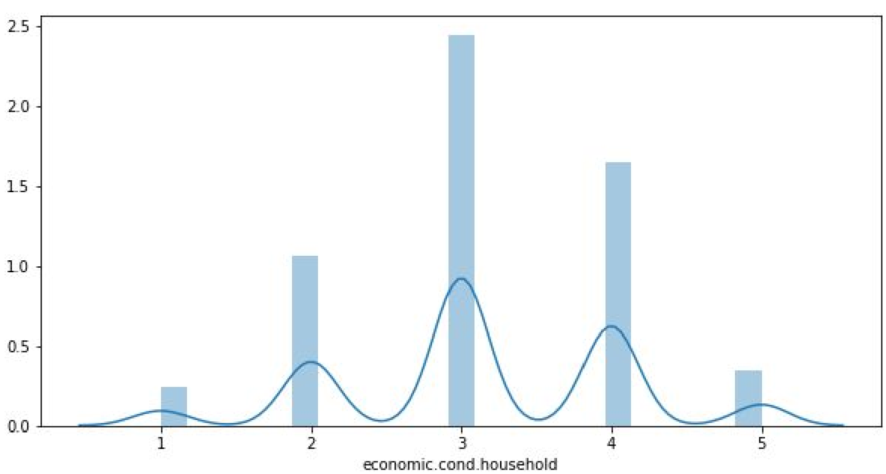
- Data Consist of **1525 Rows and 10 Columns**
- Data consists of **Duplicate** rows - 8 duplicate rows
- Data does not consist of **Missing Values - Null Value check**
- Data consist of **70%** votes in favour of **Labour** and 30% in Conservative
- There 5 types of categories of economic cond national and Political growth

Data Visualization

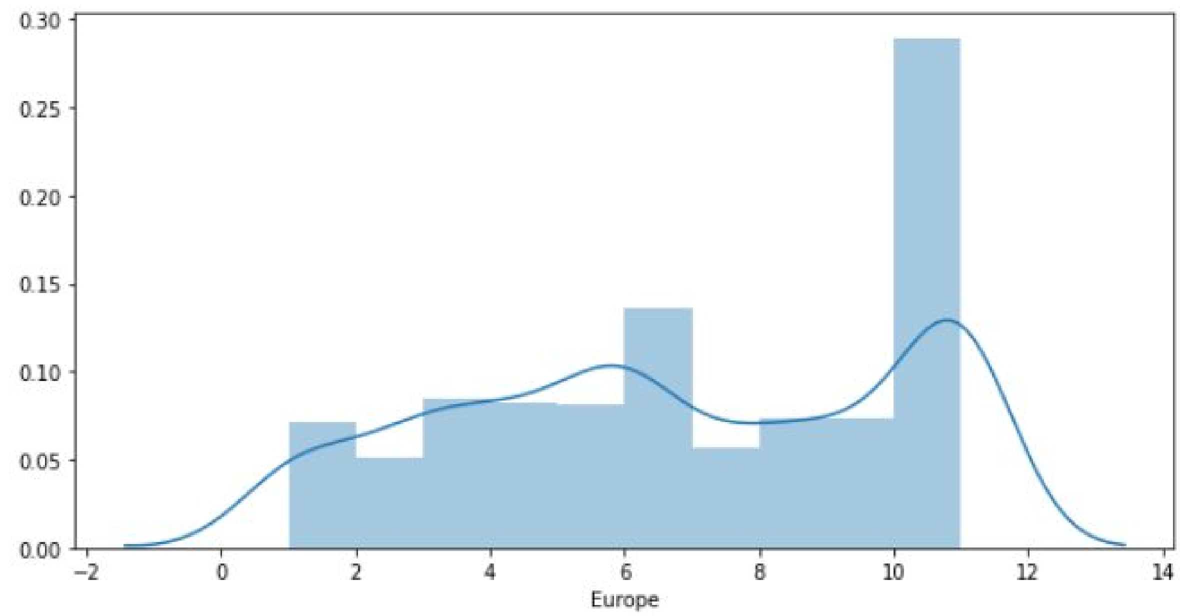
Maximum data is present for category of **Type 2** of **Political knowledge**



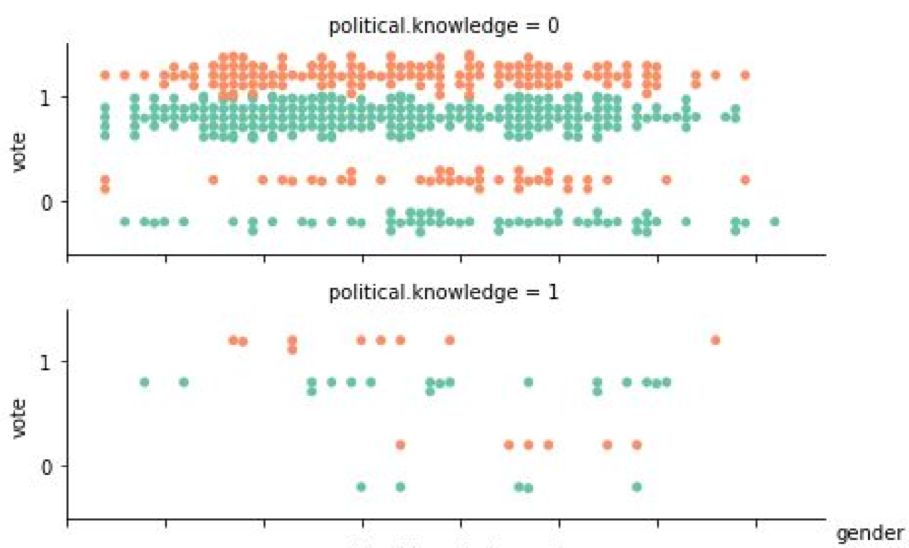
Maximum data is present for **Type 3 of Economic condition house hold**

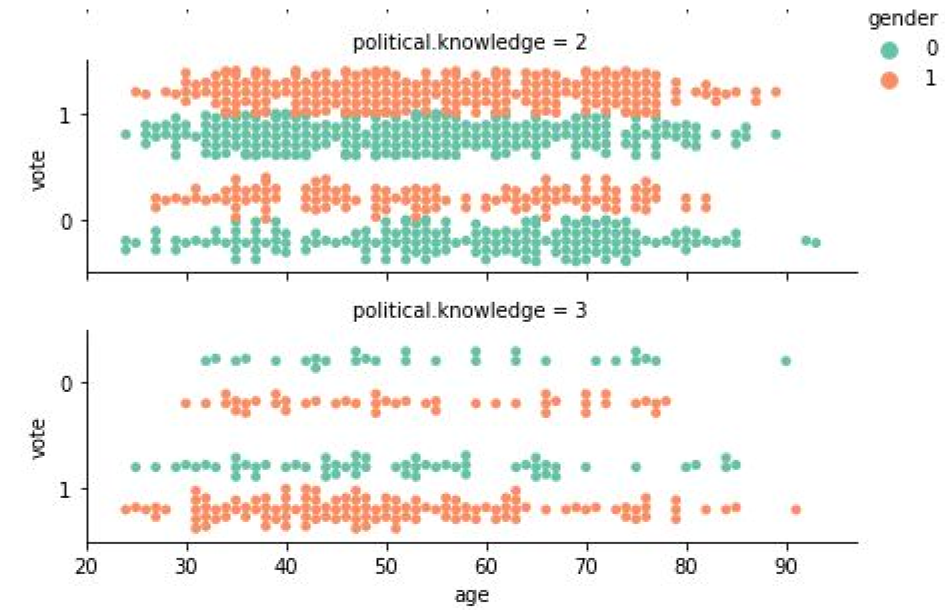


We can observe the data has high value of attitude wrt to European intergation



From the below plot we can understand the how the vote very to the 2 parties 0-Conservative and 1 - labour on based on the political knowledge and gender 0 -Female, 1- male.

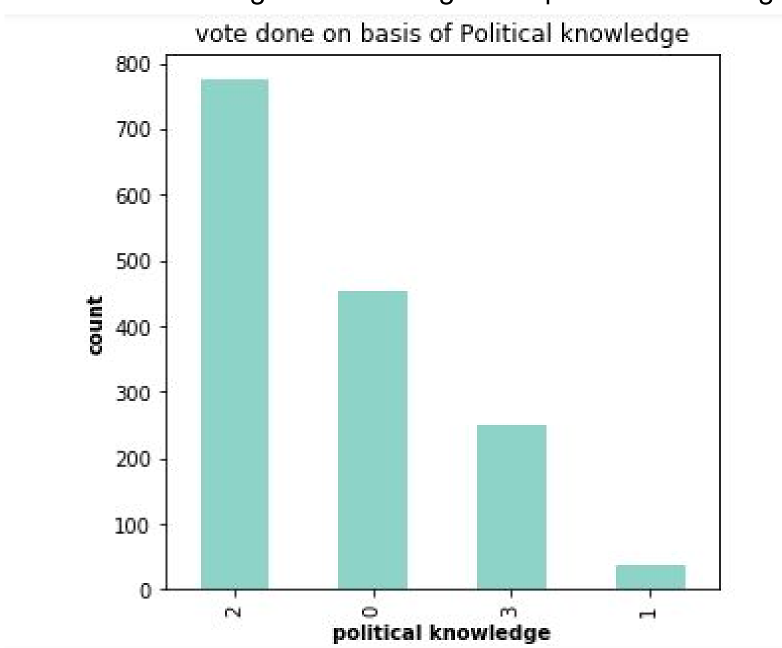




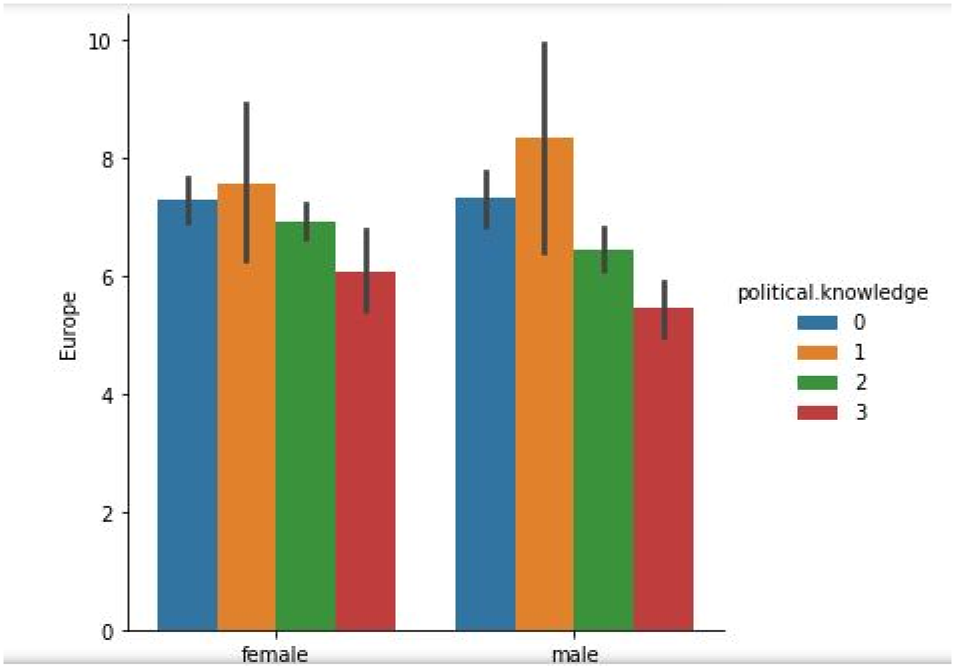
We can clearly see that maximum population is of political knowledge = 2
Where as in

- political knowledge = 0 is of male and have voted for labour
- political knowledge = 1 is very sparsely populated
- political knowledge = 2 has almost balance voting tending towards Labour
- political knowledge = 3 has voted for Labour

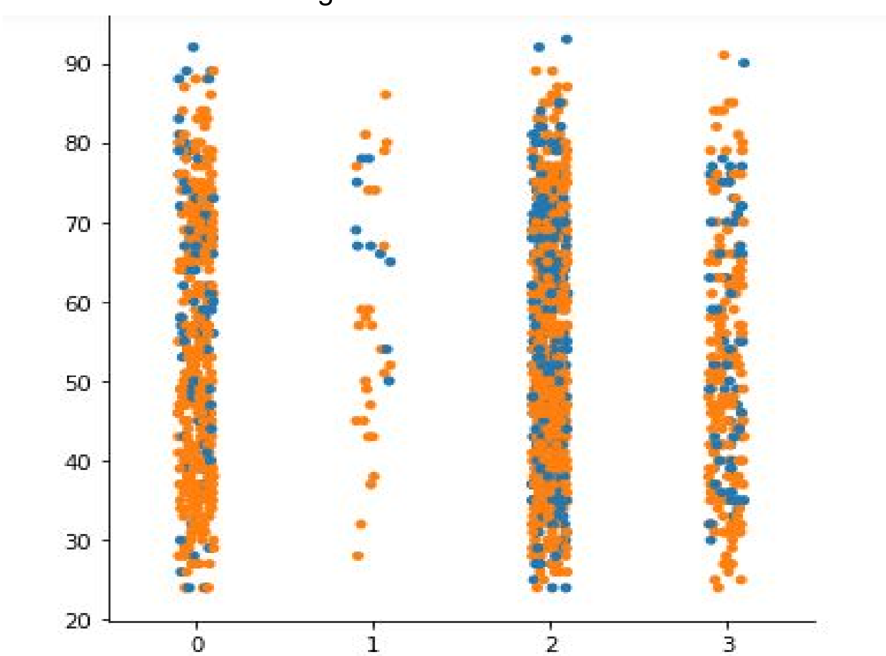
The count of votes given according to the political knowledge



Highest attitude of respondents to European Integration can be seen in males and females of **Political knowledge 1**
Lowest attitude of respondents to European Integration can be seen in males and females of **Political knowledge 3**

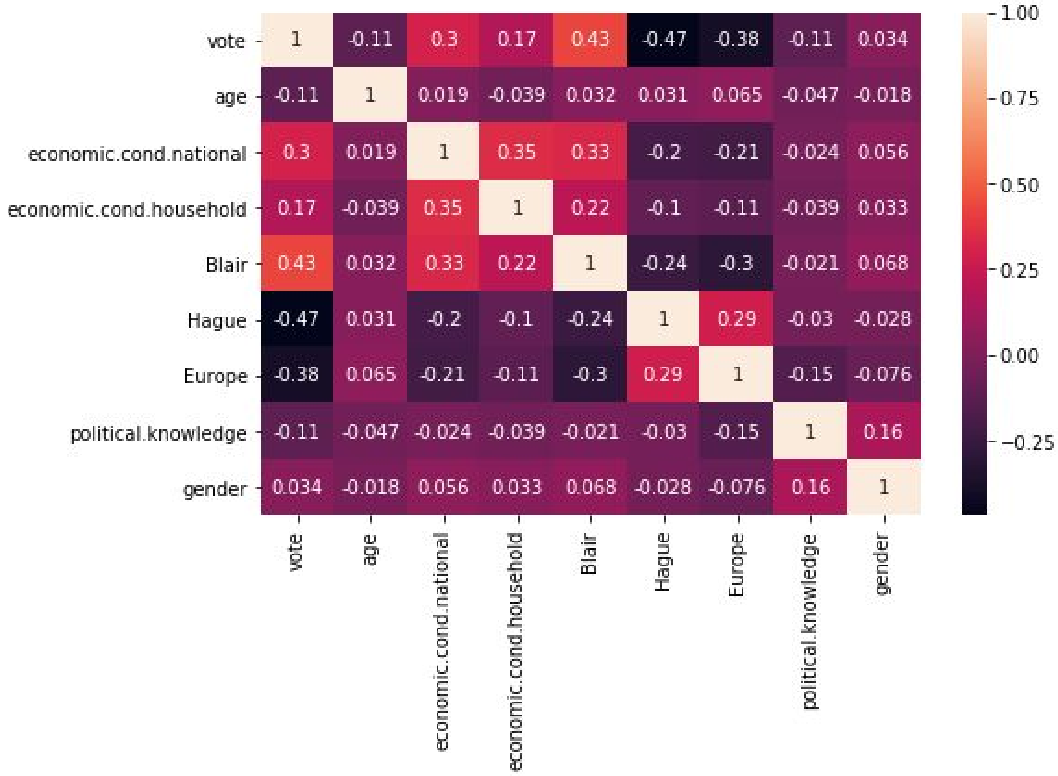


In the below graph, x axis is the political knowledge of people while y axis is the age.
The hue shows the vote given



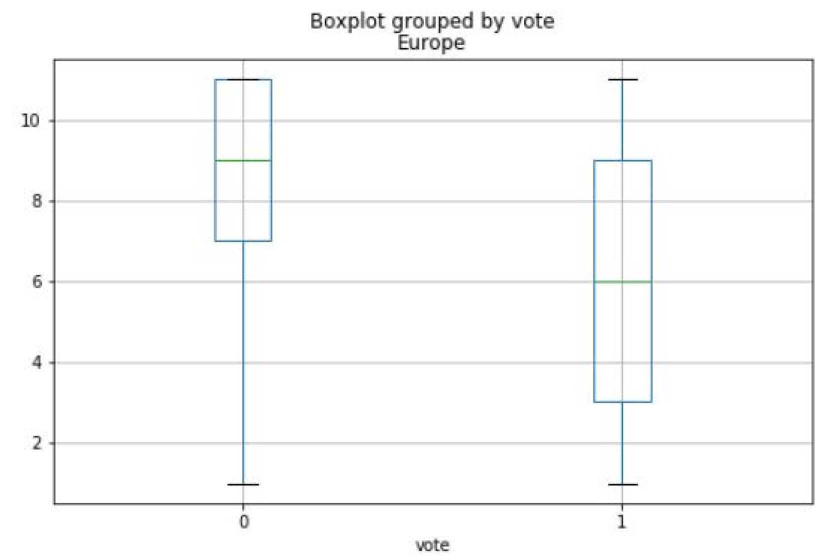
Heat Map

Correlation between the features of the data is shown by the heat map:



We can conclude there is **no positive correlation** between the features of the data and **negative 0.50 correlation between Hague and Vote** .

With increase in attitude (Europe) Vote for conservative party has increased

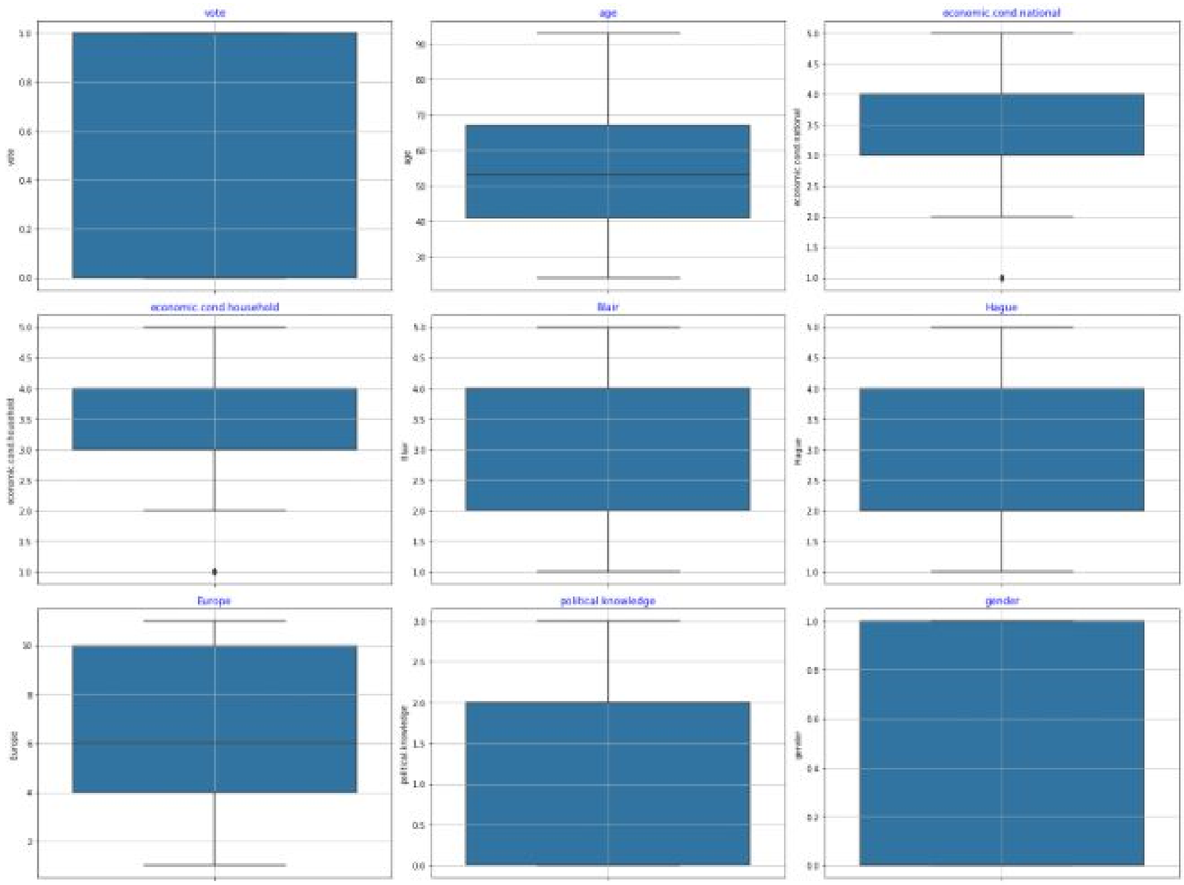


<u>Displot with hue as votes</u>

- We can interpret that **age is bimodal** whereas **national and household economic conditions are multimodal**
- Also with increase in national economic conditions, maximum vote tends to go in favour to labour

- As the attitude of respondents to European Integration increases we can see the vote going in favor of Conservative party.

Treatment of Outliers



We can observer outliers in economic national and household features of the table.
We can treat that by finding the IQR and reducing the range by replacing it .

Scaling

As data is Categorical scaling is not required.
This can be proved by implementing Logistic Regression with scaled data .

There was no significant difference in AUC score or confusion matrix of the scaled and unscaled data. Hence Scaling is not necessary in this case.

Data Encoding and Spliting

This needs to be done in order to convert Object datatype to int. Spliting is required to as to not over fit the model.

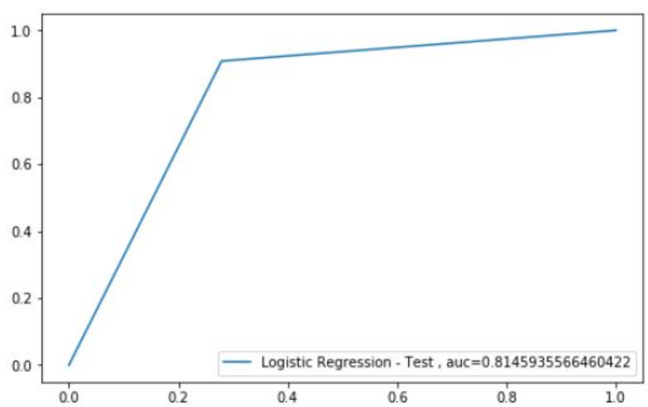
Logistic Regression and LDA

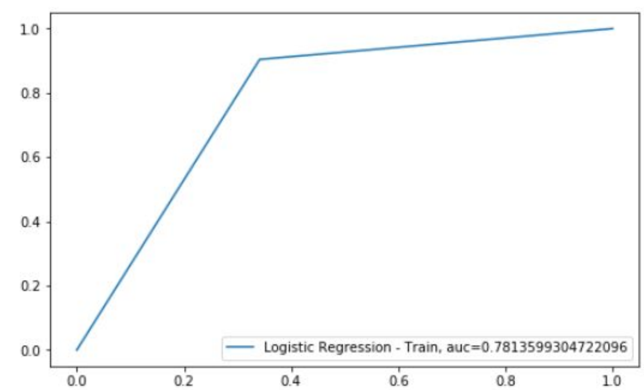
LDA

```
Model Score for Test Data  0.8464912280701754
Model Score for Train Data  0.8226027397260274
--------------------Test LDA SMOTE Model------------------------
[[108  21]
 [ 49 278]]
              precision    recall  f1-score   support

           0       0.69      0.84      0.76       129
           1       0.93      0.85      0.89       327

    accuracy                           0.85       456
   macro avg       0.81      0.84      0.82       456
weighted avg       0.86      0.85      0.85       456

--------------------Train LDA SMOTE Model------------------------
[[609 121]
 [138 592]]
              precision    recall  f1-score   support

           0       0.82      0.83      0.82       730
           1       0.83      0.81      0.82       730

    accuracy                           0.82      1460
   macro avg       0.82      0.82      0.82      1460
weighted avg       0.82      0.82      0.82      1460
```

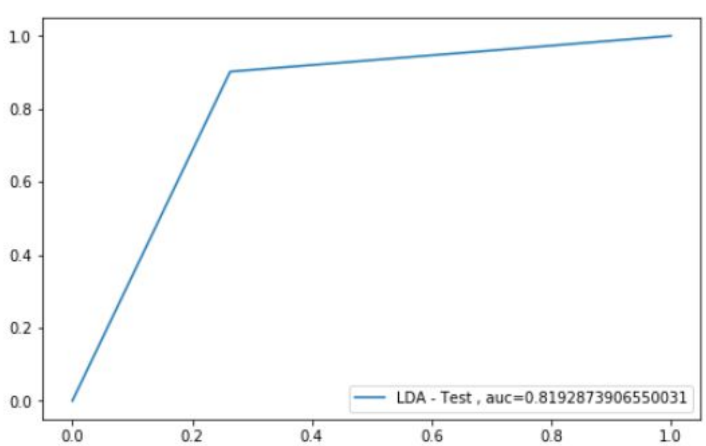
Logistic Regression

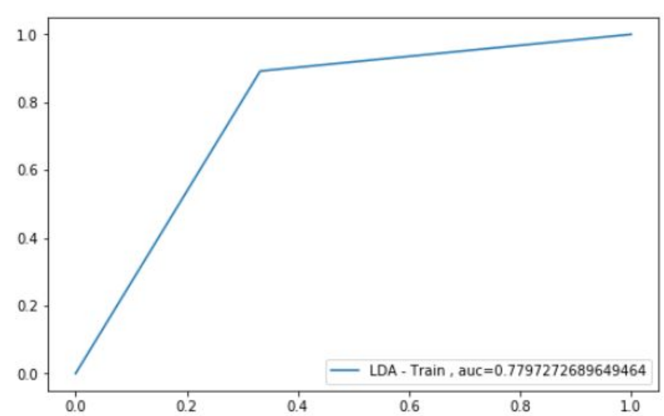
```
Model Score for Test Data  0.8464912280701754
Model Score for Train Data  0.8226027397260274
--------------------Test LR SMOTE Model------------------------
[[108  21]
 [ 49 278]]
              precision    recall  f1-score   support

           0       0.69      0.84      0.76       129
           1       0.93      0.85      0.89       327

    accuracy                           0.85       456
   macro avg       0.81      0.84      0.82       456
weighted avg       0.86      0.85      0.85       456

--------------------Train LR SMOTE Model------------------------
[[607 123]
 [136 594]]
              precision    recall  f1-score   support

           0       0.82      0.83      0.82       730
           1       0.83      0.81      0.82       730

    accuracy                           0.82      1460
   macro avg       0.82      0.82      0.82      1460
weighted avg       0.82      0.82      0.82      1460
```



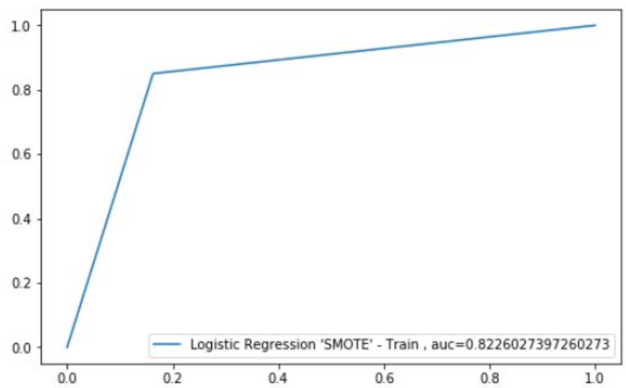


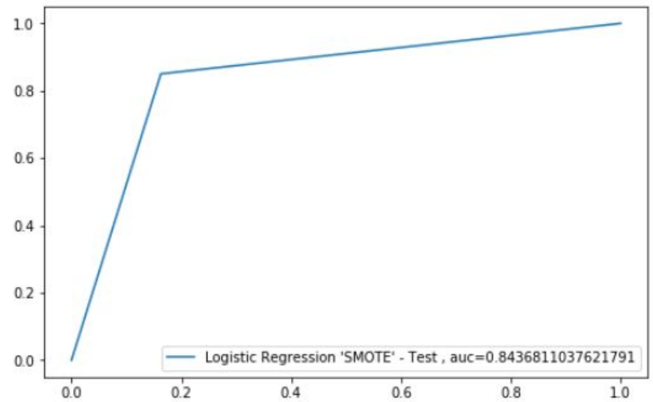
LDA Train Test AUC score



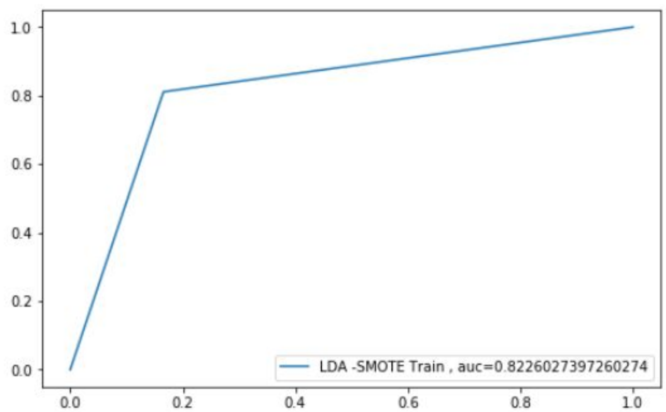


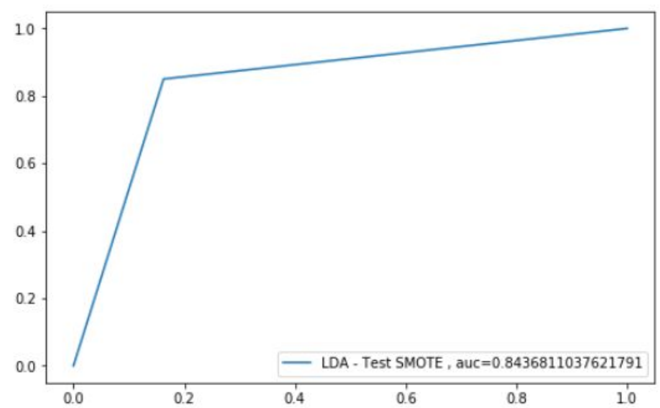
LR Smote Train and Test - AUC score





LDA Train and Test - AUC score





When comparing both the models, we can understand

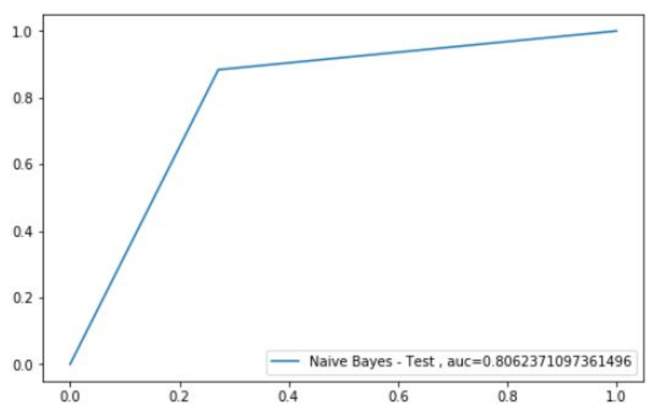
- With **SMOTE** data , precision, accuracy , auc score of test as well as train data is **nearly same for both**
- While general data we can analysis, auc score of LDA - test data is comparatively better than Logistic Regression
- The AUC score was drastically increased with SMOTE

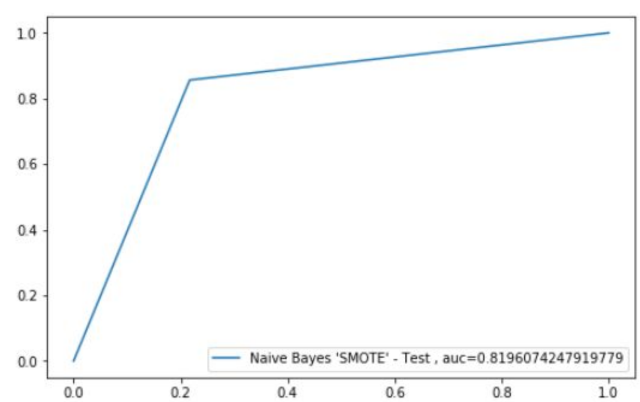
<u>**Naive Bayes**</u>

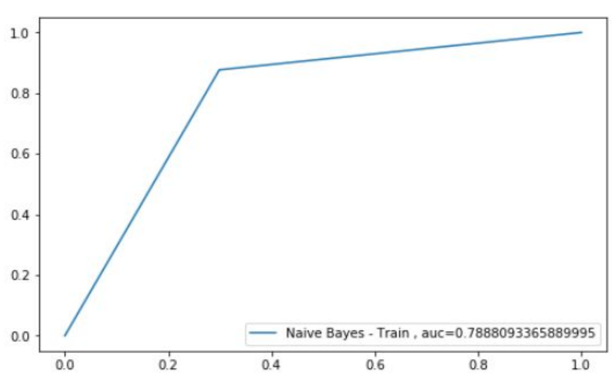
We can see a significant increase in AUC score when used NB with **SMOTE** data

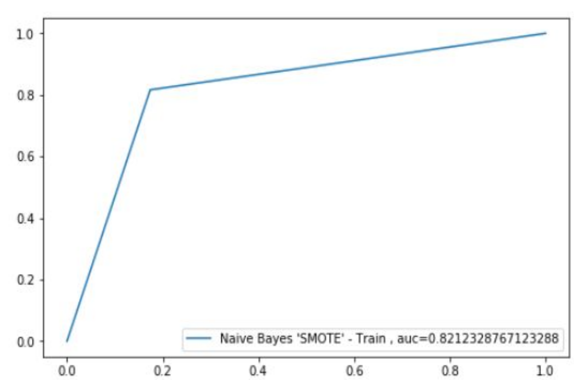
```
Model Score for Train Data  0.82186616399623
---------------------Test Naive Bayes Model---------------
[[ 94  35]
 [ 38 289]]
              precision    recall  f1-score   support

           0       0.71      0.73      0.72       129
           1       0.89      0.88      0.89       327

    accuracy                           0.84       456
   macro avg       0.80      0.81      0.80       456
weighted avg       0.84      0.84      0.84       456

---------------------Train Naive Bayes Model--------------
[[232  99]
 [ 90 640]]
              precision    recall  f1-score   support

           0       0.72      0.70      0.71       331
           1       0.87      0.88      0.87       730

    accuracy                           0.82      1061
   macro avg       0.79      0.79      0.79      1061
weighted avg       0.82      0.82      0.82      1061
```

Naive Bayes | **Naive Bayes SMOTE**









KNN

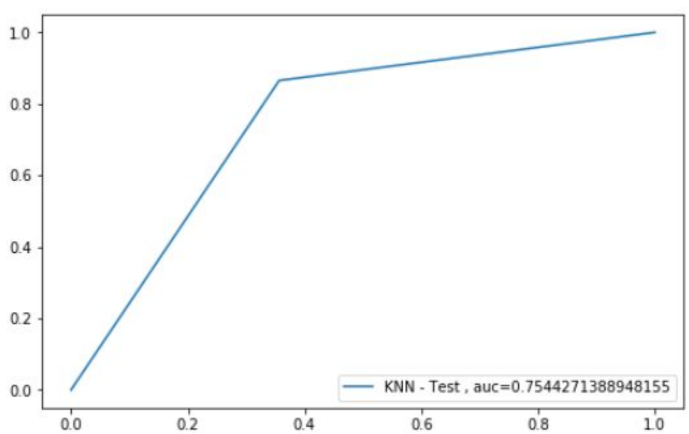
KNN

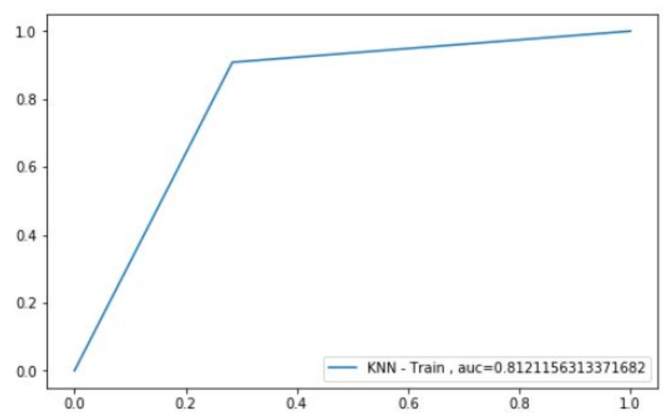
```
Model Score for Test Data  0.8026315789473685
Model Score for Train Data  0.8482563619227145
---------------------Test KNN Model--------------------
[[ 83  46]
 [ 44 283]]
              precision    recall  f1-score   support

           0       0.65      0.64      0.65       129
           1       0.86      0.87      0.86       327

    accuracy                           0.80       456
   macro avg       0.76      0.75      0.76       456
weighted avg       0.80      0.80      0.80       456

---------------------Train KNN Model-------------------
[[237  94]
 [ 67 663]]
              precision    recall  f1-score   support

           0       0.78      0.72      0.75       331
           1       0.88      0.91      0.89       730

    accuracy                           0.85      1061
   macro avg       0.83      0.81      0.82      1061
weighted avg       0.85      0.85      0.85      1061
```

KNN - SMOTE

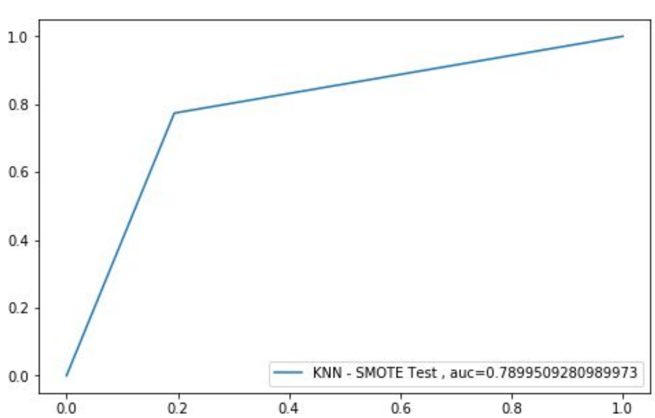
```
Model Score for Test Data  0.7828947368421053
Model Score for Train Data  0.8821917808219178
--------------------Test KNN Model SMOTE------------------
[[104  25]
 [ 74 253]]
              precision    recall  f1-score   support

           0       0.58      0.81      0.68       129
           1       0.91      0.77      0.84       327

    accuracy                           0.78       456
   macro avg       0.75      0.79      0.76       456
weighted avg       0.82      0.78      0.79       456

--------------------Train KNN Model SMOTE-----------------
[[685  45]
 [127 603]]
              precision    recall  f1-score   support

           0       0.84      0.94      0.89       730
           1       0.93      0.83      0.88       730

    accuracy                           0.88      1460
   macro avg       0.89      0.88      0.88      1460
weighted avg       0.89      0.88      0.88      1460
```

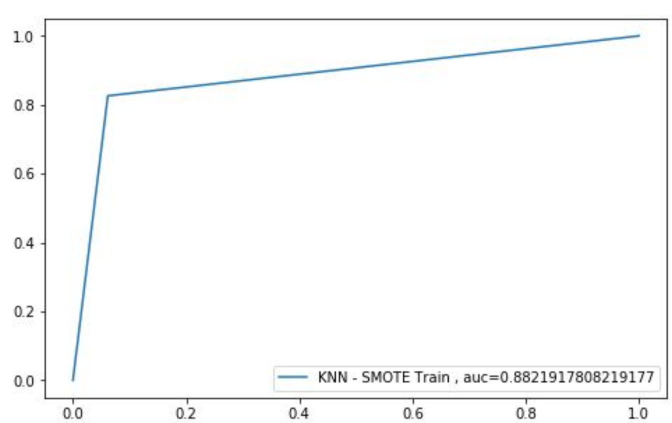
KNN -Train Test AUC





KNN -Train Test AUC - SMOTE





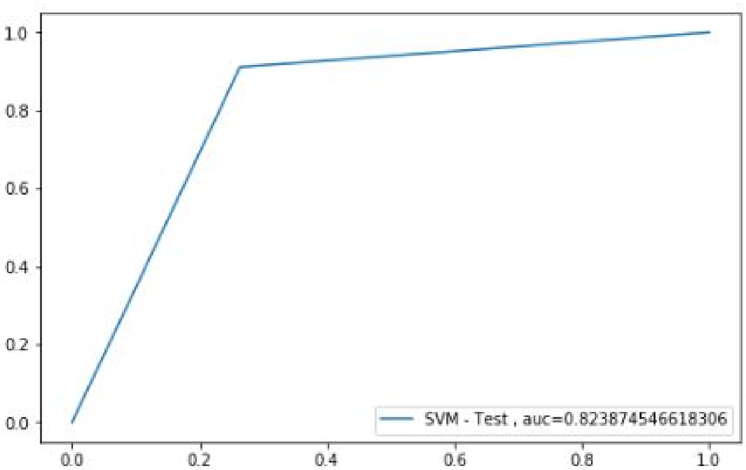
SVM

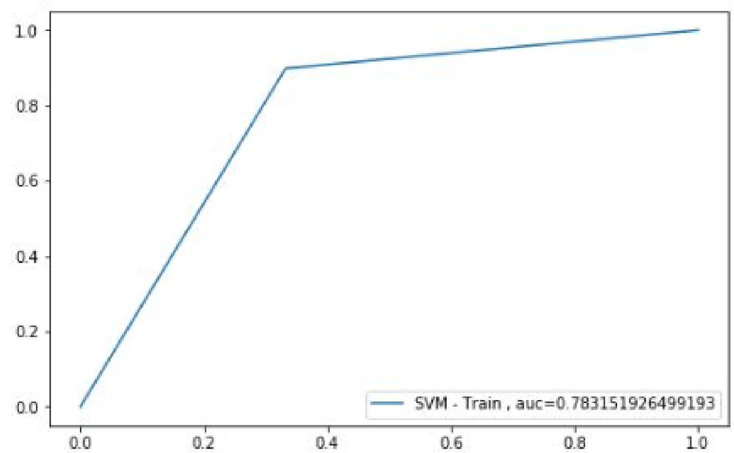
SVM

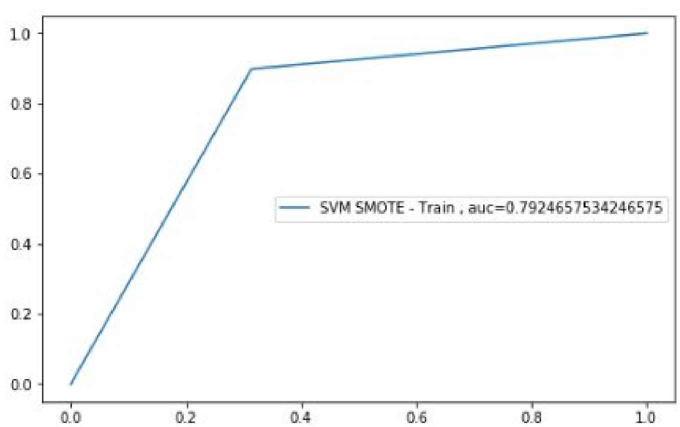
```
Model Score for Test Data  0.8618421052631579
Model Score for Train Data  0.8265786993402451
---------------------Test SVM Model ----------------------
[[ 95  34]
 [ 29 298]]
              precision    recall  f1-score   support

           0       0.77      0.74      0.75       129
           1       0.90      0.91      0.90       327

    accuracy                           0.86       456
   macro avg       0.83      0.82      0.83       456
weighted avg       0.86      0.86      0.86       456

---------------------Train SVM Model ---------------------
[[221 110]
 [ 74 656]]
              precision    recall  f1-score   support

           0       0.75      0.67      0.71       331
           1       0.86      0.90      0.88       730

    accuracy                           0.83      1061
   macro avg       0.80      0.78      0.79      1061
weighted avg       0.82      0.83      0.82      1061
```

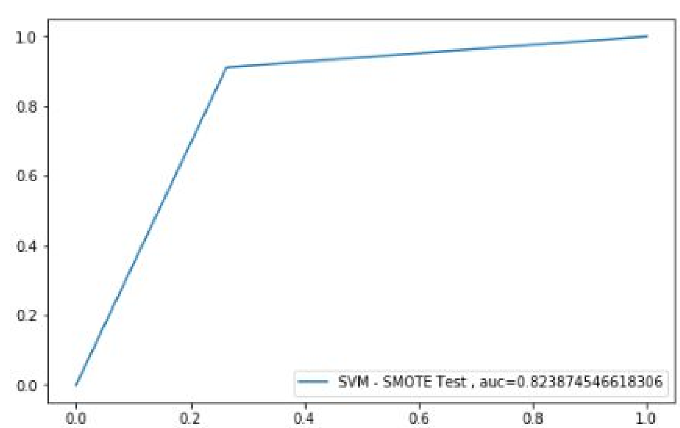
SVM SMOTE

```
Model Score for Test Data  0.8618421052631579
Model Score for Train Data  0.7924657534246575
---------------------Test SVM Model SMOTE----------------
[[ 95  34]
 [ 29 298]]
              precision    recall  f1-score   support

           0       0.58      0.81      0.68       129
           1       0.91      0.77      0.84       327

    accuracy                           0.78       456
   macro avg       0.75      0.79      0.76       456
weighted avg       0.82      0.78      0.79       456

---------------------Train SVM Model SMOTE---------------
[[501 229]
 [ 74 656]]
              precision    recall  f1-score   support

           0       0.87      0.69      0.77       730
           1       0.74      0.90      0.81       730

    accuracy                           0.79      1460
   macro avg       0.81      0.79      0.79      1460
weighted avg       0.81      0.79      0.79      1460
```



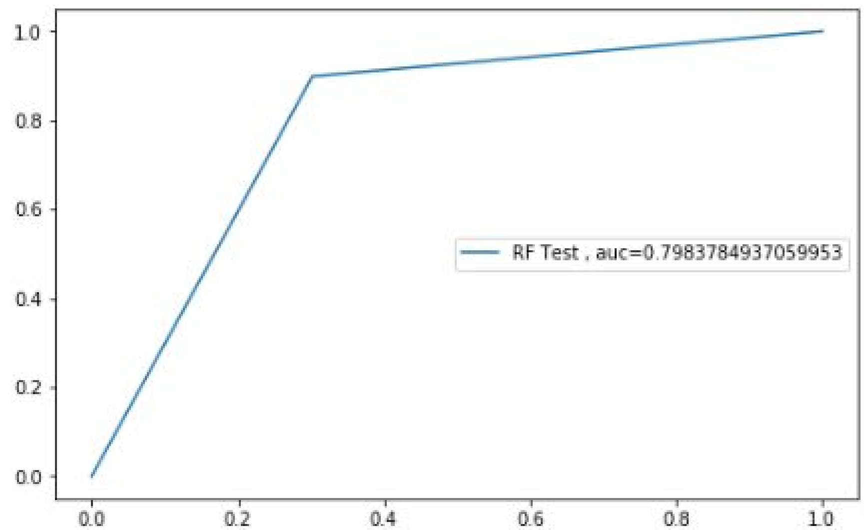






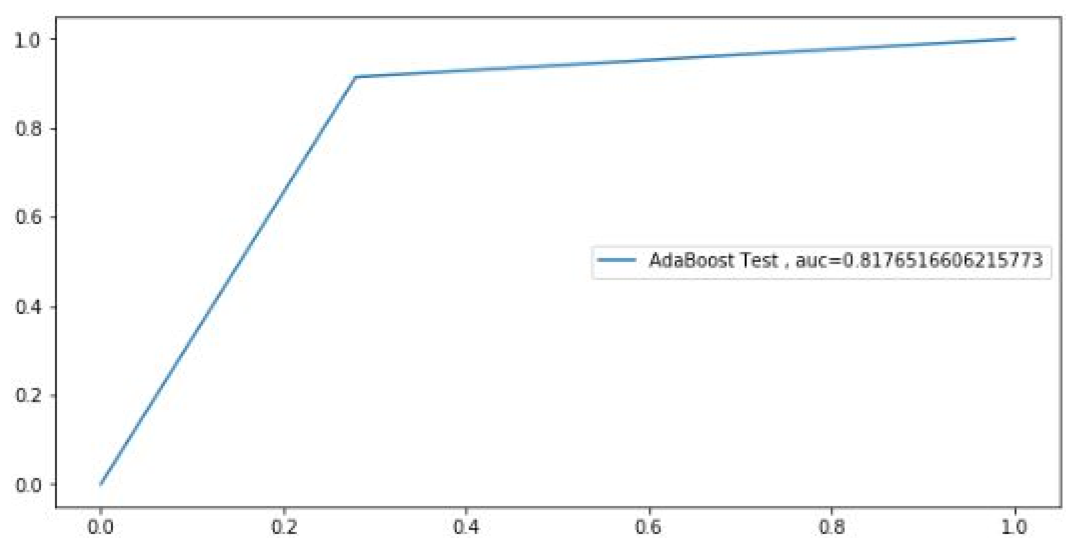
Random Forest

```
[[ 90  39]
 [ 33 294]]
```

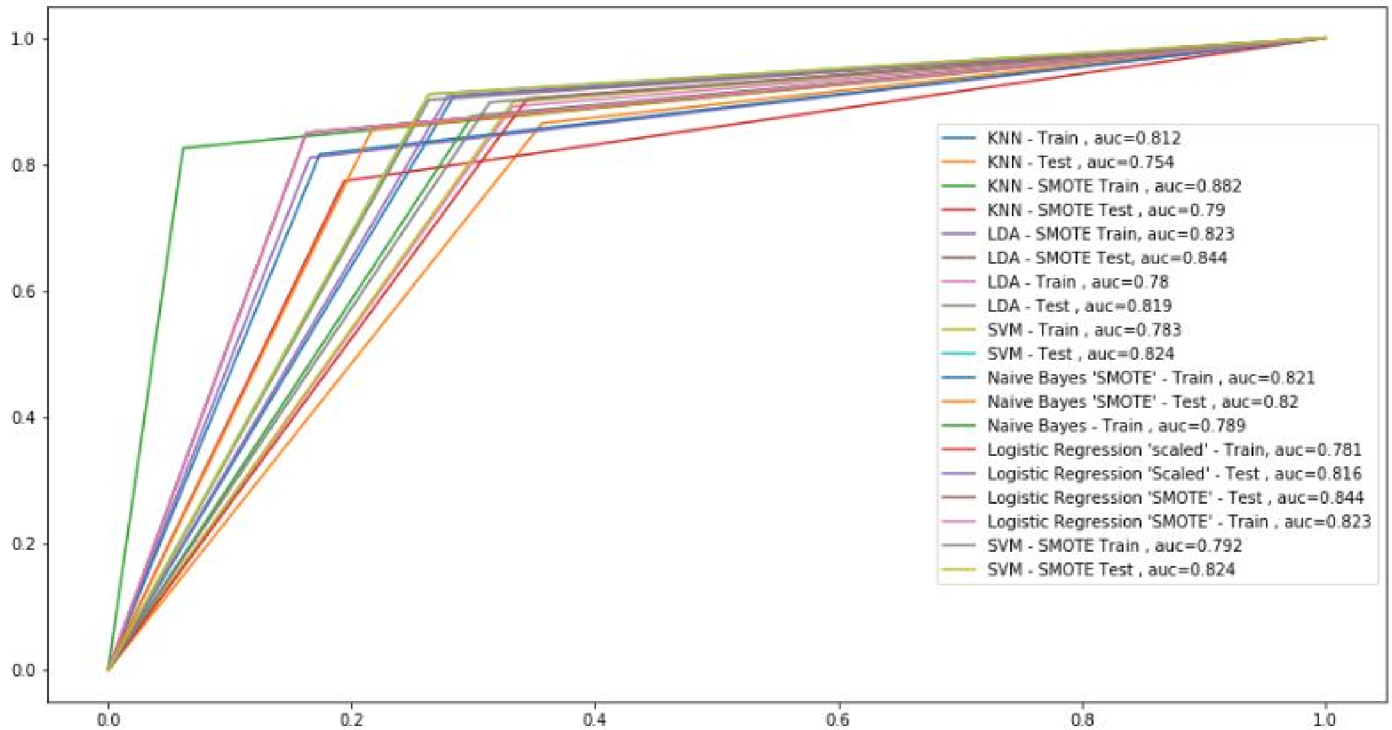


AdaBoost

```
[[ 93  36]
 [ 28 299]]
```



Compare - Find the best Model



The maximum AUC Score is for LDA and Logistic Regression with Smote AUC = 0.844

Hence we can use them for our prediction.

Problem 2

- The number of sentences - Roosevelt= 69, Kennedy =55, Nixon =70
- Number of words - Roosevelt= 1328, Kennedy =1359, Nixon =1783
- Number of Characters - Roosevelt= 6146, Kennedy =6152, Nixon =8106
- Removed all the stopwords from all the three speeches
- Which word occurs the most number of times in his inaugural address for each president? Mention the top three words.
- Roosevelt= know , us , sprit
- Kennedy = let, us, new
- Nixon = us, let, new