
# FINAL PRESENTATION

Project name: Bank Marketing Campaign

Team: Data Science Master

Date: January 12th, 2022

# Agenda

Executive Summary

Data Understanding

Data Transformation

Data Dependency

Model Building

Model Results and Cross Validation

Recommendations

# Team Member Details


GROUP NAME: DATA SCIENCE MASTER
NAME : ABHIMANYU GANGANI
EMAIL : Agangani97@gmail.com
COUNTRY : UNITED KINGDOM
COLLEGE : ANGLIA RUSKIN UNIVERSITY
SPECIALIZATION : DATA SCIENCE

# Executive Summary

- **Client :**

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps them understand whether a particular customer will buy their product or not.

- **Problem Statement :**

Build a Classification ML model to shortlist customers who are most likely to buy the term deposit product. This would allow the marketing team to target those customers through various channels.

- **Analysis :**

The Analysis of this data is divided into the following parts:

1. Data Understanding
2. Univariate analysis
3. Bivariate analysis
4. Model recommendations

# Data Understanding :

- **Dataset Description :**

Four Different datasets provided.

1. bank-additional-full: 20 inputs (+1 target variable) and 41119 observations
2. bank-additional: 20 inputs (+1 target variable) and 4119 observations b
3. bank-full: 17 inputs (+1 target variable) and 45211 observations
4. bank: 17 inputs (+1 target variable) and 4521 observations

# Data Understanding :

Data columns (total 21 columns):

| # | Column | Dtype | Description |
|---|---|---|---|
| 0 | age | int64 | Age of Client. |
| 1 | job | object | Type of Job. |
| 2 | marital | object | Marital Status. |
| 3 | education | object | Level of Education. |
| 4 | default | object | Has credit in default? |
| 5 | housing | object | Has housing loan? |
| 6 | loan | object | Has personal loan? |
| 7 | contact | object | How client has been communicated? |
| 8 | month | object | last contacted month. |
| 9 | day_of_week | object | last contacted day. |
| 10 | duration | int64 | duration of communication(seconds). |
| 11 | campaign | int64 | number of contacts performed in Campaign. |
| 12 | pdays | int64 | number of days passed after contact. |
| 13 | previous | int64 | number of total contacts performed. |
| 14 | poutcome | object | outcome of the previous campaign. |

# Data Understanding :

Data columns (total 21 columns):

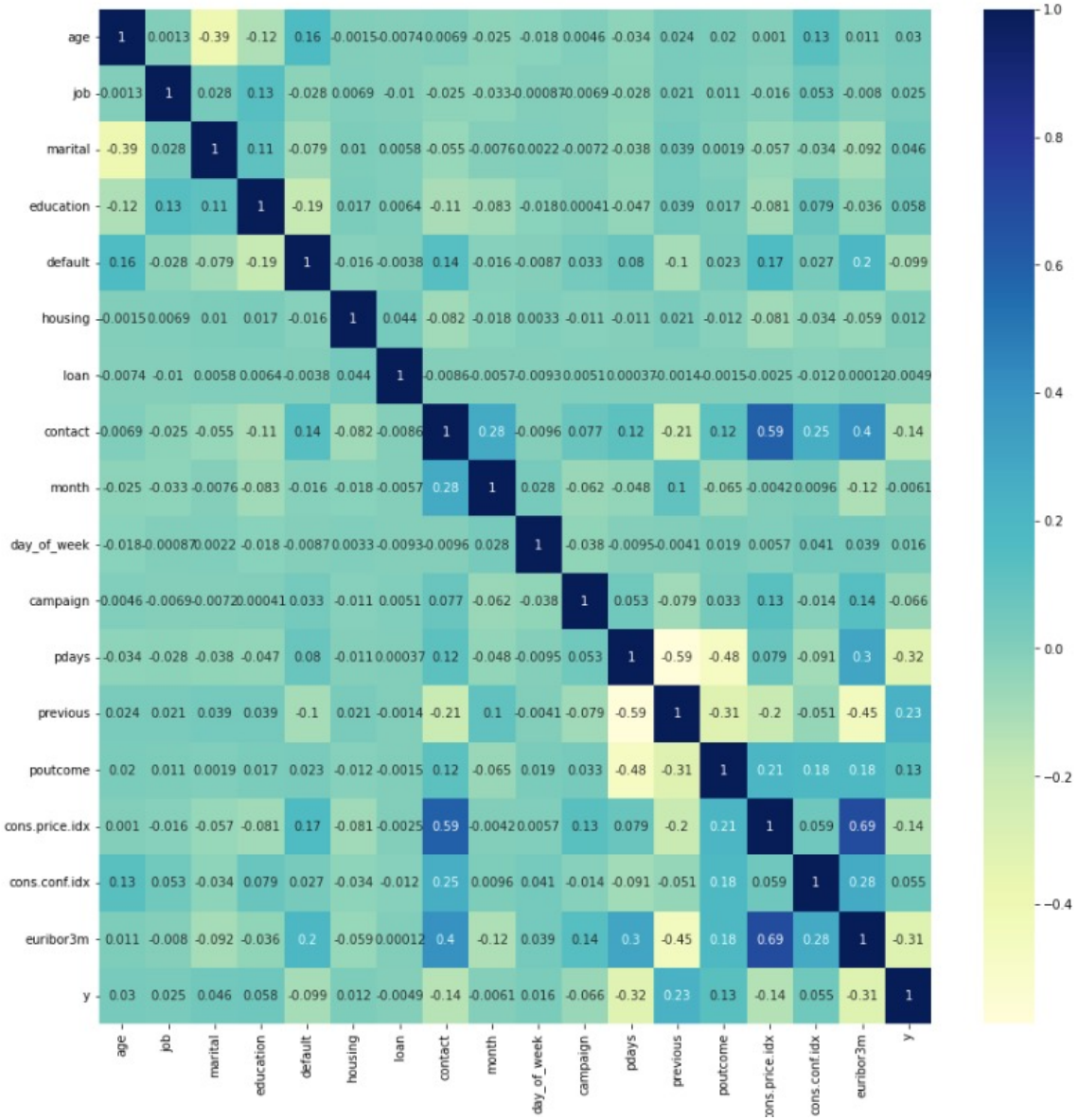
| # | Column | Dtype | Description |
|---|---|---|---|
| 15 | emp.var.rate | float64 | Employment variation rate. |
| 16 | cons.price.idx | float64 | Consumer price index. |
| 17 | cons.conf.idx | float64 | Consumer confidence index. |
| 18 | euribor3m | float64 | Euribor 3 months rate. |
| 19 | nr.employed | float64 | number of employees. |
| 20 | y | object | has the client subscribed product. |

# Data Transformation :

- **Assumptions and transformations :**

1. Timeline of observations - May 2008 to November 2010.
2. Dropping duration feature as it highly affects the target variable y, if the call is not performed the duration will have value 0 and this makes the target variable 0 as well for corresponding entry. This will hinder the realistic predictive model.
3. A frequently occurring missing value ‘unknown’ is considered as another category for the categorical features.
4. Duplicated rows were deleted from the dataset.
5. Outliers are not dropped as these seems to be realistic values for age and campaign feature.
6. Heat map shows high correlation between ‘emp.var.rate’, nr.employed’ and ‘euribor3m’. We will drop two features ‘emp.var.rate’, nr.employed’ as euribor3m shows us the money strength in the current market.

# Data Dependency (Correlation)




After changing all the variables to numerical we will plot heatmap to check correlation between variables. We consider 0.8 as the thresholds which is not observed between any features in correlation matrix.

Please Increase size to get better understanding.

# Model Building

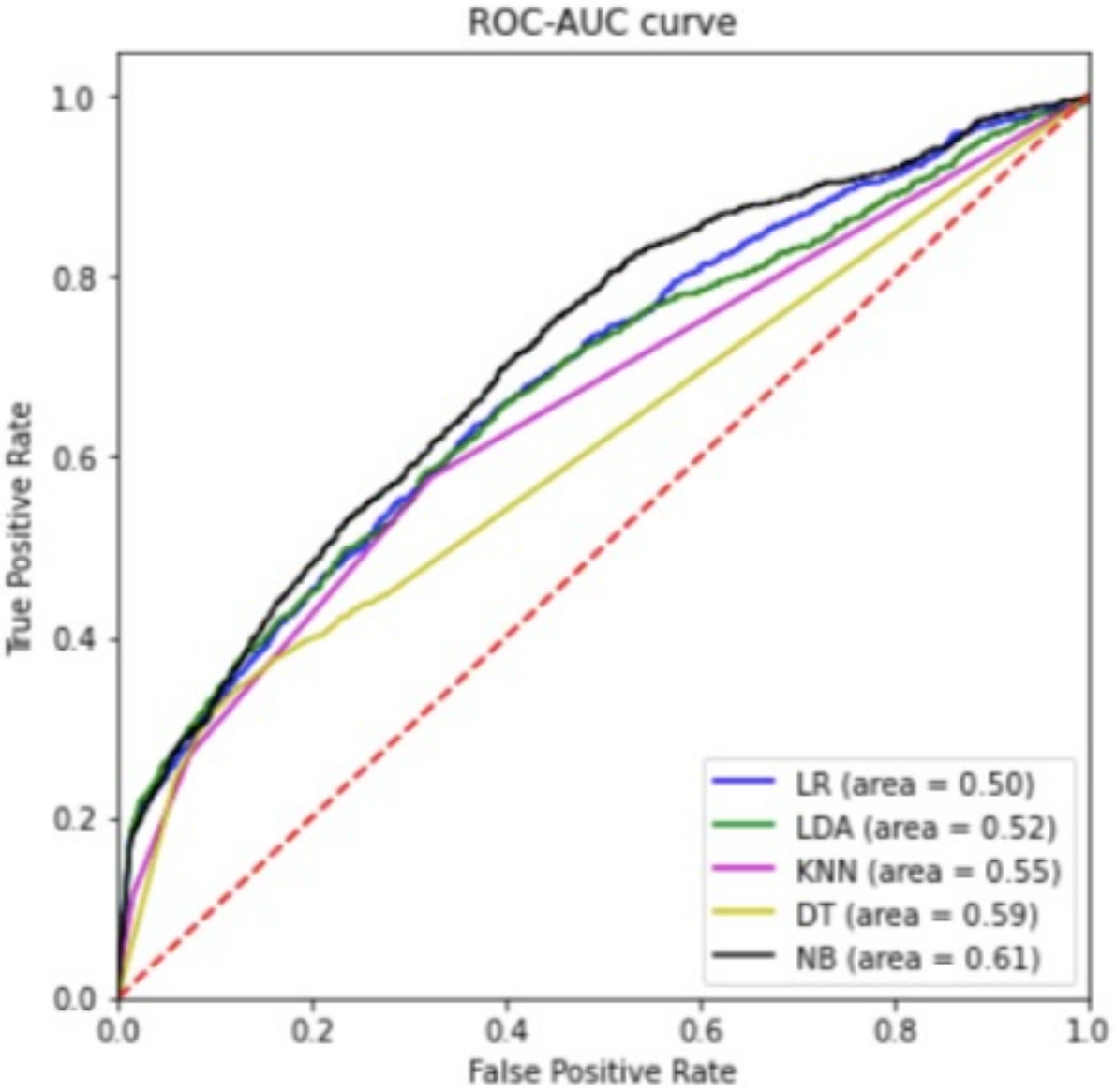
In order to predict the client subscription for a deposit term, we will use a predictive ML model to helps us identify potential customers. We will split our data to train and test with ratio of 0.75 and 0.25 respectively.

We choose to test out the following set of models since we don't know yet what algorithms will do well on this dataset.

The Following algorithm selected include:

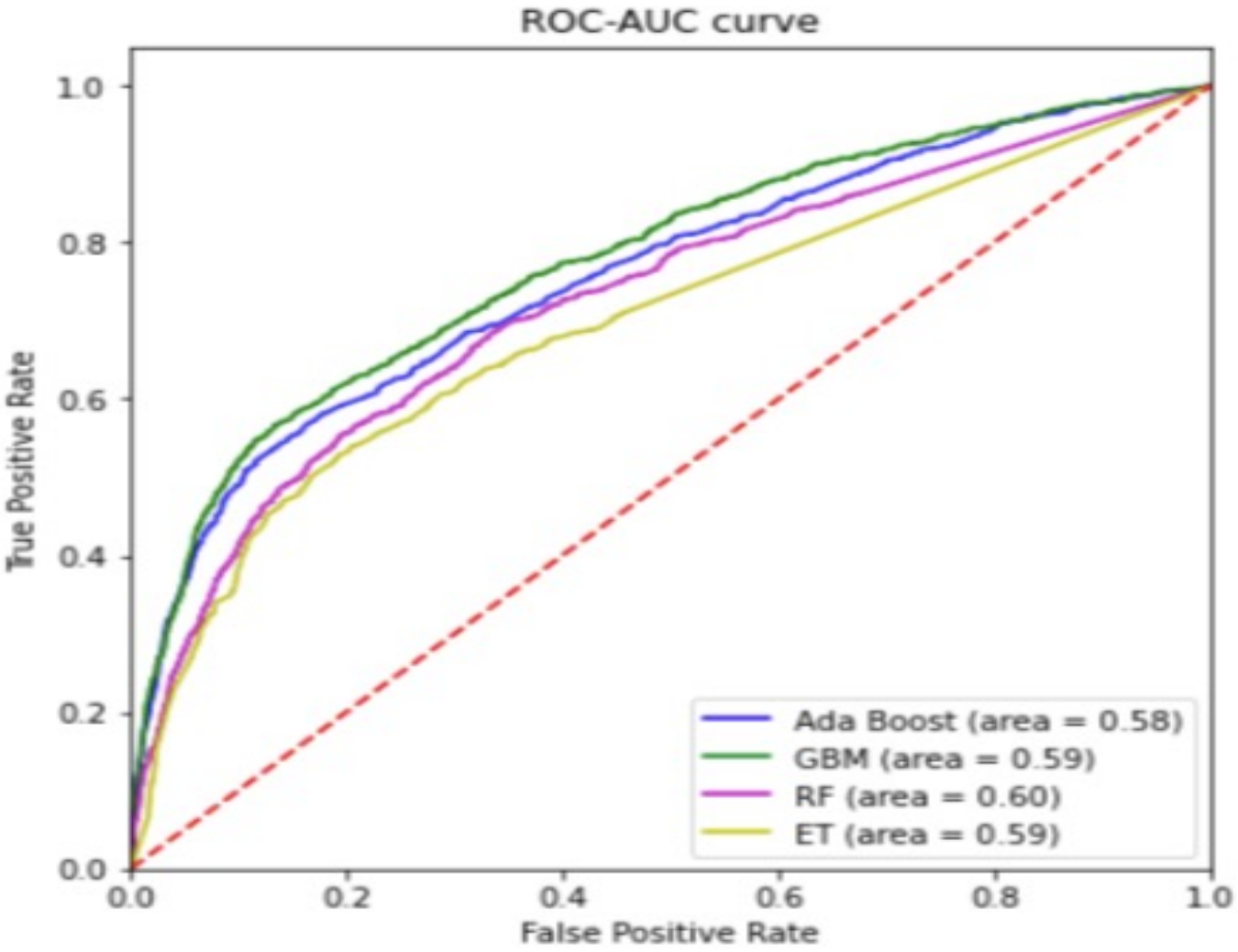
- **Linear Algorithms :**
  Logistic Regression (LR)
  Linear Discriminant Analysis (LDA)
- **Ensemble Methods :**
  Boosting methods: AdaBoost (AB) and Gradient Boosting (GBM)
  Bagging methods: Random Forests (RF) and Extra Trees (ET).
- **Non Linear Algorithms :**
  Classifications and Regression Trees (CART).
  Support Vector Machines (SVM)
  Gaussian Naive Bayes (NB)
  K-nearest Neighbours (KNN)

# Model Results (Linear and Non-Linear)




- Here we can observe that Naive Bayes Classifier(0.61) is giving us the highest ROC_AUC score

# Model Results (Ensemble)




- Here we observe that random forest method is returning highest ROC_AUC score and all four models shows almost same ROC_AUC score.

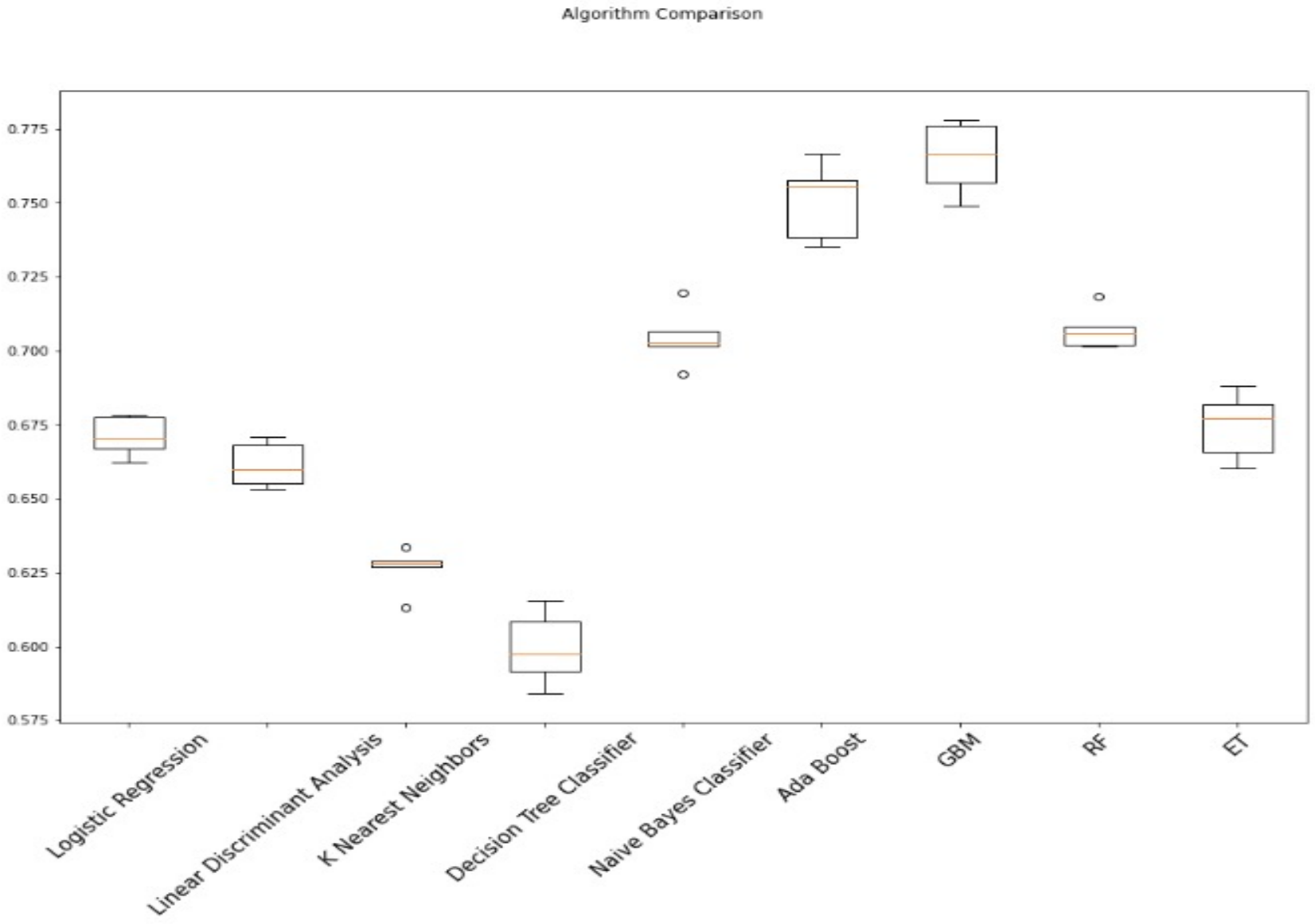
# Cross Validation

- Cross-validation is a technique for evaluating a machine learning model and testing its performance. CV is commonly used in applied ML tasks. It helps to compare and select an appropriate model for the specific predictive modelling problem.

- K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data. K refers to the number of groups the data sample is split into. For example, if you see that the k-value is 5, we can call this a 5-fold cross-validation.

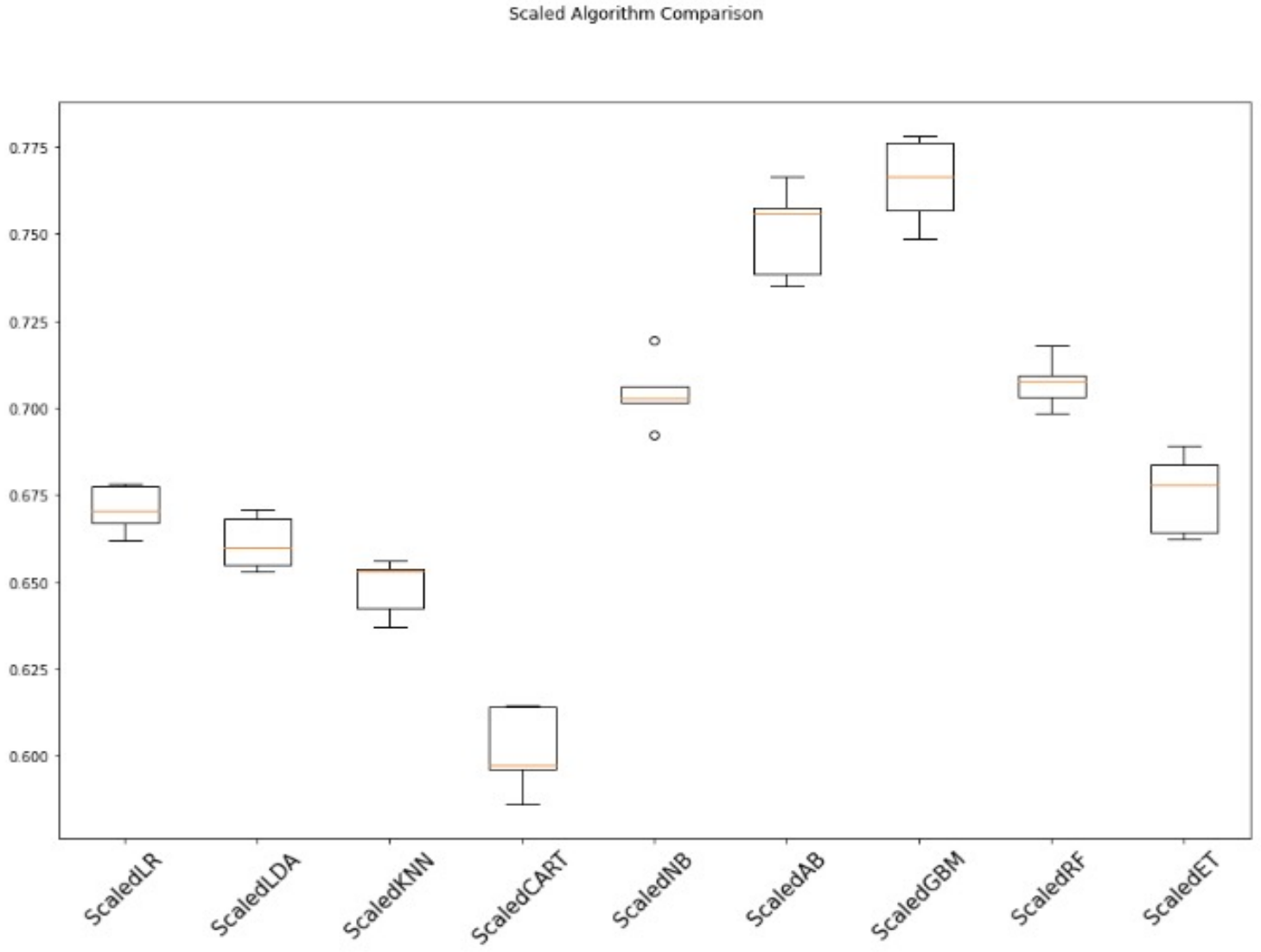
- Area under ROC Curve (or AUC for short) is a performance metric for binary classification problems. The AUC represents a model's ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model that is as good as random.

# Cross Validation : (Mean ROC_AUC score and Standard Deviations without standardising data )




```
LR: 0.671185 (0.006165)
LDA: 0.661438 (0.007034)
KNN: 0.626143 (0.006812)
DT: 0.599433 (0.011311)
NB: 0.704496 (0.008885)
ADB: 0.750668 (0.011948)
GBM: 0.765256 (0.011226)
RF: 0.707093 (0.006232)
ET: 0.674725 (0.010318)
```

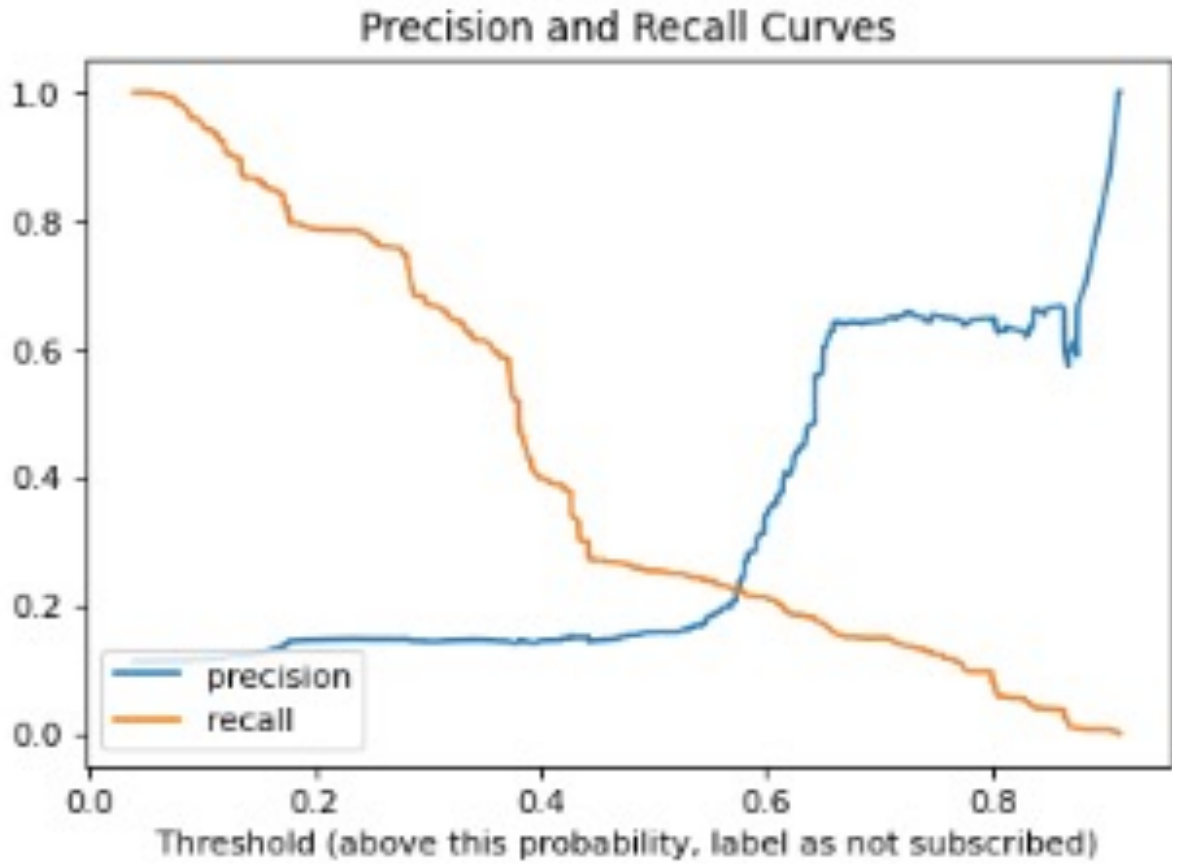
# Cross Validation : (Mean ROC_AUC score and Standard Deviations after standardising data )

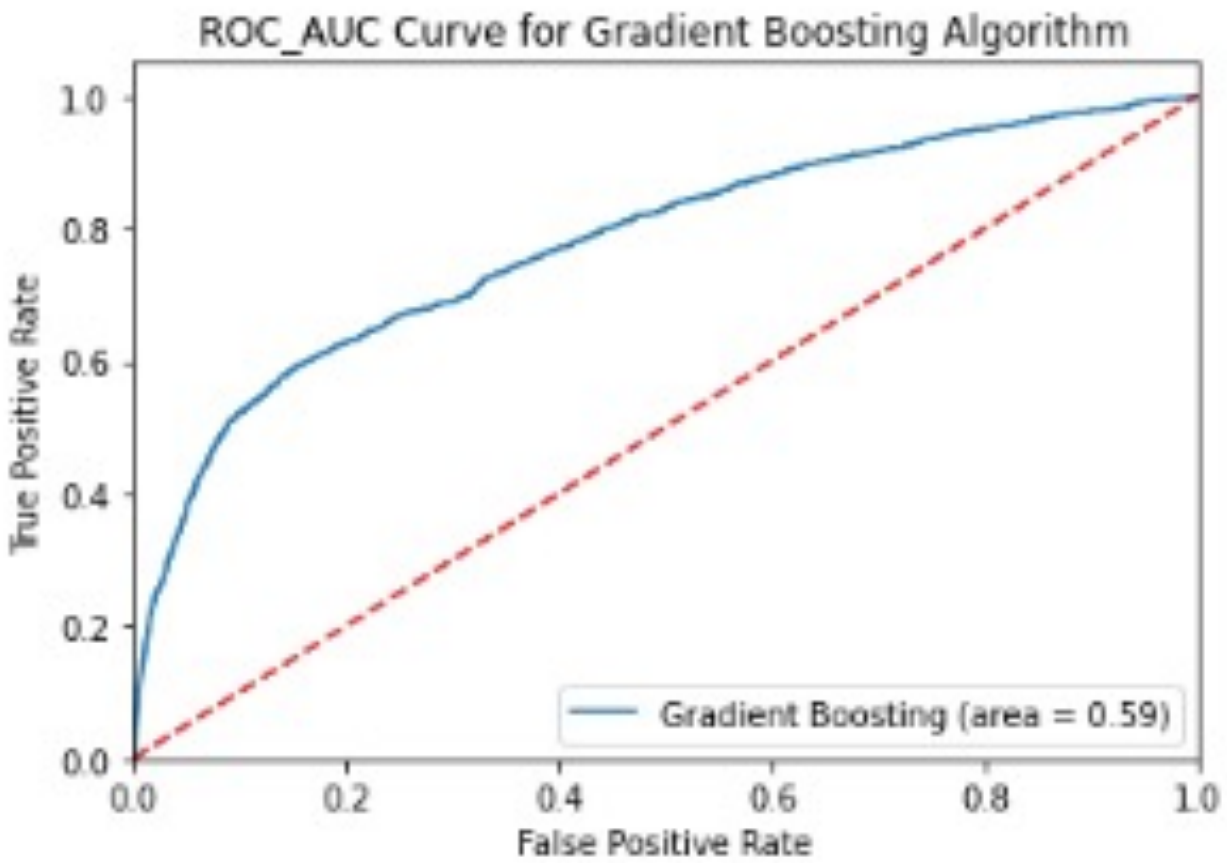



- We can also see that the standardizati on of the data has lifted the skill of KN N but still the GBM model is the most accurate algorithm tested so far. Stand ardising the dataset have also reduced the variance in the roc_auc score.

- It was observed that the best configuration was n estimators=150 resulting in a mean squared error of 0.766885.

# Model Results : (Precision, recall and ROC_AUC Curve for gradient Boosting model :)



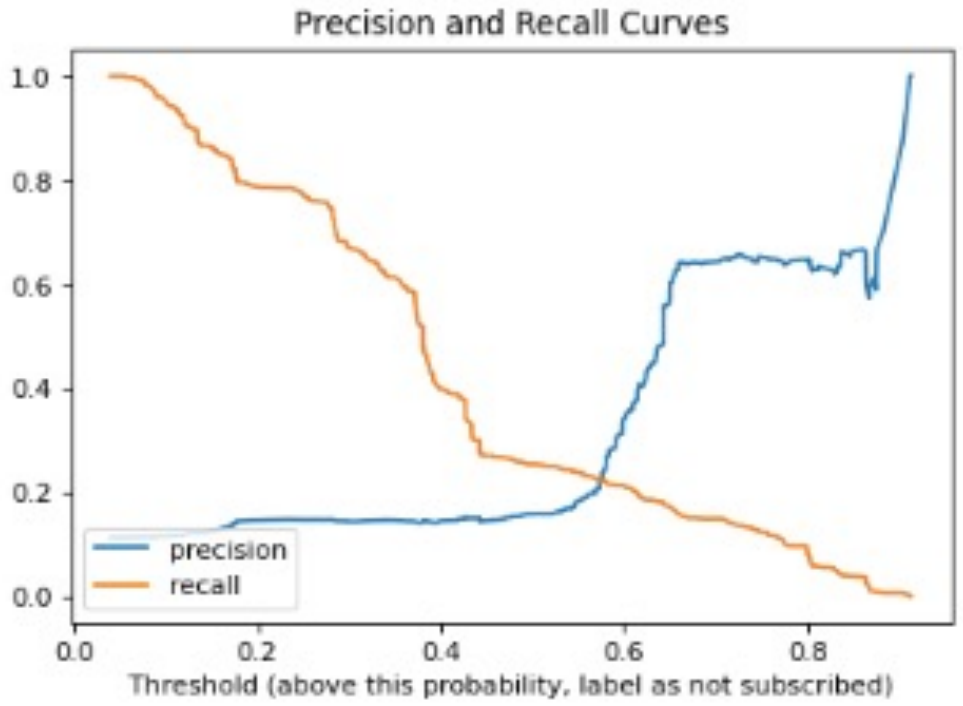

- **Final Result**: From all the above models GBM performed better Scored well on training and test data.

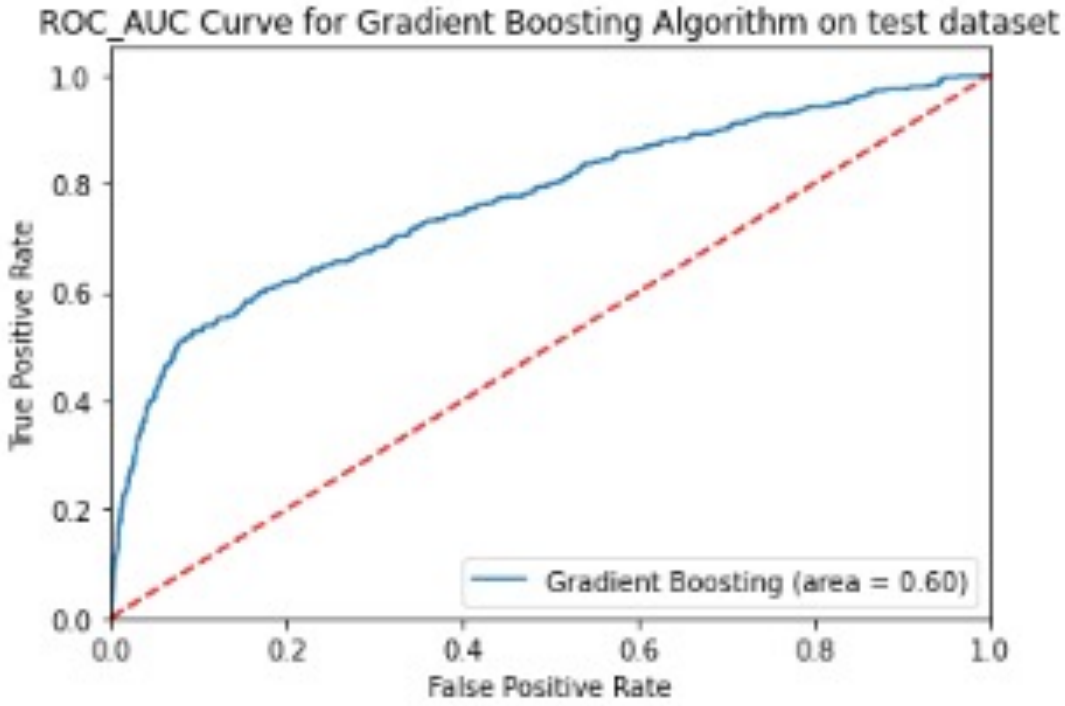
# Model Results : (Precision, recall and ROC_AUC Curve for gradient Boosting model :)

- **Mean Squared Error :** 0.10
- **ROC_AUC Score :** 0.58
- **Accuracy :** 0.896
- **Confusion Matrix :** [[9005 127]
  [ 941 221]]

# Model Results : (Precision, recall and ROC_AUC Curve for gradient Boosting model on test data:)

- **Mean Squared Error :** 0.09
- **ROC_AUC Score :** 0.60
- **Accuracy :** 0.9016



# Model Recommendation :

We can see that both boosting techniques provide strong accuracy scores in the high 70s (%). The GBM model is the best model compared to the other ones. Therefore we will consider that model for production.

# GITHUB LINK :

## GITHUB LINK :

https://github.com/AbhimanyuGangani/Week_7_Bank_Marketing/tree/main/final_week_bank_marketing

# Thank You