أكاديمية سدايا
SDAIA Academy

Comparison of classifier Algorithms on bank marketing Dataset

Presented by :
Arwa AlBassam & Nouf Alsaeed

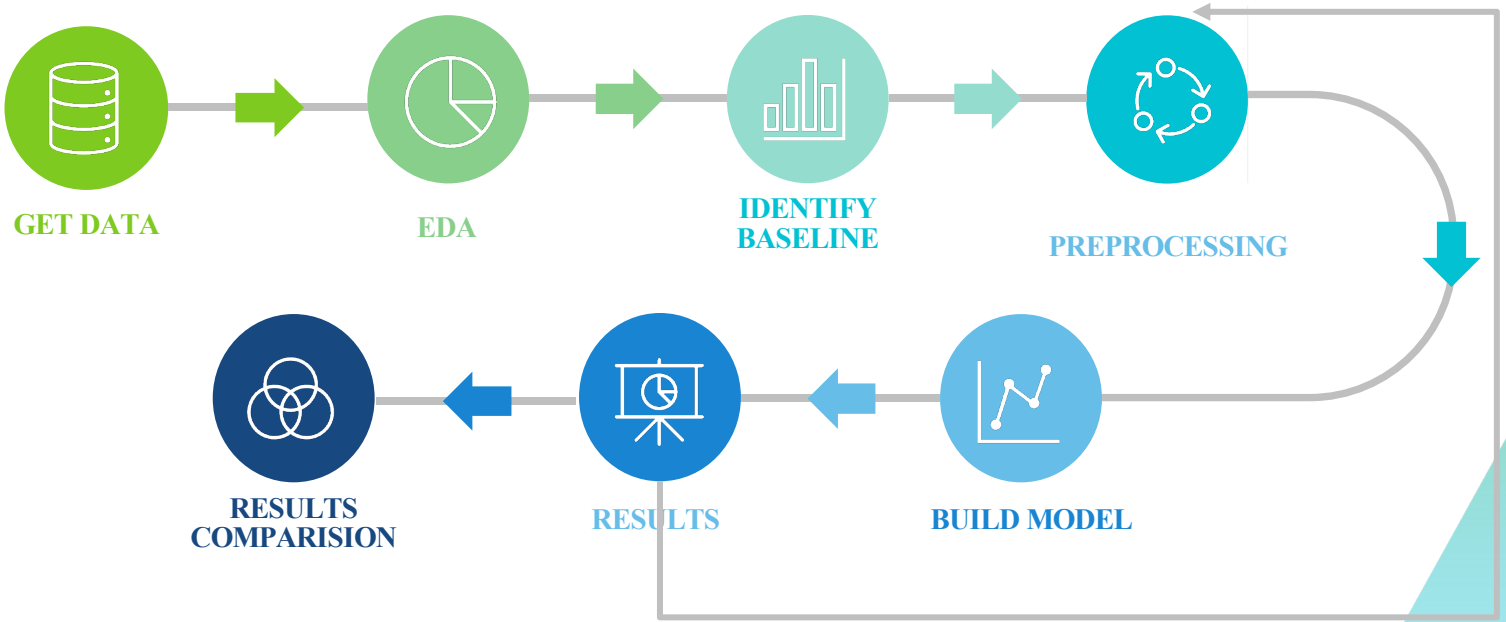
10/7/21

OUTLINE

Introduction

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Workflow


GET DATA
EDA
IDENTIFY BASELINE
PREPROCESSING
BUILD MODEL
RESULTS
RESULTS COMPARISION

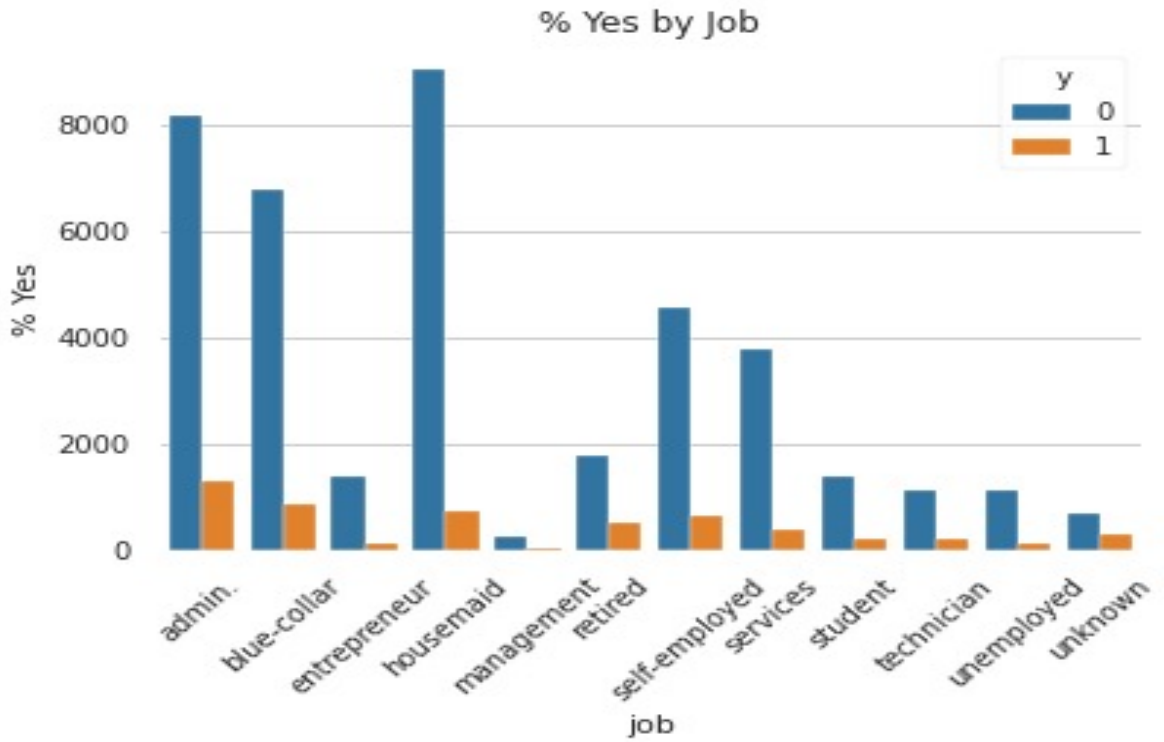
Dataset Description

From UCI
45211 Rows X 17 Columns

Feature	Description
Age	Age of clints
Job	Type of job
Marital	Marital status
Education	(Categorical: "unknown", "secondary", "primary", "tertiary")
Balance	Average yearly balance, in euros (numeric)
Housing	Has housing loan? (Binary: "yes", “no")
Loan	Has personal loan? (Binary: "yes", “no")
Poutcome	Outcome of the previous marketing campaign
Y	Has the client subscribed to a term deposit? (Binary: "yes","no")

EDA

Visualizations



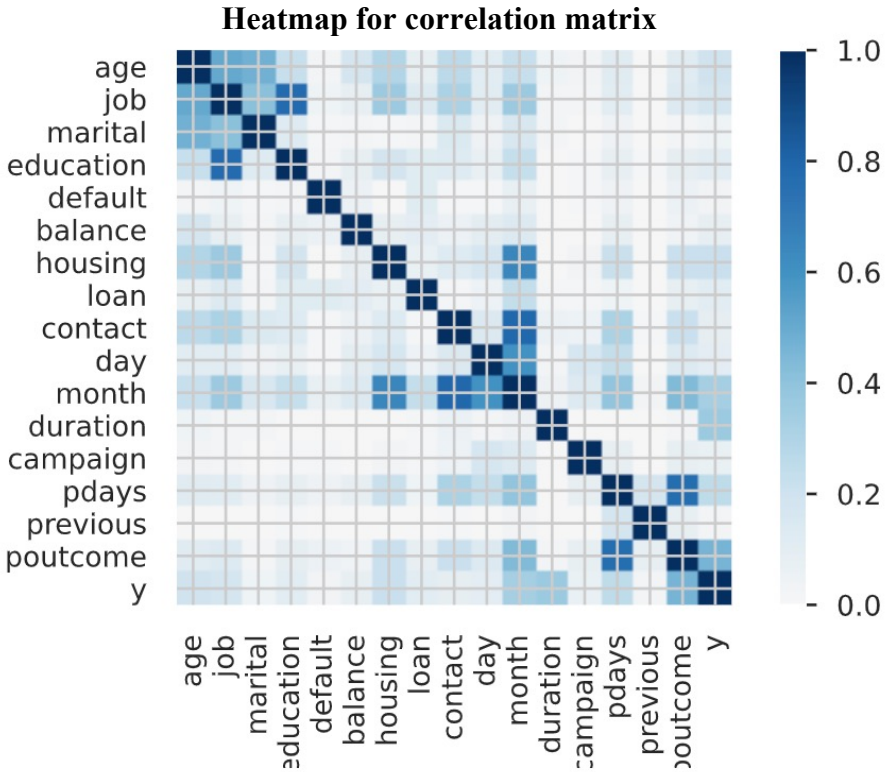
EDA (Cont.)

Dataset statistics

Number of variables	17
Number of observations	45211
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	29.2 MiB
Average record size in memory	677.2 B

EDA (Cont.)

Visualizations



Base Model (Linear Regression)

ACCURACY

RECALL

Data Preprocessing

- Creating Dummy Variables.
- Balance Data.
- Feature Scaling.

Data Preprocessing (cont.)

Balancing the Model

y
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	353.3 KiB

Toggle details

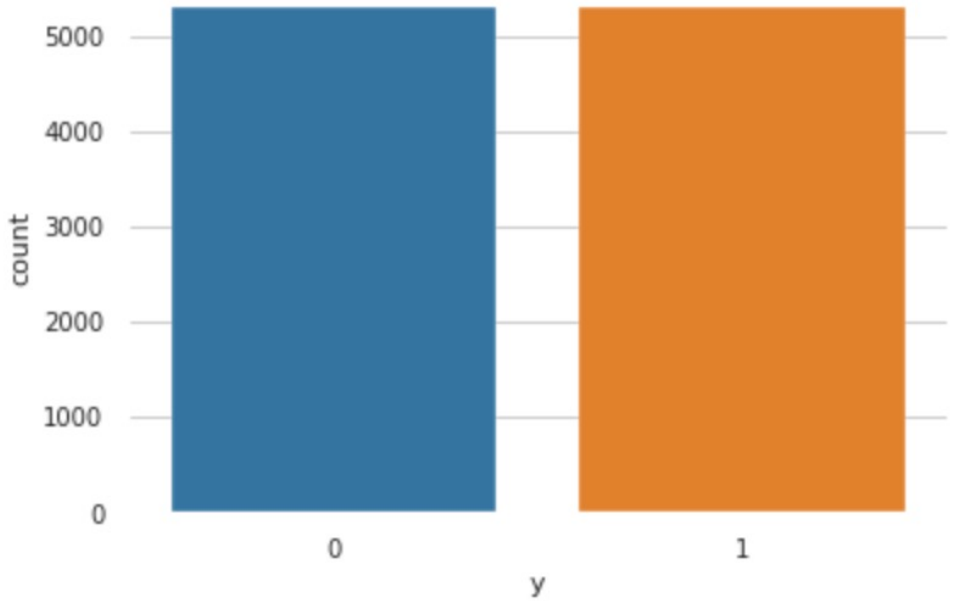
Frequency Table

Value	Count	Frequency (%)
no	39922	88.3%
yes	5289	11.7%

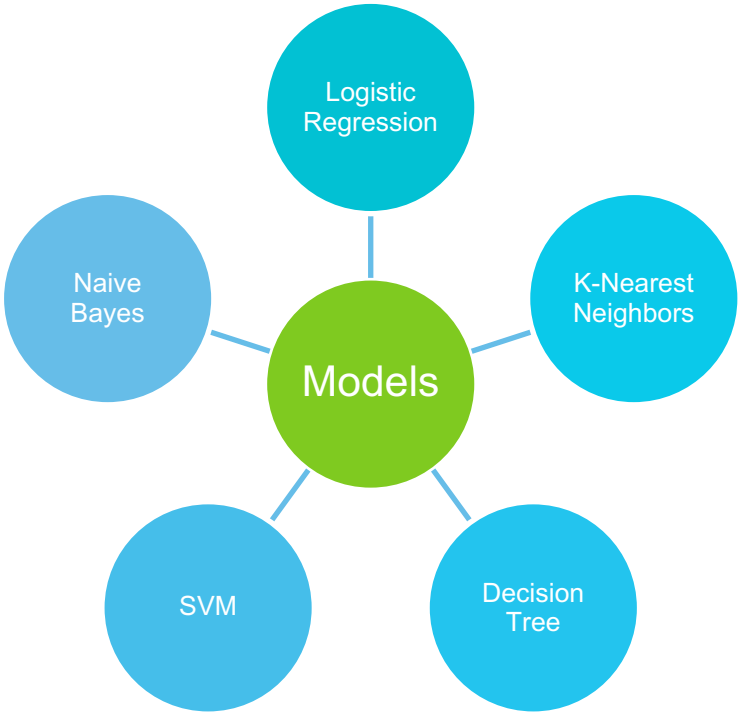
Data Preprocessing (cont.)

Balancing the Model

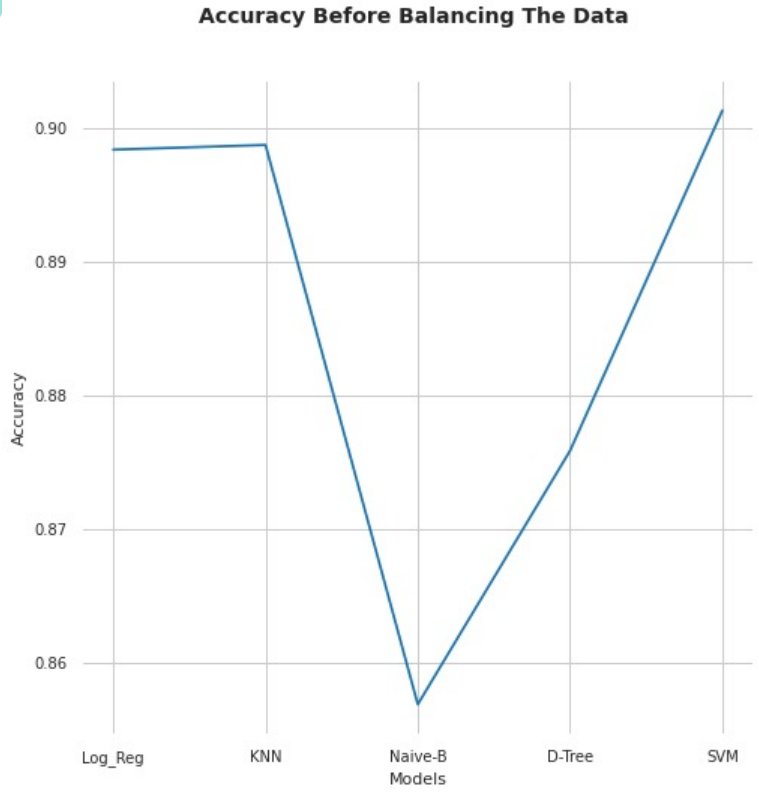
```
1    5289
0    5289
Name: y, dtype: int64
```

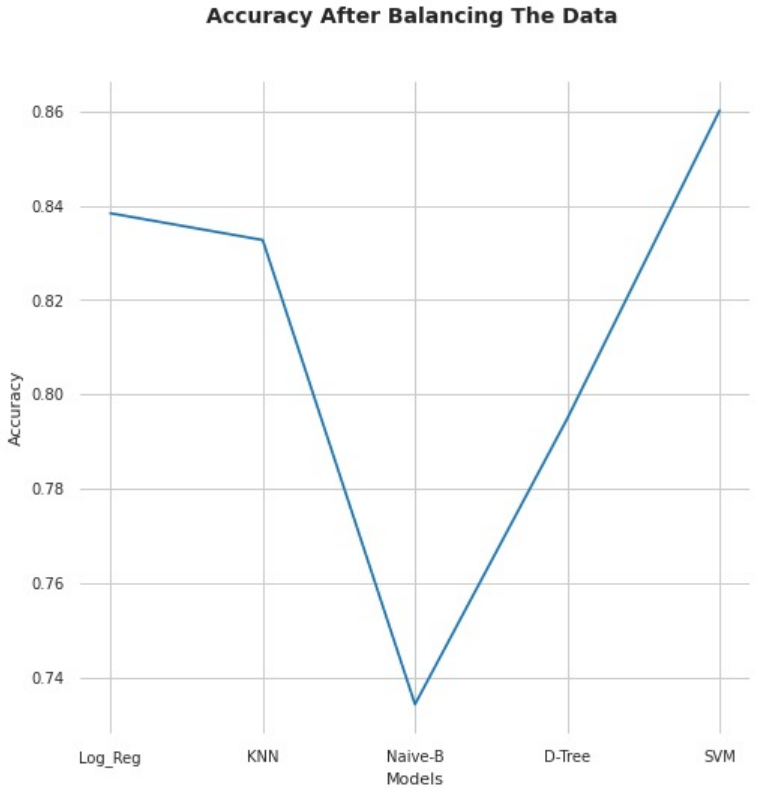


Modelling

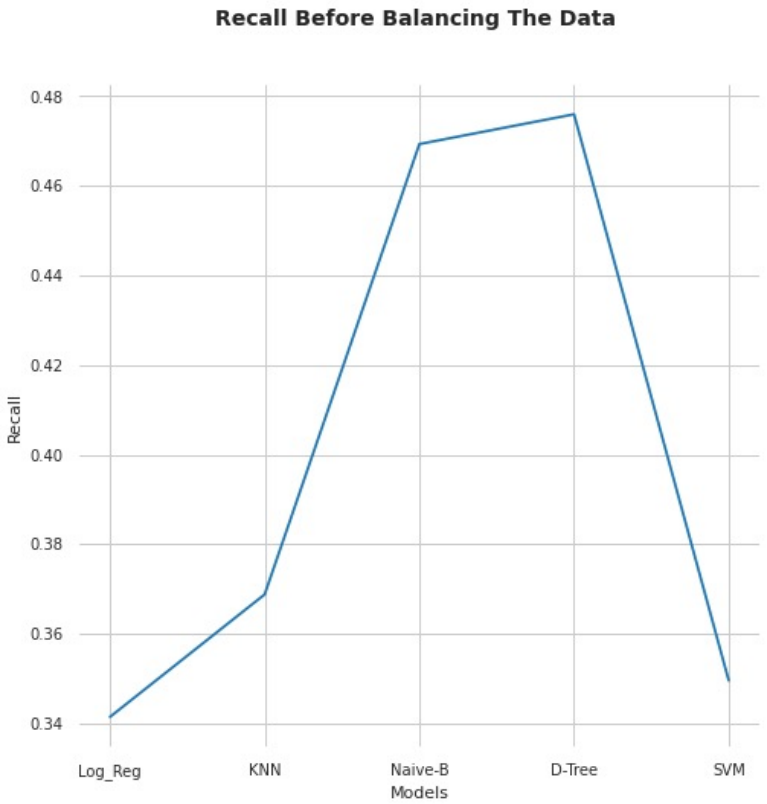

Logistic Regression
Naive Bayes
K-Nearest Neighbors
Models
SVM
Decision Tree

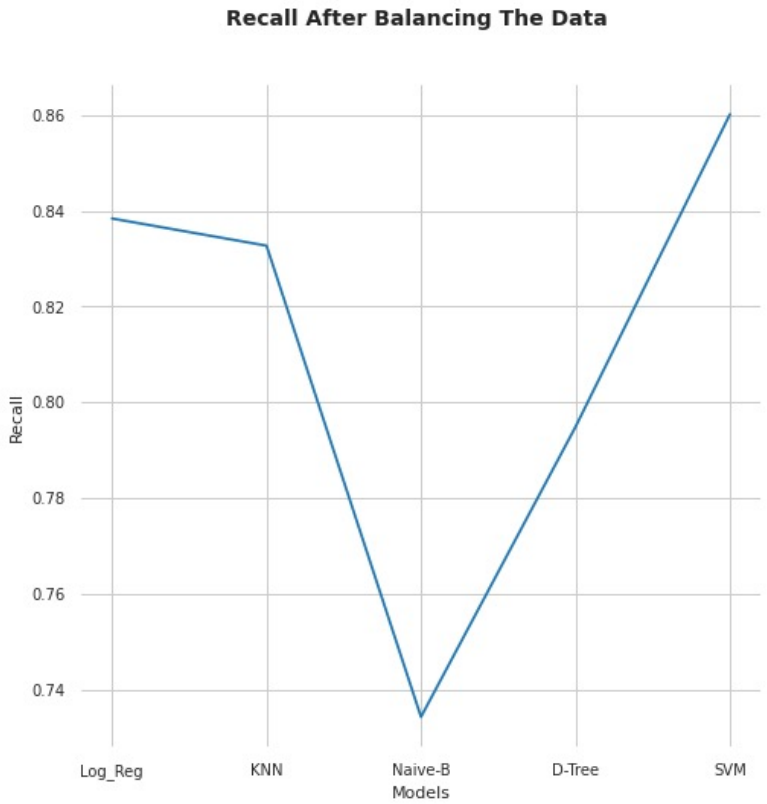
Models Evaluation



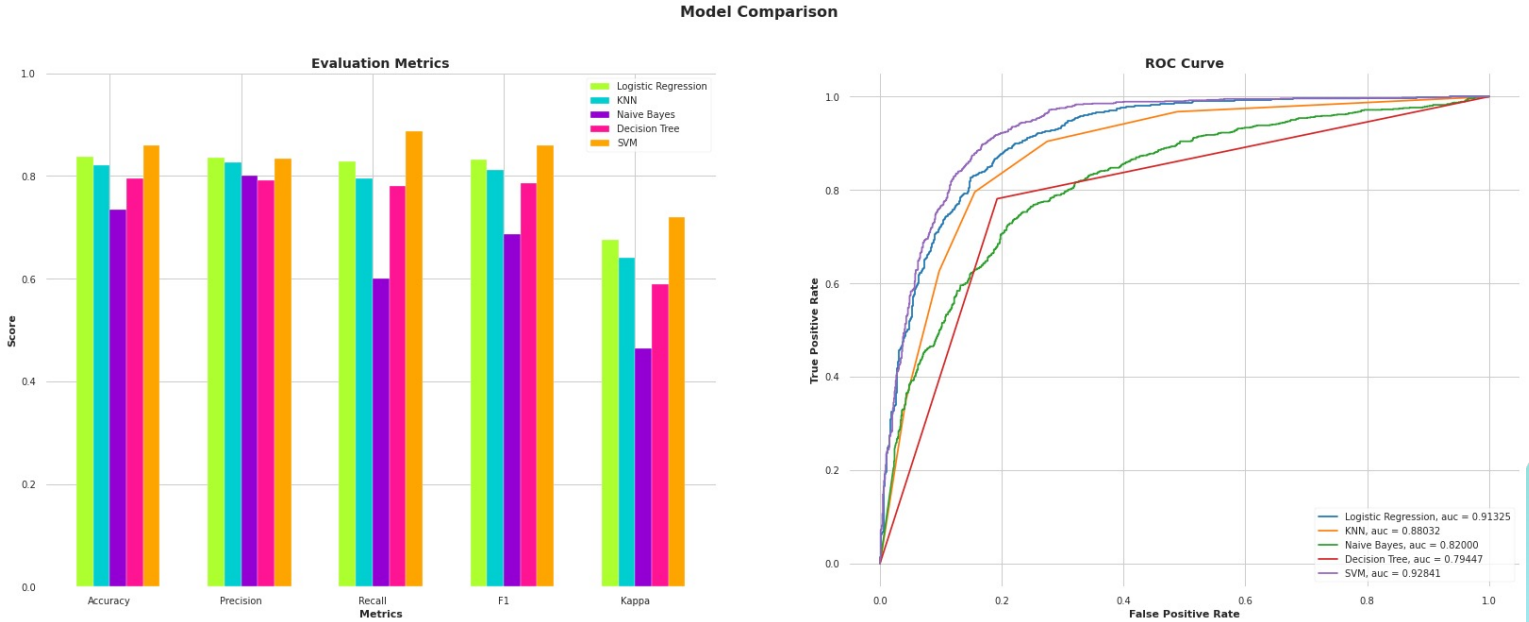


Models Evaluation

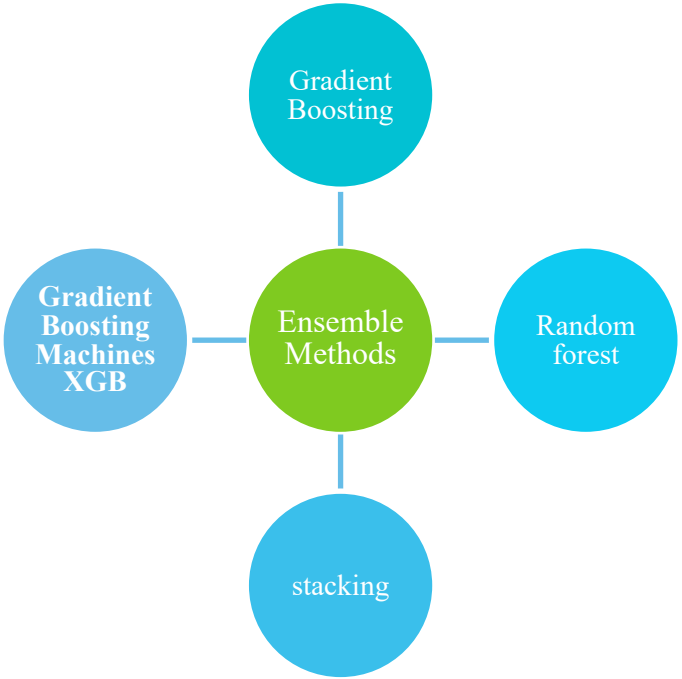




Models Evaluation



Ensemble Methods



Stacking Ensemble Family

Average Voting

01

ACCURACY: 86.48 %

Max Voting

02

ACCURACY: 86.24%

Weighted Average Ensemble

03

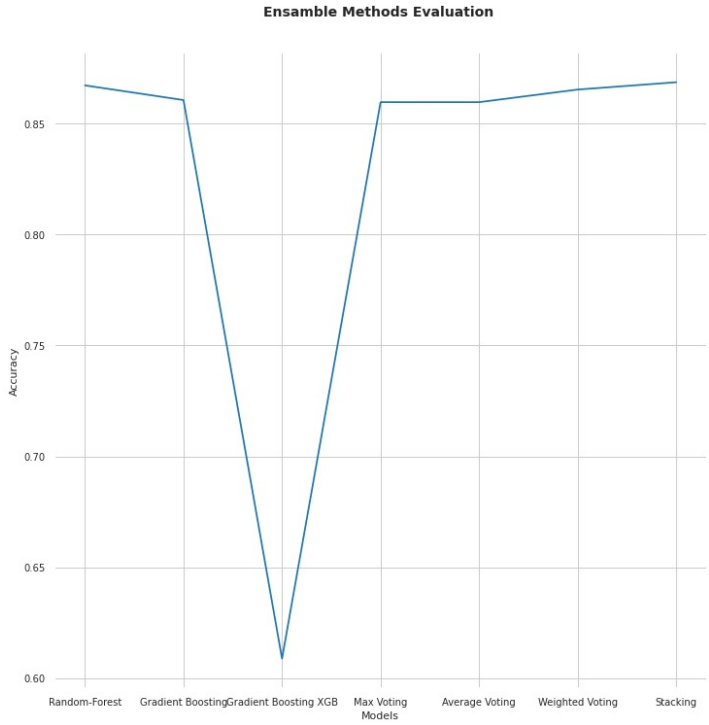
ACCURACY: 86.06%

Stacking

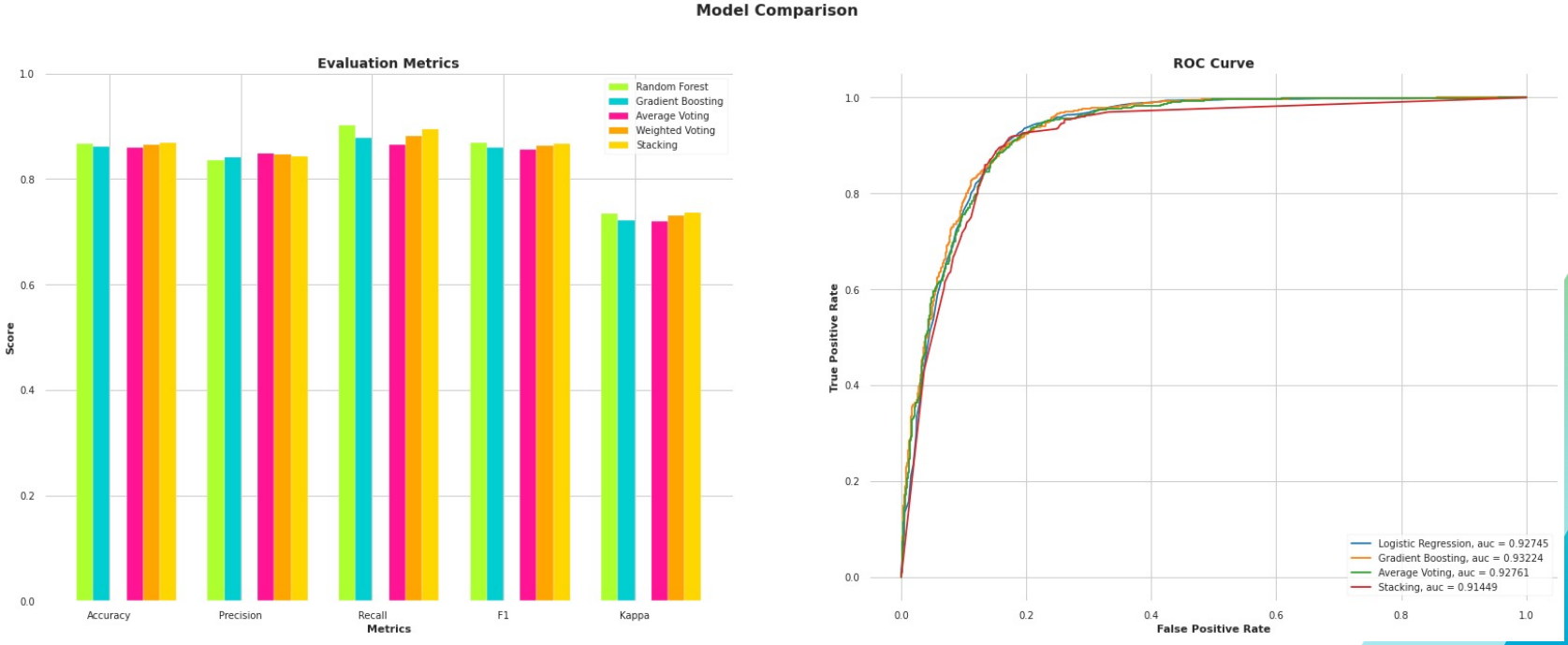
04

ACCURACY: 86.7 **%**

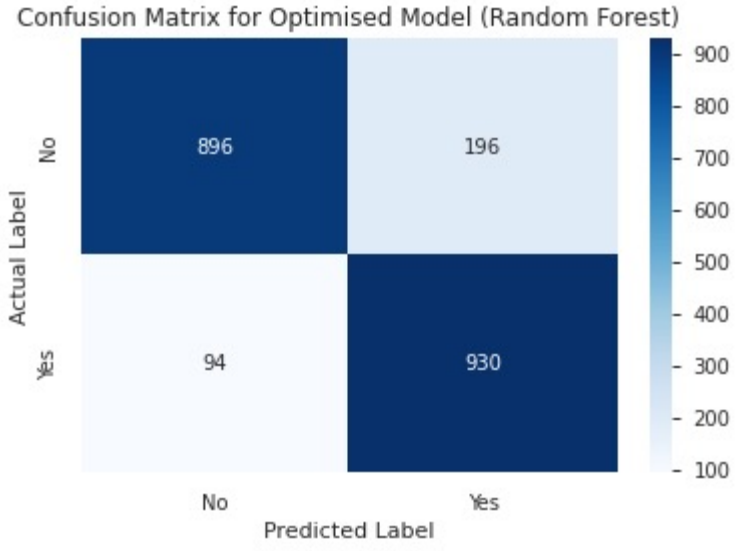
Ensemble Methods Evaluation



Ensemble Methods Evaluation

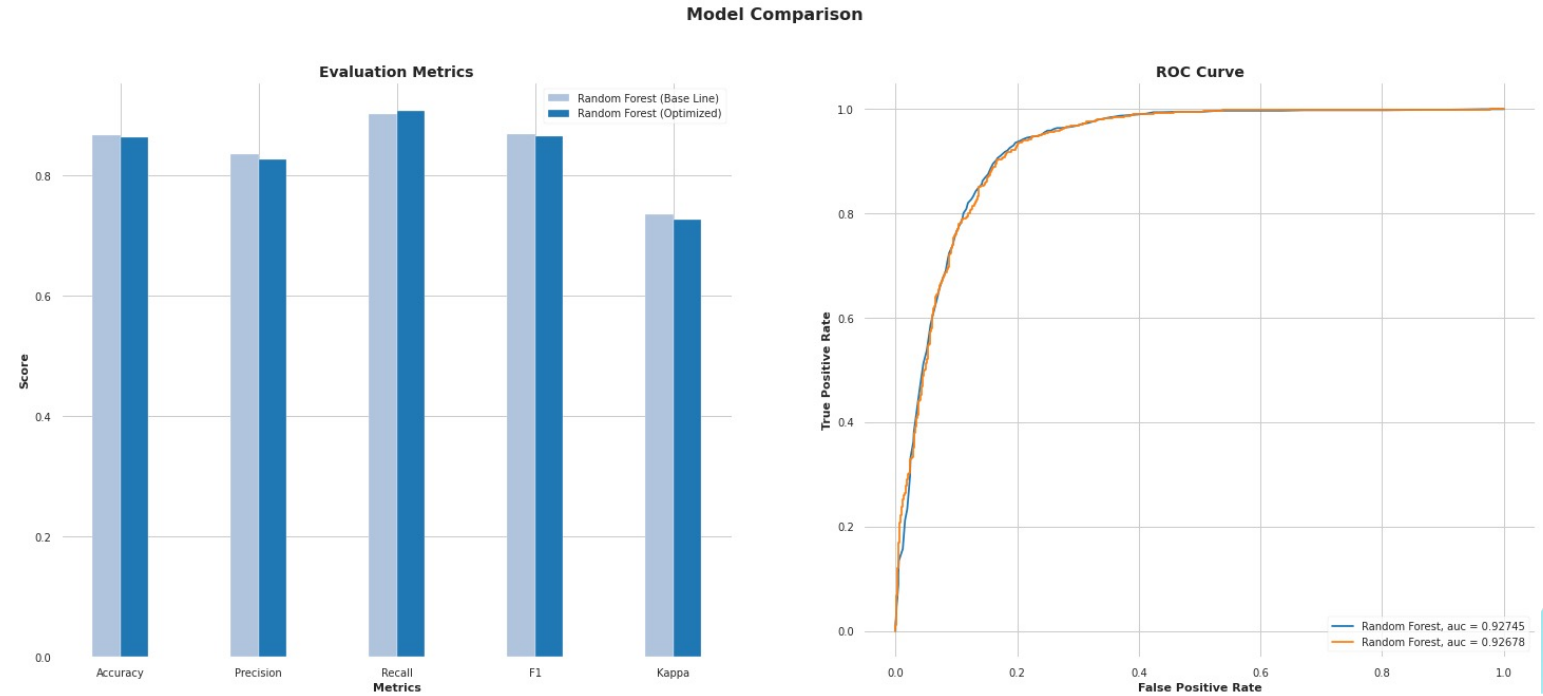


Best model for our project (Random Forest)



ACCURACY: 86.8%

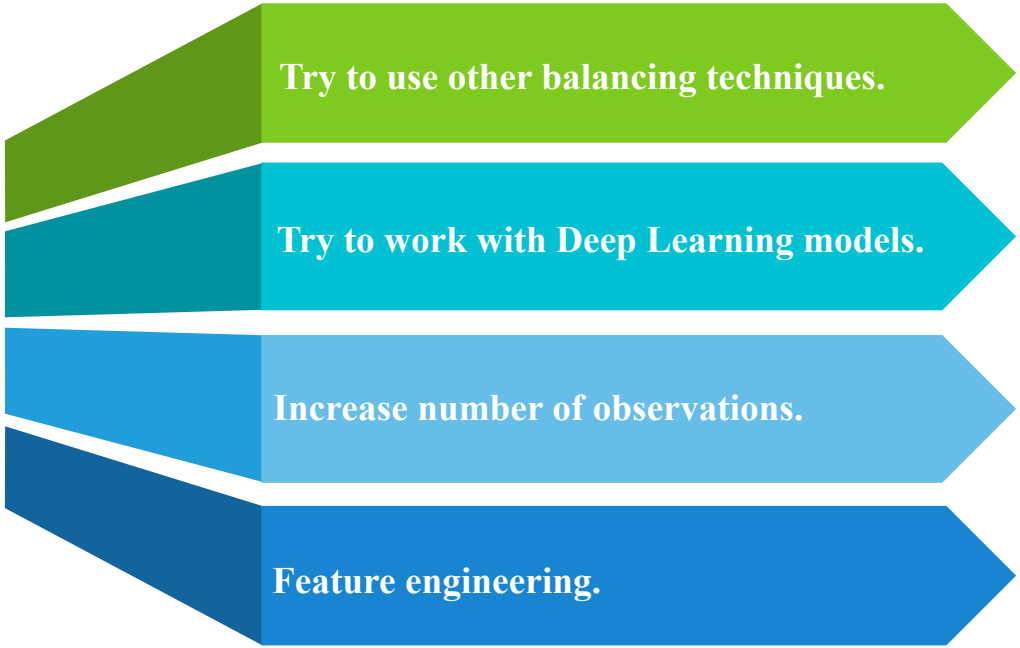
Final results (Model Optimization)



Conclusion

The result is not that much different after optimizing the model using GridSearchCV which can mean that we hit our limit with this model.

Future Work


Try to use other balancing techniques.
Try to work with Deep Learning models.
Increase number of observations.
Feature engineering.

Thank you!