Capstone Project

Bank Marketing Effectiveness Prediction

Content

- Problem Statement
- Data Summary
- Data Analysis
- Challenges
- Conclusions
- Q&A

Problem Statement

- The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Summary

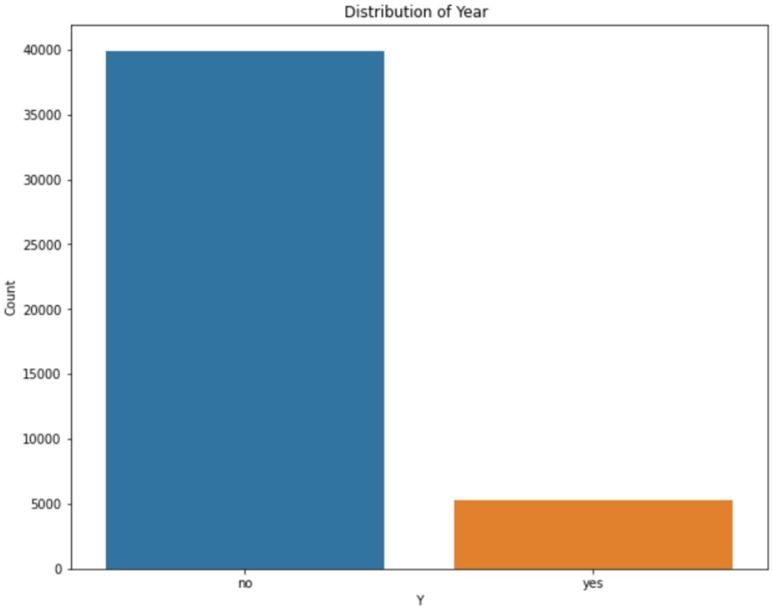
- job : type of job (categorical: 'admin.' , 'blue-collar' , 'entrepreneur' , 'housemaid' , 'management' , 'retired' , 'self employed' , 'services' , 'student' , 'technician' , 'unemployed' , 'unknown')
- marital : marital status (categorical: 'divorced' , 'married' , 'single')
- education : (categorical: 'primary' , 'secondary' , 'tertiary' , 'unknown')
- default: has credit in default? (categorical: 'no' , 'yes')
- housing: has housing loan? (categorical: 'no' , 'yes')
- loan: has personal loan? (categorical: 'no' , 'yes')
- contact: contact communication type (categorical: 'cellular' , 'telephone' , 'unknown')
- month: last contact month of year (categorical: 'jan' , 'feb' , 'mar', ..., 'nov', 'dec')
- poutcome: outcome of the previous marketing campaign (categorical: 'failure' , 'success')

Contd ...

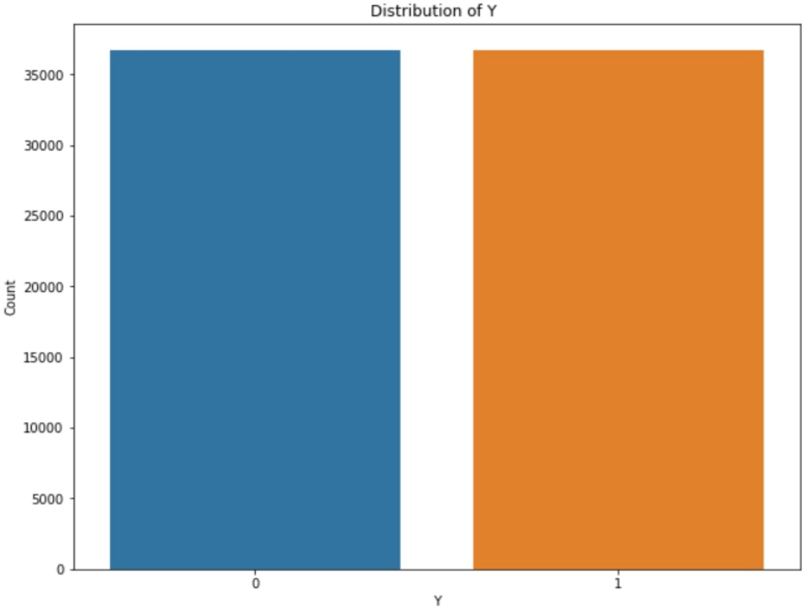
- day : last contact day of year(categorical: 1 , 2 , . . . , 31)
- age : (numeric)
- duration: last contact duration, in seconds (numeric)
- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- pdays: number of days that passed by after the client was last contacted from a previous campaign
- previous: number of contacts performed before this campaign and for this client (numeric)
- y - has the client subscribed a term deposit? (binary: 'yes','no')

Count plot of Dependent Feature(Y)

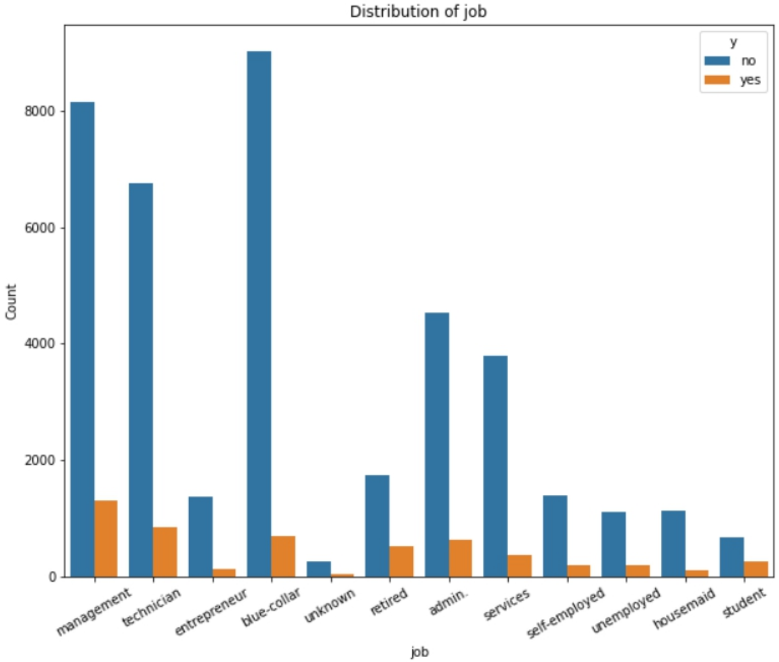
Before



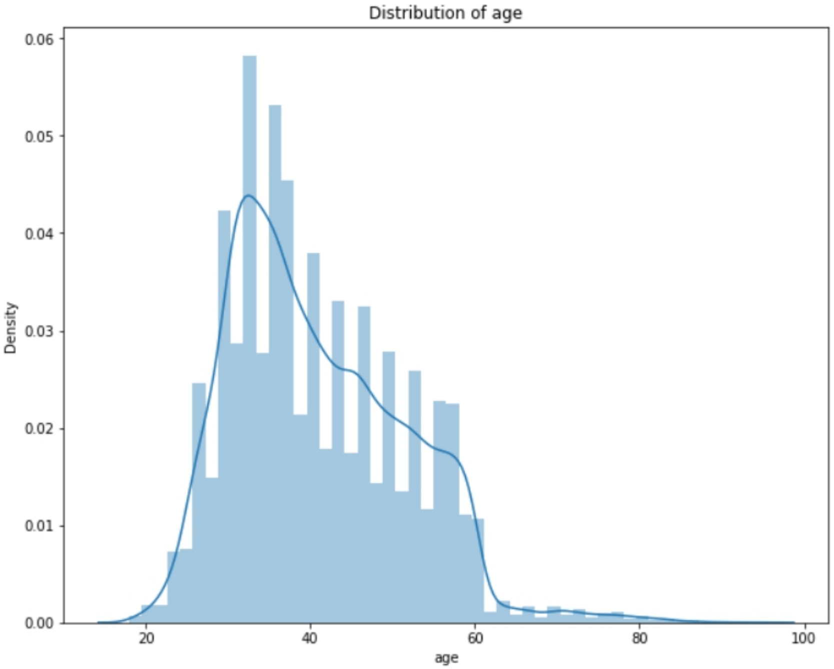
After



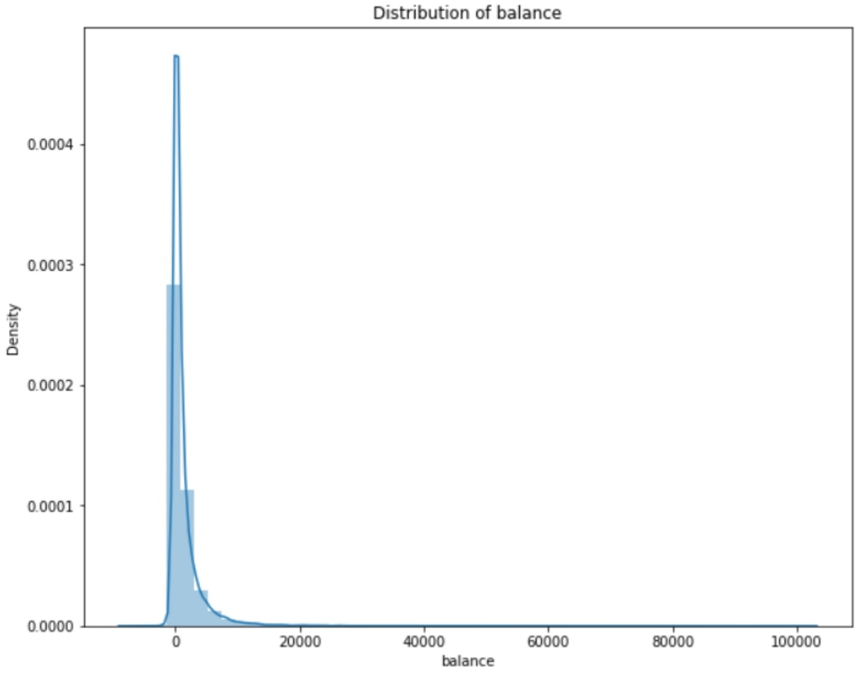
Count plot of Job



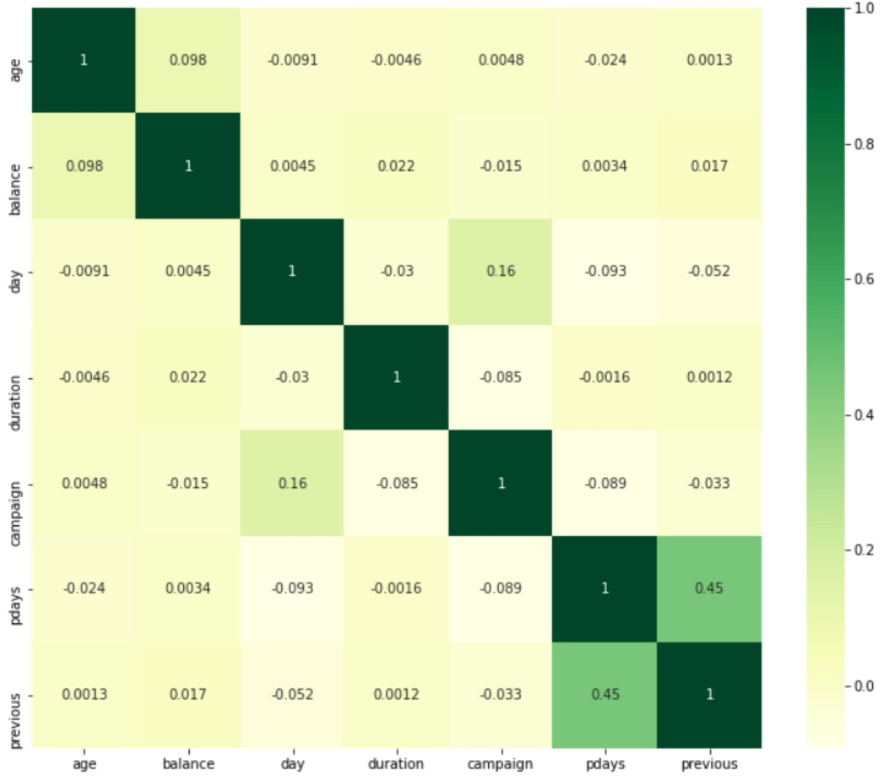
Distribution plot of Age



Distribution plot of Balance



Correlation



	age	balance	day	duration	campaign	pdays	previous
age	1	0.098	-0.0091	-0.0046	0.0048	-0.024	0.0013
balance	0.098	1	0.0045	0.022	-0.015	0.0034	0.017
day	-0.0091	0.0045	1	-0.03	0.16	-0.093	-0.052
duration	-0.0046	0.022	-0.03	1	-0.085	-0.0016	0.0012
campaign	0.0048	-0.015	0.16	-0.085	1	-0.089	-0.033
pdays	-0.024	0.0034	-0.093	-0.0016	-0.089	1	0.45
previous	0.0013	0.017	-0.052	0.0012	-0.033	0.45	1

Logistic Regression

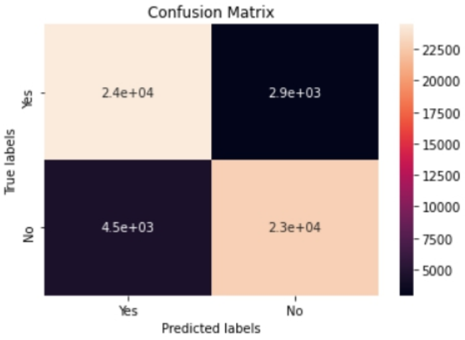
Best Parameters

C : 0.1

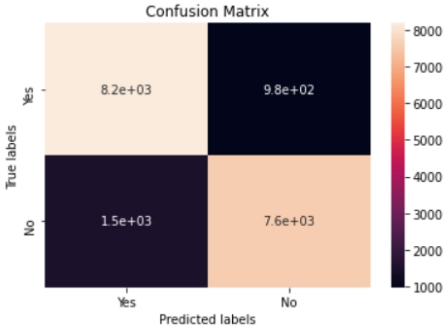
ROC-AUC score

Train Data	0.93
Test Data	0.93

Confusion matrix of Train Data



Confusion matrix of Test Data



K-Nearest Neighbors

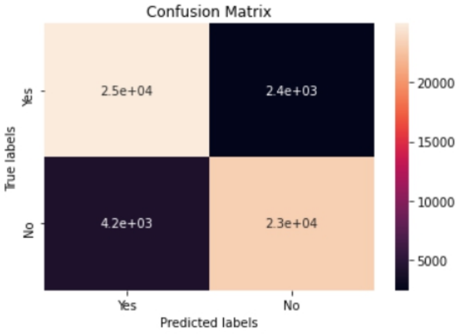
Best Parameters

n_neighbors : 27

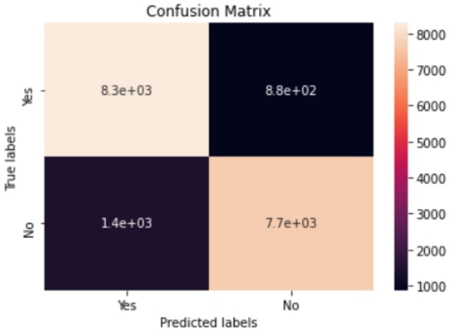
ROC-AUC score

Train Data	0.95
Test Data	0.93

Confusion matrix of Train Data



Confusion matrix of Test Data



Decision Tree

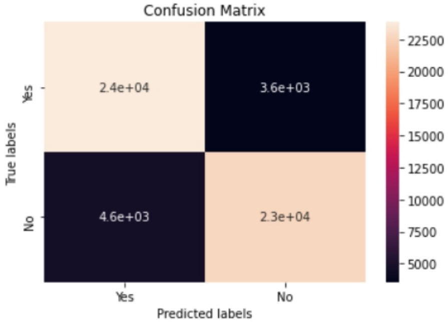
Best Parameters

max_depth : 10
min_samples_leaf : 10
min_samples_split : 20

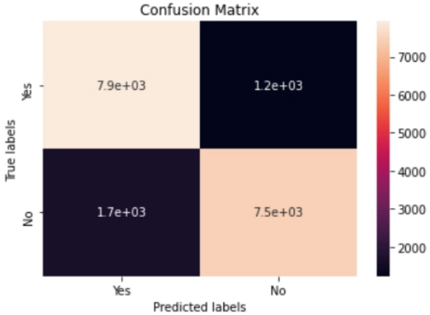
ROC-AUC score

Train Data	0.95
Test Data	0.93

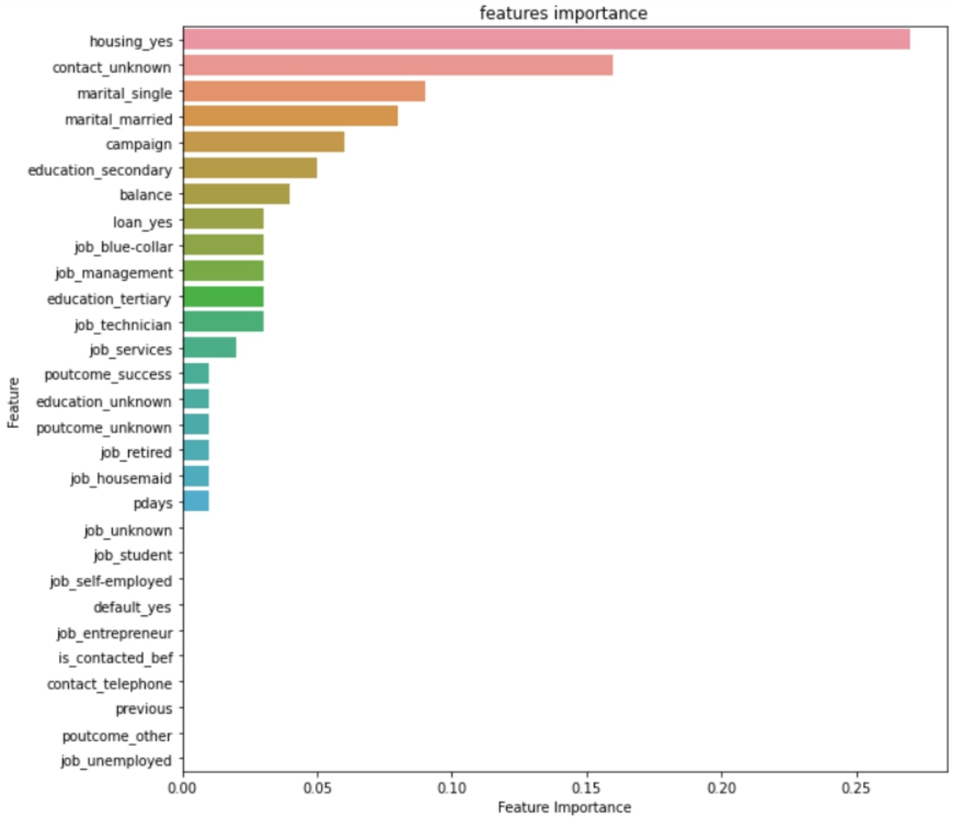
Confusion matrix of Train Data



Confusion matrix of Test Data



Decision Tree Feature Importance



RandomForest

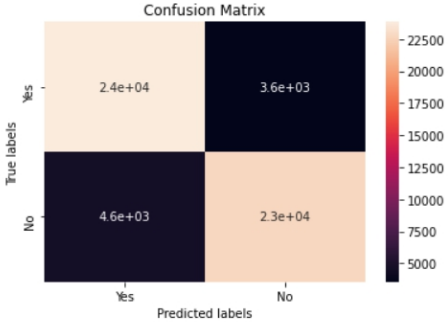
Best Parameters

max_depth : 10
min_samples_leaf : 10
min_samples_split : 20

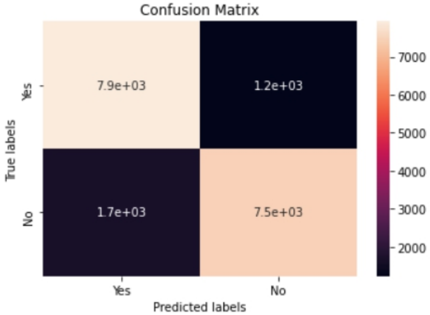
ROC-AUC score

Train Data	0.92
Test Data	0.92

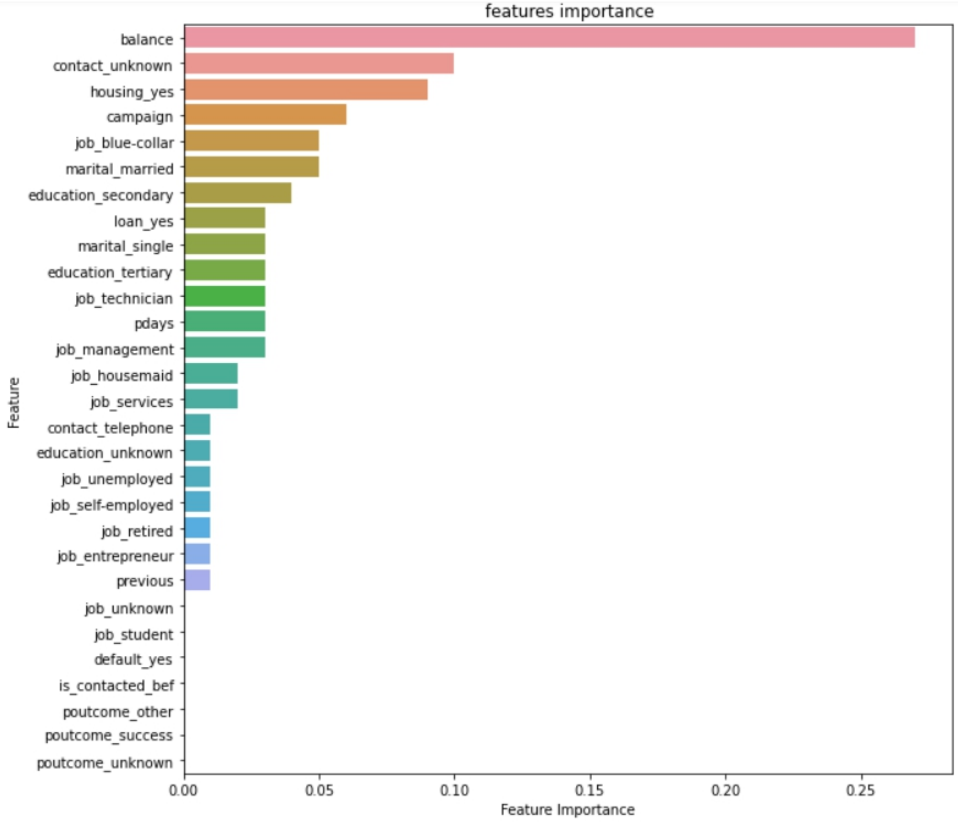
Confusion matrix of Train Data



Confusion matrix of Test Data



RandomForest Feature Importance



XGBoost

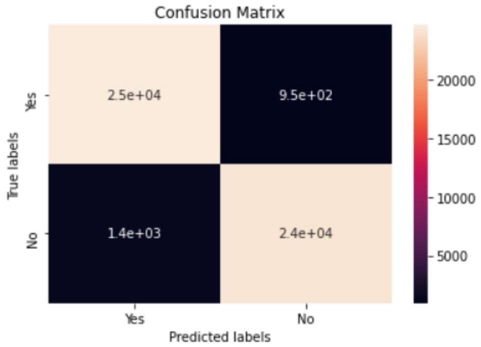
Best Parameters

learning_rate : 0.5
max_depth : 9
n_estimators : 125

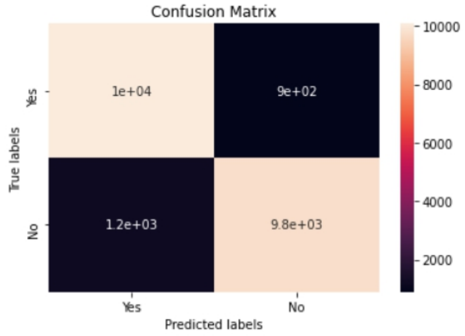
ROC-AUC score

Train Data	0.99
Test Data	0.95

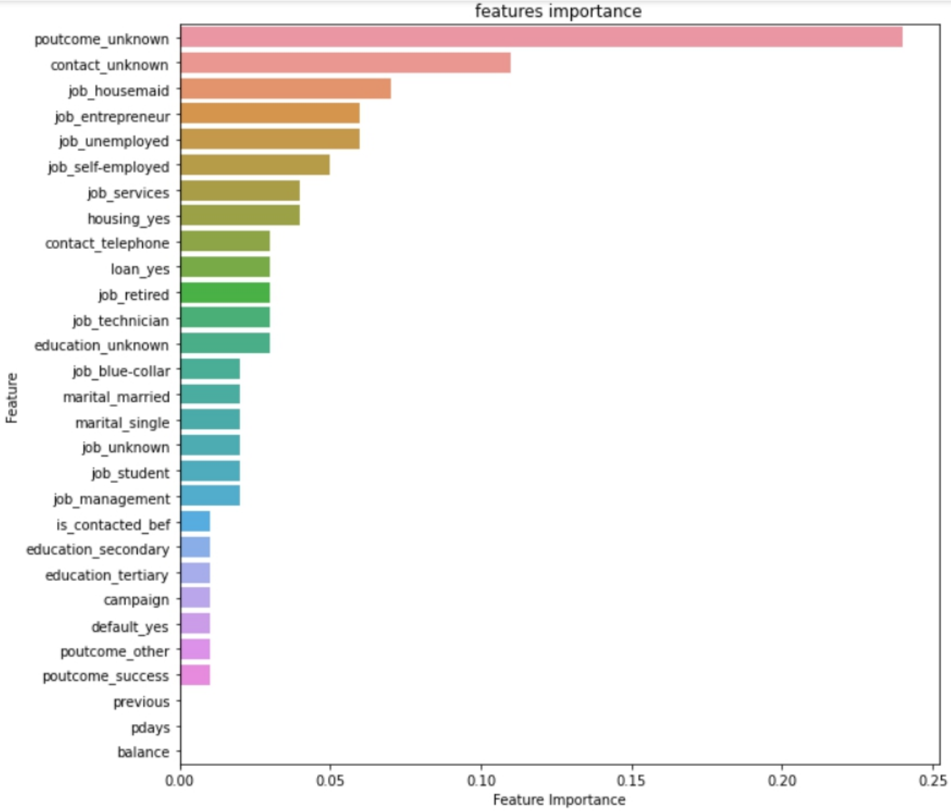
Confusion matrix of Train Data



Confusion matrix of Test Data



XGBoost Feature Importance



Challenges

- Handling imbalanced datset
- Feature Engineering
- Optimising the Model

Conclusion

- For age , most of the customers are in the age range of 30-40.
- For balance , above 1000$ is like to subscribe a term deposite .
- Comparing to all algorithms XGboost algorithm has best accuracy score and ROC-AUC score . So it is concluded as optimal model.

Q & A

Thank You