Bank Marketing Prediction
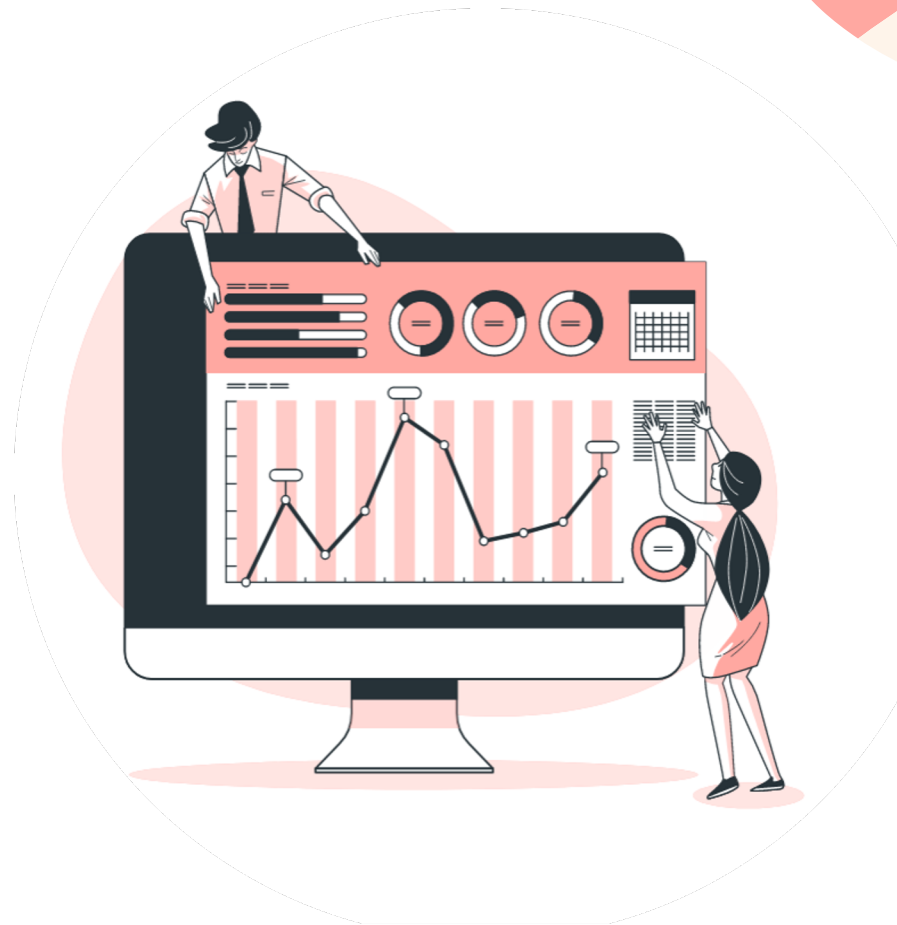
Hernan Trujillo

July 2021

01

Introduction

Goal

Predict if a client will subscribe, or not, to a term deposit account

Data Source

Target

Regular clients of the bank

Predict Target

Potential clients to subscribe.

Data

UCI Machine Learning Repository

Instances: 41.188   Attributes: 20

Method
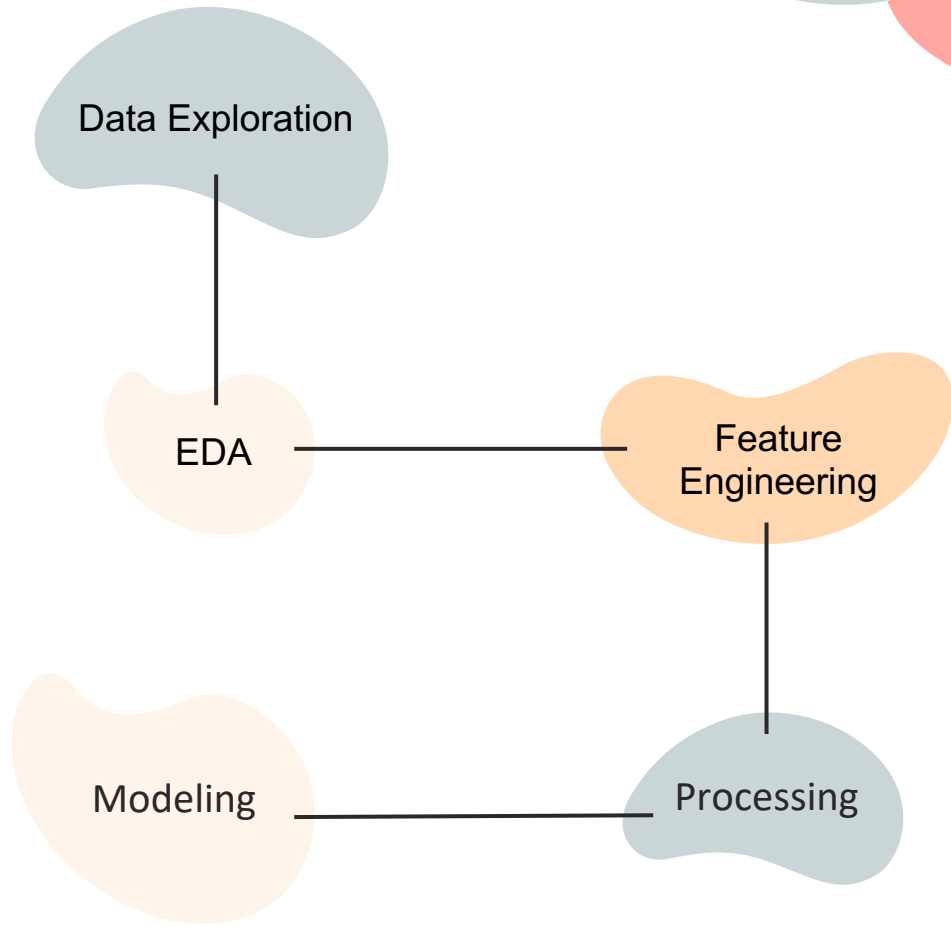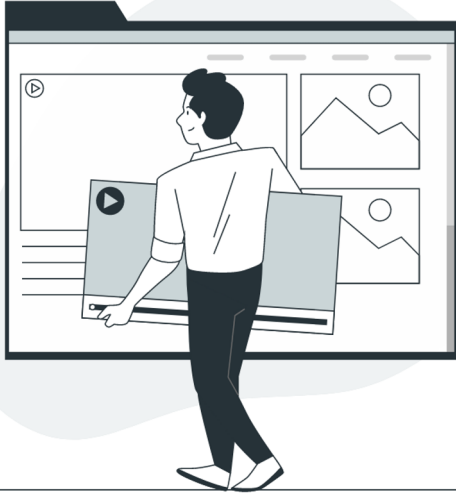
Phone calls gathered through direct marketing campaigns.

02
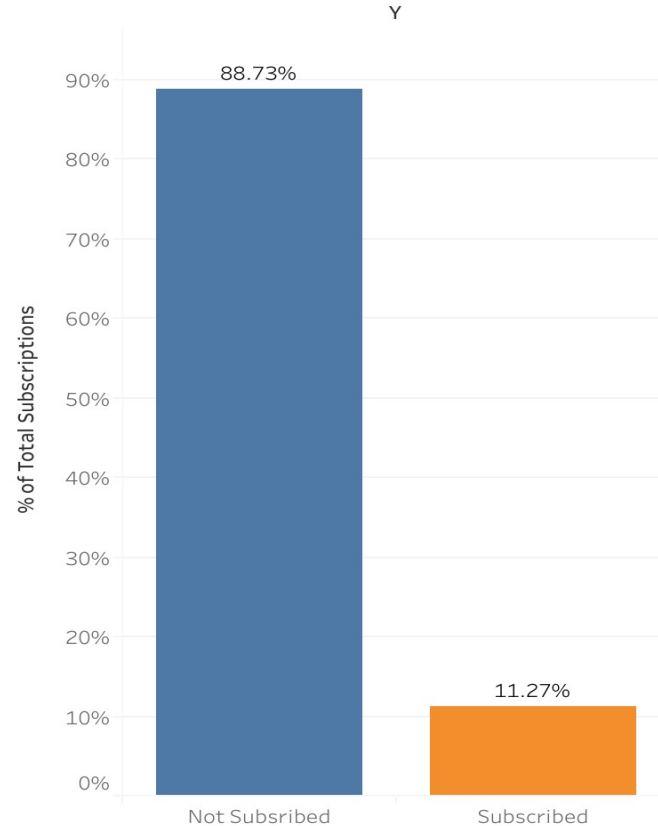
Approach

Classification Model

Classification Workflow

Data Exploration

EDA

Feature Engineering

Modeling

Processing

Data Exploration

Imbalanced Target

- 89% Not Subscribed
- 11% Subscribed

Subscription Outcome

Y

88.73%

11.27%

% of Total Subscriptions

90%

80%

70%

60%

50%

40%

30%

20%

10%

0%

Not Subsribed

Subscribed

Exploratory Data Analysis (EDA)

AVG phone call duration by month

Month	Not Subsribed	Subscribed
mar	181.2	309.3
apr	247.2	471.9
may	230.7	704.8
jun	205.9	554.5
jul	230.1	709.0
aug	194.8	549.7
sep	226.4	385.2
oct	240.7	349.4
nov	212.3	543.5
dec	270.7	423.9

AVG phone call duration

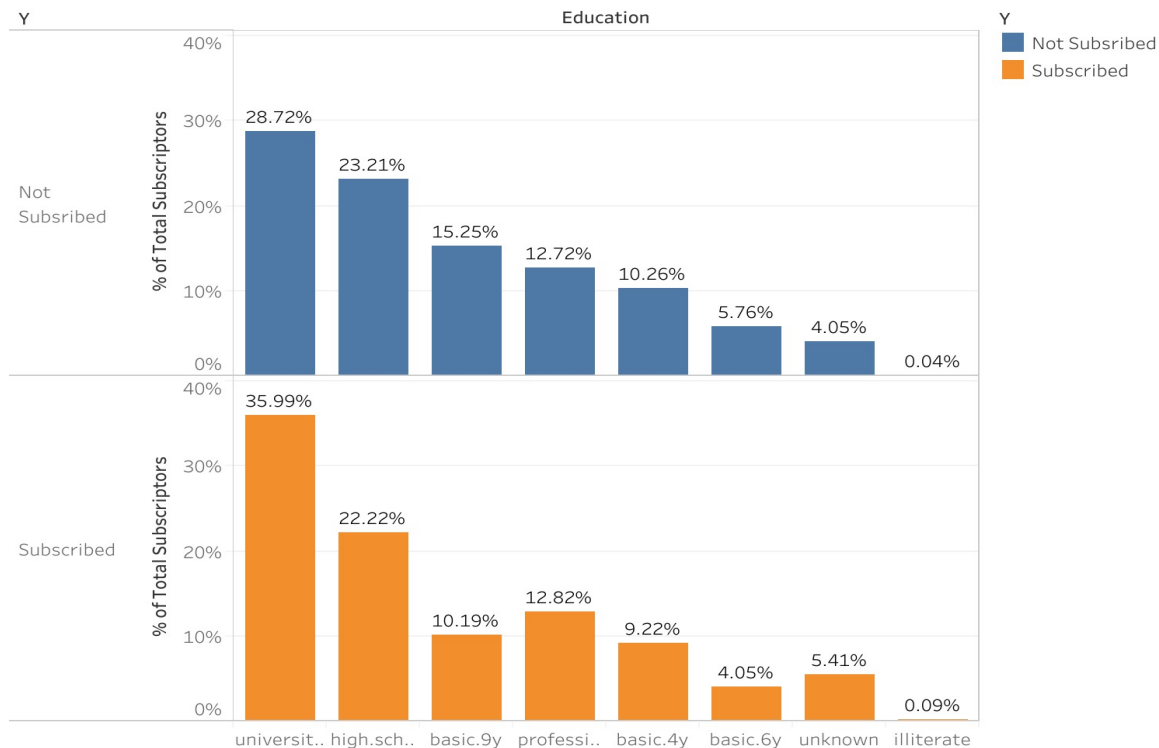|	Not Subsribed	Subscribed
Avg. Duration	220.8	553.2
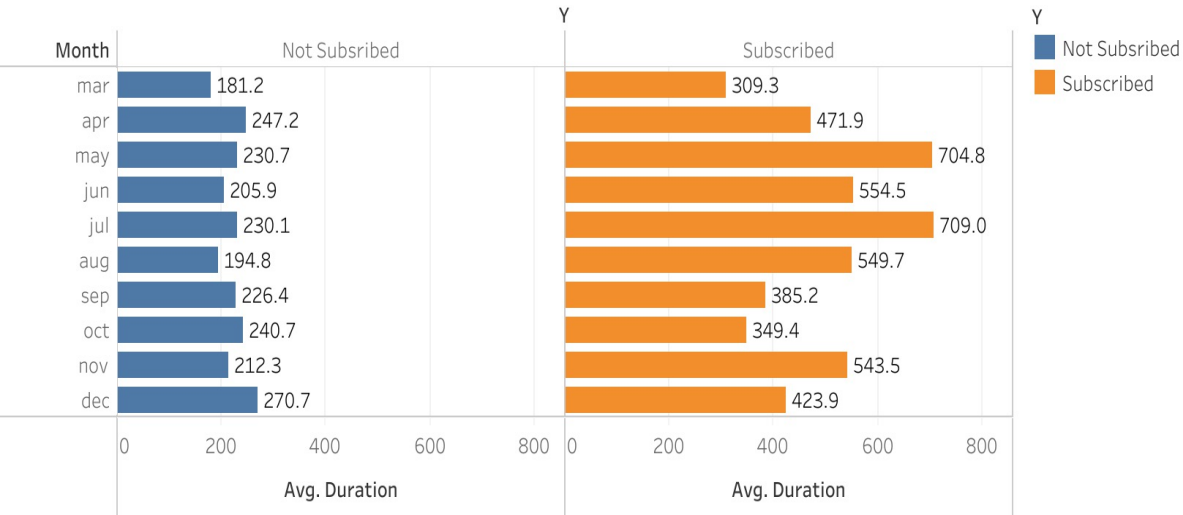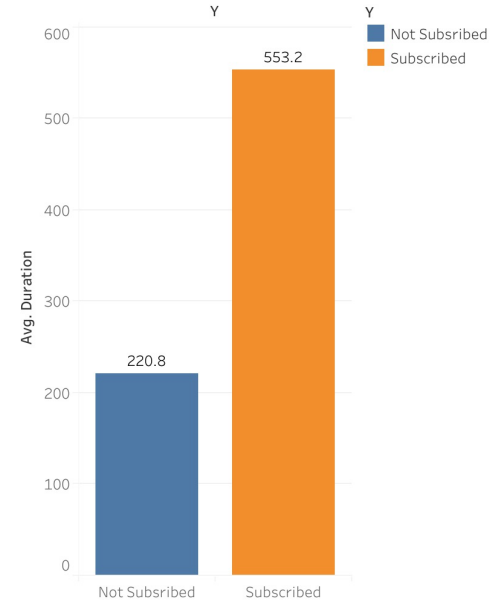
Exploratory Data Analysis (EDA)

Subscribersn by Education

Focus on highly educated target

Feature Engineering Outliers
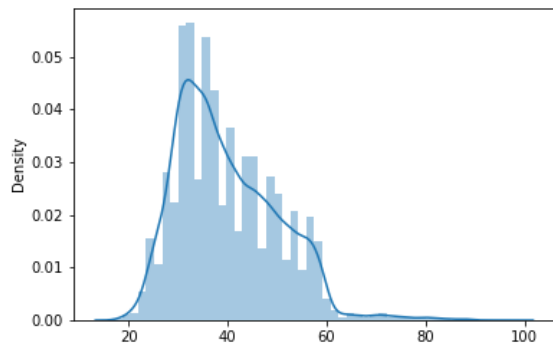
Age

Campaign
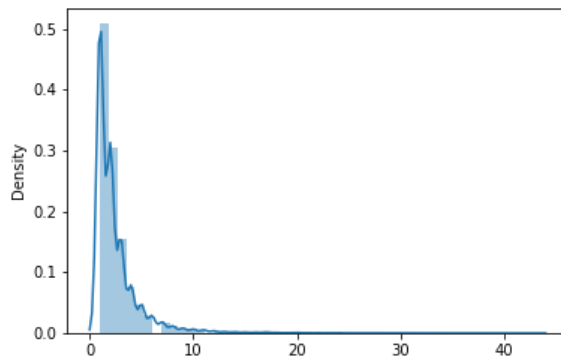
|	age	campaign
count	38205.000000	38205.000000
mean	39.855935	2.567334
std	10.286042	2.768519
min	17.000000	1.000000
25%	32.000000	1.000000
50%	38.000000	2.000000
75%	47.000000	3.000000
max	98.000000	43.000000

Feature Correlations

Dropping highly correlated features for:
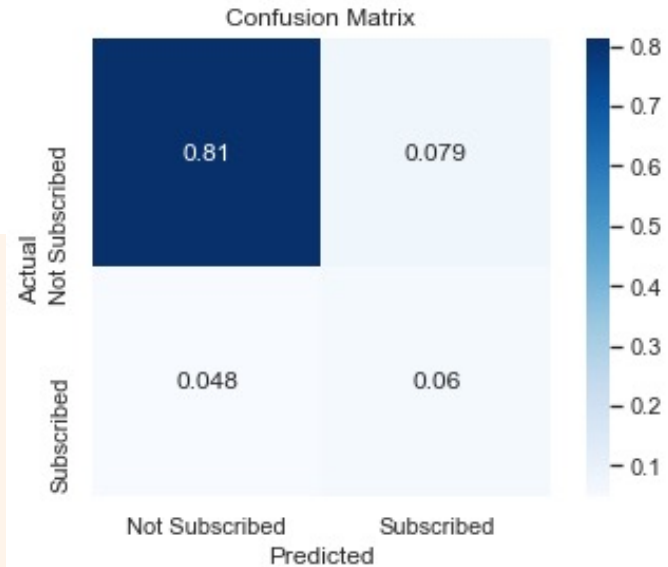
- Avoid overfitting

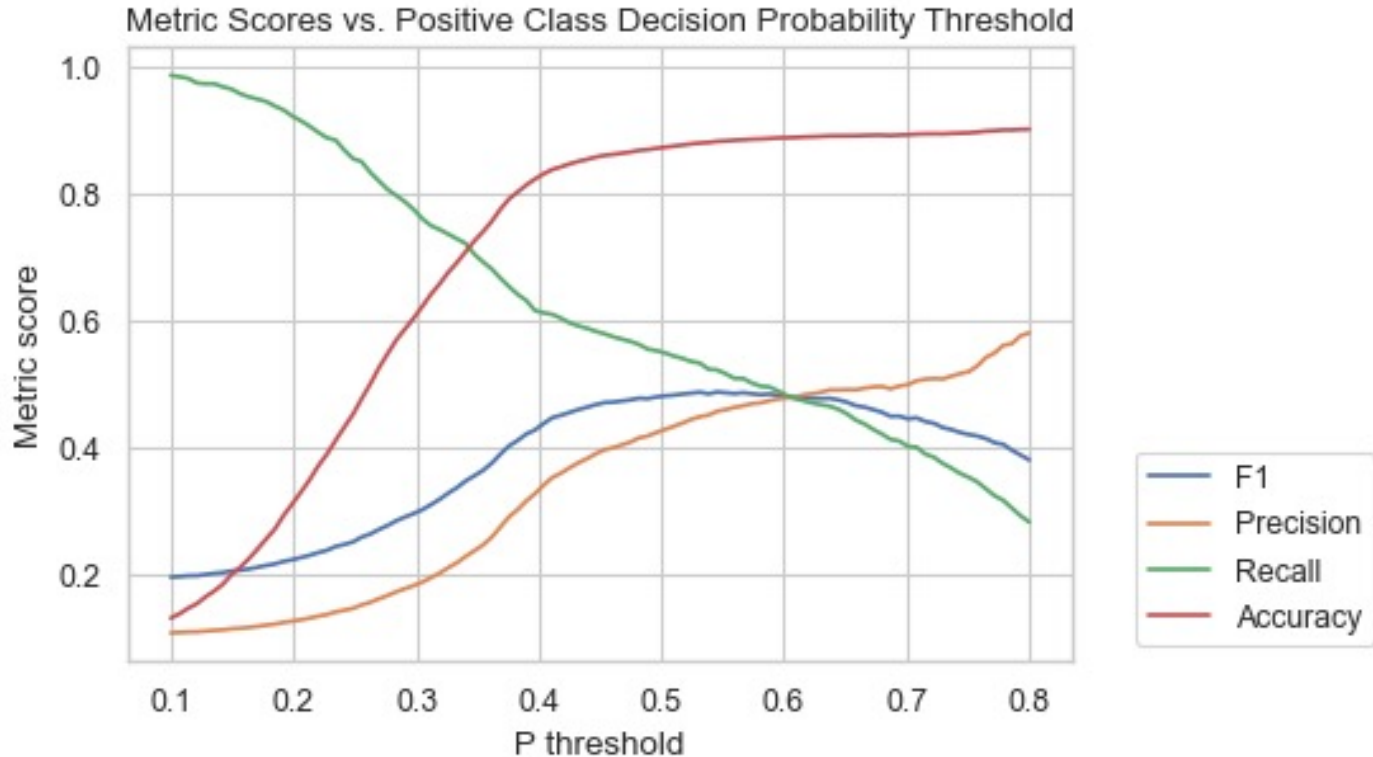- Easy for interpretation

Keeping "euribor3m"

Confusion Matrix

Logistic Classification
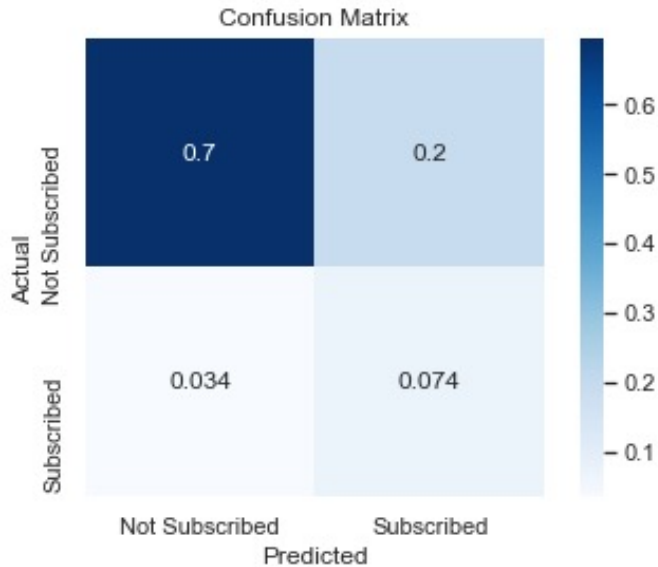
Random Forest Classification

Random Forest Threshold

Metric Scores vs. Positive Class Decision Probability Threshold

03

Results

Modeling

Classification Models Scores

Base Model: Logistic Regression	F1 0.47
Logistic Regression (Class Weight)	F1 0.48
Random Forest	F1 0.49

F1=0.49

Conclusion:

Random Forest Model

ROC-AUC Comparison
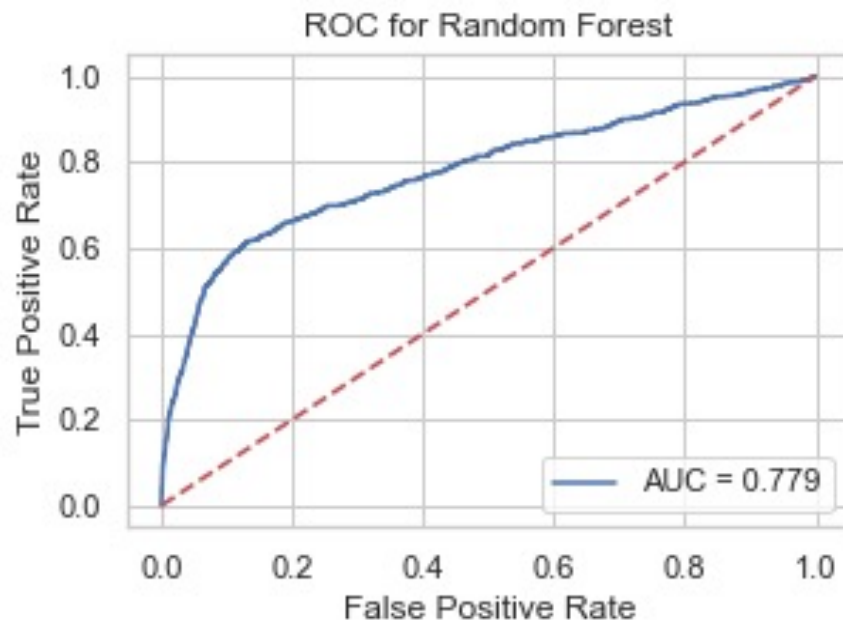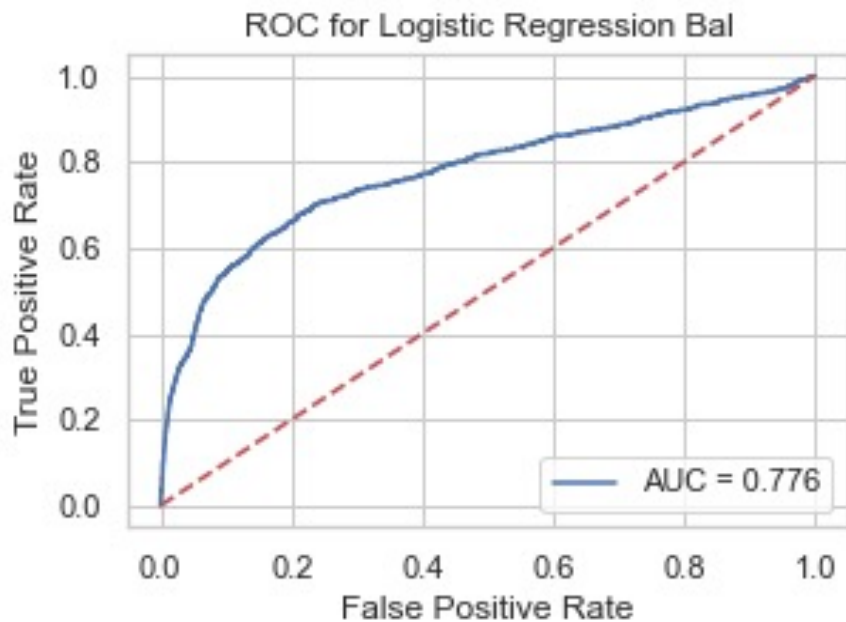
Log Loss Cross-Entropy Scores

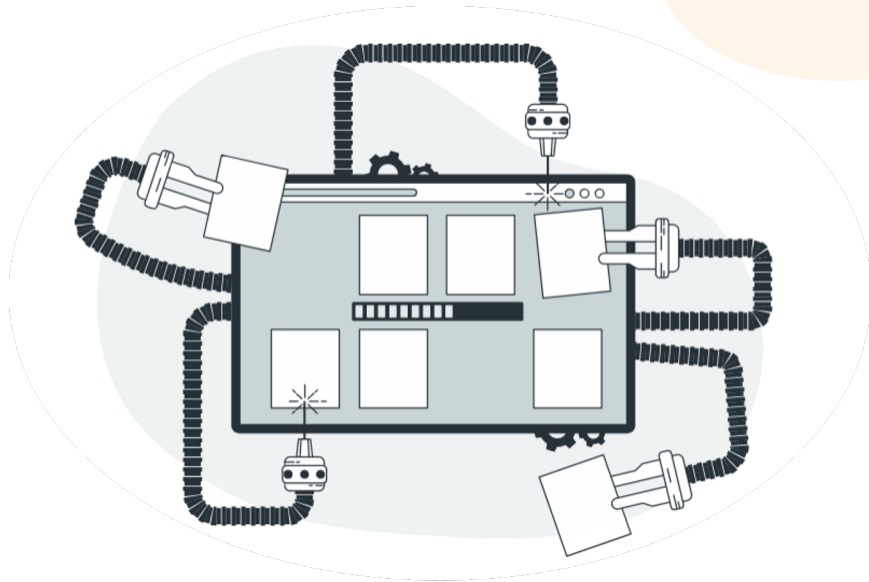|	
Logistic Regression (balanced)	F1 0.54
Random Forest Classifier (balanced)	F1 0.44

Log Loss=0.44

Prediction Confidence Score

Random Forest Model

Future work

Try more sophisticated models

Model 1

Gradient Boosting
Machine

Model 2

XGBoost

Thanks!

hernantru943@gmail.com

https://public.tableau.com/app/profile/hernan.trujillo/viz/BankMarketing

https://github.com/hernantru943/CLS_PROJECT