Marketing Analytics - Project

Personal Finance - Customer Segmentation

Group K:
Tanushree **Balaji**, Vivek **Dhulipalla**, Pratik **Gawli**,
Snehal **Naravane**, Aishwarya **Rajeev**

Table of Contents

Introduction

Project Description

Problem Statement:

Examine campaigns run by a financial institution to promote their term deposit product and identify strategies for effective promotion & higher conversion

Approach:

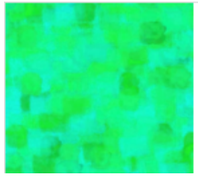
Identify key drivers for subscription using Logistic Regression and spot viable target segments using K-means.

Data Introduction

Data:
Banking Customer Campaign Response and Subscription Rate

Features:

- Demographics
- Campaign Response
- Economic Factors
- Product Usage



Bank Customers - Segmentation

Direct marketing campaigns of a Portuguese banking institution

kaggle.com

Exploratory Data Analytics

Pre-processing

DROP DUPLICATES

The rows with duplicate values for the following combination were removed: [age, job, marital, education, default, housing, loan, contact, duration, campaign]

CREATE DUMMY VARIABLES

The categorical features of job type, marital status, education type, contact type etc. were dummy encoded.

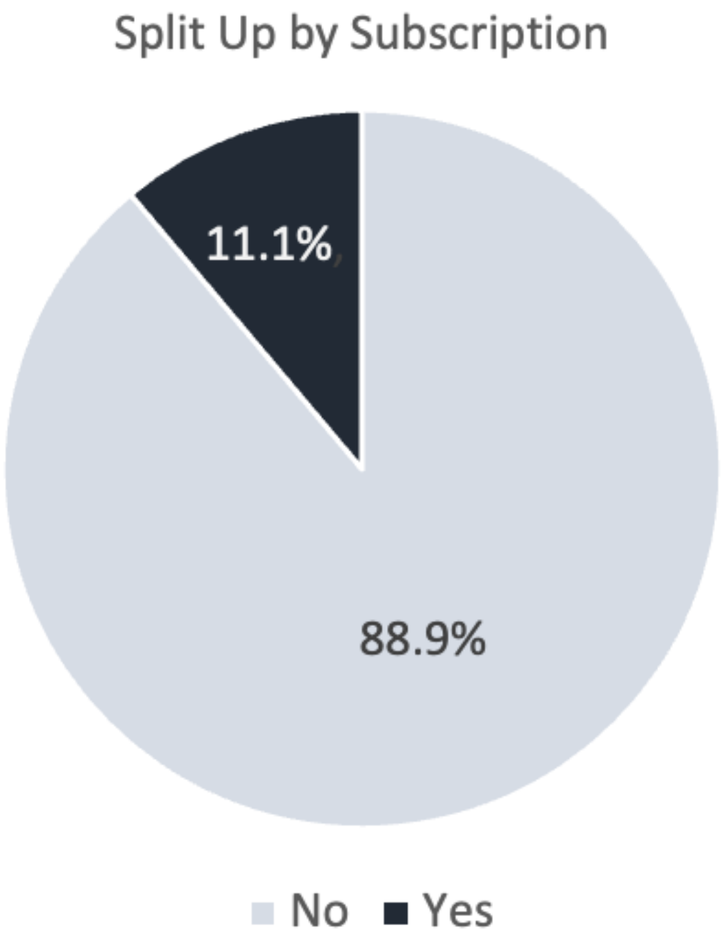
SCALE NUMERIC VARIABLES

All the numerical fields were scaled between the values 0 and 1 for equal scaling and better comparison of the ranges.

REMOVE NULLS

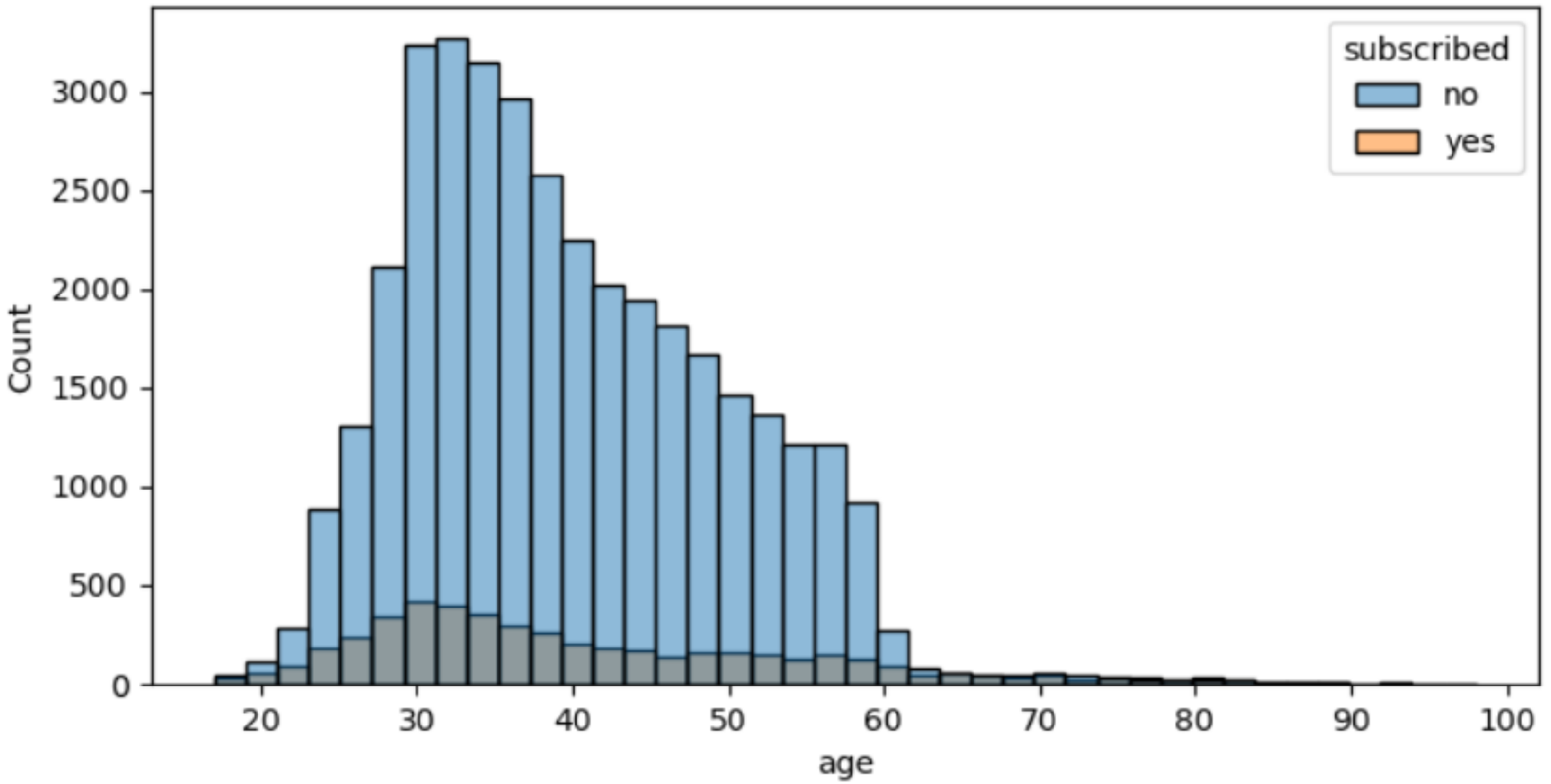
All the records with missing/null values were dropped.

Target Variable - Subscription



It's an imbalanced data set with approximately 10% subscribers

Age Group



Age brackets	Total Proportion	Subscribed Proportion
10 - 20	0.34%	0.73%
20 - 30	17.59%	19.14%
30 - 40	39.78%	36.96%
40 - 50	24.86%	17.97%
50 - 60	15.22%	15.02%
60 - 70	1.18%	5.41%
70 - 80	0.74%	3.10%
80 - 90	0.26%	1.55%
90 - 100	0.02%	0.11%
Total	**100.00%**	**100.00%**

Subscribers are heavily concentrated in the middle age range (40 years), possibly due to higher customer volumes, but also a greater inclination to invest.

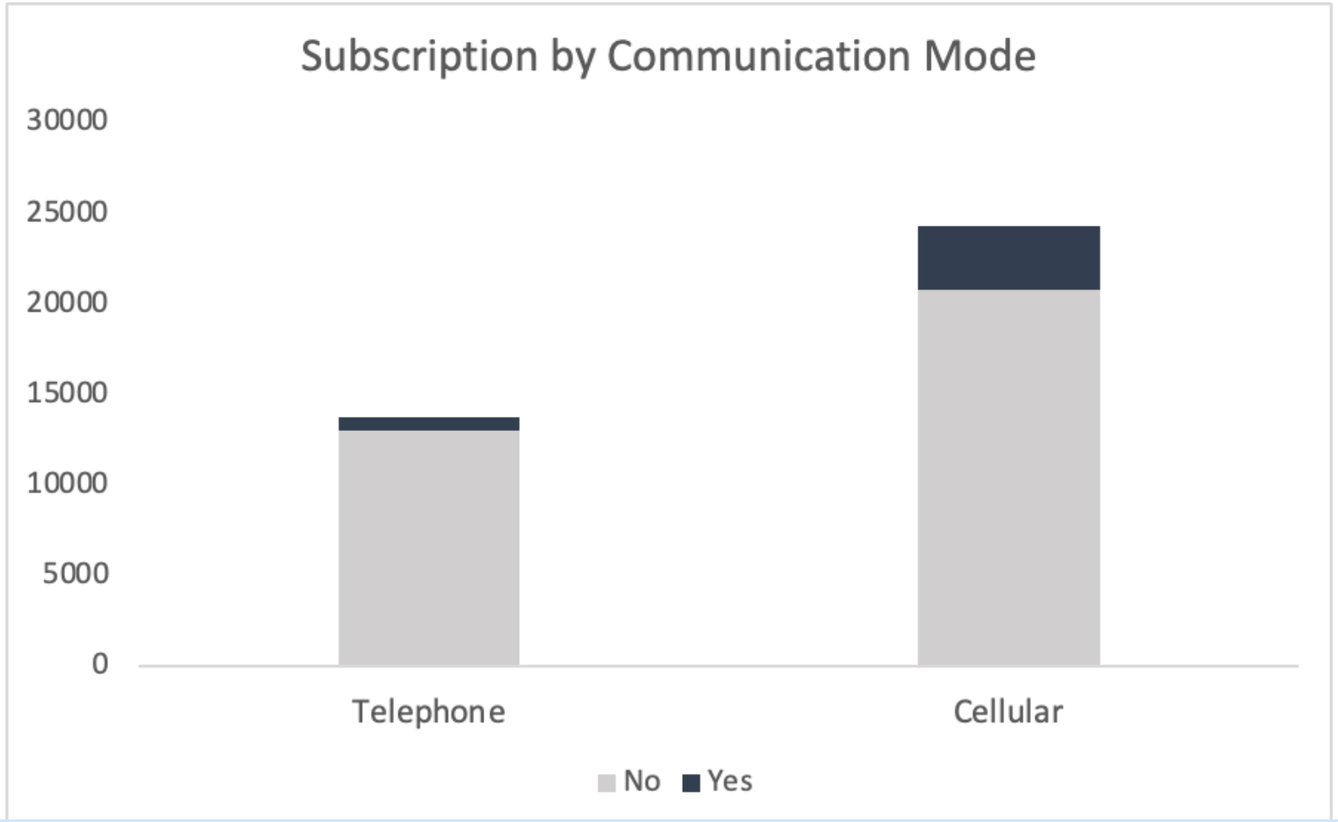
Previous Campaign Touch Points

# Prev contact	Number of customers	Conversion Rate
0	35481	8.83%
1	4541	21.20%
2	753	46.42%
3	215	59.26%
4+	94	57.45%

Despite lack of active marketing communication (0 contact segment), ~9% converted organically.

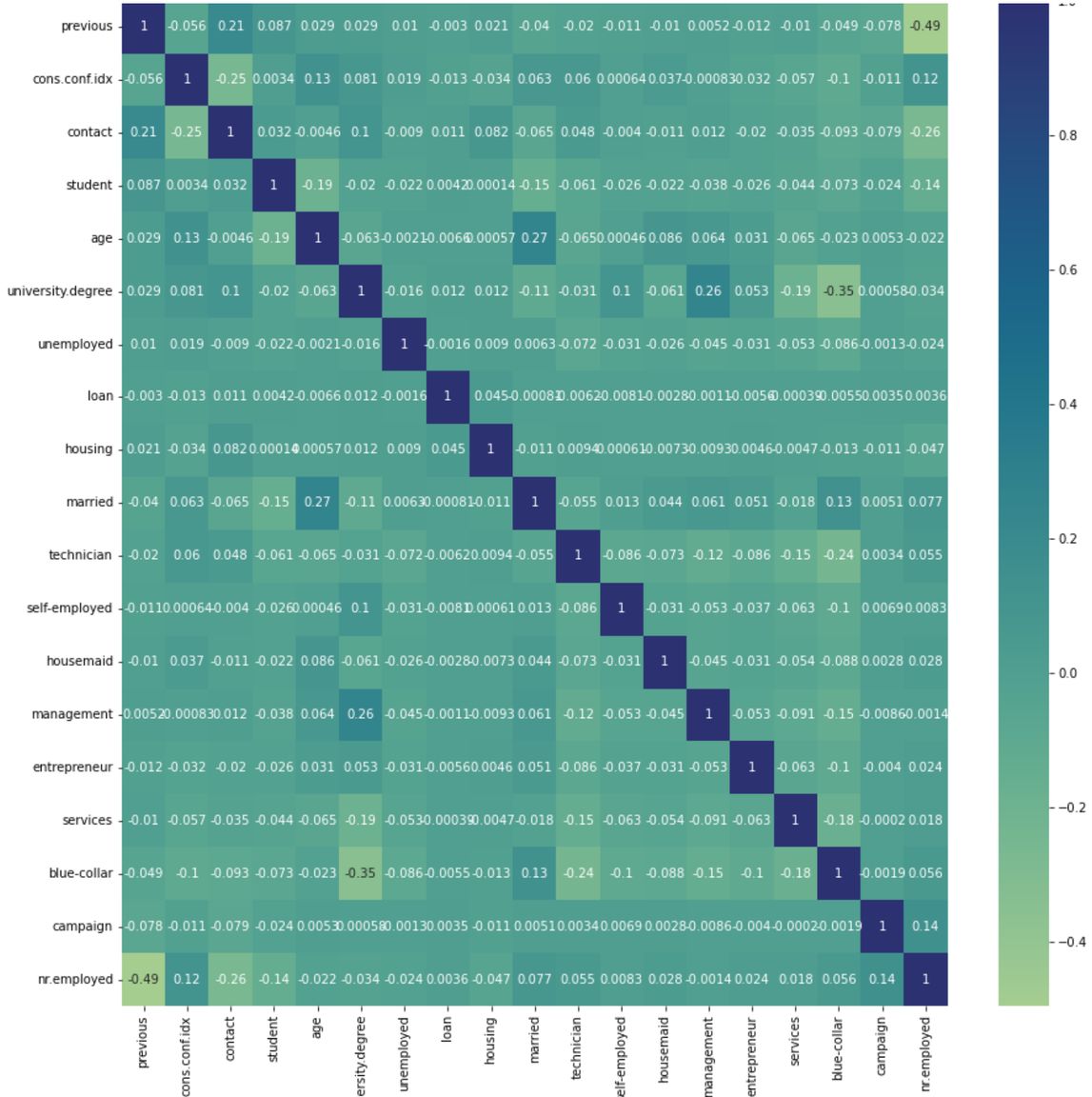
High touch point customers have an average conversion of 46%, indicating better customer engagement

Communication Mode



About 63% of the customers are contacted through cellular with a 9% hit rate

Feature Selection

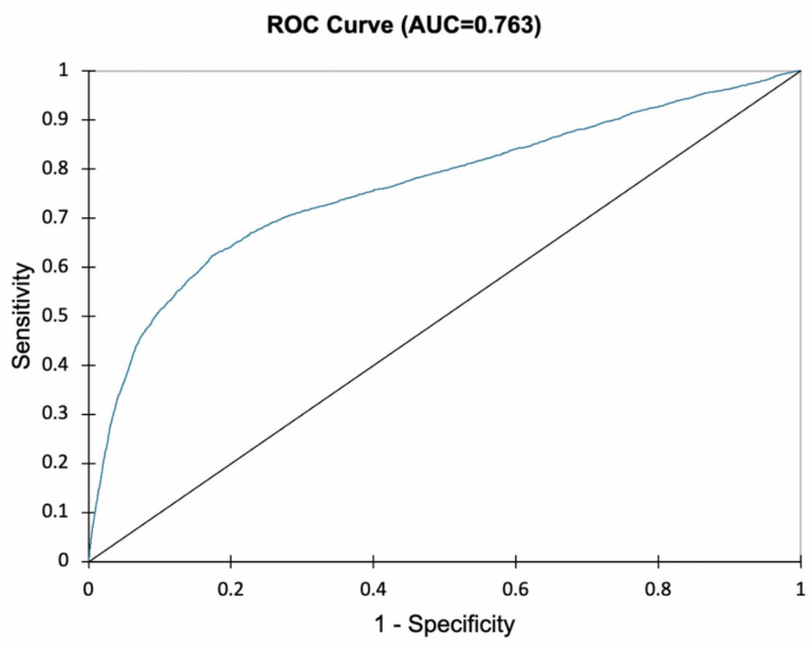


A correlation threshold of +/- 30% was used to select features prior to modeling in order to have statistically significant coefficients.

Logistic Regression & Segmentation

Logistic Regression

Source	Value	Pr > Chi_
Intercept	-0.786	<0.0001
age	0.26	0.055
housing	-0.044	0.215
loan	-0.041	0.409
contact	0.596	<0.0001
campaign	-1.922	<0.0001
previous	0.789	<0.0001
cons.conf.idx	0.752	<0.0001
nr.employed	-2.923	<0.0001
blue-collar	-0.339	<0.0001
entrepreneur	-0.193	0.07
housemaid	-0.147	0.225
management	-0.162	0.027
self-employed	-0.11	0.265
services	-0.286	0
student	0.342	0.001
technician	-0.058	0.283
unemployed	0.006	0.958
married	-0.056	0.148
university.degree	0.125	0.004



Given the binary nature of the target variable, we used logistic regression to identify drivers for subscription.

AUC for this model is 0.76, indicating that it's identified the coefficients rather accurately

Significant Variable Selection

Source	Value
previous	0.789
cons.conf.idx	0.752
contact	0.596
student	0.342
age	0.26
university.degree	0.125
management	-0.162
blue-collar	-0.339
Intercept	-0.786
campaign	-1.922
nr.employed	-2.923

Prior exposure to other campaigns seems to a have a positive effect on subscription

Contact through cellphones works better than landlines. Do people equate landlines to spam?

Older people are more drawn to term deposits

Highlighted variables are hence some important drivers for term deposit subscription

Segmentation

Having chosen some intuitive/ actionable attributes from logistic regression, it'd be interesting to see if patterns emerge when users are segmented

Iterating between cluster options, 7 seemed like a desirable choice

Results


Sales Performance
Nuda
Jenet
Brian
Steven
Robert
Anna
Satid
Aphisit
Naomi
Target
Sales
Sales Performance
Nuda
Jenet
Brian
Steven
Robert
Anna
Satid
Aphisit
Naomi
Target
Sales

Clustering Results

Cluster	age	contact	previous	Sum of weights	Within-cluster variance	Cluster Category
1	0.283	0.000	0.005	13786	0.014	Late 20s, Telephone Contact, No prior campaign exposure
2	0.248	1.000	0.012	14174	0.006	Early 20s, Cellphone Contact, No prior campaign exposure
3	0.444	1.000	0.017	4640	0.004	Middle Age, Cellphone Contact, No prior campgain exposure
4	0.151	1.000	0.109	4591	0.019	Teens, Cellphone Contact, Minimal campaign exposure
5	0.575	1.000	0.030	467	0.006	Nearing Retirement, Cellphone Contact, No prior exposure
6	0.639	1.000	0.216	405	0.033	Retired, Cellphone Contact, Minimal Contact Exposure
7	0.847	1.000	0.055	81	0.012	Elderly, Cellphone Contact, No Campaign Exposure

These were the categories that emerged, breaking down the data into 7 parts. Seems like cellphone contact leads to a lot of variety.

Clustering Results Contd

11.1%

Average
Subscription Rate

3.2%

Average
Success Rate

We chose clusters that performed better than average

Clustering Results Contd

Cluster	age	contact	previous	Cluster Category	Subscription Rate	Previous Success Rate
1	0.283	0.000	0.005	Late 20s, Telephone Contact, No prior campaign exposure	5.2%	0.7%
2	0.248	1.000	0.012	Early 20s, Cellphone Contact, No prior campaign exposure	10.8%	1.3%
3	0.444	1.000	0.017	Middle Age, Cellphone Contact, No prior campaign exposure	11.6%	2.4%
4	0.151	1.000	0.109	Teens, Cellphone Contact, Minimal campaign exposure	22.7%	14.2%
5	0.575	1.000	0.030	Nearing Retirement, Cellphone Contact, No prior campaign exposure	36.4%	7.5%
6	0.639	1.000	0.216	Retired, Cellphone Contact, Minimal campaign exposure	50.9%	38.8%
7	0.847	1.000	0.055	Elderly, Cellphone Contact, No prior campaign exposure	45.7%	16.0%

Users in clusters 4-7 have decent success rates despite minimal exposure to prior campaigns.

Clustering Results Contd

Clusters 4 & 5 - Not Similar

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²
Intercept	-5.577	1.201	21.561	<0.0001
age	8.285	2.012	16.954	<0.0001
contact	0.000	0.000		
previous	7.404	1.855	15.940	<0.0001
clus_4	4.407	1.205	13.382	0.000
clus_$*age	-12.433	2.117	34.479	<0.0001
clus_4*prev	-2.628	1.885	1.944	0.163
clus4_contac	0.000	0.000		

Clusters 4 & 6 - Not Similar

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²
Intercept	-1.330	0.828	2.580	0.108
age	1.156	1.134	1.040	0.308
contact	0.000	0.000		
previous	2.929	0.825	12.595	0.000
clus_4	0.160	0.833	0.037	0.848
clus_4*age	-5.304	1.311	16.357	<0.0001
clus_4*prev	1.847	0.891	4.296	0.038
clus4_contact	0.000	0.000		

Clusters 5 & 6 - Not Similar

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²
Intercept	-1.330	0.828	2.580	0.108
age	1.156	1.134	1.040	0.308
contact	0.000	0.000		
previous	2.929	0.825	12.595	0.000
clus_5	-4.247	1.459	8.477	0.004
clus_5*age	7.129	2.309	9.528	0.002
clus_5*prev	4.475	2.030	4.860	0.027
clus5_contac	0.000	0.000		

Clusters 5 & 7 - Similar

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²
Intercept	-0.786	4.252	0.034	0.853
age	0.534	5.022	0.011	0.915
contact	0.000	0.000		
previous	2.923	2.356	1.539	0.215
clus_5	-4.791	4.418	1.176	0.278
clus_5*age	7.750	5.410	2.052	0.152
clus_5*prev	4.482	2.998	2.235	0.135
clus5_contact	0.000	0.000		

While 4 and 6 are pretty distinct, segments 5 and 7 can be combined into 1 while targeting

Summary / Checks

Observations & Leads for future campaigns

26.2%

Average
Subscription Rate

15.4%

Average
Success Rate

By restricting marketing to chosen segments, the average rates will shoot up.

$ Value Check

Assume a more conservative success rate of **3.2%** (also our earlier baseline) (i.e.) **177** people subscribe:

Based on market trends:
Minimum balance for term deposit = **$1000**
Average annual interest rate = **3%** ~ **$30**

Let's the average price of contacting a customer is about **$2**
Let's also assume that the bank rotates the money and makes an average of **5%** on it

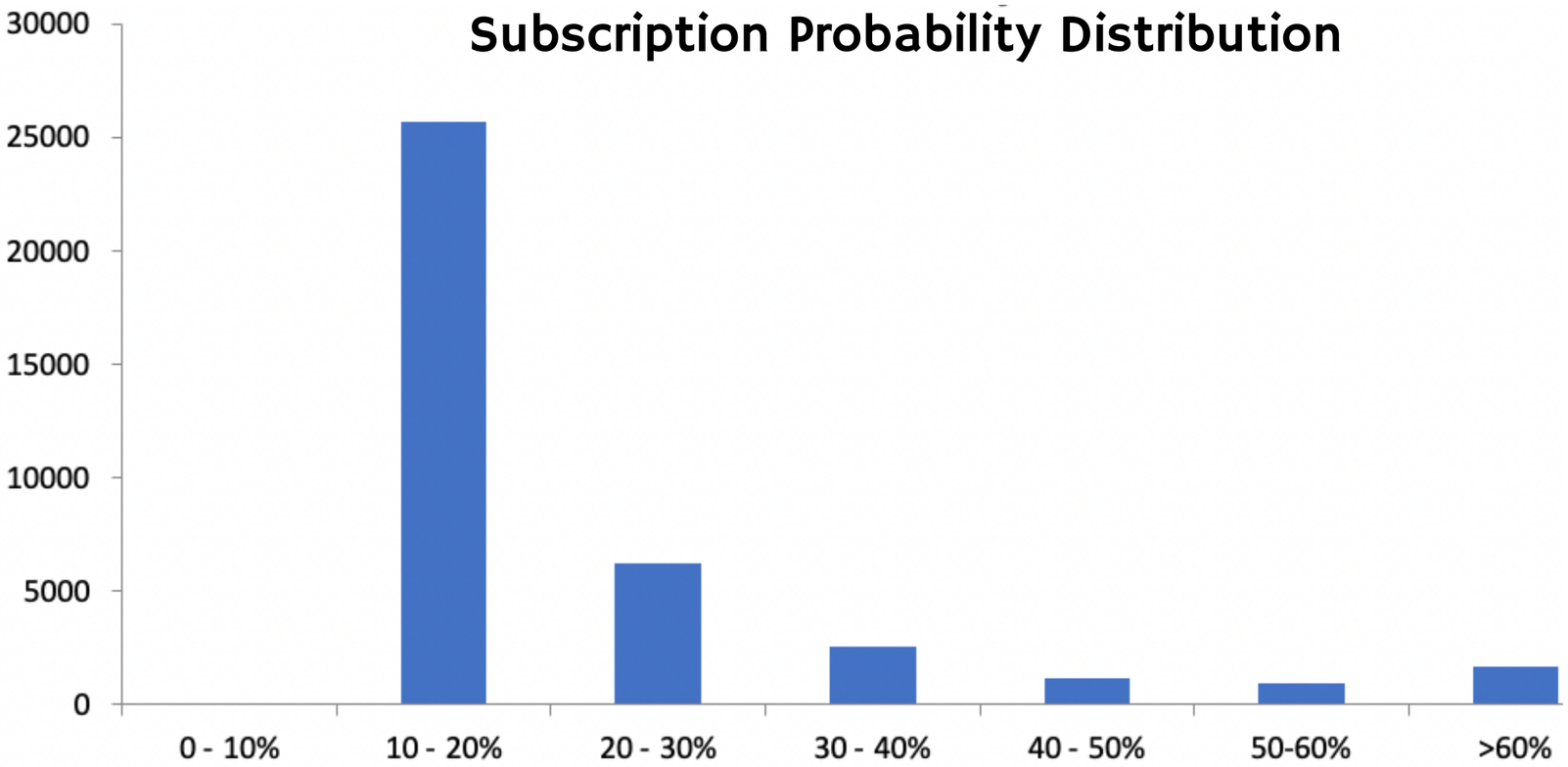
Total revenue to the bank = 177 * ($1000 + 5% of 1000) = **$185850**
Total cost if there are 10 touch points for each customer = $2 * 10 * 5544 = **$110,880**
Interest payment to subscribed customers = $177 * 30 = **$5310**

Profit = Total Revenue - Total Cost = **$69660**

Therefore, at 3.2% success rate, the bank makes an additional profit of **$69660**
At 15.4% success rate, the bank would make a profit of **$870k** at an additional interest payment of only **$20k**

Thus, high success rates significantly drive up profit, making targeting a worthwhile activity at not a lot of extra cost

Sanity Check



If we rely on just the posterior probability with a threshold of around 50% we would only target 1,600 customers with an uncertainty about the performance. However, with segmentation, we establish better defined target segments

Recommendations

Recommendations

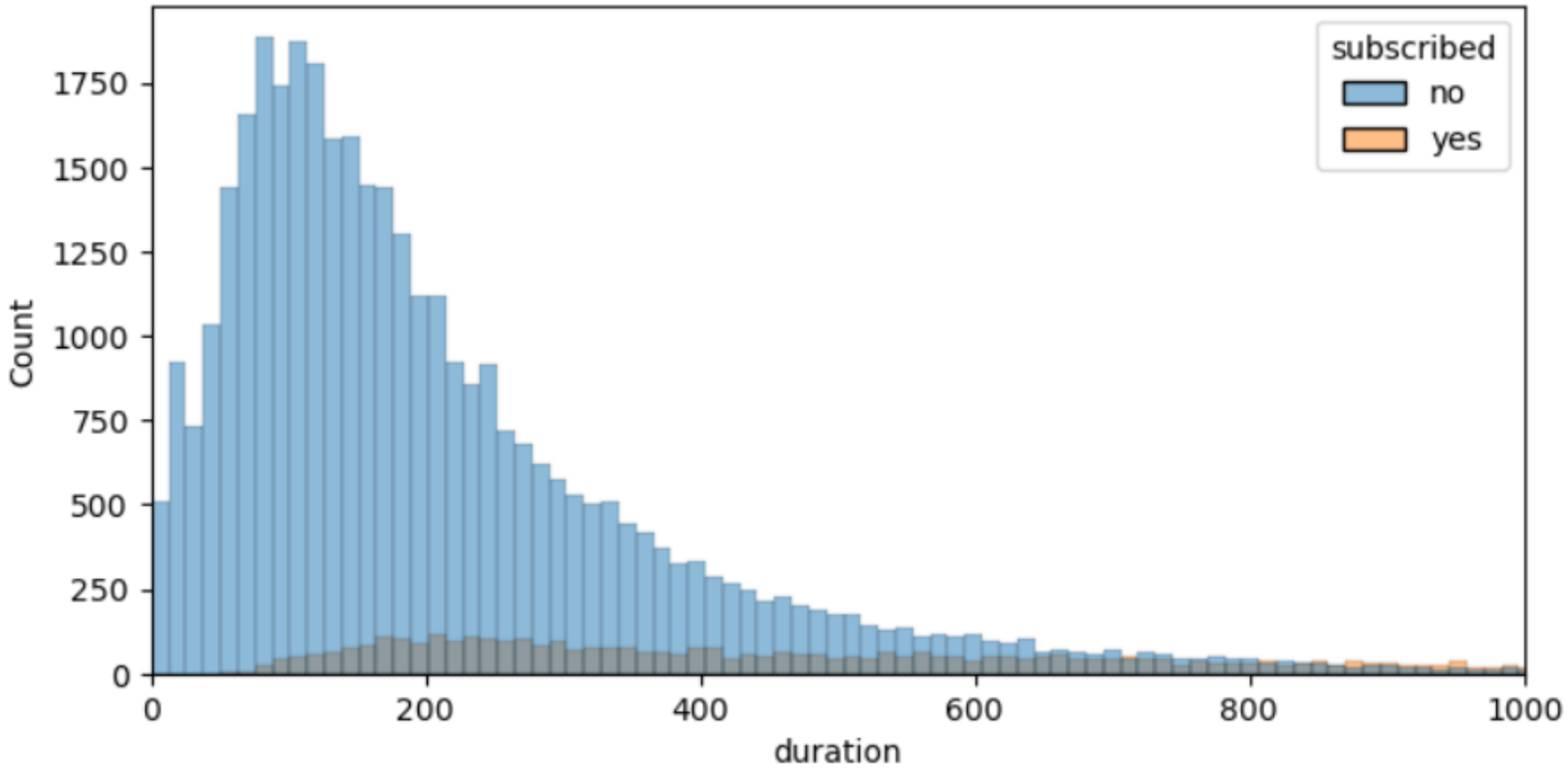
- Segments that are most eager to convert include the retired/elderly. This makes them a viable target to start off any experiment with
- The elderly segment has a small sample size. This opens up the space to find similar customers outside of this group to target
- It is also a good idea to test this out on other Banking datasets
- While this is definitely a better strategy than targeting all customers at once, it's important to always take modelling results with a pinch of salt. Hence, it would help to start small & scale up, as in the case of A/B tests

Thank You!

Any Questions?

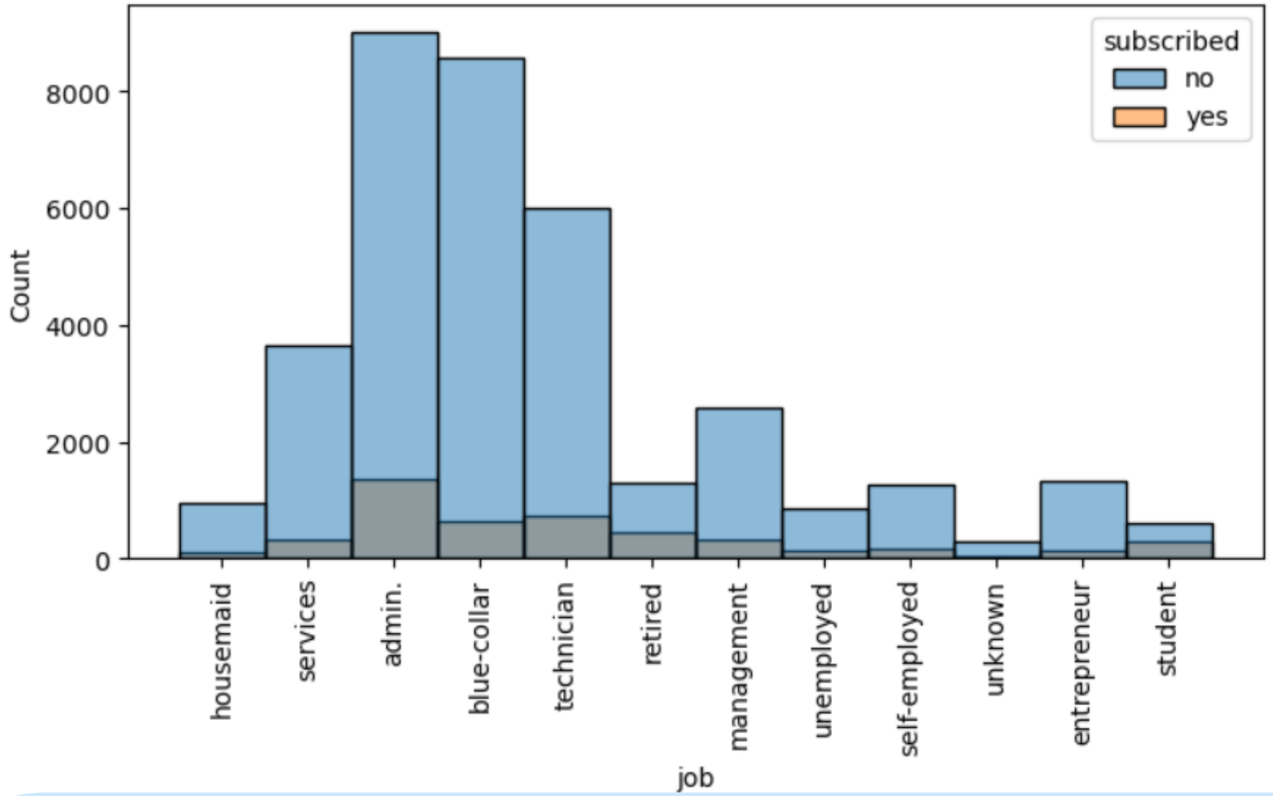
Appendix

Duration of last contact



The mode duration of the last contact for all was at 90 seconds, while the mode duration of the last contact for subscribed customers was 301 seconds, which makes sense as the customers who subscribed actually stayed on the call longer.

Job



Total	
Job type	Proportion
admin.	25.47%
blue-collar	22.64%
technician	16.49%
services	9.73%
management	7.16%
retired	4.21%
entrepreneur	3.57%
self-employed	3.49%
housemaid	2.60%
unemployed	2.49%
student	2.14%

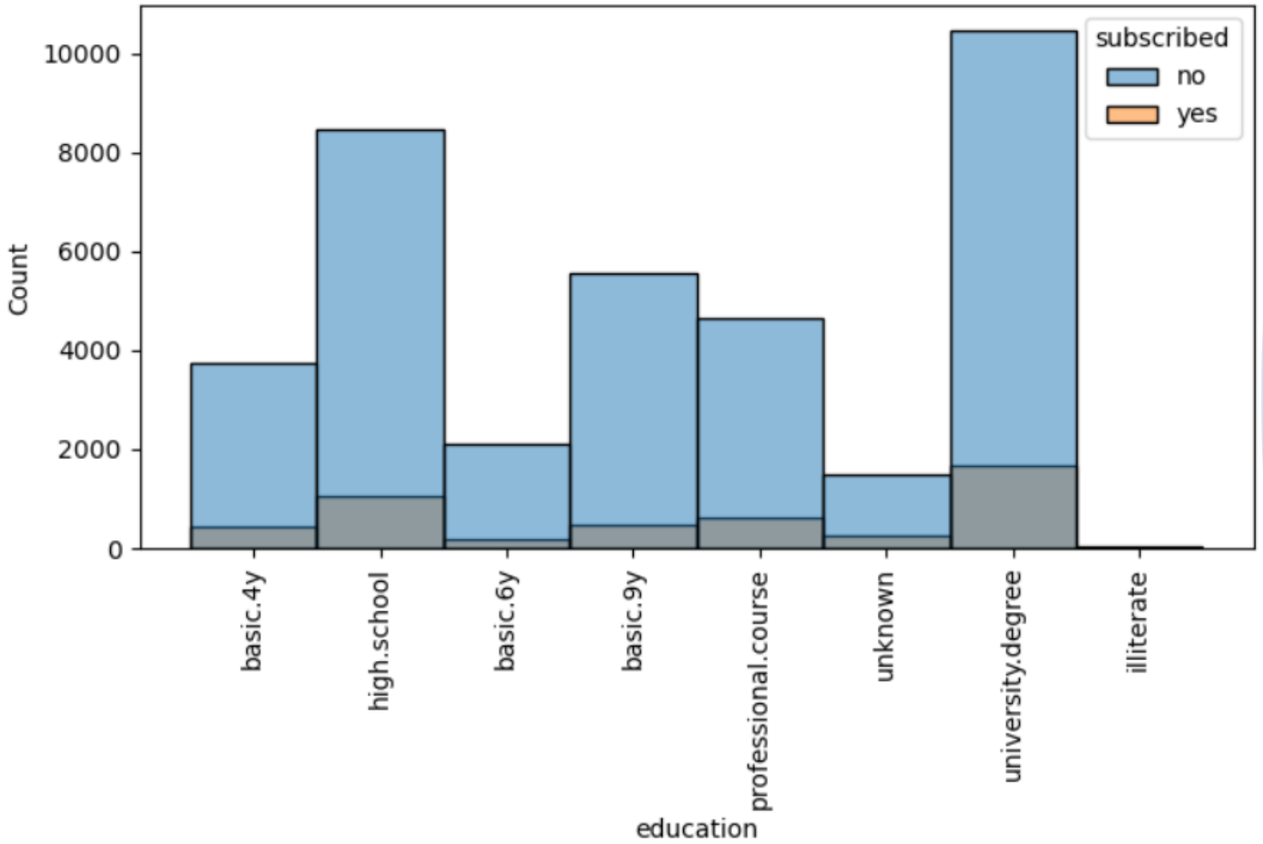
Subscribed	
Job type	Proportion
admin.	29.33%
technician	15.88%
blue-collar	13.86%
retired	9.40%
management	7.14%
services	7.03%
student	5.98%
self-employed	3.24%
unemployed	3.13%
entrepreneur	2.70%
housemaid	2.31%

The job type of most of the subscribers were admin, technician or blue-collar, which are also the most common job types overall (though with different proportions)

Education

Total	
Education level	Proportion
university.degree	30.80%
high.school	24.15%
basic.9y	15.32%
professional.course	13.28%
basic.4y	10.59%
basic.6y	5.82%
illiterate	0.05%

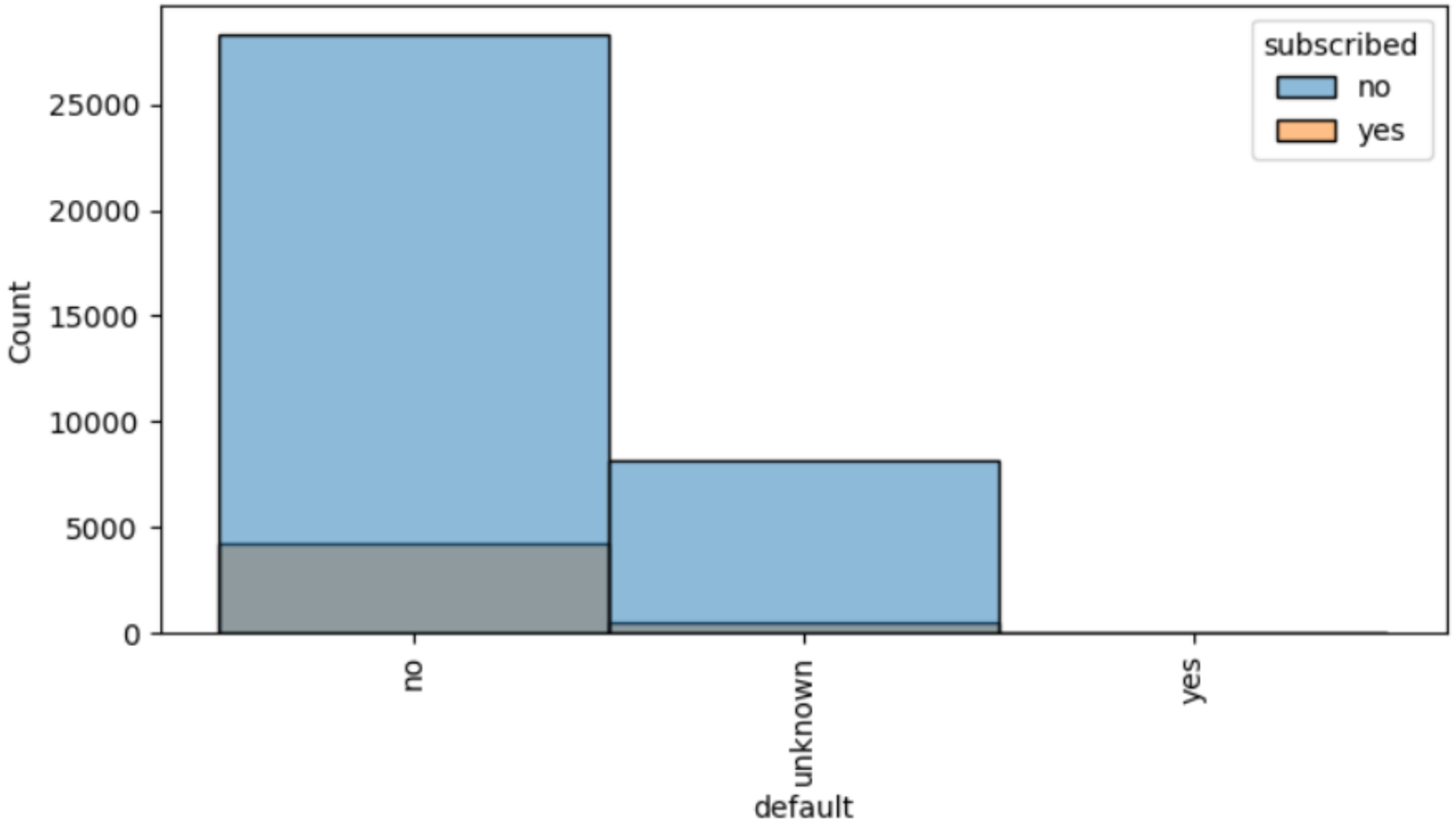
Subscribed	
Education level	Proportion
university.degree	30.80%
high.school	24.15%
professional.course	15.32%
basic.9y	13.28%
basic.4y	10.59%
basic.6y	5.82%
illiterate	0.05%



The education level of most of the subscribers were university degree, high school and basic 9y, with professional course coming after, which are also the most common job types overall (though with different proportions).

So, it seems that the higher the education level, the higher the propensity to invest in a term deposit.

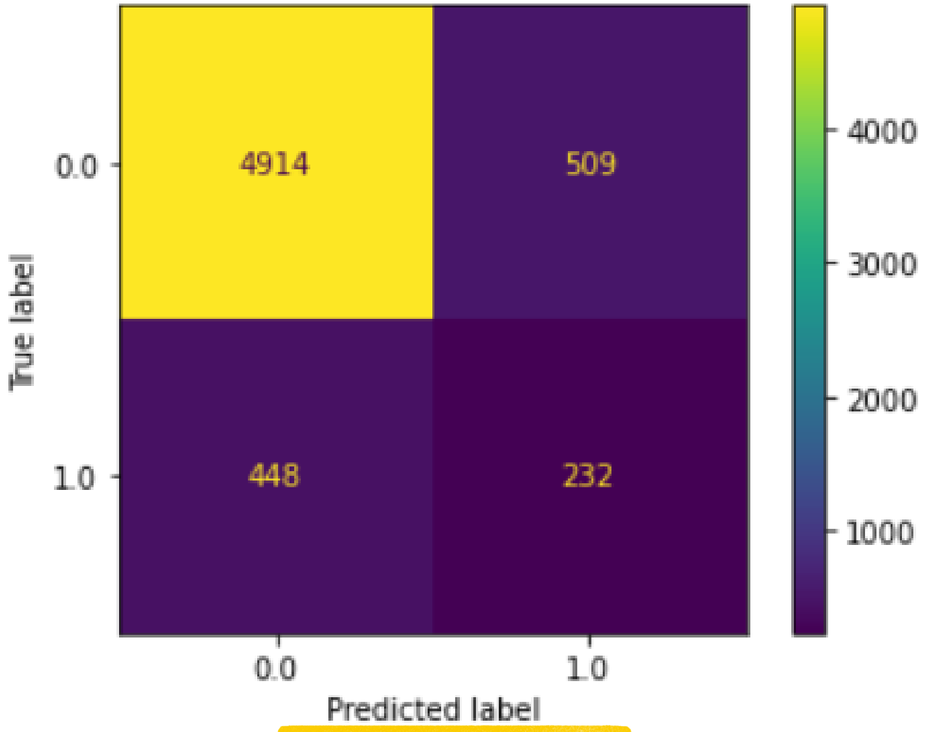
Default



Maximum subscribers are from the 'no default' segment

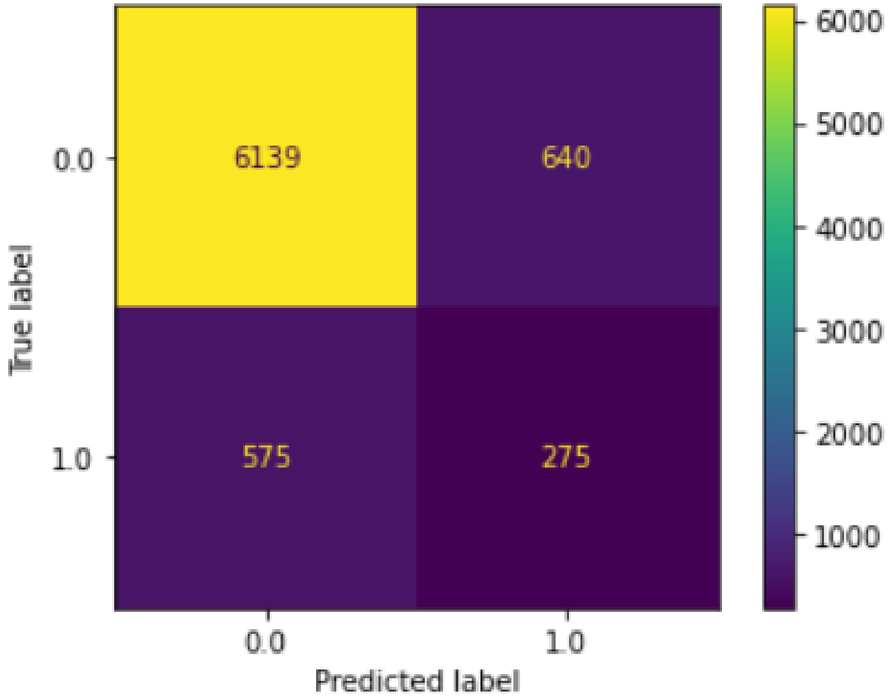
Confusion Matrix: Decision Tree

Confusion Matrix for Validation set using Decision Tree classifier:



F1 Score: 0.33

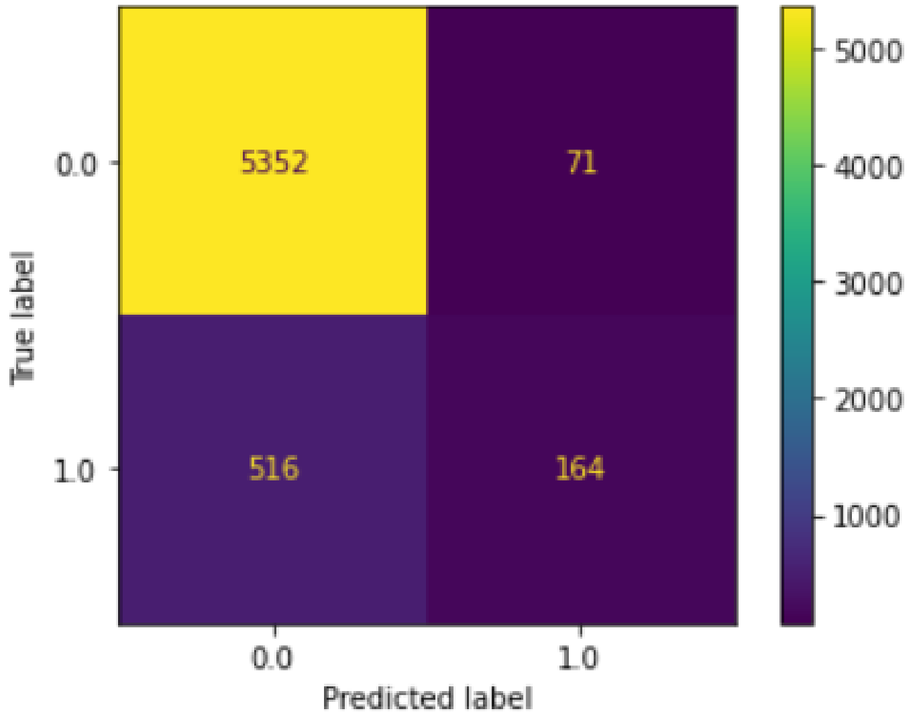
Confusion Matrix for Test set using Decision Tree classifier:



F1 Score: 0.32

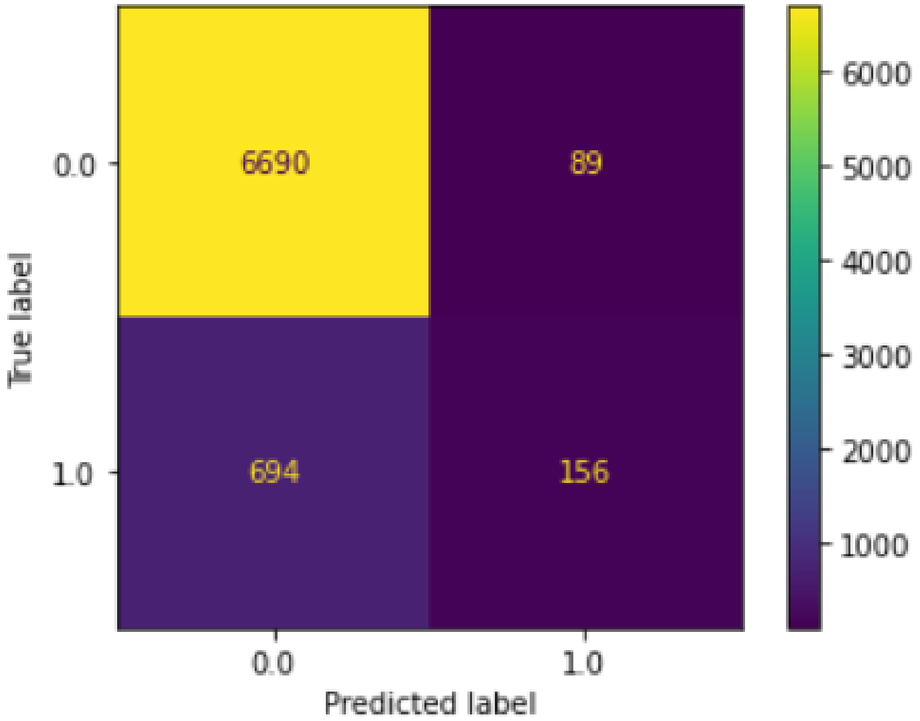
Confusion Matrix: Logistic Regression

Confusion Matrix for Validation set using Logistic Regression classifier:



F1 Score: 0.36

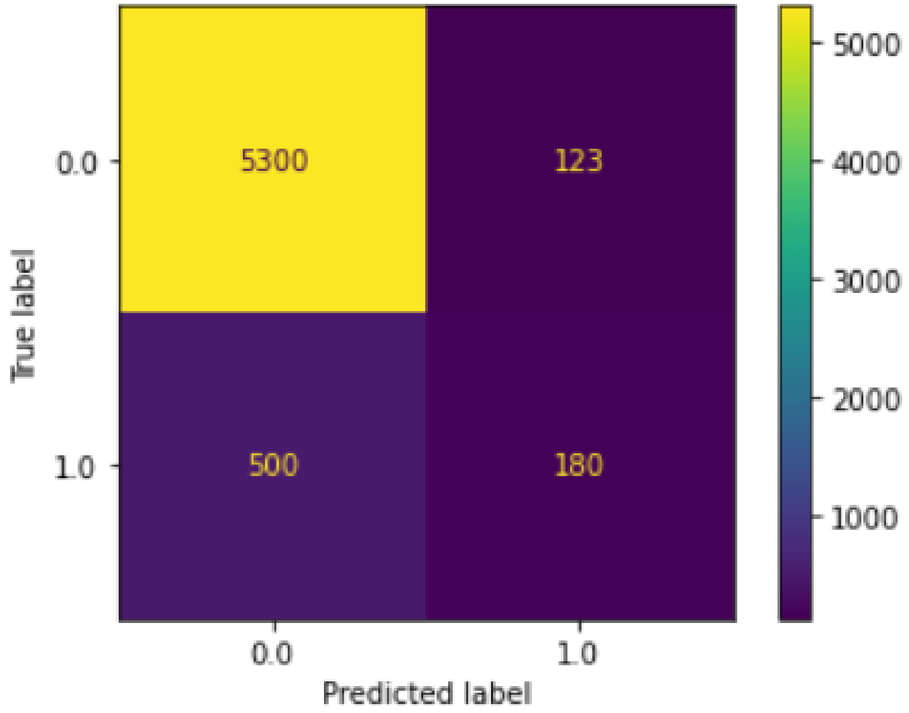
Confusion Matrix for Test set using Logistic Regression classifier:



F1 Score: 0.28

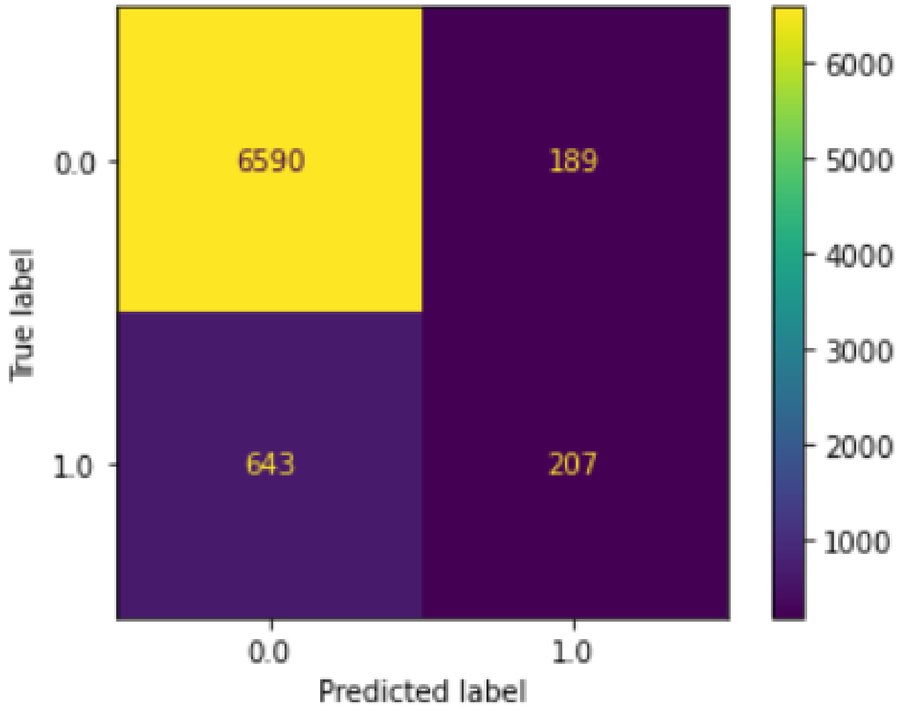
Confusion Matrix: MLP Classifier

Confusion Matrix for Validation set using MLP classifier:



F1 Score: 0.37

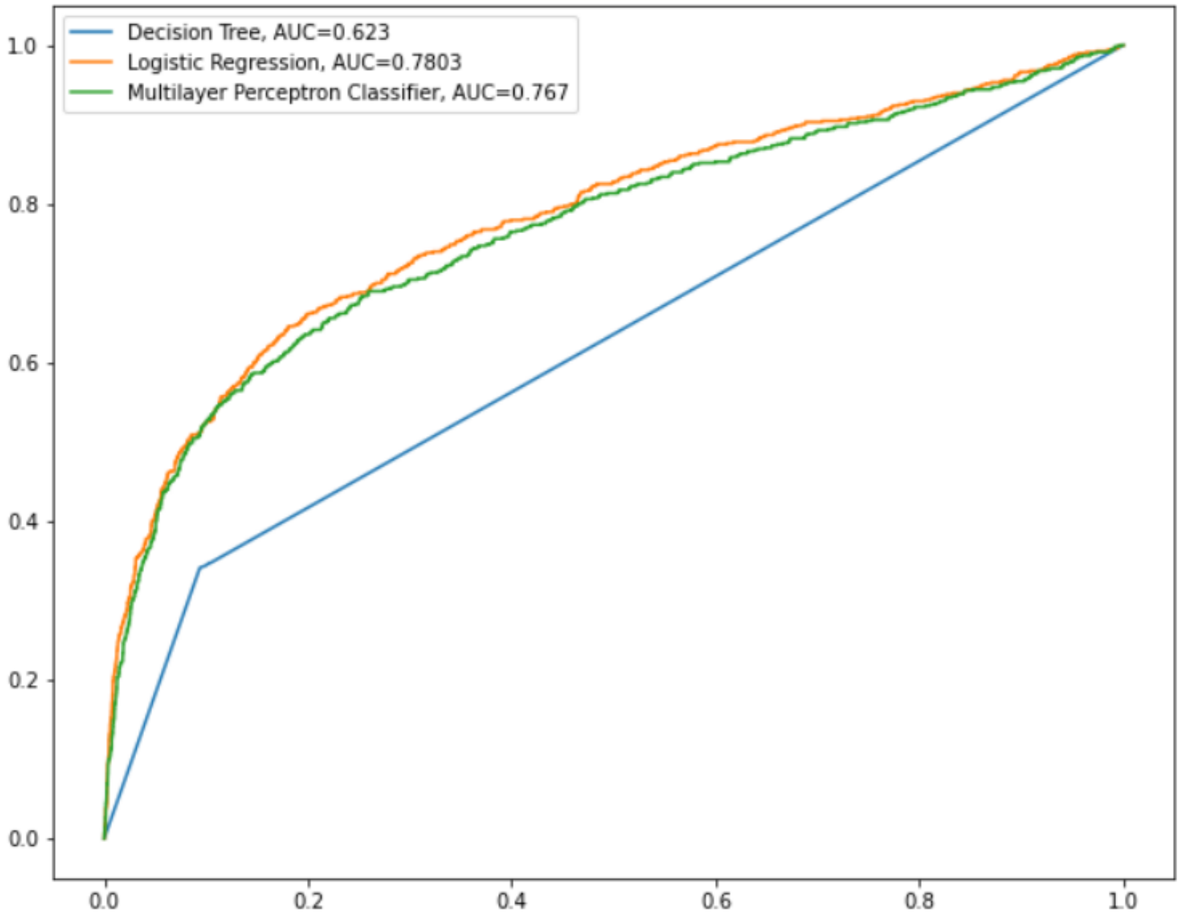
Confusion Matrix for Test set using MLP classifier:



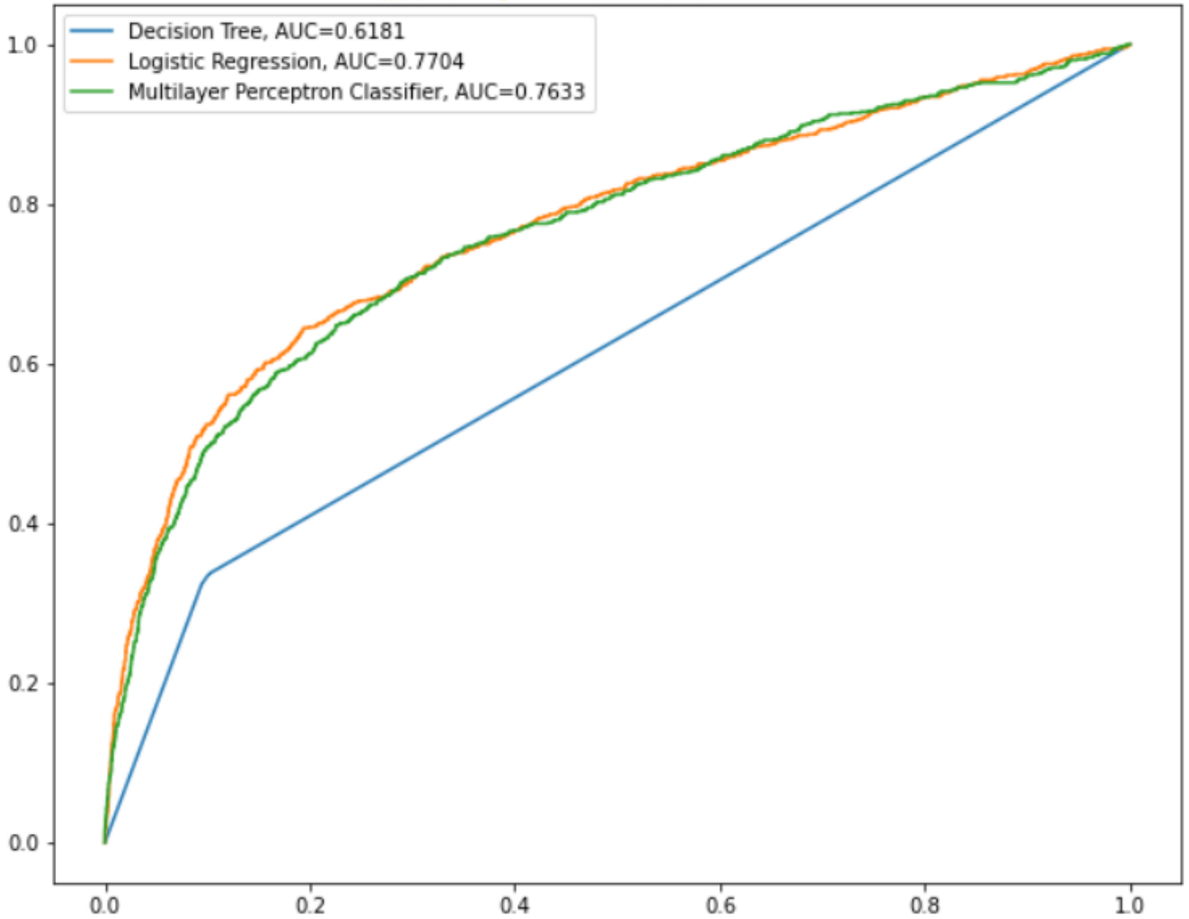
F1 Score: 0.33

ROC Curves

Validation Set

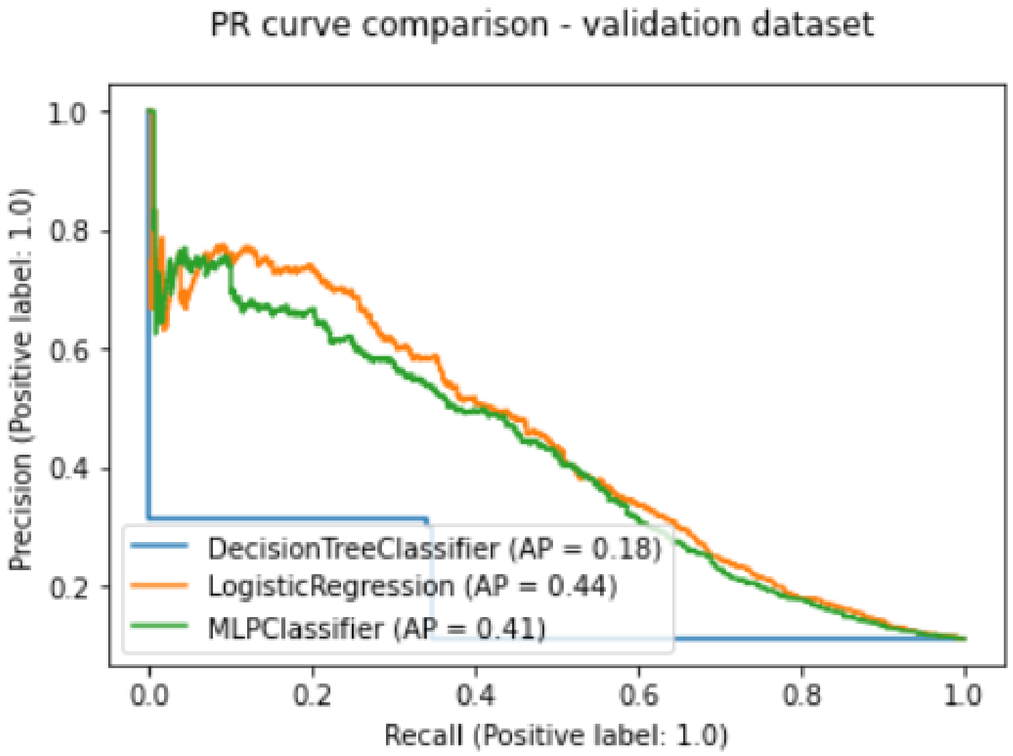


Test Set

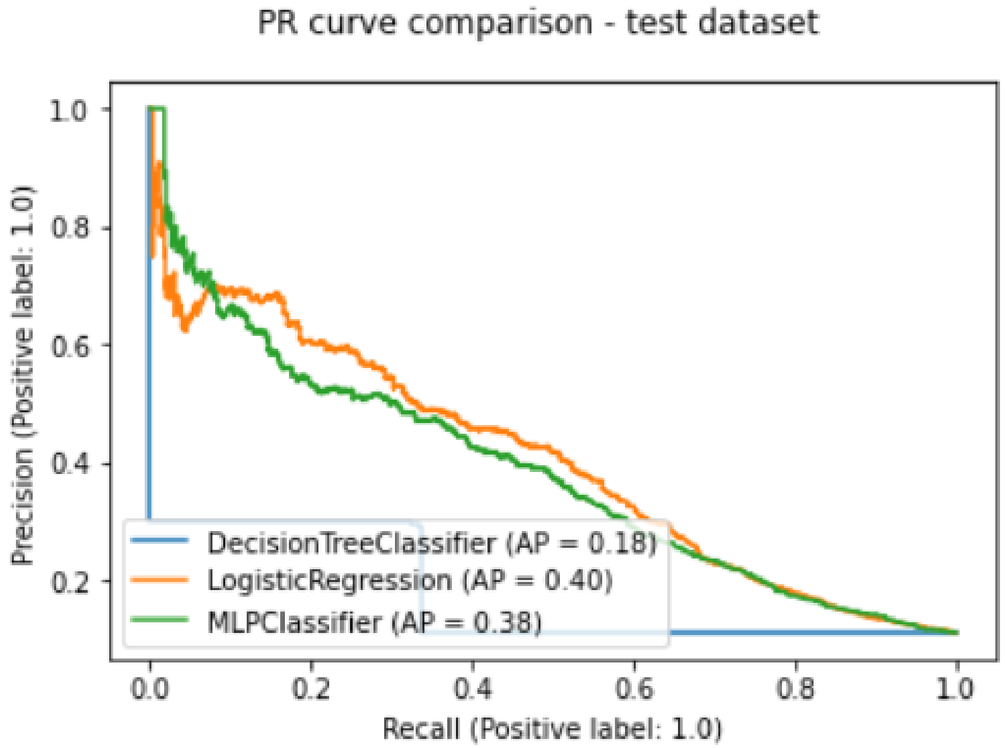


PR Curves

Validation Set



Test Set



The End