BANK MARKETING

Suthatta Dontriros (Fongnam)

Term Deposit

- Term deposits is an interest-bearing bank account and the bank makes a revenue from this.
- Telemarketing campaigns are the one of the high productive way to contact with people.
- However, the bank needs to spend more cost for telemarketing such as labour cost and phone bill for the large call centers investment.
- Hence, it is important to identify the customers most likely to subscribe the campaign to reduce the cost.

Problem Statement

- The goal for this project is to develop the model that predicts the success of a bank marketing campaign based on the features in the dataset from UCI.

- This model should help the bank to identify potential customers who will be interested in the term deposit campaign.

Dataset

41,188 rows 21 columns

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no

Client data

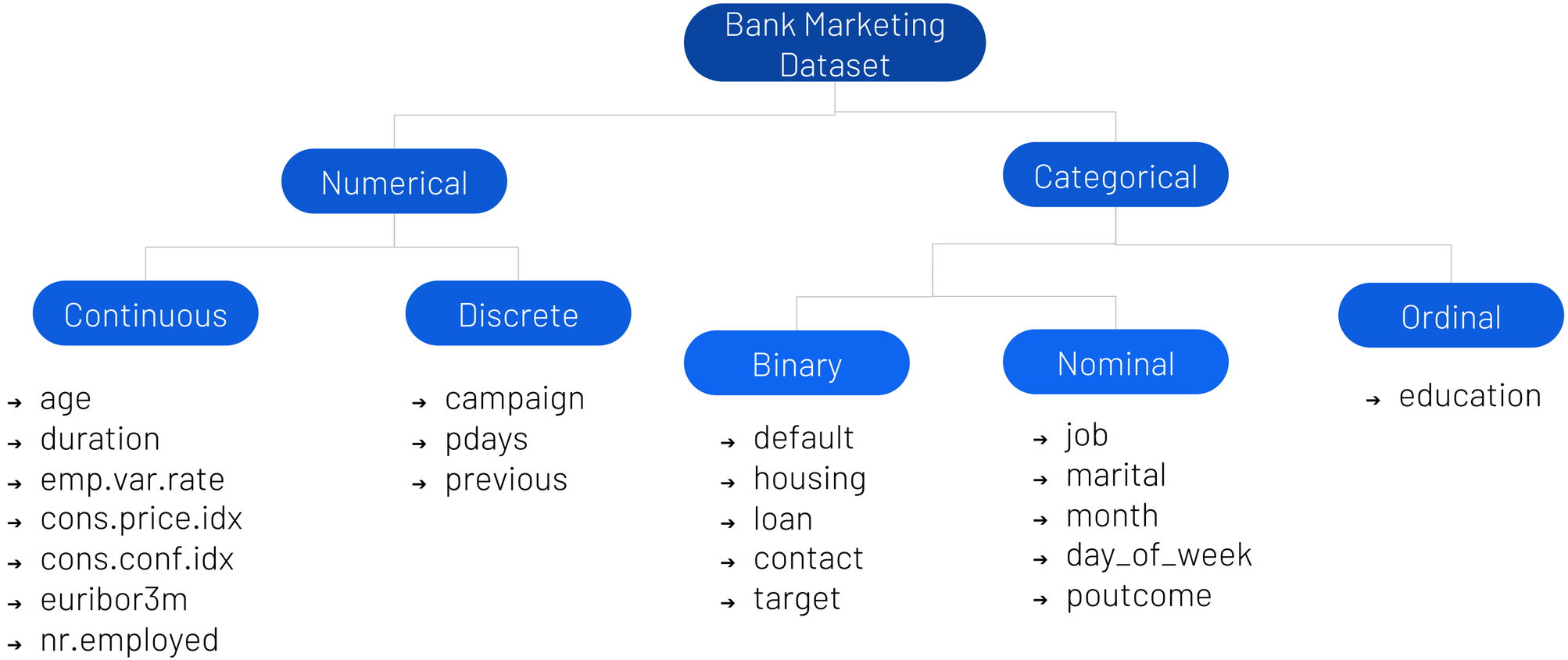
Current Campaign

Previous campaign

Social and economic
Context attributes

Dataset

41,188 rows 21 columns



7 columns

3 columns

5 columns

5 columns

1 column

Work Process

Data Cleaning

EDA

Model

Flask API

Data cleaning

Dealing with unexpected value

- Fill 'unknown’
- Replace 999 by 0

Drop Duplicate rows

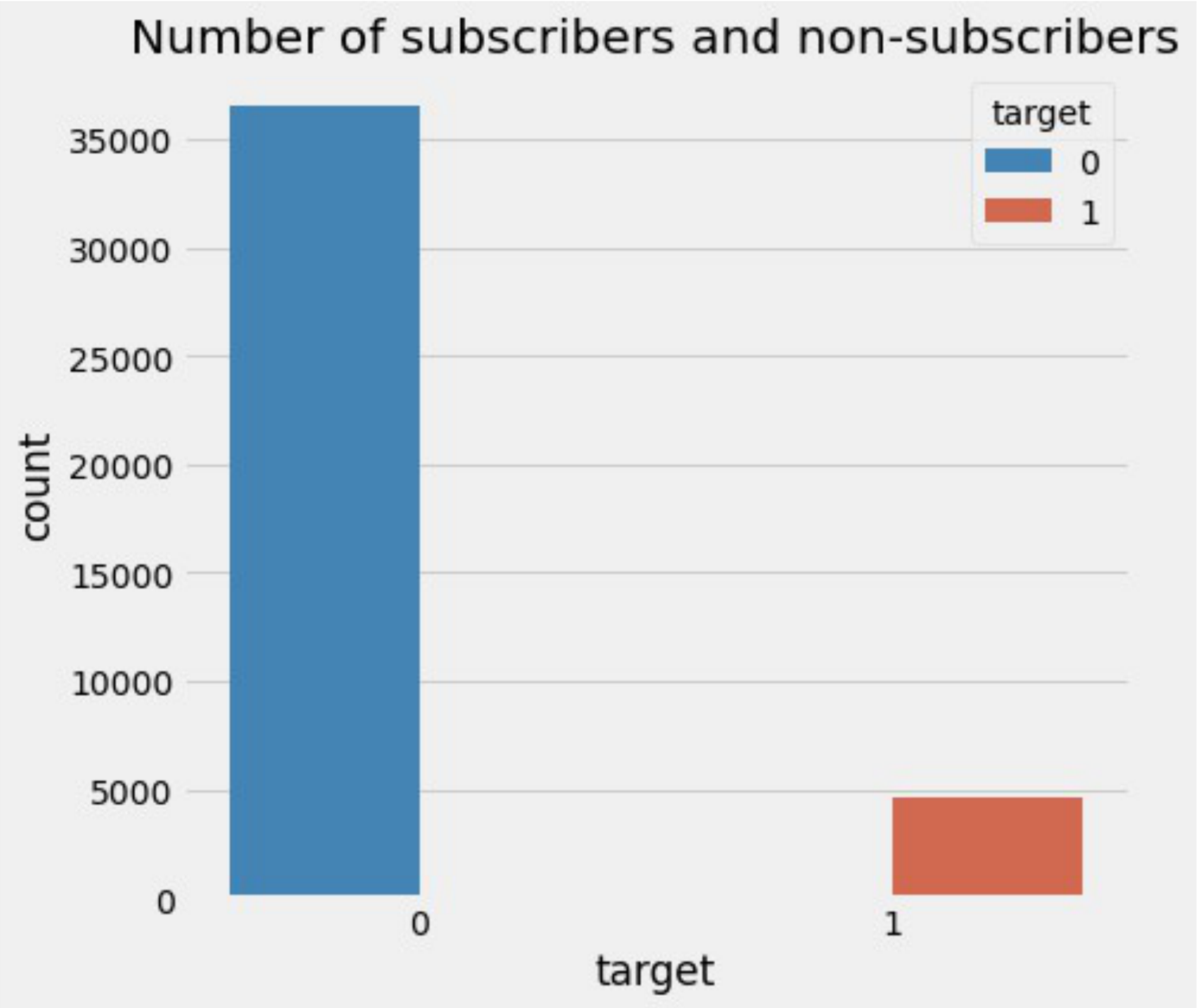
- Drop 11 duplicate rows

Feature Engineering

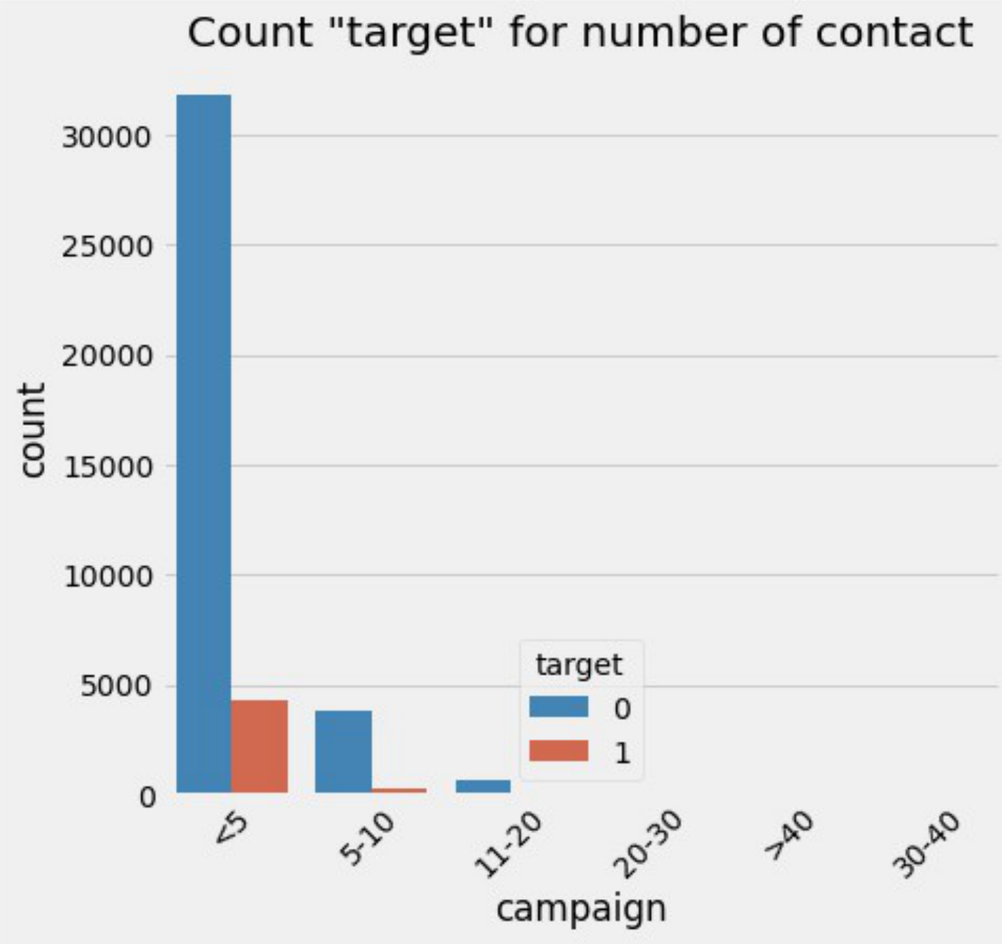
Binary Feature Encoding

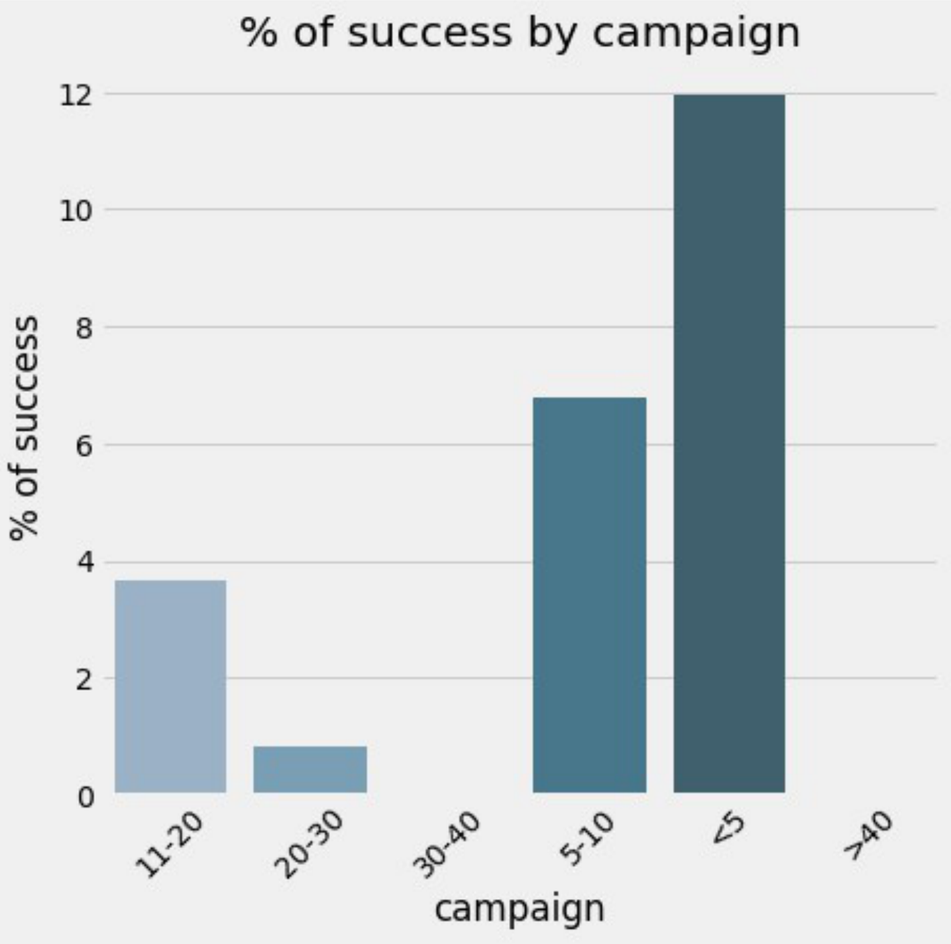
- yes = 1
- no = 0

Data contains only 11% of positive class

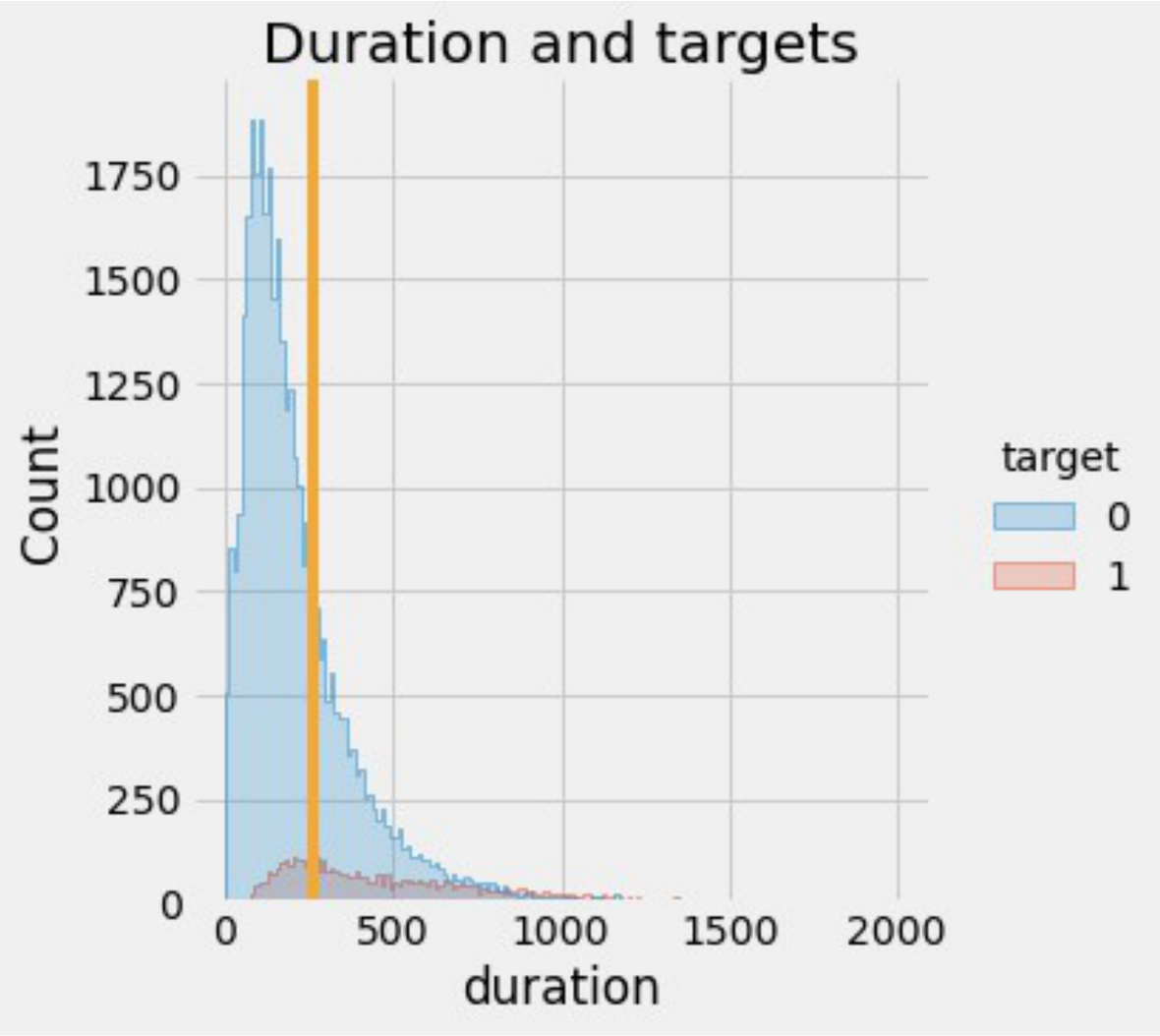


More number of contact, less success rate





Successful case took more time per call

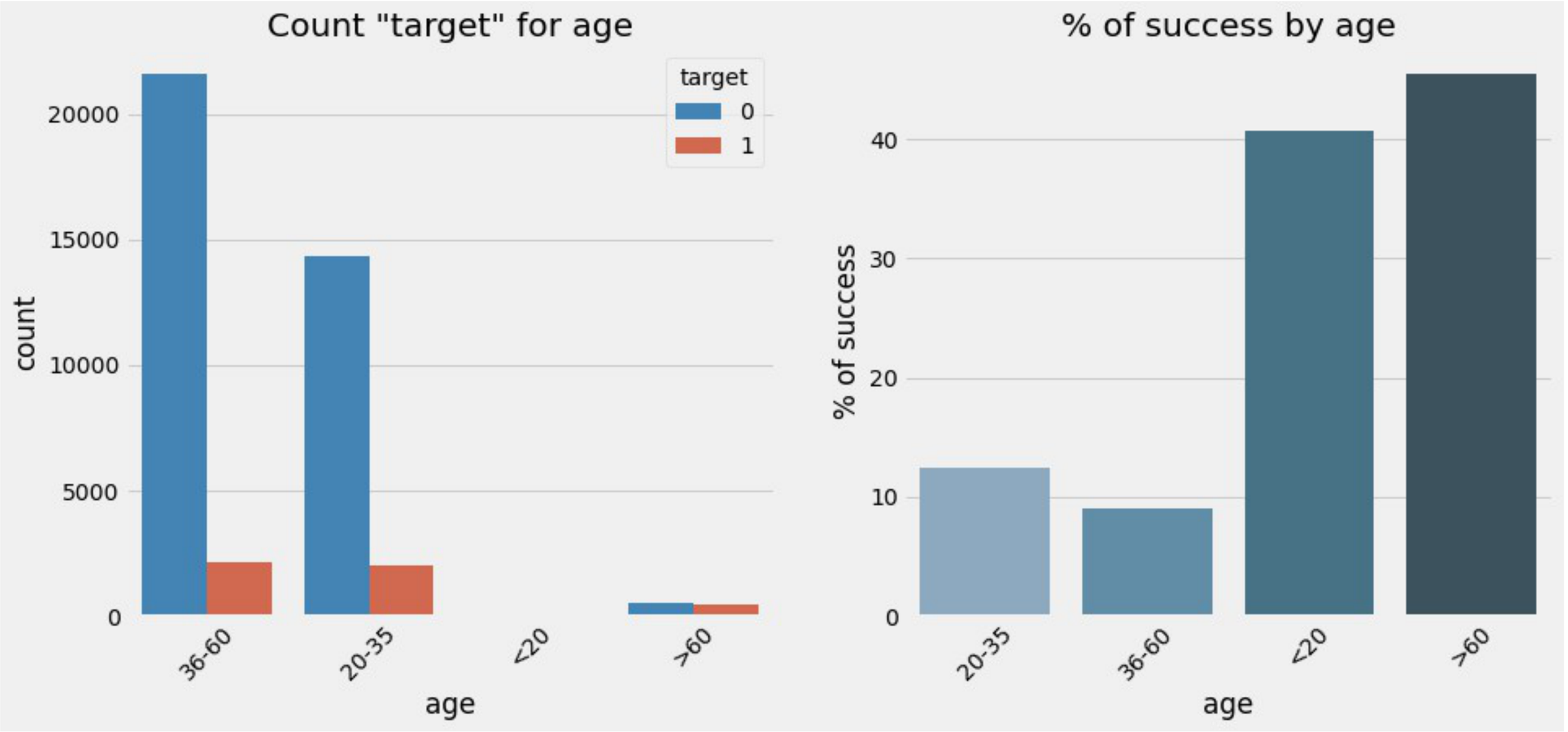


Average duration = 4.3 mins

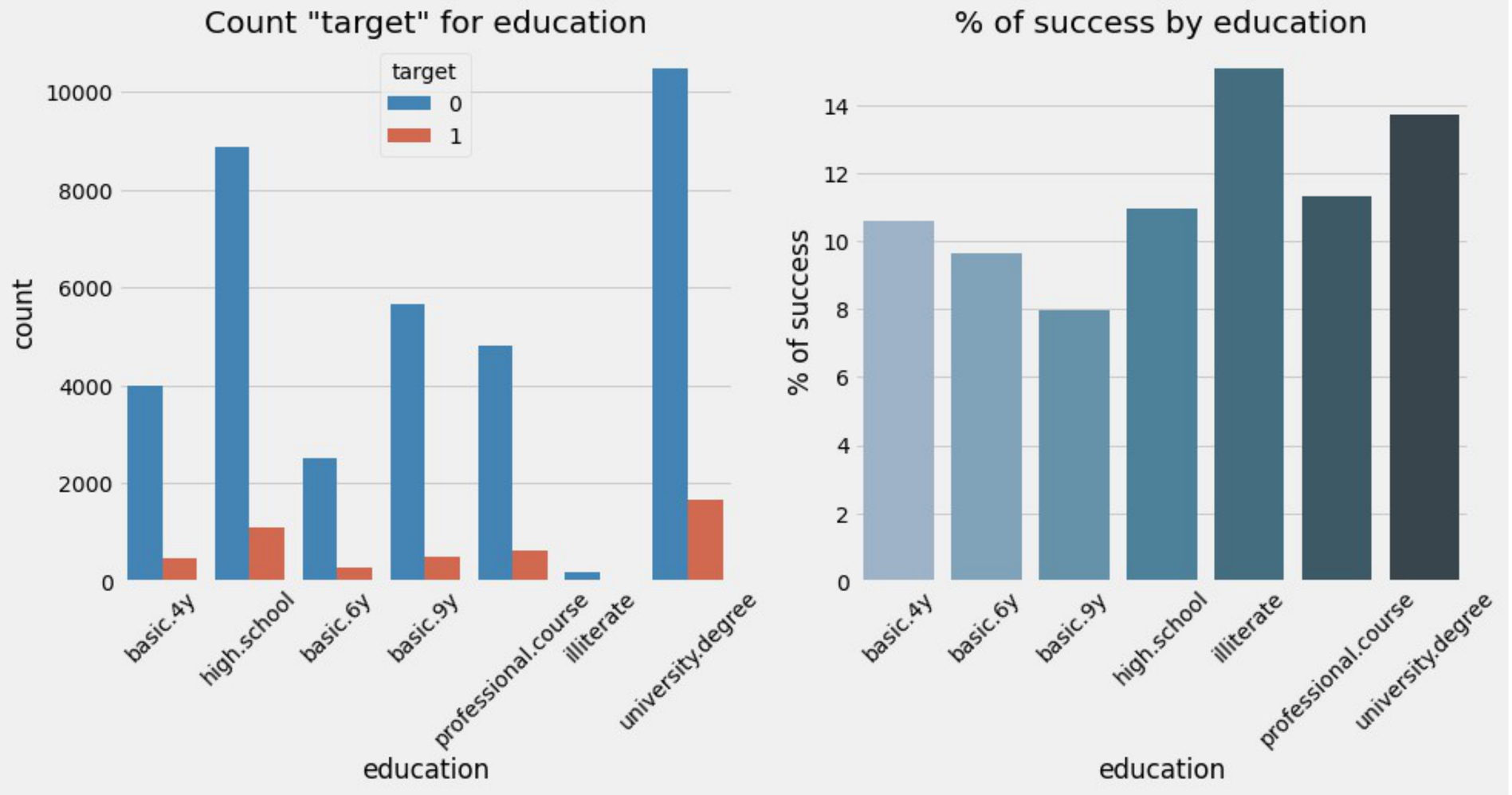
Average duration for successful case = 9.2 mins

Average duration for unsuccessful case = 3.7 mins

Young adults & middle-aged adults have low success rate

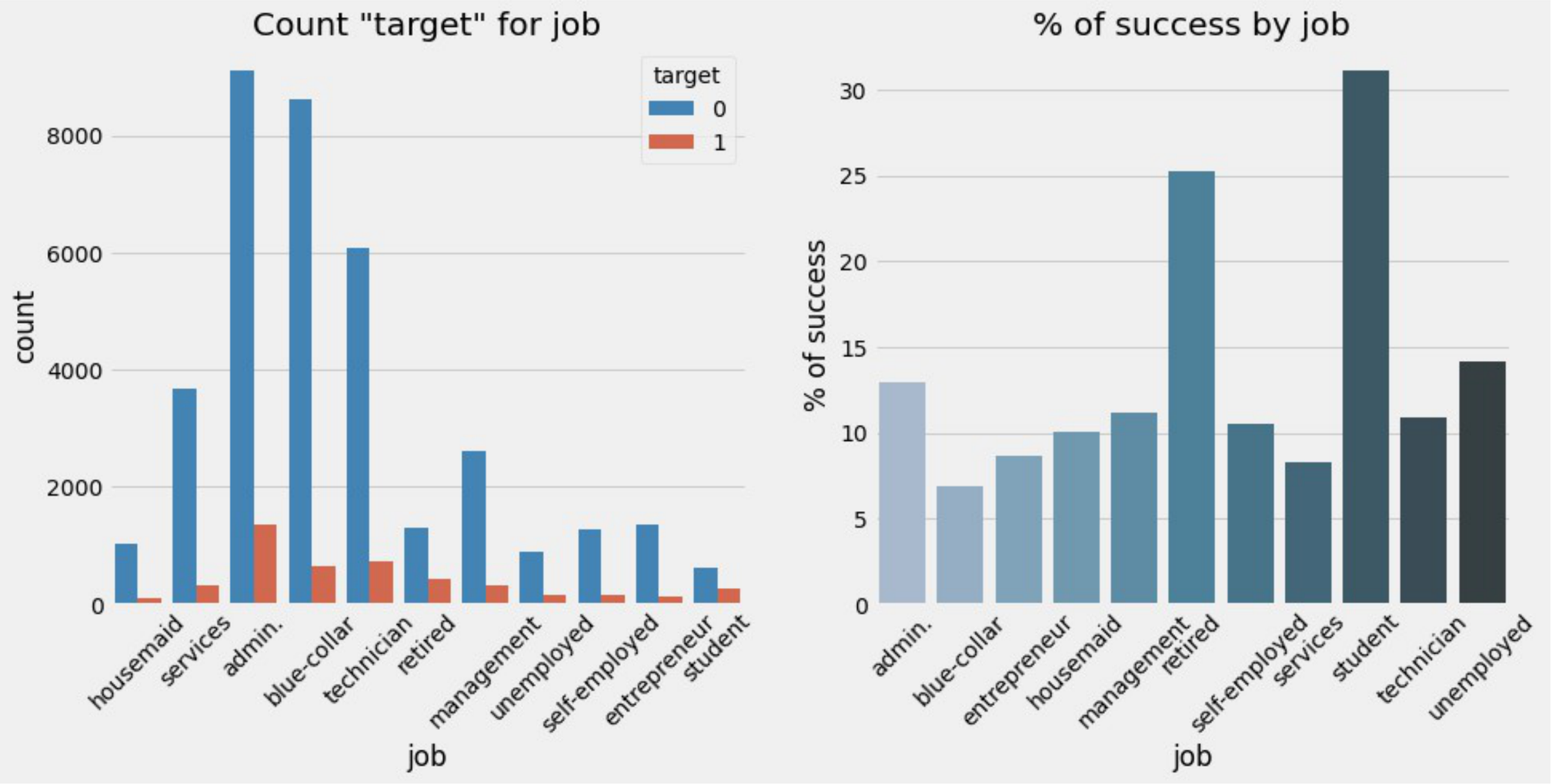


Customer with university degree is the high value customer

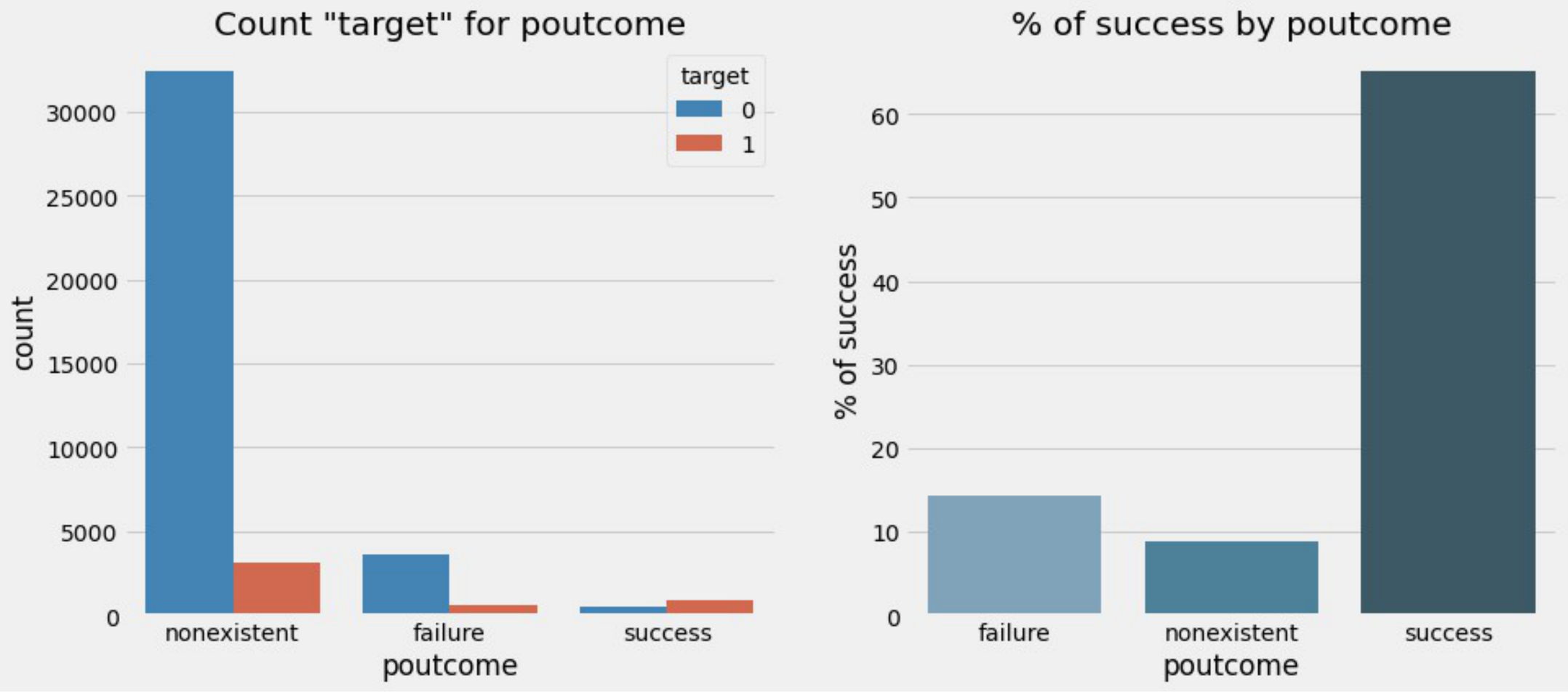


Admin gives the high success case,

but student shows the high success rate



Customers who agreed with the previous campaign tend to subscribe the term deposit



Model

Method 1

Method 2

Method 3

Pipeline

- Classification model
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - XGBoost
- SMOTE

Best model

Model

Method 1

- All features
- Dummy
- SMOTE

Feature Engineering (Method 1)

Dummy

Nominal columns

- Take get_dummies() on 8 columns
  (job, education, contact, month, day_of_week, marital, poutcome, default)

Model

Method 1

- All features
- Dummy

Method 2

- Feature Selection
- Dummy
- More feature engineering

Feature Engineering (Method 2)

Feature Selection

- Drop columns (housing, previous, loan, emp.var.rate)

Dummy Nominal columns

- Take get_dummies() on 7 columns (job, contact, marital, poutcome, default, day_of_week, month)

Remove Outliers

- Remove outliers in column named campaign by IQR

Ordinal Feature Encoding

- Convert string to integer in education column

Model

Method 1

- All features
- Dummy
- SMOTE

Method 2

- Feature Selection
- Dummy
- More feature engineering

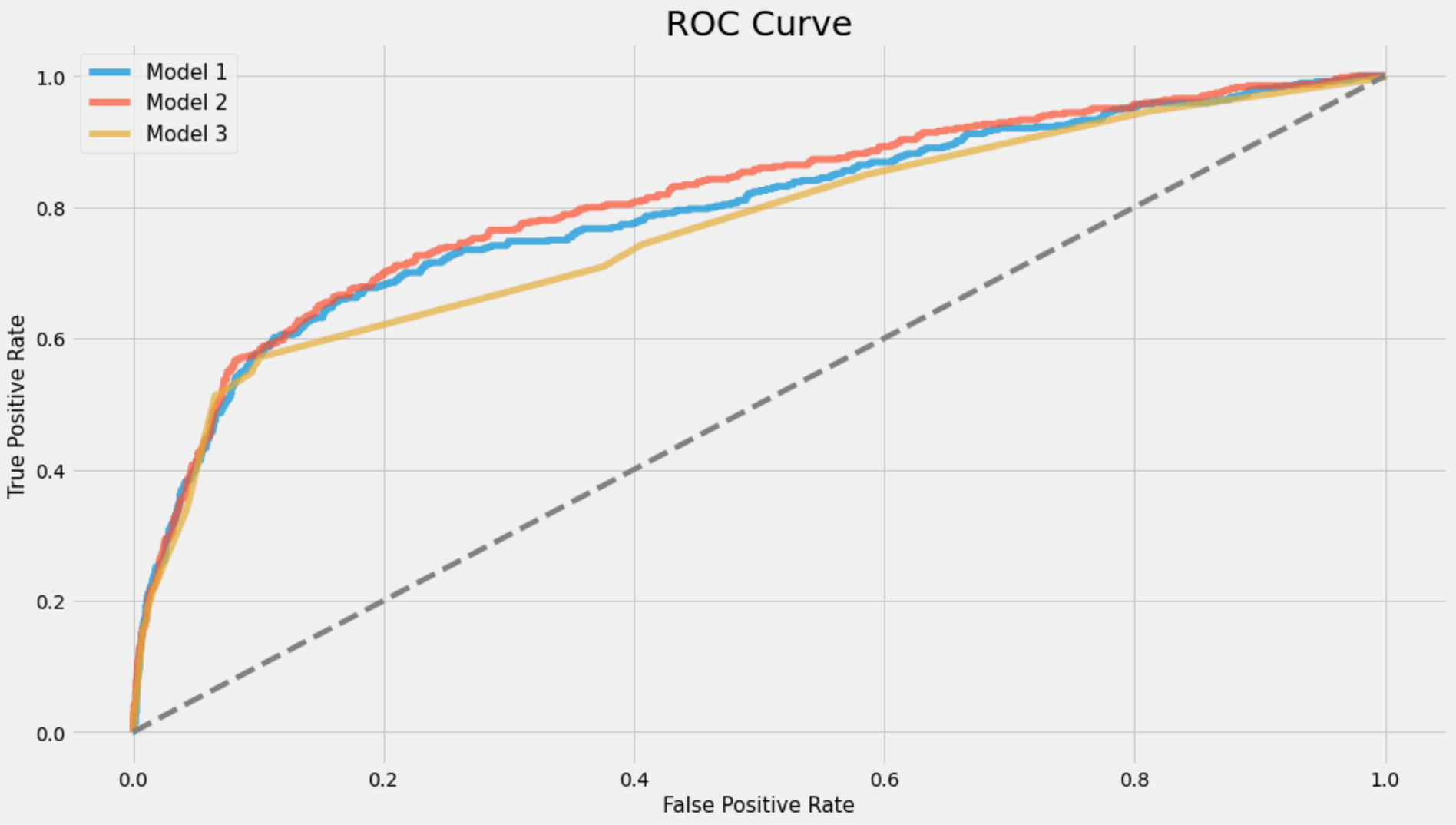
Method 3

- Feature Selection
- Dummy
- Feature engineering
- PCA

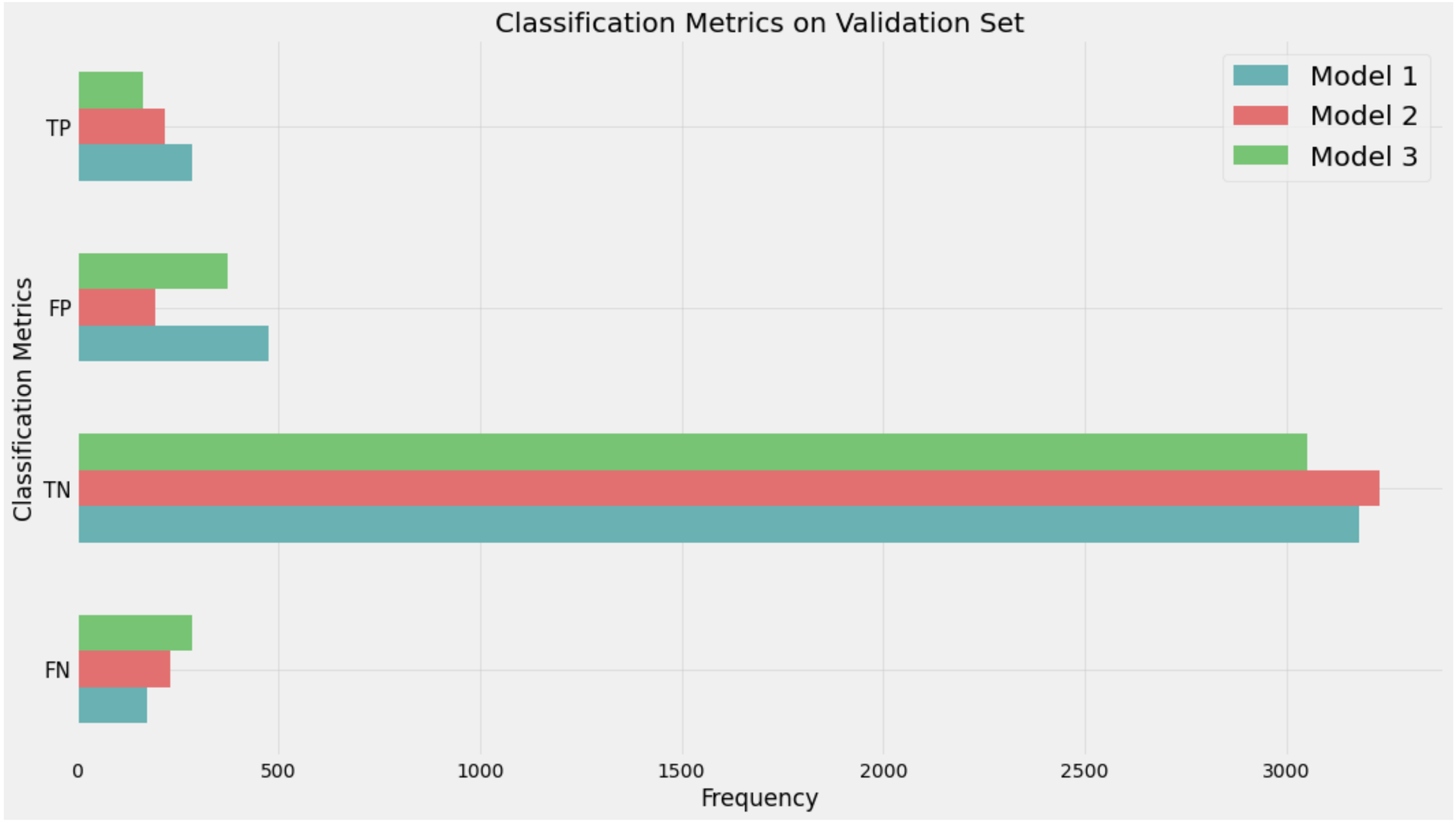
Model

	Method 1	Method 2	Method 3
Best Model	Random Forest	XGBoost	Logistic Regression
Training ROCAUC score	0.795	0.799	0.791
CV ROCAUC score	0.790	0.789	0.788
Validation ROCAUC score	0.808	0.815	0.815
Best Parameters	'rf__max_depth': 5 'rf__min_samples_split': 4 'rf__n_estimators': 50 'sm__k_neighbors': 5	'sm__kneighbors': 5 'xg__learning_rate': 0.1 'xg__max_depth': 5 'xg__min_child_weight': 4	'lr__C': 0.01 'lr__penalty': 'l2' 'sm__k_neighbors': 3

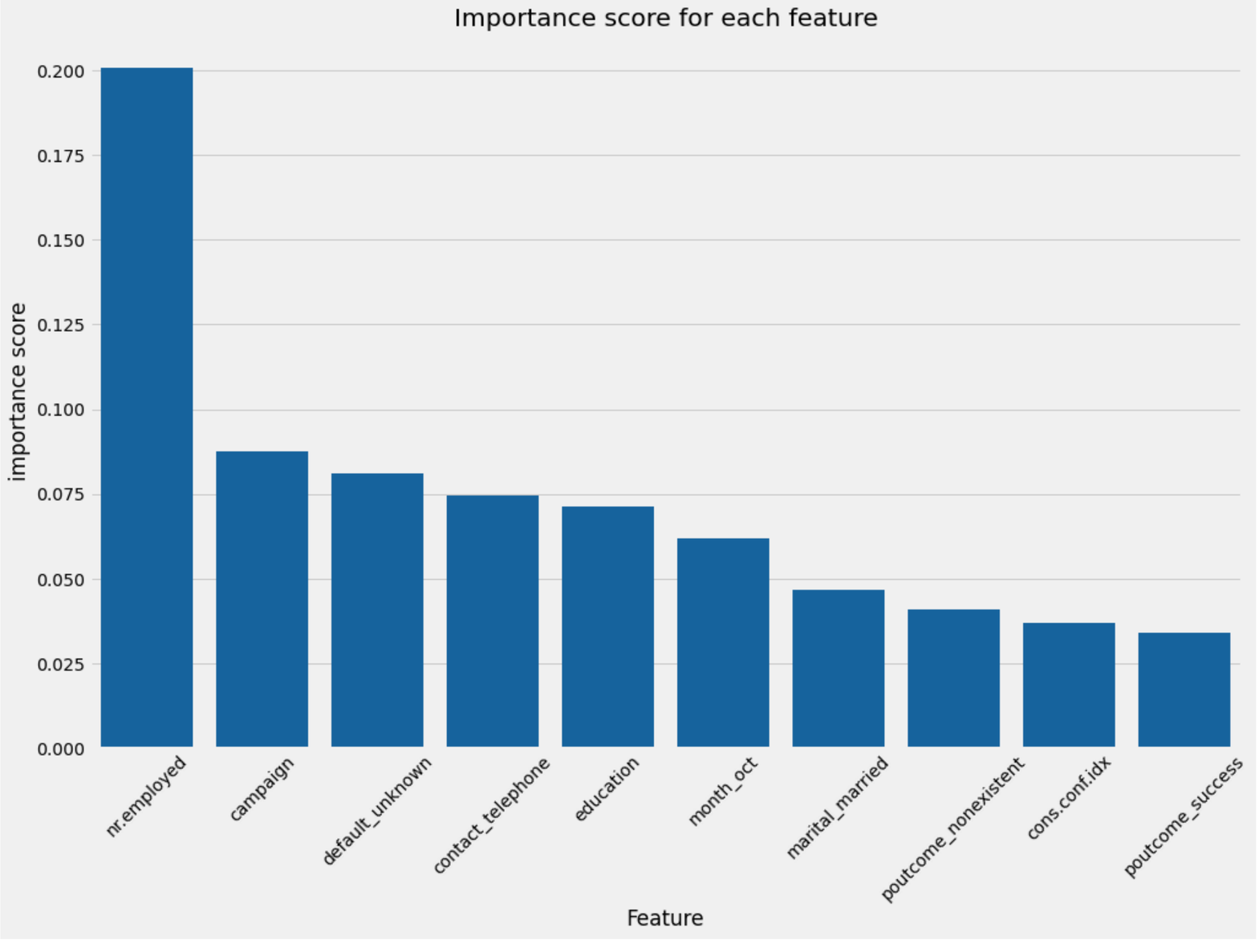
Model


ROC Curve
Model 1
Model 2
Model 3
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

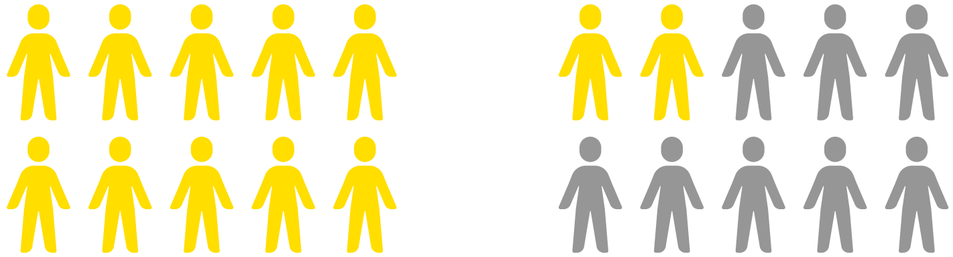
Model


Classification Metrics on Validation Set
Model 1
Model 2
Model 3
TP
FP
TN
FN
Classification Metrics
0
500
1000
1500
2000
2500
3000
Frequency

Feature Importance



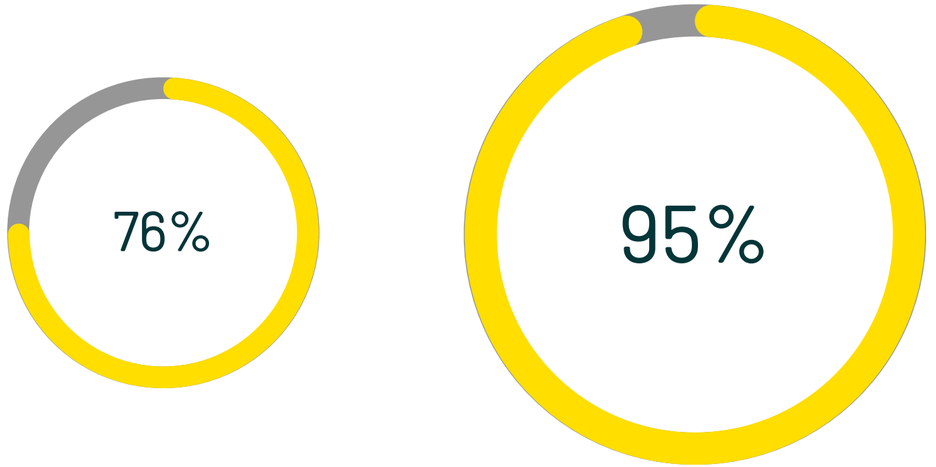
Cost-Benefit Analysis



$1,074,000 **$147,095**

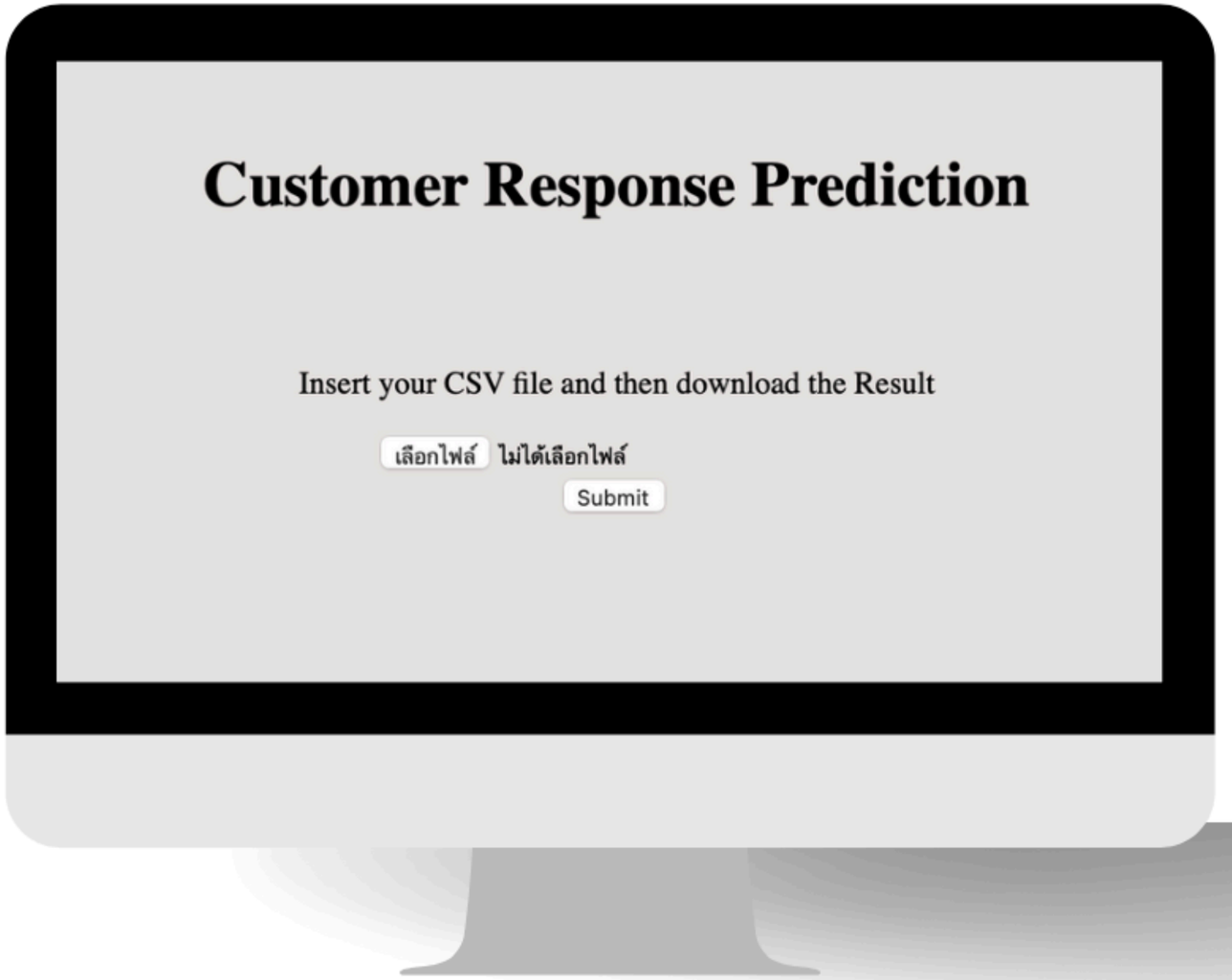
Reduce cost

* Model lets the bank miss some successful cases

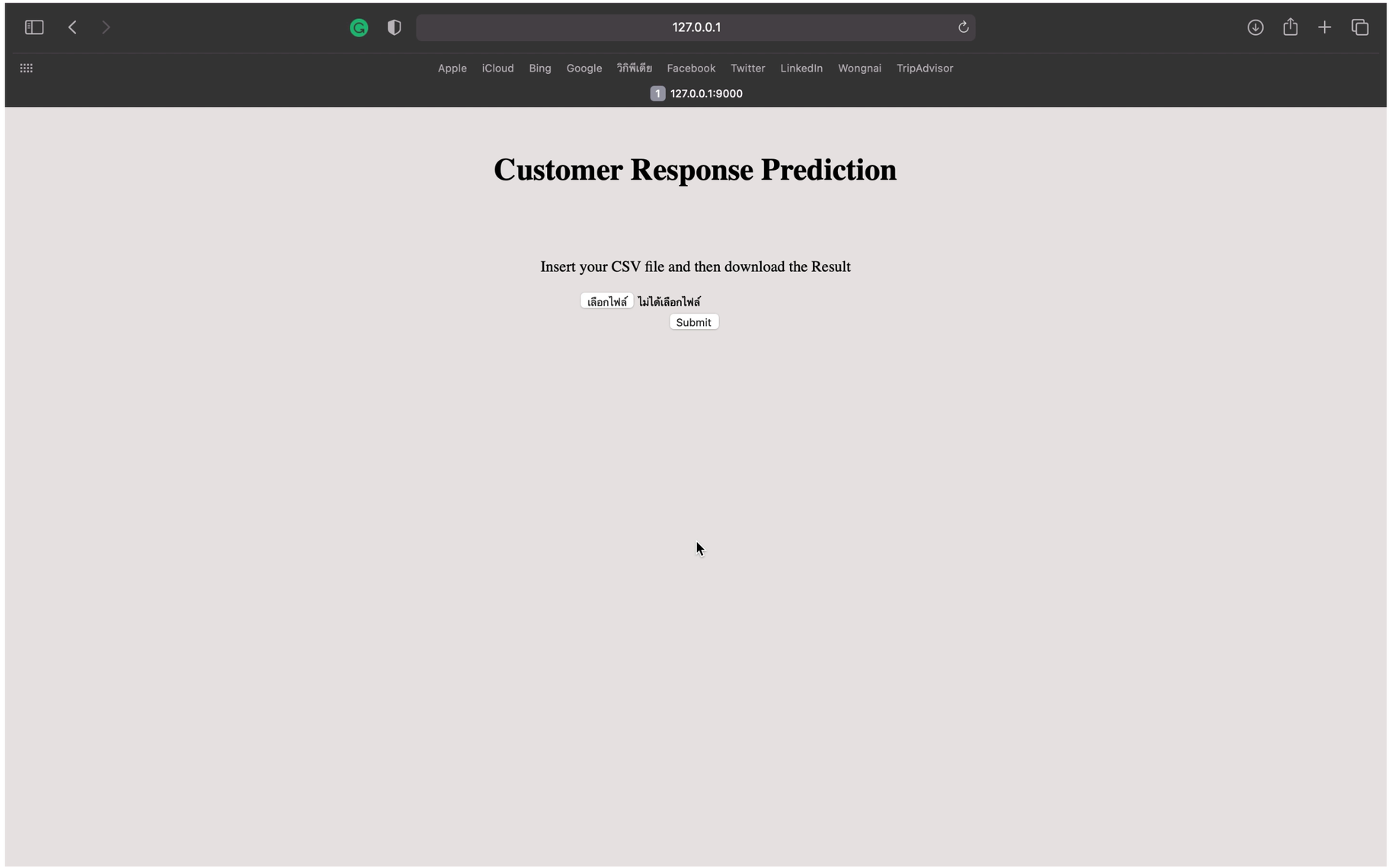


Increase % of Profit

Deploy model using Flask API



Deploy model by Flask API



Conclusion

- The best performing model is XGBoost with the ROCAUC score at 0.815.
- Number of employees was the top feature importance for tree splitting.
- This model should reduce the cost around 80% and increase the percent of profit from 76% to 95%. However, this model had some FN that means the bank will miss up some customers.

Next steps

Further improve the model:

- more data collection
- more data about the customer such as incomes, account balance and location
- Try to build Deep learning model

Recommendations

Digital marketing can increase the success rate.

Thank you

Q&A