Data Glacier

Your Deep Learning Partner

Project: Bank Marketing Campaign

Name: Najma Abdi,Leena Garnta,Adama Sal

Batch Code: LISUM45

Specialization: Data Science

Agenda

Problem Statement

Business Understanding

Objective

Dataset

EDA

Recommendation

Model building

Evaluation

Results

Problem Statement

ABC Bank is planning to launch a new term deposit product.Before investing in a large-scale marketing campaign,the bank wants to use machine learning model to identify customers who are highly likely to subscribe ,based on their previous interactions with marketing campaigns

Business Understanding

- Predicting marketing campaign outcomes and identifying key influencing features helps the organization run more efficient campaigns. By segmenting customers likely to subscribe to term deposits, machine learning models can highlight high-potential customers, allowing marketing channels (telemarketing, SMS/email, etc.) to focus efforts effectively—saving both time and resources.

Objective

- Build a Classification ML model to shortlist customers who are most likely to buy the term deposit product. This would allow the marketing team to target those customers through various channels.

Dataset

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

bank-additional-full.csv: 20 inputs (+1 target variable) and 41118 observations

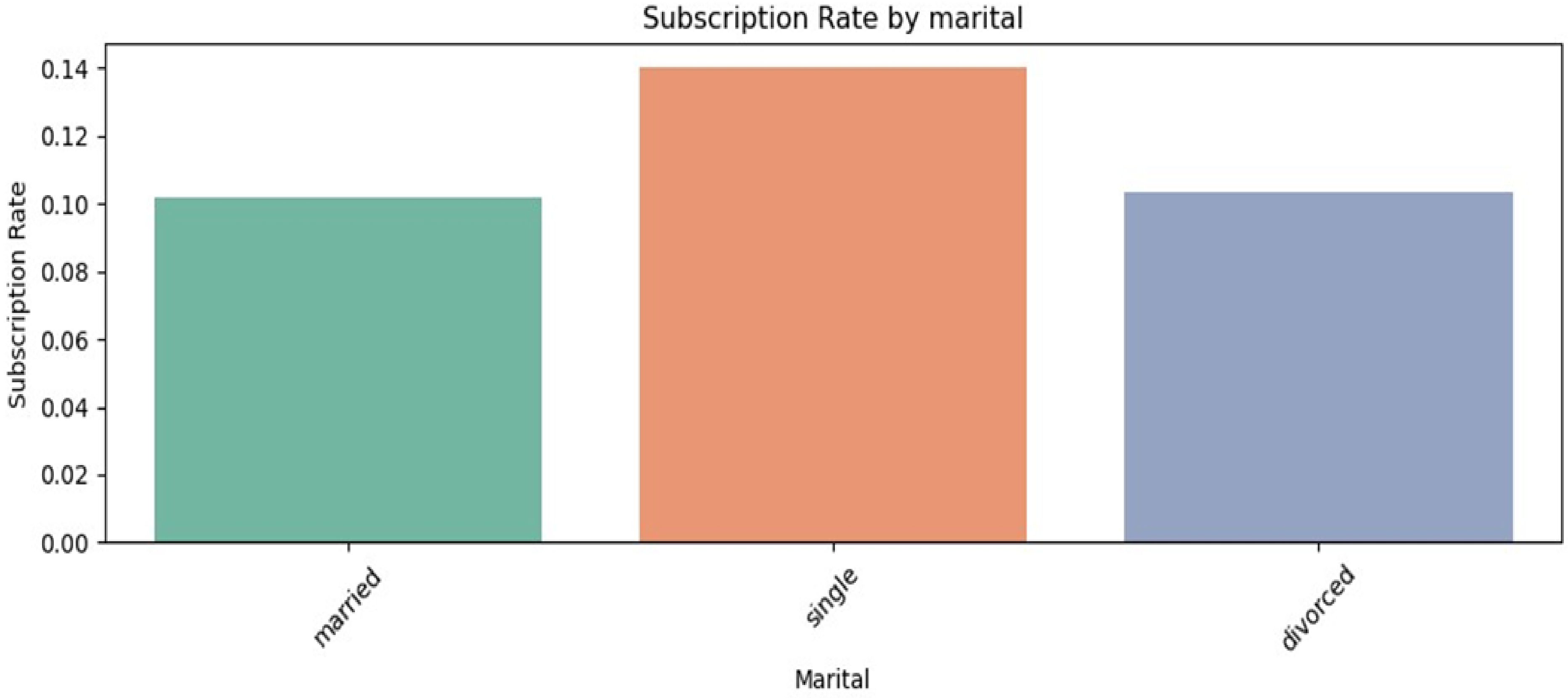
Assumptions:

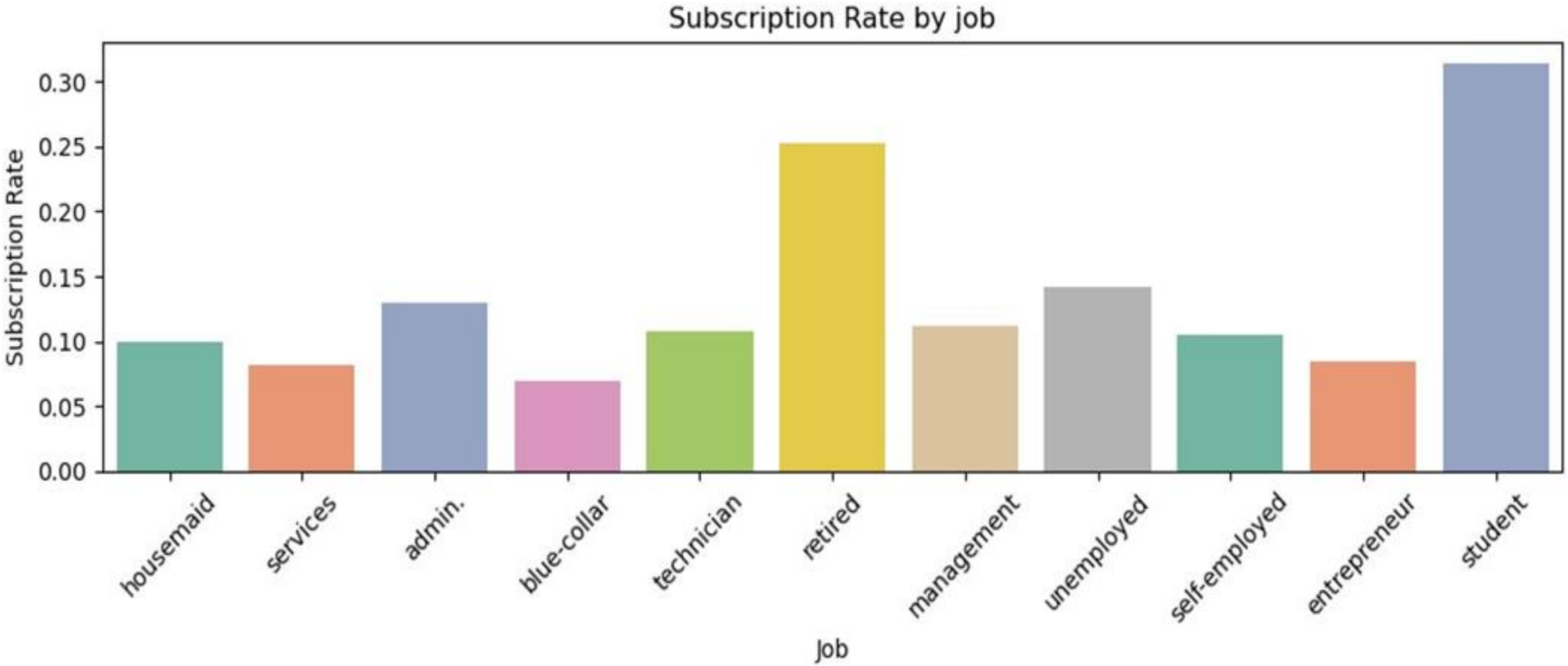
- ‘Duration’ feature is dropped as suggested in the dataset description
- A frequently occurring missing value ‘unknown’ is considered as another category for the categorical features.
- Duplicate rows were deleted from the dataset.

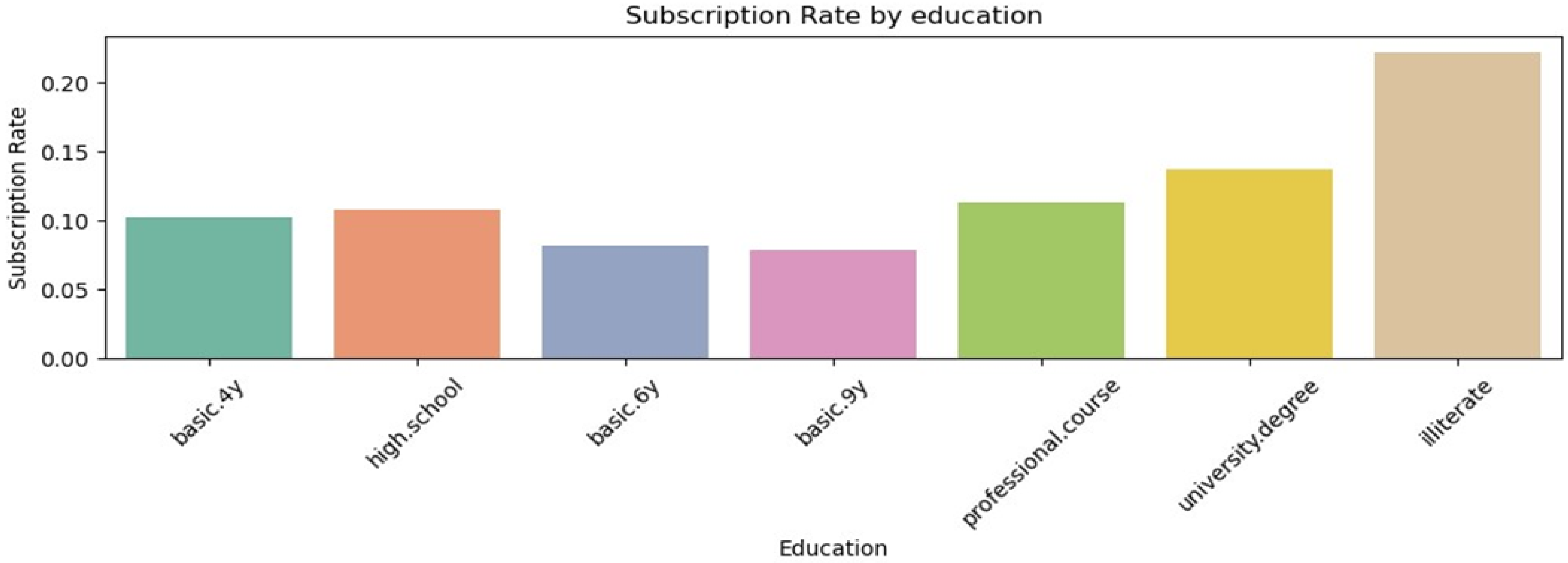
EDA

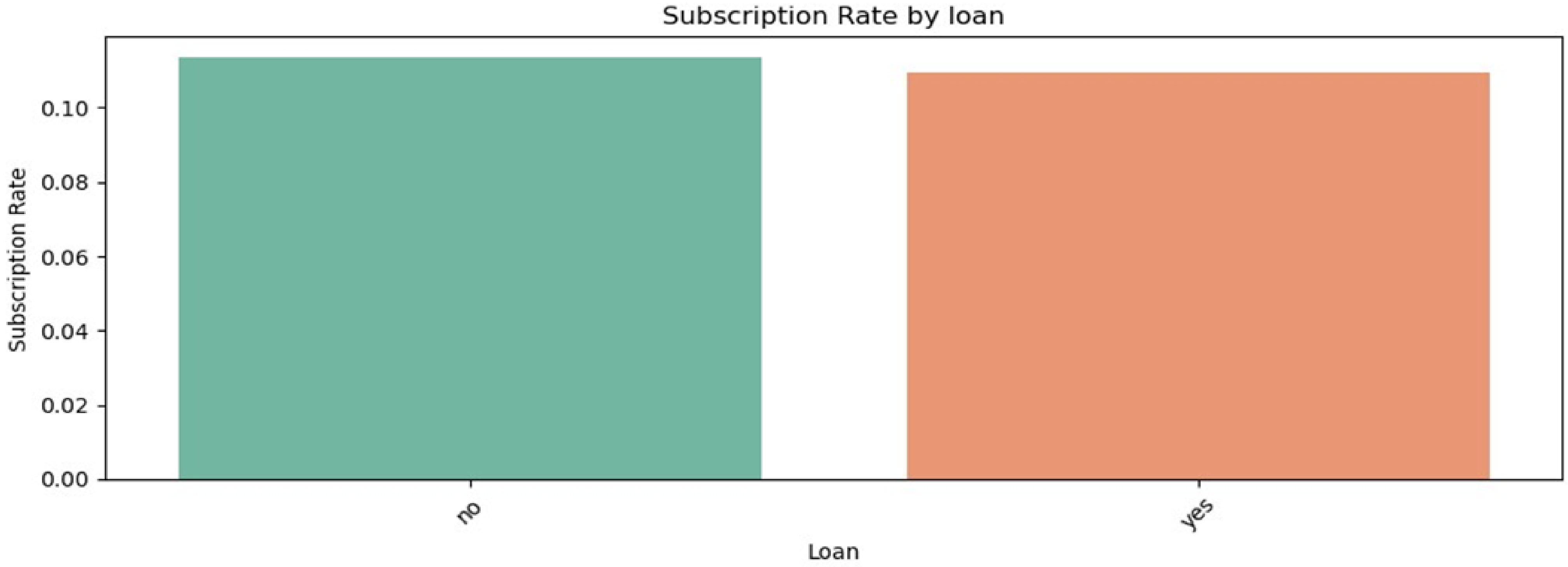
EDA
1. Univariate Analysis

- In Jobs, most of the customers are administrative staffs and technicians.
- Married customers have been communicated more for subscription to the deposit more compared to single

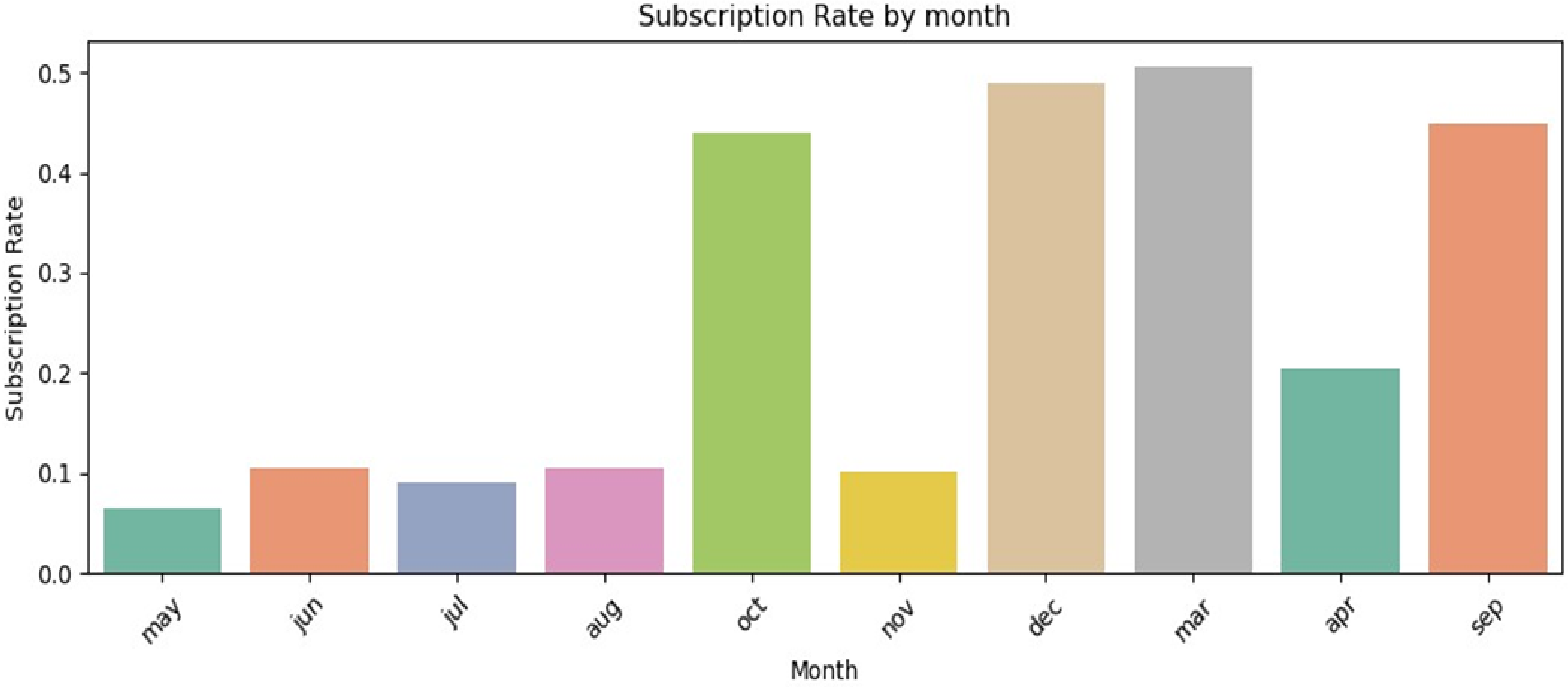


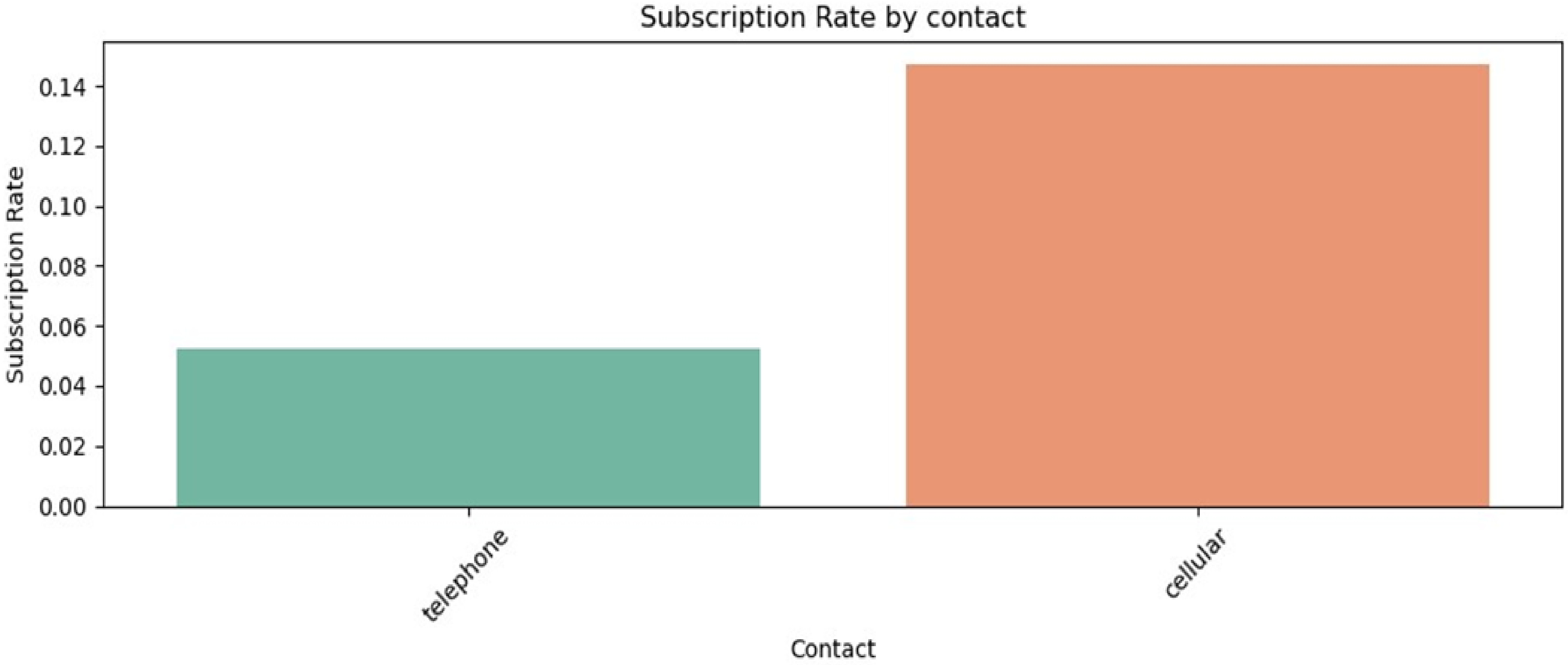

Subscription Rate by job
Subscription Rate
0.30
0.25
0.20
0.15
0.10
0.05
0.00
housemaid
services
admin.
blue-collar
technician
retired
management
unemployed
self-employed
entrepreneur
student
Job


Subscription Rate by education
Subscription Rate
0.00
0.05
0.10
0.15
0.20
basic.4y
high.school
basic.6y
basic.9y
professional.course
university.degree
illiterate
Education

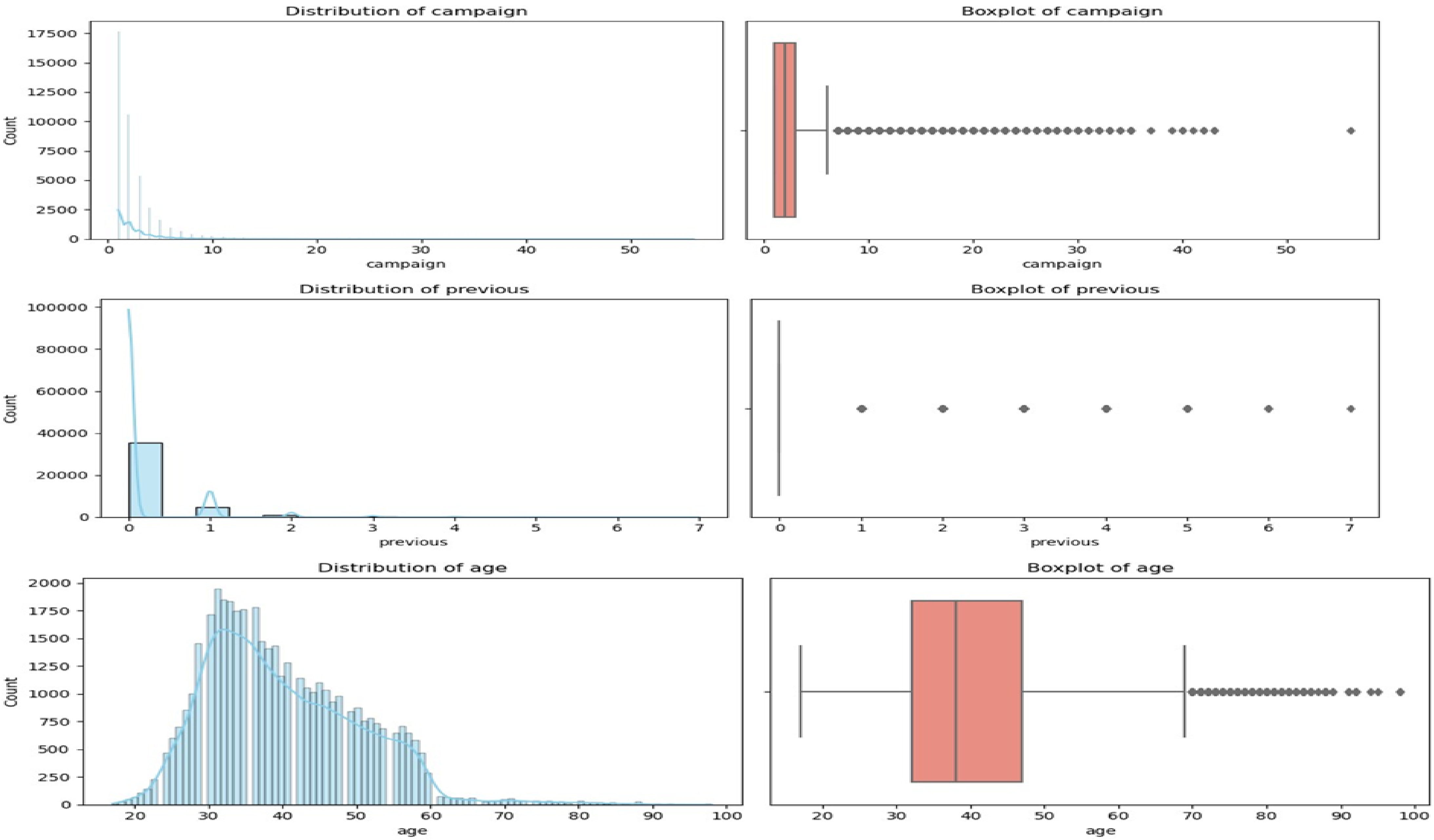


- Most of the customers are university graduates followed by high school degree
- Majority of the customers do not have a personal loan


Subscription Rate by month
Subscription Rate
0.0
0.1
0.2
0.3
0.4
0.5
may
jun
jul
aug
oct
nov
dec
mar
apr
sep
Month


Subscription Rate by contact
Subscription Rate
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
telephone
cellular
Contact

- The type of communication used to contact customers is mostly cellular compared to telephone
- May seems to be the month with most contacts made



Distribution of campaign

Boxplot of campaign

Distribution of previous

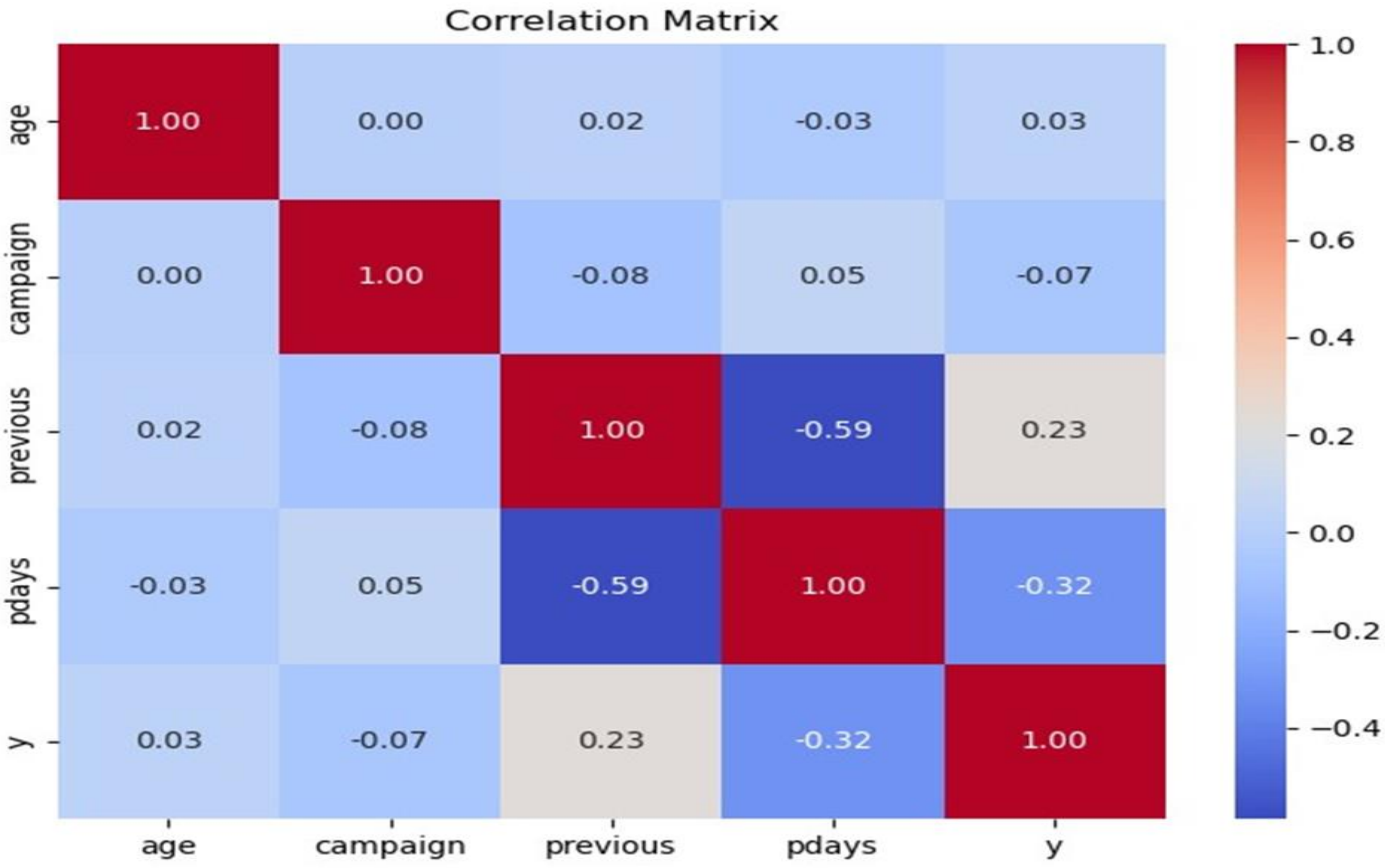
Boxplot of previous

Distribution of age

Boxplot of age

- The age group of customers contacted mostly fall between 20 to 60
- Number of follow ups made to a customer is less in the previous campaign

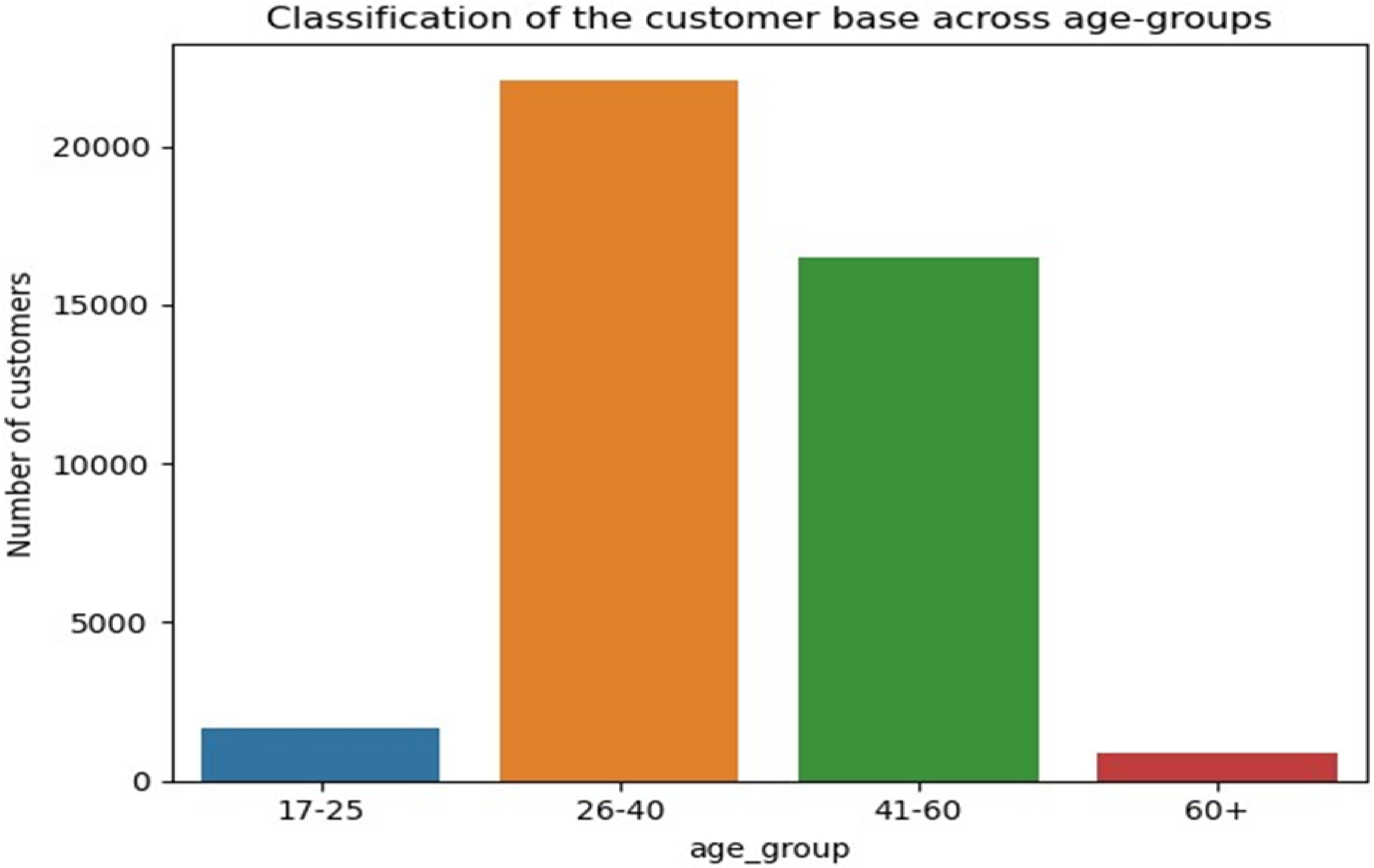
Correlation map



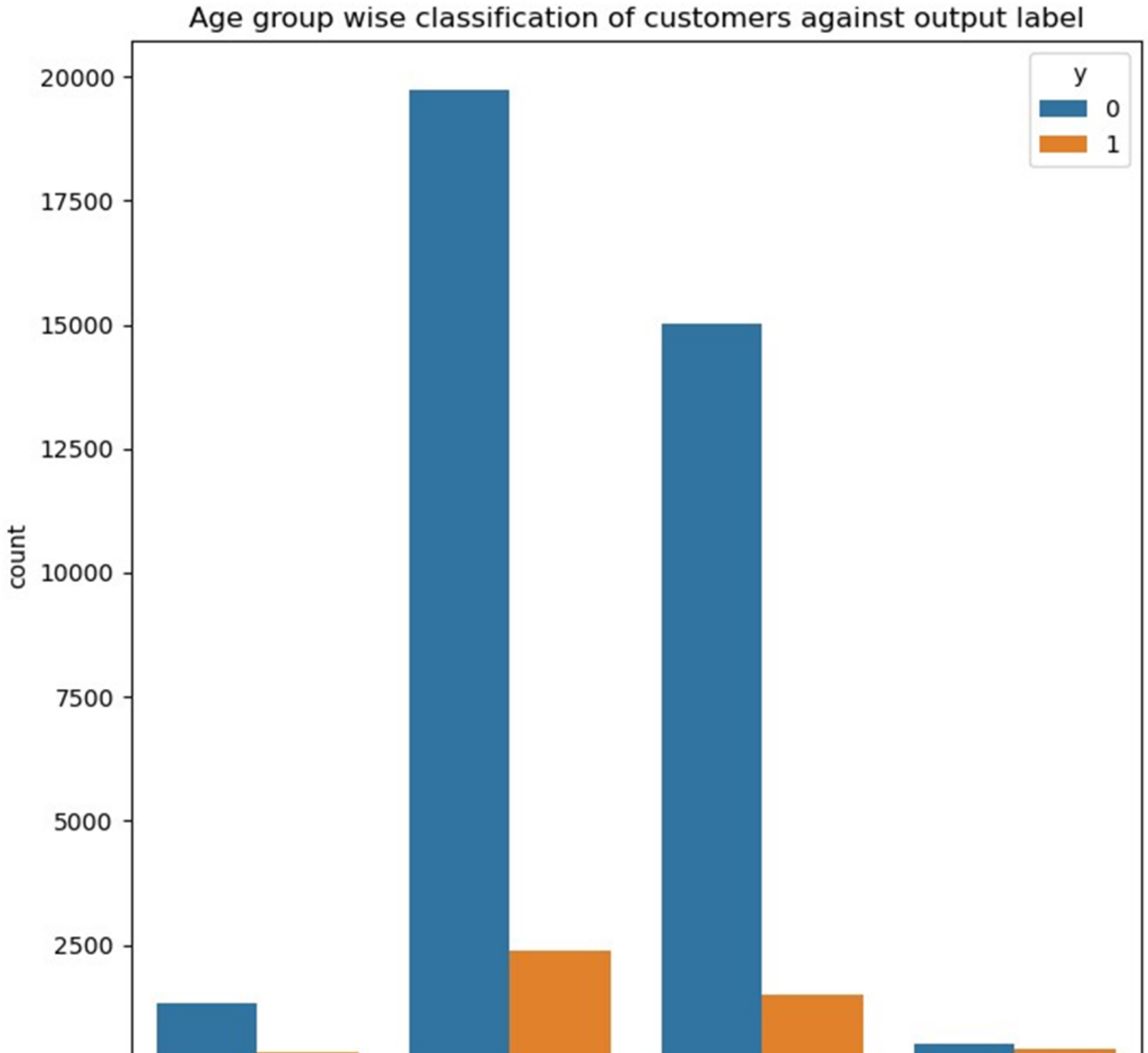
	age	campaign	previous	pdays	y
age	1.00	0.00	0.02	-0.03	0.03
campaign	0.00	1.00	-0.08	0.05	-0.07
previous	0.02	-0.08	1.00	-0.59	0.23
pdays	-0.03	0.05	-0.59	1.00	-0.32
y	0.03	-0.07	0.23	-0.32	1.00

EDA
2. Bivariate Analysis

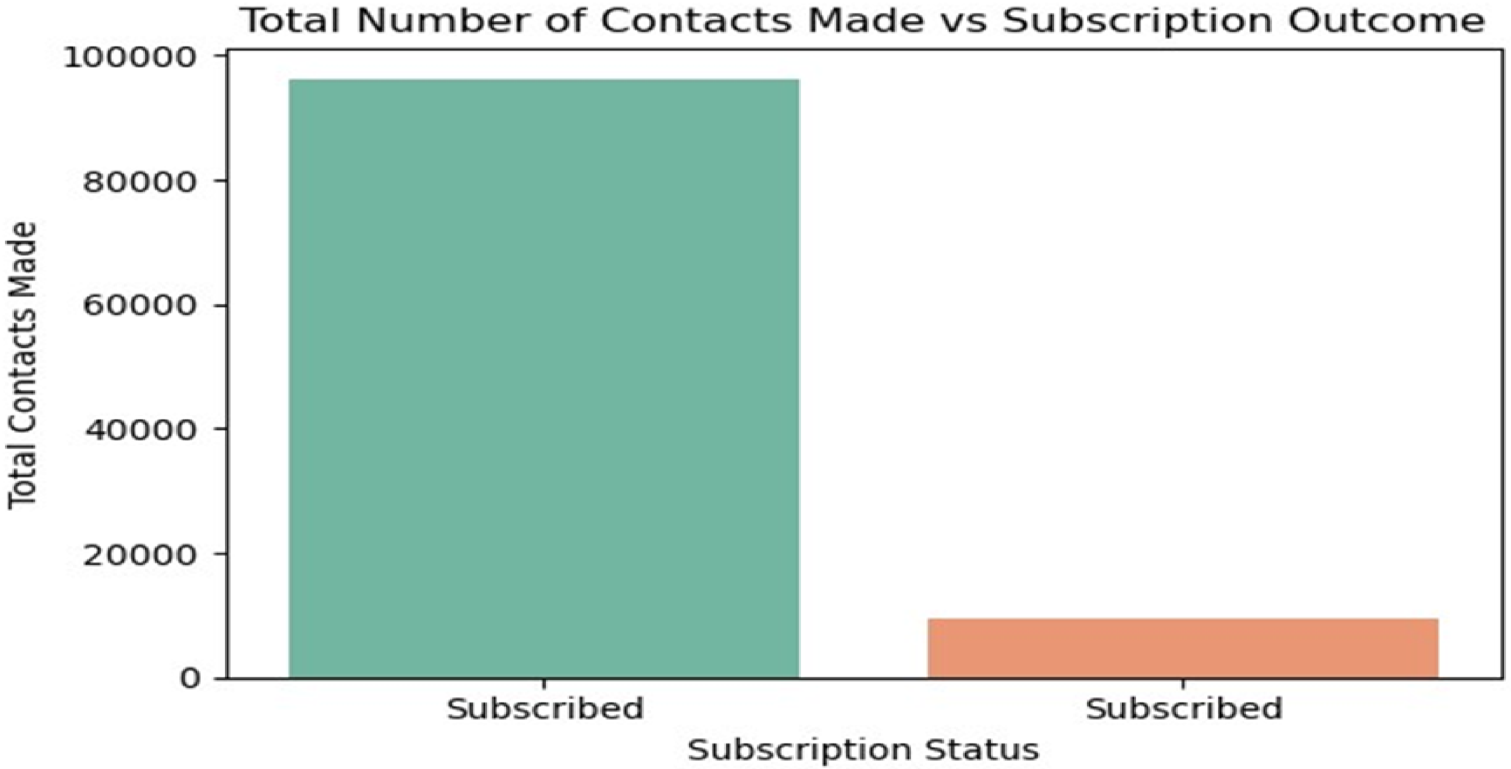
- Classification of the customer base across age-groups



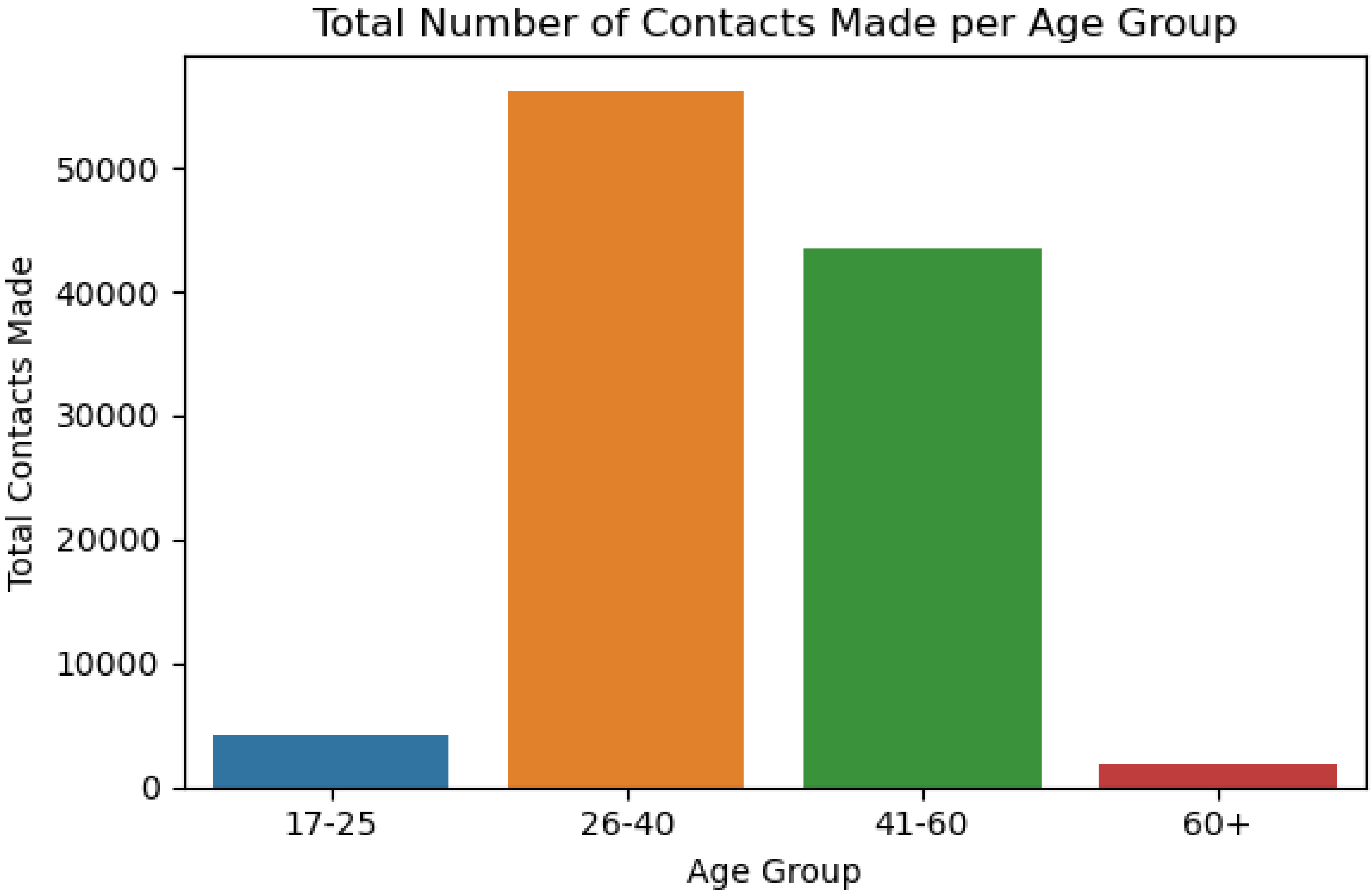
- Looking at relation between different age groups and the output label y
- In the age-groups of 26-40 and 41-60 yrs, majority of the people are not subscribed to the term deposit plan


Age group wise classification of customers against output label
count
20000
17500
15000
12500
10000
7500
5000
2500
y
0
1

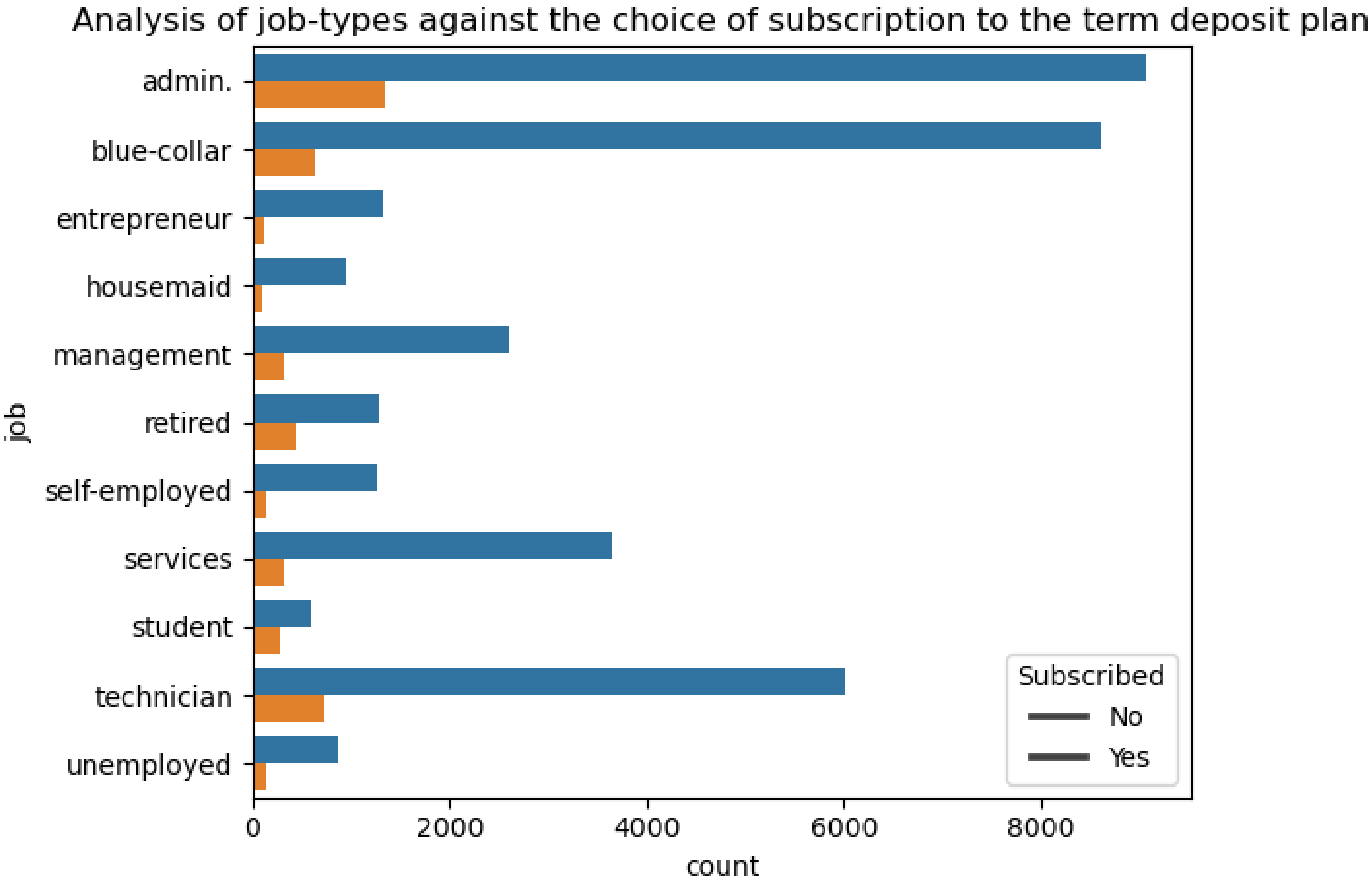
- Looking at relation between Number of contacts made to the customer (campaign) and the output label y
- When a greater number of contacts is made to the customer, they haven't subscribed to the term deposit plan



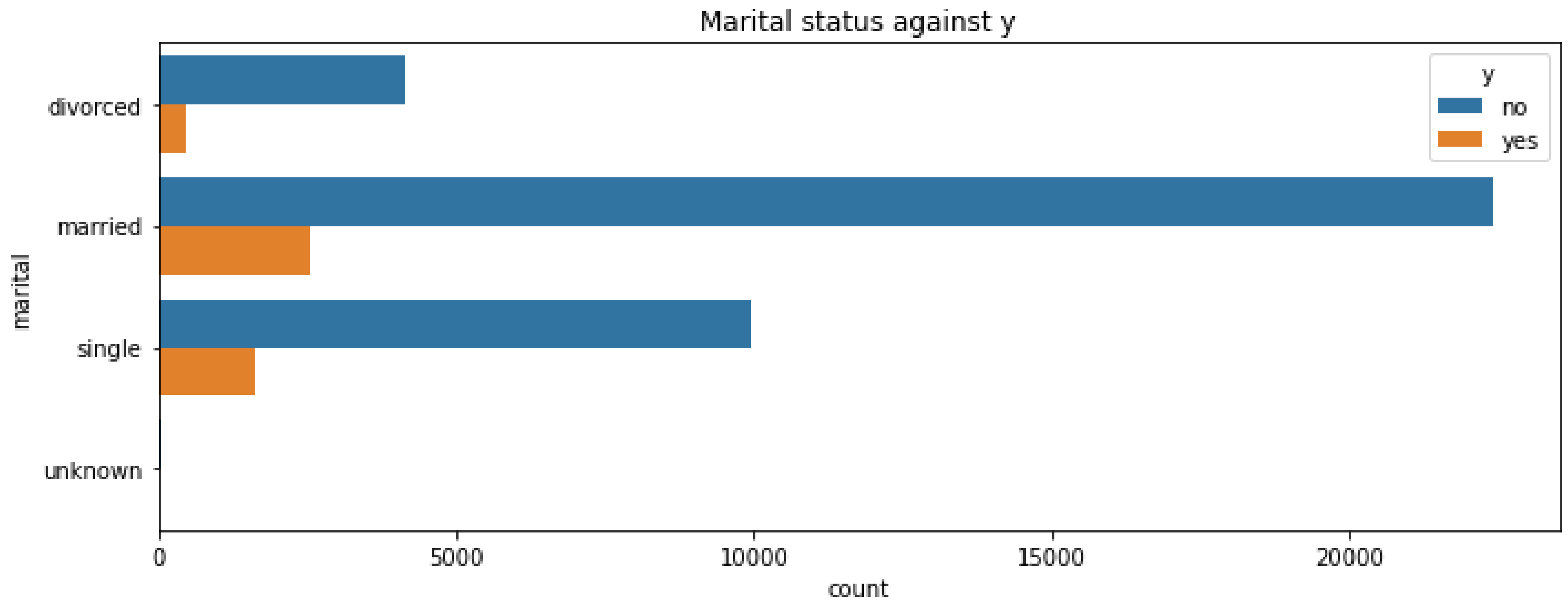
- Looking at relation between 'age_group' and 'campaign' that is number of contacts performed for each age group
- The 26-40 and 41-60 age-groups witness majority of the contacts made in this campaign. These two age-groups seem to the target groups for the bank.

Total Number of Contacts Made per Age Group
Total Contacts Made
0
10000
20000
30000
40000
50000
17-25
26-40
41-60
60+
Age Group

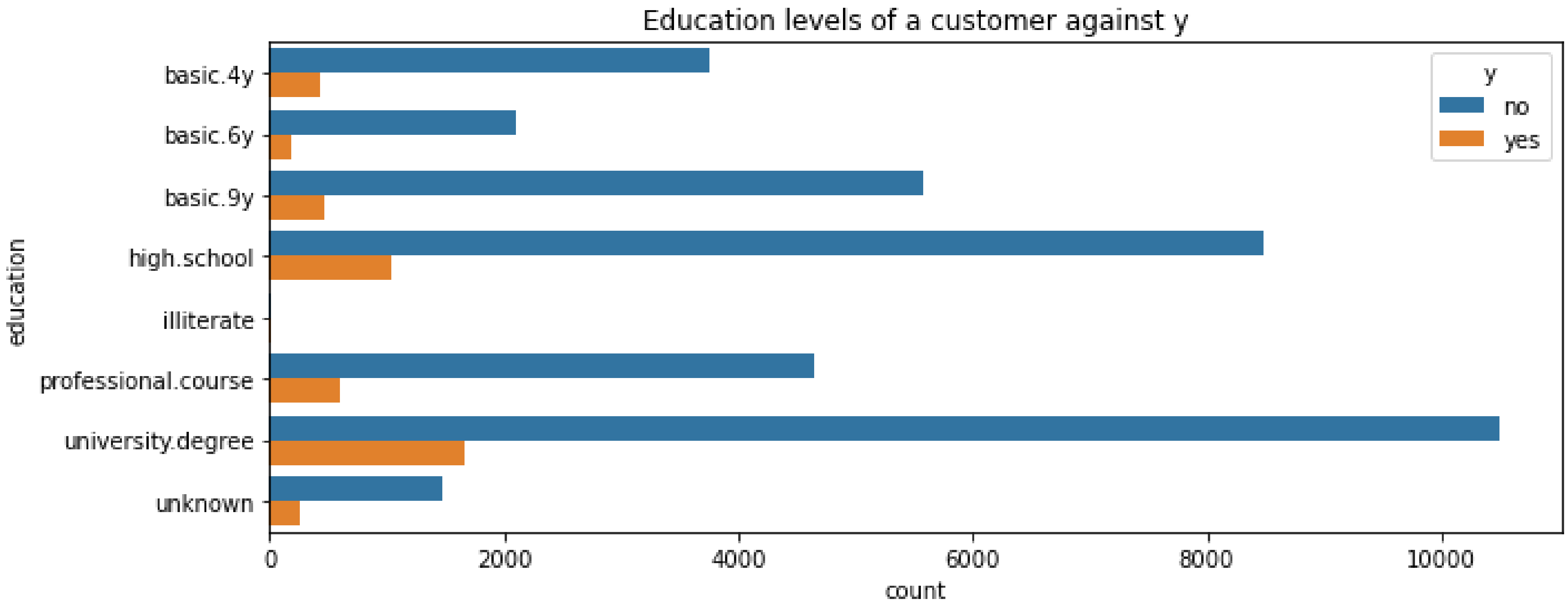
- Looking at relation between job and the output label y
- Looking at the jobs, 'admin', 'blue-collar' and 'technician' are the prominent jobs and most of the customers in these jobs have rejected the term deposit plan.


Analysis of job-types against the choice of subscription to the term deposit plan
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
job
0
2000
4000
6000
8000
count
Subscribed
No
Yes

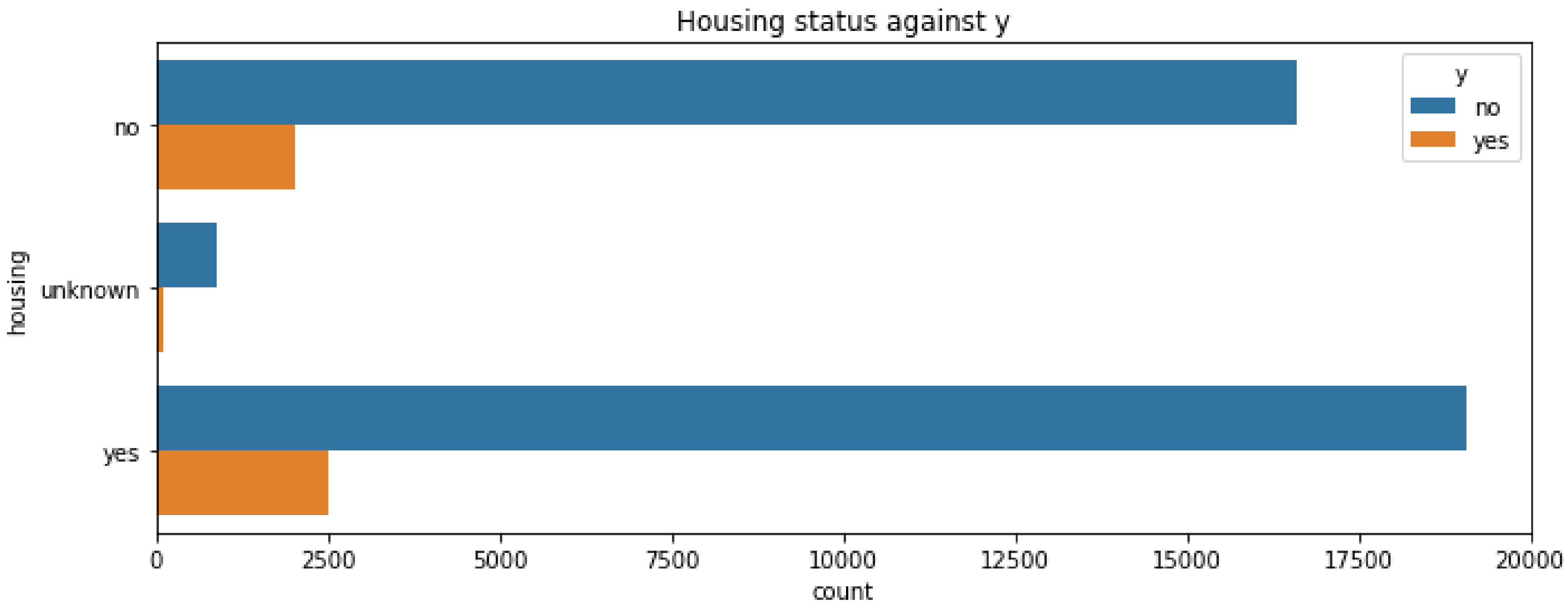
- Analysing marital status and the output label
- married and single customers are the majority of the customer base and comparatively married customers have taken the term deposit



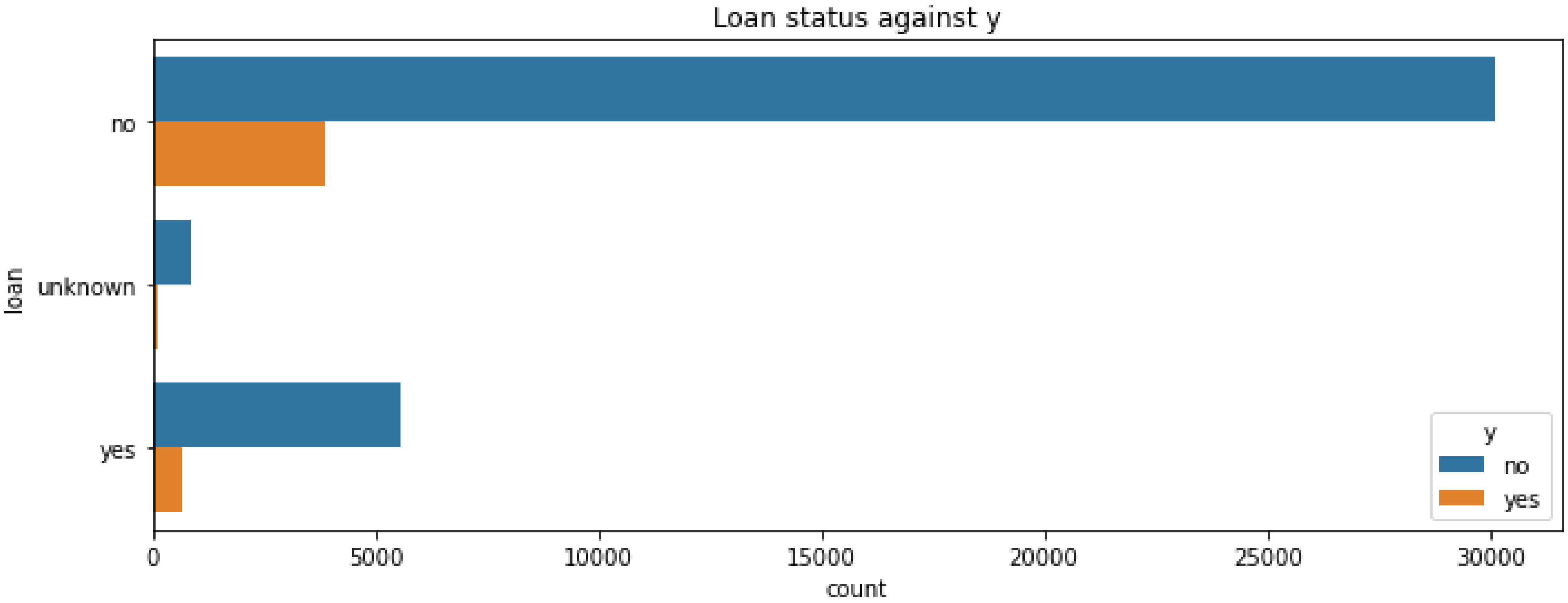
- Analysing the different education levels of a customer against the choice of subscription
- Customers with university degree have subscribed to the term deposit more



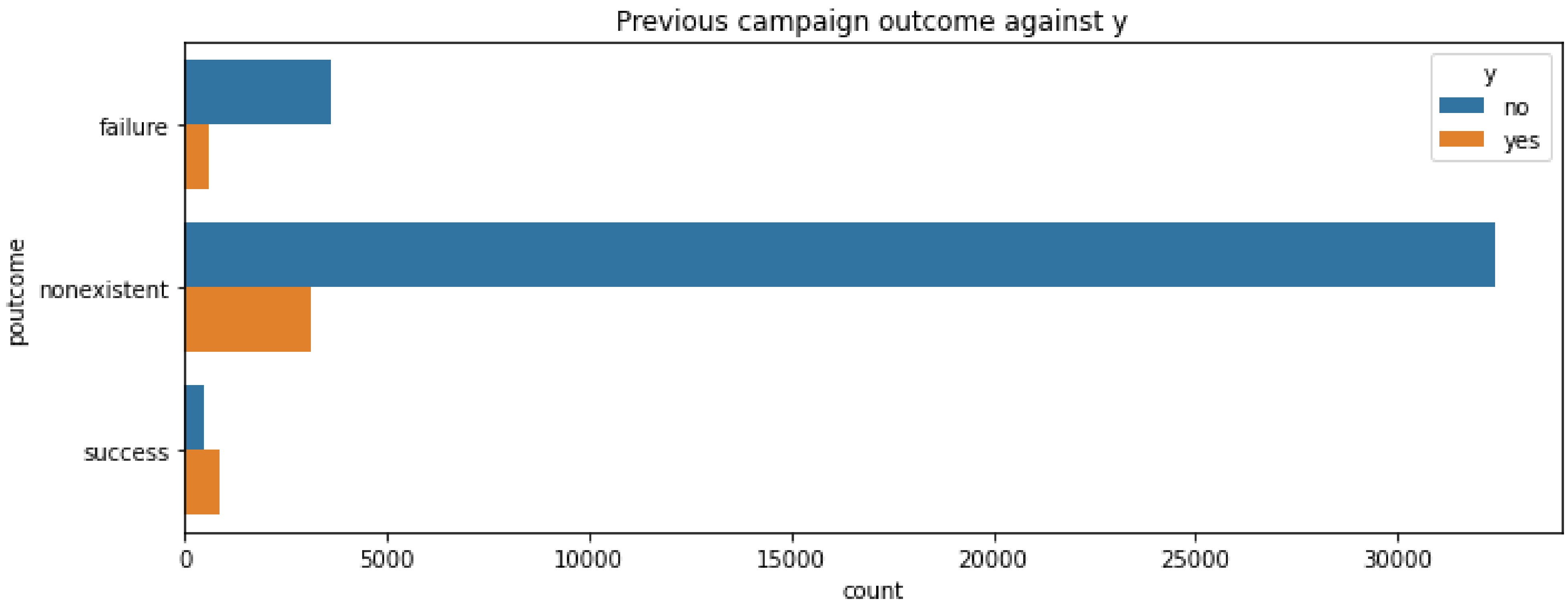
- Analysing housing status and y
- Number of customers who have subscribed to the term deposit is comparatively more for those with housing loan



- Analysing loan status and y
- Number of customers who have subscribed to the term deposit is comparatively less for those with personal loan



- Analysing poutcome and y
- The success rate of previous marketing campaign has resulted in more number of people subscribing to the term deposit



Recommendation to improve campaign

1. **May is the most effective month to contact customers**
2. **Increase the time of contacts made per customer**
3. **Give more focus on university graduate students and high school degree students**
4. **age-groups of 26-40 and 41-60 have a higher proportion among customers, therefore these groups present a profitable target for the marketing team.**
5. **Target the admins and technicians for more subscriptions**

Recommended models for this dataset

1. **Logistic Regression**
2. **Naïve Bayes**
3. **Decision Tree**
4. **Random Forest**
5. **Gradient Boosting**

Hyper parameter tuning and model evaluation will be performed in order to determine the best model and the important features

Model Building

- Categorical features like 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome', 'campaign2' and ‘y’ are converted to numerical features
- Feature scaling is performed on all features to achieve global minimum fast in gradient descent using StandardScaler()
- Dataset is split into train and test using train_test_split()
- The dataset is trained using the five recommended models, Logistic Regression (Linear), Decision Tree (Linear), Naïve Bayes (Linear), Random Forest (Ensemble) and Gradient Boosting (Boosting)
- For evaluating the model, cross validation testing is used with the number of folds as 10 and accuracy as the metric.

Accuracy results

Model	**Accuracy**
Logistic Regression	**90.02%**
Decision Tree	84.13%
Naïve Bayes	80.28%
Random Forest	89.21%
Gradient Boosting	90.03%

From all the above Models, Gradient Boosting performed the best with an accuracy of 90.03% therefore I recommend this model for production purpose

Hyper parameter tuning

- **Since Gradient Boosting gave better performance, hyper parameter tuning was performed on it to identify best features.**
- **The model was run using the following parameters**

Parameter	**Values**
classifier__n_estimators	100, 200
classifier__learning_rate	0.01,0.1,0.2
Classifier__max_depth	3,5,7
classifier__subsample	0.8,1.0

- **The best parameters values were identified as**

Parameter	**Values**
classifier__n_estimators	100
classifier__learning_rate	0.1
Classifier__max_depth	3
classifier__subsample	0.8

- **With Hyperparameter tuning the accuracy increased from 90% to 90%.**

Confusion matrix and classification report

The Gradient Boosting model achieved 11,125 correct predictions and 1,228 incorrect predictions,

performing very well at identifying non-subscribers (class 0) but struggling with subscribers (class 1).

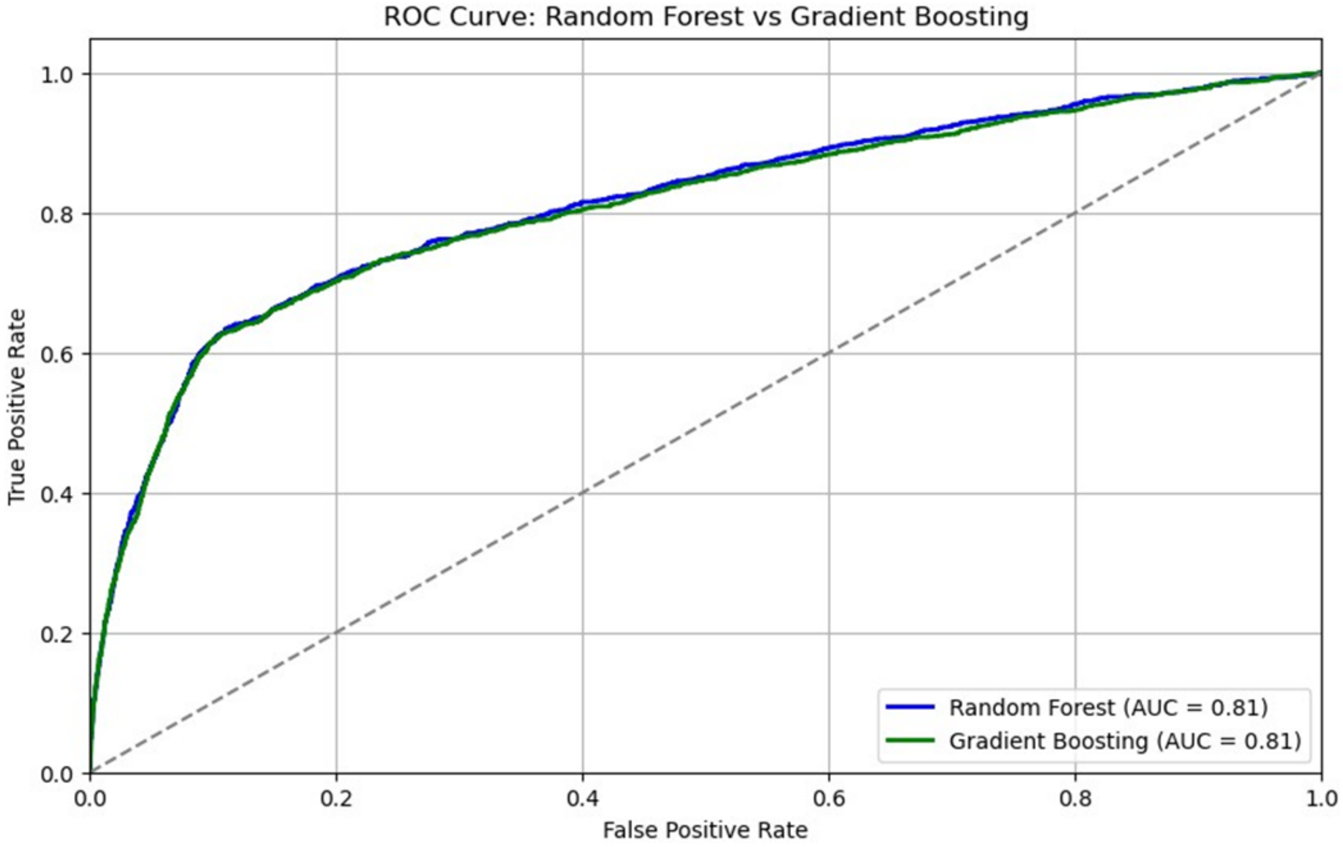
The classification report shows an overall accuracy of 90%, with a precision of 90%, meaning the model is highly accurate when predicting subscribers. However, the recall for class 1 is low at 23%,

indicating it misses many actual subscribers, resulting in a low F1-score of 35% for this class.

These results suggest that while the model performs strongly overall, the class imbalance is affecting its ability to detect subscribers effectively.

AUC-ROC curve

- An ROC curve is a graph showing the performance of a classifier. ROC is a probability curve plotted with True Positive Rate (also called Recall or Sensitivity) on the y-axis against False Positive Rate (also called as Precision) on the x-axis. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative.


ROC Curve: Random Forest vs Gradient Boosting

Data Glacier
Your Deep Learning Partner

GITHUB REPO LINK : **https://github.com/leenarganta/bank_marketing_campaign/tree/main/presentation**

Thank You

Data Glacier

Your Deep Learning Partner