Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

16th Juillet 2025

Project: Bank Marketing Campaign

Name: Najma Abdi, Leena Ganta, Adama Sall

Batch Code: LISUM45

Specialization: Data Science

Agenda

Problem Statement

Dataset

EDA

Recommendations

Data Glacier

Your Deep Learning Partner

Problem Statement

- ABC Bank is planning to launch a new term deposit product and wants to predict which customers are likely to subscribe based on their past interactions with previous marketing campaigns. By developing a machine learning model, the bank aims to identify high-probability prospects in advance, allowing the marketing team to focus its efforts where they are most likely to succeed. This targeted approach will help reduce overall campaign costs while improving conversion rates.

- The primary business objective is to predict whether a client will subscribe to a term deposit. This binary classification problem helps optimize marketing efforts and reduce unnecessary contact with clients

Dataset

ata Set UCI Link

After downloading the data and uploaded it in JupyterNotebook along with import pandas as pd" module, we have found the following:

Dataset: bank-additional-full.csv, Date Analyzed: 2025-06-19
Tool Used: Python (pandas)

Data Overview:

- Shape: 41188 rows × 21 columns
- Target Variable: 'y' (binary: yes/no)
- Class Distribution (Target)
  - No: 36548 (88.73%)
  - Yes: 4640 (11.27%)

- housing: 990
- loan: 990

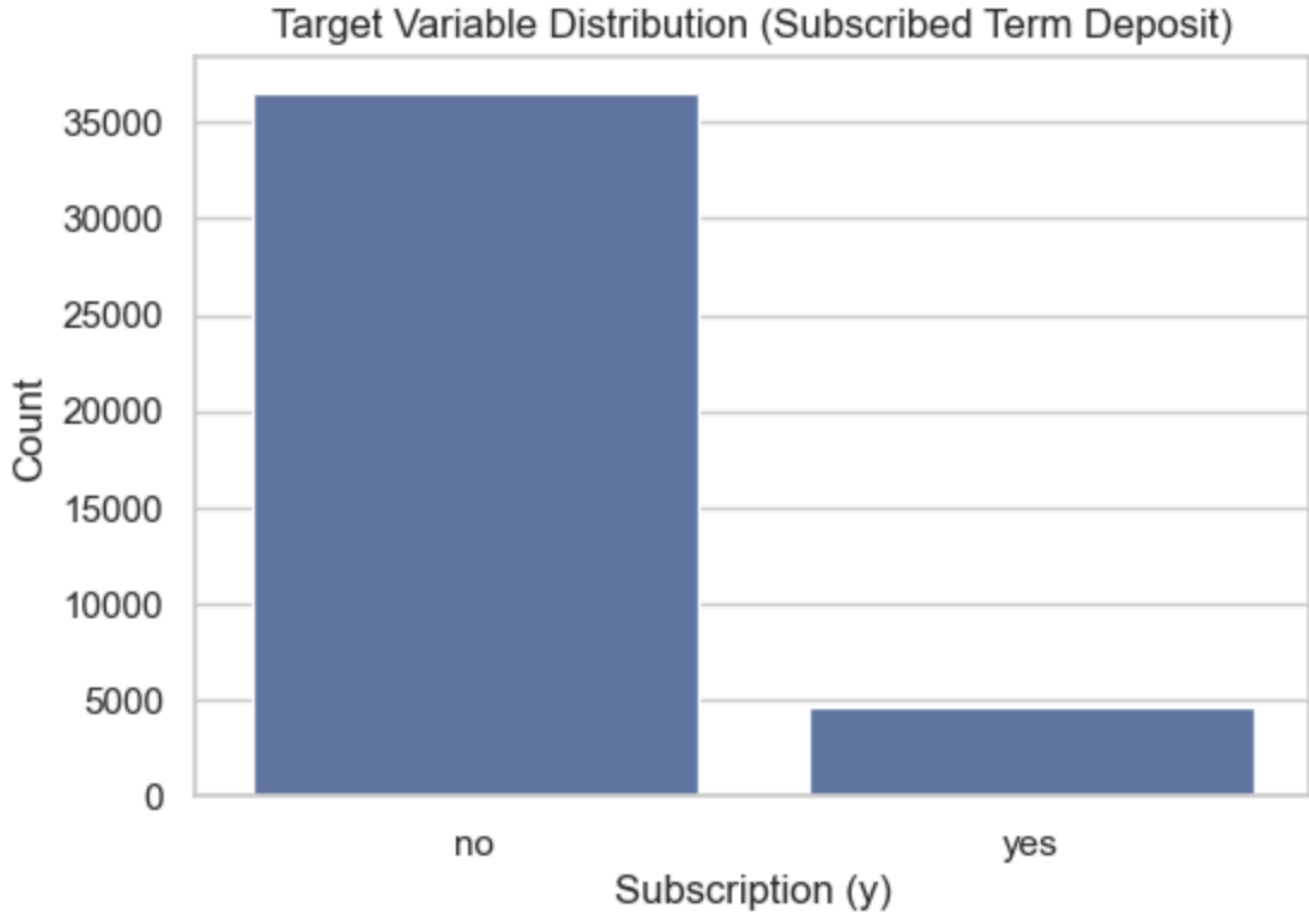
Duplicate rows: 12

EDA

EDA

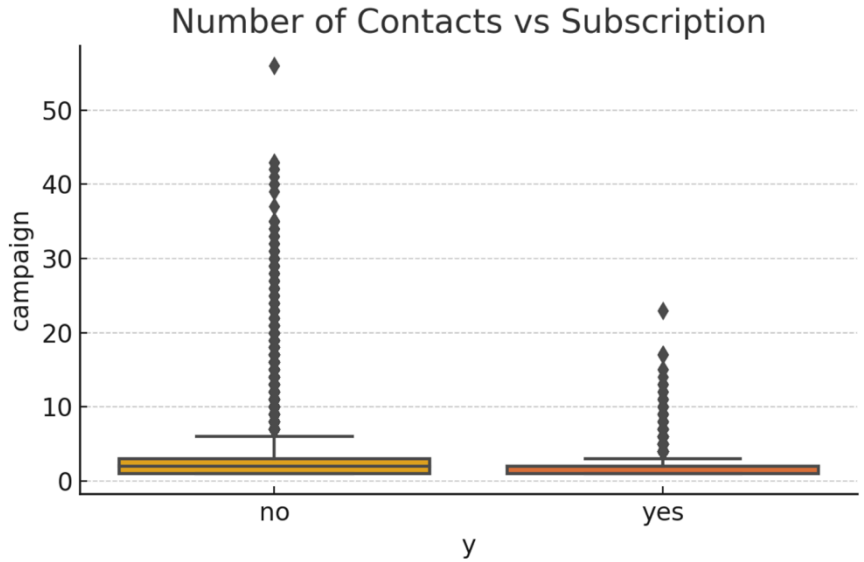
1. Univariate Analysis

- Majority (~89%) of clients do not subscribe.
- Success rate is ~11%.
- This indicates class imbalance and campaign challenges.

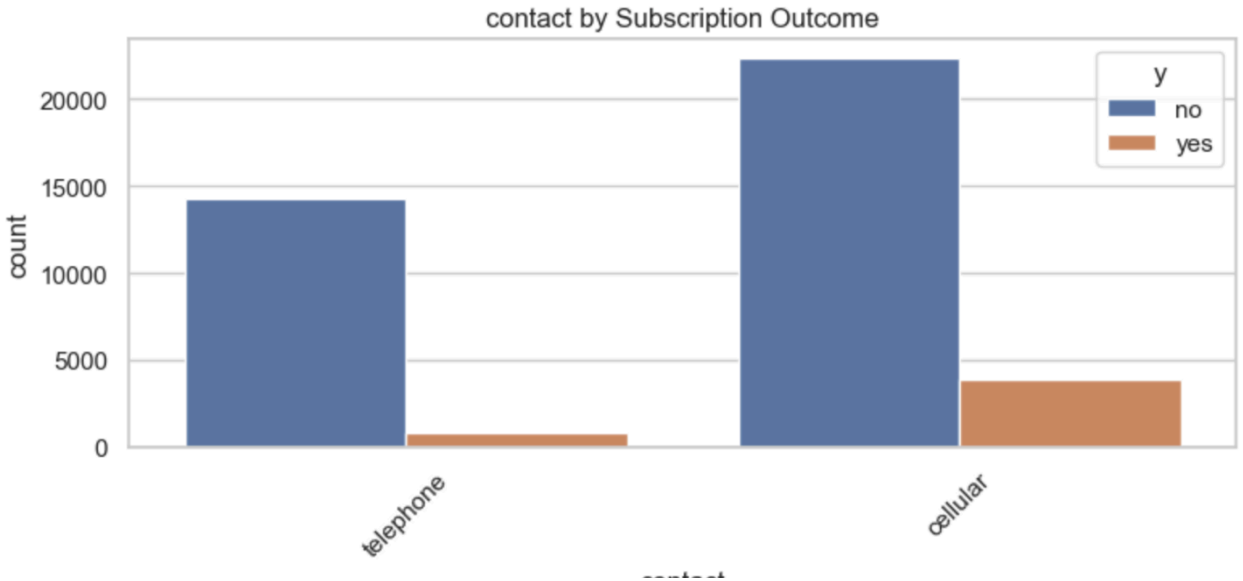

Target Variable Distribution (Subscribed Term Deposit)
Count
35000
30000
25000
20000
15000
10000
5000
0
no
yes
Subscription (y)

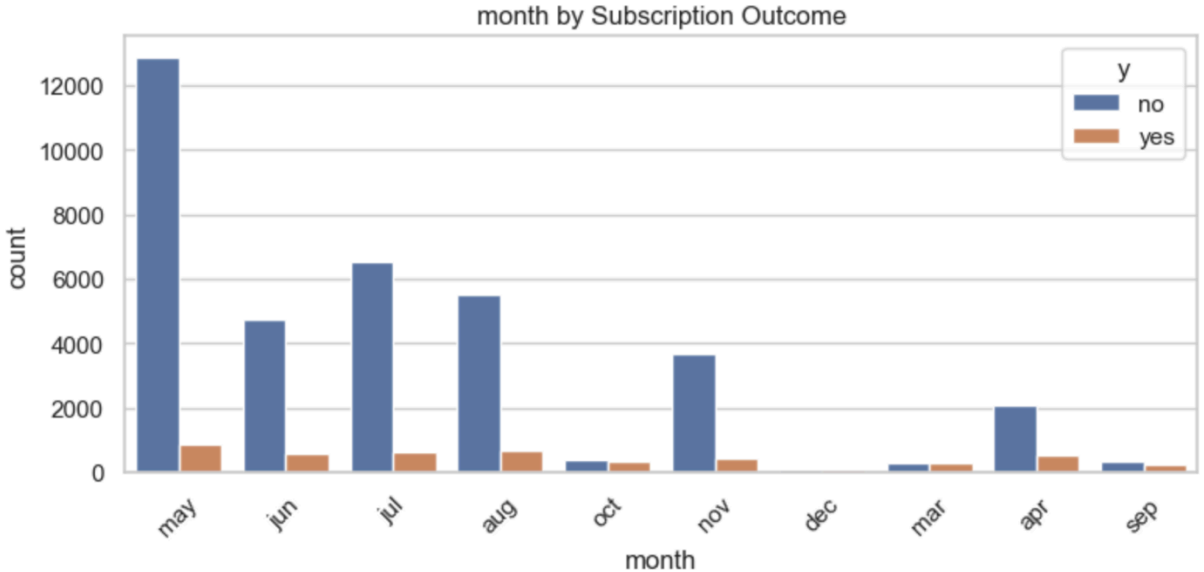
👥 Profile of Likely Subscribers

- Subscribers are often students or retired individuals.
- More likely with no personal/housing loans.
- Longer call durations correlate with success.

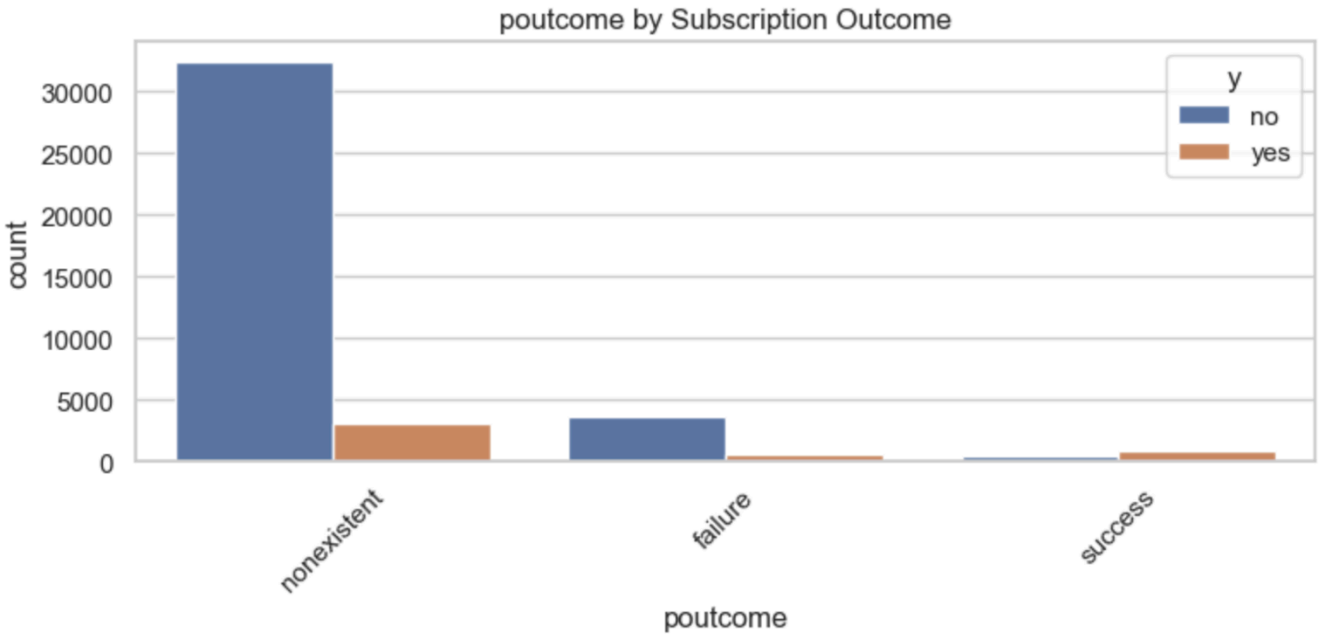

Number of Contacts vs Subscription
campaign
0
10
20
30
40
50
no
yes
y

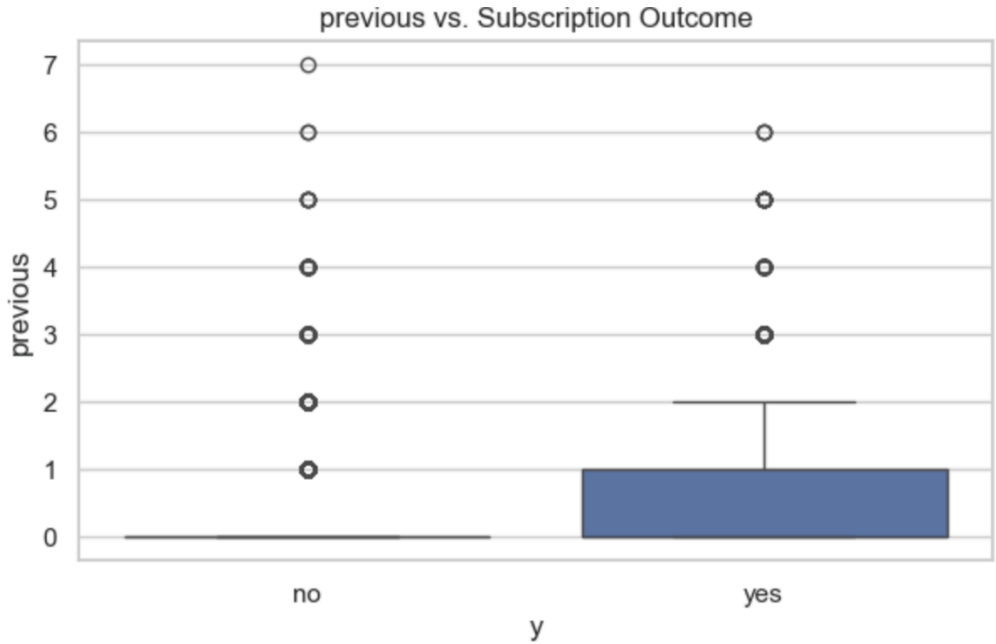
- The type of communication used to contact customers is mostly cellular compared to telephone.
- May seems to be the month with most contacts made.



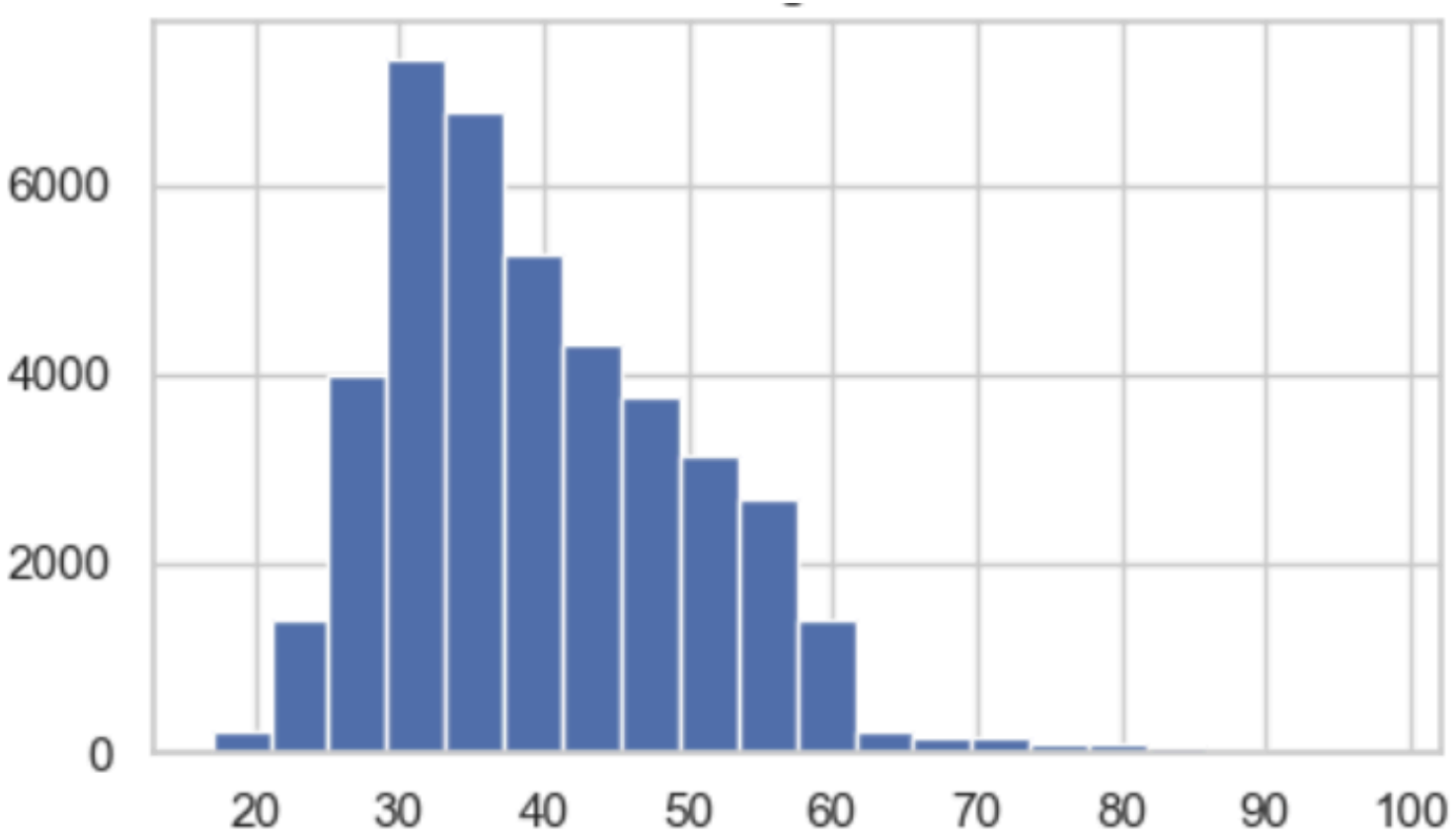

month by Subscription Outcome
y
no
yes
count
12000
10000
8000
6000
4000
2000
0
may
jun
jul
aug
oct
nov
dec
mar
apr
sep
month

- Most of the previous campaign results whether succeeded or failed is nonexistent.
- In the previous campaign, the percentage of people who subscribed is less to those who did not.



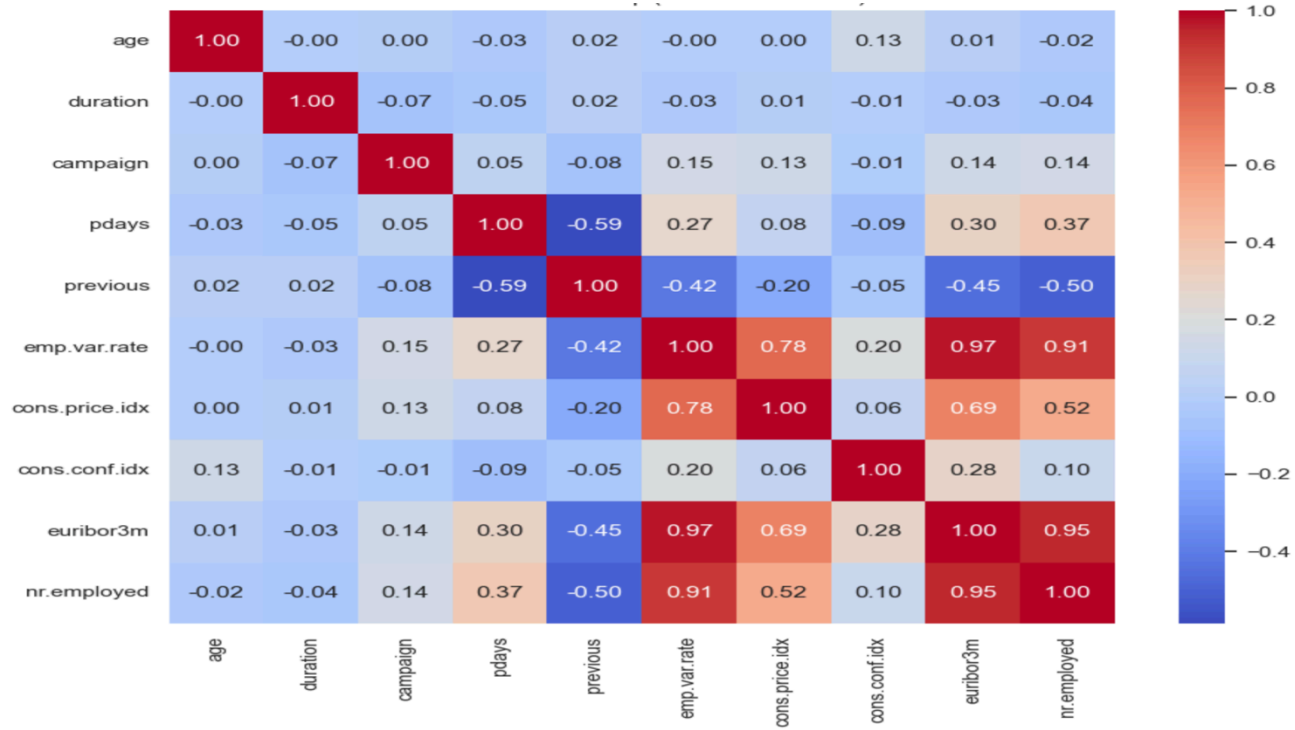

previous vs. Subscription Outcome
previous
7
6
5
4
3
2
1
0
no
yes
y

- 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

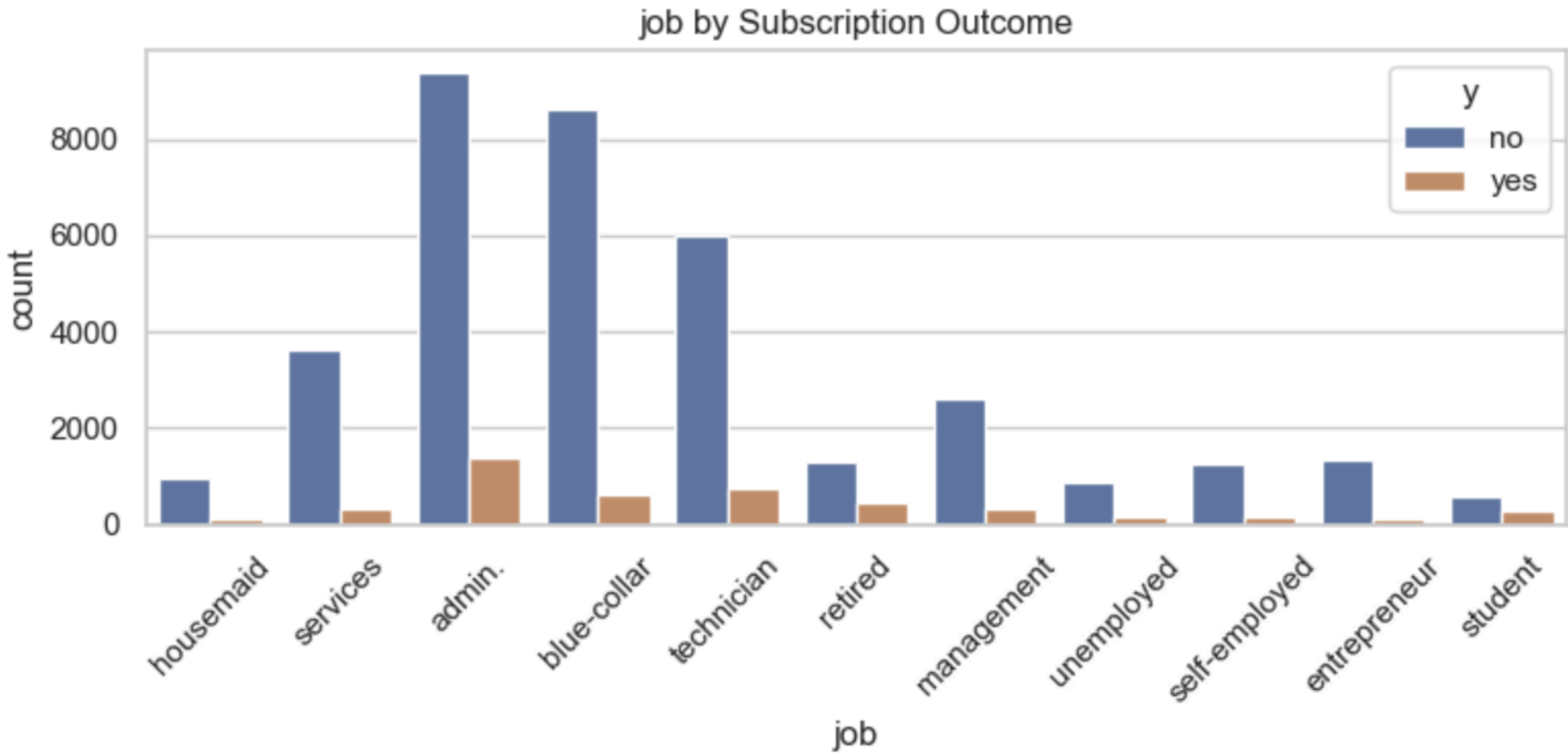
- Most features are not highly correlated, reducing multicollinearity concerns.


age 1.00 -0.00 0.00 -0.03 0.02 -0.00 0.00 0.13 0.01 -0.02
duration -0.00 1.00 -0.07 -0.05 0.02 -0.03 0.01 -0.01 -0.03 -0.04
campaign 0.00 -0.07 1.00 0.05 -0.08 0.15 0.13 -0.01 0.14 0.14
pdays -0.03 -0.05 0.05 1.00 -0.59 0.27 0.08 -0.09 0.30 0.37
previous 0.02 0.02 -0.08 -0.59 1.00 -0.42 -0.20 -0.05 -0.45 -0.50
emp.var.rate -0.00 -0.03 0.15 0.27 -0.42 1.00 0.78 0.20 0.97 0.91
cons.price.idx 0.00 0.01 0.13 0.08 -0.20 0.78 1.00 0.06 0.69 0.52
cons.conf.idx 0.13 -0.01 -0.01 -0.09 -0.05 0.20 0.06 1.00 0.28 0.10
euribor3m 0.01 -0.03 0.14 0.30 -0.45 0.97 0.69 0.28 1.00 0.95
nr.employed -0.02 -0.04 0.14 0.37 -0.50 0.91 0.52 0.10 0.95 1.00
age duration campaign pdays previous emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

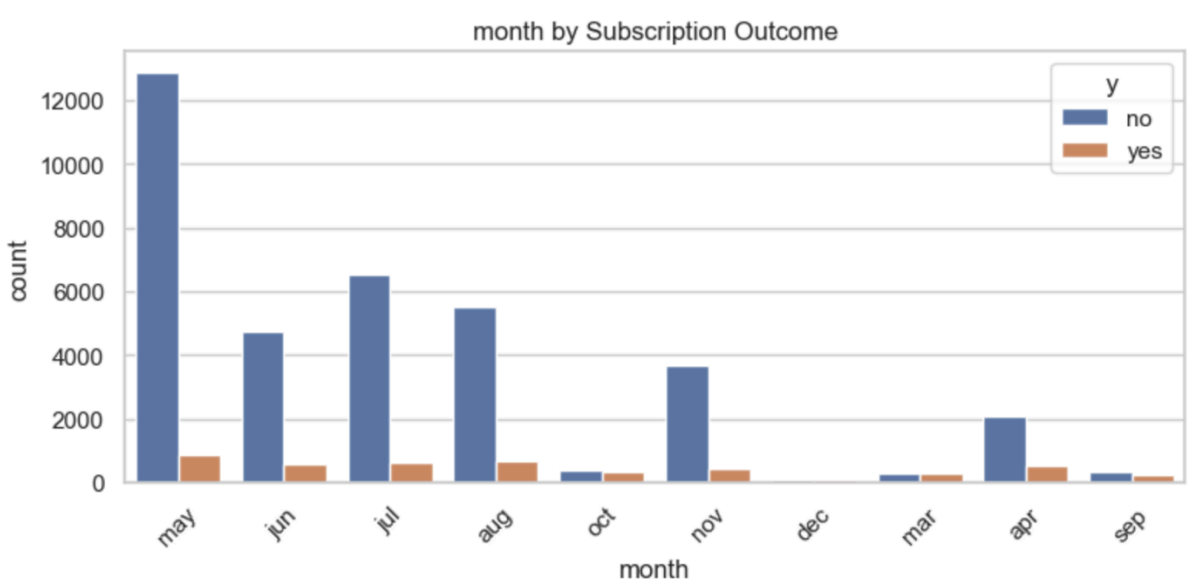
EDA

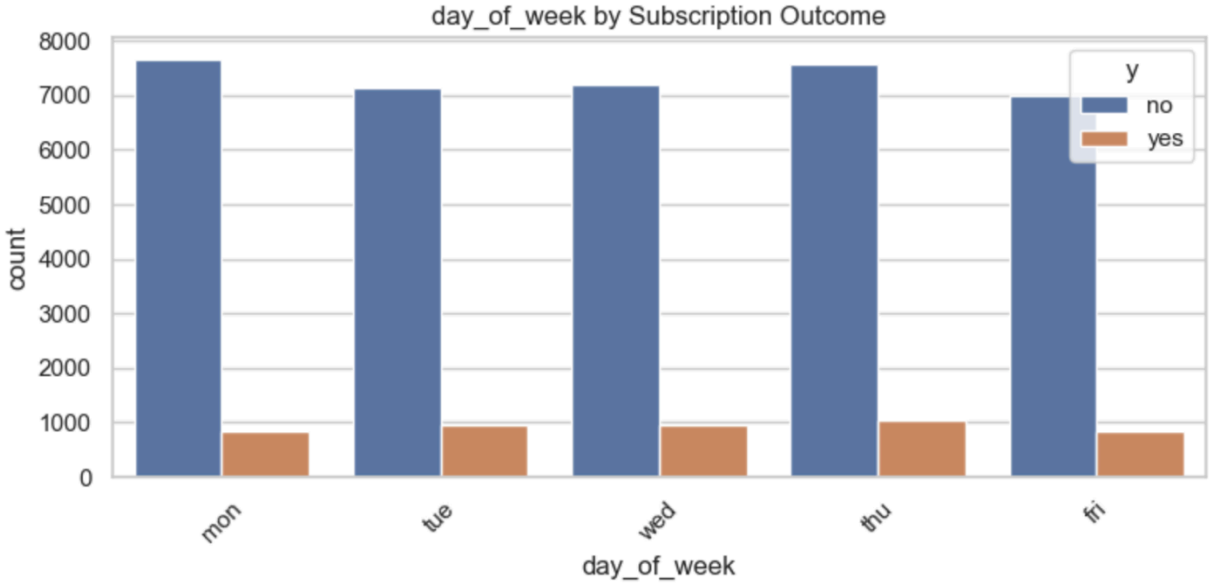
2. Bivariate Analysis

- Classification of the Job Subscription Outcome.

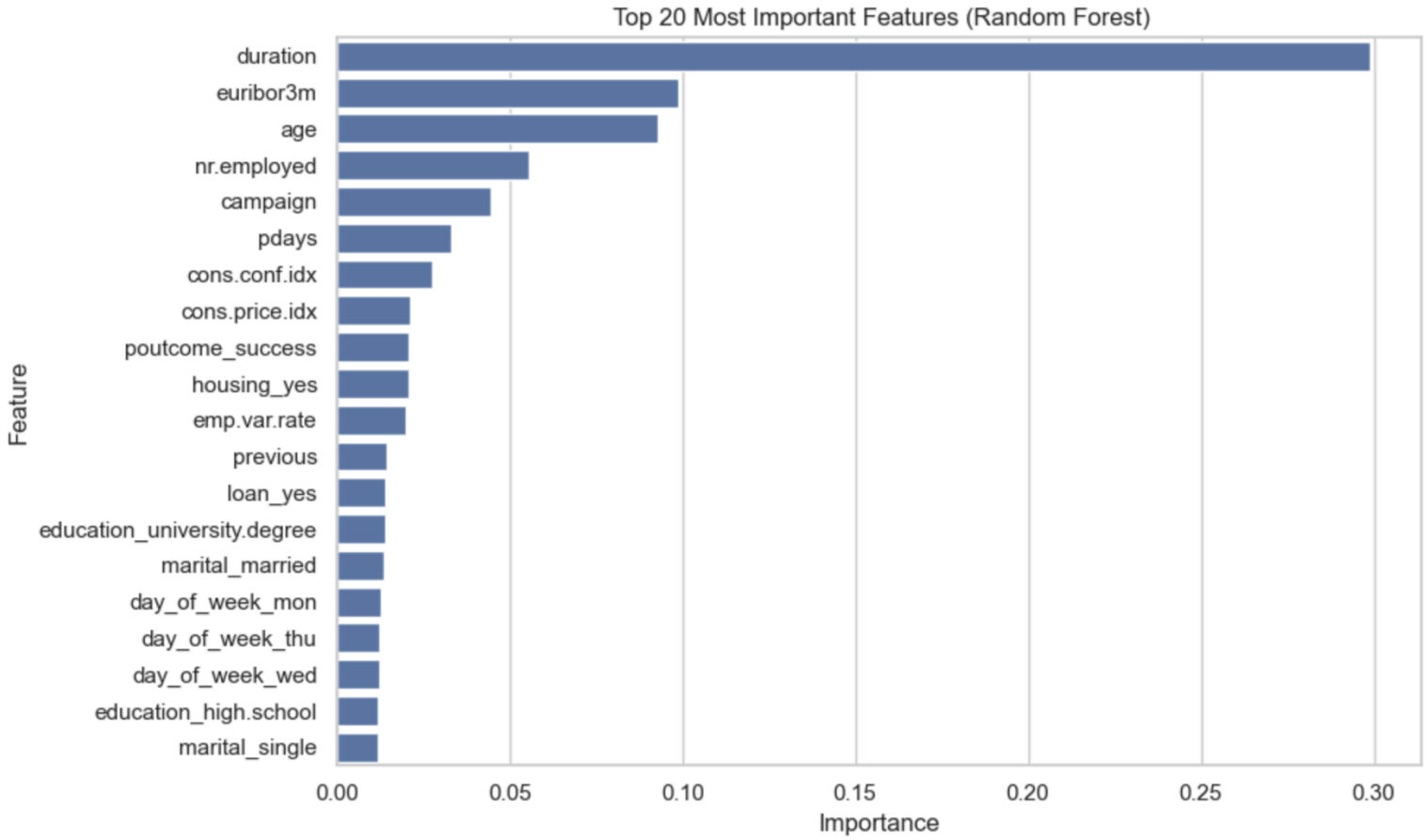

job by Subscription Outcome
y
no
yes
count
8000
6000
4000
2000
0
housemaid
services
admin.
blue-collar
technician
retired
management
unemployed
self-employed
entrepreneur
student
job

- Looking at the correlation between month by Subscription outcome and day-of-week.

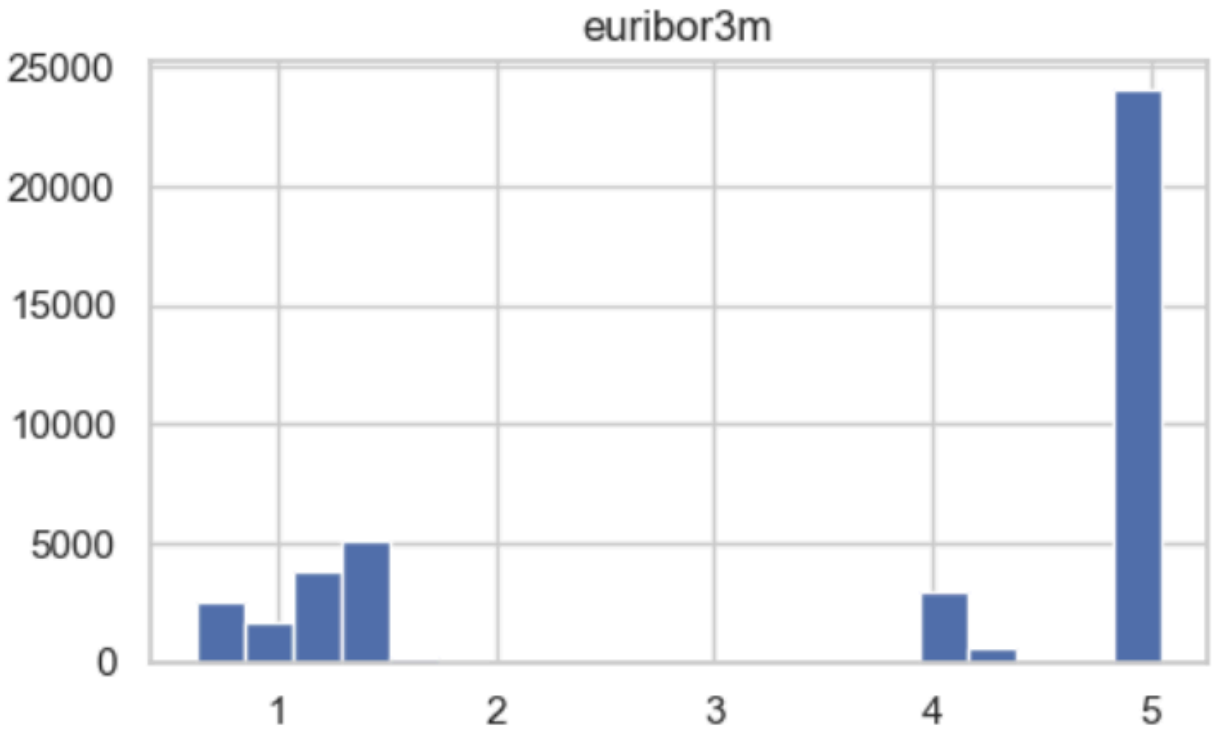


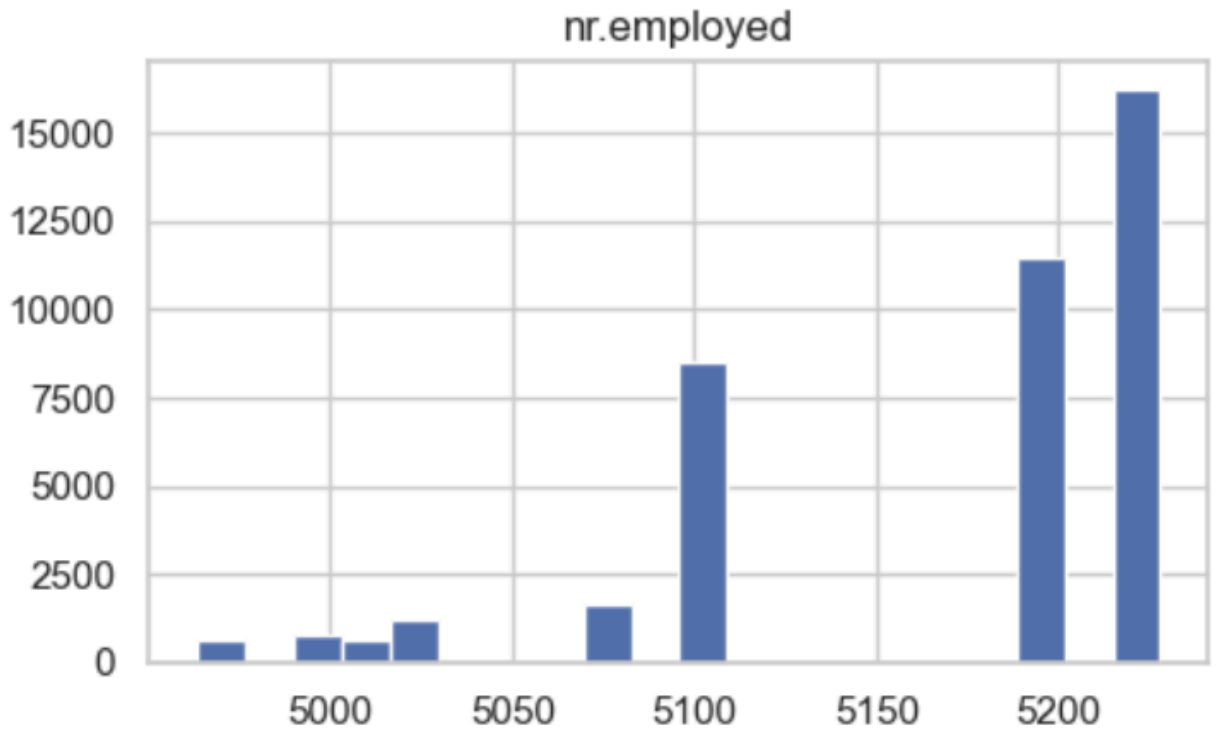


- Looking at relation The Top 20 Most Important Features



- Economic variables (e.g., euribor3m, nr.employed) have strong correlations and predictive potential.


euribor3m
25000
20000
15000
10000
5000
0
1
2
3
4
5


nr.employed
15000
12500
10000
7500
5000
2500
0
5000
5050
5100
5150
5200

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**

Data Glacier
Your Deep Learning Partner

profitable target for the marketing team.

5. Target the admins and technicians for more subscriptions

Recommended models for this dataset

Data Glacier
Your Deep Learning Partner

Problem Type: Binary classification with imbalance.

1. **Random Forest: Robust, interpretable**
2. **XGBoost: High accuracy, handles imbalance Decision Tree**
3. **Logistic Regression: Good, baselined Boost**

Techniques:

4. **- SMOTE or class weighting**
5. **- Exclude 'duration' for real-time prediction**
6. **Deploy with: sklearn, xgboost, joblib, Fast API**

Thank You