

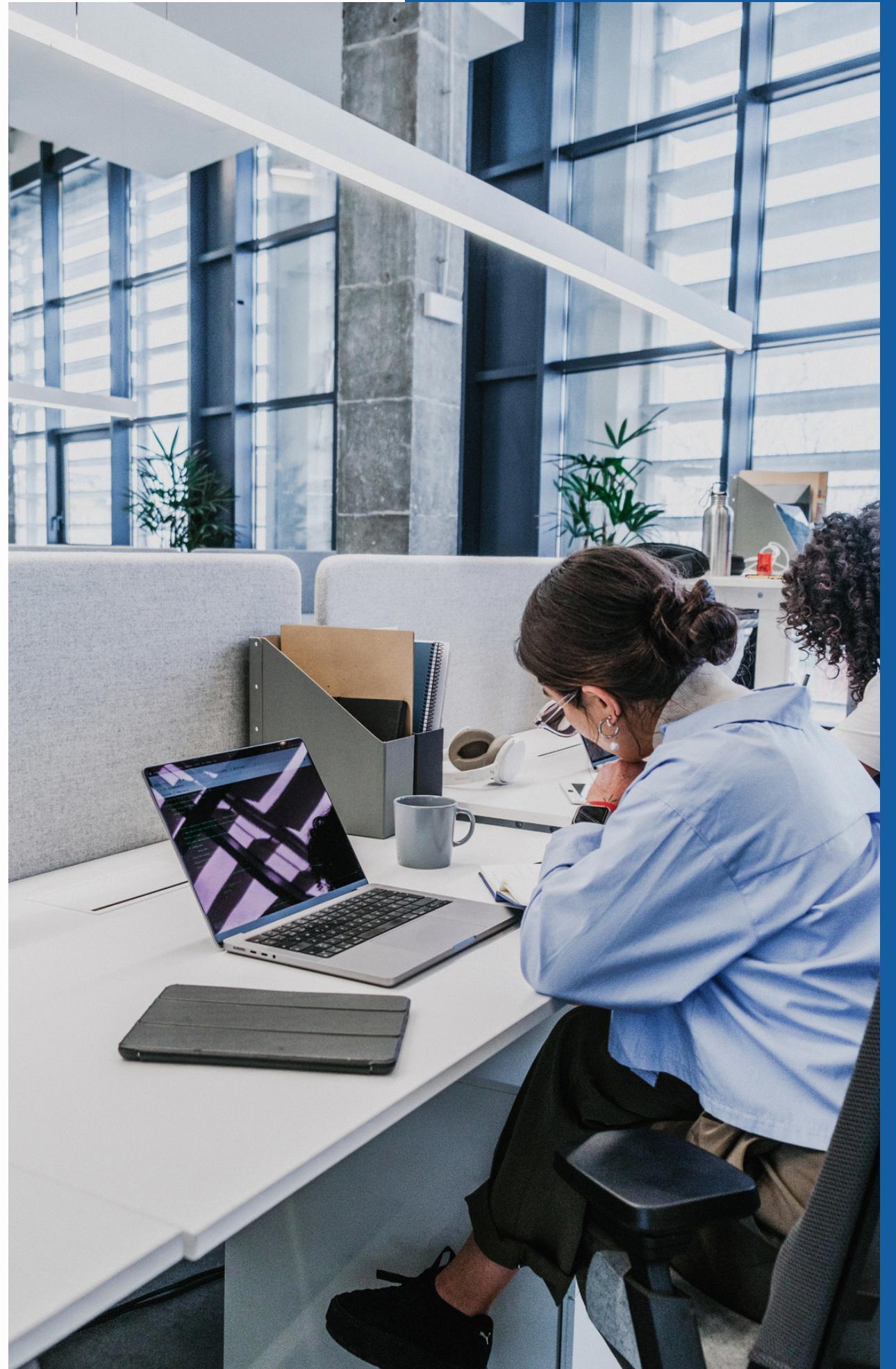
Predictive Analysis of Bank Deposits

By: Dao Phuoc Thinh



Overview

- ▶ I. BUSSINES PROBLEM
- ▶ II. OVERVIEW AND CLEAN DATA
- ▶ III. EDA AND VISUALIZATION
- ▶ IV. MODEL PREDICT
- ▶ V. CONCLUSION





I. Business Problem

Lợi nhuận ngân hàng phụ thuộc vào tiền gửi dài hạn. Ngân hàng đã tiến hành một chiến dịch tiếp thị dựa trên các cuộc gọi điện thoại nhưng không hiệu quả. Các giám đốc tiếp thị chịu áp lực phải thuyết phục khách hàng mua tiền gửi dài hạn. Ngân hàng cần thêm tiền gửi dài hạn để tăng cường dự trữ tiền mặt và nâng cao lợi nhuận.



Realistic Challenges

- **Phản Hồi Tiêu Cực Cao:** Hầu hết khách hàng có xu hướng từ chối chiến dịch tiếp thị.
- **Tiếp Cận Khách Hàng:** Liên hệ với tất cả khách hàng tốn rất nhiều thời gian và công sức.

Main target

- **Phân tích dữ liệu:** Giúp ngân hàng tối ưu hóa chiến lược tiếp thị, tiết kiệm nguồn lực và tăng tỷ lệ thành công của các chiến dịch tiếp theo.
- **Xây Dựng Mô Hình Dự Đoán:** Dự đoán liệu khách hàng mới có đăng ký gửi tiền có kỳ hạn hay không, dựa trên dữ liệu từ các chiến dịch tiếp thị trước đó.



III. Overview and clean Data

Lorem ipsum dolor sit amet, consectetur adipiscing
elit. Duis vulputate nulla at ante rhoncus, vel
vitae ante imperdiet odio.

Dataset Name

- Bank Target Marketing
(<https://www.kaggle.com/datasets/mountboy/online-store-customer-data/>).

Overview Data

Dataset gồm 1 table (17 cột và 56373 dòng) chứa:

- Thông tin khách hàng
- Thông tin chiến dịch trước
- 7 cột là dạng số and 10 cột là dạng phân loại.
- Dữ liệu không có giá trị rỗng.

Clean Data

- Loại bỏ 19.8% dữ liệu trùng lặp (còn 45211 dòng)

Dataset Detail

Biến số	Giá trị
age	Tuổi của khách hàng.
job	Loại công việc.
marital	Tình trạng hôn nhân.
education	Trình độ học vấn.
default	Khách hàng có nợ quá hạn không.
balance	Số dư trung bình hàng năm (tính bằng euro).
housing	Khách hàng có khoản vay mua nhà không.
loan	Khách hàng có khoản vay cá nhân không.
contact	Phương thức liên lạc cuối cùng.
day	Ngày liên lạc cuối cùng trong tháng.
month	Tháng liên lạc cuối cùng trong năm.
duration	Thời lượng cuộc gọi cuối cùng (tính bằng giây).
campaign	Số lần liên lạc trong chiến dịch này.
pdays	Số ngày kể từ lần liên lạc trước.
previous	Số lần liên lạc trước chiến dịch này.
poutcome	Kết quả của chiến dịch tiếp thị trước đó.
deposit	Biến mục tiêu (khách hàng có đăng ký tiền gửi có kỳ hạn hay không).

III. EDA and Visualization



oooo

III. EDA AND VISUALIZATION

01

General Analysis

02

Rate Deposit

03

Analysis of
categorical variables

04

Analysis of numerical
variables

1. General Analysis

	job	marital	education	default	housing	loan	contact	month	poutcome	deposit
count	56373	56373	56373	56373	56373	56373	56373	56373	56373	56373
unique	12	3	4	2	2	2	3	12	4	2
top	management	married	secondary	no	yes	no	cellular	may	unknown	no
freq	12024	33565	28678	55390	30411	47669	37327	16590	45285	45795

Categorical Data insights

- Job: Hầu hết khách hàng thuộc danh mục quản lý.
- Marital: Đa số khách hàng đã kết hôn.
- Education: Đa số khách hàng có bằng cấp trung học.
- Housing: Đa số có khoản vay mua nhà.
- Loan: Đa số không có vay cá nhân.
- Contact: Phương thức liên lạc phổ biến nhất là 'cellular'.
- Month: Tháng 5 có số lần liên hệ cuối cùng cao nhất.
- Poutcome: Đa số kết quả chiến dịch trước đó là 'unknown'.
- Deposit: Đa số không đăng ký gửi tiền gửi có kỳ hạn.

Descriptive Statistics:

- Pdays (Ngày tính):
 - Nhiều khách hàng có giá trị -1, cho thấy không có liên hệ trước đó.
- Previous (Trước đó):
 - Sự phổ biến của các giá trị 0 cho thấy nhiều khách hàng mới.
- Balance (Số dư):
 - Các giá trị âm cho thấy có thể là thấu chi hoặc nợ, gợi ý về khó khăn tài chính.
- Outliers (Các giá trị ngoại lệ):
 - Số lượng nhỏ các giá trị ngoại lệ trong **balance** và **duration** ảnh hưởng đến phân phối dữ liệu, như được biểu hiện qua khoảng cách giữa phân vị 99.9 và các giá trị tối đa.

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
80%	51.000000	1859.000000	24.000000	368.000000	4.000000	-1.000000	0.000000
85%	53.000000	2539.000000	27.000000	437.000000	4.000000	102.000000	1.000000
90%	56.000000	3574.000000	28.000000	548.000000	5.000000	185.000000	2.000000
95%	59.000000	5768.000000	29.000000	751.000000	8.000000	317.000000	3.000000
96%	59.000000	6572.600000	30.000000	823.000000	8.000000	337.000000	4.000000
97%	60.000000	7777.900000	30.000000	914.700000	10.000000	349.000000	5.000000
98%	63.000000	9439.400000	30.000000	1051.000000	12.000000	360.000000	6.000000
99%	71.000000	13164.900000	31.000000	1269.000000	16.000000	370.000000	8.900000
99.9%	83.000000	32892.770000	31.000000	2091.740000	32.000000	650.370000	22.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

2. Rate deposit

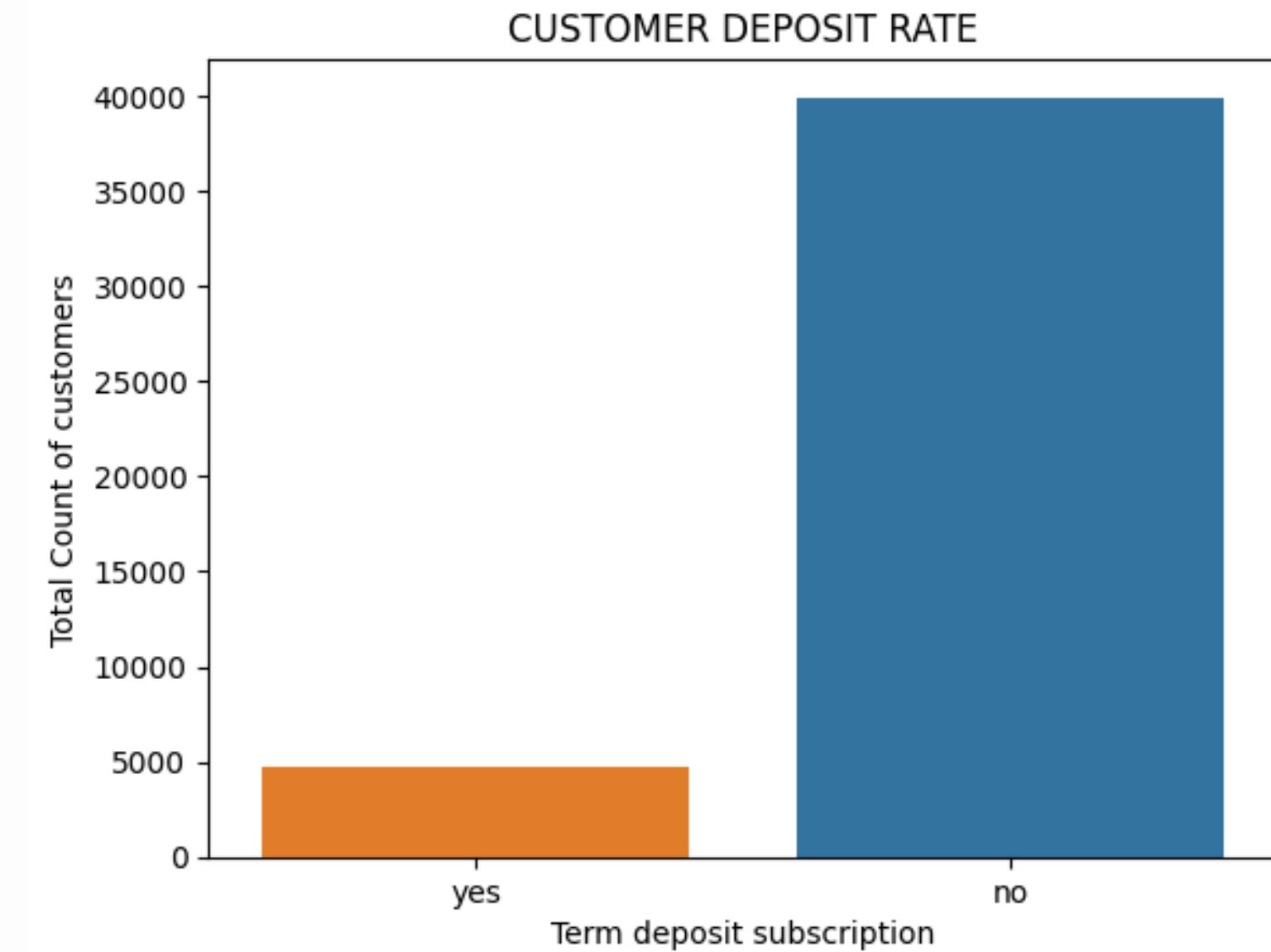
Dữ liệu cũng bị mất cân bằng, có thể ảnh hưởng đến độ chính xác của mô hình dự đoán.

88%

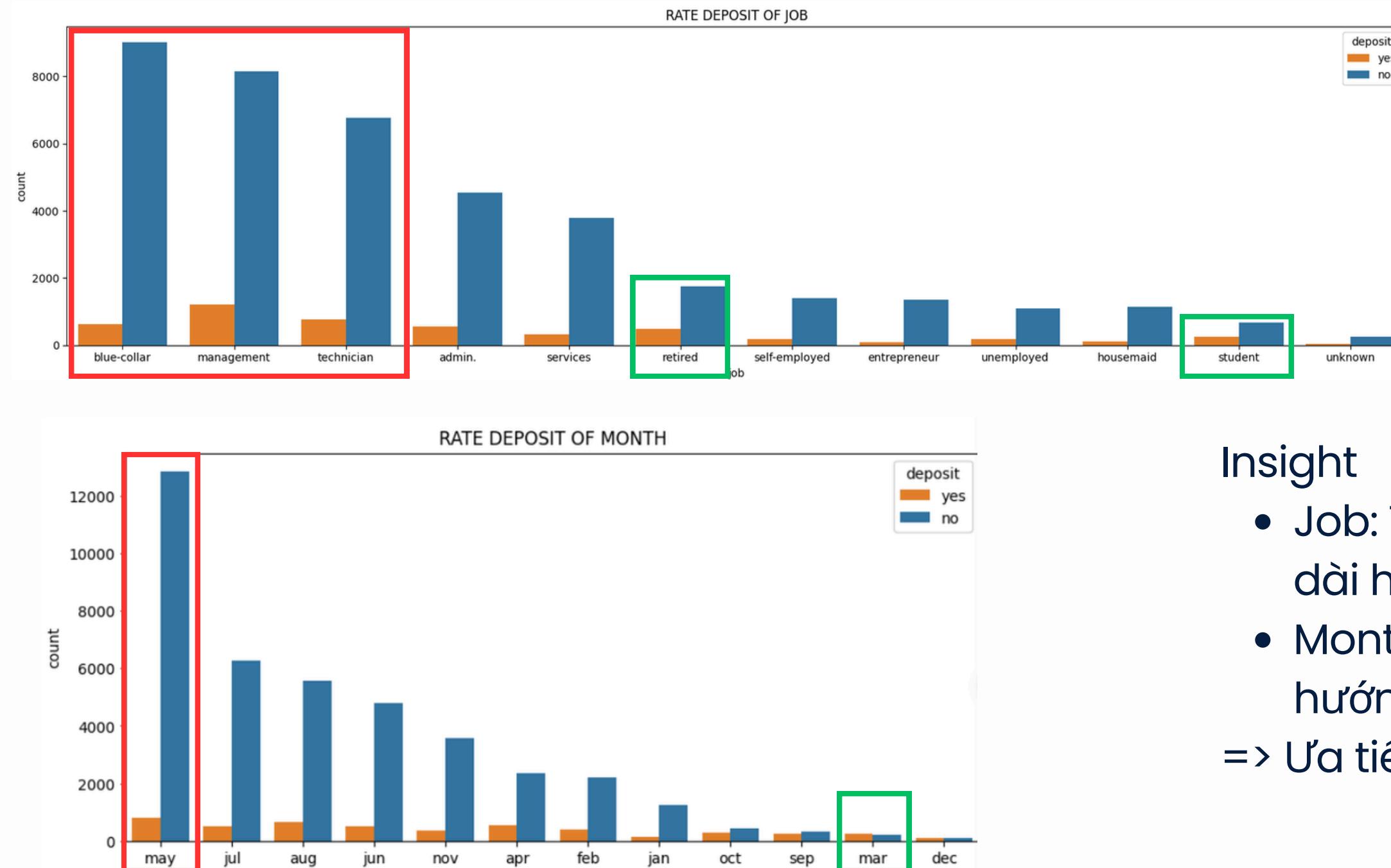
Khách hàng không
gửi tiền

12%

Khách hàng gửi
tiền



3. Analysis of categorical variables

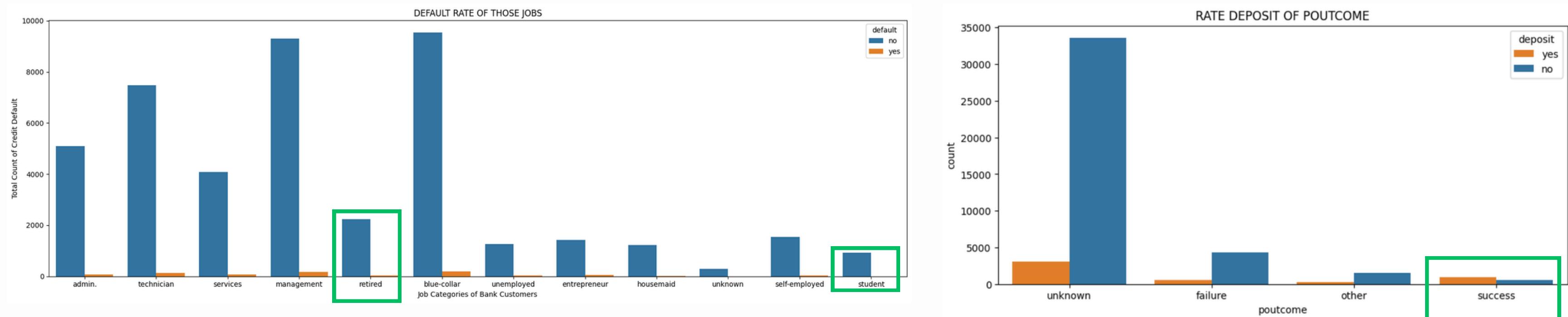


Insight

- Job: Tỉ lệ sinh viên, người thất nghiệp gửi dài hạn cao.
- Month: Tháng cuối năm khách hàng xu hướng gửi cao.
=> Ưa tiên phân bổ

3. Analysis of categorical variables

- Dựa trên phân tích trước đó, có thể thấy rằng khách hàng sinh viên và đã nghỉ hưu có khả năng đăng ký gửi tiền có kỳ hạn cao hơn. Điều này có thể là do họ có tỷ lệ vỡ nợ tín dụng thấp hơn cũng như có nhiều thời gian và nguồn lực hơn để xem xét đầu tư dài hạn.



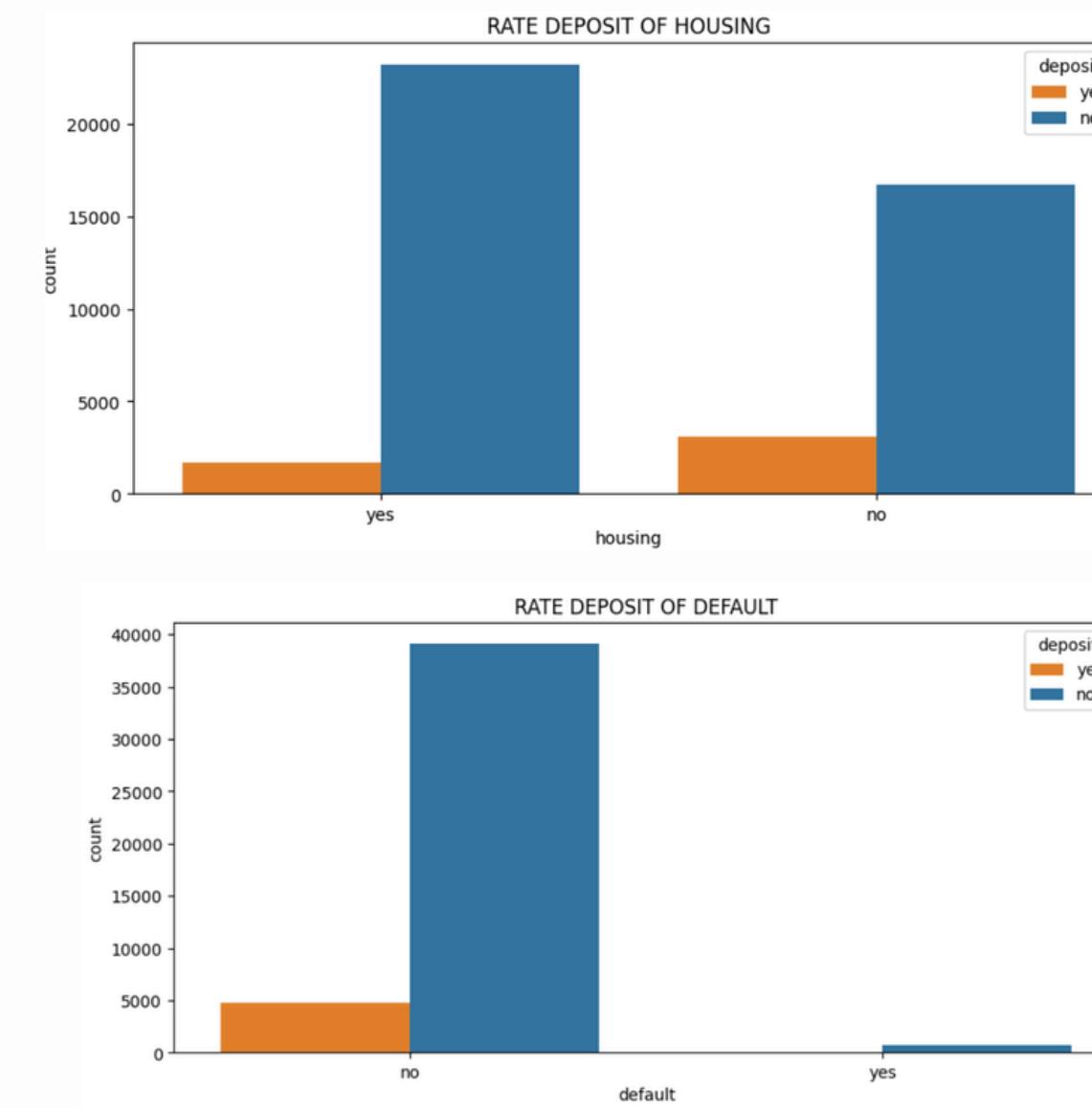
Insight:

- Poutcome: Những khách hàng đã đăng ký gửi tiền có kỳ hạn trong chiến dịch tiếp thị trước đó có nhiều khả năng đăng ký lại hơn. Điều này cho thấy cơ hội thành công cao trong việc nhắm mục tiêu vào những khách hàng này.

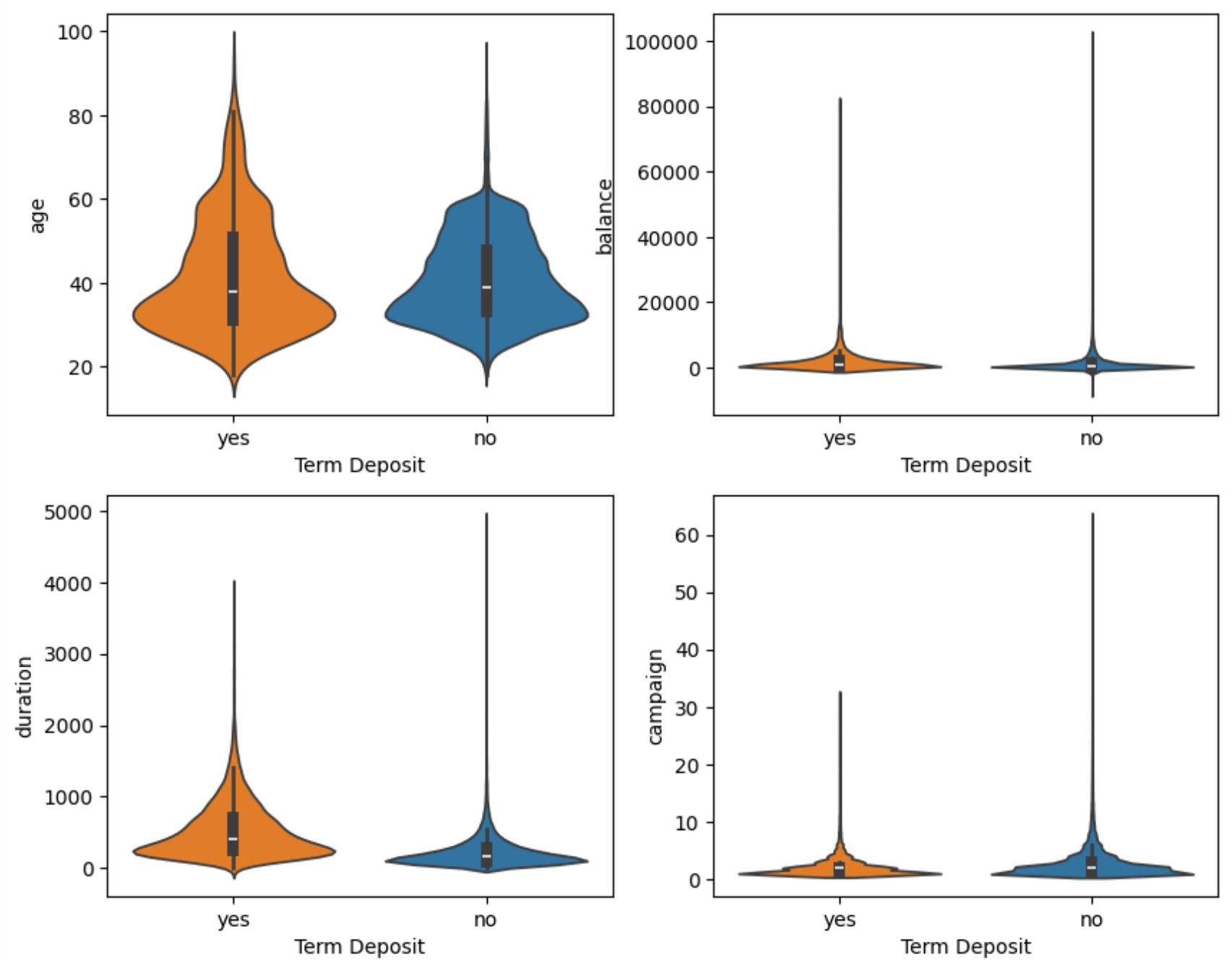
3. Analysis of categorical variables

Housing, Default, Loan:

- Những khách hàng không có khoản vay cá nhân, nhà ở hoặc vỡ nợ tín dụng có tỷ lệ đăng ký gửi tiền có kỳ hạn cao hơn so với những khách hàng có khoản vay hoặc nợ xấu.



4. Analysis of numerical variables



AGE

- Khách hàng trong độ tuổi từ 20 đến 30 có khả năng đăng ký tiền gửi kỳ hạn cao hơn.
- Sau độ tuổi 60, cũng có tỷ lệ cao khách hàng đăng ký, mặc dù số lượng khách hàng liên hệ trong nhóm tuổi này ít hơn.

BALANCE

- Phân bố số dư cho thấy khách hàng có số dư ít hơn ít có khả năng đăng ký tiền gửi kỳ hạn.

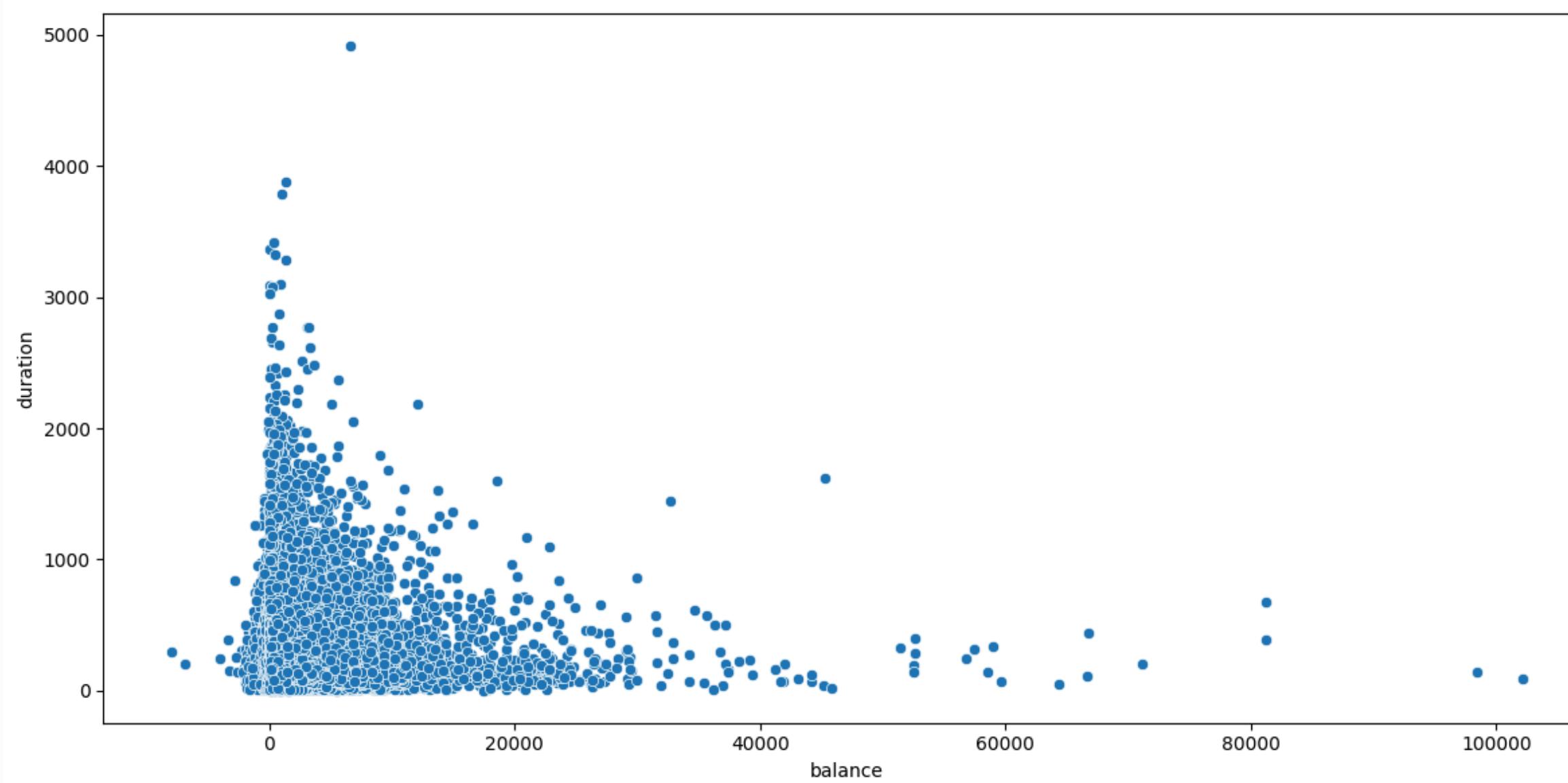
DURATION

- Khi thời lượng liên hệ cuối cùng với khách hàng dài hơn, khả năng khách hàng đăng ký tiền gửi kỳ hạn rất cao.

CAMPAIGN

- Khách hàng được liên hệ 5 lần hoặc ít hơn trong chiến dịch tiếp thị hiện tại có nhiều khả năng đăng ký tiền gửi kỳ hạn hơn.

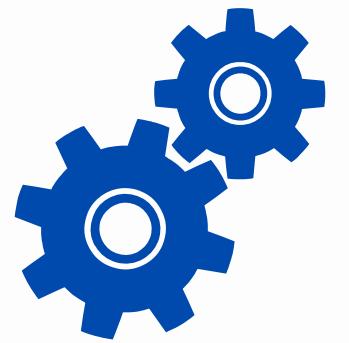
4. Analysis of numerical variables



- Từ biểu đồ phân tán, chúng ta có thể quan sát thấy rằng những khách hàng có số dư ngân hàng thấp hoặc bằng 0 được ngân hàng liên hệ thường xuyên hơn. Chiến lược này có thể không hiệu quả nhất.
- Do đó, ngân hàng nên tập trung vào những khách hàng có số dư ngân hàng trung bình và cao khi liên hệ với họ để tăng khả năng đạt được kết quả tích cực.

IV. Model Predict

Dự đoán liệu khách hàng mới có đăng ký gửi tiền có kỳ hạn hay không, dựa trên dữ liệu từ các chiến dịch tiếp thị trước đó.



Preprocessing



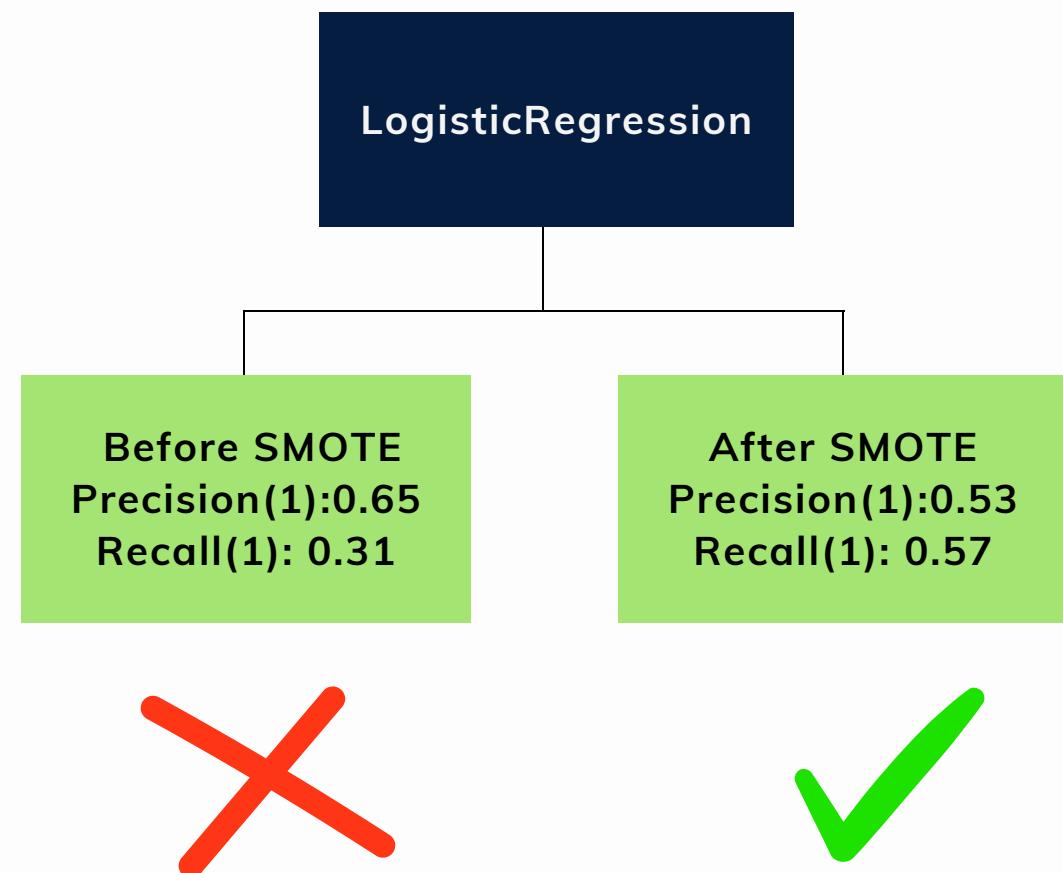
GridSearchCH



Evaluate

1. Preprocessing

- **Label coder** column: education, default, housing, loan, deposit
- **One hot** column: job, marital, contact, month, poutcome
 - (44709, 46)
- **Scaler data**: age, balance, day, duration, campaign, pdays, previous
- **Train test split**: X_train, X_test, y_train, y_test (test_size = 0.25, random_state = 42)
- SMOTE



3. GridSearchCV

- Tìm hyperparameter tối ưu nhất cho 6 model:

Model	param_grid	best_param
RandomForestClassifier	n_estimators: [10 ,50, 100, 200]	n_estimators: 200
KNeighborsClassifier	n_neighbors: [10 ,50, 100, 200]	n_neighbors: 10
AdaBoostClassifier	'n_estimators': [10 ,50, 100, 200]	n_estimators: 200
GradientBoostingClassifier	n_estimators: [10 ,50, 100, 200]	n_estimators: 200
LGBMClassifier	n_estimators: [10 ,50, 100, 200]	n_estimators: 50
XGBClassifier	n_estimators: [10 ,50, 100, 200]	n_estimators: 100

4. Evaluation

Model	Precision (1)	Recall (1)	F1-score (1)	Accuracy
LogisticRegression	0.53	0.57	0.55	0.89
DecisionTreeClassifier	0.45	0.56	0.50	0.88
RandomForestClassifier	0.57	0.62	0.59	0.90
KNeighborsClassifier	0.4	0.79	0.53	0.85
AdaBoostClassifier	0.54	0.6	0.57	0.90
GradientBoostingClassifier	0.54	0.75	0.63	0.90
LGBMClassifier	0.54	0.72	0.62	0.90
XGBClassifier	0.57	0.62	0.59	0.91

Tiêu chí chọn:

- Ưa tiên độ bao phủ recall (nhận biết mọi khách hàng gửi tiền).
- Thứ hai là độ chính xác: Precision

Chọn model GradientBoostingClassifier:

- Recall cao thứ hai: 0.75
- Precision cao thứ hai: 0.51
- F1-score cao nhất: 0.63



V. Conclude

Insight for customers

- Job: Người thất nghiệp và học sinh có tiềm năng gửi cao.
 - Age: Độ tuổi từ 20 -30, đặc biệt là từ 60 trở lên.
 - Housing, Loan, Default: Khách hàng có những mục này nên tránh.
 - Balance: Tránh tập trung khách hàng số dư âm hoặc bằng 0.
-

Insights for the campaign

- Month: Những tháng cuối năm có xu hướng gửi tiếp kiệm hơn.
 - Poutcome: Những khách hàng thành công chiến dịch trước thì chiến dịch tiếp theo xu hướng cũng vậy.
 - Duration: Thời lượng nghe càng lâu thì khả năng thành công càng cao.
 - Campaign: Khách hàng được liên hệ trên 5 lần khả năng gửi tiền rất thấp.
-

Model Predict

- Chọn model GradientBoostingClassifier
- Chọn n_estimators: 200

THANK YOU!

