



# FINAL PROJECT DATA SCIENCE DIGIHUB ADVANCE PPI DUNIA



## **OPTIMIZING BANK CUSTOMER TARGETING FOR PHONE CAMPAIGNS**

by Mufid Habiburrahman



# Introducing **ABOUT ME**

Lulusan Teknik Sipil dari Universitas Sebelas Maret Surakarta (GPA 3.18). Memiliki kemampuan analitis, berpikir terstruktur, serta kepemimpinan melalui pengalaman beberapa organisasi selama di kampus. Supportive, Loyal, berusaha menjadi Long Life Learner, dengan minat pada pengembangan diri dan analisis data.

Disclaimer : Project yang dilakukan adalah mengisi code yang rumpang dan melakukan modeling berdasarkan contoh yang sudah pernah dipelajari.



**Mufid Habiburrahman**

[hr.mufid98@gmail.com](mailto:hr.mufid98@gmail.com) | [linkedin.com/in/mufid-habiburrahman/](https://linkedin.com/in/mufid-habiburrahman/) | [github.com/hrmufid](https://github.com/hrmufid)





### Source

### Banking Dataset – Marketing Targets :

<https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets>

### Context

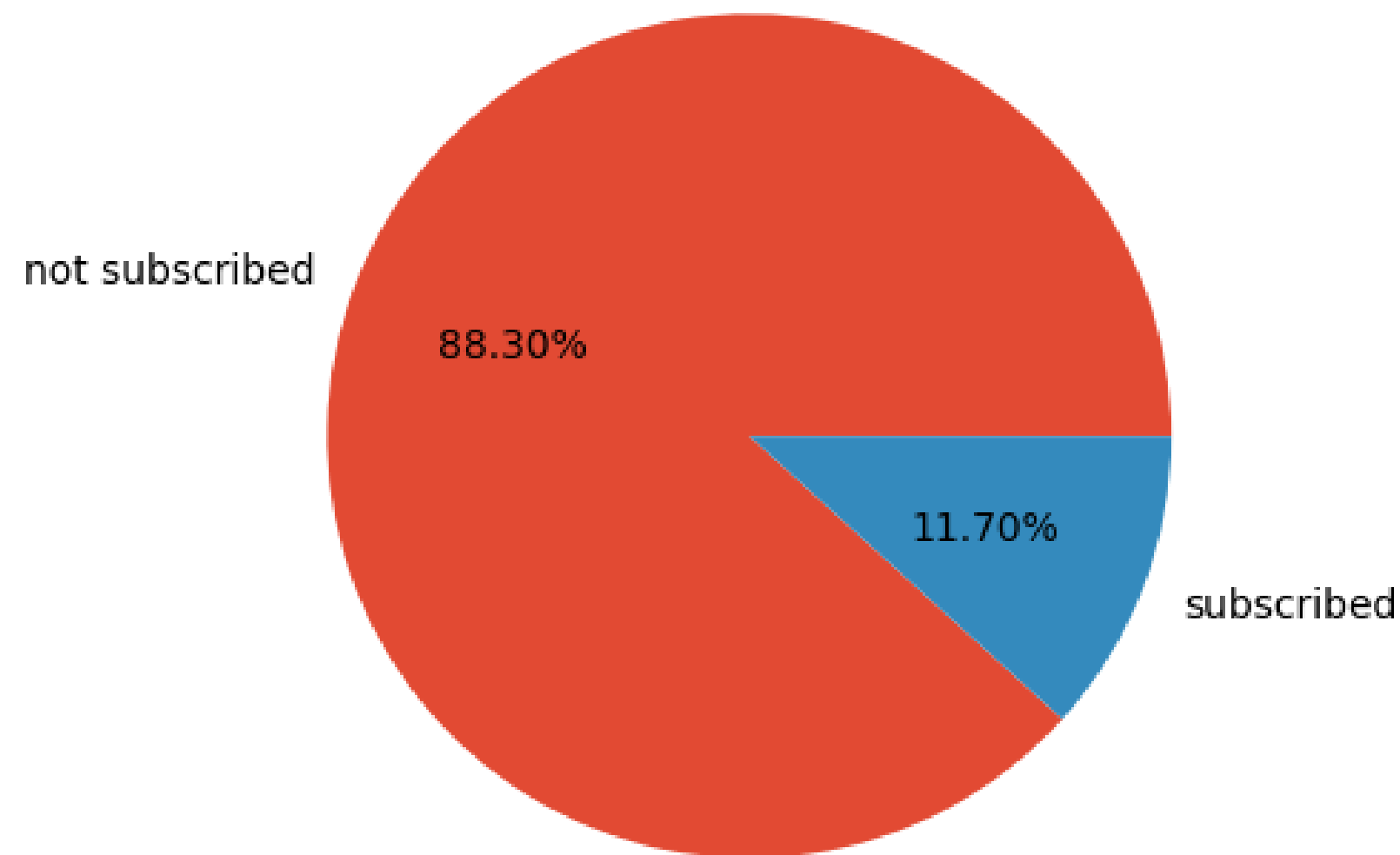
**Deposito berjangka** merupakan salah satu sumber pendapatan utama bagi bank. Deposito berjangka adalah investasi tunai yang disimpan di lembaga keuangan, di mana uang Anda diinvestasikan untuk jangka waktu tertentu dengan tingkat bunga yang telah disepakati.

Bank memiliki berbagai strategi pemasaran untuk menjual produk deposito berjangka kepada nasabah, seperti melalui email marketing, iklan, telemarketing, dan pemasaran digital.

Dari berbagai metode tersebut, **kampanye pemasaran melalui telepon (telemarketing)** masih menjadi salah satu cara paling efektif untuk menjangkau calon nasabah. Namun, metode ini memerlukan investasi besar karena melibatkan penggunaan pusat panggilan (call center) dalam jumlah besar untuk menjalankan kampanye tersebut.

# THE PROBLEM

Subscribed vs Not Subscribed



Dari seluruh nasabah yang sudah dikontak dengan telepon untuk promosi campaign berkaitan dengan deposito berjangka (Term Deposito), hanya sebagian kecil yaitu 11.70% yang akhirnya memutuskan untuk berlangganan (subscribed).

Sehingga diperlukan optimasi untuk **mengidentifikasi calon nasabah yang paling berpotensi** untuk berlangganan deposito sebelum kampanye dijalankan, agar upaya pemasaran lewat telepon bisa difokuskan hanya pada target yang tepat dan biaya bisa ditekan.

# WHAT WE CAN DO?

## GOALS

- Membuat model Machine Learning (ML) untuk memprediksi calon nasabah yang potensial berlangganan Deposito Berjangka (Term Deposito)
- Membantu tim marketing mengurangi cost campaign.
- 

## OBJECTIVES

- Meningkatkan persentase nasabah baru untuk berlangganan lebih dari 8% dari 11.70% menjadi sekitar 20%.

## BUSSINESS METRICS

- Subscription nasabah
- Jumlah berapa kali nasabah dikontak

**STAGE 1**

---

**DATA EXPLORATION**

# DESCRIBE THE DATASET

Features dan Target

age	housing	campaign
job	loan	pdays
marital	contact	previous
education	day	poutcome
default	month	y
balance	duration	

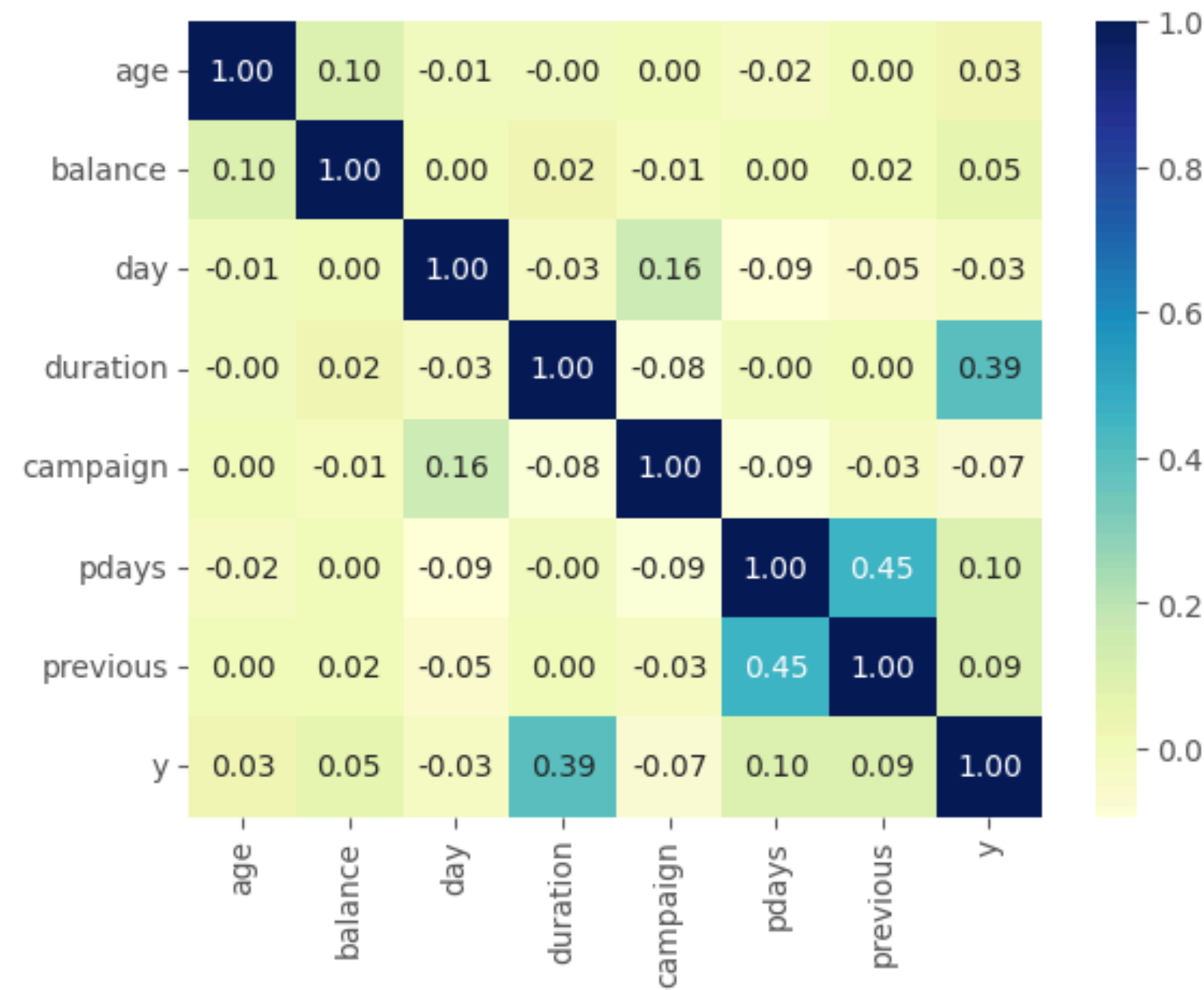
- Jumlah baris 45211.
- Kolom tipe numerik sebanyak 7 & kolom tipe categoric (termasuk kolom target) 10.
- Kolom target 'y' diubah namanya menjadi 'subscribed' supaya lebih informatif.
- Kolom 'y' atau 'subscribed' dijadikan kolom target.
- Tidak ada null values.
- Tidak ada data yang duplikat.
- 
- df\_test merupakan duplikat dari df\_train.



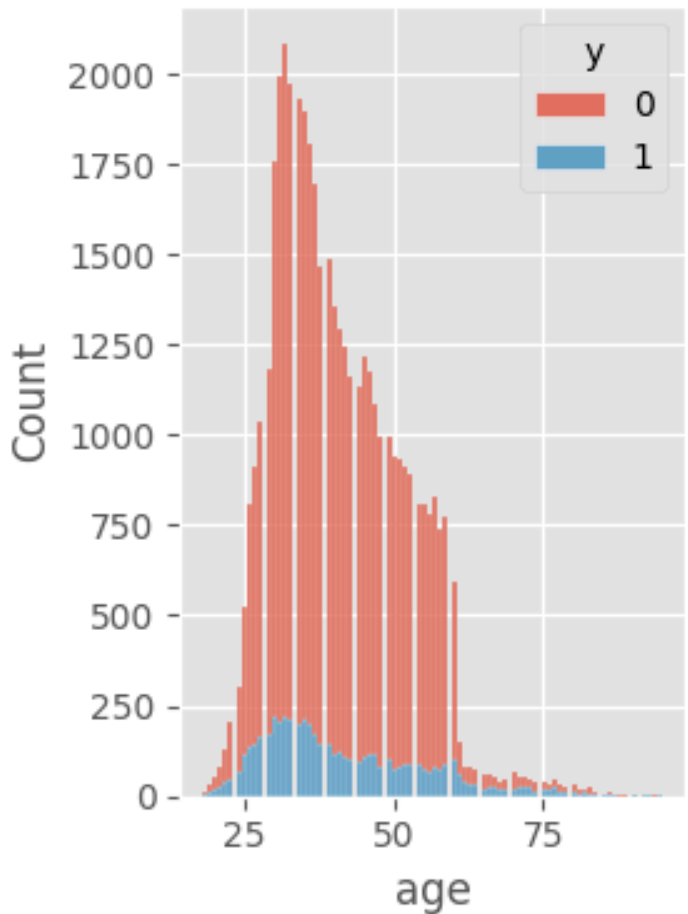
# EXPLORATORY DATA ANALYSIS

Correlation Feature & Distributions Data by Target 'subscribed'

Heatmap Correlation - Numerical Features



- Tidak ada korelasi yang kuat antara target y dengan feature numerikal yang ada, mengindikasikan penggunaan model non-linear lebih tepat untuk dataset ini. Nilai korelasi tertinggi target-feature ada pada y dengan duration, yaitu 0.39 (tidak cukup kuat).

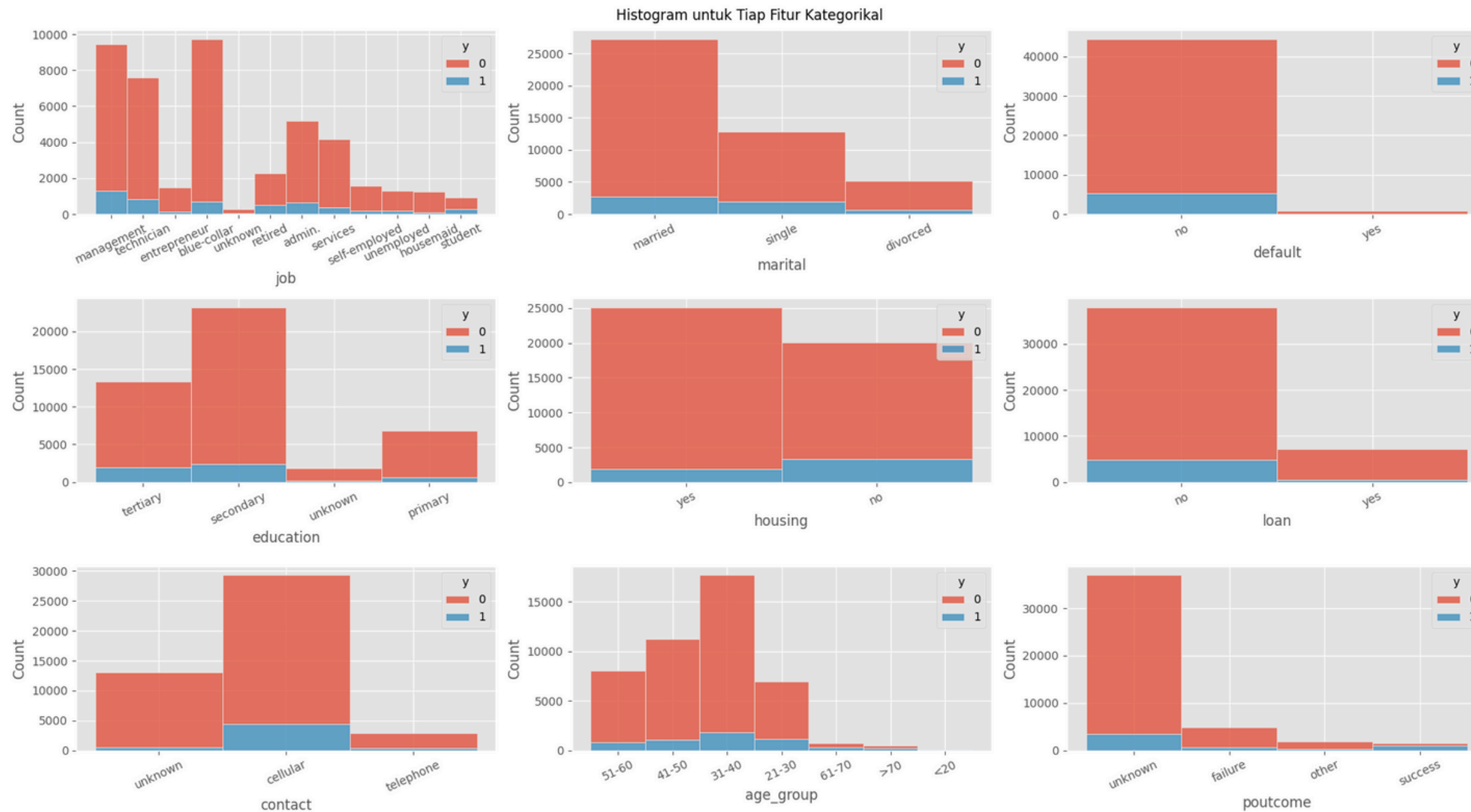


- Untuk selanjutnya, kolom 'age' dibuat menjadi kolom kategorik untuk melihat sebaran subscription.
- Beberapa kolom lain juga di-grouping menjadi categorical columns untuk mendapatkan insight.



# EXPLORATORY DATA ANALYSIS

## Correlation Feature & Distributions Data by Target 'subscribed'



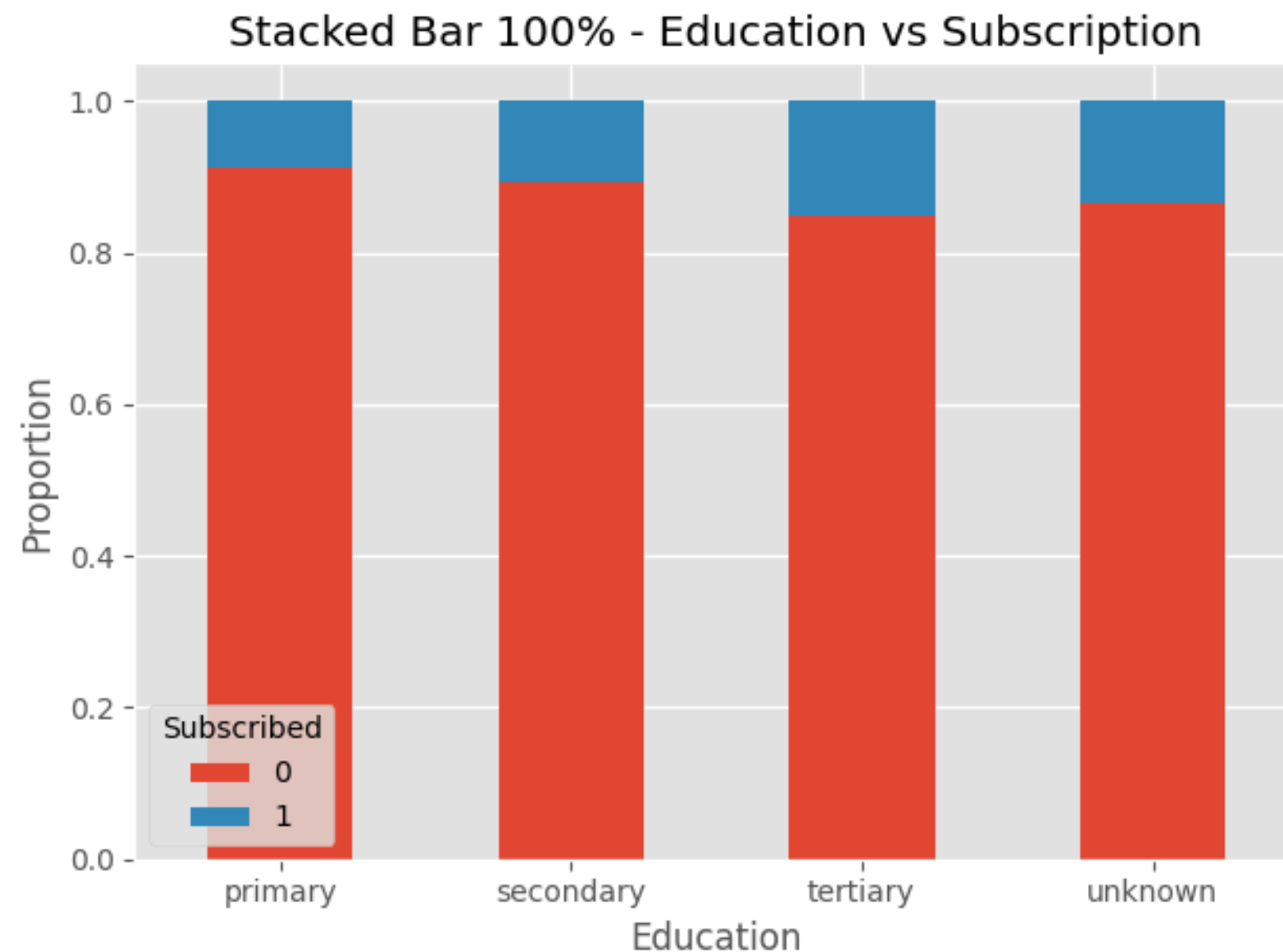
Kecenderungan nasabah yg

membuka Deposito Berjangka :

- memiliki job management, technician dan blue-collar
- memiliki status menikah
- tidak pernah gagal membayar cicilan (default = 'no')
- memiliki tingkat pendidikan lanjutan: secondary dan tertiary
- tidak memiliki cicilan personal
- dikontak melalui telepon selular
- age group 21-60 (usia produktif)
- memiliki rata-rata balance positif (negative balance mungkin mengindikasikan ybs memiliki hutang, rata-rata saldo zero dapat mengindikasikan ybs menggunakan akun banknya hanya untuk transaksi dan tidak ada uang berlebih)

# EXPLORATORY DATA ANALYSIS

## Business Insights



## Target Utama Nasabah Potensial

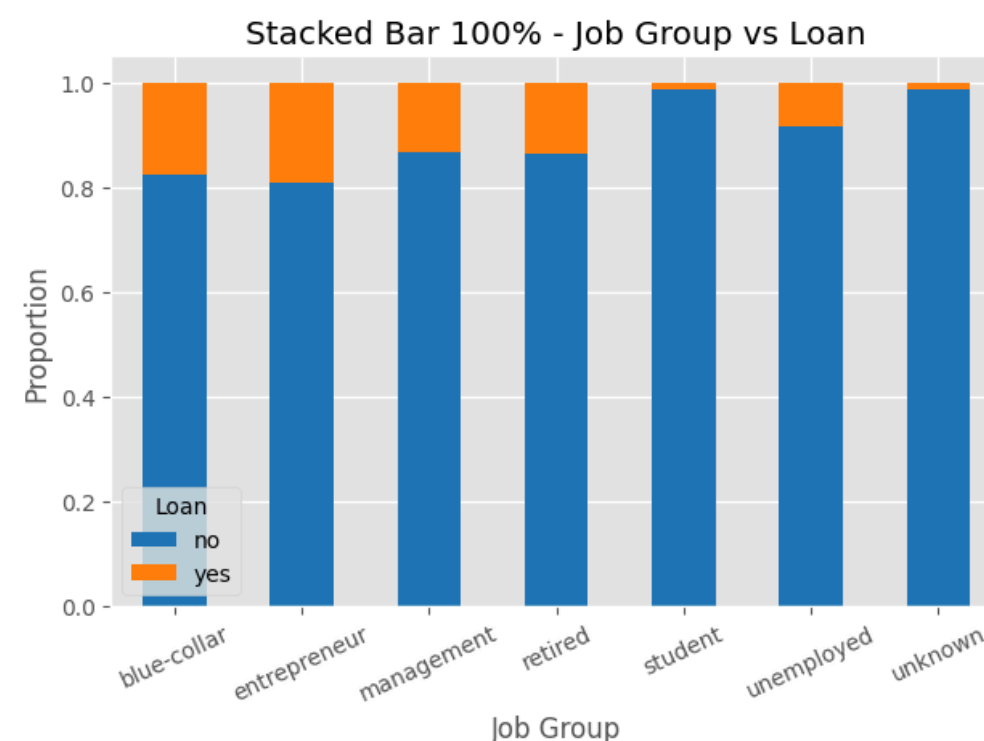
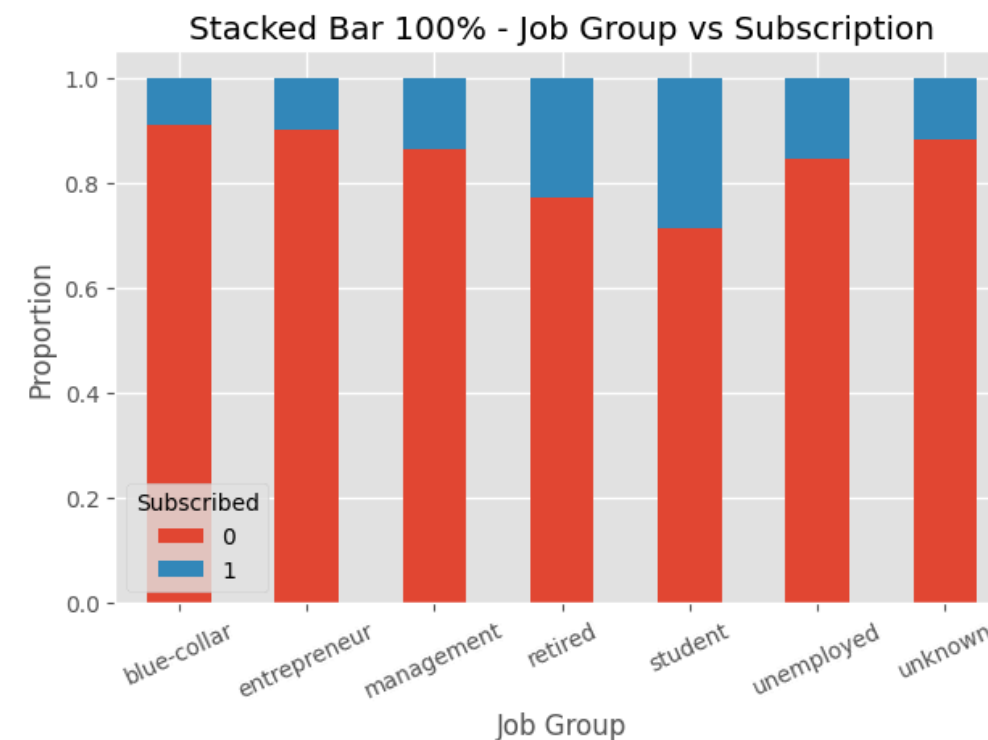
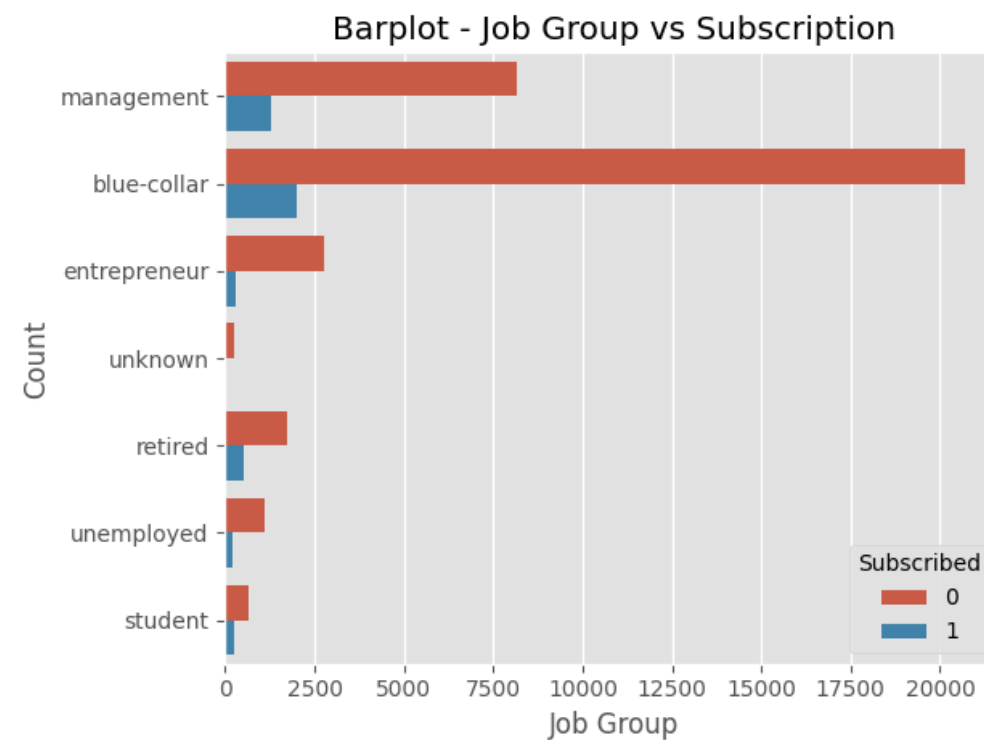
1. Usia produktif (20–60 tahun)
2. Status menikah
3. Pendidikan menengah & tinggi (secondary & tertiary)
4. Tidak memiliki cicilan rumah (housing loan)
5. Tidak memiliki hutang pribadi (personal loan). Fokus telemarketing dapat diarahkan ke segmen ini.

## Pendidikan

- Mayoritas yang ambil term deposit punya pendidikan secondary atau tertiary.
- Semakin tinggi pendidikan, semakin besar peluang mengambil term deposit.
- Saran: Gunakan pendekatan yang lebih sederhana dan mudah dimengerti untuk nasabah berpendidikan rendah.

# EXPLORATORY DATA ANALYSIS

## Business Insights

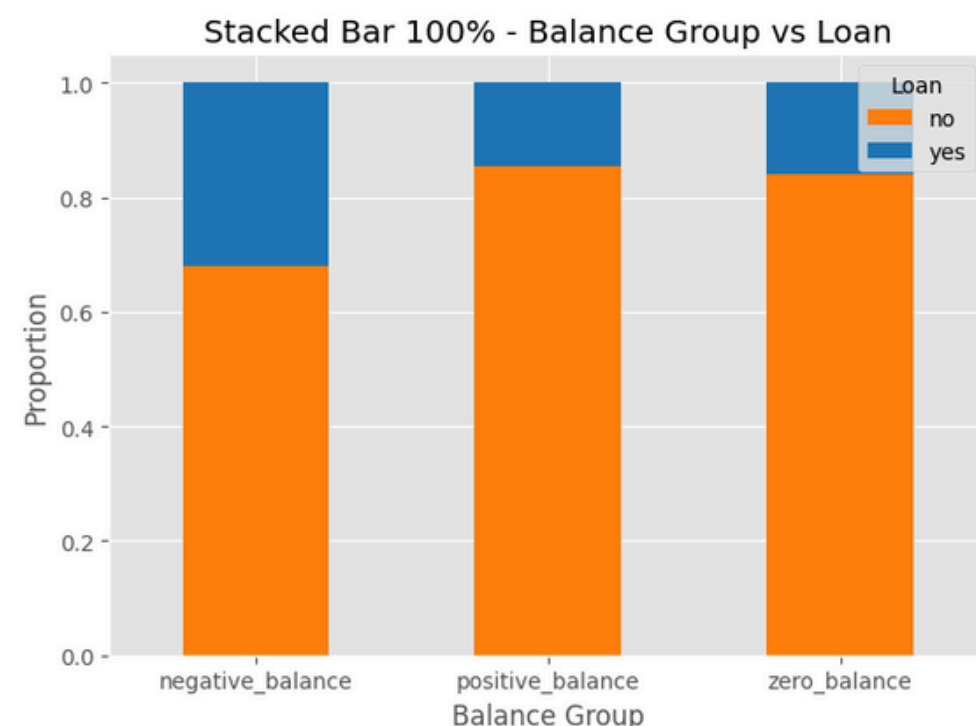
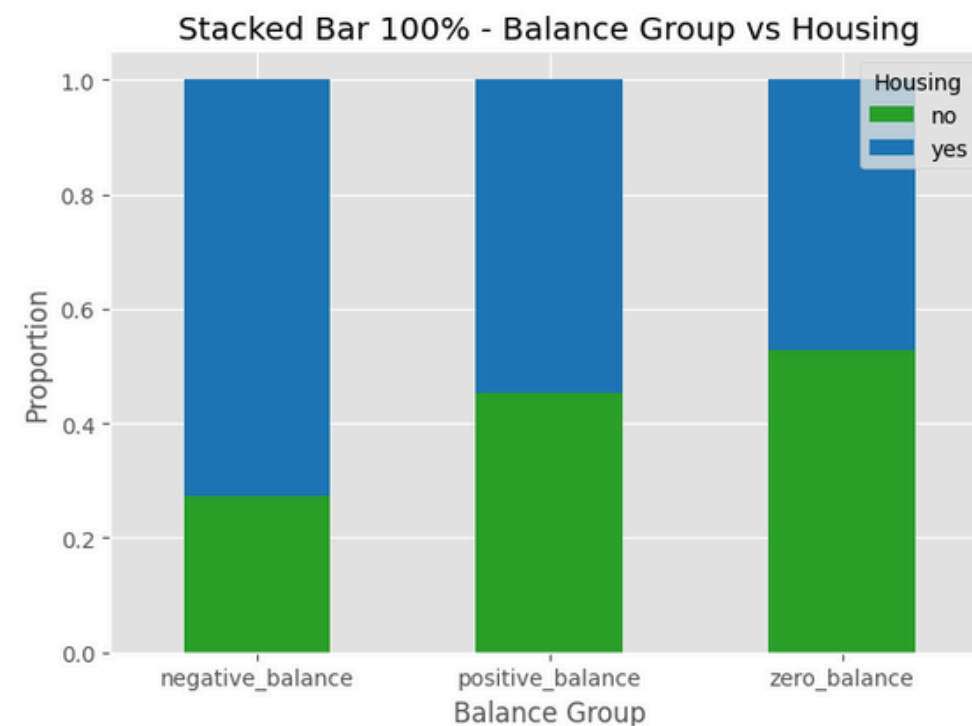
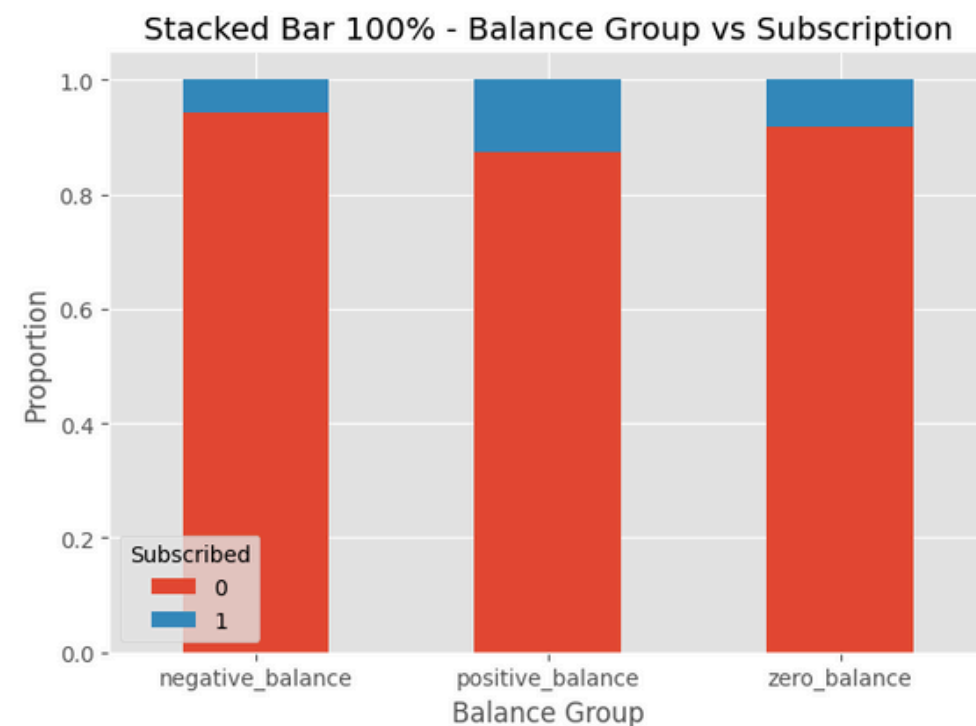


Insight dari Jenis Pekerjaan dan Konversi Term Deposit:

- Pekerjaan terbanyak yang ambil term-deposit: Mayoritas berasal dari blue-collar dan management.
- Tingkat konversi rendah pada blue-collar:
  - Meskipun banyak jumlahnya, persentase yang mengambil term deposit (yes) rendah.
  - Perlu evaluasi ulang pendekatan marketing untuk segmen ini.
- Tingkat konversi tinggi pada student dan retired:
  - Persentase 'yes' tinggi di segmen mahasiswa dan pensiunan.
  - Disarankan buat produk & pendekatan khusus untuk dua segmen ini agar potensi mereka maksimal.
- Entrepreneur: konversi rendah, banyak pinjaman pribadi:
  - Kemungkinan lebih butuh modal usaha daripada simpanan.
  - Mungkin lebih cocok ditawarkan produk pinjaman (loan) daripada term-deposit.
- Rekomendasi utama:
  - Fokus pendekatan personal & produk yang sesuai untuk masing-masing segmen agar kampanye lebih efektif.

# EXPLORATORY DATA ANALYSIS

## Business Insights



## Insight dari Average Yearly Balance:

### 1. Saldo Tahunan Negatif:

- Sebaiknya tidak menjadi target telemarketing.
- Mereka memiliki % konversi term-deposit paling rendah.
- Sekitar 70% memiliki kredit rumah dan 30% memiliki kredit pribadi.
- Hal ini menunjukkan mereka tidak memiliki dana menganggur (cold money) untuk diinvestasikan.

### 2. Saldo Tahunan = 0:

- Akun mereka cenderung digunakan hanya untuk transaksi harian (transfer, bayar tagihan).
- Kemungkinan mereka menyimpan dana di institusi keuangan lain.
- Jika memungkinkan, gunakan data tambahan seperti dari SLIK untuk mengetahui potensi aset mereka.
- Buatlah program atau penawaran khusus agar mereka tertarik memindahkan investasinya ke bank.

### 3. Rekomendasi:

- Fokuskan kampanye telemarketing ke nasabah dengan saldo positif dan tidak memiliki banyak pinjaman.

**STAGE 2**

---

**DATA**

**PREPROCESSING**

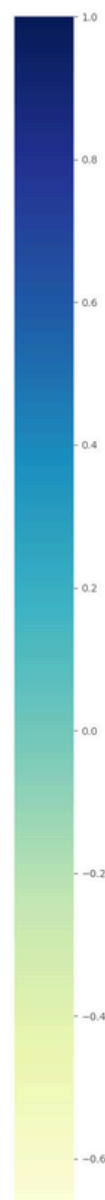
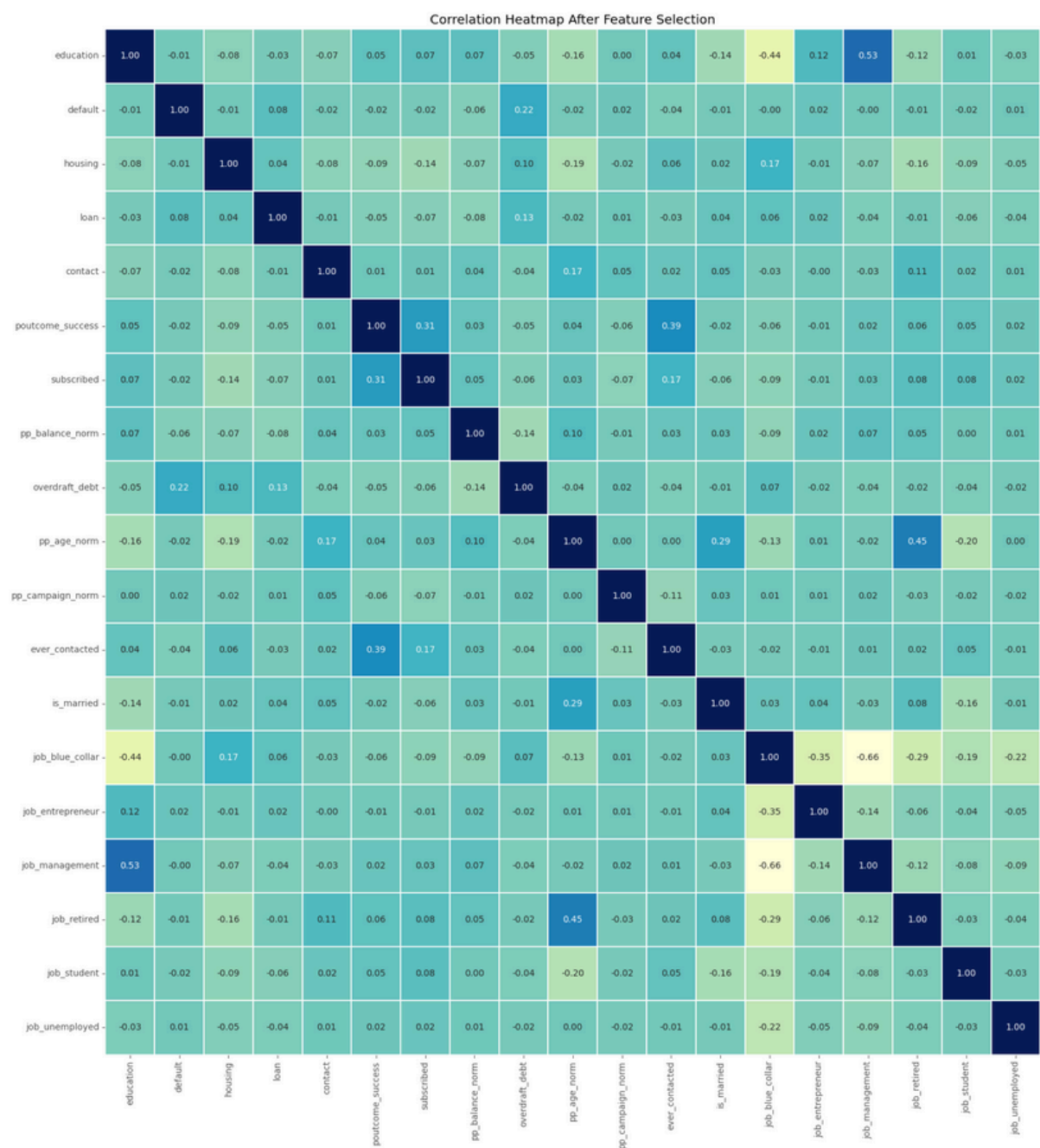
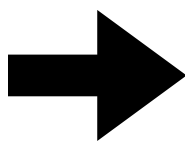
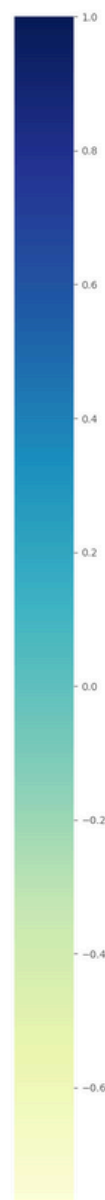
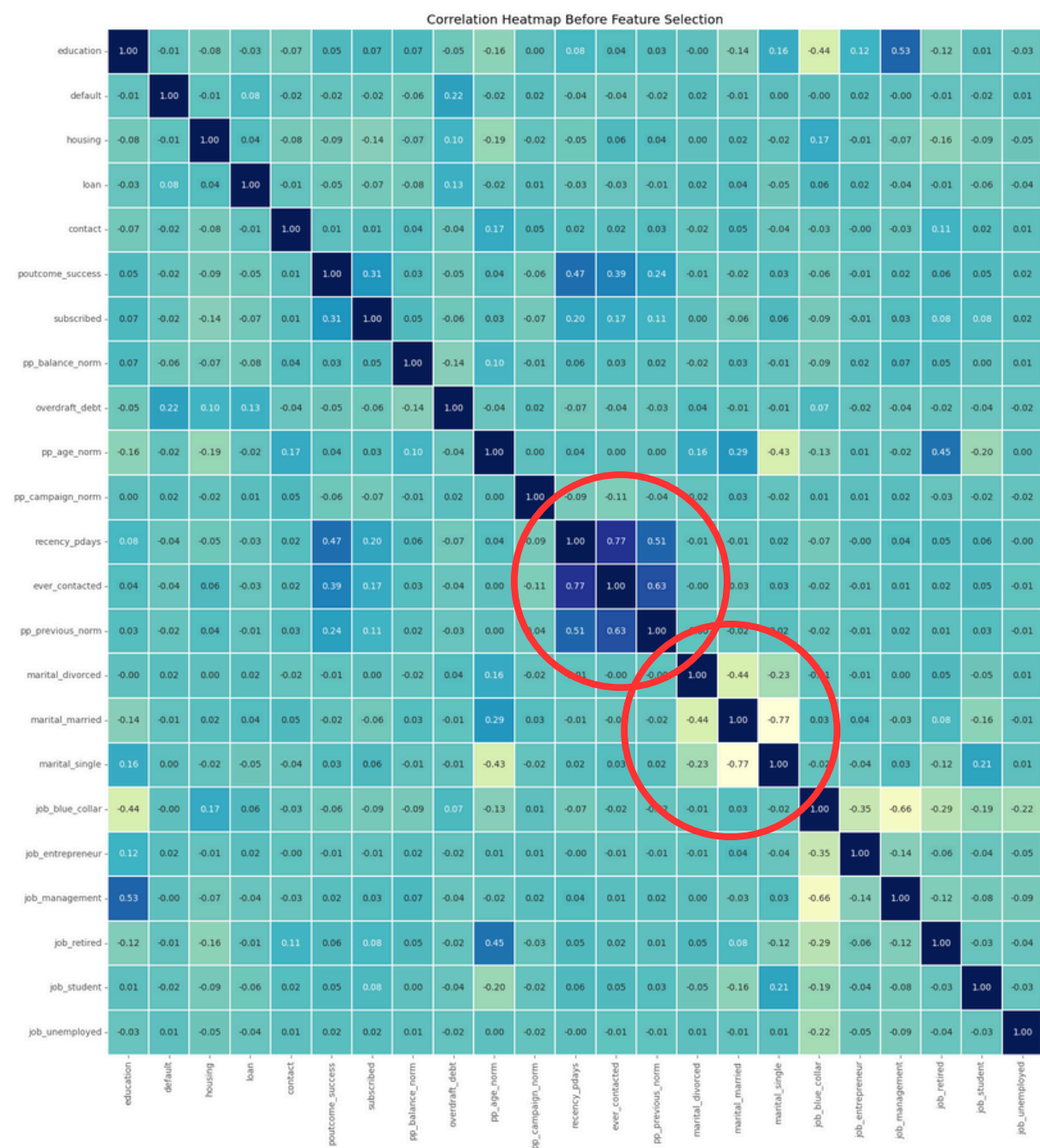
# DATA PREPROCESSING

Feature	Type	Preprocessing I	Preprocessing II	Preprocessing III
age	num	Handle outliers	Feature transformation (MinMaxScaler)	
job	cat	Handle missing values	Feature extraction (job_group)	Feature encoding (OHE job_group)
marital	cat	Feature Encoding (Label encoding)		
education	cat	Handle missing values	Feature encoding	
default	cat	Feature Encoding (Label encoding)		
balance	num	Handle outliers	Feature transformation	
housing	cat	Feature Encoding (Label encoding)		
loan	cat	Feature Encoding (Label encoding)		
contact	cat	Handle missing values	Feature Encoding (Label encoding)	
day	num	Handle outliers	Feature transformation	DROP FEATURE
month	cat	Feature Encoding (Label encoding)	Feature extraction	DROP FEATURE
duration	num	Handle outliers	Feature transformation	Feature extraction + DROP
campaign	num	Handle outliers	Feature transformation	
pdays	num	Feature selection (Drop feature)		DROP FEATURE
previous	num	Feature selection (Drop feature)		DROP FEATURE
poutcome	cat	Feature Encoding (Label encoding)		
y	cat	Feature encoding	Handle class imbalance	

- Drop beberapa feature ('day', 'month', 'duration', 'pdays', 'previous') karena bisa menyebabkan multikolinearitas



# DATA PREPROCESSING



Drop beberapa feature ('day', 'month', 'duration', 'pdays', 'previous') karena bisa menyebabkan multikolinearitas



# DATA PREPROCESSING

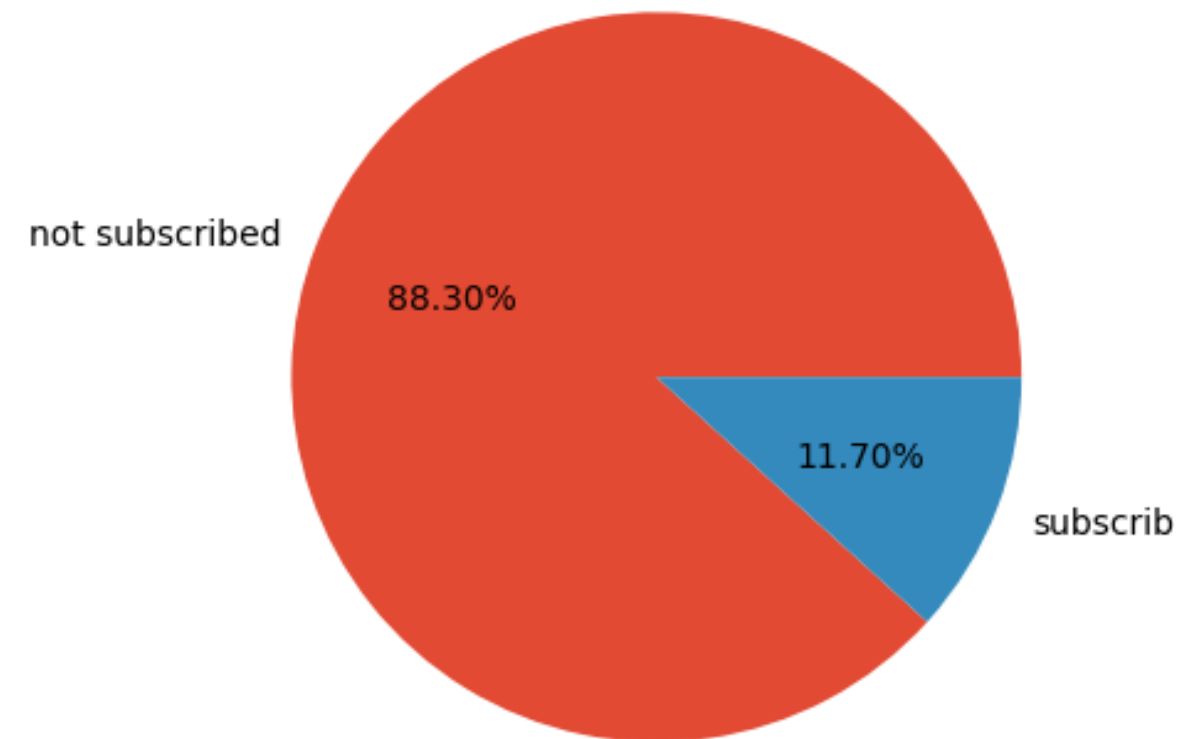
Original  
y\_class  
False 39922  
True 5289  
Name: count, dtype: int64

UNDERSAMPLING  
y\_class  
False 10578  
True 5289  
Name: count, dtype: int64

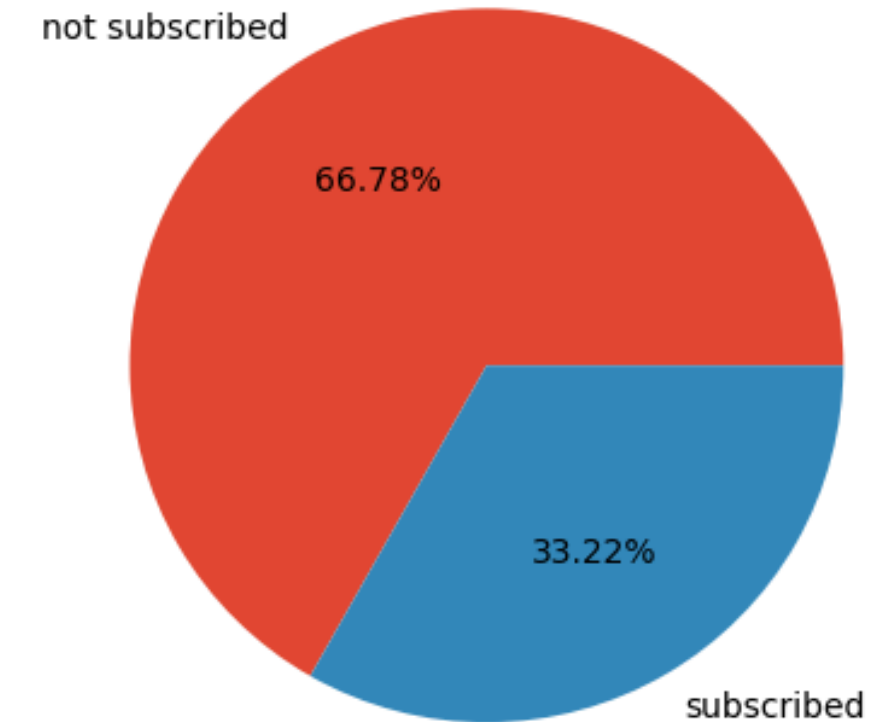
OVERSAMPLING  
y\_class  
False 39922  
True 19961  
Name: count, dtype: int64

SMOTE  
y\_class  
False 39922  
True 19961  
Name: count, dtype: int64

Subscribed vs Not Subscribed



Subscribed vs Not Subscribed (After Handle Class Imbalance)



handle class imbalance untuk mendapatkan data target yang lebih baik

**STAGE 3**

---

**MODELING**

# TRIAL MODEL

Metrik yang digunakan : Recall

	Model	Accuracy_train	Accuracy_test	Precision_train	Precision_test	Recall_train	Recall_test	F1-Score_train	F1-Score_test	ROC AUC_test
0	Decision Tree	0.9980	0.7928	1.0000	0.6924	0.9941	0.6956	0.9971	0.6940	0.7696
1	Random Forest	0.9980	0.8431	0.9984	0.8099	0.9956	0.6997	0.9970	0.7507	0.8944
2	Logistic Regression	0.7321	0.7297	0.7369	0.7440	0.3011	0.3046	0.4275	0.4323	0.7254
3	KNN	0.8551	0.7792	0.7782	0.6717	0.7888	0.6773	0.7835	0.6745	0.8398
4	XGBoost	0.8600	0.8417	0.9054	0.8777	0.6462	0.6174	0.7542	0.7249	0.8752
5	Adaboost	0.7901	0.7877	0.8086	0.8210	0.4826	0.4750	0.6044	0.6018	0.8057

Model XGBoost memiliki nilai Recall yang cukup besar dan tidak overfit antara data train dengan data test

# TUNNING MODEL XGBOOST

Metrik yang digunakan : Recall

```
# Predict & Evaluation  
eval_classification(xg_tuned)
```

```
⇒ Accuracy (Train Set): 0.9045  
Accuracy (Test Set): 0.8771
```

```
Precision (Train Set): 0.9420  
Precision (Test Set): 0.8988
```

```
Recall (Train Set): 0.7595  
Recall (Test Set): 0.7169
```

```
F1-Score (Train Set): 0.8409  
F1-Score (Test Set): 0.7976  
roc_auc (train-proba): 0.9545  
roc_auc (test-proba): 0.9075
```

Setelah dilakukan hyperparameter tuning, didapatkan nilai prediksi yang lebih baik, khususnya dari metrik Recall

# TRIAL MODEL

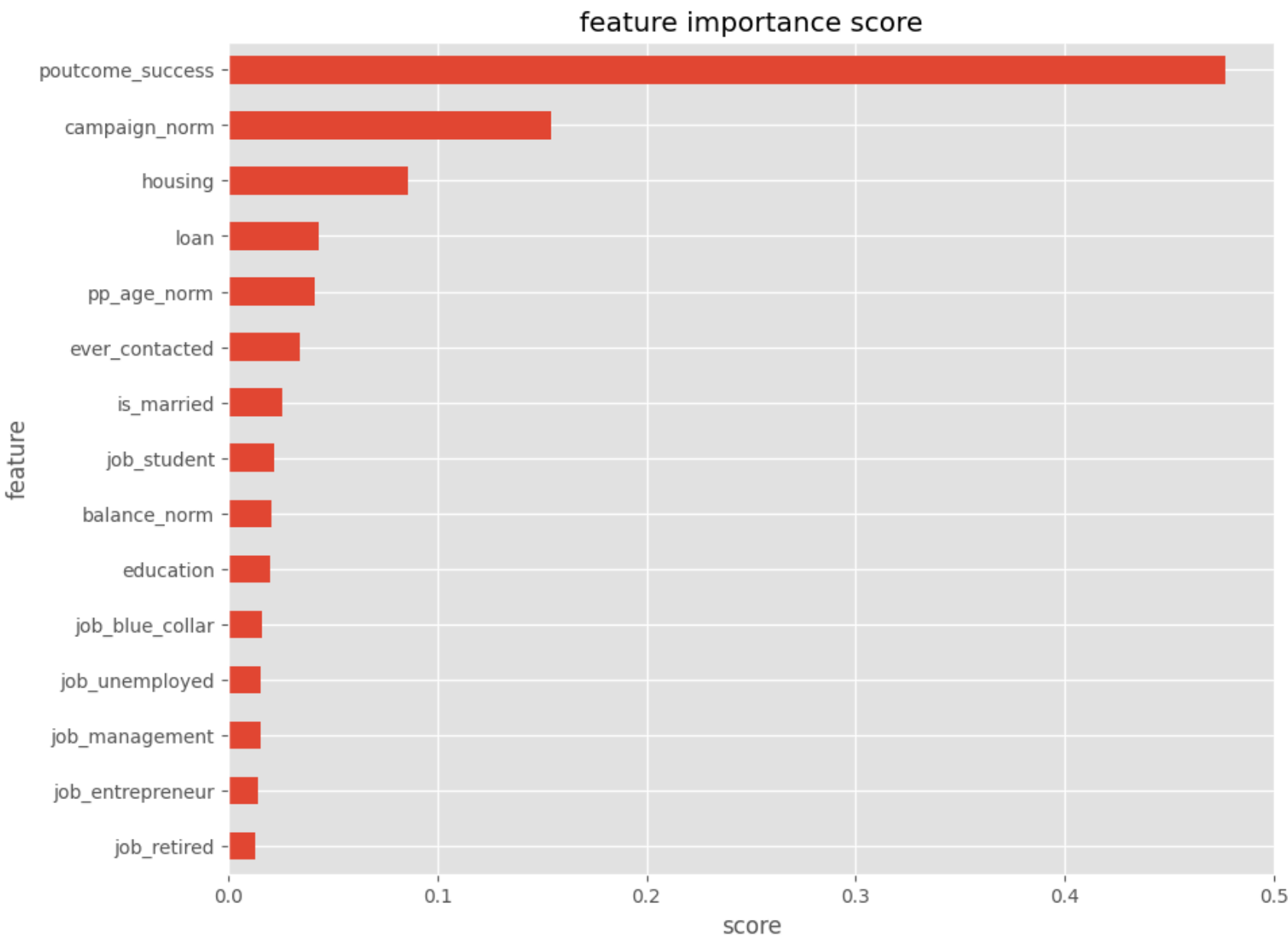
## Metrik yang digunakan : Recall

	Model	Accuracy_train	Accuracy_test	Precision_train	Precision_test	Recall_train	Recall_test	F1-Score_train	F1-Score_test	ROC AUC_test
0	Decision Tree PCA	0.9980	0.7568	1.0000	0.6369	0.9941	0.6509	0.9971	0.6438	0.7315
1	Random Forest PCA	0.9980	0.8034	0.9979	0.7331	0.9962	0.6570	0.9971	0.6930	0.8597
2	Logistic Regression PCA	0.7253	0.7188	0.7053	0.6935	0.2976	0.2998	0.4185	0.4186	0.7036
3	KNN PCA	0.8528	0.7748	0.7761	0.6653	0.7825	0.6707	0.7793	0.6680	0.8330
4	XGBoost PCA	0.8024	0.7610	0.8119	0.7338	0.5273	0.4590	0.6393	0.5647	0.8034
5	Adaboost PCA	0.7247	0.7186	0.7882	0.7749	0.2345	0.2351	0.3615	0.3607	0.7085

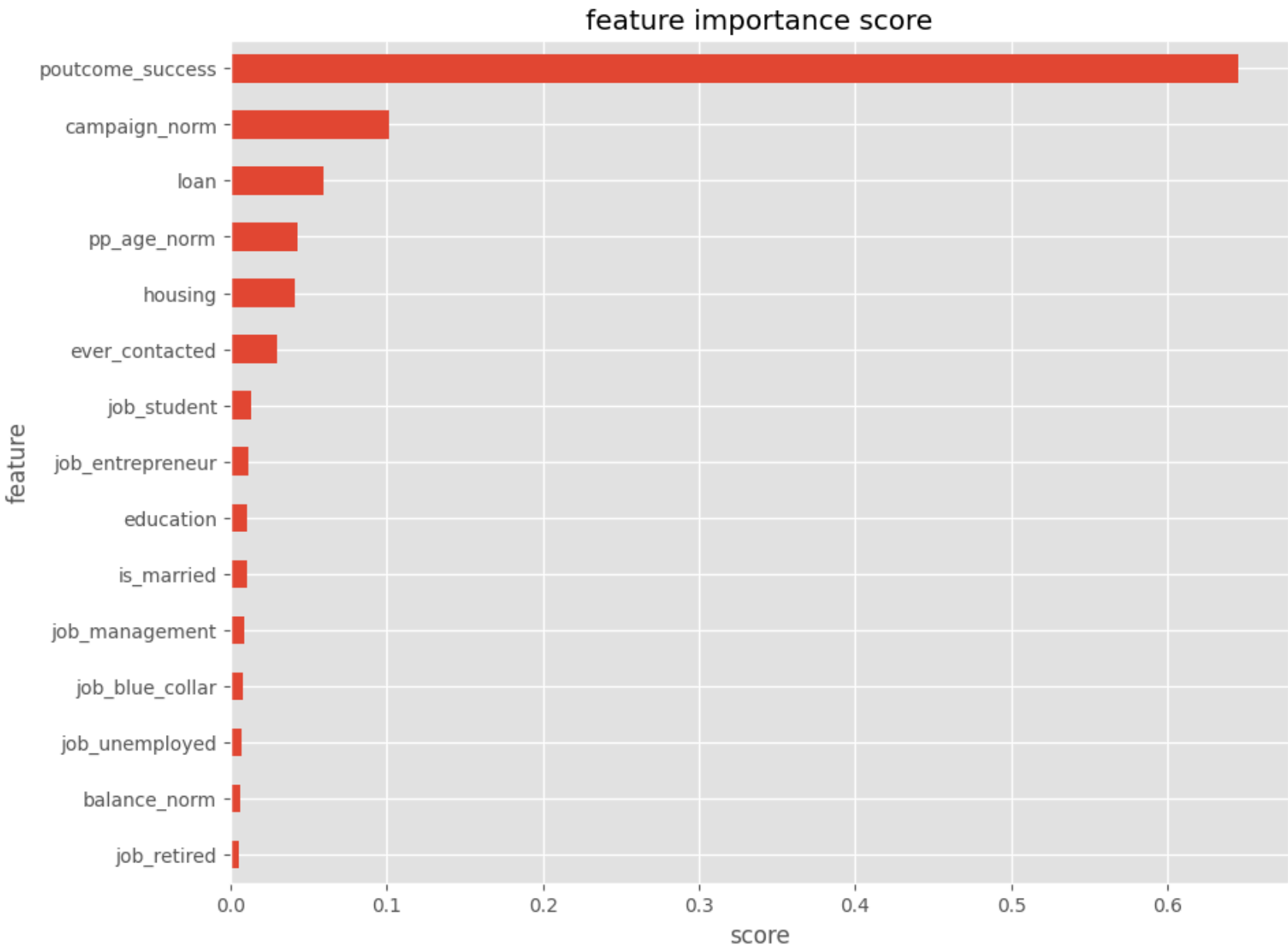
Trial model dengan menggunakan PCA (Principal Component Analysis) untuk melakukan dimensionality reduction, namun ternyata nilai Recall justru turun dan mengalami overfit. Sehingga model XGBoost yang di-PCA tidak digunakan.

# FEATURE IMPORTANCE MODEL XGBOOST

Sebelum Tuning



Setelah Tuning



# MODEL RESULTS

Simulasi dari data yg ada

Total nasabah	45211	
Persentase subscription	11.70	%
Total subscribed	5290	
Recall	71.69	%
Total nasabah yang diprediksi potensial "subscribed"	32412	



# RECOMMENDATIONS

---



# THAT'S ALL

Thank you

