

Exploration of Bank Direct Marketing Dataset

Predicting Subscription to Term Deposits



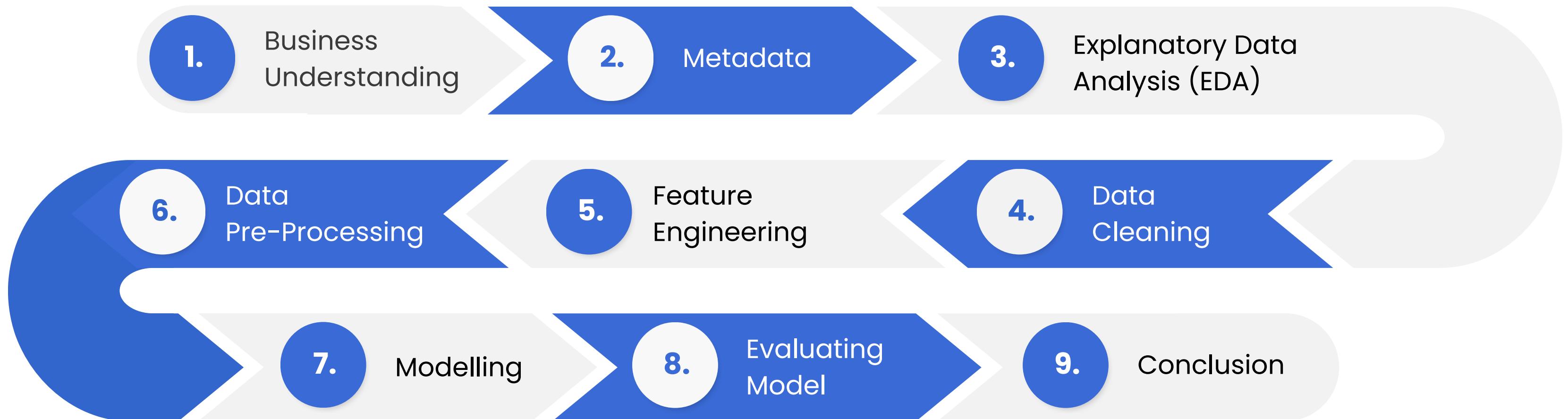
Dataset about bank direct marketing campaigns, containing client and campaign-related attributes used to predict whether a customer will subscribe to a term deposit.

by Data Snipers Team

tinyurl.com/DashboardDataSnipers



Outline



1

Business Understanding



Data Nasabah

**age** (float64)

usia nasabah

**job** (object)

jenis pekerjaan nasabah: 12 kategori

**marital** (object)

status pernikahan: married, divorce, single

**education** (object)

tingkat pendidikan: unknown, secondary, primary, tertiary

**loan** (object)

memiliki pinjaman pribadi? (ya/tidak)

**housing** (object)

memiliki pinjaman rumah? (ya/tidak)

**balance** (float64)

Saldo rata-rata tahunan (dalam Euro)

**default** (object)

memiliki kredit macet? (ya/tidak)

2

Metadata

Data Kontak Terakhir dalam Kampanye Saat Ini

**contact** (object)

jenis komunikasi: unknown, celuler, telephone

**duration** (float64)

durasi kontak terakhir (dalam detik)

**day** (float64)

hari terakhir kontak dalam bulan tersebut

**month** (object)

bulan terakhir kontak

**campaign** (float64)

jumlah kontak selama kampanye ini

2

Metadata

Atribut Tambahan

**previous (float64)**

jumlah kontak sebelum kampanye ini

**pdays (float64)**

jumlah hari sejak kontak terakhir dari kampanye sebelumnya (-1 jika belum pernah dikontak)

**poutcome (object)**

hasil kampanye sebelumnya: unknown, other, failed, success

Variabel Target

**Subscription Status (object)**

status langganan nasabah terhadap deposito berjangka (ya/tidak)

3

Explanatory Data Analysis (EDA)

Statistik Deskriptif

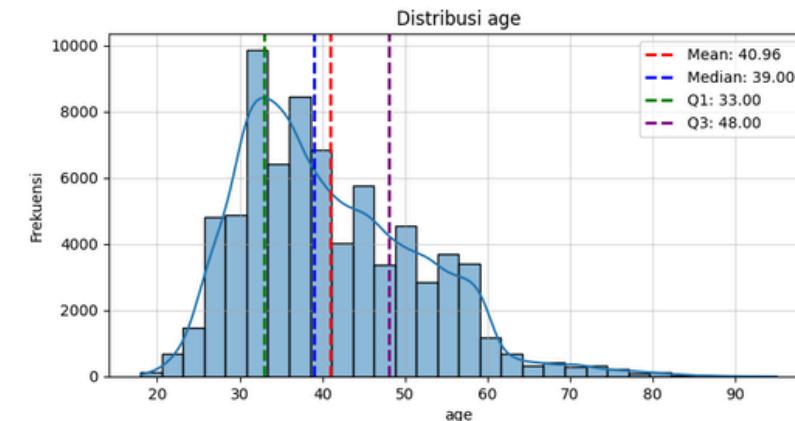
Variabel Numerik

	age	balance	day	duration	campaign	pdays	previous
mean	40,95586649	1500,089164	15,41475601	353,3474836	2,43586877	49,32791349	0,71146701
std	10,86696309	3116,42686	7,90918208	336,530792	2,6999849	106,3617404	2,16683426
mode	32	0	20	124	1	-1	0
min	18	-8019	1	0	1	-1	0
25%	33	113	9	134	1	-1	0
50%	39	533	15	239	2	-1	0
75%	48	1685	21	461	3	4	0
max	95	102127	31	4918	63	871	275

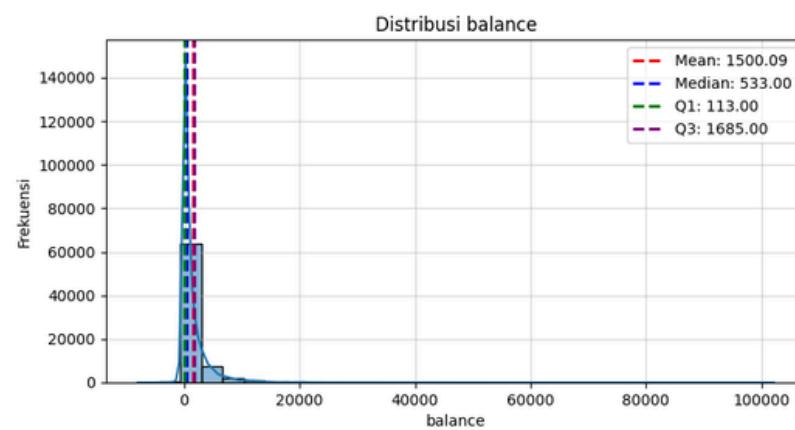
Variabel Kategorik

	job	marital	education	default	housing	loan	contact	month	poutcome	subscription status
unique	12	3	4	2	2	2	3	12	4	2
mode	management	married	secondary	no	yes	no	cellular	may	unknown	no
freq	17085	47964	41821	74184	41947	67446	53876	18604	57058	43921

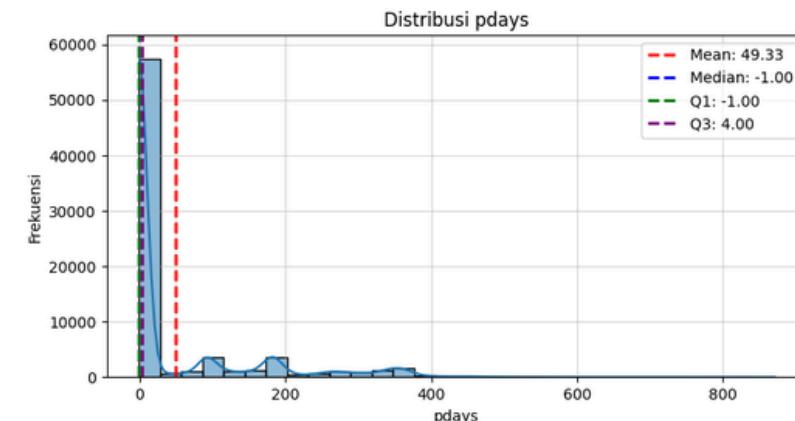
Distribusi Variabel Numerik



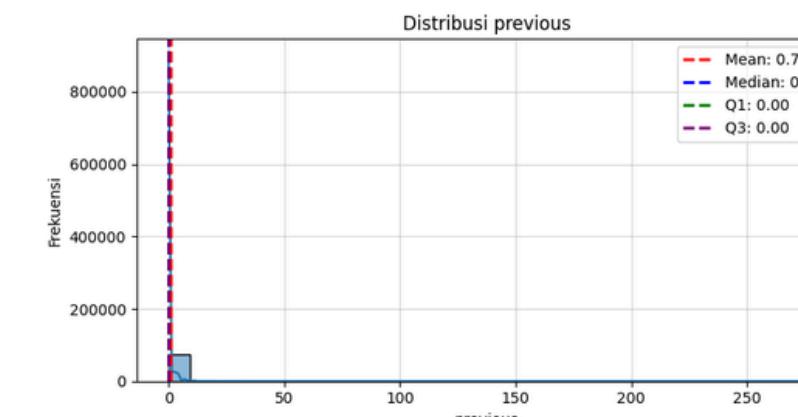
- Distribusi **age** *positively skewed*
- rentang usia nasabah: 18–95 tahun
- Mayoritas usia: 33–48 tahun
- Rata-rata: kisaran 40 tahun



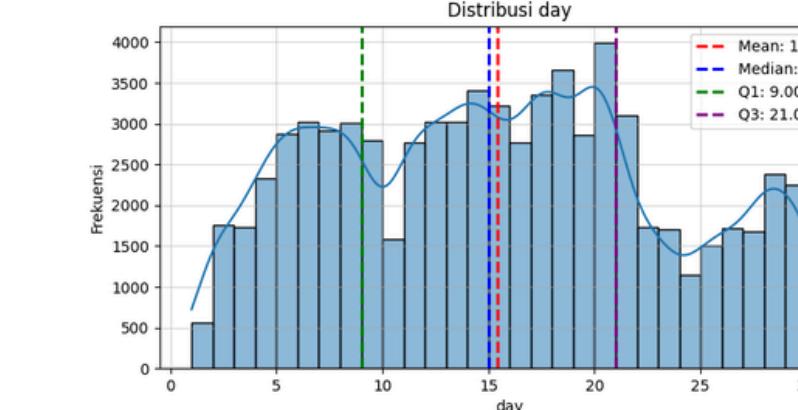
- Distribusi **balance**, *positively skewed*
- Mayoritas nasabah memiliki saldo rendah
- ada beberapa outlier dengan saldo yang sangat tinggi, menunjukkan potensi segmen nasabah premium



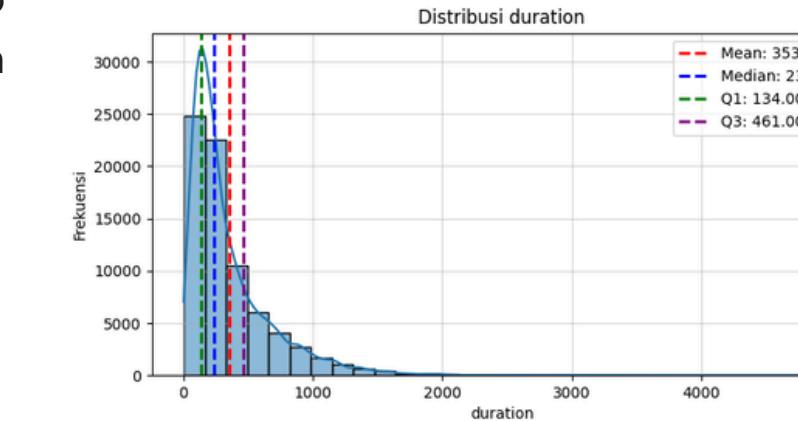
- Distribusi **pdays**, *truncated* pada -1
- Mayoritas nasabah belum pernah dihubungi sebelumnya, yang berarti sebagian besar adalah prospek baru bagi bank.



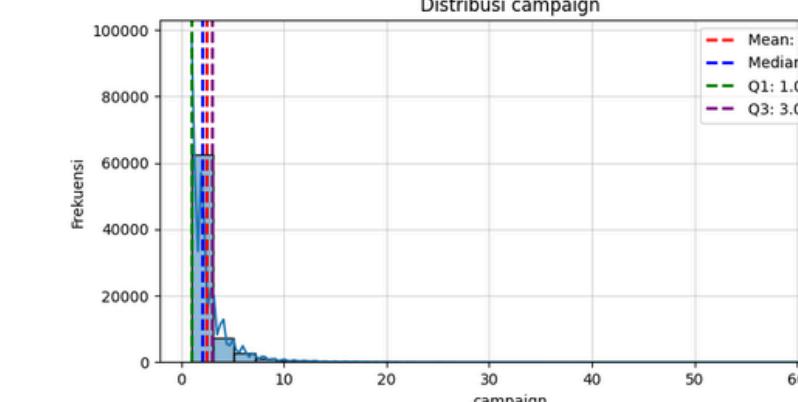
- Distribusi **previous**, *truncated* pada 0
- Sebagian besar nasabah tidak memiliki riwayat kontak dalam kampanye sebelumnya, menandakan bahwa pendekatan pemasaran belum banyak mengandalkan retensi pelanggan.



- Distribusi **day** tidak beraturan
- Kontak dilakukan secara merata sepanjang bulan
- Sedikit cenderung lebih banyak terjadi di pertengahan bulan.



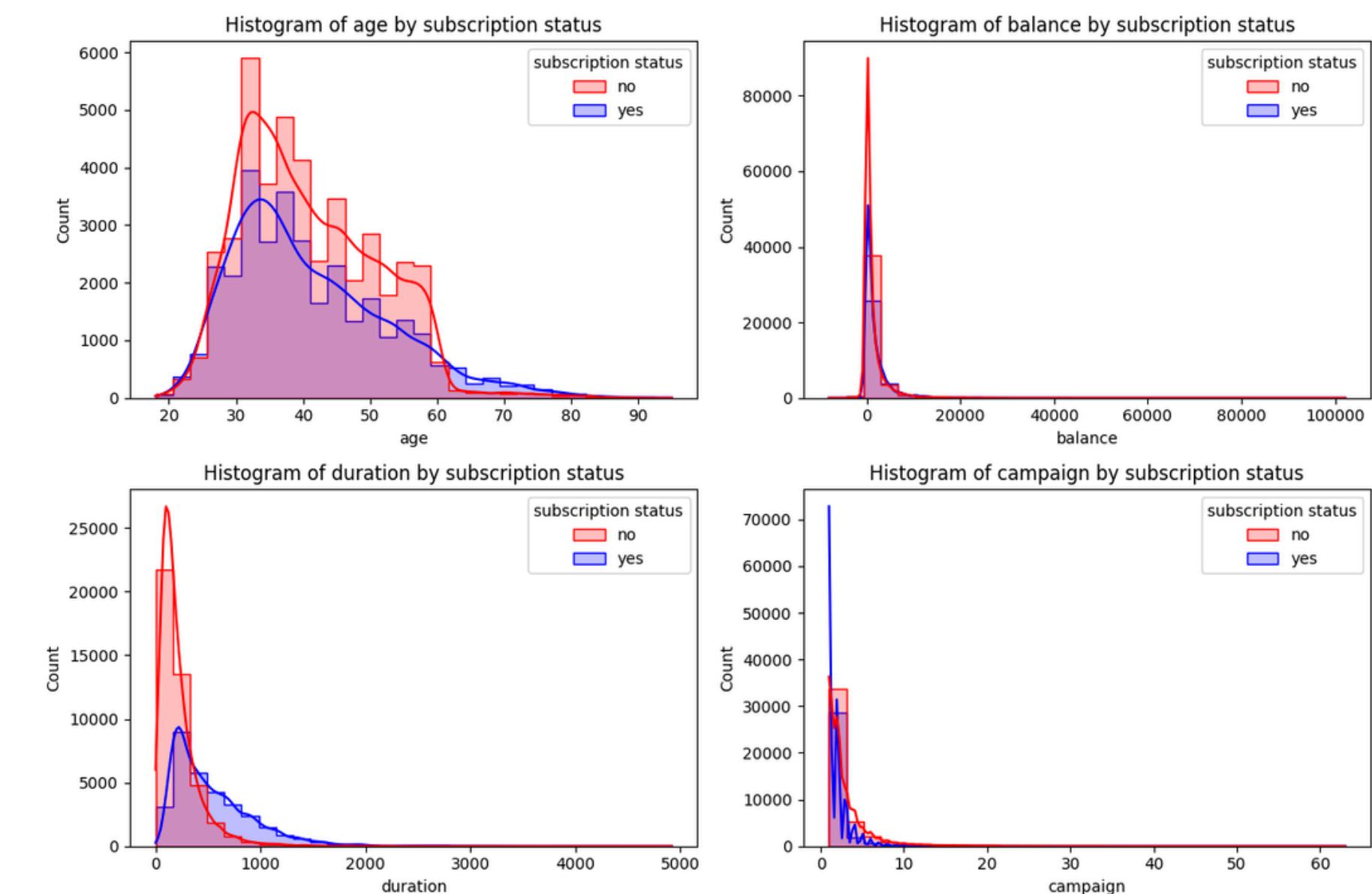
- Distribusi **duration** *positively skewed*
- Sebagian besar panggilan selama kampanye berlangsung singkat sekitar di bawah 500 detik atau di bawah 8–9 menit.



- Distribusi **campaign** *truncated* pada 1
- Mayoritas nasabah dihubungi hanya 1–4 kali dalam satu kampanye, menunjukkan pendekatan pemasaran yang relatif terbatas.

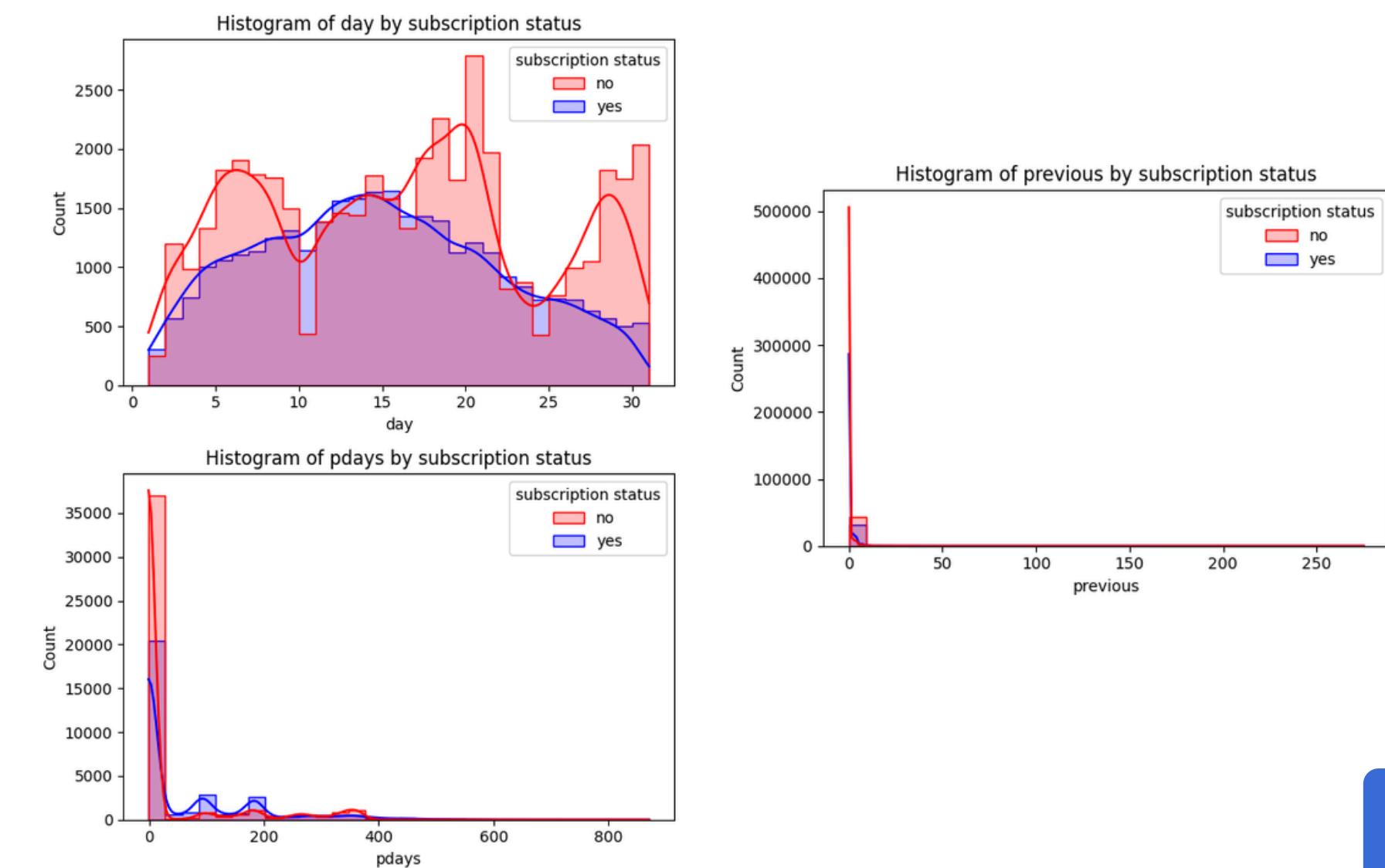
Distribusi Variabel Numerik Berdasarkan Target

- **Histogram Age by Label :** Kelompok 60-80 tahun lebih cenderung berlangganan deposito berjangka
- **Histogram Balance by Label :** Sebagian besar saldo berada di nilai rendah, dengan distribusi yang sangat miring ke kanan. Tidak ada perbedaan signifikan antara pelanggan yang berlangganan dan tidak.
- **Histogram Duration by Label :** Durasi panggilan lebih lama meningkatkan peluang berlangganan. Namun, durasi ekstrem (>3000 detik) justru menurunkan kemungkinan tersebut.
- **Histogram Campaign by Label :** Frekuensi kontak yang tinggi (>30) mayoritas tidak berlangganan

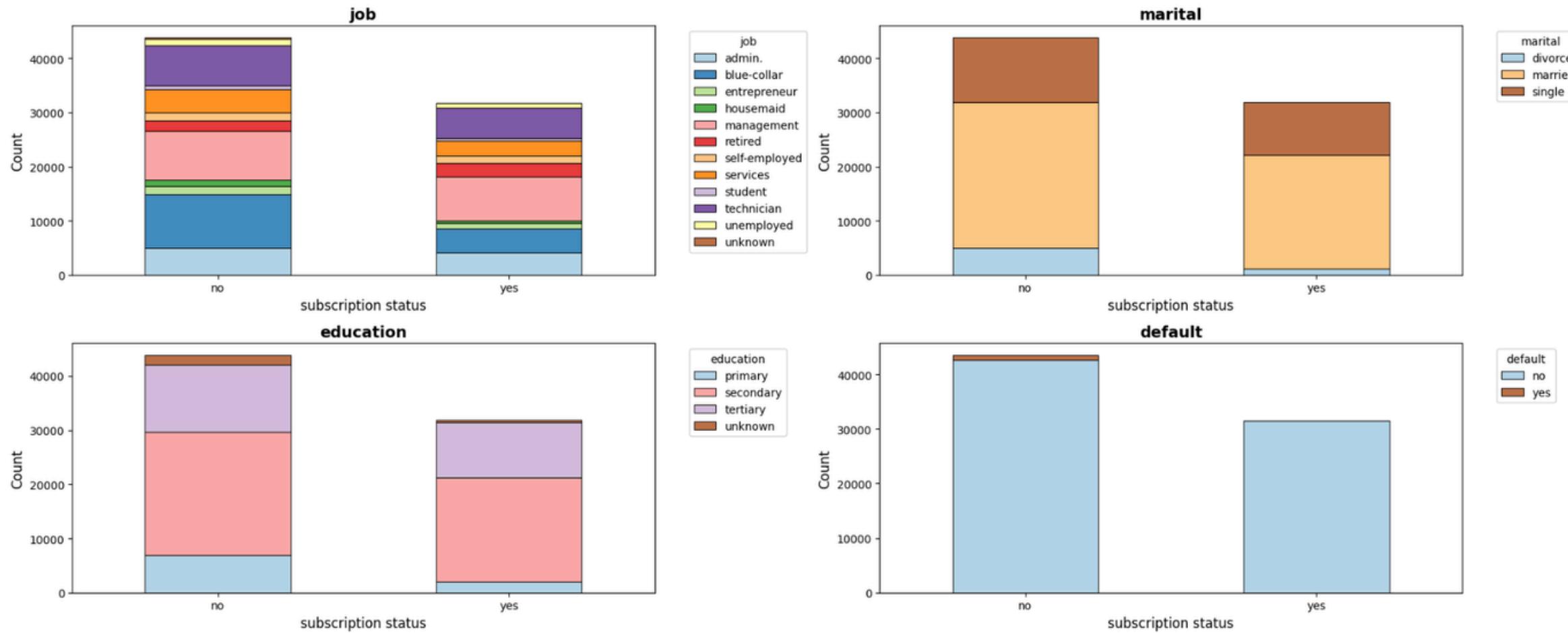


Distribusi Variabel Numerik Berdasarkan Target

- **Histogram Day by Label** : Nasabah yang berlangganan cenderung dihubungi di pertangahan bulan
- **Histogram Pdays by Label** : Nasabah yang dihubungi lagi dalam waktu terlalu dekat cenderung tidak berlangganan, sedangkan yang dihubungi lagi dalam selang waktu 60-200 hari cenderung berlangganan
- **Histogram Previous by Label** : Nasabah dengan riwayat kontak sebelumnya sedikit, cenderung tidak berlangganan



Distribusi Variabel Kategorik Berdasarkan Target



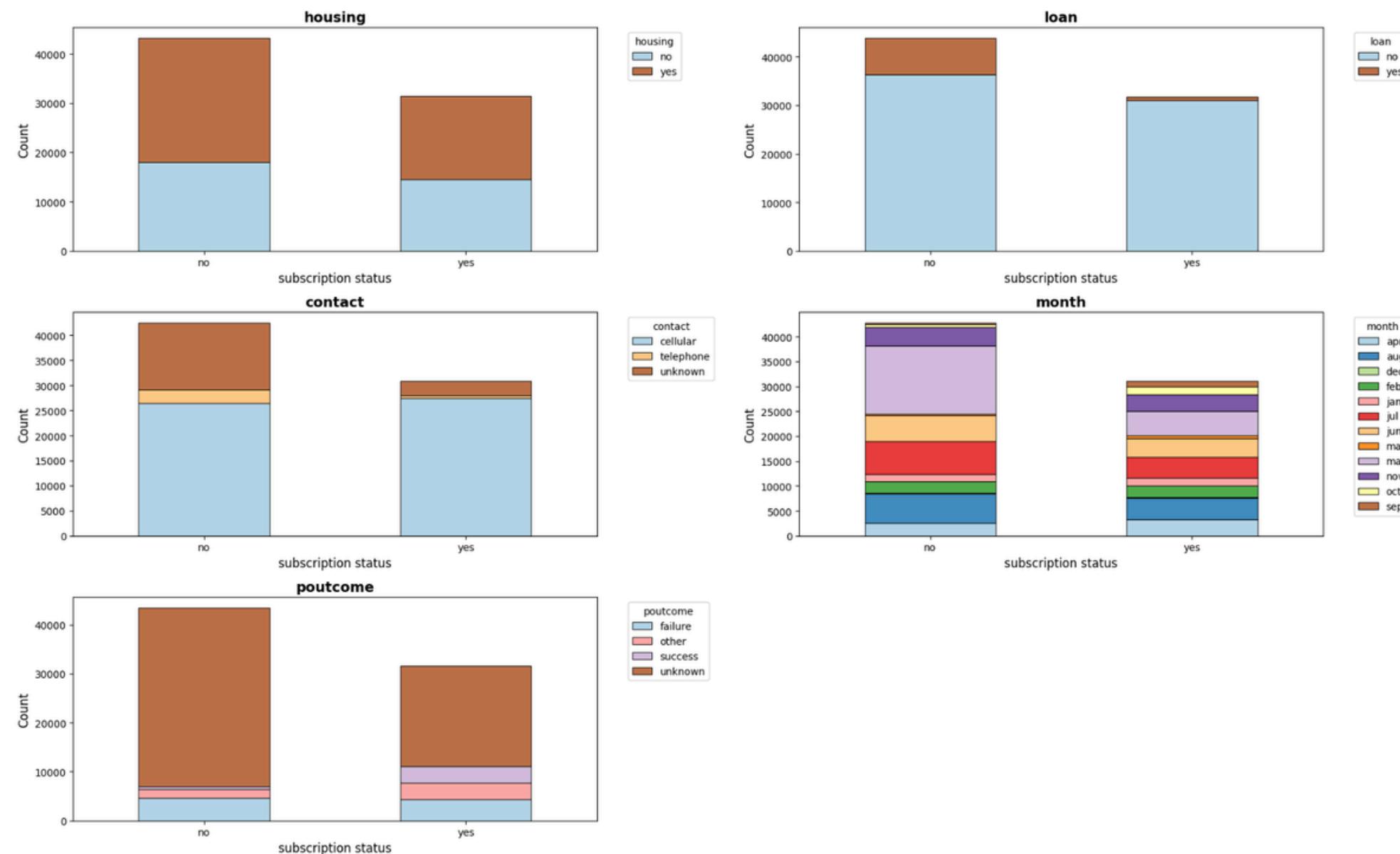
- Mayoritas nasabah yang berlangganan deposito berjangka bekerja di sektor manajemen ('management'), sedangkan mayoritas nasabah yang tidak berlangganan bekerja sebagai pekerja kasar ('blue collar')
- Proporsi nasabah dengan status pernikahan 'divorced' lebih tinggi pada kelompok yang tidak berlangganan deposito berjangka dibandingkan dengan kelompok yang berlangganan.
- Distribusi tingkat pendidikan ('education') antara kedua kelompok nasabah tidak menunjukkan perbedaan yang signifikan
- Nasabah yang berlangganan deposito berjangka cenderung tidak memiliki riwayat kredit macet ('default').



Distribusi Variabel Kategorik Berdasarkan Target



- Nasabah yang tidak berlangganan deposito berjangka cenderung memiliki kredit perumahan (KPR)
- Nasabah yang berlangganan deposito berjangka cenderung tidak memiliki pinjaman (loan)
- Nasabah yang berlangganan deposito berjangka lebih sering dihubungi melalui saluran seluler ('cellular') dibandingkan dengan mereka yang tidak berlangganan.
- Sebagian besar panggilan terakhir dilakukan pada bulan Mei, dengan distribusi bulan yang serupa di kedua kelompok.
- Nasabah yang berlangganan deposito berjangka memiliki proporsi keberhasilan yang lebih tinggi dalam kampanye sebelumnya.

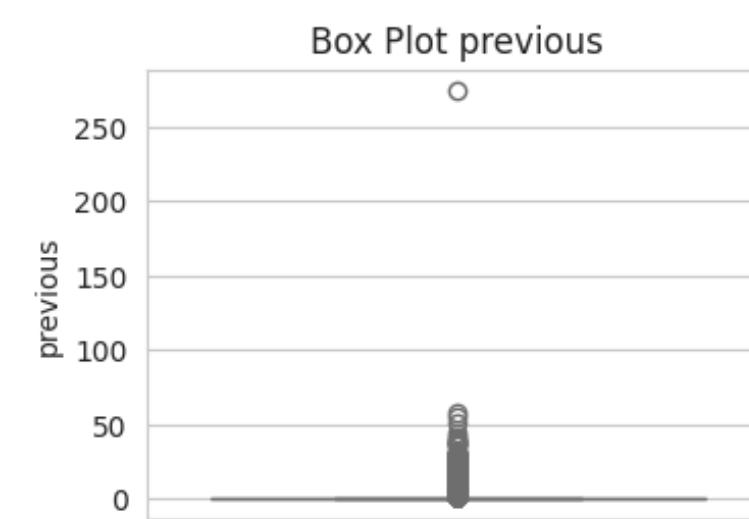
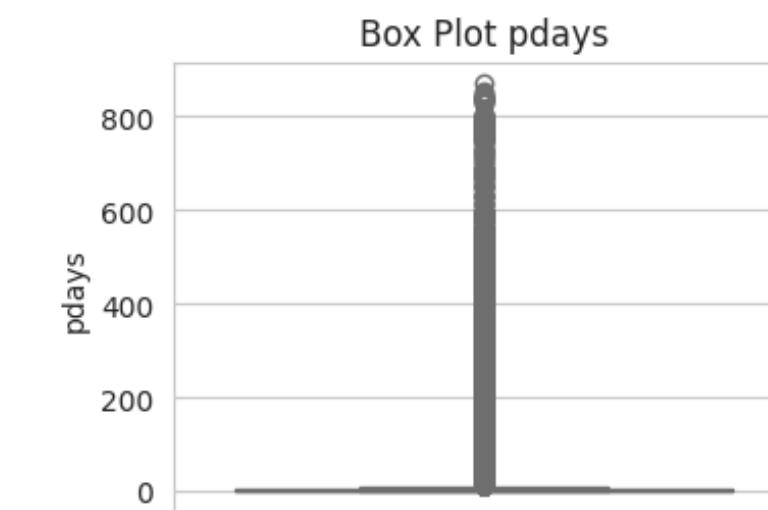
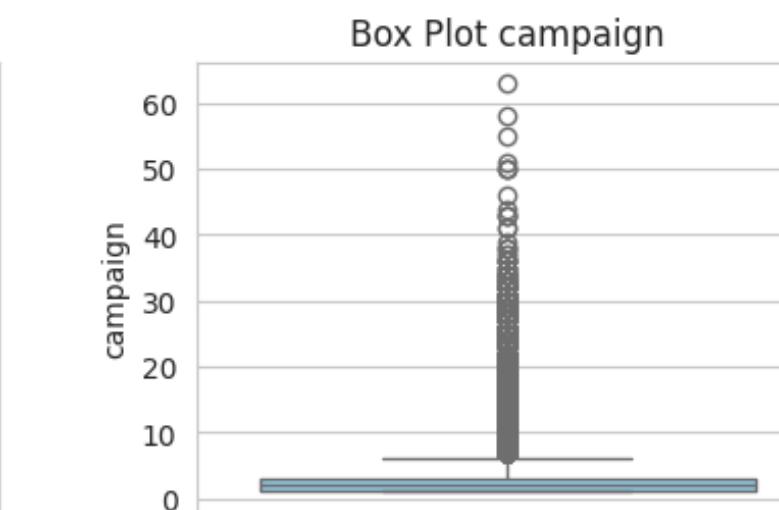
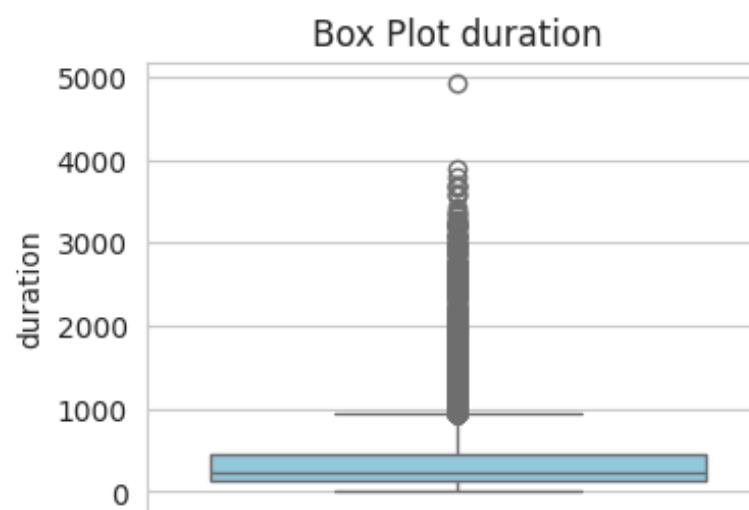
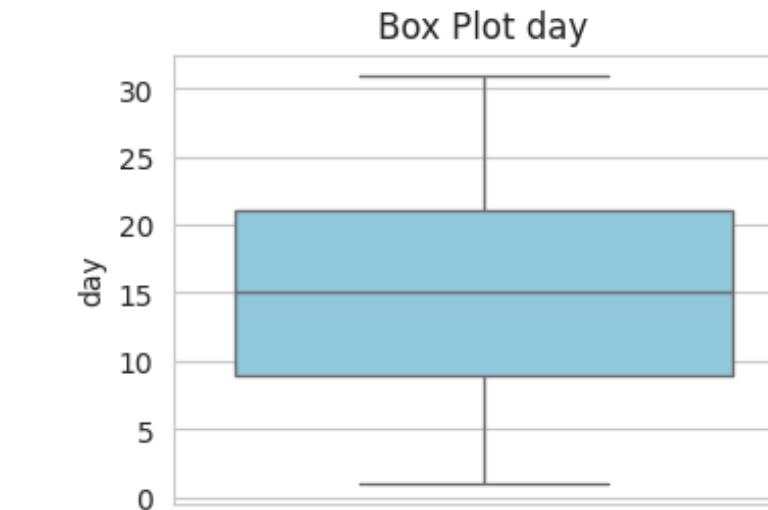
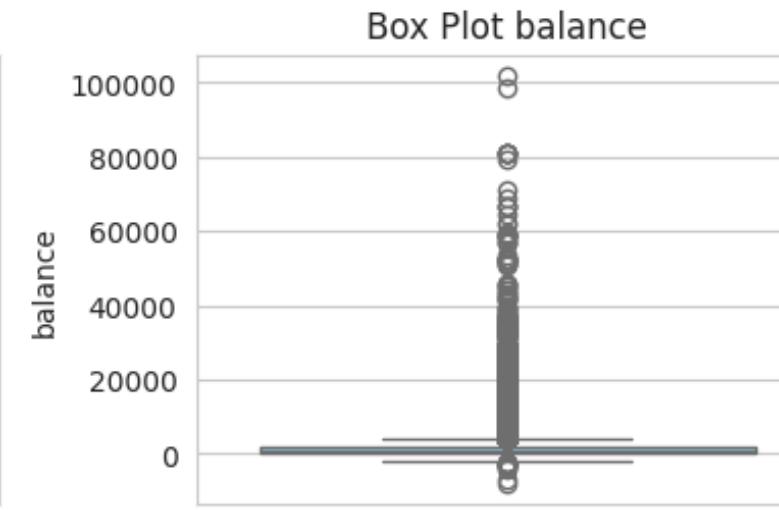
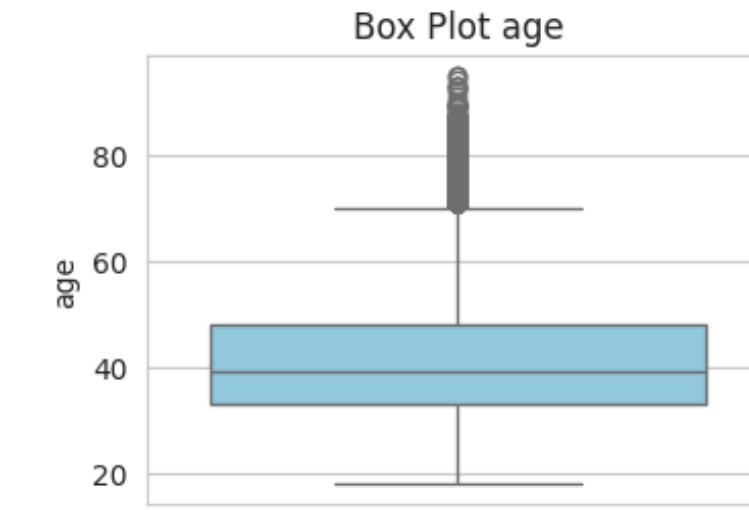


Missing Value

features	null count	null ratio (%)
age	678	0,894
job	74	0,0976
marital	70	0,0924
education	194	0,2560
default	686	0,9051
balance	974	1,2851
housing	1206	1,5912
loan	68	0,0897
contact	2520	3,3249
day	1340	1,7680
month	1991	2,6270
duration	763	1,0067
campaign	656	0,8655
pdays	143	0,1887
previous	1238	1,6334
poutcome	771	1,0173
Subscription Status	0	0

- Semua fitur kecuali variabel target (**Subscription Status**) memiliki nilai *missing value*.
- Dari total 75.791 data (klien), setiap fitur memiliki missing value dalam rentang >0-3%.

Outlier



Terdapat **6 variabel** yang memiliki *outliers*

1. Age
2. Duration
3. Balance
4. Campaign
5. Pdays
6. Previous

4

Data Cleaning

Missing Data Handling

1. Domain Knowledge-Based Imputation, untuk variabel:

- education
- default
- previous

2. Imputasi Statistical-Analysis Based, untuk variabel:

- | | |
|-----------|------------|
| ◦ age | ◦ day |
| ◦ job | ◦ month |
| ◦ marital | ◦ duration |
| ◦ balance | ◦ campaign |
| ◦ housing | ◦ pdays |
| ◦ loan | ◦ poutcome |
| ◦ contact | |

- Iterative Imputer
- KNN Imputer



yang akan dijelaskan lebih lanjut di tahap Pipeline pada **Data Pre-processing**

education

Distribusi Job per Education (%)

job	education			
	primary	secondary	tertiary	unknown
admin.	4,179006	77,690030	15,992103	2,138861
blue-collar	30,717157	58,007193	7,679289	3,596361
entrepreneur	11,124260	44,457594	40,828402	3,589744
housemaid	42,060811	39,583333	15,709459	2,646396
management	3,232240	27,828944	67,308031	1,630785
retired	24,201333	49,735693	22,776373	3,286601
self-employed	6,871795	45,299145	46,358974	1,470085
services	7,252181	74,996424	15,334001	2,417394
student	5,996622	55,827703	22,888514	15,287162
technician	2,751729	71,129900	23,897002	2,221368
unemployed	17,720307	56,321839	24,425287	1,532567
unknown	19,122257	24,137931	12,852665	43,887147

job	education
admin	NaN
services	NaN
technician	NaN

job	education
admin	secondary
services	secondary
technician	secondary

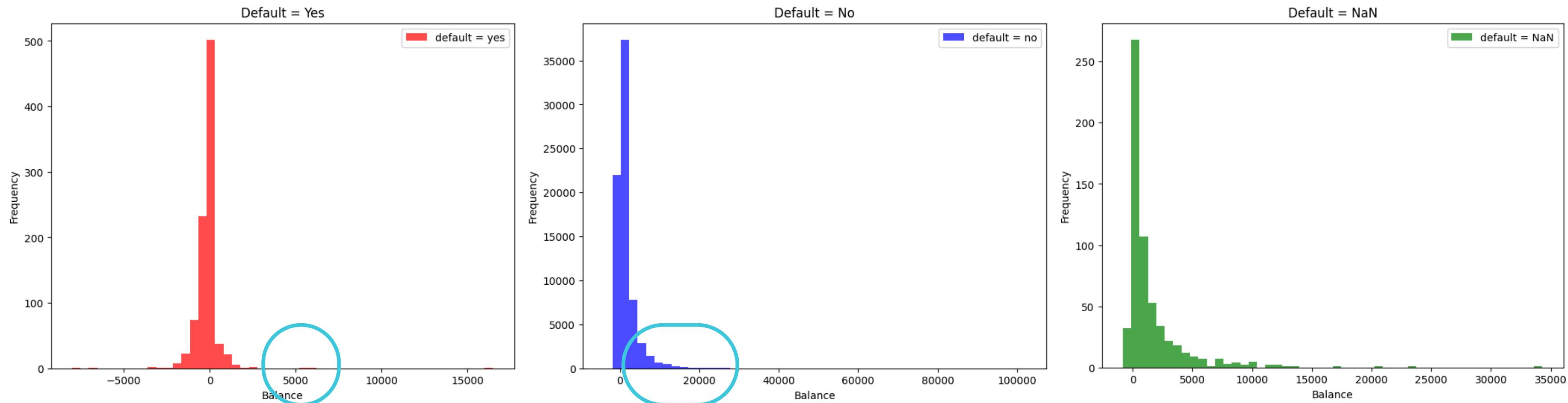
Notes:

Untuk nilai education yang lain, akan ditangani dalam pipeline di tahap Data Pre-processing.

default

variabel status kredit macet nasabah

Distribusi Balance Berdasarkan Default



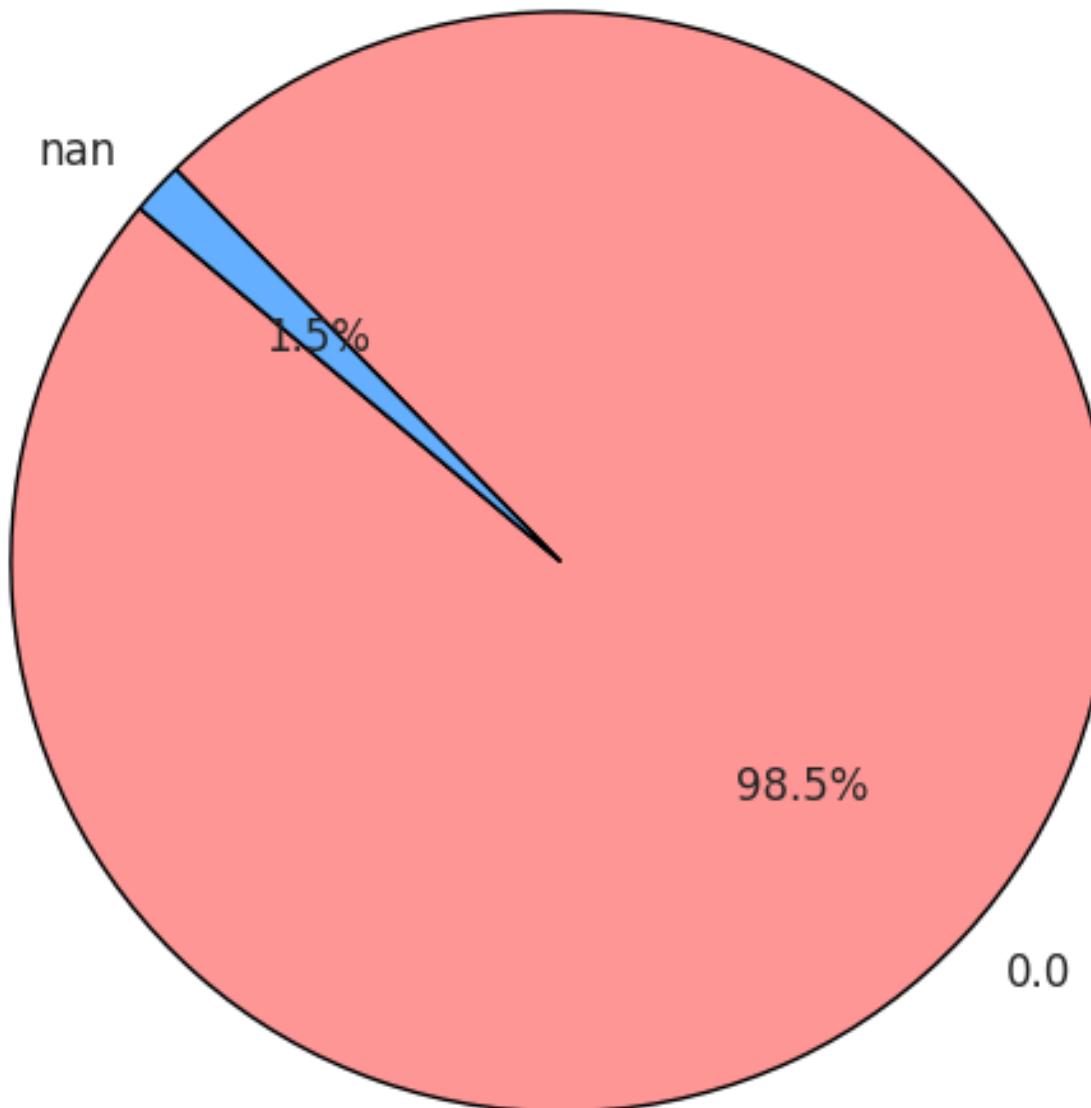
- hanya terdapat sedikit kasus untuk kategori default = "yes" dengan balance > 5000
- mayoritas balance > 5000 hanya terdapat pada kategori default = "no"

- Sehingga, data *missing* dengan balance > 5000 akan langsung dikategorikan sebagai "no".
- Sementara itu, untuk data *missing* dengan balance ≤ 5000 , akan ditangani dalam pipeline di tahap Data Pre-processing.

previous

variabel jumlah kontak pada kampanye sebelumnya

Distribusi Previous Ketika Pdays = -1



Saat Pdays = -1,

Previous hanya bernilai 0, sehingga missing value pada previous yang memiliki pdays bernilai -1, akan diisi 0

pdays	previous
-1	NaN
-1	NaN
-1	NaN

pdays	previous
-1	0
-1	0
-1	0

Notes:

Untuk nilai previous yang lain, akan ditangani dalam pipeline di tahap Data Pre-processing.

4

Data Cleaning



Outlier Data Handling

1. Age
2. Duration
3. Balance
4. Campaign
5. Pdays
6. Previous

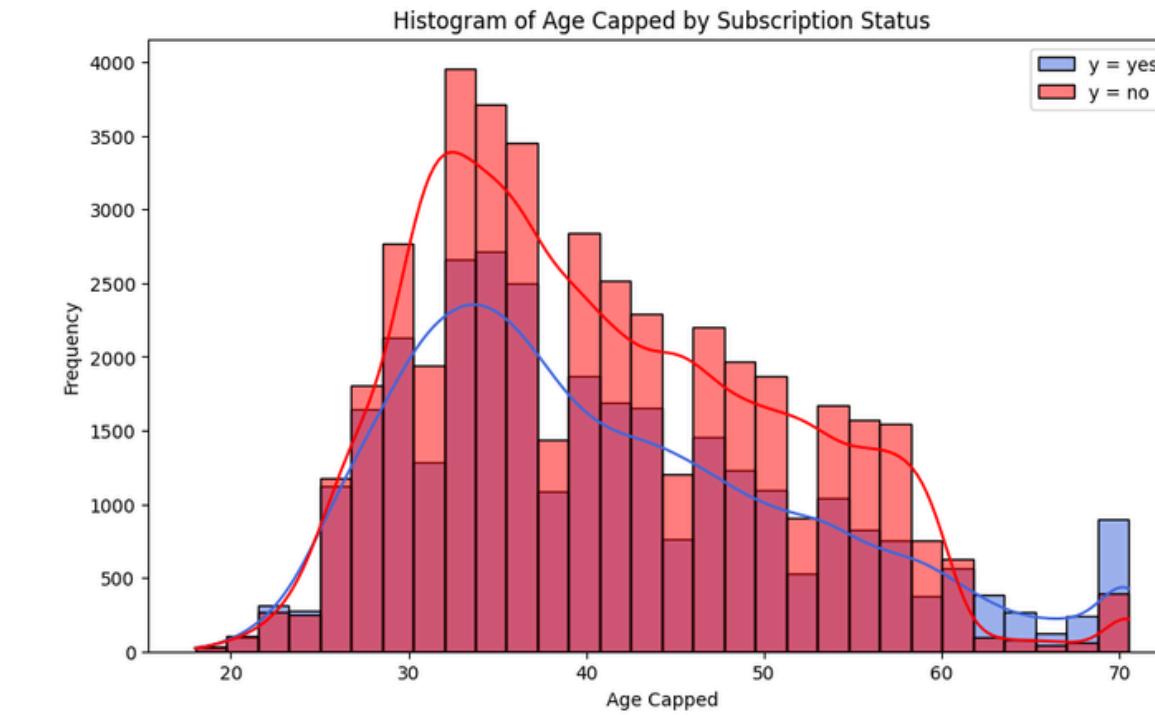
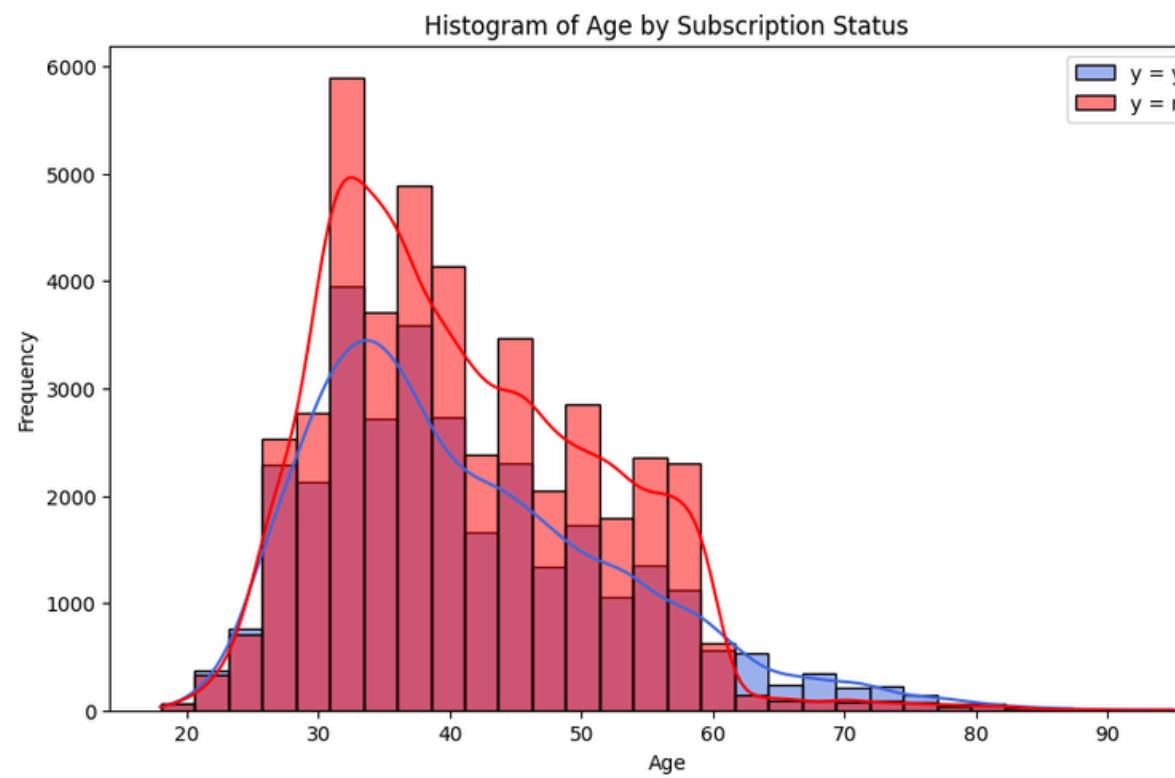
akan di atasi dengan

- Metode Capping
- Metode Winsorizing

Notes:

Penanganan outlier dilakukan dengan memperhatikan efek penanganan terhadap klasifikasi variabel target

Variabel Age



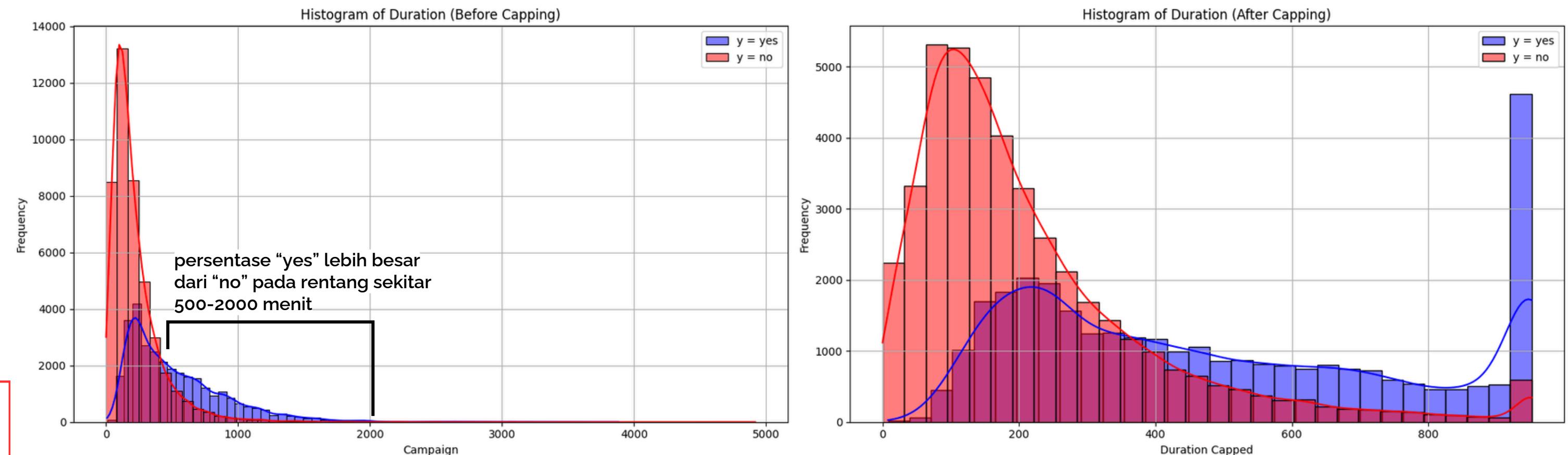
Range Age	no	yes
12	72-77	29.291045
13	78-83	39.676113
14	84-89	33.333333
15	90-95	16.666667
		83.333333

Outlier pada variabel age ditangani dengan capping,
Penanganan capping **tidak menghilangkan informasi age > 70**, di data capping maupun
asli nasabah dengan age > 70 cenderung berlangganan

Variabel Duration

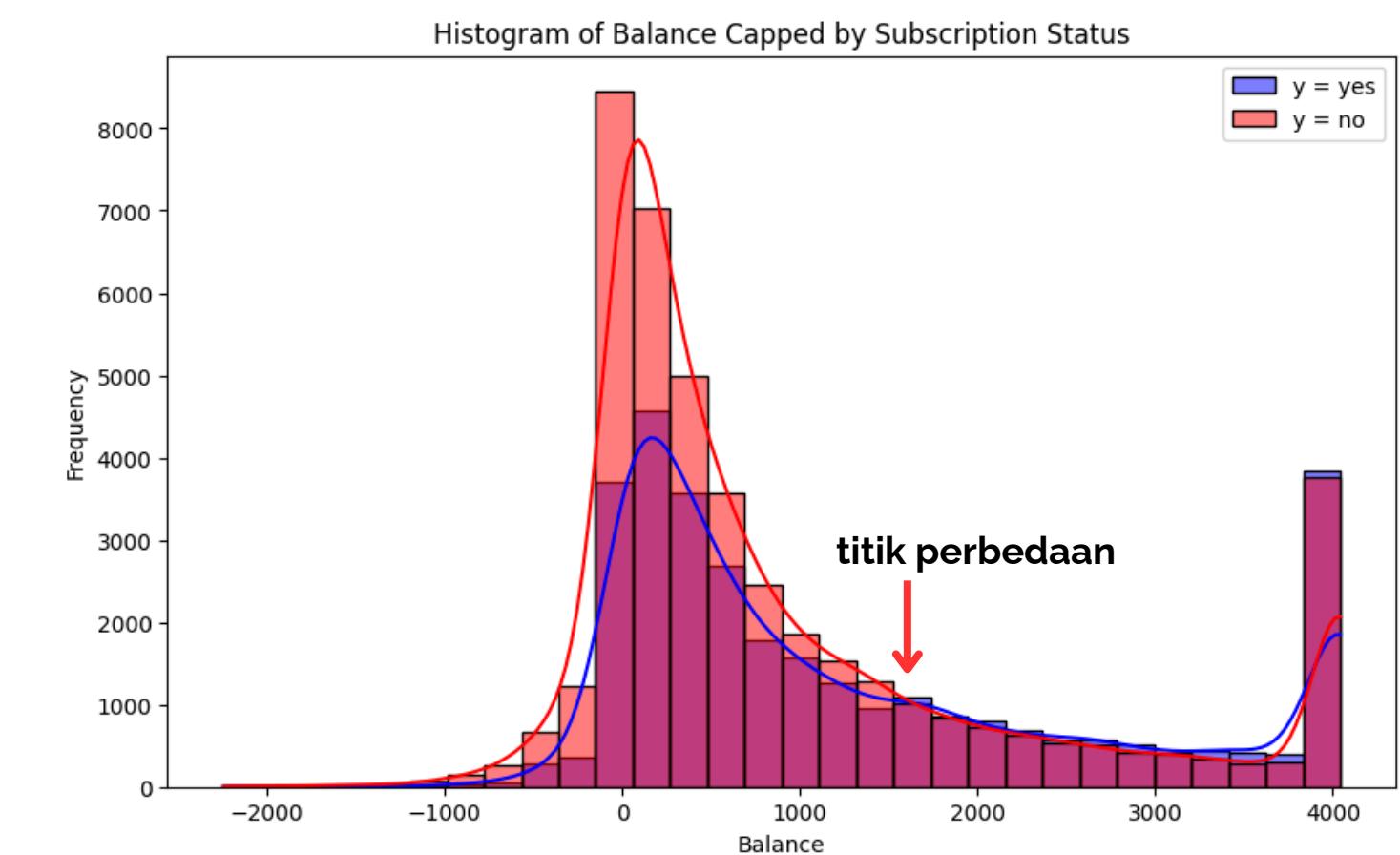
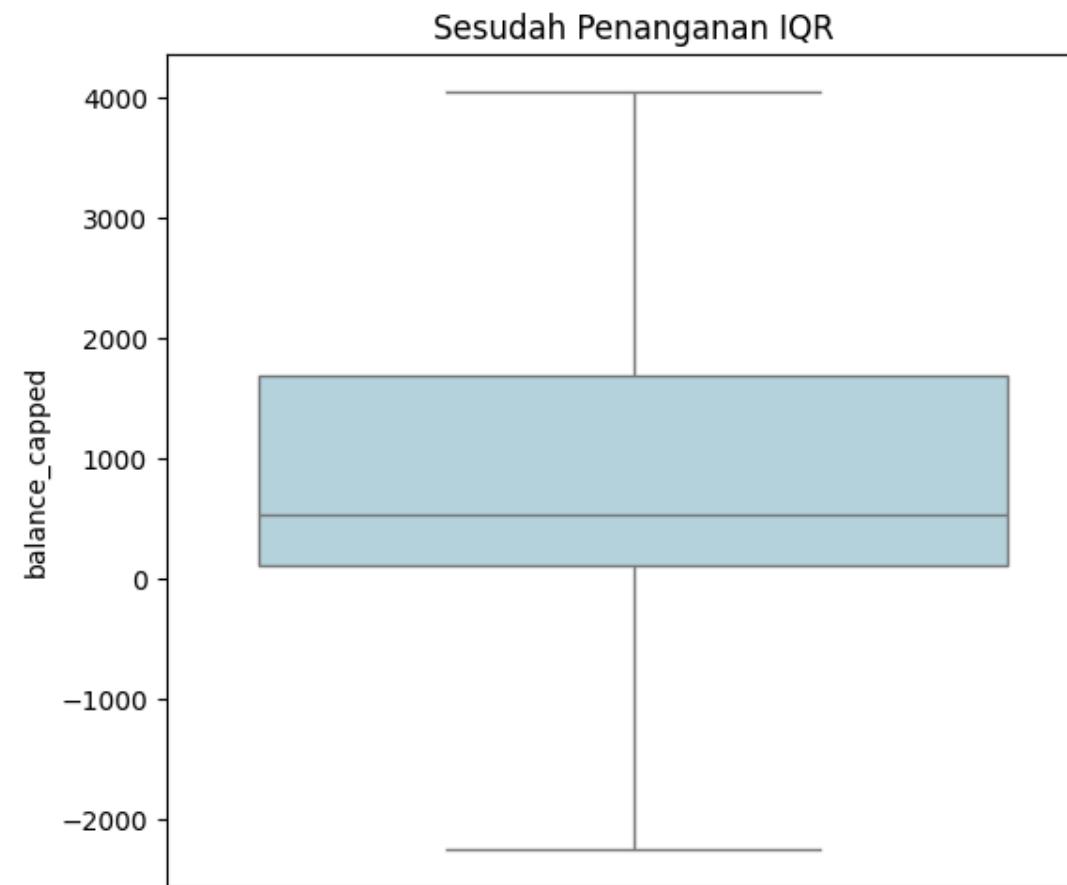
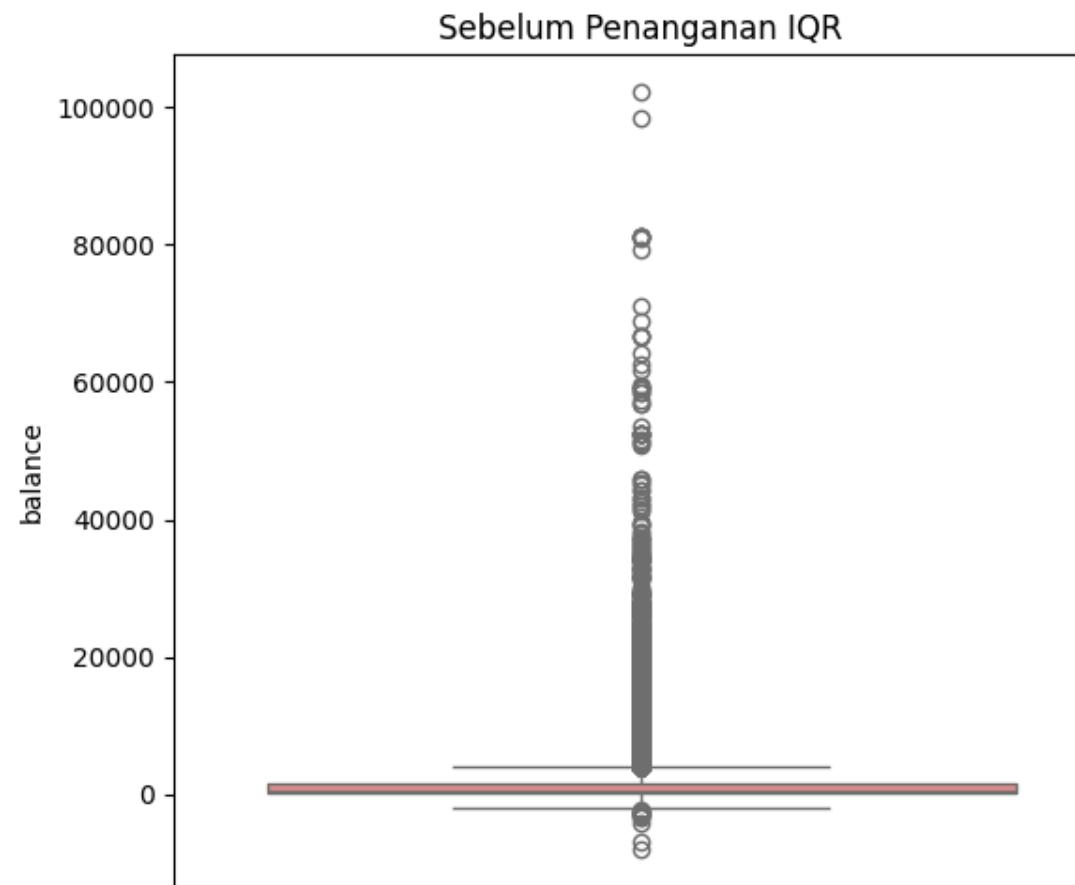
variabel durasi panggilan

y	Condition	no	yes
0	0-250	78.378585	21.621415
1	251-501	49.449570	50.550430
2	502-752	26.488804	73.511196
3	753-1003	15.306593	84.693407
4	1004-1254	11.873589	88.126411
5	1255-1505	10.404040	89.595960
6	1506-1756	7.972665	92.027335
7	1757-2007	12.107623	87.892377
8	2008-2258	28.070175	71.929825
9	2259-2509	19.354839	80.645161
10	2510-2760	2.272727	97.727273
11	2761-3011	3.703704	96.296296
12	3012-3262	8.333333	91.666667
13	3263-3513	57.142857	42.857143
14	3514-3764	NaN	100.000000
15	3765-4015	50.000000	50.000000
16	4016-4266	NaN	NaN
17	4267-4517	NaN	NaN
18	4518-4768	NaN	NaN
19	4769-4918	100.000000	NaN



- Outlier pada variabel duration **tidak akan ditangani** karena akan menghilangkan informasi penting.
- Oleh karena itu, outlier pada variabel duration **akan tetap digunakan** karena ada informasi yang berarti.
- Berdasarkan histogram, terdapat rentang duration tertentu, sekitar 500-2000 menit, yang menunjukkan adanya kemungkinan lebih besar nasabah untuk berlangganan dibandingkan rentang durasi lainnya.

Variabel Balance

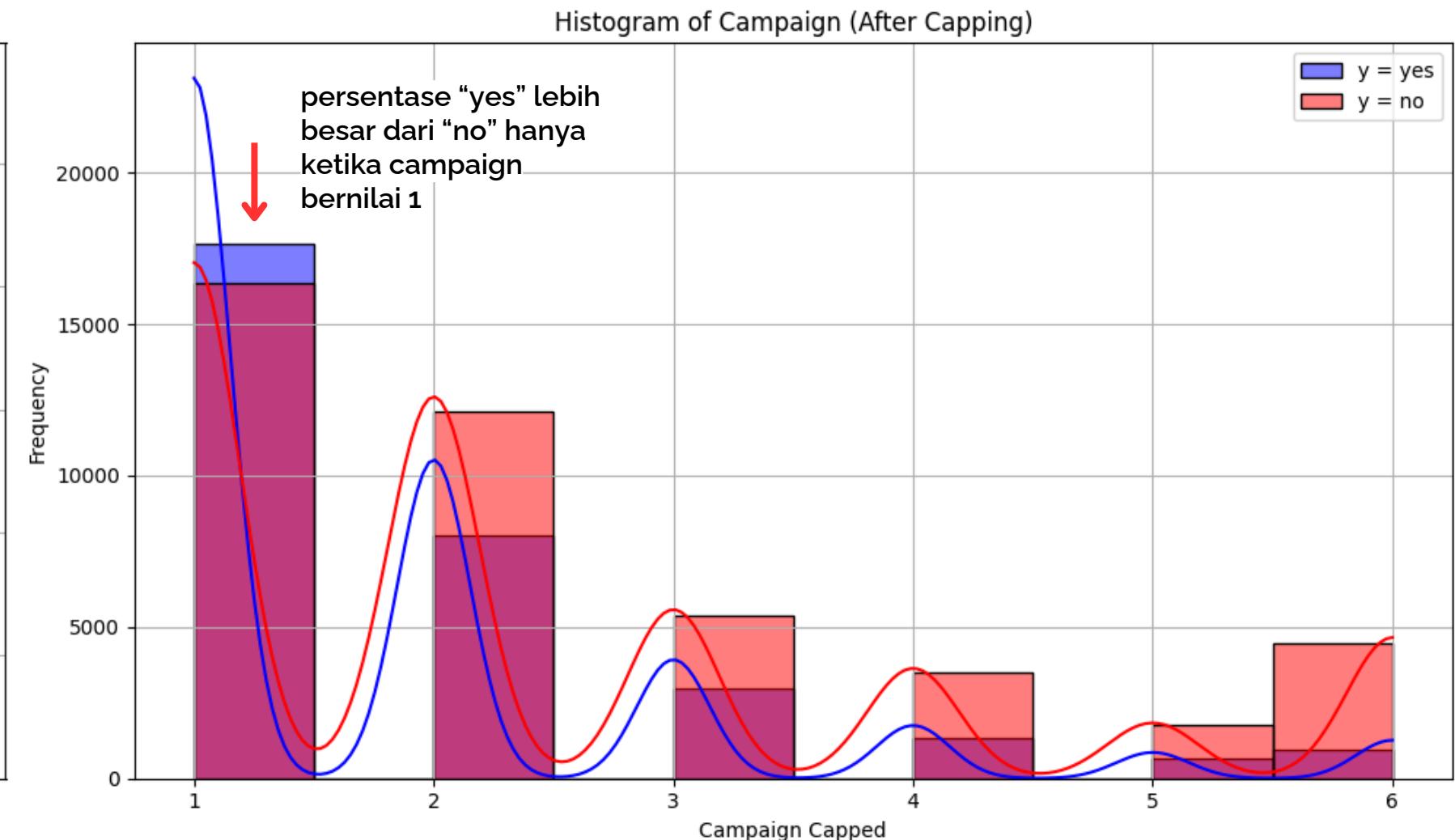
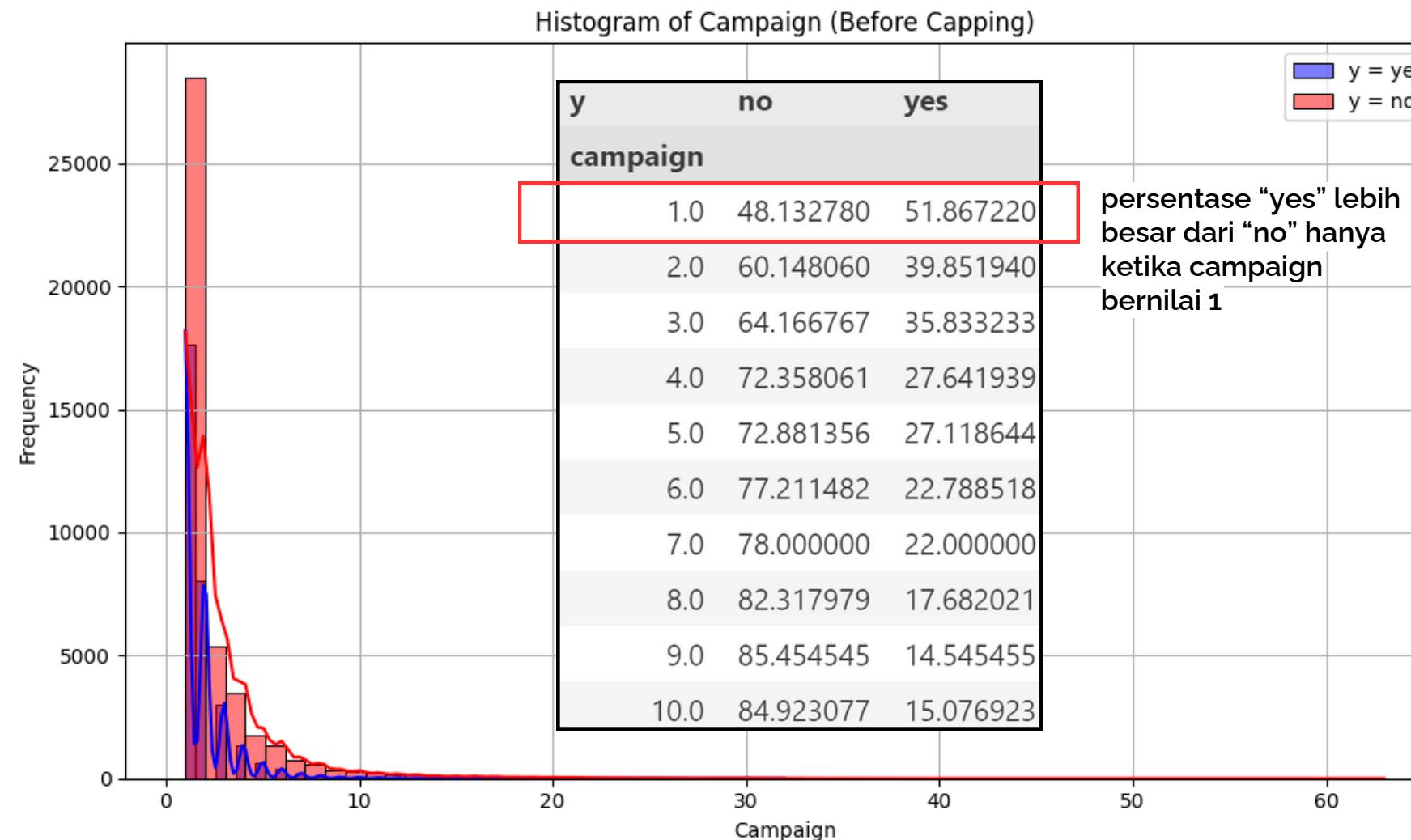


- Outlier pada variabel balance ditangani dengan **metode capping**, yaitu menetapkan batas atas dan batas bawah menggunakan *Interquartile Range* (IQR).
- Dari boxplot di atas, dapat dilihat bahwa metode capping berhasil menangani outlier dengan baik.

- Pola distribusi antara kedua kelompok lebih jelas terlihat setelah penanganan outlier.
- Terdapat titik perbedaan yang ditandai dengan panah merah, yang menunjukkan kemungkinan hubungan antara saldo nasabah dengan keputusan mereka untuk berlangganan (subscription).

Variabel Campaign

variabel durasi jumlah kontak pada campaign sekarang



- Outlier pada variabel campaign ditangani dengan **metode capping**, yaitu menetapkan batas atas dan batas bawah menggunakan *Interquartile Range* (IQR).
- Dari histogram dan kde plot sebelum dilakukan capping, campaign bernilai dari 0 hingga lebih dari 60, tetapi mostly bernilai 1.

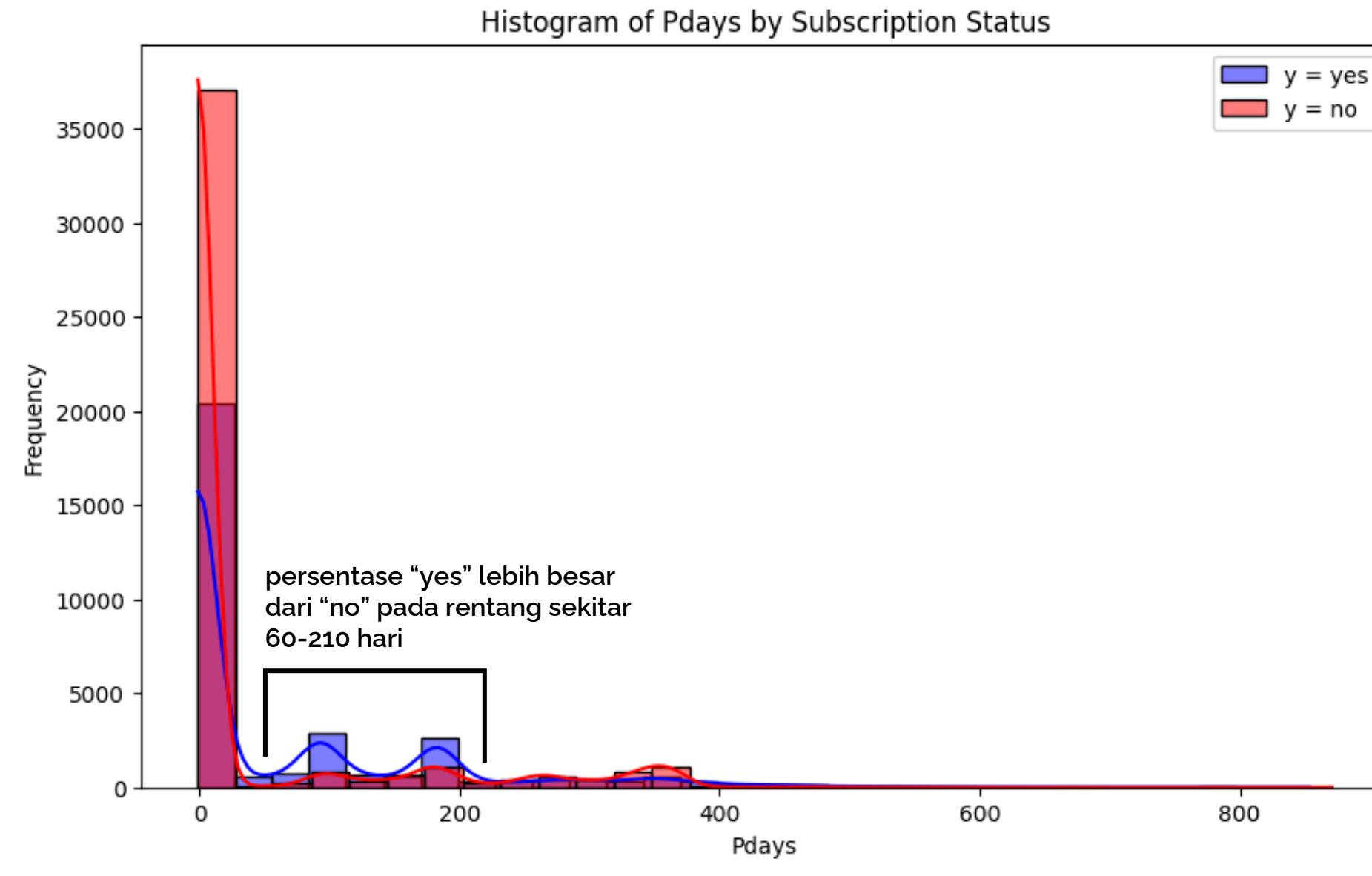
- Pola distribusi antara kedua kelompok lebih jelas terlihat setelah penanganan outlier.
- Pola yang sama tetap terlihat, di mana untuk campaign lebih dari 1 kali, tingkat keberhasilan semakin menurun.



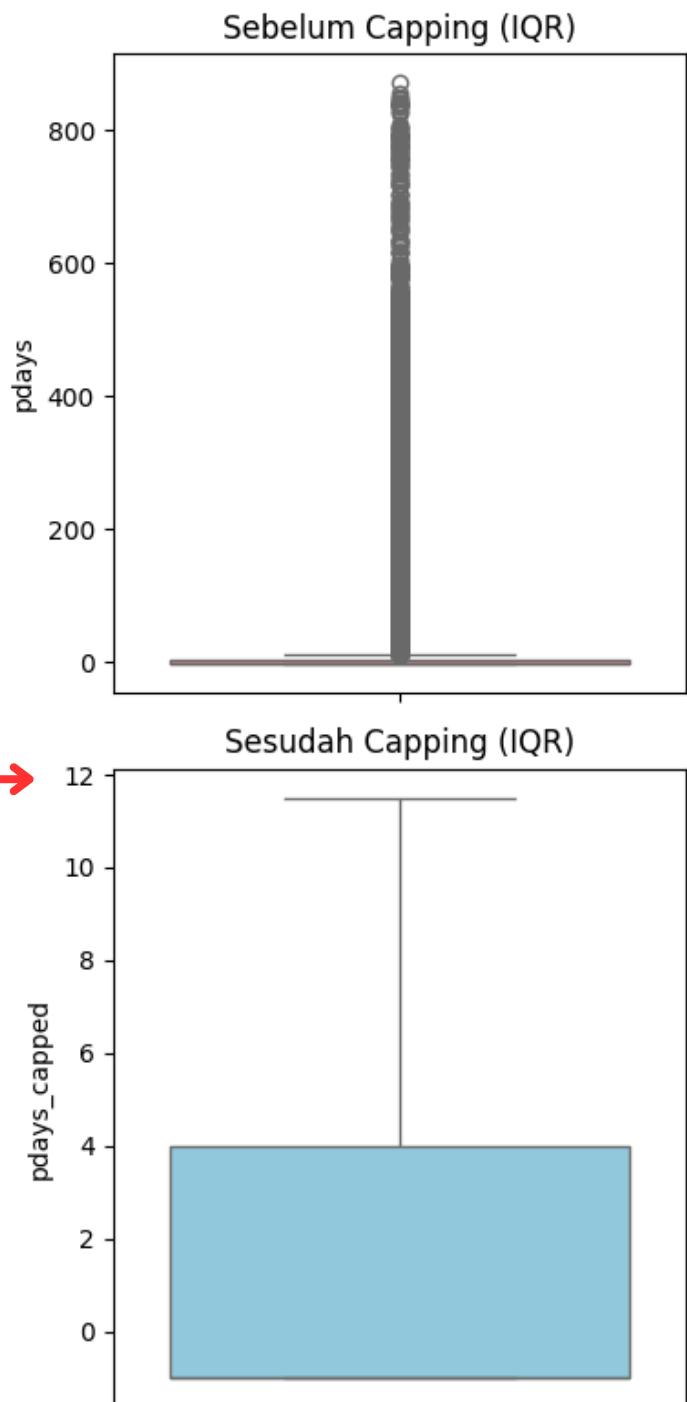
Variabel Pdays

variabel selisih hari sejak panggilan sebelumnya

- Outlier pada variabel pdays **tidak akan ditangani** karena akan menghilangkan informasi penting.
- Berdasarkan histogram, terdapat rentang pdays tertentu, sekitar 60-210 hari, yang menunjukkan adanya kemungkinan lebih besar nasabah untuk berlangganan dibandingkan rentang hari lainnya.
- Padahal jika di lakukan salah satu penanganan, misalkan metode capping, pdays akan terbatas maksimum pada 12 hari.

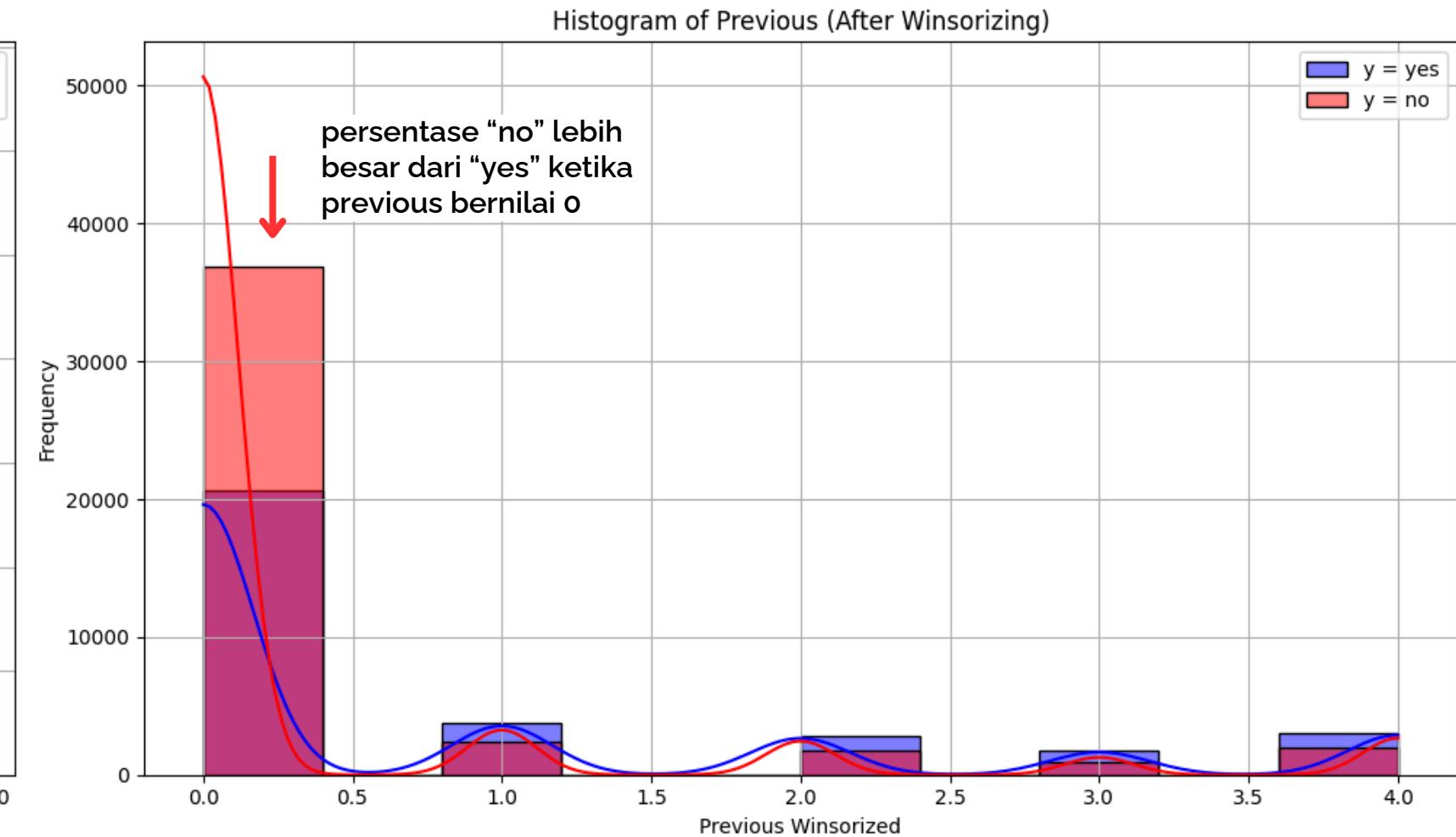
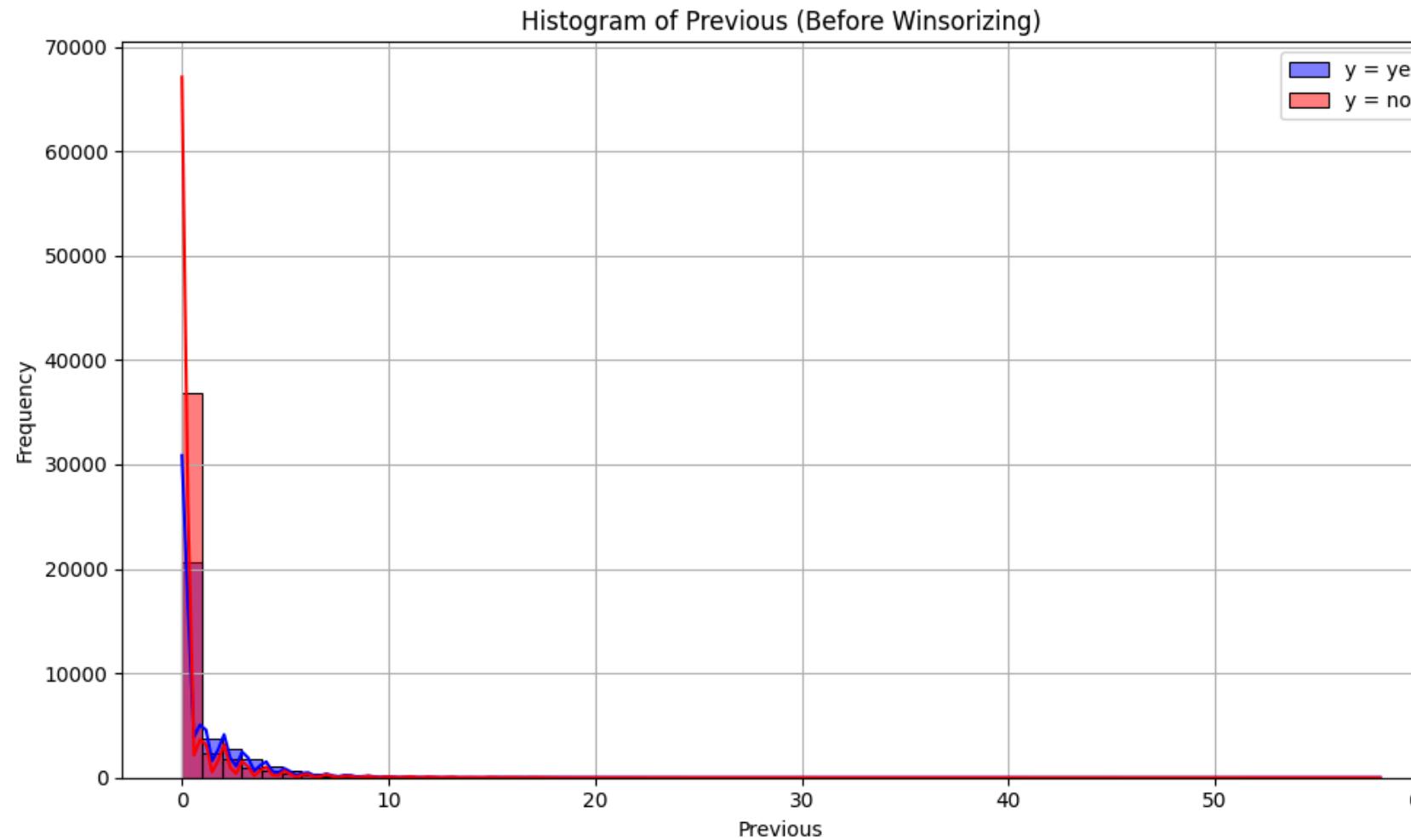


- Oleh karena itu, outlier pada variabel pdays **akan tetap digunakan** karena ada informasi yang berarti.

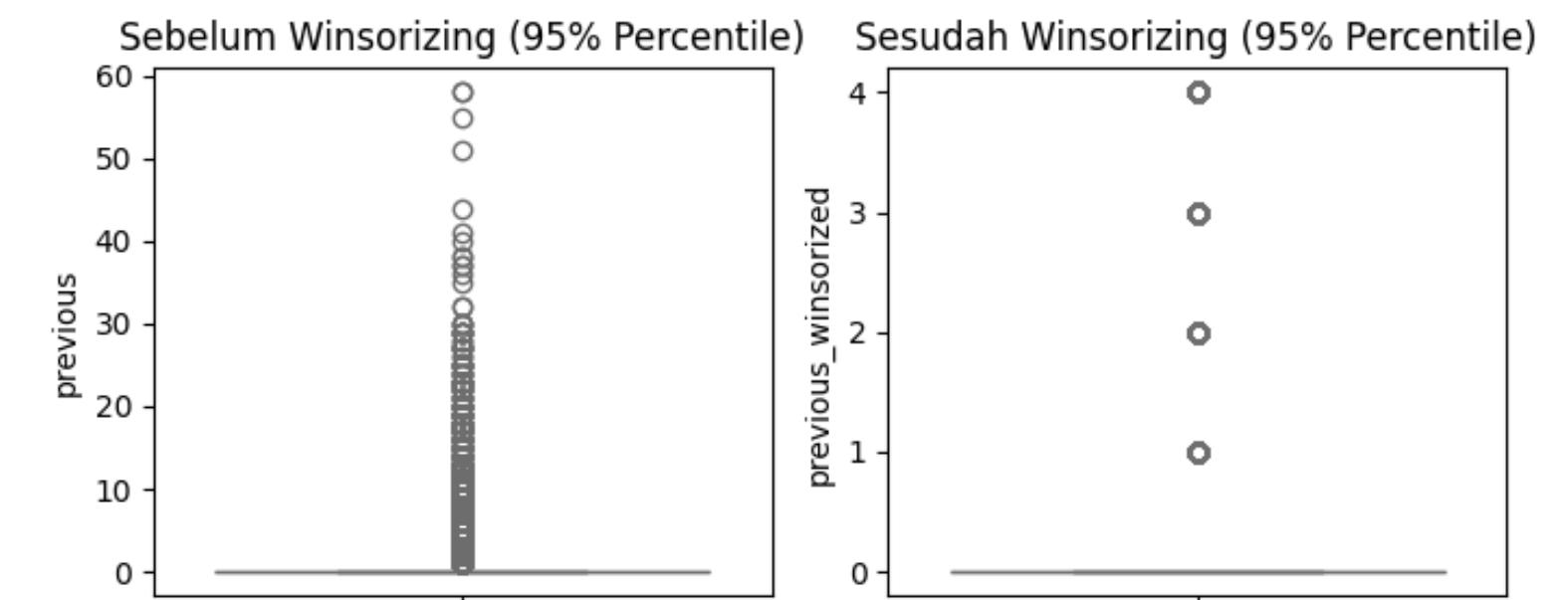


Variabel Previous

variabel jumlah kontak pada campaign sebelumnya



- Outlier pada variabel campaign ditangani dengan **metode winsorizing**, yaitu menetapkan batas atas dan batas bawah menggunakan persentil tertentu.
- Metode ini tidak sepenuhnya mengatasi outlier, tetapi jika dilihat dari histogram, pola distribusi antara kedua kelompok deposito lebih jelas terlihat setelah winsorizing..



5 Feature Engineering

Pembuatan Fitur Baru



week (object)

keterangan minggu ke-berapa dalam bulan tersebut

week akan bernilai "1st" (day = 1-7), "2nd" (day = 8-14), "3rd" (day=15-21), "4th" (day 22-31)



housing_loan (object)

kepemilikan loan dan housing loan

housing_loan akan bernilai "yes" jika nasabah memiliki loan dan housing loan, selain kondisi tersebut fitur ini akan bernilai "no"



previous_contacted (object)

keterangan apakah pernah dihubungi sebelumnya

previous_contacted bernilai "no" jika pdays bernilai -1 dan benilai "yes" jika pdays != -1



season (object)

musim di Negara Portugal ketika dilakukan campaign

season berisi kategori empat musim berdasarkan month

- spring : mar - may
- summer : jun - aug
- autumn : sep - nov
- winter : dec - feb

5

Feature Engineering

Feature Encoding

Mengubah variabel kategorikal menjadi numerik

- **job : 12 kategori**

'technician': 0, 'management': 1, 'entrepreneur': 2, 'blue-collar': 3, 'unknown': 4, 'admin.': 5, 'retired': 6, 'housemaid': 7, 'services': 8, 'student': 9, 'unemployed': 10, 'self-employed': 11

- **marital : 3 kategori**

'divorced': 0, 'single': 1, 'married': 2

- **education : 4 kategori**

'primary': 0, 'secondary': 1, 'tertiary': 2, 'unknown': 3

- **contact : 3 kategori**

'unknown': 0, 'cellular': 1, 'telephone': 2

- **month : 12 kategori**

'jan': 0, 'feb': 1, 'mar': 2, 'apr': 3, 'may': 4, 'jun': 5, 'jul': 6, 'aug': 7, 'sep': 8, 'oct': 9, 'nov': 10, 'dec': 11

- **month : 12 kategori**

'jan': 0, 'feb': 1, 'mar': 2, 'apr': 3, 'may': 4, 'jun': 5, 'jul': 6, 'aug': 7, 'sep': 8, 'oct': 9, 'nov': 10, 'dec': 11

- **poutcome : 4 kategori**

'failure': 0, 'unknown': 1, 'other': 2, 'success': 3

- **season : 4 kategori**

'Winter': 0, 'Spring': 1, 'Summer': 2, 'Autumn': 3

- **week : 4 kategori**

'1st': 0, '2nd': 1, '3rd': 2, '4th': 3

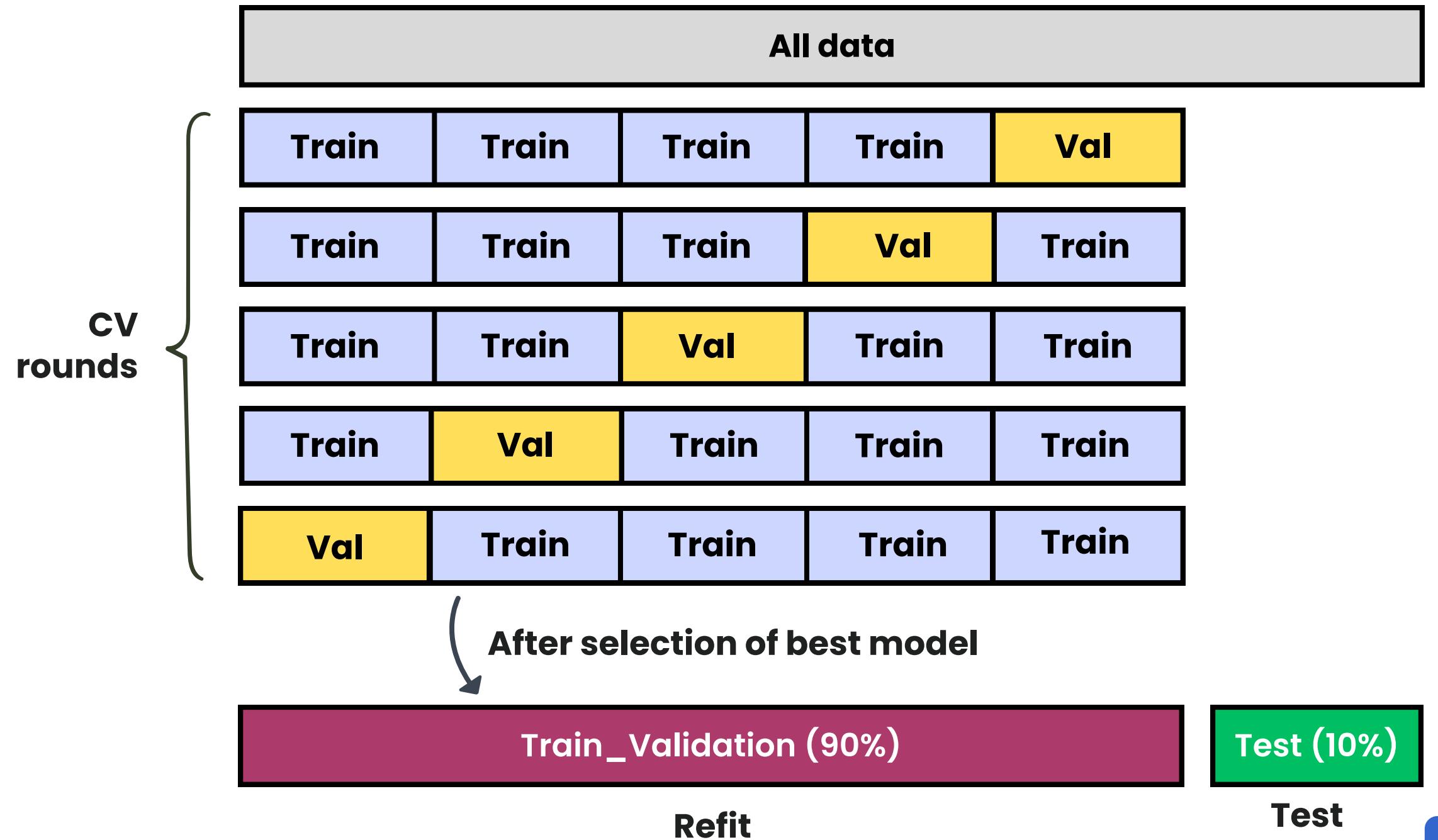
- **default, housing, loan, housing_loan, previous_contacted : 2 kategori**

'no': 0, 'yes': 1

6

Data Pre-processing

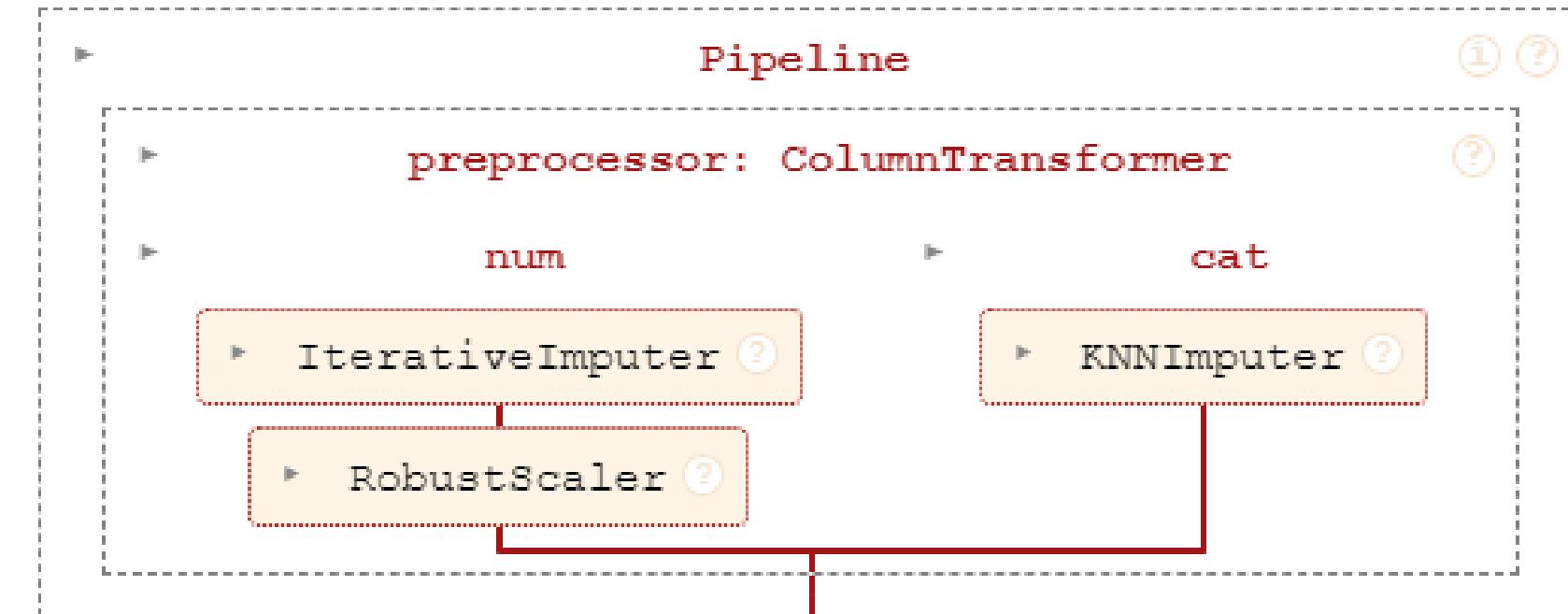
Splitting Data



6

Data Pre-processing

Pipeline



Processed X_train shape: (60632, 21)

Processed X_test shape: (15159, 21)

6 Data Pre-processing

Feature Selection

Cek **Mutual Information Score**

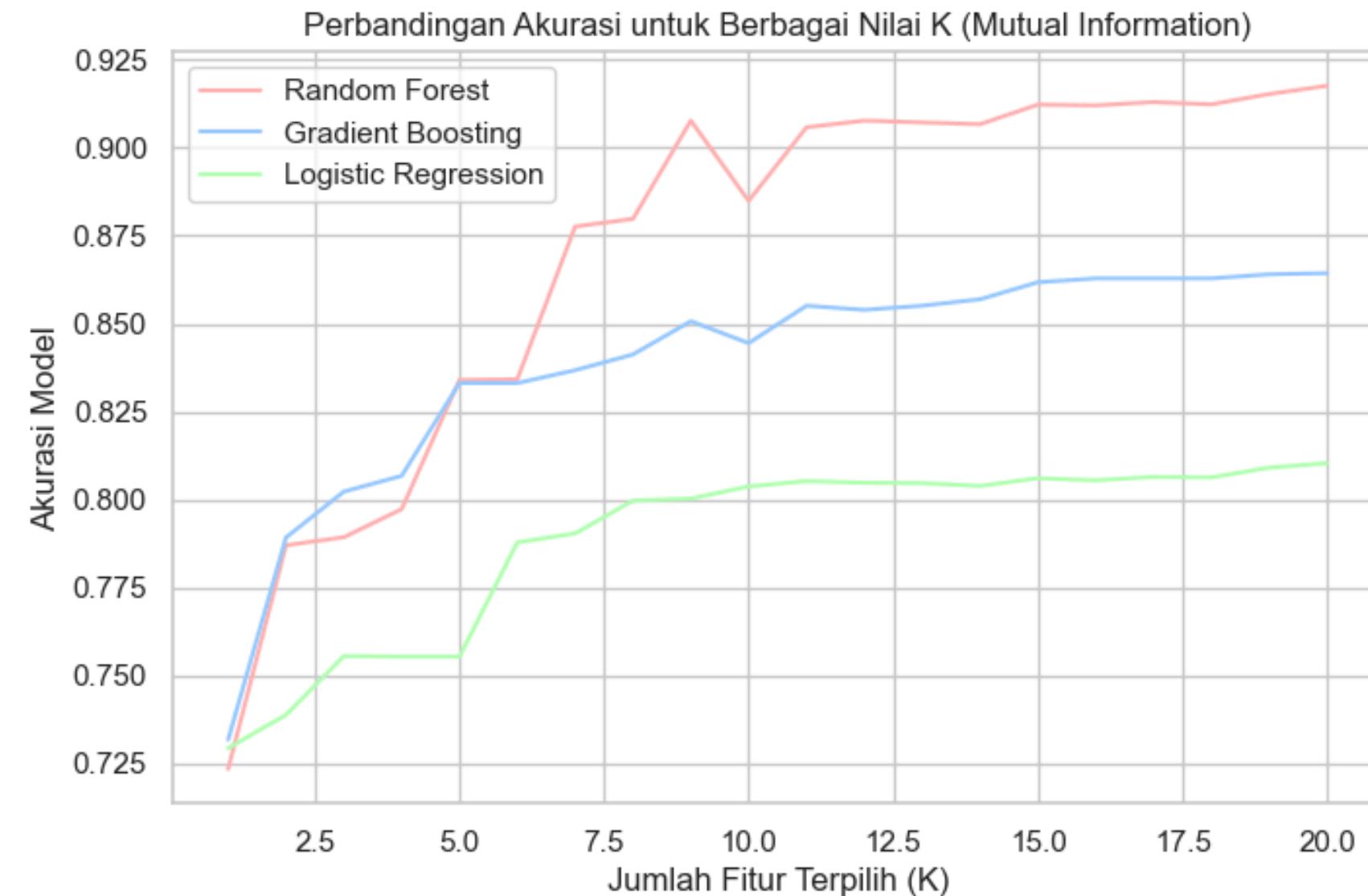
Feature	Mutual Information Score	Feature	Mutual Information Score
num__duration	0,165184	num__campaign_capped	0,021844
num__pdays	0,057832	cat__housing_loan	0,017353
cat__contact	0,052799	cat__education	0,017129
cat__poutcome	0,042234	cat__marital	0,014989
cat__previous_contacted	0,038333	cat__season	0,014924
cat__month	0,033714	cat__job	0,010065
cat__loan	0,032031	num__age_capped	0,008762
num__balance_capped	0,028167	cat__Week	0,007747
num__previous_winsorized	0,024256	cat__housing	0,005378
num__day	0,024178	cat__default	0,003747

6

Data Pre-processing

Feature Selection

Menentukan nilai K optimal



Dari plot di atas, diperoleh bahwa semakin tinggi K yang digunakan, semakin tinggi pula nilai akurasi model. Selanjutnya, pemodelan akan menggunakan nilai K tertinggi (seluruh feature).

7

Modelling

Perbandingan Performa Model (dengan 5-Cross Validation)

Model	Tanpa Balancing Data			Balancing dengan SMOTE		
	mean F1 Score	std dev F1 Score	mean running time	mean F1 Score	std dev F1 Score	mean running time
Random Forest	0.916108	0.001244	7.430152	0.914475	0.001367	9.811358
Extra Tree	0.912138	0.000917	5.666335	0.911604	0.001368	6.945771
LGBM	0.893896	0.002299	0.927199	0.893188	0.001085	0.372155
XGBoost	0.905597	0.001362	0.570212	0.904841	0.001996	0.420876
Gradient Boosting	0.857562	0.002675	0.002244	0.858594	0.000921	12.008532
Stacked Model	0.919618	0.001189	72.64	0.918413	0.001479	156.18



base models :
1. Random Forest
2. Extra Tree
3. LGBM
4. XGBoost
meta model:
Logistic Regression

Model dengan performa terbaik adalah **Random Forest** dan **Stacked Model** yang diterapkan tanpa balancing data

7

Modelling

Fit Model pada Data Train Keseluruhan dan Predict Data Test

Model	Tanpa Balancing Data	
	F1 Score	Running Time
Random Forest	0.9185	39.7
Stacking Model	0.9247	68.13

Hyperparameter Tuning

Model	Tanpa Balancing Data	
	F1 Score	Running Time
Random Forest	0.9186	633.8
Stacking Model	0.9290	768.78

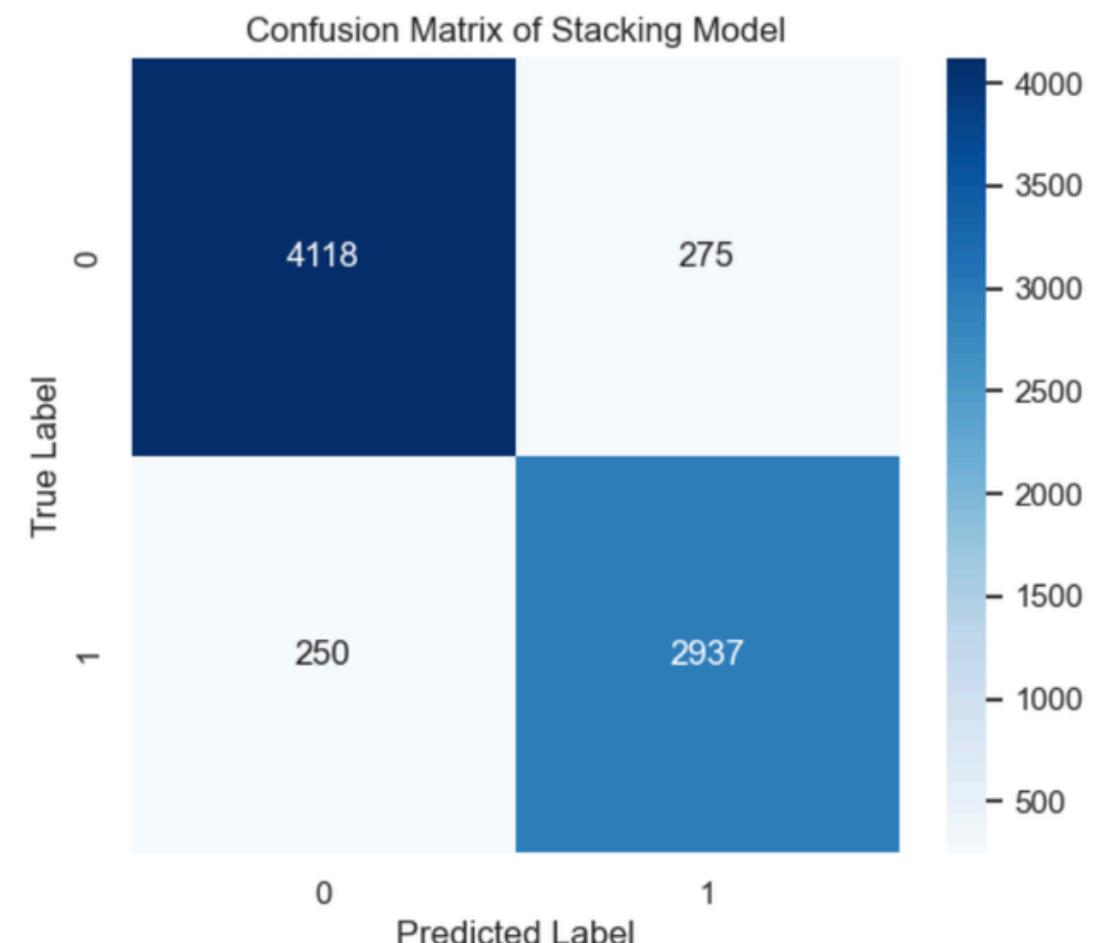
8

Evaluasi

Classification Report

	precision	recall	f1-score	support
0	0.94277	0.93740	0.94008	4393
1	0.91438	0.92156	0.91796	3187
accuracy			0.93074	7580
macro avg	0.92857	0.92948	0.92902	7580
weighted avg	0.93083	0.93074	0.93078	7580

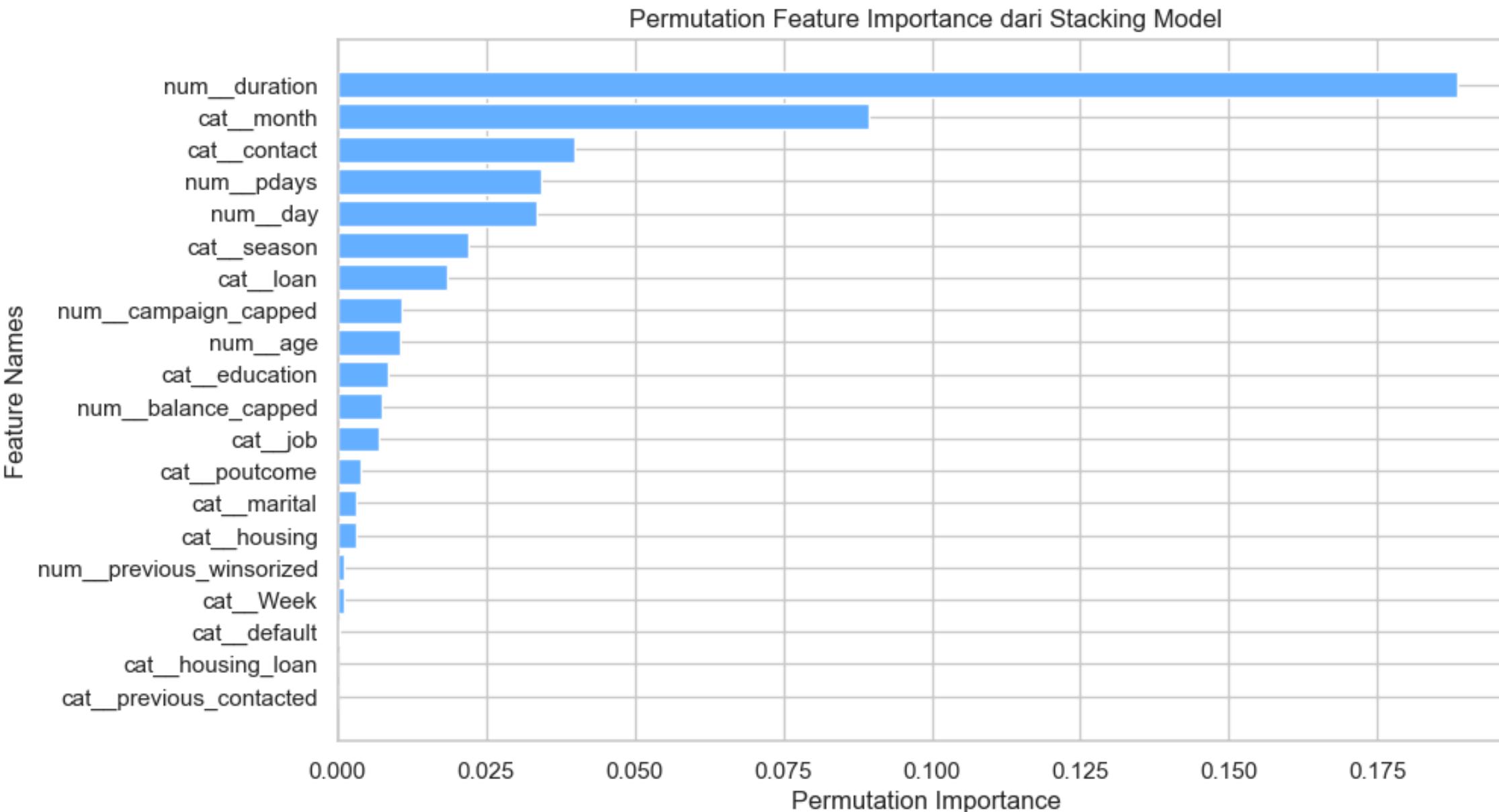
Confusion Matrix



8

Evaluasi

Feature Importance

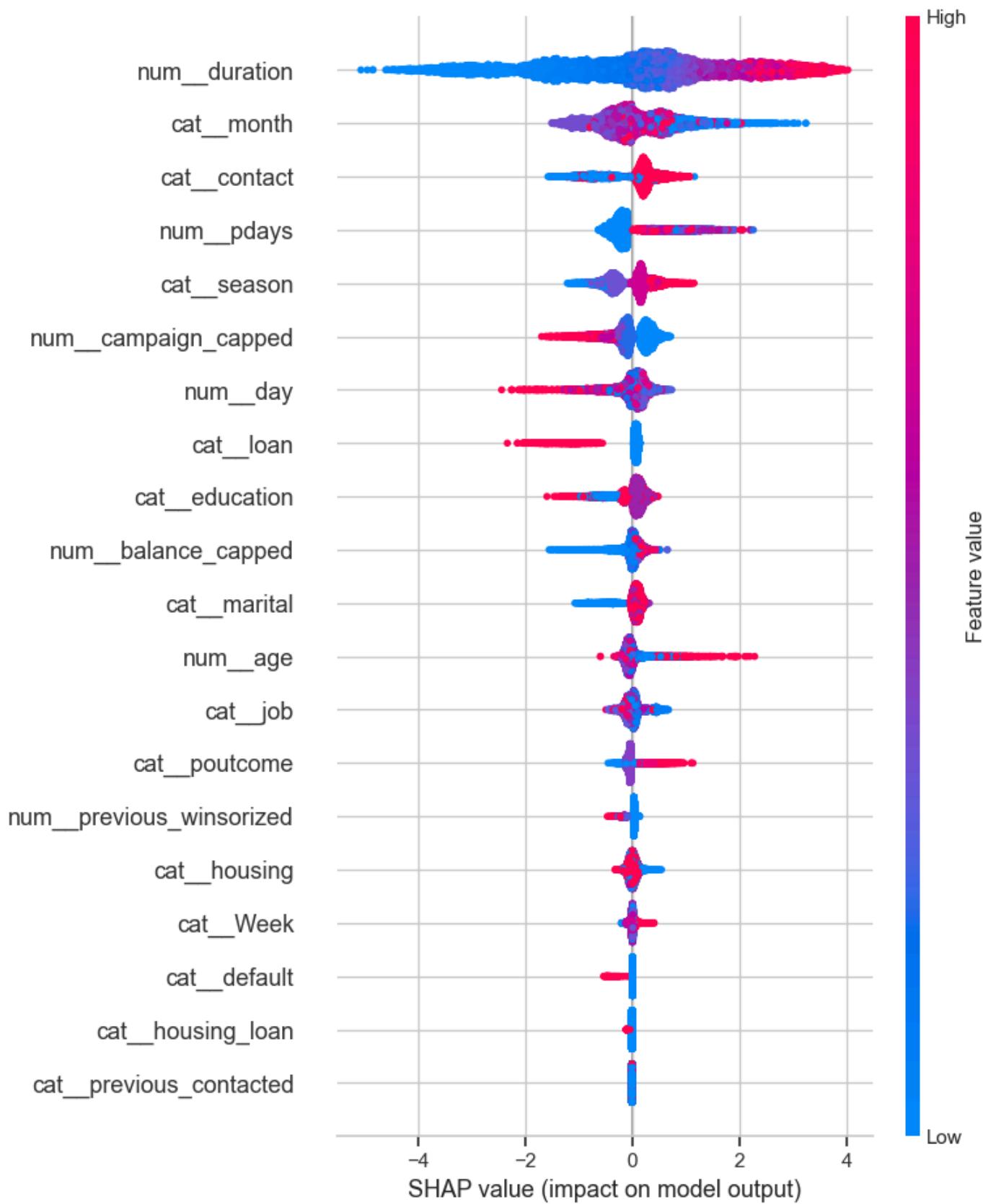


Fitur yang berkaitan dengan cara melakukan panggilan sangat memengaruhi keputusan nasabah untuk berlangganan deposito berjangka. Sementara itu, fitur terkait demografi dan aset yang dimiliki nasabah relatif kurang berpengaruh

8

Evaluasi

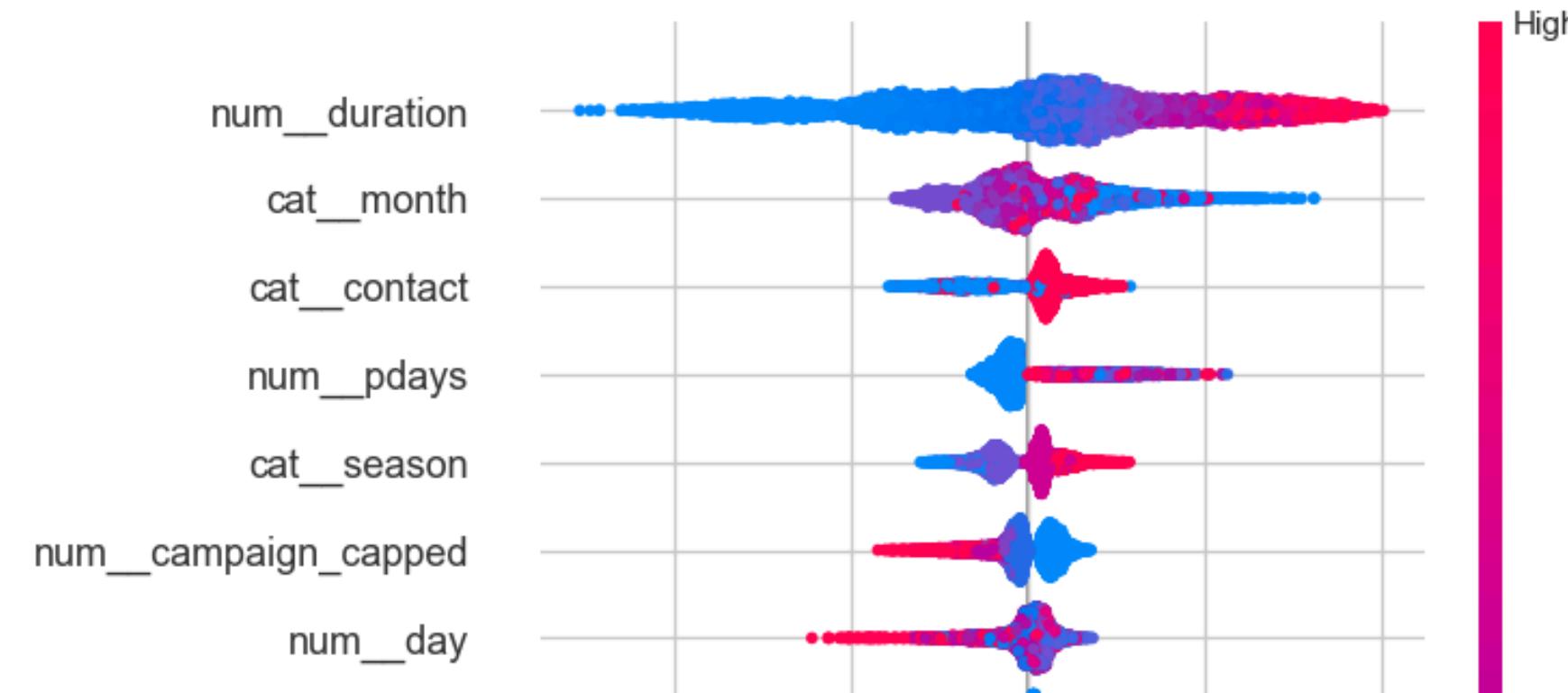
SHAP Values



8

Evaluasi

SHAP Values

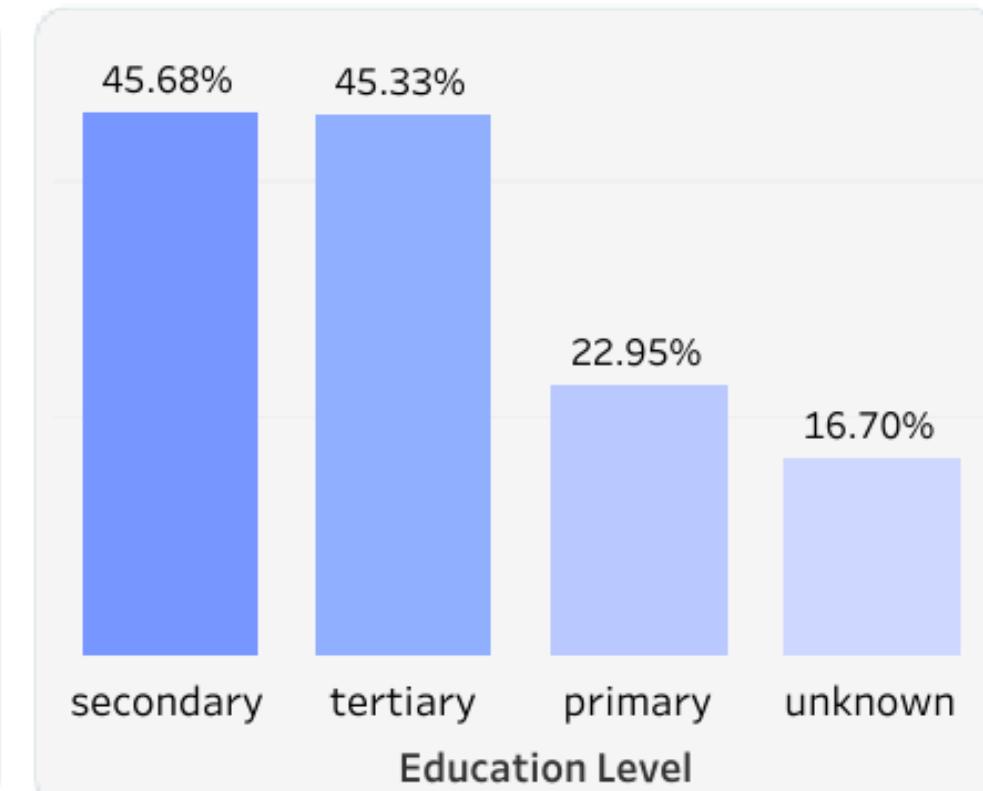


- Secara umum semakin lama durasi panggilan, semakin besar kemungkinan nasabah berlangganan deposito berjangka
- Nasabah yang sudah lama tidak dihubungi (pdays besar) lebih mungkin berlangganan. Sebaliknya, nasabah yang baru saja dihubungi (pdays kecil) cenderung tidak berlangganan, dimungkinkan akibat kejemuhan
- Frekuensi kontak yang terlalu tinggi dapat berdampak negatif menurunkan peluang berlangganan
- Menghubungi nasabah pada akhir bulan cenderung mengurangi probabilitas berlangganan

9

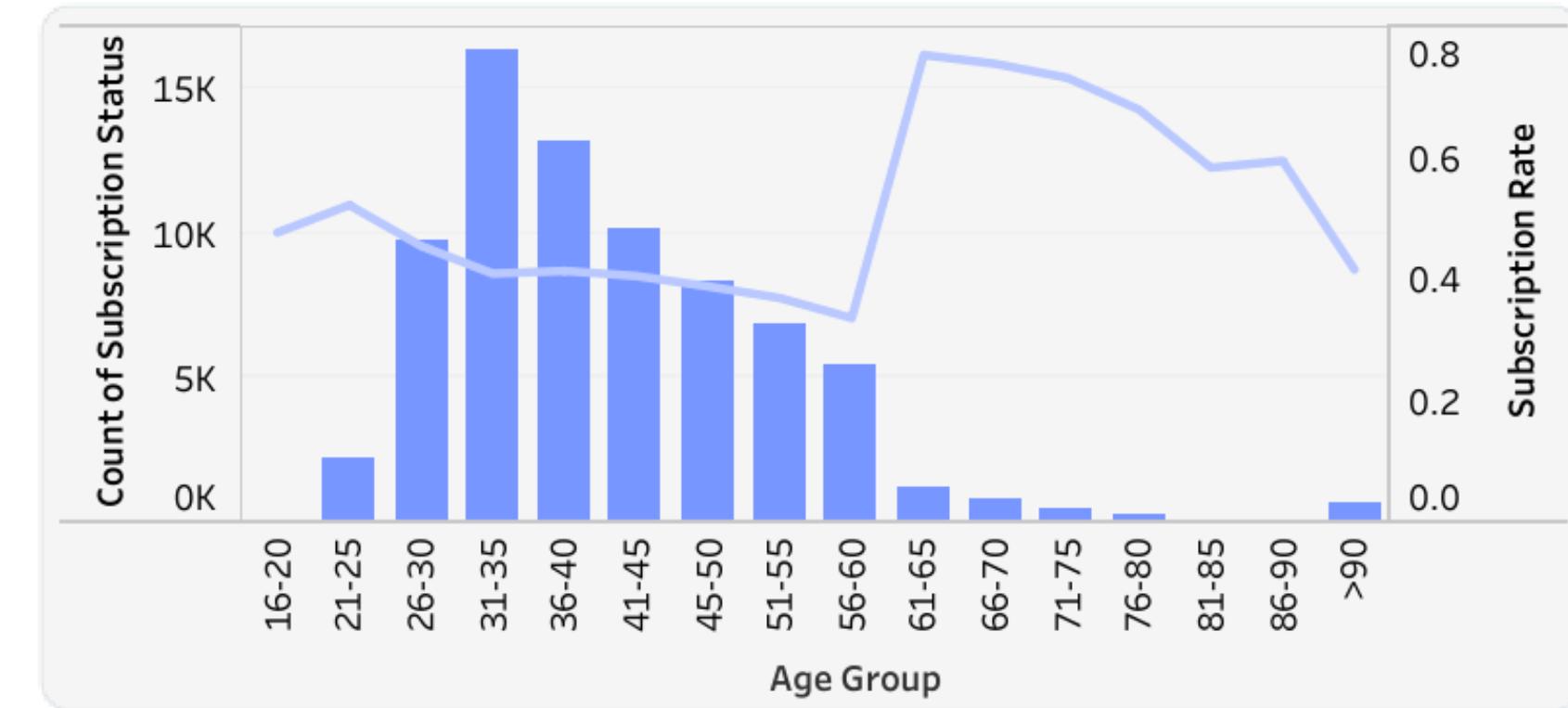
Kesimpulan

subscription rate

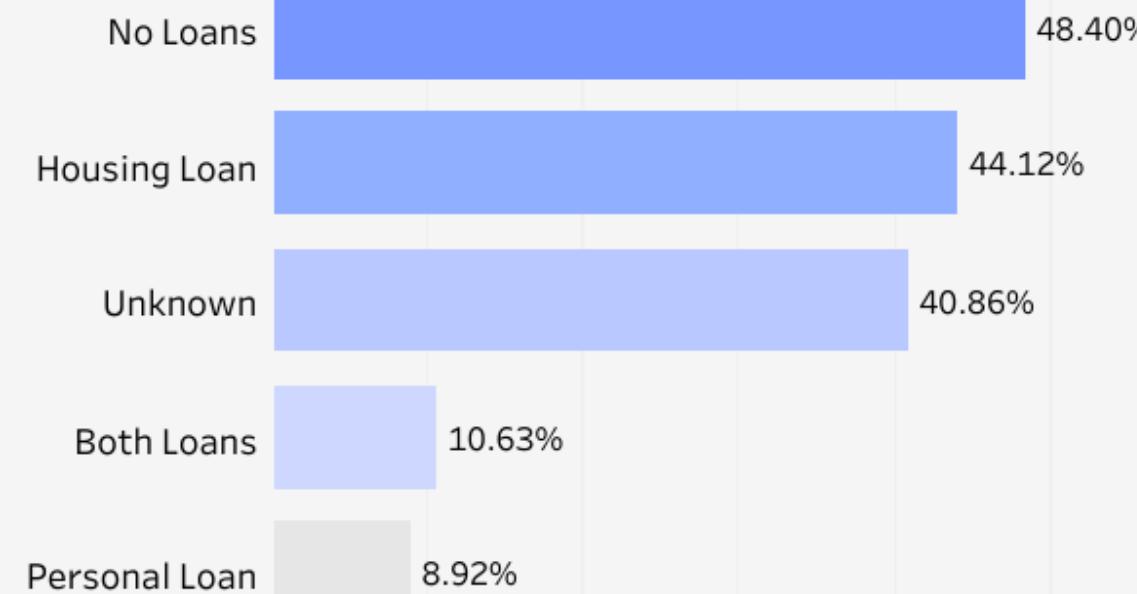


Who To Call

Menghubungi nasabah dengan **usia tinggi** (>60 tahun), **tanpa memiliki pinjaman** dan **berpendidikan paling tidak secondary** akan meningkatkan peluang nasabah berlangganan



Loan Type



9

Kesimpulan

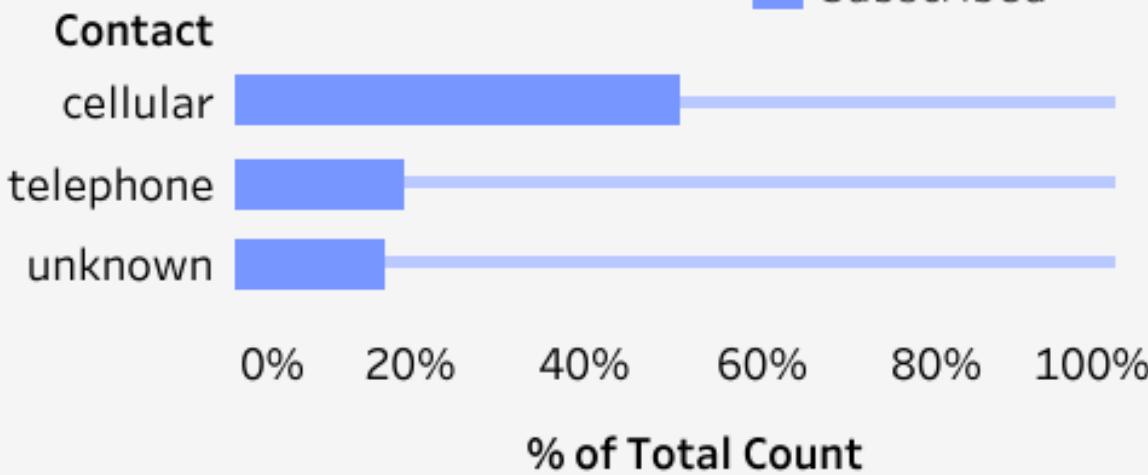
How to Call

Menghubungi dengan **cellular** dengan **durasi** yang tidak terlalu panjang (**8-9 menit**) dengan **menghindari waktu pertengahan tahun dan akhir bulan** akan meningkatkan peluang nasabah berlangganan

Succesful Calls
535.4

Avg Call Duration (Sec)

No Subscribed
Subscribed



Subscription Rate Based On Week and Month

Week	Month											
	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1st	0,9750	0,1985	0,6849	0,6369	0,2256	0,3054	0,4624	0,5066	0,6553	0,7523	0,8874	0,4143
2nd	0,9198	0,7323	0,7740	0,7296	0,2769	0,5692	0,4650	0,4658	0,8333	0,7976	0,9083	0,4615
3rd	0,9678	0,9355	0,7262	0,3797	0,2903	0,4060	0,4686	0,4808	0,8759	0,7869	0,2538	0,5556
4th	0,2551	0,8109	0,6190	0,7259	0,2403	0,5850	0,2473	0,2558	0,7143	0,7947	0,8671	0,5000

Terima Kasih



https://public.tableau.com/app/profile/cintya.kusumawardhani/viz/DataSnipers_DashboardAcademy/CampaignPerformance#1