

UNO

Laporan Final Project Bank UNO Marketing Targets

Ketua : Daniel Rowin

Anggota :

Febby Maghfirani Aziz

Amodya Subagio

Nur Rahman Shalahudin

Fakhry Abdurrohman

Daviro Yota Nagasan Wahyudi

Putri Vina Fajriyani

Asmiyeni Islamiati



UNO

Laporan Final Project
Bank UNO Marketing Targets

Business Understanding

Final Project - Stage 0



01 Problem Statement

Jumlah nasabah yang membuka deposito berjangka pada UNO Bank dari total 45.211 nasabah yang ada, hanya sekitar 5.289 nasabah atau 11,7 % nasabah saja yang membuka deposito berjangka.

Permintaan dari manajemen UNO Bank itu minimal sebesar 15% dari total nasabah UNO bank yang membuka deposito berjangka.

Diantaranya terdapat 4.369 nasabah atau 9.66% nasabah yang dihubungi melalui telepon cellular dan 390 nasabah atau 0.86% nasabah yang dihubungi melalui telepon rumah.

02 Role

Sebagai tim data scientist dalam suatu perusahaan bernama UNO Bank, kami diminta oleh manajemen UNO Bank untuk memprediksi apakah nasabah akan berlangganan deposito berjangka berdasarkan data yang tersedia guna meningkatkan performa dari perusahaan tersebut.

03 Goal

Meningkatnya jumlah nasabah yang membuka deposito berjangka menjadi sebesar 15%. Dimana ini berdasarkan permintaan manajemen UNO Bank.

04 Objective

1. Menganalisis profil nasabah berdasarkan variabel-variabel seperti usia, pekerjaan, status perkawinan, pendidikan, dan fitur lainnya sehingga nasabah tertarik untuk membuka deposito berjangka di UNO Bank.
2. Membangun model yang dapat memprediksi dengan akurat apakah seorang nasabah akan berlangganan deposito berjangka setelah kampanye pemasaran telepon dilakukan berdasarkan klasifikasi nasabahnya.

05 Business Metrics

Conversion Rate : Persentase nasabah UNO Bank yang membuka deposito berjangka

UNO

Laporan Final Project
Bank UNO Marketing Targets

Exploratory Data Analysis

Final Project - Stage 1



1. Descriptive Statistics (15 poin)

Gunakan function `info` dan `describe` pada dataset final project kalian. Tuliskan hasil observasinya, seperti:

- A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?
- B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?
- C. Apakah ada kolom yang memiliki nilai summary agak aneh?
(min/mean/median/max/unique/top/freq)

* Untuk masing-masing jenis observasi, tuliskan juga jika tidak ada masalah, misal untuk A: "Semua tipe data sudah sesuai"

01. Descriptive Statistics

```
df.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column      Non-Null Count Dtype  
--- 
 0   age         45211 non-null    int64  
 1   job          45211 non-null    object  
 2   marital     45211 non-null    object  
 3   education   45211 non-null    object  
 4   default     45211 non-null    object  
 5   balance     45211 non-null    int64  
 6   housing     45211 non-null    object  
 7   loan         45211 non-null    object  
 8   contact     45211 non-null    object  
 9   day          45211 non-null    int64  
 10  month        45211 non-null    object  
 11  duration    45211 non-null    int64  
 12  campaign    45211 non-null    int64  
 13  pdays       45211 non-null    int64  
 14  previous    45211 non-null    int64  
 15  poutcome    45211 non-null    object  
 16  y           45211 non-null    object  
dtypes: int64(7), object(10)
memory usage: 5.9+ MB

```

```
df['duration'].describe()
```

```
count    45211.000000
mean     258.163080
std      257.527812
min      0.000000
25%     103.000000
50%     180.000000
75%     319.000000
max     4918.000000
Name: duration, dtype: float64
```

```
df[nums].describe()
```

	age	balance	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	2.763841	40.197828	0.580323
std	10.618762	3044.765829	3.098021	100.128746	2.303447
min	18.000000	-8019.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	63.000000	871.000000	275.000000

```
df[cats].describe()
```

	job	marital	education	default	housing	loan	contact	poutcome	y
count	45211	45211	45211	45211	45211	45211	45211	45211	45211
unique	12	3	4	2	2	2	3	4	2
top	blue-collar	married	secondary	no	yes	no	cellular	unknown	no
freq	9732	27214	23202	44396	25130	37967	29285	36959	39922

```
df.isna().sum()
```

```
age  
job  
marital  
education  
default  
balance  
housing  
loan  
contact  
day  
month  
duration  
campaign  
pdays  
previous  
poutcome  
y  
dtype: int64
```

01. Descriptive Statistics

A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

Sebaiknya dtype untuk kolom day adalah object karena menunjukkan tanggal.

B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

Semua kolom tidak memiliki nilai kosong, karena nilai kosong pada dataset sudah dikonversikan menjadi unknown, oleh karena itu terdapat beberapa feature yang memiliki value unknown.

C. Apakah ada kolom yang memiliki nilai summary agak aneh?

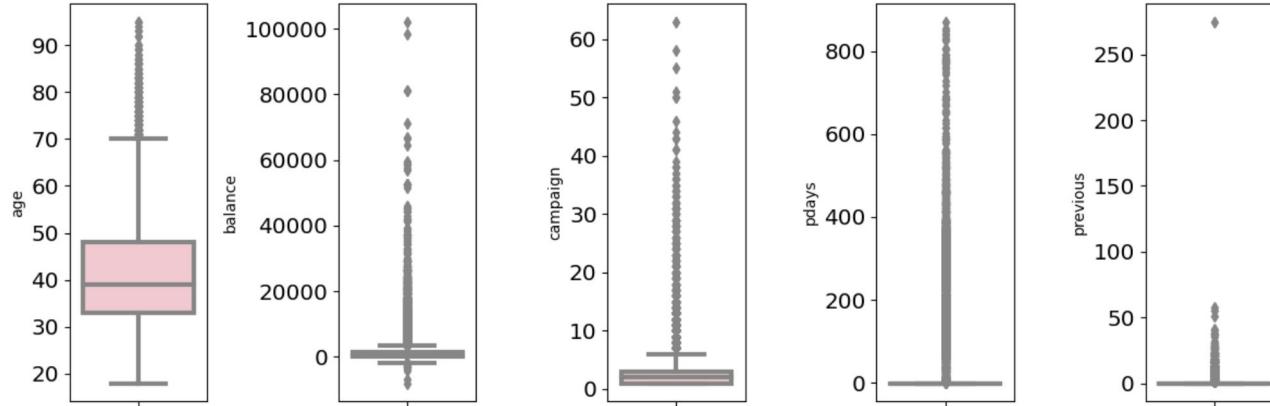
(min/mean/median/max/unique/top/freq)

- Balance terendah adalah -8019, mungkin diperlukan pemeriksaan lanjutan.***
- Terdapat contact dan outcome yang unknown***
- Nilai minimum pada kolom "duration" adalah 0***

2. Univariate Analysis (25 poin)

Gunakan visualisasi untuk melihat distribusi masing-masing kolom (feature maupun target). Tuliskan hasil observasinya, misalnya jika ada suatu kolom yang distribusinya menarik (misal skewed, bimodal, ada outlier, ada nilai yang mendominasi, kategorinya terlalu banyak, dsb). Jelaskan juga apa yang harus di-follow up saat data pre-processing.

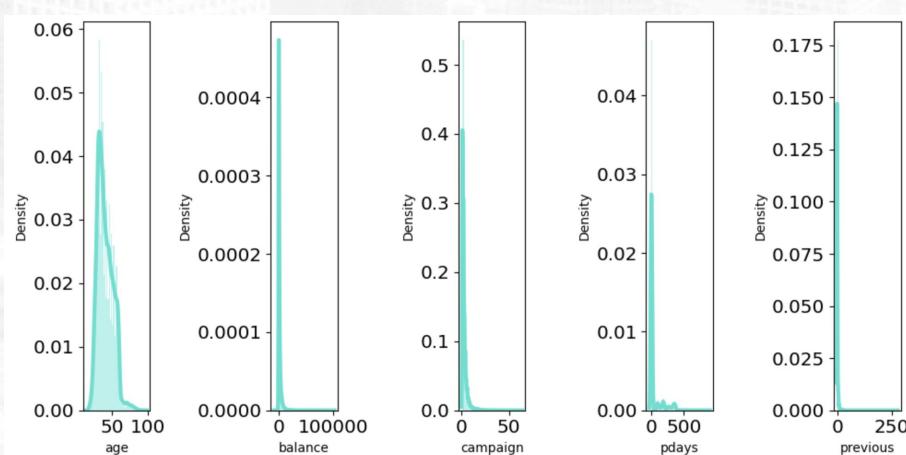
02. Univariate Analysis



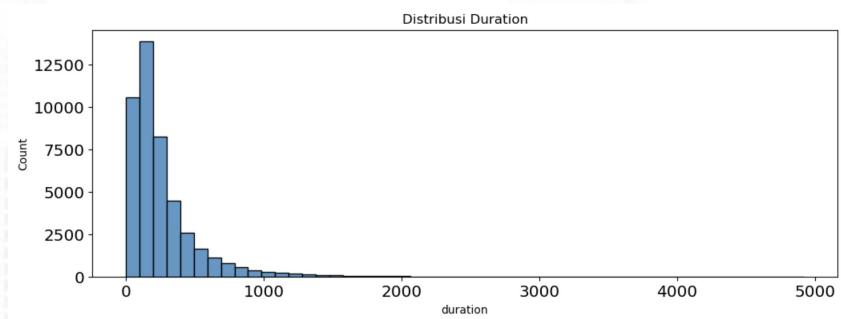
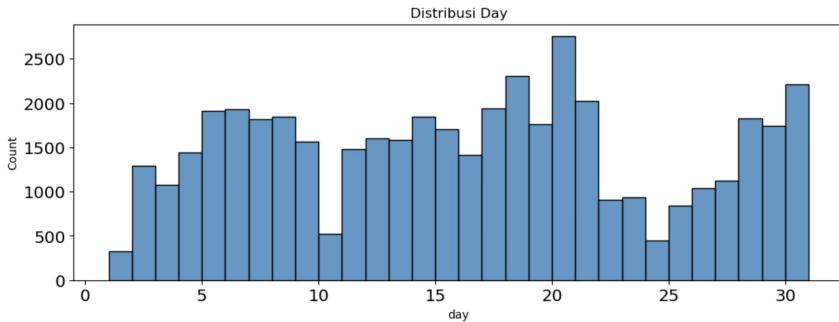
- Terdapat banyak outliers dan skewed pada feature balance, campaign, pdays, dan previous. Menangani nilai outlier pada kolom balance, campaign, dan pdays jika diperlukan.
- Feature age sudah hampir mendekati distribusi normal.

Follow-up untuk Data Pre-processing:

- Menangani nilai outlier pada feature balance, campaign, pdays dan previous jika diperlukan.
- Mungkin perlu melakukan normalisasi atau transformasi pada kolom yang memiliki skewness signifikan untuk mendapatkan distribusi yang lebih normal, terutama pada kolom balance.

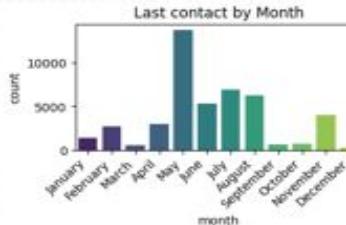
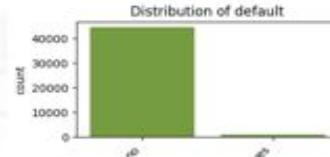
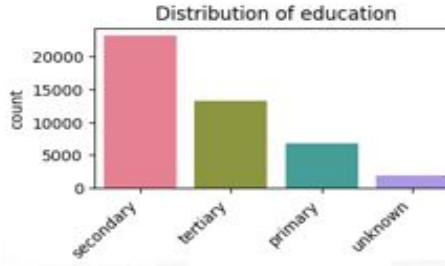
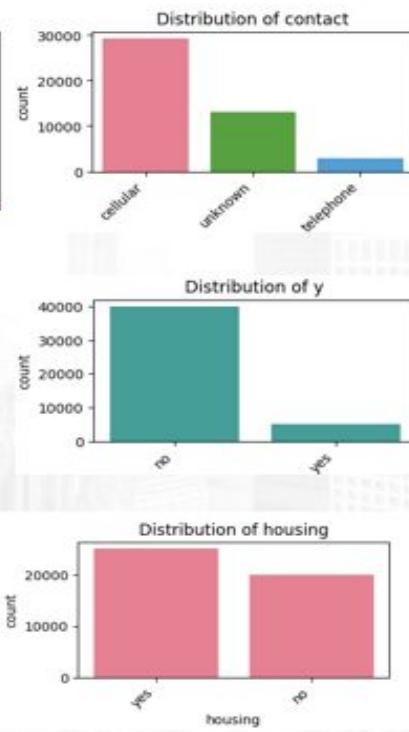
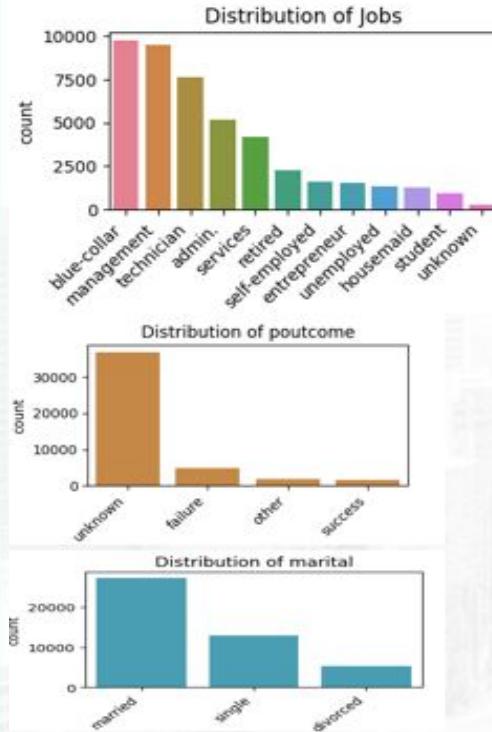


02. Univariate Analysis



- Kebanyakan data menunjukkan last contact day of the month adalah tanggal 20-21.
- Last contact duration kebanyak berlangsung kurang dari 1000 detik.

02. Univariate Analysis



- "Blue-collar", "management", "technician", "admin", dan "services" memiliki distribusi data dengan jumlah terbanyak
- Kebanyakan dari data menunjukkan status "married"
- "Secondary" education menunjukkan jumlah hasil terbanyak
- Kebanyakan dari data menunjukkan "no" default
- Kebanyakan nasabah dihubungi via cellular dan tidak memiliki loan
- Last contact month terbanyak adalah bulan May
- Untuk kolom job, pendidikan (education), dan kontak (contact) yang memiliki nilai "unknown", jika dianggap perlu, pada tahap pre-prosesing kemungkinan bisa diganti dengan nilai modusnya.
- Kebanyakan outcome marketing campaign (poutcome) memiliki nilai "unknown".

3. Multivariate Analysis (15 poin)

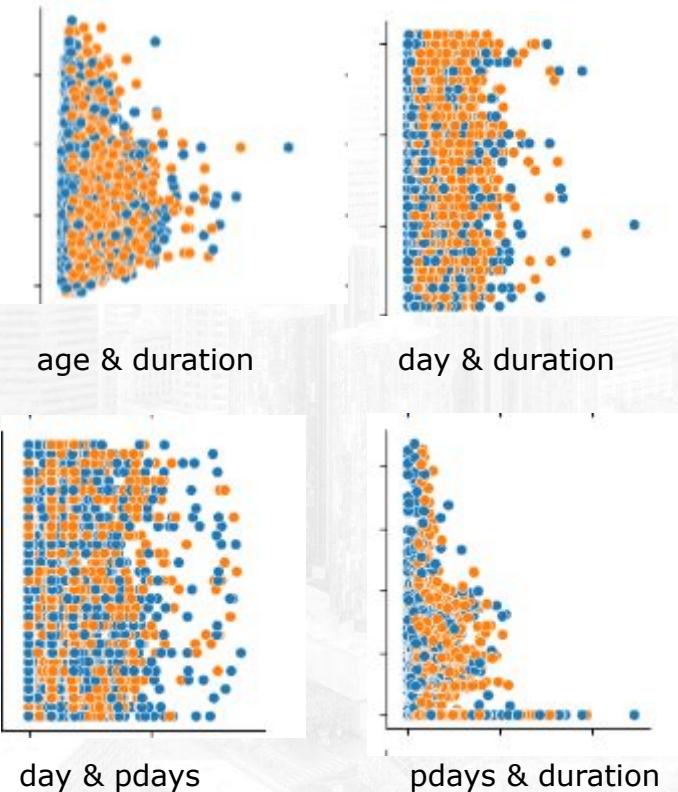
Lakukan multivariate analysis (seperti correlation heatmap dan category plots, sesuai yang diajarkan di kelas). Tuliskan hasil observasinya, seperti:

- A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?
- B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

* Tuliskan juga jika memang tidak ada feature yang saling berkorelasi

3. Multivariate Analysis

A. Korelasi antar feature dan label

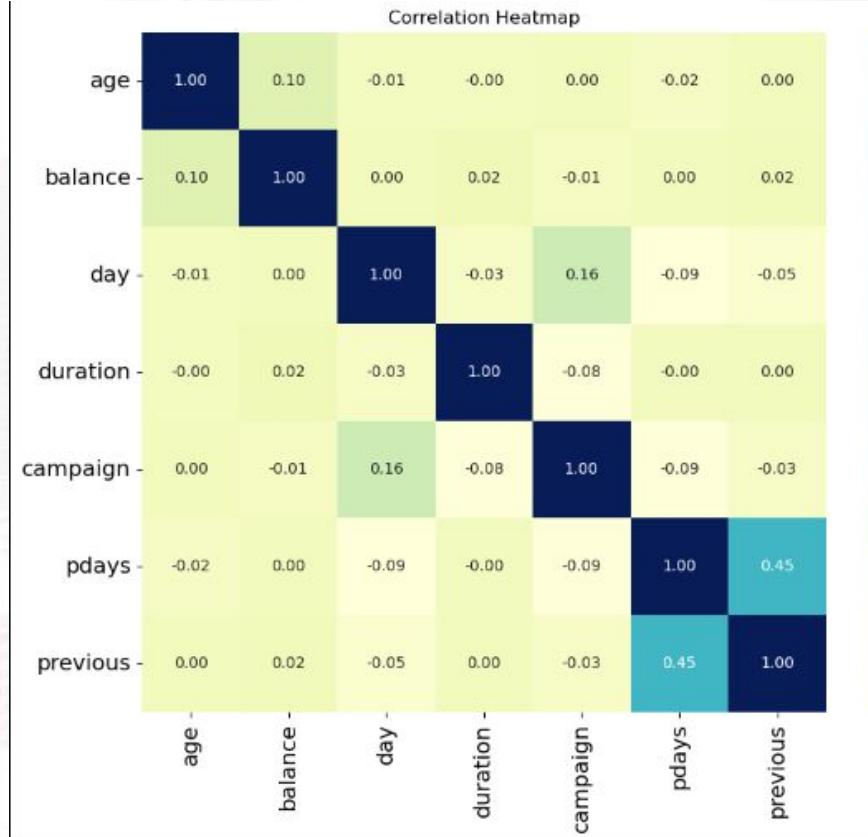


Korelasi antara feature dan label yang mengindikasikan kombinasi yang baik :

- age & duration
- day & duration
- day & pdays
- pdays & duration

Untuk korelasi antara feature dan label apabila dalam scatter plot terdapat kecenderungan terpisahnya antara plot feature dan label mengindikasikan kombinasi yang baik.

3. Multivariate Analysis



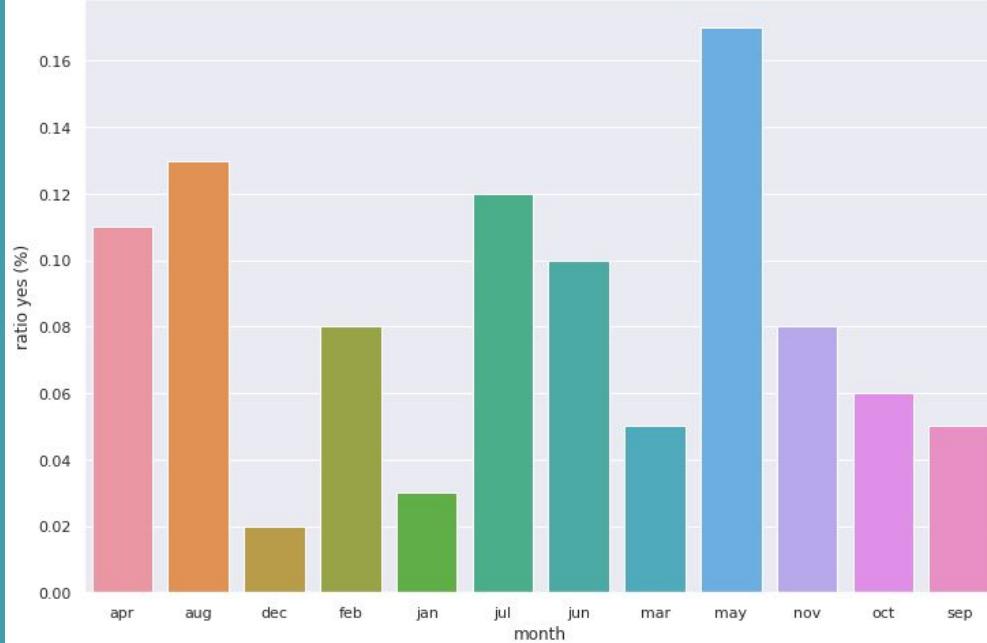
B. Korelasi antar feature :

Berdasarkan heatmap ini, dapat dilihat bahwa korelasi antar feature tidak terlalu kuat karena sebagian besar nilai korelasi tidak melebihi angka 0,7. Namun ada beberapa fitur yang memiliki nilai yang lebih tinggi dengan yang lain yaitu korelasi antara pdays & previous yang memiliki korelasi positif sebesar (0.45), day & campaign (0.16), age & balance (0.10), campaign & pdays, pdays & days dengan nilai korelasi negatif (-0,09)

Tindakan yang perlu dilakukan adalah menjaga fitur fitur yang memiliki nilai korelasi tinggi untuk dilakukan analisis lebih lanjut guna memahami hubungan hubungan antar fitur tersebut.

3. Multivariate Analysis

Hubungan Antara Fitur "month" Dengan Nasabah Yang Memilih "yes"



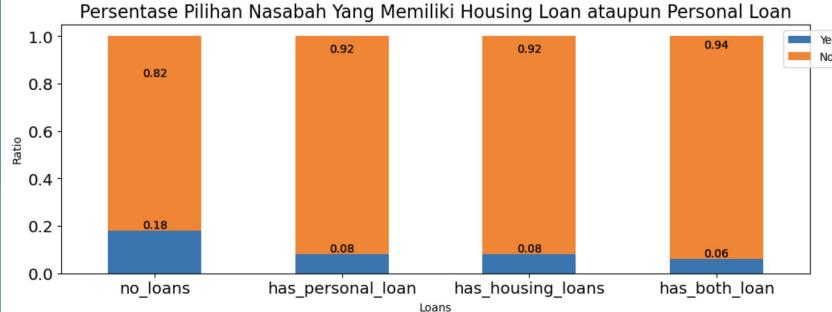
C. Ratio Yes (%)

Dapat dilihat bahwa pada bulan may persentase customer yang melakukan pendaftaran deposito sedikit lebih tinggi dibandingkan bulan-bulan lainnya. Namun jika kita melihat secara keseluruhan pada setiap bulannya jumlah customer yang mendaftar tidaklah berbeda jauh antara satu bulan ke bulan lainnya, bahkan semua bulan tidak ada yg memiliki persentase "yes" lebih besar dari 0,2%. Hal ini dapat menjadi landasan jika kita memutuskan untuk tidak menggunakan feature month terhadap model yang akan kita buat, dikarenakan feature month tidak memiliki pengaruh yang signifikan perihal seorang customer tertarik mendaftar deposito berjangka.

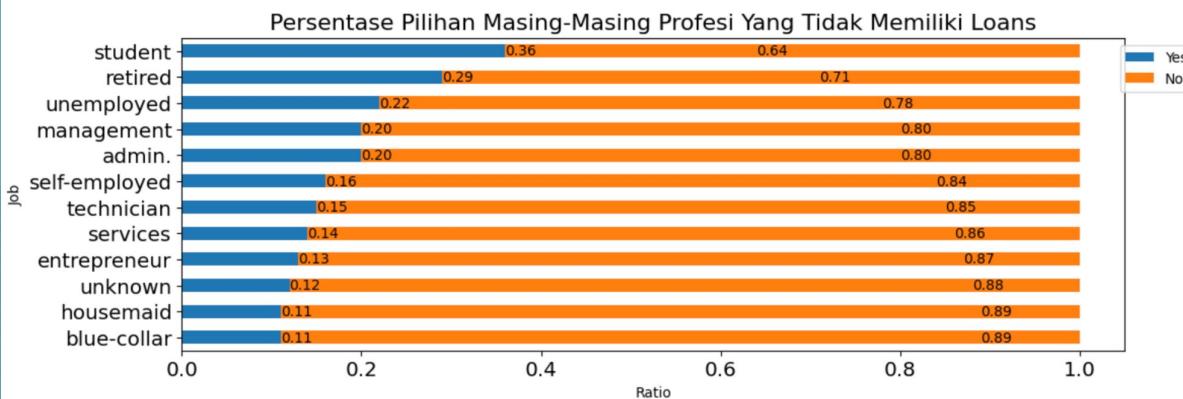
4. Business Insight (30 poin)

Selain EDA, lakukan juga beberapa analisis dan visualisasi untuk menemukan suatu business insight. Tuliskan minimal 3 insight, dan berdasarkan insight tersebut jelaskan rekomendasinya untuk bisnis.

04. Business Insight



Nasabah yang membuka term deposit didominasi oleh nasabah yang **tidak memiliki housing loans maupun personal loans** dan jika dilihat berdasarkan pekerjaannya. Nasabah yang membuka term deposit kebanyakan beprofesi sebagai **student, retired, unemployed, admin, dan management**. Direkomendasikan kepada pihak marketing untuk mencari pendekatan yang paling baik bagi nasabah yang berprofesi sebagai blue-collar dan housemaid karena memiliki rasio terendah.



04. Business Insight

y campaign

0 no 2.846350

1 yes 2.141047

```
import scipy.stats as st
# Hypothesis Testing using mann whitney
stat, p_value= st.mannwhitneyu(yes['campaign'],no['campaign'],alternative='two-sided')
p_value

1.9484904873905108e-71

alpha = 0.05
print('P-Value :',p_value)

if p_value >= alpha:
    print('Tidak cukup bukti jumlah campaign mampu membedakan user untuk membuka akun atau tidak')
else:
    print('cukup bukti jumlah campaign mampu membedakan user untuk membuka akun atau tidak')
```

P-Value : 1.9484904873905108e-71

cukup bukti jumlah campaign mampu membedakan user untuk membuka akun atau tidak

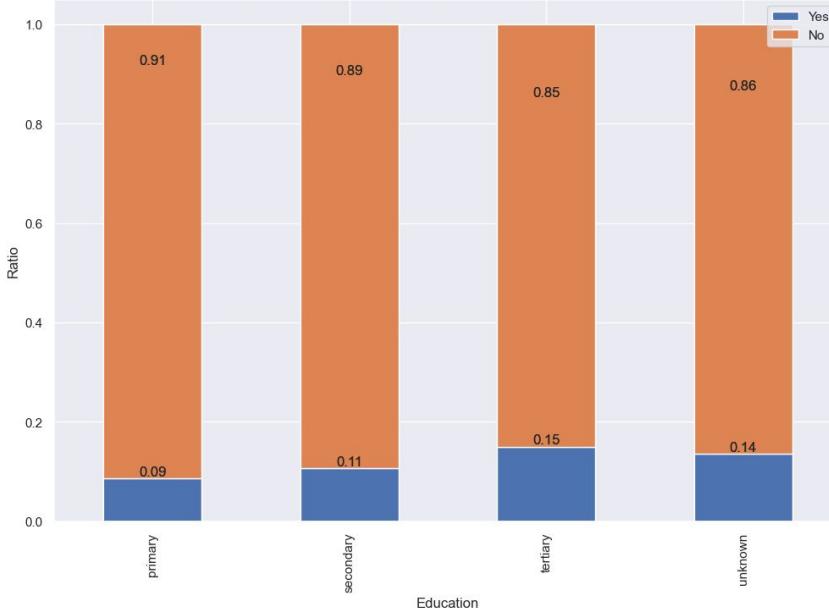
Hasil uji hipotesis menunjukkan bahwa $p\text{-value} < \alpha$, maka kita akan mengambil keputusan bahwa jumlah campaign berpengaruh terhadap nasabah untuk membuka akun atau tidak secara signifikan.

Namun berdasarkan berdasarkan rata-rata ternyata semakin banyak campaign yang diberikan ternyata user akan semakin menolak membuka akun.

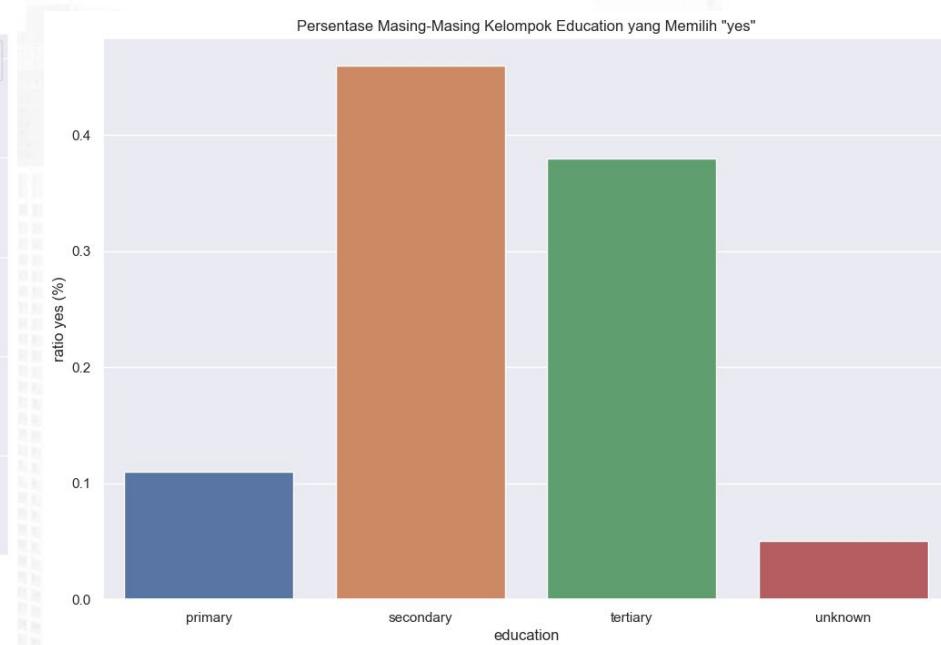
Jadi kesimpulannya, jumlah campaign berhubungan terbalik dengan user membuka akun.

04. Business Insight

Percentase Masing-Masing Kelompok Education yang Memilih "yes"



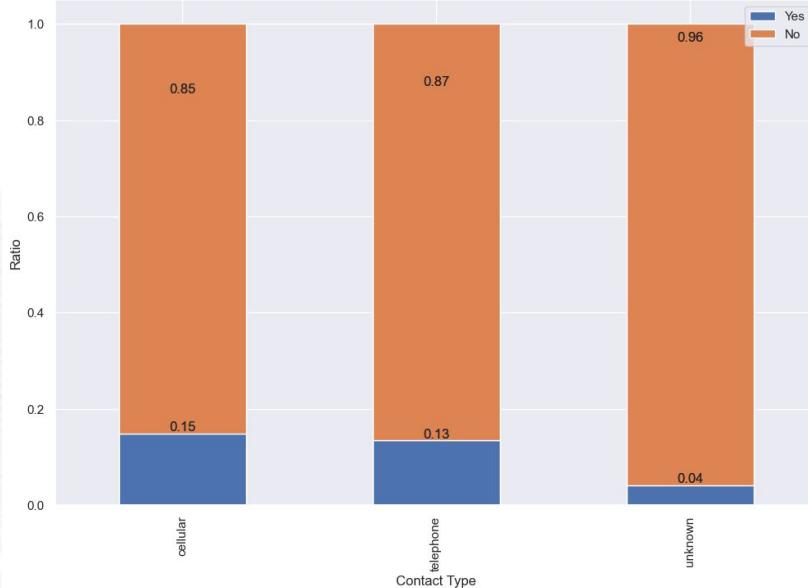
Percentase Masing-Masing Kelompok Education yang Memilih "yes"



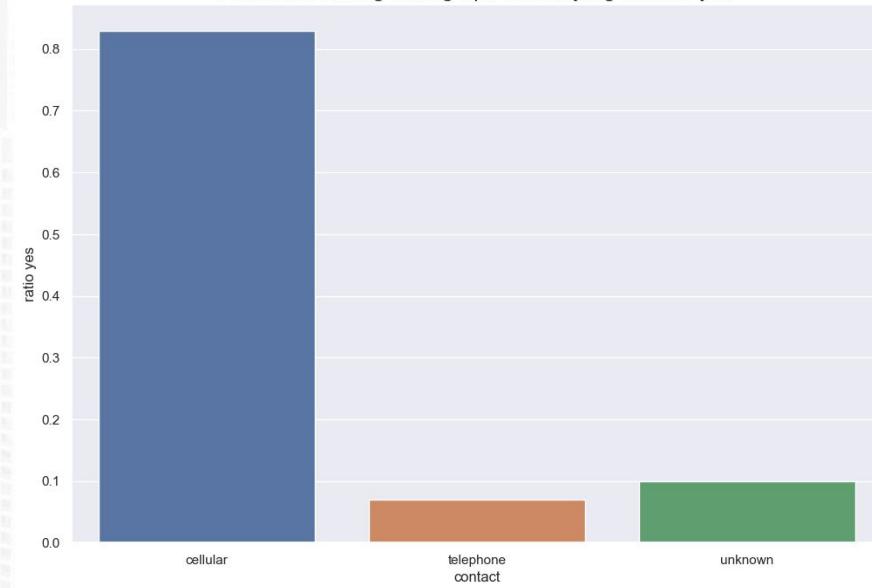
- Pada visualisasi diatas dapat dilihat bahwa kelompok education **tertiary** (lulusan S1 atau di atasnya) paling banyak memilih untuk mendaftar deposito berjangka sebanyak 15% dibanding total yang mendaftar di kelompoknya, yang diikuti oleh kelompok education **yang tidak diketahui**, lalu **secondary** (lulusan SMP dan SMA) dan terakhir **primary** (Lulusan SD ke bawah).
- Dengan membandingkan terhadap mereka yang mengambil yes saja, terlihat bahwa kelompok education **secondary** paling banyak memilih yes, diikuti dekat kelompok **tertiary**, lalu **primary** dan kelompok Education **yang tidak diketahui**.
- Dari insight tersebut, rekomendasi bisnis yang dapat diberikan adalah menargetkan campaign terhadap kelompok **tertiary**. Bisa di daerah kampus, universitas, mengajak investasi sebagai kesempatan untuk mendapatkan passive income untuk uang yang mungkin telah mereka tabung.
- Rekomendasi ke dua, bisa ditargetkan untuk mereka yang masih di jenjang sekolah, utamanya **secondary**. Bisa melakukan campaign ke sekolah-sekolah, melibatkan orang tua, untuk membuka rekening dan deposito dini.

04. Business Insight

Percentase Pilihan Masing-Masing Cara Contact



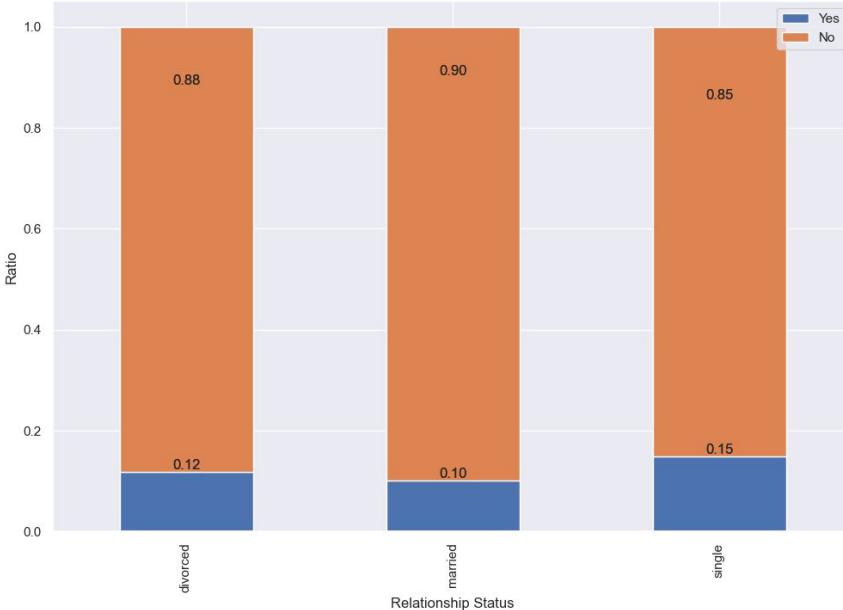
Percentase Masing-Masing Tipe Contact yang Memilih "yes"



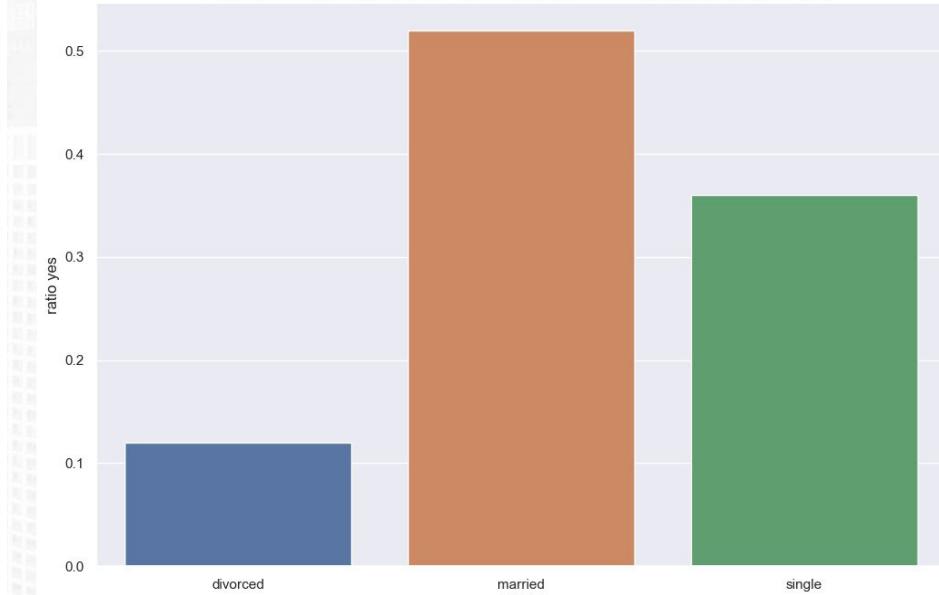
Bisa dilihat melalui visualisasi diatas bahwa tipe contact **cellular** (HP) merupakan tipe contact yang paling banyak menghasilkan customer untuk mendaftar deposito berjangka. Hal ini bisa terjadi karena di era sekarang ini orang-orang lebih banyak melakukan komunikasi melalui telepon cellular dibandingkan telepon biasa (telepon rumah), dan hampir setiap orang pasti mempunyai telepon cellular. Dari sini pihak bank **direkomendasi** mulai merubah strategi campaign untuk memprioritaskan melakukan contact melalui telepon cellular. Selain itu pihak bank sebelum melakukan panggilan secara langsung ke nomor telepon cellular, bisa juga bisa melakukan campaign melalui email atau melalui whatsapp sebagai tahapan awal dalam menawarkan deposito berjangka. Dengan demikian peluang customer yang mendaftar deposito berjangka akan meningkat.

04. Business Insight

Percentase Pilihan Masing-Masing Status Pernikahan



Percentase Masing-Masing Tipe Status Pernikahan yang Memilih "yes"



- Pada visualisasi diatas dapat dilihat bahwa **yang masih single** paling banyak memilih untuk mendaftar deposito berjangka sebanyak 15% dibanding total yang mendaftar di antara nasabah yang masih single, lalu diikuti oleh yang sudah **bercerai**, dan terakhir mereka yang sudah **menikah** dalam 10%.
- Dengan membandingkan terhadap mereka yang mengambil yes saja, terlihat bahwa **yang sudah menikah** paling banyak memilih yes, diikuti mereka yang masih **single**, dan terakhir mereka yang sudah **ceraia**.
- Rekomendasi untuk meningkatkan mereka yang mendaftar deposito berjangka pihak bank saat melakukan campaign kepada masing-masing individu dapat melakukan promosi dengan strategi pendekatan yang berbeda. Misalkan kepada kelompok "**single**" bisa melakukan promosi seperti "dengan bunga deposito dalam setahun adalah x% maka kira-kira dalam y tahun dana menikah akan dapat terkumpul". Sedangkan untuk kelompok "**married**" bisa melakukan promosi seperti "dengan bunga deposito dalam setahun adalah x% maka dana pendidikan untuk anak/dana pensiun saat masa tua akan terjamin"

5. Git (15 poin)

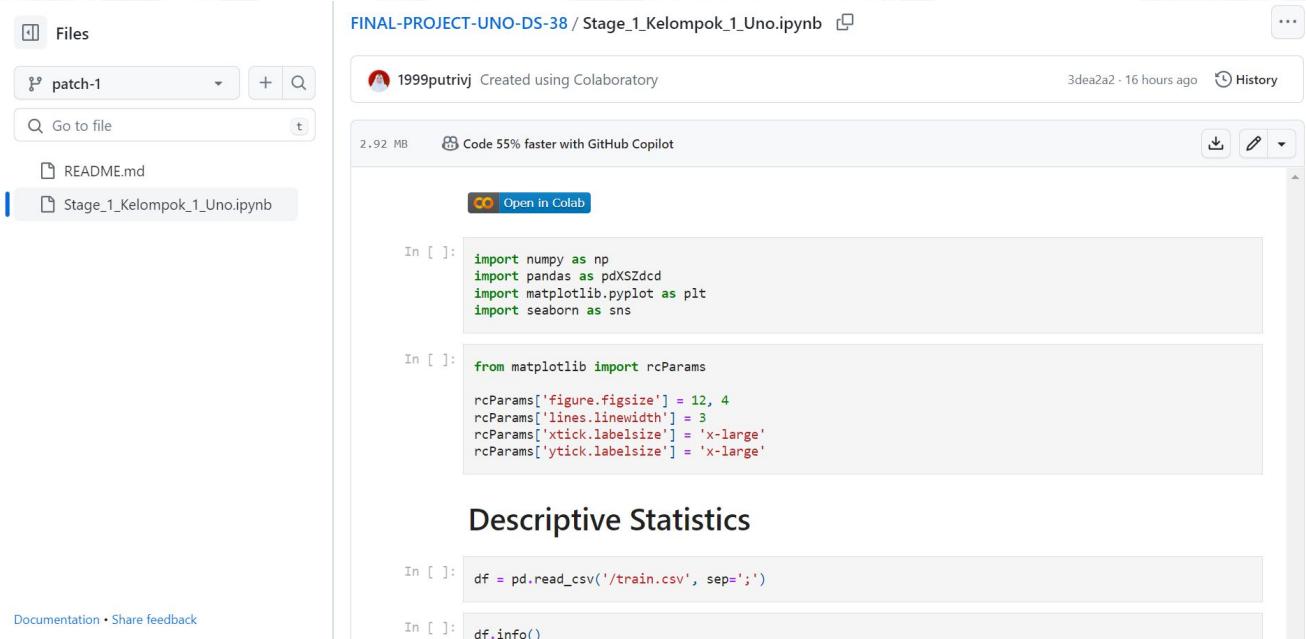
Upload project teman-teman di sebuah repository git. Berkolaborasilah di Git jika ada perubahan version dari waktu ke waktu.

- A. Buat Repository Git
- B. Upload file notebook atau file penggerjaan lainnya pada repository tersebut

Untuk file README, dapat merupakan summary insight yang telah didapatkan dari EDA.

5. Git (15 poin)

Git - Uno



The screenshot shows a GitHub repository interface. On the left, there's a sidebar with a 'Files' section containing a 'patch-1' folder, a 'README.md' file, and the 'Stage_1_Kelompok_1_Uno.ipynb' notebook, which is currently selected. The main area displays the contents of the notebook:

FINAL-PROJECT-UNO-DS-38 / Stage_1_Kelompok_1_Uno.ipynb

Created using Colaboratory · 16 hours ago · History

Code 55% faster with GitHub Copilot

[Open in Colab](#)

```
In [ ]:  
import numpy as np  
import pandas as pdXSZdc  
import matplotlib.pyplot as plt  
import seaborn as sns
```

```
In [ ]:  
from matplotlib import rcParams  
  
rcParams['figure.figsize'] = 12, 4  
rcParams['lines.linewidth'] = 3  
rcParams['xtick.labelsize'] = 'x-large'  
rcParams['ytick.labelsize'] = 'x-large'
```

Descriptive Statistics

```
In [ ]:  
df = pd.read_csv('/train.csv', sep=';')
```

```
In [ ]:  
df.info()
```

Laporan Final Project
Bank UNO Marketing Targets

Data Pre-Processing

Final Project - Stage 2



Estimasi Waktu Penggerjaan

 **3 - 5 jam**

Jumlah Soal

 **2 Soal**

Total Point

 **100 poin**

Teknis Pengerjaan

1. Pekerjaan dilakukan secara **berkelompok, sesuai kelompok Final Project**
2. Masing-masing anggota kelompok tetap perlu submit ke LMS (jadi bukan perwakilan)
3. File yang perlu dikumpulkan:
 - o File **jupyter notebook** (.ipynb) yang berisi source code.
 - o File **laporan homework** (.pdf) yang berisi rangkuman dari apa saja yang telah dilakukan.
4. Upload hasil pengerjaanmu melalui LMS.
 - o Masukkan semua file ke dalam **1 file** dengan format **ZIP**.
 - o Nama File:
Preprocessing - <Nama Kelompok>.zip

1. Data Cleansing (50 poin)

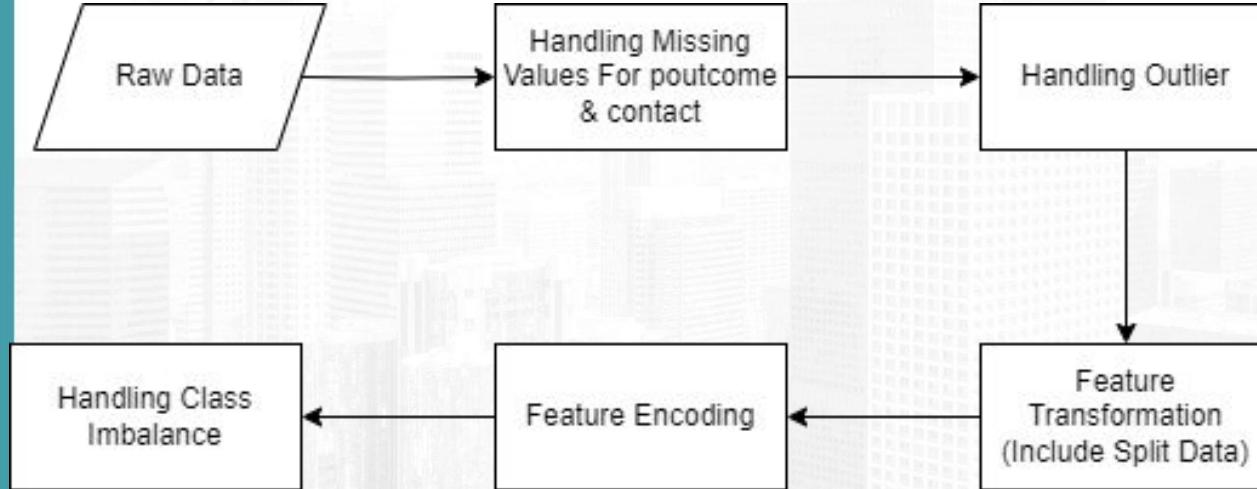
Lakukan pembersihan data, sesuai yang diajarkan di kelas, seperti:

- A. Handle missing values
- B. Handle duplicate data
- C. Handle outliers
- D. Feature transformation
- E. Feature encoding
- F. Handle class imbalance

Di laporan homework, tuliskan apa saja yang telah dilakukan dan metode yang digunakan.

* Tetap tuliskan jika memang ada tidak yang perlu di-handle (contoh: "Tidak perlu feature encoding karena semua feature sudah numerical" atau "Outlier tidak di-handle karena akan fokus menggunakan model yang robust terhadap outlier").

Diagram Pre-Processing



Feature Transformation

StandardScaler()

Split Train-Test Data

random state = 42
Split = 0.2

Oversampling

SMOTE

Number of Train Data:
35828

Number of Test Data: 8957

A. Handle missing values

```
df.isna().sum()
```

```
age          0  
job          0  
marital      0  
education    0  
default       0  
balance      0  
housing      0  
loan          0  
contact       0  
day           0  
month         0  
duration      0  
campaign      0  
pdays         0  
previous      0  
poutcome      0  
y              0  
dtype: int64
```

```
[ ] df['poutcome'].replace({'unknown': 'never'}, inplace=True)
```

Tidak terdapat missing values pada dataset, namun untuk fitur poutcome yang memiliki value unknown dilakukan replacement menjadi never yang menunjukkan bahwa customer tidak pernah dihubungi karena memiliki pdays = -1.

Pada contact, terdapat unknown yang diartikan tidak diketahui cara kontak melalui apa, diikuti dengan cara seluler dan telefon (rumah). Di sini, value unknown akan dijadikan seluler disebabkan terdapat feature duration, yang berarti durasi campaign yang dilakukan, dan lebih mudah untuk mengasumsi suatu nomor telefon itu seluler dibandingkan telefon rumah (sebab terdapat distinct dari 3 atau 4 angka pertama [seperti 031, 021, dst]).

B. Handle duplicated data

```
[180] df.duplicated().sum()
```

```
0
```

Tidak terdapat duplicated values pada dataset

C. Handle Outliers

```
[ ] nums2 = ['age', 'campaign']
for num in nums2:
    df[num] = np.log(df[num])
```

```
▶ from scipy import stats

print("Before removing outlier: ", len(df))

for num in nums2:
    z_scores = np.abs(stats.zscore(df[num]))
    df = df[z_scores < 3]

print("After removing outlier: ", len(df))
```

```
➡ Before removing outlier: 45211
After removing outlier: 44790
```

Beberapa fitur numerik pada dataset memiliki sebaran yang right skewed (long right tailed), oleh karena itu dilakukan log transformation terlebih dahulu sebelum me-remove outliers. Log transformation hanya dilakukan pada fitur age dan campaign saja, karena jika semua fitur numerik dilakukan log transformation jumlah data setelah remove outlier akan hilang secara keseluruhan atau berjumlah 0. Setelah log transformation pada fitur age dan campaign, selanjutnya menghapus outliers menggunakan z-scores yang mana jumlah data berkurang dari 45.211 menjadi 44.790.\

D. Feature transformation

```
[ ] from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split

X = df.drop(['y'], axis=1)
y = df['y']

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42, test_size=0.2)

print(f'Number of Train Data: {y_train.shape[0]}')
print(f'Number of Test Data: {y_test.shape[0]}')

Number of Train Data: 35832
Number of Test Data: 8958
```

Sebelum dilakukan feature transformation pada, dataset dibagi menjadi train data dan test data terlebih dahulu. Yang mana 80% data merupakan train data yang berjumlah 35.835 dan test data sebanyak 20% dari keseluruhan dataset yang berjumlah 8.958

D. Feature transformation

```
▶ from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
columns_to_standardize = ['age', 'balance', 'campaign', 'pdays', 'previous']
X_train[columns_to_standardize] = scaler.fit_transform(X_train[columns_to_standardize])
X_test[columns_to_standardize] = scaler.transform(X_test[columns_to_standardize])
print("DataFrame setelah distandardisasi:")
X_train.head()
```

Selanjutnya dilakukan feature scaling pada train data dan test data dengan standarization untuk fitur numerik.

E. Feature Encoding

```
[ ] mapping_default = {
    'no' : 0,
    'yes' : 1,
}
X_train['default'] = X_train['default'].map(mapping_default)
X_test['default'] = X_test['default'].map(mapping_default)

[ ] mapping_housing = {
    'no' : 0,
    'yes' : 1,
}
X_train['housing'] = X_train['housing'].map(mapping_housing)
X_test['housing'] = X_test['housing'].map(mapping_housing)

▶ mapping_loan = {
    'no' : 0,
    'yes' : 1,
}
X_train['loan'] = X_train['loan'].map(mapping_loan)
X_test['loan'] = X_test['loan'].map(mapping_loan)
```

```
[ ] X_train_encoded_education = pd.get_dummies(X_train['education'], prefix = 'pendidikan')
X_test_encoded_education = pd.get_dummies(X_test['education'], prefix = 'pendidikan')

[ ] X_train_encoded_kerja = pd.get_dummies(X_train['job'], prefix = 'kerja')
X_test_encoded_kerja = pd.get_dummies(X_test['job'], prefix = 'kerja')

[ ] X_train_encoded_marital = pd.get_dummies(X_train['marital'], prefix = 'status')
X_test_encoded_marital = pd.get_dummies(X_test['marital'], prefix = 'status')

[ ] X_train_encoded_contact = pd.get_dummies(X_train['contact'], prefix = 'contact')
X_test_encoded_contact = pd.get_dummies(X_test['contact'], prefix = 'contact')

[ ] X_train_encoded_poutcome = pd.get_dummies(X_train['poutcome'], prefix = 'poutcome')
X_test_encoded_poutcome = pd.get_dummies(X_test['poutcome'], prefix = 'poutcome')
```

- Feature Encoding merupakan proses mengubah feature categorical menjadi feature numeric.
- Pada data yang bertipe ordinal dan distinct values = 2 (ya/tidak) diubah menggunakan Label Encoding, sisanya diubah menggunakan one hot encoding

F. Handle class imbalance

```
▶ y_train.value_counts()
```

```
→ no      31603  
yes     4229  
Name: y, dtype: int64
```

```
[ ] # OVERSAMPLING
```

```
from imblearn import over_sampling  
X_oversampling , y_oversampling = over_sampling.SMOTE(random_state=42).fit_resample(X_train_combined,y_train)  
print(pd.Series(y_oversampling).value_counts())
```

```
no      31603  
yes    31603  
Name: y, dtype: int64
```

- Oversampling SMOTE pada data train yang memiliki ketimpangan pada ditribusi target

2. Feature Engineering (35 poin)

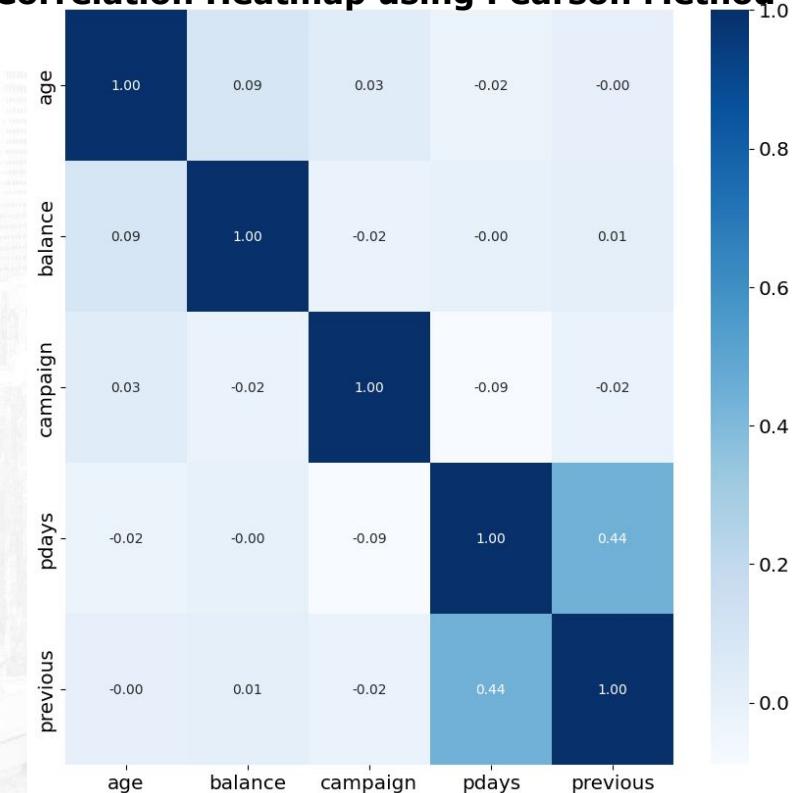
Cek feature yang ada sekarang, lalu lakukan:

- A. Feature selection (membuang feature yang kurang relevan atau redundan)
- B. Feature extraction (membuat feature baru dari feature yang sudah ada)
- C. Tuliskan minimal 4 feature tambahan (selain yang sudah tersedia di dataset) yang mungkin akan sangat membantu membuat performansi model semakin bagus (ini hanya ide saja, untuk menguji kreativitas teman-teman, tidak perlu benar-benar dicari datanya dan tidak perlu diimplementasikan)

* Untuk 2A & 2B, tetap tuliskan jika memang tidak bisa dilakukan (contoh: "Semua feature digunakan untuk modelling (tidak ada yang dihapus), karena semua feature relevan")

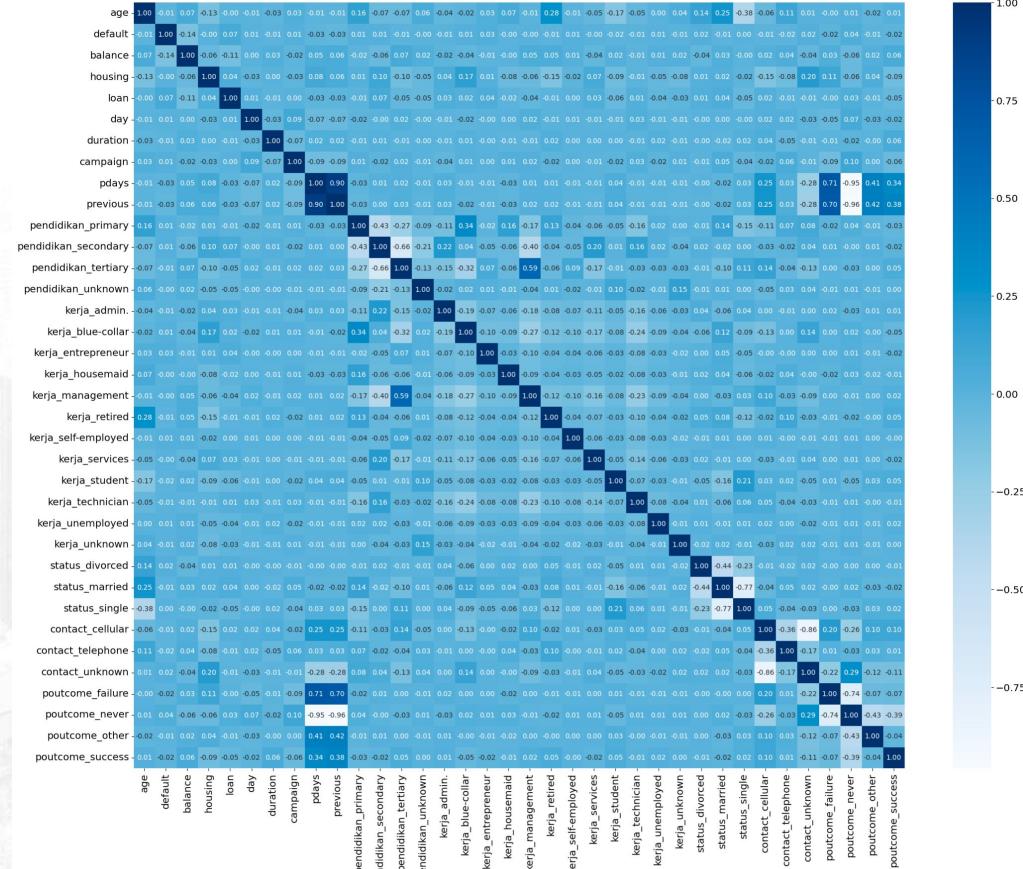
Feature Engineering

Numerical-Numerical Correlation Heatmap using Pearson Method

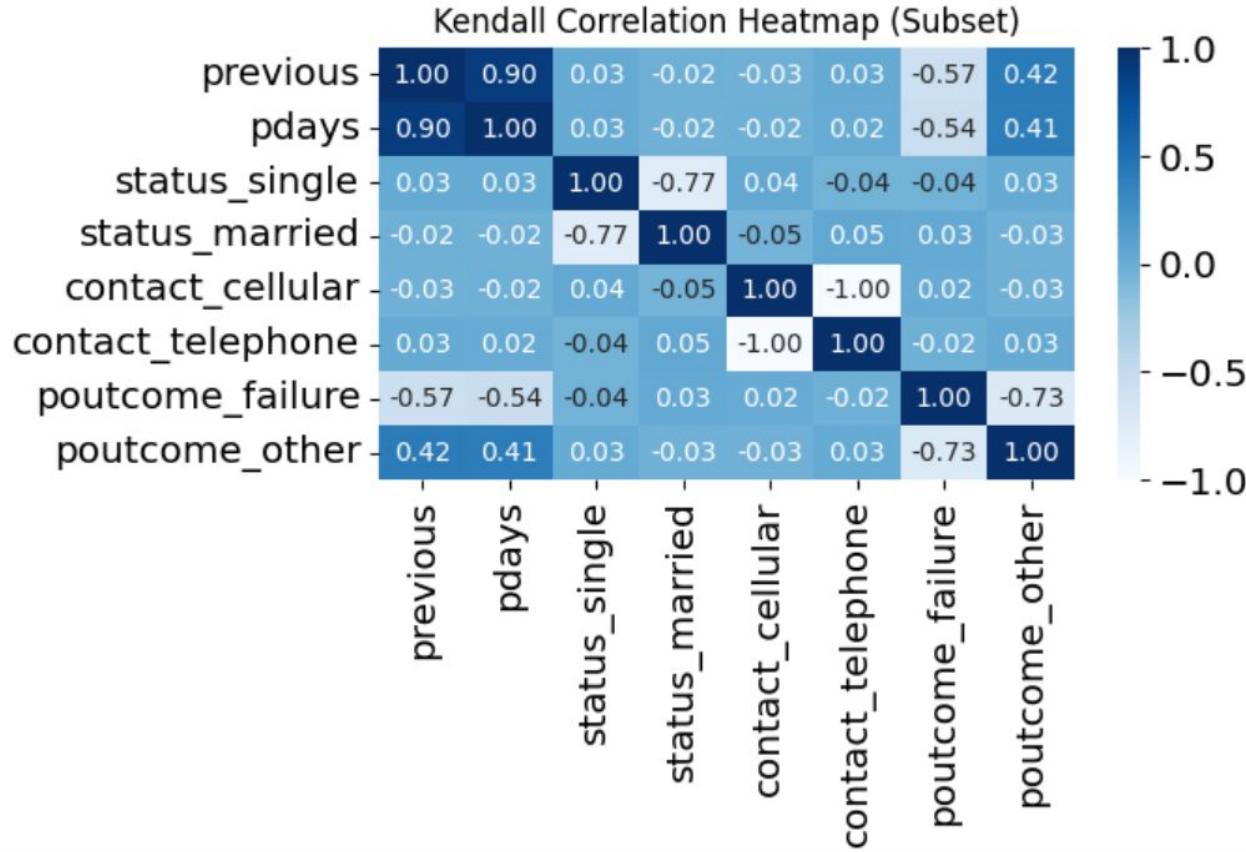


Feature Engineering

Categorical-Numerical and Categorical-Categorical Correlation Heatmap using Kendall Method



Feature Engineering



Feature Engineering

Delete Unnecessary Columns

Sebab tidak ada feature tanggal yang lengkap, hanya akan didrop '**day**' sebab korelasinya terhadap **y** kecil, serta '**month**' sebab dari apa yang telah disimpulkan pada EDA Multivariate Analysis. Ditambah, day dan month tanpa time series year dan tanpa adanya info kapan campaign dilaksanakan, menjadi sulit untuk menarik kesimpulan dari hal tersebut.

Delete Redundant Data

Feature '**previous**' dengan '**pdays**', '**status_single**' dengan '**status_married**', '**contact_cellular**' dengan '**contact_unknown**', '**poutcome_failure**' dengan '**poutcome_never**', memiliki korelasi di atas 0.7 yang menyebabkan mereka redundant untuk dijadikan feature bersama, sehingga akan digunakan salah satu saja.

Feature Engineering

Feature Extraction

Dengan adanya 32 Feature, dirasa telah cukup feature yang dibutuhkan, dengan feature ini memiliki relevansi tersendiri terhadap 'target' yang ingin dicapai. Sehingga kami **tidak akan mengeluarkan feature baru.**

Ide Feature *for future references*

1. Jumlah anak/tanggungan
2. Memiliki produk deposito berjangka pada bank lain (y/n)
3. Sudah berapa lama menjadi nasabah bank tersebut
4. Memiliki produk investasi lain selain deposito berjangka (y/n)
5. Durasi setiap campaign (dengan informasi waktu yang lebih memadai)

Feature Engineering

17 Feature Raw Data

Age
Job
Marital
Education
Default
Balance
Housing
Loan
Contact
Day
Month
Duration
Campaign
Pdays
Previous
Poutcome
Y

35 Engineered Feature

age
default
balance
housing
loan
day
duration
campaign
pdays
previous
pendidikan_primary
pendidikan_secondary
pendidikan_tertiary
pendidikan_unknown
kerja_admin
kerja_blue-collar
kerja_entrepreneur
kerja_housemaid
kerja_management
kerja_retired
kerja_self-employed
kerja_services
kerja_student
kerja_technician
kerja_unemployed
kerja_unknown
status_divorced
status_married
status_single
contact_cellular
contact_telephone
poutcome_failure
poutcome_never
poutcome_other
poutcome_success

3. Git (15 poin)

Upload project teman-teman di sebuah repository git. Berkolaborasilah di Git jika ada perubahan version dari waktu ke waktu.

- A. Buat Repository Git
- B. Upload file notebook atau file penggerjaan lainnya pada repository tersebut

Untuk file README, dapat merupakan summary dari proses data preproses yang telah dilakukan. Boleh menggunakan repositori yang sama atau membuat baru.

Git

README

B. Handle Duplicate Data

Tidak terdapat duplicated values pada dataset

C. Handle Outliers

Beberapa fitur numerik pada dataset memiliki sebaran yang right skewed (long right tailed), oleh karena itu dilakukan log transformation terlebih dahulu sebelum me-remove outliers. Log transformation hanya dilakukan pada fitur age dan campaign saja, karena jika semua fitur numerik dilakukan log transformation jumlah data setelah remove outlier akan hilang secara keseluruhan atau berjumlah 0. Setelah log transformation pada fitur age dan campaign, selanjutnya menghapus outliers menggunakan z-scores yang mana jumlah data berkurang dari 45.211 menjadi 44.790.

D. Feature Transformation

- Sebelum dilakukan feature transformation, dataset dibagi menjadi train data dan test data terlebih dahulu. Yang mana 80% data merupakan train data yang berjumlah 35.835 dan test data sebanyak 20% dari keseluruhan dataset yang berjumlah 8.958
- Selanjutnya dilakukan feature scaling pada train data dan test data dengan standarization untuk fitur numerik.

README

Feature Engineering

Delete Unnecessary Columns

Sebab tidak ada feature tanggal yang lengkap, hanya akan didrop 'day' sebab korelasinya terhadap y kecil, serta 'month' sebab dari apa yang telah disimpulkan pada EDA Multivariate Analysis. Ditambah, day dan month tanpa time series year dan tanpa adanya info kapan campaign dilaksanakan, menjadi sulit untuk menarik kesimpulan dari hal tersebut.

Delete Redundant Data

Feature 'previous' dengan 'pdays', 'status_single' dengan 'status_married', 'contact_cellular' dengan 'contact_unknown', 'poutcome_failure' dengan 'poutcome_never', memiliki korelasi di atas 0.7 yang menyebabkan mereka redundant untuk dijadikan feature bersama, sehingga akan digunakan salah satu saja.

Feature Extraction

Dengan adanya 32 Feature, dirasa telah cukup feature yang dibutuhkan, dengan feature ini memiliki relevansi tersendiri terhadap 'target' yang ingin dicapai. Sehingga kami tidak akan mengeluarkan feature baru.

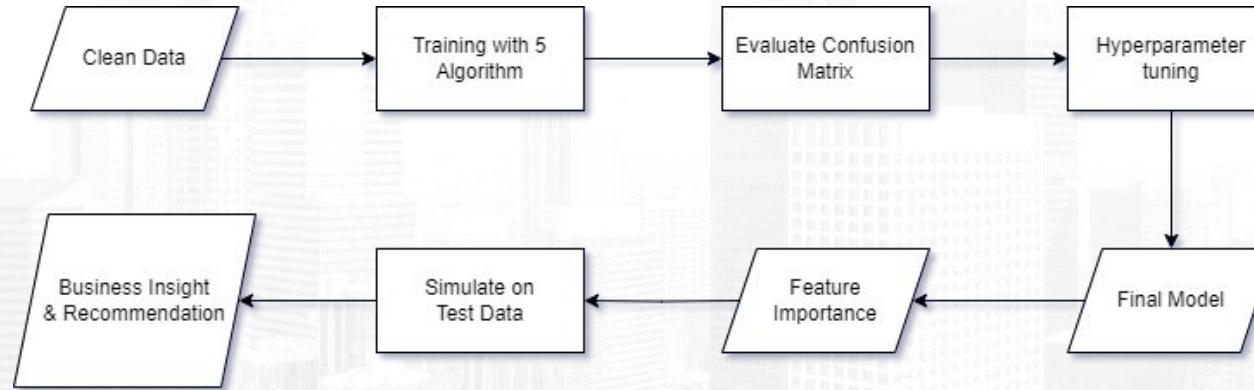
[Link Github Final Project Stage 2 - Data Pre-Processing](#)

Laporan Final Project Bank UNO Marketing Targets **Modeling**

Final Project - Stage 3



Modelling



Algorithm

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. AdaBoost
5. XGBoost

Modelling

Algorithm Supervised Machine Learning

Untuk Modelling, kami menggunakan beberapa Algoritma yang telah dipelajari, yaitu,

1. Logistic Regression

- Interpretability: Logistic Regression adalah model yang relatif mudah diinterpretasikan, yang membuatnya berguna untuk memahami hubungan antara fitur dan probabilitas kejadian target.
- Efisiensi: Cocok untuk skenario ketika asumsi bahwa hubungan antara fitur dan target bersifat linier.
- Ketahanan terhadap Overfitting: Logistic Regression cenderung kurang rentan terhadap overfitting pada data yang relatif sederhana.

2. Decision Tree

- Kemampuan Memprediksi Nonlinearitas: Decision Tree dapat menangani hubungan non-linier antara fitur dan target.
- Keputusan yang Mudah Diinterpretasikan: Decision Tree menghasilkan struktur pohon yang mudah diinterpretasikan, memungkinkan pemahaman yang baik tentang faktor-faktor yang mempengaruhi keputusan.

Modelling

Algorithm Supervised Machine Learning

3. Random Forest

- Kombinasi Kelebihan Decision Tree: Random Forest menggabungkan keunggulan Decision Tree dan mengatasi kelemahan-kelemahan seperti overfitting dengan menggunakan ensemble learning.
- Mampu Menangani Banyak Fitur: Cocok untuk data dengan banyak fitur karena dapat menangani sejumlah besar variabel prediktif.

4. AdaBoost

- Peningkatan Kinerja: AdaBoost dapat meningkatkan kinerja model dengan fokus pada kasus yang sulit diidentifikasi oleh model sebelumnya, sehingga efektif untuk meningkatkan recall.
- Kemampuan Menyesuaikan: Mampu menyesuaikan diri dengan model lemah yang lebih baik untuk menangani ketidakseimbangan kelas.

5. XGBoost

- Performa yang Tinggi: XGBoost merupakan algoritma yang sangat efisien dan sering kali memberikan kinerja yang sangat baik, terutama dalam hal akurasi dan waktu komputasi.
- Ketahanan terhadap Overfitting: Memiliki teknik penanganan overfitting dan kemampuan untuk menangani kompleksitas model.

Modelling

Langkah-langkah yang akan diterapkan:

1. Melakukan modelling untuk masing-masing algorithm model
2. Mengevaluasi Model dengan parameter confusion matrix + f1 score
3. Menerapkan hyperparameter tuning, akan digunakan metode cross-validation
4. Mengevaluasi Model yang telah di-tune
5. Mengambil Feature Importance, dan Insight Bisnis

Evaluasi Confusion Matrix

Train-Test Split						
Algorithm	AUC-Score Train	AUC-Score Test	Recall	Precision	F1	Accuracy
Logistic Regression	0.89	0.88	0.65	0.78	0.69	0.90
Decision Tree	1.00	0.67	0.67	0.66	0.66	0.86
Random Forest	1.00	0.88	0.65	0.76	0.68	0.90
AdaBoost	0.90	0.89	0.66	0.77	0.69	0.90
XGBoost	0.97	0.89	0.69	0.75	0.71	0.90

Evaluasi Confusion Matrix

Dari tabel hasil evaluasi model menggunakan beberapa metrik kinerja, terutama AUC-Score Train, AUC-Score Test, Recall, Precision, F1, dan Accuracy, tampaknya XGBoost memberikan kinerja yang baik secara keseluruhan. Berikut adalah beberapa alasan mengapa kami memilih XGBoost berdasarkan hasil tersebut:

AUC-Score (Area Under the Curve):

XGBoost memiliki nilai AUC-Score yang tinggi baik pada data latih maupun data uji. AUC adalah metrik yang baik untuk mengukur kemampuan model memisahkan kelas positif dan negatif. Semakin tinggi AUC, semakin baik model dalam membedakan antara kelas.

Recall:

Recall pada XGBoost juga relatif tinggi. Recall mengukur sejauh mana model dapat mengidentifikasi keseluruhan kasus positif. Pada konteks kampanye pemasaran, recall yang tinggi berarti model mampu mendeteksi sebanyak mungkin pelanggan yang benar-benar berlangganan, yang merupakan hal yang penting.

Precision:

Precision XGBoost terbilang baik, menunjukkan bahwa dari prediksi positif yang dilakukan model, sebagian besar adalah benar. Ini penting untuk meminimalkan jumlah kontak yang tidak perlu kepada pelanggan yang sebenarnya tidak berlangganan.

Evaluasi Confusion Matrix

F1-Score:

F1-Score yang baik pada XGBoost menunjukkan keseimbangan yang baik antara recall dan precision. F1-Score menggabungkan kedua metrik tersebut menjadi satu skor, memberikan gambaran komprehensif tentang kinerja model.

Accuracy:

Tingkat akurasi XGBoost juga cukup tinggi, menunjukkan seberapa baik model dalam membuat prediksi yang benar secara keseluruhan.

Konsistensi Kinerja (Train dan Test):

Model XGBoost menunjukkan konsistensi kinerja baik pada data latih maupun data uji, yang menunjukkan kemampuan umumnya dalam menggeneralisasi dari data pelatihan ke data baru.

Dengan kombinasi nilai yang tinggi pada berbagai metrik, XGBoost dapat dianggap sebagai pilihan yang baik untuk model pada kasus ini

Evaluasi Confusion Matrix

Hyperparameter Tuning						
Algorithm	AUC-Score Train	AUC-Score Test	Recall	Precision	F1	Accuracy
Logistic Regression	0.88	0.89	0.65	0.78	0.68	0.90
Decision Tree	0.77	0.78	0.65	0.77	0.69	0.90
Random Forest	0.89	0.92	0.58	0.82	0.61	0.90
AdaBoost	0.90	0.90	0.66	0.77	0.69	0.90
XGBoost	0.90	0.92	0.68	0.77	0.72	0.90

Evaluasi Confusion Matrix

Meskipun hasil evaluasi model setelah penyetelan hyperparameter menunjukkan kinerja yang tinggi untuk beberapa model, ada beberapa perbedaan yang dapat memengaruhi pemilihan model. Berikut adalah beberapa alasan mengapa kami memilih XGBoost berdasarkan hasil tersebut:

AUC-Score (Area Under the Curve):

XGBoost memiliki nilai AUC-Score yang konsisten baik pada data latih maupun data uji, menunjukkan kemampuan model untuk membedakan antara kelas positif dan negatif. AUC-Score yang tinggi adalah indikator keunggulan model dalam hal ini.

Recall:

Recall pada XGBoost lebih tinggi dibandingkan dengan model lainnya. Recall yang tinggi merupakan faktor penting dalam konteks kampanye pemasaran, karena kita ingin mendeteksi sebanyak mungkin pelanggan yang benar-benar berlangganan.

F1-Score:

XGBoost memiliki F1-Score yang lebih tinggi dibandingkan dengan model lainnya. F1-Score menyatukan recall dan precision, dan nilai yang tinggi menunjukkan keseimbangan yang baik antara kedua metrik tersebut.

Evaluasi Confusion Matrix

Konsistensi Kinerja (Train dan Test):

Model XGBoost menunjukkan konsistensi kinerja baik pada data latih maupun data uji, menandakan kemampuan umumnya untuk menggeneralisasi dari data pelatihan ke data baru. Konsistensi ini dapat membantu mengurangi risiko overfitting.

Presisi:

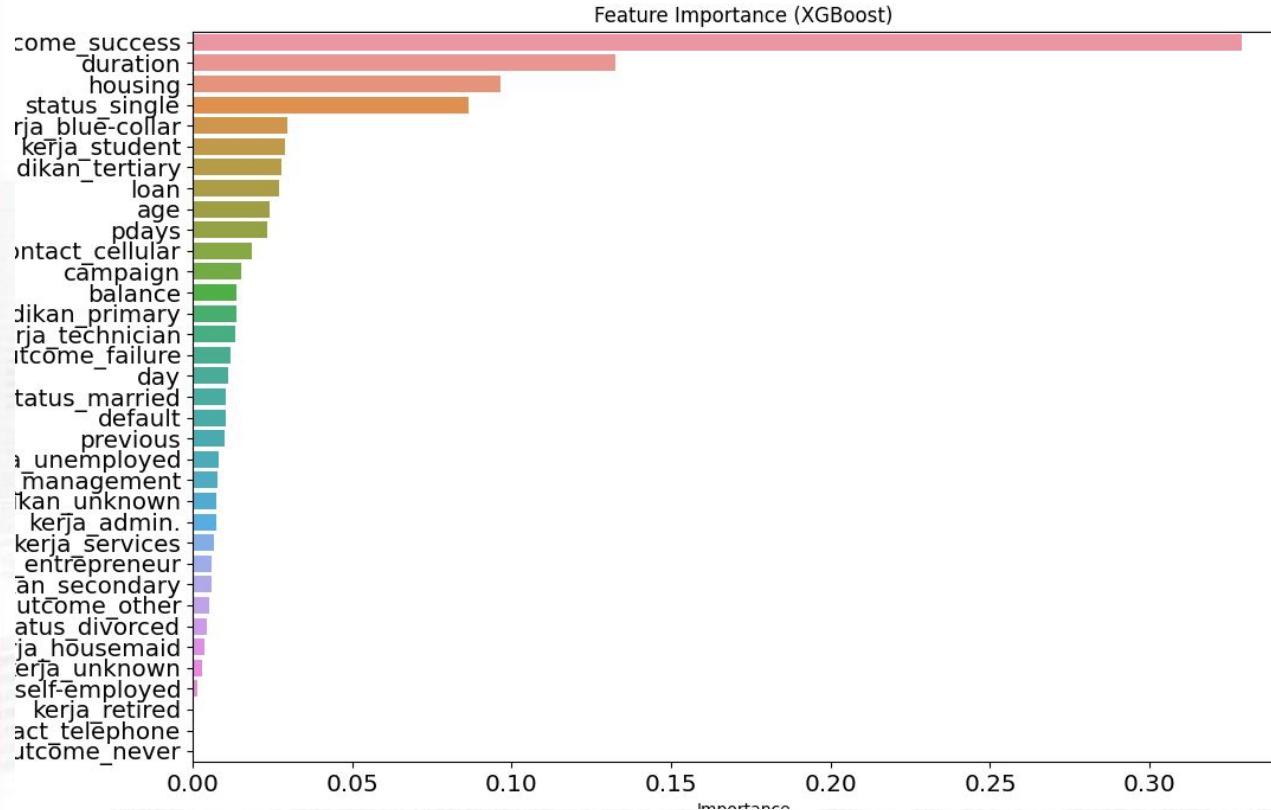
Meskipun presisi XGBoost sama dengan Logistic Regression dan AdaBoost, kinerja secara keseluruhan yang lebih baik pada metrik lainnya menjadikan XGBoost sebagai pilihan yang menarik.

AUC-Score Test yang Tinggi:

XGBoost memiliki nilai AUC-Score pada data uji yang cukup tinggi, menunjukkan bahwa model mampu memberikan prediksi yang baik bahkan pada data yang belum pernah dilihat sebelumnya.

Dengan mempertimbangkan aspek-aspek tersebut, XGBoost tetap menjadi pilihan yang baik dan konsisten untuk kasus ini. Meskipun performa model Random Forest cukup baik dalam beberapa metrik, XGBoost masih unggul dalam hal F1-Score dan Recall, yang relevan dalam skenario kampanye deposito untuk meningkatkan identifikasi pelanggan berpotensi.

Feature Importance XGBoost



Feature-feature yang memiliki importance kurang dari 0,030 akan dihapus, guna mengurangi feature 'noise' yang ada

Setelah itu dilakukan modelling menggunakan XGBoost dengan menggunakan feature-feature yang sudah dikurangi.

Final Model XGBoost

Test Accuracy: 0.90

Test Classification Report:

	precision	recall	f1-score	support
0	0.92	0.97	0.95	7925
1	0.63	0.34	0.44	1032
accuracy			0.90	8957
macro avg	0.78	0.66	0.69	8957
weighted avg	0.89	0.90	0.89	8957

Train AUC Score: 0.88

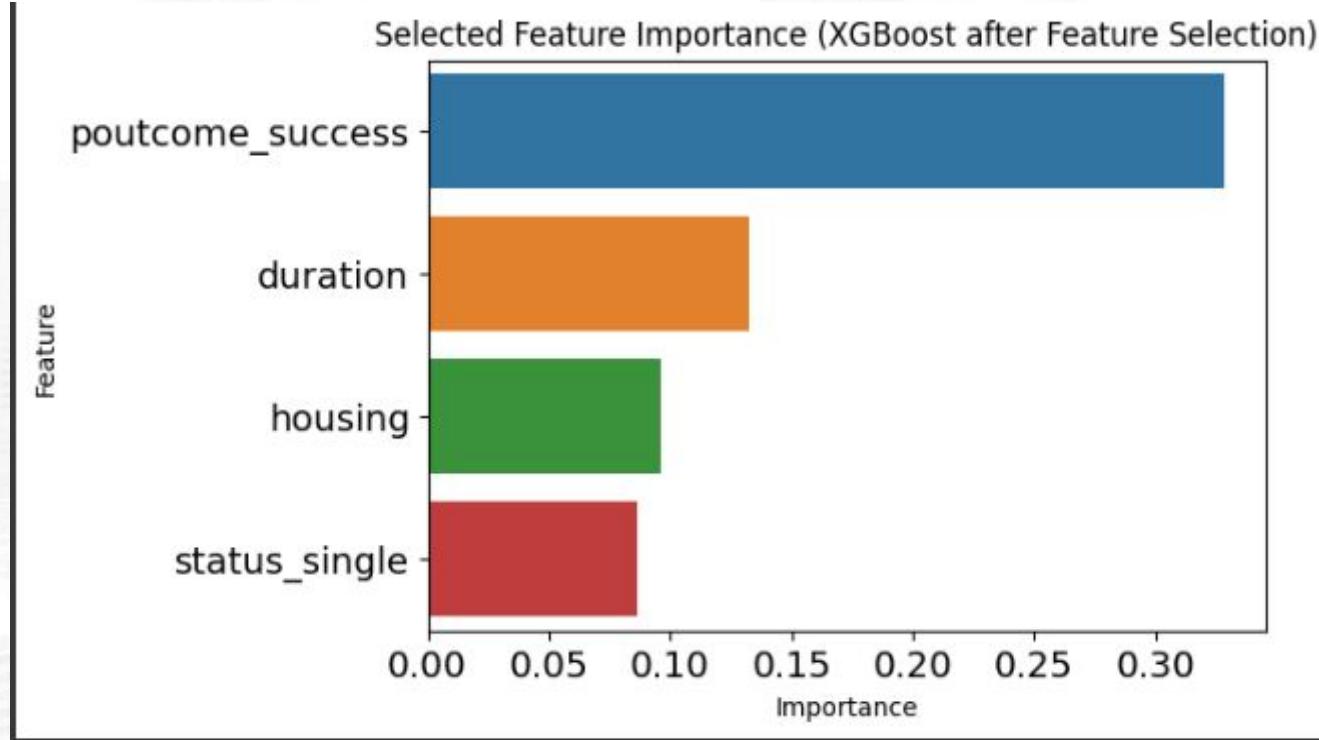
Test AUC Score: 0.86

Didapatkan model final menggunakan XGBoost dengan akurasi 90%

dengan macro average untuk precision sebesar 78%, recall 66%, f1-score 69%.

Dengan AUC-Score untuk data train sebesar 88%, dan data test sebesar 86%

Feature Importance XGBoost



Feature-feature yang memiliki pengaruh besar pada model yaitu,

Poutcome_success

Duration

Housing

status_single

Feature Importance

Review definisi feature:

poutcome_success: diambil dari feature encoding untuk feature "poutcome" yang berarti hasil/outcome dari campaign previous(p) / sebelumnya, yang mana hasilnya berupa success (bukan failure, unknown, other)

Duration: feature berkaitan tentang durasi campaign yang dilakukan pada waktu sebelumnya, dalam detik

Housing: feature yang merepresentasikan apakah pengguna memiliki pinjaman uang untuk rumah atau tidak (y/n)

Status_single : feature yang mempresentasikan pengguna yang statusnya belum menikah (y/n)

Business Insight

1. Poutcome_success:

Insight: Keberhasilan kampanye pemasaran sebelumnya memiliki dampak yang signifikan. Bisnis dapat fokus pada strategi yang **telah terbukti berhasil** dalam kampanye sebelumnya, memperkuat taktik yang menghasilkan tingkat keberhasilan yang tinggi.

2. Duration:

Insight: Durasi kontak memiliki dampak besar. Dengan mengoptimalkan durasi panggilan menjadi **minimal 4 menit**, bank dapat meningkatkan efektivitas kampanye pemasaran teleponnya, memastikan bahwa setiap interaksi memberikan nilai maksimal bagi pelanggan dan meningkatkan peluang konversi.

3. Housing:

Insight: Informasi tentang kepemilikan rumah mempengaruhi keputusan pelanggan. Ini bisa dikaitkan dengan kondisi keuangan pelanggan atau prioritas finansial mereka. Bisnis dapat menyusun strategi pemasaran khusus untuk kelompok **pelanggan yang tidak memiliki kredit rumah**, menawarkan produk atau layanan yang lebih sesuai dengan kebutuhan mereka.

4. status_single:

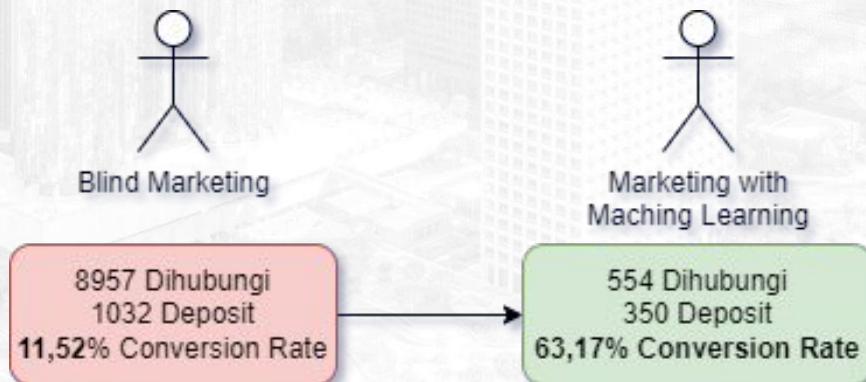
Insight: **Pelanggan dengan status sudah menikah atau divorce sangat berpotensi** untuk menerima campaign, pasangan yang menikah/divorce lebih fokus pada kestabilan finansial jangka panjang bagi keluarga mereka, termasuk tabungan untuk pendidikan anak, pensiun, atau pembelian properti, membuat produk seperti deposito berjangka menjadi pilihan yang menarik.

Simulasi

Dari Simulasi Pada Model, didapatkan hasil sebagai berikut.

Simulated on Data Test	Predicted Yes	Predicted No
Actual Yes	350	682
Actual No	204	7721

Kami merekomendasikan untuk **menghubungi yang terprediksi Yes saja**, dengan conversion rate sebesar 63,17%. **Goal tercapai dengan minimal meningkatkan sampai 15%**.



Rekomendasi Bisnis

1. Optimalkan Kampanye Pemasaran:

Identifikasi faktor-faktor **keberhasilan dari kampanye pemasaran sebelumnya** yang menghasilkan poutcome_success tinggi. Tingkatkan strategi pemasaran dengan mempertimbangkan pendekatan yang sama untuk meningkatkan kesuksesan kampanye mendatang.

2. Peningkatan Durasi Kontak:

Implementasikan **minimal durasi panggilan 4 menit** sebagai standar untuk semua interaksi dengan pelanggan dalam kampanye pemasaran telepon. Ini berdasarkan analisis data yang menunjukkan durasi interaksi yang lebih lama berkorelasi dengan peningkatan peluang konversi pelanggan.

3. Penargetan Kelompok Tanpa Kredit Rumah:

Fokus pada kelompok klien yang **tidak memiliki kredit rumah**. Pertimbangkan kondisi keuangan pelanggan dalam menyesuaikan penawaran. Pelanggan dengan pinjaman perumahan besar mungkin lebih berhati-hati dalam membuat komitmen finansial baru. Rancang kampanye khusus atau tawarkan insentif yang sesuai untuk menarik perhatian kelompok ini.

4. Pemasaran Berfokus Keluarga:

Gunakan materi pemasaran yang menyoroti bagaimana deposito berjangka dapat membantu dalam perencanaan masa depan keluarga, seperti pendidikan anak atau dana darurat keluarga. Tawarkan paket atau promosi khusus untuk **pasangan yang menikah**, mungkin dengan suku bunga yang lebih baik atau manfaat tambahan ketika membuka deposito bersama.

TERIMA KASIH

[Link Google Collab Stage 1](#)

[Link Google Collab Stage 2](#)

[Link Google Collab Stage 3](#)

[Link Google Collab Stage 4](#)