Banking Telesales Marketing Optimization

Rahmad Ramadhan Laska

Problem Statement

Deposito berjangka merupakan sumber pendapatan utama bagi bank. Di antara berbagai saluran yang digunakan untuk menjangkau nasabah, kampanye penjualan lewat telepon tetap menjadi salah satu yang paling efektif. Akan tetapi, biayanya tinggi. Dari referensi daring, pusat panggilan internal (4 tenaga penjualan) dapat menelan biaya sekitar 460.000 USD per tahun, sementara pusat panggilan yang dialihdayakan menelan biaya 25-65 USD per jam per tenaga penjualan untuk AS dan 8-18 USD per jam per tenaga penjualan untuk Amerika Selatan dan Asia Tenggara.

Dengan demikian, sangat penting untuk mengidentifikasi nasabah yang paling mungkin berkonversi sebelumnya sehingga mereka dapat ditargetkan secara khusus melalui panggilan. Klasifikasi ini akan membantu bank untuk mengoptimalkan kampanye penjualan lewat telepon: meningkatkan tingkat penerimaan dan/atau mengurangi biaya. Salah satu lembaga perbankan Portugis mempekerjakan Chang Corp untuk membantu proyek pengoptimalan penjualan lewat telepon mereka.

1. https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets
2. https://www.cloudtask.com/blog/how-much-does-it-cost-to-outsource-a-call-center

Goal

Buatlah model yang membantu mengklasifikasikan pelanggan berdasarkan kemungkinan mereka untuk menerima tawaran penjualan lewat telepon, sehingga kami dapat membantu klien untuk:

1. Meningkatkan tingkat penerimaan penjualan lewat telepon, dan/atau
2. Mengurangi biaya penjualan lewat telepon

Business Metrics

- Take up rate (Tingkat Penerimaan)
- Telesales cost (Biaya Penjualan Tele)

Objectives

1. Cari tahu karakteristik dan faktor lain dari nasabah yang akan tertarik dan mendaftar pada produk deposito berjangka yang ditawarkan
2. Bangun model pembelajaran mesin yang dapat memprediksi calon nasabah yang akan ditawarkan kampanye penjualan melalui telepon
3. Berikan rekomendasi kepada tim pemasaran untuk meningkatkan efektivitas kampanye bagi calon nasabah deposito berjangka

Stage 01

Link Colab

Feature List

1 - age (numeric)
2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown","secondary","primary","tertiary")
5 - default: has credit in default? (binary: "yes","no")
6 - balance: average yearly balance, in euros (numeric)
7 - housing: has housing loan? (binary: "yes","no")
8 - loan: has personal loan? (binary: "yes","no")
related with the last contact of the current campaign:
9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12 - duration: last contact duration, in seconds (numeric)
other attributes:
13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):
17 - y - has the client subscribed a term deposit? (binary: "yes","no")

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

Insight Descriptive Statistics (1/3)

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

Dari semua informasi di atas, terlihat bahwa data tersebut memiliki 17 feature/ kolom. Tipe data untuk masing-masing kolom sudah sesuai. Antara nama kolom dengan isinya juga sudah sesuai.

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

Dari 17 kolom tidak ada satupun yang memiliki nilai kosong, sehingga tidak perlu ada preprocessing untuk missing values.

Insight Descriptive Statistics (2/3)

C. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

Kolom Numerik

Age: Untuk kolom age tidak terlihat potensi skewed karena mean (40.9) dan mediannya (39.0) tidak berbeda signifikan.

Balance: Untuk kolom balance terlihat nilai mean (1362.3) sangat jauh di atas mediannya (448.0). Nilai minimum balance memiliki nilai aneh, yaitu nilai yang minus sebesar -8019. Perlu dilakukan investigasi lanjutan terkait balance negatif.

Day: Untuk kolom day terlihat mean (15.8) dan median (16.0) hampir sama. Kemudian nilai minimum, maksimum, kuartil 1 dan 3 terlihat normal.

Duration: Untuk kolom duration terlihat bahwa nilai mean (258.2) lebih besar dari nilai median (180.0), ada potensi positive skew.

Campaign: Untuk kolom campaign nampak adanya kecenderungan positive skew, di mana nilai mean (2.7) lebih besar dibanding nilai median (2.0).

Pdays: Untuk kolom pdays terlihat data berkumpul di angka -1 (36954 dari 45211 records), yang artinya hampir semua client tidak pernah di hubungi sebelumnya.

Previous: Untuk kolom previous terlihat kebanyakan client (36954 dari 45211 records) memiliki nilai 0, yang artinya belum pernah menerima campaign sebelumnya.

Insight Descriptive Statistics (3/3)

C. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

Kolom Kategori

Job : terdapat 12 pekerjaan berbeda, mayoritas 'blue-collar' (21.5%).

Marital : lebih dari 50% klien sudah menikah.

Education : lebih dari 50% secondary.

Housing : lebih dari 50% client memiliki pinjaman rumah.

Loan : mayoritas client tidak memiliki personal loan.

Contact : lebih dari 50% client menggunakan telepon seluler.

Month : paling banyak di bulan May saat melakukan campaign.

Poutcome : ~80% memiliki value 'unknown'

y : 88.3% client tidak membeli deposito. Kolom yang merupakan target ini memiliki class-imbalance.

Insight Univariate Analysis (1/4)

Numerical

Age :

1. slight skew positif = mean > modus/median
2. tidak ditemukan low outlier
3. ditemukan high outlier = ada yang nilainya jauh diatas normal

Day :

1. multimodal distribution
2. tidak ditemukan low outlier dan high outlier

Numerical

Balance :

1. skew positif = mean > modus/median
2. ditemukan low outlier = ada yang nilainya jauh dibawah normal
3. ditemukan high outlier = ada yang nilainya jauh diatas normal

Duration :

1. skew positif = mean > modus/median
2. tidak ditemukan low outlier = tidak ada yang nilainya jauh dibawah normal
3. ditemukan high outlier = ada yang nilainya jauh diatas normal

Insight Univariate Analysis (2/4)

Numerical

Campaign :

1. skew positif = mean > modus/median
2. tidak ditemukan low outlier = tidak ada yang nilainya jauh dibawah normal
3. ditemukan high outlier = ada yang nilainya jauh diatas normal

Pdays :

1. skew positif = mean > modus/median
2. tidak ditemukan low outlier = tidak ada yang nilainya jauh dibawah normal
3. ditemukan high outlier = ada yang nilainya jauh diatas normal

Numerical

Previous :

1. skew positif = mean > modus/median
2. tidak ditemukan low outlier = tidak ada yang nilainya jauh dibawah normal
3. ditemukan high outlier = ada yang nilainya jauh diatas normal

Insight Univariate Analysis (3/4)

Categorical

Job : Terdapat 12 kategori, kemungkinan terlalu banyak (perlu grouping). Top 3 nya ‘blue-collar’, ‘management’, dan ‘technician’ (masing-masing memiliki nilai di atas 7000 records). Terdapat ‘unknown’ job dengan jumlah records di bawah 500

Marital : ‘married’ menempati posisi tertinggi dengan jumlah records > 25,000, ‘single’ dan ‘divorced’ berada pada posisi kedua dan ketiga. Possibly pada tahap pre-processing dilakukan grouping menjadi married_flag 'yes' and 'no'.

Education : ‘secondary’ menempati posisi tertinggi dengan > 20,000 records, diikuti dengan ‘tertiary’, ‘primary’, dan ‘unknown’. ‘unknown’ < 2,500 records. Pada tahap preprocessing kemungkinan dilakukan label encoding.

Default : hampir semua records memiliki value ‘no’

Housing : posisi pertama ‘yes’ dengan jumlah record sekitar 25,000, sisanya ‘no’. distribusi keduanya cukup seimbang

Loan : ‘no’ menempati posisi pertama (jumlah record > 35,000), sisanya ‘yes’

Insight Univariate Analysis (4/4)

Categorical

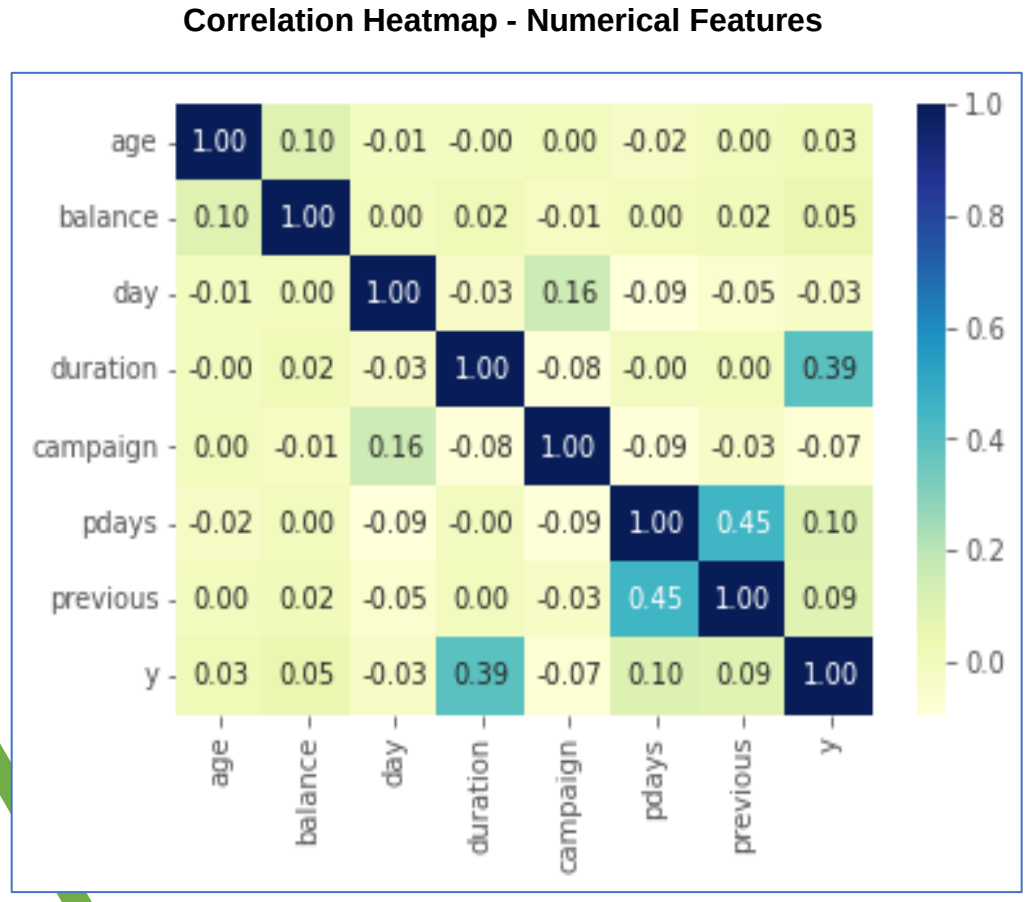
Contact : ‘cellular’ menempati posisi pertama (jumlah record > 25,000), siikuti ‘unknown’ dan ‘telephone’. Jumlah ‘unknown’ cukup banyak, sekitar 13,000 records, kemungkinan memerlukan penyesuaian pada tahap pre-processing.

Month : ‘may’ menempati posisi pertama (jumlah record > 12,000), diikuti ‘jul’ dan ‘aug’

Poutcome : ‘unknown’ menempati posisi tertinggi dengan jumlah record > 35,000. Jumlah ‘success’ < 2,500

y : ‘no’ berada di posisi tertinggi, dengan jumlah record sekitar 40,000. ‘yes’ sekitar 5,000 record. Kolom yang merupakan target ini memiliki class-imbalance, kemungkinan besar akan dilakukan oversampling pada tahap selanjutnya.

Insight Multivariate Analysis (1/2)



A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?

Korelasi antar feature numerik dengan target

- y sebagai variabel target / variabel dependent / label, y adalah variabel yang ingin diprediksi dari variabel bebas lainnya.
- 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).

Hubungan antar feature kategorikal dengan target

- Klien yang paling tertarik dengan produk deposito adalah klien dengan status menikah, pendidikan lanjutan 'secondary' dan 'tertiary', tidak memiliki pinjaman KPR, tidak memiliki hutang di bank (personal loan), dikontak menggunakan celullar/HP, tidak pernah gagal membayar hutang (default = 'no'), memiliki rata-rata balance positif

Dari penjabaran plot multivariat numerik dan kategorikal di atas, semua feature numerik dan kategorikal akan dipertahankan untuk modeling iterasi pertama.

Insight Multivariate Analysis (2/2)

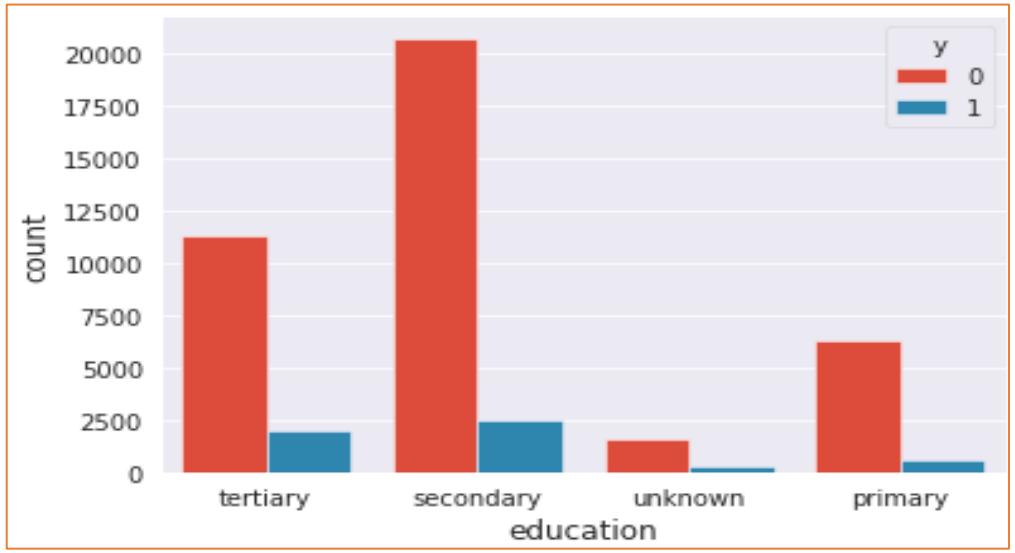
B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

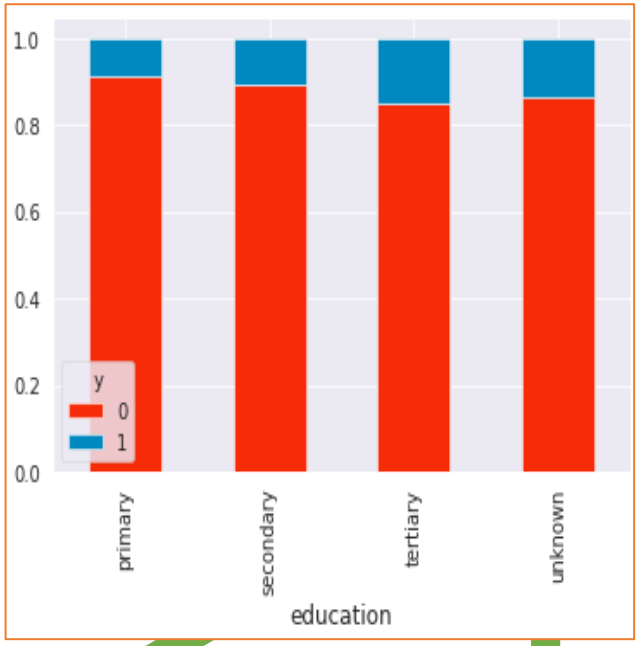
- Terdapat korelasi positif dari feature numerik yang cukup kuat (0.45), yaitu antara pdays dengan previous. Dari hasil univariate analysis sebelumnya, diketahui pada sample data mayoritas client belum mendapat campaign, sehingga ber-value -1 pada pdays dan 0 pada previous. Kemungkinan besar hal ini menyebabkan tingginya nilai korelasi antara kedua feature tersebut
- Dari heatmap korelasi feature numerik tidak ada feature pasangan dengan nilai korelasi >0.7, kemungkinan besar tidak ada data yang redundant.

Business Insights (1/3)

Dari hasil EDA yang dilakukan, terdapat beberapa quick insights yang bisa dipertimbangkan dalam upaya meningkatkan performa campaign term-deposit oleh telesales:

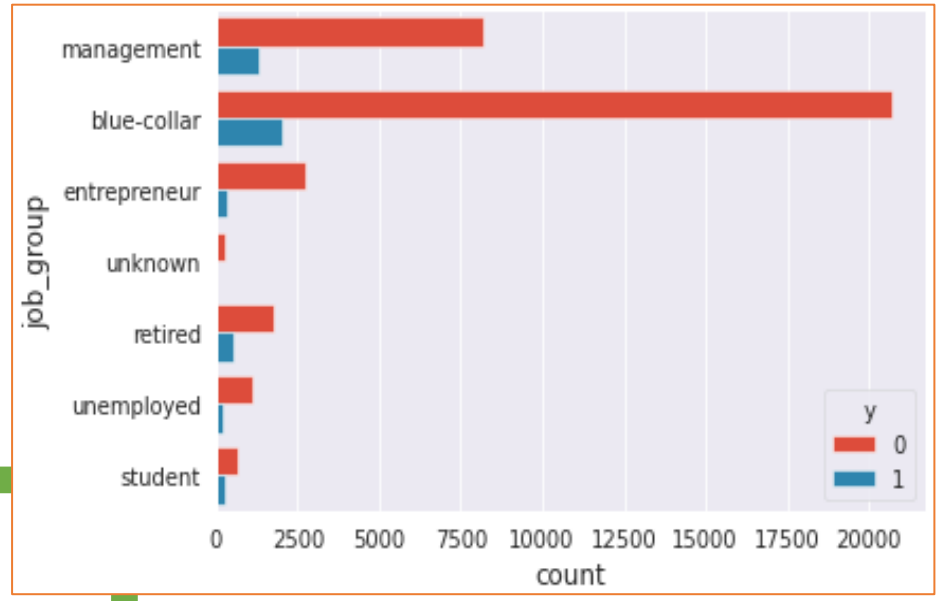
1. Secara umum, client yang tertarik dengan produk deposito adalah klien dengan karakteristik
   a. usia produktif: 20-60 b. menikah
   c. pendidikan lanjutan 'secondary' dan 'tertiary'
   d. tidak memiliki cicilan rumah
   e. tidak memiliki hutang di bank
   Tim Marketing dapat memusatkan effort terhadap orang-orang dengan karakteristik seperti di atas.
2. Pendidikan. Dari grafik di samping dapat dilihat bahwa mayoritas client yang sign-up untuk term-deposit memiliki pendidikan lanjutan (secondary dan tertiary). Jika dilihat stacked bar 100% (proporsi yang mengambil term-deposit untuk tiap jenjang pendidikan), % yes meningkat seiring dengan bertambah tingginya pendidikan. Tim marketing mungkin bisa memberikan approach yang berbeda kepada client dengan pendidikan lebih rendah, yaitu dengan menjelaskan manfaat term-deposit menggunakan bahasa yang mudah dipahami.



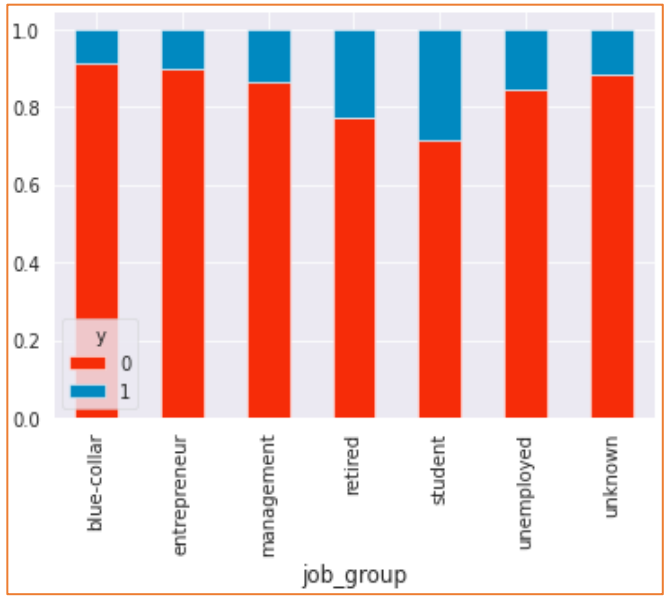


Business Insights (2/3)

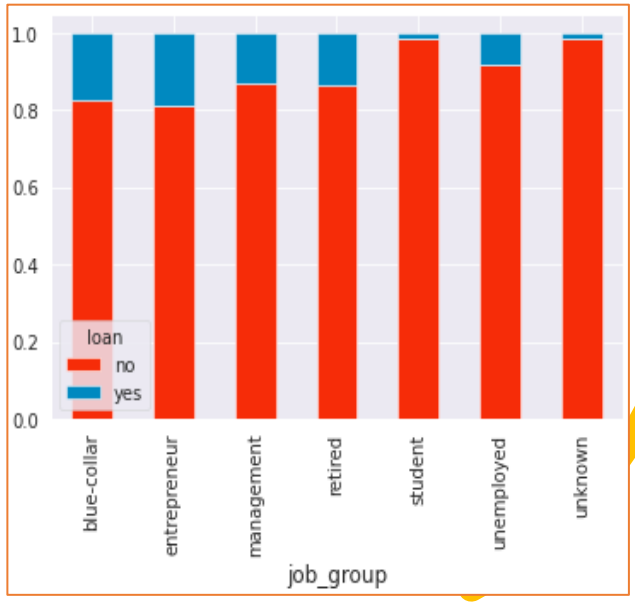
3. Jenis pekerjaan.
   a. Setelah dilakukan pengelompokan jenis pekerjaan, dari plot di bawah dapat dilihat client yang mengambil term-deposit mayoritas memiliki pekerjaan: blue-collar, management. Namun demikian, jika dilihat % yes dari masing-masing pekerjaan, % yes blue-collar relatif rendah. Jika sample menggambarkan populasi, maka kemungkinan banyak client bank yang memiliki pekerjaan blue-collar. Tim marketing perlu re-assess apakah product term-deposit dan approach yang dilakukan sudah tepat untuk segment ini.
   b. Terlihat % yes untuk student dan retired lebih tinggi dibanding job group lain. Pengamatan ini dapat dijadikan basis agar tim marketing dapat merancang product dan marketing approach yang lebih menarik lagi untuk student dan retired. Diharapkan dengan adanya product yang lebih tailored untuk 2 segment ini, akan ada akuisisi baru dan kedua segment bisa bertambah besar
   c. % yes untuk entrepreneur pun relatif rendah. Segment ini kemungkinan besar memerlukan modal usaha. Nampak dari visualisasi 3C bahwa hampir 20% entrepreneur memiliki personal loan (relatif lebih tinggi dibanding segment lain). Sebaiknya tim marketing mempertimbangkan apakah lebih menguntungkan dan menarik bila segment ini difokuskan untuk ditawarkan loan



3A



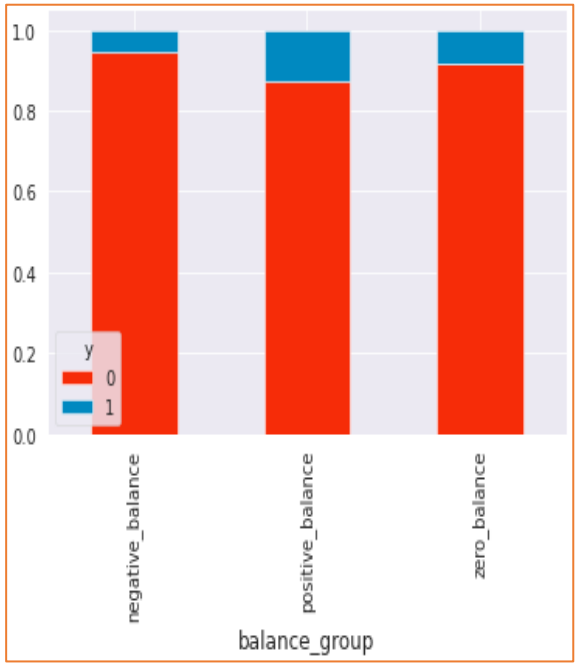
3B



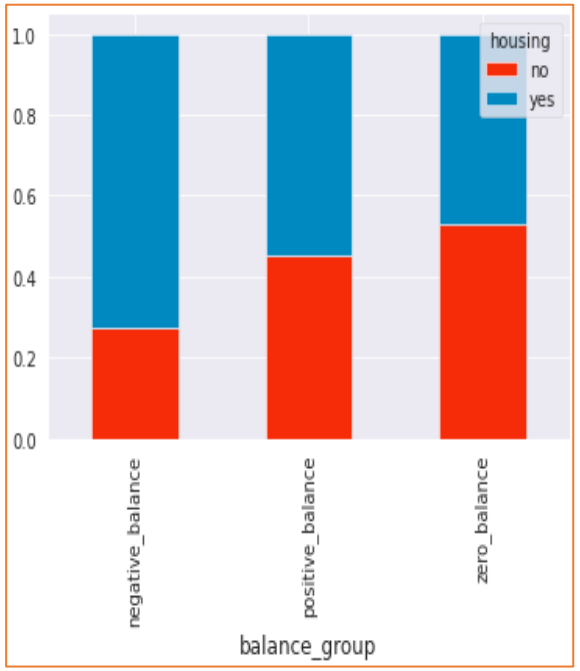
3C

Business Insights (3/3)

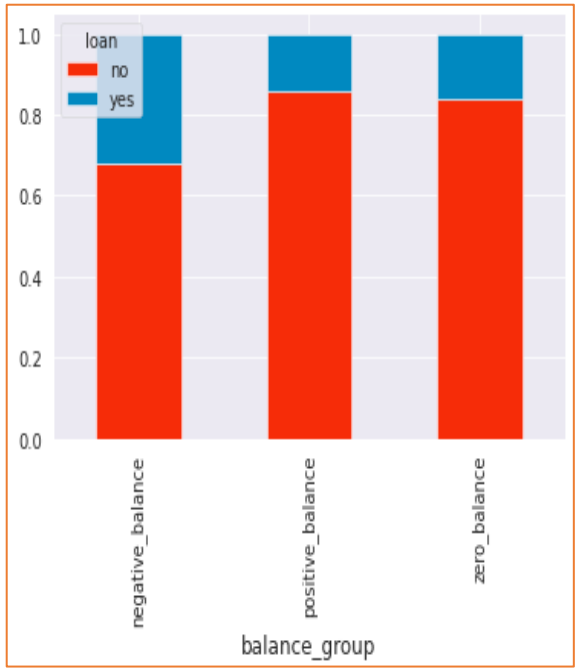
4. Average yearly balance. Untuk performa campaign yang lebih baik, sebaiknya orang-orang dengan average yearly balance negative diexclude dari whitelist telemarketing campaign. Selain % yes take up term-depositnya paling rendah dibanding 2 segment lain (visualisasi 4A), sekitar 70% dari orang2 ini memiliki housing loan dan 30% memiliki personal loan (vis 4B dan 4C). Adanya loan mengindikasikan rendah atau tidak adanya 'cold money' yang bisa digunakan untuk investasi (seperti pada term-deposit).
   Sementara itu, client dengan average yearly balance = 0 mengindikasikan bahwa akun mereka lebih bersifat transaksional. Client hanya akan menambah saldo ketika perlu melakukan pembayaran atau transfer. Client seperti ini mungkin memiliki 'cold money' di insititusi lain. Jika kita bisa memperoleh data biro (seperti SLIK) yang mengcompile data asset dan liability client di insititusi lain, kita dapat membuat program untuk menarik minat mereka agar berinvestasi pada bank kita.



4A



4B



4C

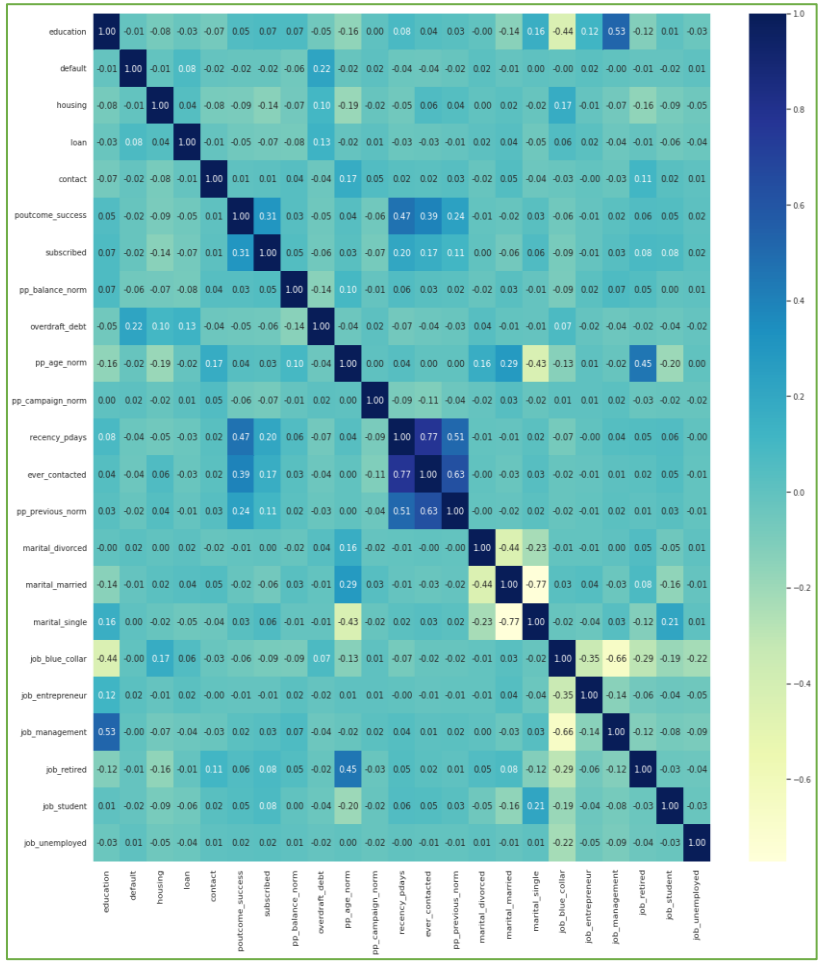
Stage 02

Link Colab

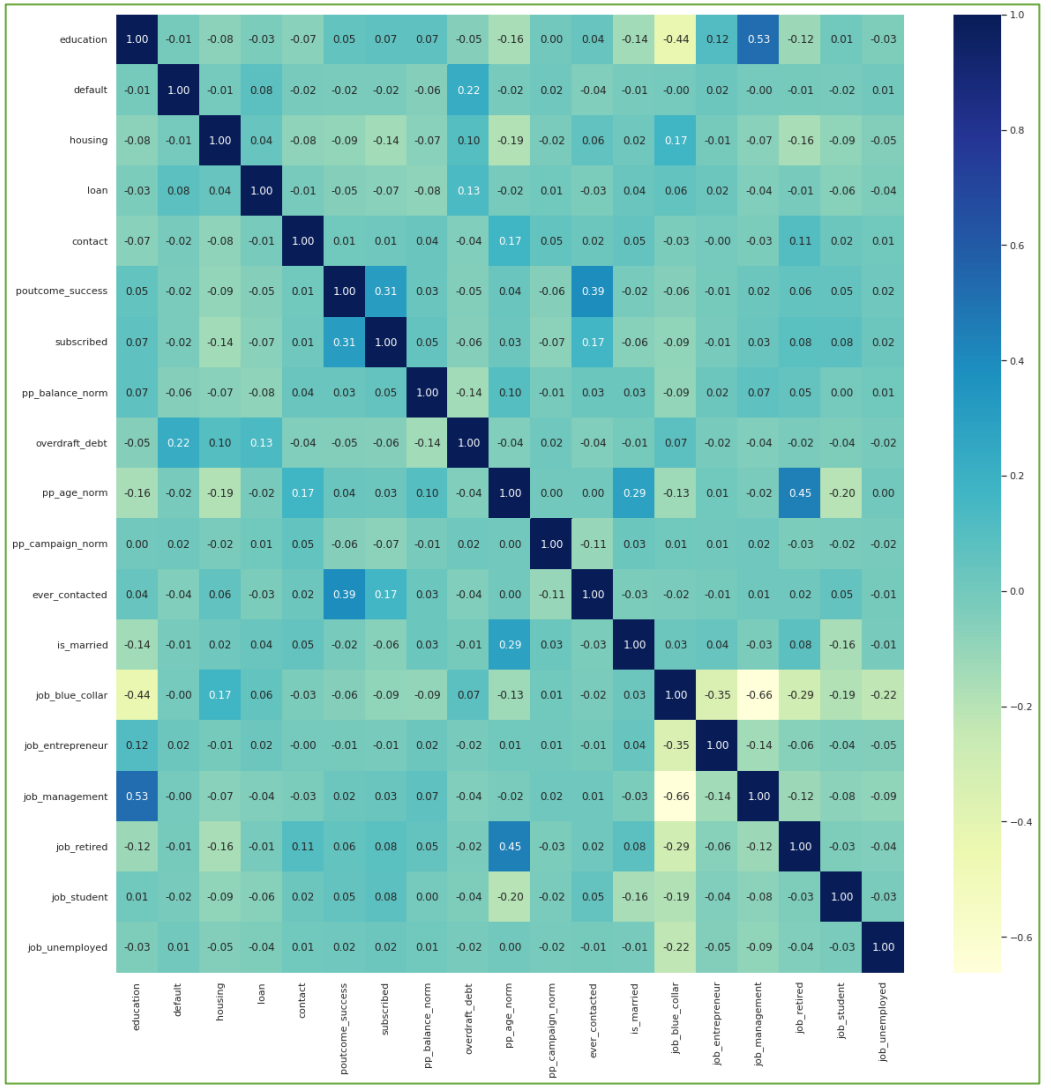
Summary Preprocessing (1A - 2B)

Feature	Type	Preprocessing
age	numerical	No outlier handling (natural, no sudden jump), Feature transformation - MinMax Scaler
job	categorical	Handle missing value (using mode from education), Feature extraction (create job group to reduce unique values), Feature encoding - OHE Job Group (non ordinal category)
marital	categorical	Feature encoding - 'is_married' (0 1)
education	categorical	Handle missing value (using mode from job), Feature encoding - Label encoding (ordinal category)
default	categorical	Feature encoding (0 1)
balance	numerical	No outlier handling, hard code negative values as 0, Feature transformation - MinMax Scaler, Feature extraction - negative balance as 'overdraft_debt'
housing	categorical	Feature encoding (0 1)
loan	categorical	Feature encoding (0 1)
contact	categorical	Handle missing value (using mode), Feature encoding (0 1)
day	numerical	Feature selection - Drop Feature (attribute of campaign call, will be available after the call is made)
month	categorical	Feature selection - Drop Feature (attribute of campaign call, will be available after the call is made)
duration	numerical	Feature selection - Drop Feature (attribute of campaign call, will be available after the call is made)
campaign	numerical	No outlier handling (natural, no sudden jump), Feature transformation - MinMax Scaler
pdays	numerical	Feature selection - Drop Feature (due to high correlation with other features)
previous	numerical	Outlier handling (hard code 1 record), MinMax Scaler, later dropped due to high correlation with other features. Feature extraction - 'ever_contacted' (0 1) (used for modeling)
poutcome	categorical	Feature encoding - 'poutcome_success' (0 1)
y	categorical	Feature encoding (0 1), handle class-imbalance

Summary Preprocessing (1A - 2B)



Feature Selection, dropping features with high correlation (related to previous contacts), label encoding marital status



Summary Preprocessing (1A - 2B)

df_train_fin

```
df_train_fin.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 19 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   education          45211 non-null  int64
 1   default            45211 non-null  int64
 2   housing            45211 non-null  int64
 3   loan               45211 non-null  int64
 4   contact            45211 non-null  int64
 5   poutcome_success   45211 non-null  int64
 6   subscribed         45211 non-null  int64
 7   pp_balance_norm    45211 non-null  float64
 8   overdraft_debt     45211 non-null  int64
 9   pp_age_norm        45211 non-null  float64
 10  pp_campaign_norm   45211 non-null  float64
 11  ever_contacted     45211 non-null  int64
 12  is_married         45211 non-null  int64
 13  job_blue_collar    45211 non-null  uint8
 14  job_entrepreneur   45211 non-null  uint8
 15  job_management     45211 non-null  uint8
 16  job_retired        45211 non-null  uint8
 17  job_student        45211 non-null  uint8
 18  job_unemployed     45211 non-null  uint8
dtypes: float64(3), int64(10), uint8(6)
memory usage: 4.7 MB
```

Handling Class-Imbalance

```
Original
False    39922
True      5289
dtype: int64


UNDERSAMPLING
False    10578
True      5289
dtype: int64


OVERSAMPLING
False    39922
True     19961
dtype: int64


SMOTE
False    39922
True     19961
dtype: int64
```

2C. Feature Engineering - Tambahan Fitur

- **Jumlah hari sejak tanggal transaksi terakhir** —> sebagai proxy untuk ‘recency’, orang yang baru melakukan transaksi kemungkinan lebih responsif dibandingkan orang yang sudah lama tidak mengakses akunnya. Pada dataset ada pdays, yang menunjukan recency contact customer, namun demikian sebagian besar customer belum pernah dikontak. Oleh karena itu, mungkin feature ini lebih berguna untuk mewakili recency
- **Jumlah tanggungan (manusia) yang dimiliki** —> bisa menjadi indikator apakah orang tersebut kemungkinan memiliki uang untuk diinvestasikan
- **Rata-rata frekuensi transaksi bulanan** —> untuk menggambarkan seberapa sering (frequency) client melakukan interaksi dengan akun. Jika interaksinya sering, bisa jadi client tersebut nyaman dengan bank kita, dan akan lebih mudah menawarkan produk deposito
- **Riwayat term-deposit client pada bank kita** —> jika pernah melakukan deposito, dilihat recency, frequency dan monetary nya. Semakin baik skor-skor tersebut, makin mungkin client berminat untuk menambah/ menaruh uang ulang pada product term deposito
- **Jumlah asset yang dimiliki di insititusi finansial lain (misalnya menggunakan data biro SLIK)** —> bisa mengetahui kekayaan (potensi memiliki uang untuk diinvestasikan), apakah akun pada bank kita hanya untuk transaksi (misal di bank lain saldonya tinggi, tapi di kita kecil), apakah punya deposito di bank lain (familiaritas dengan produk deposito, lebih mudah untuk ditawari)
- **Jumlah liabilitas yang dimiliki di insititusi finansial lain (misalnya menggunakan data biro SLIK)** —> orang-orang yang memiliki banyak hutang di tempat lain kemungkinan tidak tertarik membeli produk deposito karena harus melunasi hutangnya

Stage 03

Link Colab

Model Evaluation and Hyperparameter Tuning

Decision Tree

Algoritma ini digunakan karena robust terhadap data yang memiliki outlier

Hasil Model Evaluation

```
Accuracy (Test Set): 0.80
Precision (Test Set): 0.70
Recall (Test Set): 0.72
F1-Score (Test Set): 0.71
roc_auc (test-proba): 0.78
roc_auc (train-proba): 1.00
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.79
Precision (Test Set): 0.75
Recall (Test Set): 0.56
F1-Score (Test Set): 0.64
roc_auc (test-proba): 0.83
roc_auc (train-proba): 0.87
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall menurun sebesar 16% tetapi tidak terjadi overfitting.

Random Forest

Algoritma ini digunakan karena robust, prediksi model lebih akurat dapat mengurangi variance / mencegah overfitting

Hasil Model Evaluation

```
Accuracy (Test Set): 0.85
Precision (Test Set): 0.80
Recall (Test Set): 0.71
F1-Score (Test Set): 0.76
roc_auc (test-proba): 0.90
roc_auc (train-proba): 1.00
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.80
Precision (Test Set): 0.82
Recall (Test Set): 0.52
F1-Score (Test Set): 0.64
roc_auc (test-proba): 0.86
roc_auc (train-proba): 0.89
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall menurun sebesar 19% tetapi tidak terjadi overfitting
- Hyperparameter tuning dilakukan untuk mengurangi overfitting: mengurangi max_depth, memperbesar min_samples_split dan min_samples_leaf. Overfitting berkurang, tetapi recall juga berkurang :(

Logistic Regression

Algoritma ini digunakan karena pada dataset yang digunakan tidak ada multikolinieritas dan algoritma ini memiliki kemampuan komputasi yang cepat.

Hasil Model Evaluation

```
Accuracy (Test Set): 0.73
Precision (Test Set): 0.75
Recall (Test Set): 0.30
F1-Score (Test Set): 0.43
roc_auc (test-proba): 0.73
roc_auc (train-proba): 0.73
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.73
Precision (Test Set): 0.75
Recall (Test Set): 0.30
F1-Score (Test Set): 0.42
roc_auc (test-proba): 0.72
roc_auc (train-proba): 0.72
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall dan precision tetap sama.

K-Nearest Neighbor

Algoritma ini digunakan karena memiliki cara kerja yang sederhana selain itu algoritma ini cocok digunakan untuk data yang bersifat non-linear.

Hasil Model Evaluation

```
Accuracy (Test Set): 0.78
Precision (Test Set): 0.67
Recall (Test Set): 0.67
F1-Score (Test Set): 0.67
roc_auc (test-proba): 0.84
roc_auc (train-proba): 0.93
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.81
Precision (Test Set): 0.71
Recall (Test Set): 0.72
F1-Score (Test Set): 0.71
roc_auc (test-proba): 0.86
roc_auc (train-proba): 1.00
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall meningkat sebesar 5% tetapi terjadi overfitting.

Naïve Bayes

Algoritma ini digunakan karena memiliki waktu komputasi yang cepat.

Hasil Model Evaluation

```
Accuracy (Test Set): 0.71
Precision (Test Set): 0.62
Recall (Test Set): 0.35
F1-Score (Test Set): 0.45
roc_auc (test-proba): 0.71
roc_auc (train-proba): 0.71
```

Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif

XGBoost

Algoritma ini digunakan karena robust, prediksi model lebih akurat dapat mengurangi bias/ mencegah underfitting.

Hasil Model Evaluation

```
Accuracy (Test Set): 0.85
Precision (Test Set): 0.89
Recall (Test Set): 0.64
F1-Score (Test Set): 0.74
roc_auc (test-proba): 0.89
roc_auc (train-proba): 0.92
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.88
Precision (Test Set): 0.90
Recall (Test Set): 0.72
F1-Score (Test Set): 0.80
roc_auc (test-proba): 0.91
roc_auc (train-proba): 0.96
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall meningkat sebesar 8% dan tidak terjadi overfitting.

AdaBoost

Algoritma ini digunakan karena robust, prediksi model lebih akurat dapat mengurangi bias/ mencegah underfitting

Hasil Model Evaluation

```
Accuracy (Test Set): 0.80
Precision (Test Set): 0.85
Recall (Test Set): 0.48
F1-Score (Test Set): 0.61
roc_auc (test-proba): 0.81
roc_auc (train-proba): 0.81
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.77
Precision (Test Set): 0.87
Recall (Test Set): 0.37
F1-Score (Test Set): 0.52
roc_auc (test-proba): 0.79
roc_auc (train-proba): 0.80
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall menurun sebesar 11% tetapi tidak terjadi overfitting.

Stage 03 with PCA

Link Colab

Principal Component Analysis (PCA)

```
Explained variance ratio: [0.31780823 0.1421783  0.12892124 0.10167672 0.09534452 0.06287705
 0.05423647]
Variansi Komponen: [0.31780823 0.45998652 0.58890776 0.69058448 0.785929   0.84880605
 0.90304252]
Total explained variance: 0.9030425176300236
```

Berdasarkan gambar tersebut, diketahui bahwa tujuh komponen utama yang dipilih dari hasil PCA mampu menjelaskan total varians sebesar **90,3%** dari data asli yang memiliki **15 fitur**. Artinya, dengan hanya menggunakan 7 komponen dari 15 fitur awal, sebagian besar informasi penting dalam data berhasil dipertahankan. Masing-masing komponen memberikan kontribusi varians yang menurun secara bertahap, di mana komponen pertama menyumbang sekitar 31,78%, dan komponen ketujuh menyumbang sekitar 5,42%. Meskipun tidak seluruhnya 100%, nilai total explained variance yang tinggi ini menunjukkan bahwa reduksi dimensi yang dilakukan cukup efisien untuk menghilangkan hampir setengah fitur asli namun tetap menjaga lebih dari 90% informasi. Pendekatan ini sangat bermanfaat untuk menyederhanakan kompleksitas data, mengurangi noise, serta meningkatkan efisiensi proses pemodelan tanpa kehilangan struktur penting dari data.

Model Evaluation and Hyperparameter Tuning

Decision Tree

Algoritma ini digunakan karena robust terhadap data yang memiliki outlier

Hasil Model Evaluation

```
Accuracy (Test Set): 0.78
Precision (Test Set): 0.67
Recall (Test Set): 0.65
F1-Score (Test Set): 0.66
roc_auc (test-proba): 0.75
roc_auc (train-proba): 1.00
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.75
Precision (Test Set): 0.67
Recall (Test Set): 0.47
F1-Score (Test Set): 0.55
roc_auc (test-proba): 0.77
roc_auc (train-proba): 0.83
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall menurun sebesar 18% tetapi tidak terjadi overfitting.

Random Forest

Algoritma ini digunakan karena robust, prediksi model lebih akurat dapat mengurangi variance / mencegah overfitting

Hasil Model Evaluation

```
Accuracy (Test Set): 0.80
Precision (Test Set): 0.72
Recall (Test Set): 0.66
F1-Score (Test Set): 0.69
roc_auc (test-proba): 0.86
roc_auc (train-proba): 1.00
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.76
Precision (Test Set): 0.72
Recall (Test Set): 0.47
F1-Score (Test Set): 0.57
roc_auc (test-proba): 0.82
roc_auc (train-proba): 0.87
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall menurun sebesar 19% tetapi tidak terjadi overfitting
- Hyperparameter tuning dilakukan untuk mengurangi overfitting: mengurangi max_depth, memperbesar min_samples_split dan min_samples_leaf. Overfitting berkurang, tetapi recall juga berkurang :(

Logistic Regression

Algoritma ini digunakan karena pada dataset yang digunakan tidak ada multikolinieritas dan algoritma ini memiliki kemampuan komputasi yang cepat.

Hasil Model Evaluation

```
Accuracy (Test Set): 0.72
Precision (Test Set): 0.72
Recall (Test Set): 0.26
F1-Score (Test Set): 0.39
roc_auc (test-proba): 0.70
roc_auc (train-proba): 0.70
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.72
Precision (Test Set): 0.72
Recall (Test Set): 0.26
F1-Score (Test Set): 0.38
roc_auc (test-proba): 0.70
roc_auc (train-proba): 0.70
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall dan precision tetap sama

K-Nearest Neighbor

Algoritma ini digunakan karena memiliki cara kerja yang sederhana selain itu algoritma ini cocok digunakan untuk data yang bersifat non-linear.

Hasil Model Evaluation

```
Accuracy (Test Set): 0.78
Precision (Test Set): 0.71
Recall (Test Set): 0.59
F1-Score (Test Set): 0.64
roc_auc (test-proba): 0.82
roc_auc (train-proba): 0.93
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.80
Precision (Test Set): 0.72
Recall (Test Set): 0.65
F1-Score (Test Set): 0.68
roc_auc (test-proba): 0.83
roc_auc (train-proba): 1.00
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall meningkat sebesar 5% tetapi terjadi overfitting.

Naïve Bayes

Algoritma ini digunakan karena memiliki waktu komputasi yang cepat.

Hasil Model Evaluation

```
Accuracy (Test Set): 0.69
Precision (Test Set): 0.55
Recall (Test Set): 0.35
F1-Score (Test Set): 0.43
roc_auc (test-proba): 0.69
roc_auc (train-proba): 0.69
```

Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif

XGBoost

Algoritma ini digunakan karena robust, prediksi model lebih akurat dapat mengurangi bias/ mencegah underfitting.

Hasil Model Evaluation

```
Accuracy (Test Set): 0.75
Precision (Test Set): 0.72
Recall (Test Set): 0.42
F1-Score (Test Set): 0.53
roc_auc (test-proba): 0.79
roc_auc (train-proba): 0.83
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.76
Precision (Test Set): 0.69
Recall (Test Set): 0.52
F1-Score (Test Set): 0.59
roc_auc (test-proba): 0.81
roc_auc (train-proba): 0.87
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall meningkat sebesar 8% dan tidak terjadi overfitting.

AdaBoost

Algoritma ini digunakan karena robust, prediksi model lebih akurat dapat mengurangi bias/ mencegah underfitting

Hasil Model Evaluation

```
Accuracy (Test Set): 0.73
Precision (Test Set): 0.74
Recall (Test Set): 0.27
F1-Score (Test Set): 0.39
roc_auc (test-proba): 0.71
roc_auc (train-proba): 0.71
```

Setelah Hyperparameter tuning

```
Accuracy (Test Set): 0.72
Precision (Test Set): 0.83
Recall (Test Set): 0.19
F1-Score (Test Set): 0.31
roc_auc (test-proba): 0.71
roc_auc (train-proba): 0.71
```

- Metrics model : Recall. Karena tujuan dari modeling ini kita ingin mendapatkan rasio prediksi data yang positif dari keseluruhan data yang benar positif
- Setelah dilakukan hyperparameter tuning nilai recall menurun sebesar 8% tetapi tidak terjadi overfitting.

Perbandingan Hasil Akhir Model Evaluation

Tanpa PCA

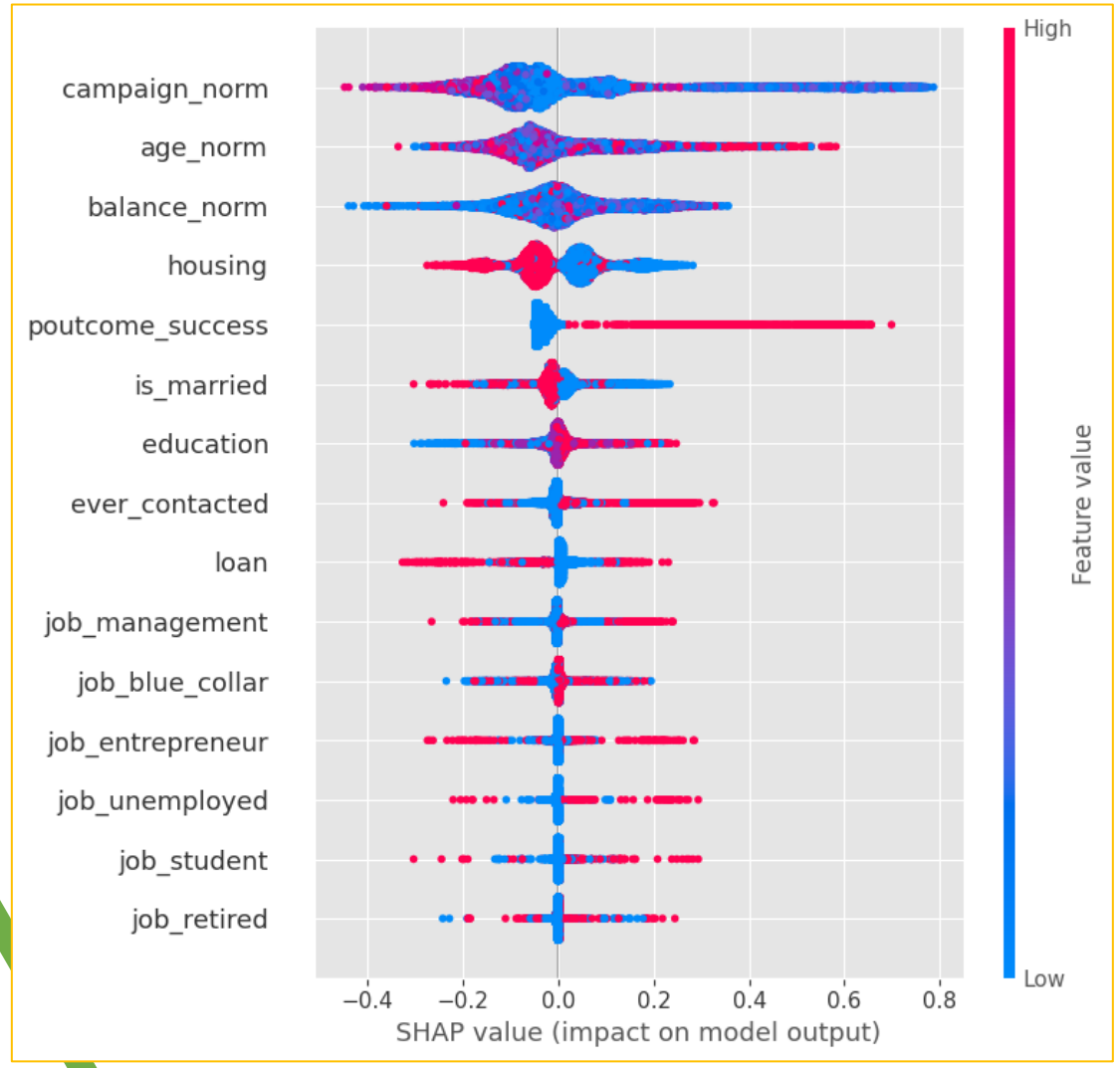
Algoritma	Hyperparameter Tuning	Score Recall
XGBoost	Yes	72%
K-Nearest Neighbor	Yes	72%
Decision Tree	No	72%
Random Forest	No	71%
Adaboost	No	48%
Naive Bayes	No	35%
Logistic Regression	No	30%

Dengan PCA

Algoritma	Hyperparameter Tuning	Score Recall
Random Forest	No	66%
K-Nearest Neighbor	Yes	65%
Decision Tree	No	65%
XGBoost	Yes	52%
Logistic Regression	No	36%
Naive Bayes	No	35%
Adaboost	No	27%

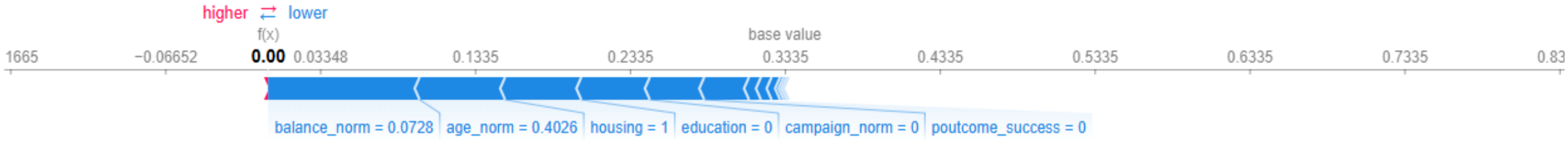
Akan dilakukan evaluasi lebih lanjut untuk XGBoost, Decision Tree dan Random Forest versi tanpa PCA

Interpretasi Model - Decision Tree (SHAP) (1/2)



1. campaign_norm semakin kecil, semakin mungkin subscribe
2. age_norm berpengaruh, namun titik merah berkumpul di dua sudut
3. Balance_norm memiliki pengaruh terhadap model, namun titik merah tersebar cukup merata. Bisa dilihat di ujung kiri semuanya biru, indikasi balance rendah maka tidak subscribe
4. Housing, titik merah terpusat di sisi kiri. Jika memiliki housing loan, tidak subscribe
5. Poutcome_success, titik merah terpusat di kanan. Jika pernah take up campaign program sebelumnya, kemungkinan akan take up lagi
6. Is_married, titik merah berpusat di kiri. Relasi negatif dengan target. Orang yang single lebih mungkin subscribe
7. Education, titik merah berpusat di kanan. Orang dengan tingkat edukasi lebih tinggi, lebih mungkin subscribe
8. Fitur-fitur lain tidak menunjukan sebaran yang konklusif

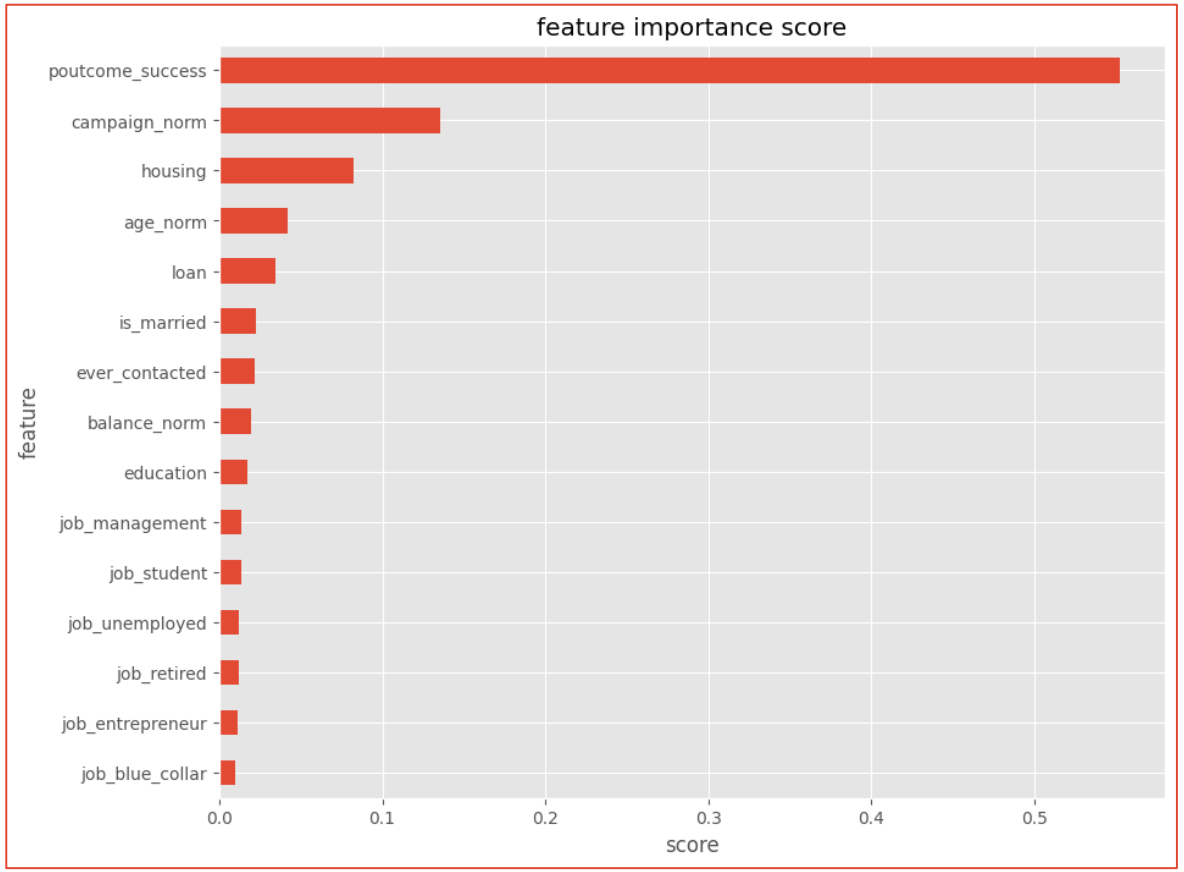
Interpretasi Model - Decision Tree (SHAP) (2/2)



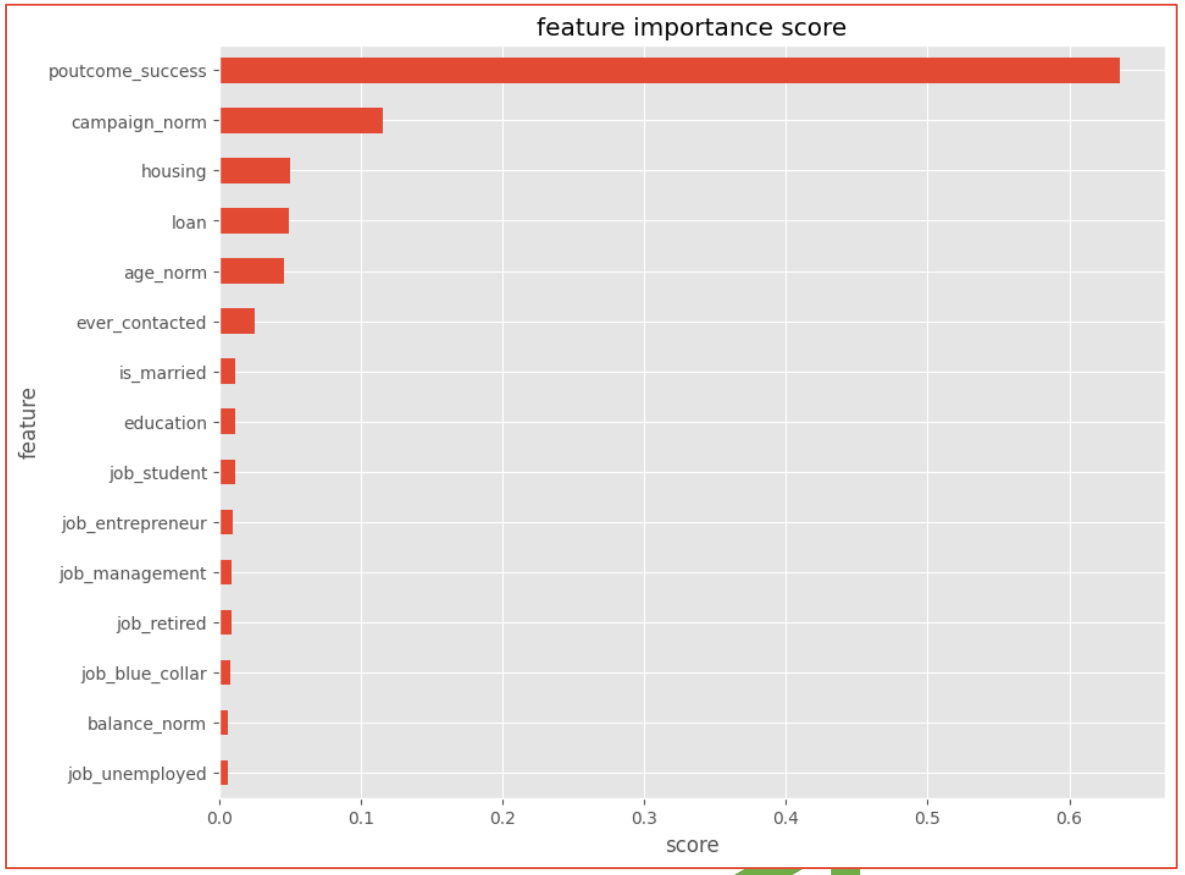
1. age_norm mempengaruhi target secara positif (warna merah). Jika dilihat summary_plot, sebenarnya titik merah berkumpul di kedua ujung, sejalan dengan hasil EDA sebelumnya di mana student dan retired memiliki kecenderungan lebih tinggi untuk subscribe.
2. Campaign_norm. Semakin tinggi jumlah campaign yang diberikan, semakin tidak mungkin orang tersebut subscribe. Aplikasi bisnis: agent tidak perlu menghabiskan effort untuk menghubungi orang berulang kali dalam 1 campaign, kemungkinan mereka memang tidak tertarik.
3. balance_norm mempengaruhi target secara negatif. Namun demikian interpretasi fitur ini sebaiknya tidak terlalu dipertimbangkan, karena warna merah tersebar (tidak mengumpul di salah satu sisi). Hal ini mungkin terjadi akibat banyaknya nilai balance yang sangat tinggi
4. housing (1 untuk memiliki house loan, 0 sebaliknya), housing = 1 mempengaruhi target secara negatif. Hal ini menunjukkan orang yang memiliki house loan cenderung tidak subscribe, kemungkinan karena mereka lebih fokus menggunakan uang untuk melunasi house loan
5. poutcome_success 0 mempengaruhi target secara negatif. Hal ini menunjukkan bahwa orang yang pernah mengambil tawaran sebelumnya, memiliki kemungkinan lebih besar untuk subscribe. Aplikasi bisnis: keep track terhadap taker dari program-program sebelumnya, karena mereka mungkin akan take up lagi. Namun demikian fitur ini sebaiknya digunakan dengan precaution, mengingat value 1 (yes) hanya cover 3% dari data sample

Interpretasi Model - XGBoost (Feature Importance) (1 / 2)

Before Hyperparameter Tuning

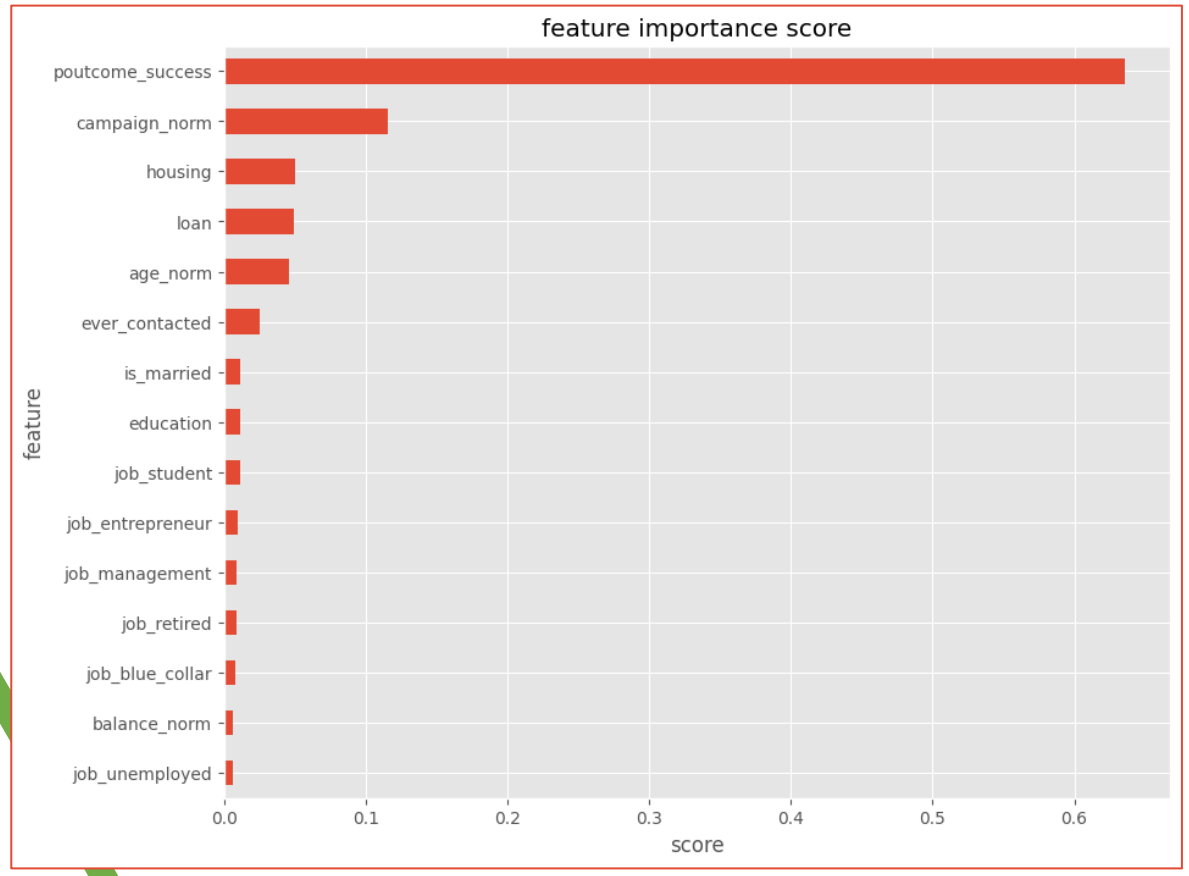


After Hyperparameter Tuning



Interpretasi Model - XGBoost (Feature Importance) (1 / 2)

After Hyperparameter Tuning



Setelah Hyperparameter Tuning, XGBoost memberikan nilai recall yang lebih tinggi (dari 0.64 ke 0.72). Model hasil tuning memiliki slight overfitting.

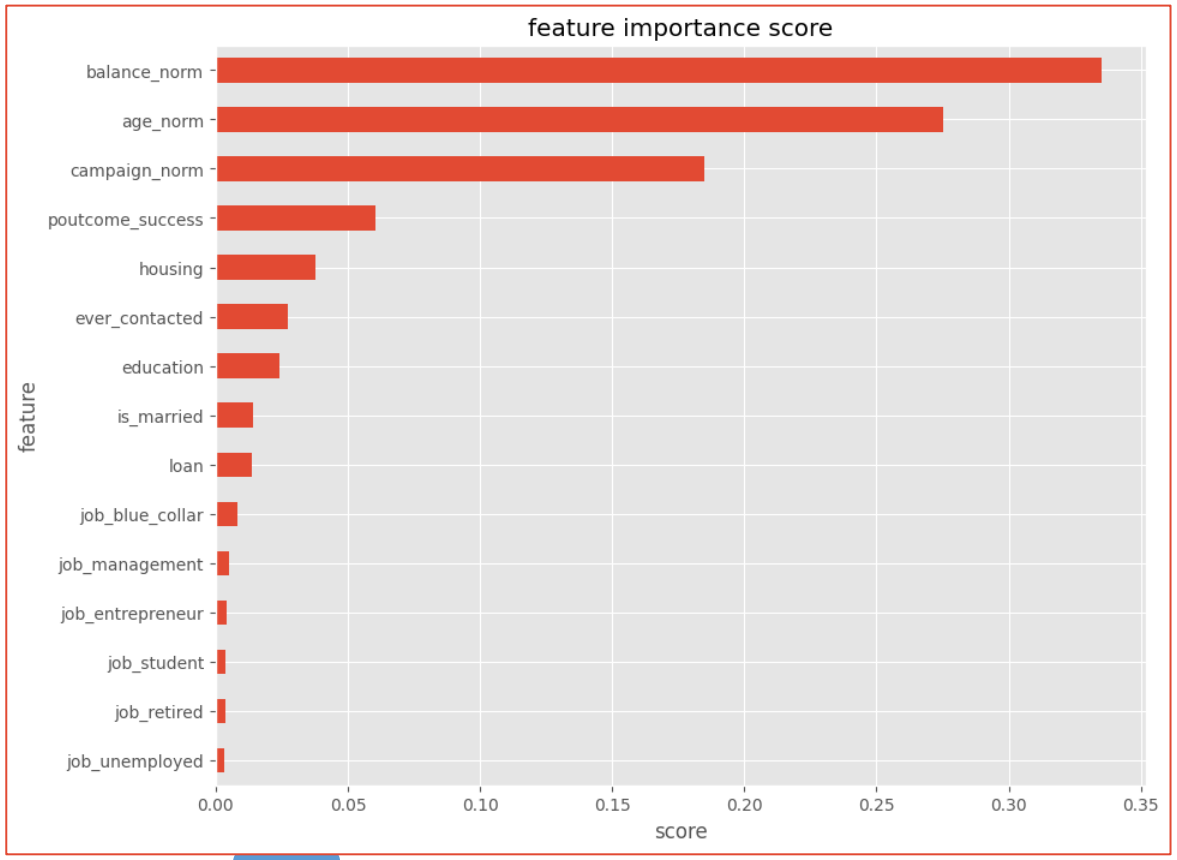
Jika kita lihat chart Feature Importance XGBoost before dan after Hyperparameter Tuning:

1. Sebelum tuning terdapat beberapa fitur yang memiliki pengaruh terhadap model
2. Setelah tuning, poutcome_success menjadi fitur yang paling dominan

Dari hasil pengamatan ini, XGBoost dengan Hyperparameter tuning mungkin memberikan nilai recall paling baik dibanding model lainnya. Namun model ini bertumpu hanya pada 1 fitur, poutcome_success, yang sebelum dilakukan oversampling, value ‘yes’ nya hanya mengcover ~3% data sample.

Interpretasi Model – Random Forest (Feature Importance)

Before Hyperparameter Tuning



Model Random Forest yang dipilih adalah yang belum mengalami Hyperparameter Tuning. Karena setelah dilakukan tuning, overfitting berkurang, namun nilai recall juga turun.

Jika kita lihat chart Feature Importance Random Forest before Hyperparameter Tuning, top 5 fitur dengan pengaruh paling besar: balance_norm, age_norm, campaign_norm, poutcome_success, housing.

Karena sama-sama tree classifier, interpretasi interaksi masing-masing fitur dengan target dapat mengikuti interpretasi pada decision tree. Top 5 fitur yang berperan penting terhadap target bersifat masuk akal.

Kesimpulan

Berdasarkan hasil evaluasi, model XGBoost dengan hyperparameter tuning menunjukkan performa terbaik, terutama dilihat dari metrik recall sebesar 0.72, yang menjadi fokus utama. Nilai recall ini menunjukkan kemampuan model dalam mendeteksi sebagian besar kasus positif, yang sangat penting dalam konteks di mana false negative harus diminimalkan. Selain itu, model ini juga memiliki akurasi 0.88, precision 0.90, dan F1-score 0.80, yang menandakan performa yang seimbang dan andal.

Model ini juga unggul dalam kemampuan diskriminasi, dengan nilai ROC AUC sebesar 0.91 untuk data uji dan 0.96 untuk data latih. Dari analisis feature importance, diketahui bahwa poutcome_success merupakan fitur paling berpengaruh, diikuti oleh campaign_norm, housing, dan loan. Dengan recall tinggi dan kinerja keseluruhan yang solid, XGBoost menjadi pilihan model paling optimal dalam studi ini.

Thank you

Rahmad Ramadhan Laska
+62 82269181935
ramadhanlaska11@gmail.com
https://www.linkedin.com/in/rmdlaska11