BANK MARKETING CAMPAIGN PREDICTION

Presented by Derielle Aisyah | JCDS 2602

TABLE OF CONTENT

BUSINESS UNDERSTANDING

LATAR BELAKANG

- Bank X memiliki produk *Term Deposit* yang dipasarkan kepada nasabah-nasabahnya
- Tingkat keberhasilan dari kampanye pemasaran produk ini cukup rendah namun terdapat biaya yang dikeluarkan untuk pelaksanaannya
- Untuk meningkatkan efisiensi dan efektivitas kampanye, perlu dilakukan analisis karakteristik nasabah yang berpotensi menerima penawaran produk

BUSINESS UNDERSTANDING

BUSINESS PROBLEM

- Bagaimana bank X dapat memprediksi nasabah yang kemungkinan besar akan berlangganan produk deposito berjangka setelah kampanye pemasaran dilakukan?

GOAL

- Meningkatkan tingkat konversi nasabah pada kampanye pemasaran dengan menargetkan nasabah yang lebih potensial berdasarkan data
- Membuat model machine learning yang dapat melakukan prediksi seorang nasabah akan berlangganan produk deposito berjangka atau tidak
- Memberikan informasi kepada tim marketing bank X, untuk mengoptimalkan strategi dan efisiensi kampanye pemasaran produk

BUSINESS UNDERSTANDING

MACHINE LEARNING PREDICTION

- Target column : `deposit`
  - `yes` = nasabah berlangganan produk deposito
  - `no` = nasabah tidak berlangganan produk deposito

METRICS EVALUATION

- False Negative (FN) : Model memprediksi tidak tertarik aslinya tertarik
- False Positive (FP) : Model memprediksi tertarik aslinya tidak tertarik
- Metrics Evaluation:
  - Precision penting untuk **menghindari FP** (menargetkan nasabah yang salah).
  - Recall penting untuk **menghindari FN (**tidak melewatkan nasabah potensial).
  - Oleh karena itu, **F1 Score** dipilih karena menyeimbangkan Precision dan Recall.

DATA UNDERSTANDING

Dataset berisi informasi dari suatu bank yang memasarkan produk *Term Deposit* atau deposito berjangka. Dataset dapat digunakan untuk memprediksi bagaimana nasabah dari bank akan menerima penawaran marketing produk deposito berjangka.

DATA UNDERSTANDING

Deskripsi kolom

Nama Kolom	Tipe Data	Deskripsi
age	int64	Usia nasabah
job	object	Jenis pekerjaan nasabah
balance	int64	Saldo rata-rata tahunan dalam rekening
housing	object	Status pinjaman perumahan (yes/no)
loan	object	Status pinjaman pribadi (yes/no)
contact	object	Jenis kontak komunikasi terakhir (misalnya: cellular, telephone)
month	object	Bulan saat terakhir kali dihubungi
campaign	int64	Jumlah kontak yang dilakukan selama kampanye pemasaran
pdays	int64	Hari sejak nasabah terakhir kali dihubungi (999 berarti belum pernah)
poutcome	object	Hasil dari kampanye pemasaran sebelumnya (misalnya: success, failure)
deposit	object	Apakah nasabah berlangganan deposito berjangka? (yes/no)

DATA UNDERSTANDING

Proporsi `deposit`

deposit	
no	0.522335
yes	0.477665

tidak termasuk dalam data yang imbalance

DATA CLEANSING

Missing Value

Pada dataset tidak terdapat missing value jadi tidak dilakukan perlakuan apapun

Duplicate Data

Terdapat 8 data duplikat, data dipertahankan karena data tersebut mencerminkan interaksi dengan pelanggan yang sah dan dianggap bagian dari perilaku pelanggan sebenarnya

EXPLORATORY DATA ANALYSIS

- Analisis distribusi numerik
- Deteksi Outlier
- Analisis Korelasi

EDA

Value pdays berdasarkan poutcome

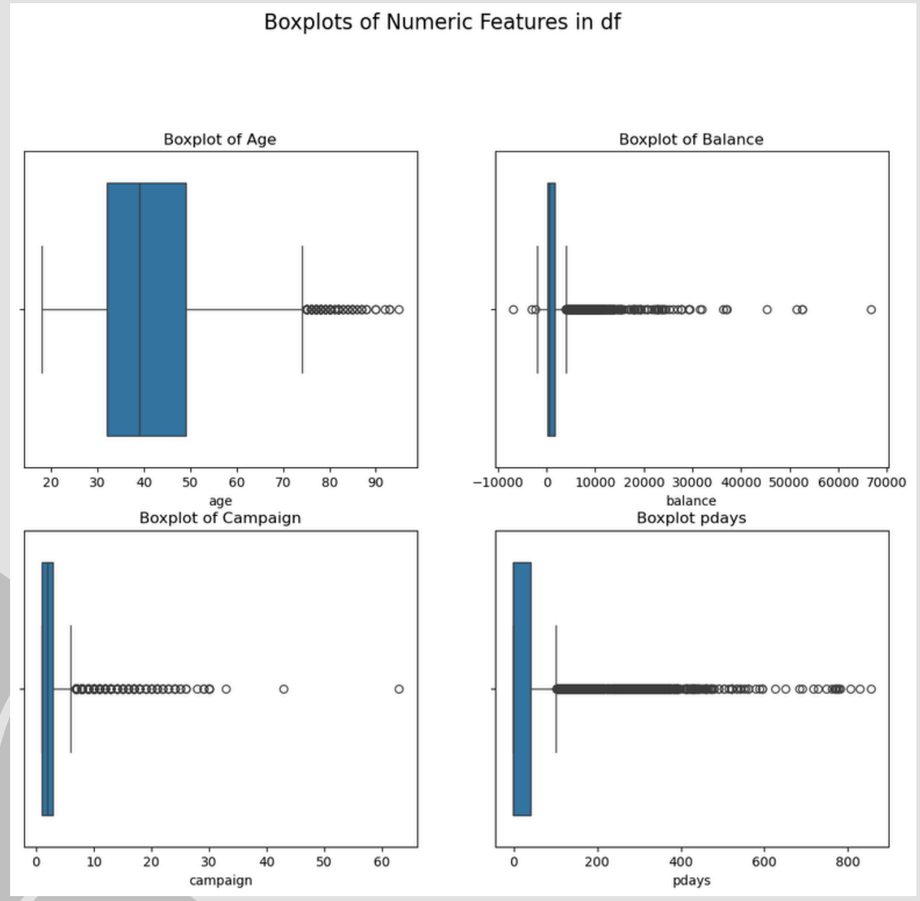
```
poutcome  pdays
failure    92        14
           181       14
           91        13
           342       13
           182       12
                    ...
success    651        1
           771        1
unknown   -1       5817
           98         1
           188        1
Name: count, Length: 742,
```

```
poutcome  pdays
failure    92        14
           181       14
           91        13
           342       13
           182       12
                    ...
success    472        1
           541        1
           651        1
           771        1
unknown   -1       5817
Name: count, Length: 740, dtype: int64
```

- Terdapat nasabah dengan `pdays` dengan value positif dan `poutcome` value = unknown.
- Ada kesalahan pencatatan nilai pada `poutcome` = unknown diganti dengan other mengartikan nasabah sudah dihubungi namun tidak ada respon.

EDA

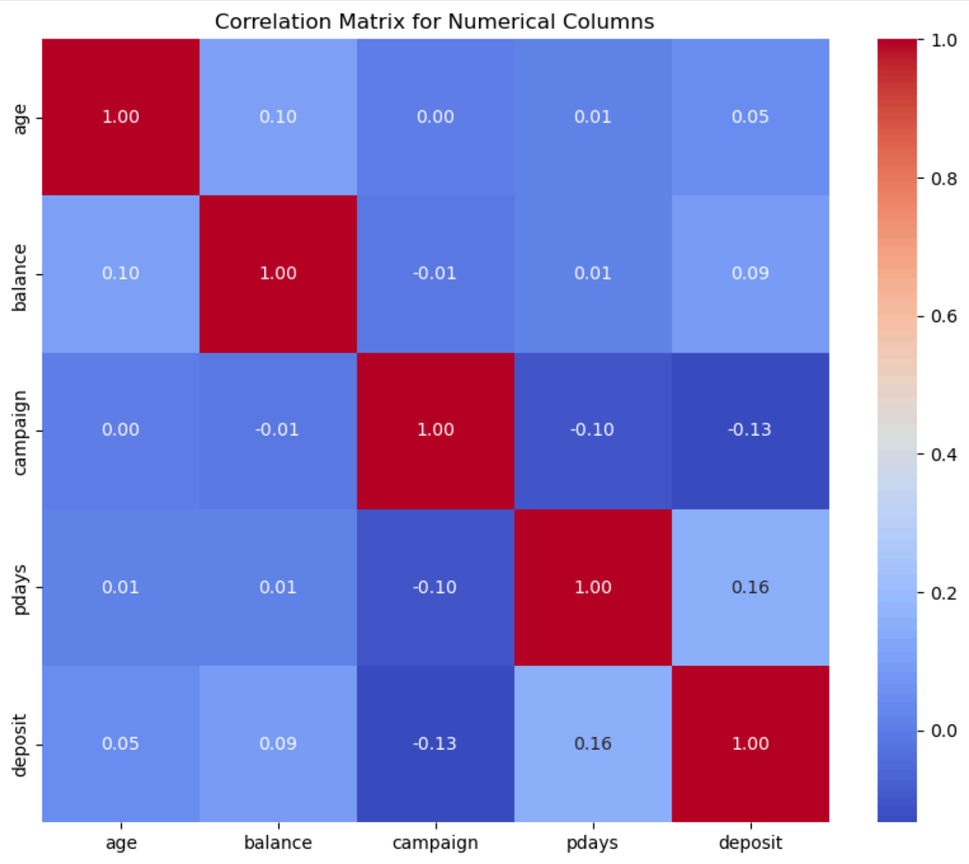
Deteksi Outlier



- Semua fitur numerik terdapat outlier, terutama untuk `balance`, `campaign`, dan `pdays`
- Namun Outlier dipertahankan karena data outlier masih masuk akal untuk digunakan dan terdapat banyak variasi dari valuenya

EDA

Analisa Korelasi : Heatmap



Korelasi antar variabel numerik terhadap `deposit`

- `pdays` : korelasi terkuat
- `campaign`: korelasi negatif
- `balance`: korelasi lemah tapi logis
- `age`: korelasi sangat lemah

DATA PREPROCESSING

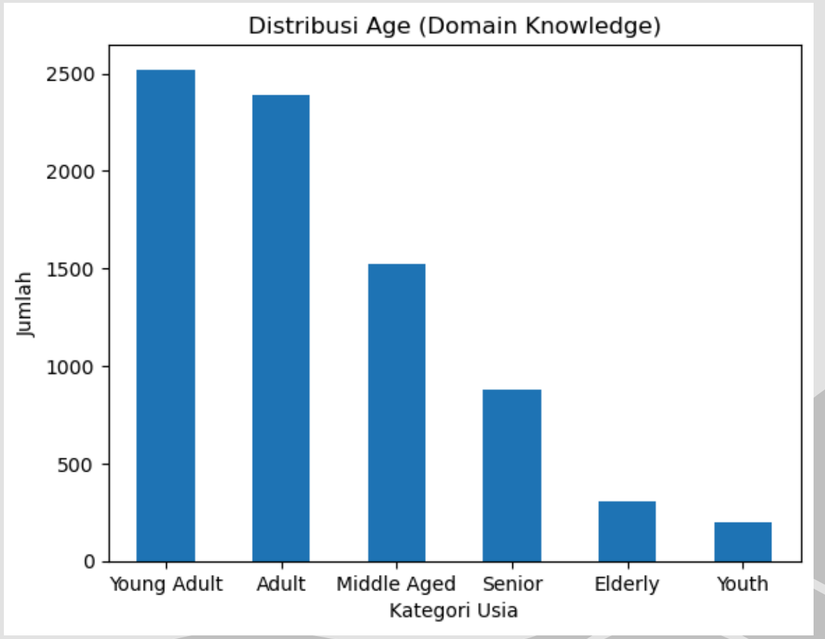
- Binning `age`
- Define X dan y
- Train Test Split
- Scaling dan Encoding

DATA PREPROCESSING

Binning ; pengelompokan data yang mengubah data numerik menjadi data kategorik dengan membaginya ke dalam beberapa interval

Binning `age` → membagi usia berdasarkan fase kehidupan atau kelompok demografis yang bisa mencerminkan perilaku keuangan

Kelompok Usia	Rentang Usia
Teen / Young Adult	< 25
Young Adult	25 - 34
Adult	35 - 44
Middle Aged	45 - 54
Senior	55 - 64
Elderly	≥ 65



DATA PREPROCESSING

Define X dan y

X adalah matriks nilai fitur, setiap kolom merupakan satu fitur. Setiap kolom X merupakan variabel independen. y adalah vektor nilai target, yang merupakan nilai yang ingin diprediksi

X = job, balance, housing, loan, contact, month, campaign, pdays, poutcomee, age_group
y = deposit

Train Test Split

Dataset dibagi menjadi 80% data train dan 20% data test.

- Train: untuk melatih model
- Test: untuk mengevaluasi performa model

DATA PREPROCESSING

Scaling dan Encoding

Scaling adalah mengubah skala data pada data numerik.

Scaler yg digunakan adalah RobustScaler() pada fitur balance, campaign, dan pdays.
RobustScaler digunakan untuk menskalakan data dengan menggunakan median dan rentang interkuartil (IQR), sehingga tahan terhadap outlier.

DATA PREPROCESSING

Scaling dan Encoding

Encoding adalah mengubah data kategorikal menjadi data numerik. Encoder yang digunakan adalah Onehot Encoding dan Ordinal Encoding.

Onehot Encoding digunakan pada kolom nominal yang tidak memiliki urutan, di mana setiap kategori direpresentasikan oleh variabel biner. Variabel biner mengambil nilai 1 jika kategori tersebut ada dan 0 jika tidak.
cat_onehot : job, housing, loan, contact, month, poutcome

Ordinal Encoding digunakan pada kolom yang memiliki urutan hierarki yang jelas. Penggunaan ordinal encoding mempertahankan informasi urutan dari value dalam kolom.

cat_ordinal : age_group

MODELING

MODELING BENCHMARK

Best Model Benchmark

Model	mean_score	std_score
GradientBoostingClassifier	0.674	0.018
StackingClassifier	0.672	0.015
RandomForestClassifier	0.670	0.016
XGBClassifier	0.667	0.019

mean_score adalah rata-rata skor cross validation dari setiap model sedangkan std_score adalah nilai standar deviasi dimana nilai rendah menunjukan variasi yang lebih konsisten

MODELING + RESAMPLING

Best Model Resampling

Model	Resampler	mean	std
Random Forest	NearMiss	0.674	0.021
Random Forest	RandomUnder	0.663	0.015
XGB	RandomUnder	0.661	0.014
Stacking	RandomOver	0.657	0.018

Model terbaik dengan resampling adalah Random Forest dengan resampler NearMiss yang menghasilkan nilai rata-rata cross validation sebesar 67%

HYPERPARAMETER TUNING

Model RandomForestClassifier, Resampler NearMiss

Before hyperparameter tuning
F1 Score = 0.670

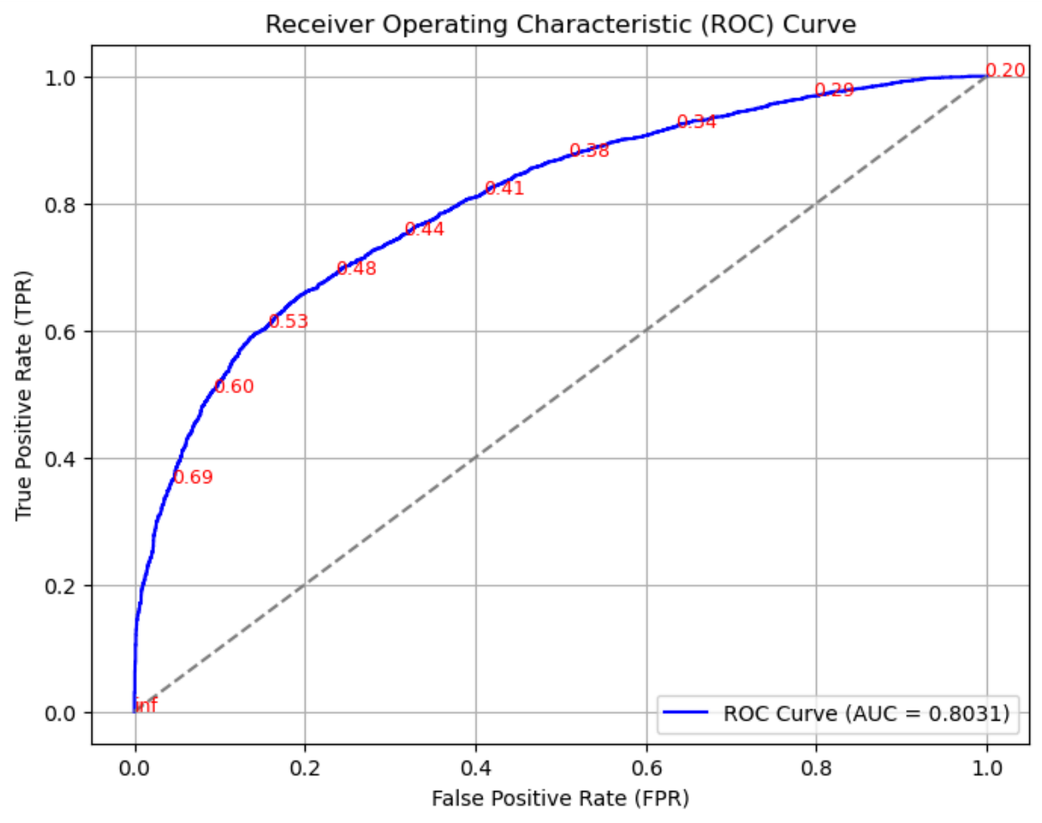
After hyperparameter tuning
F! Score = 0.679

Predict terhadap train
F1 Score = 0.70

Hasil model setelah tuning memiliki nilai train 70% dan test setelah tuning sebesar 67%. Selisih performa sudah cukup baik tidak menunjukkan overfitting di mana model belajar terlalu bagus dari data training tapi kurang generalisasi ke data test walaupun hasil dari test masih kurang baik.

ROC CURVE & PR CURVE

ROC Curve

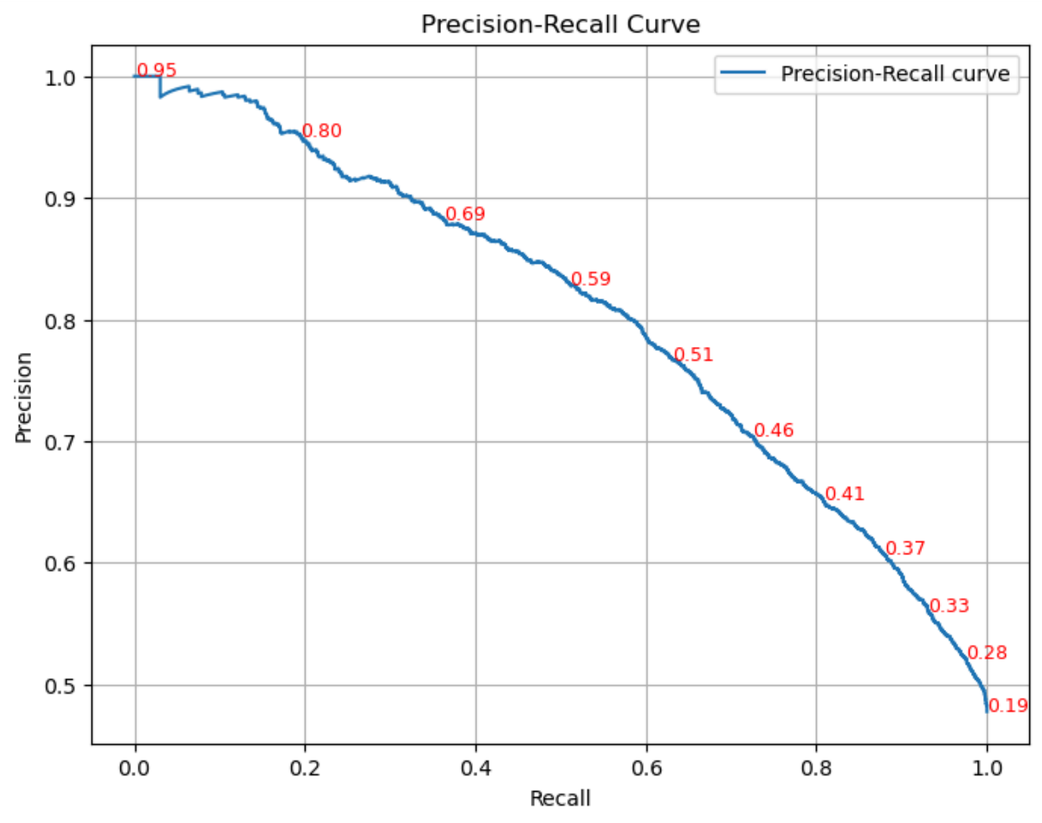


AUC = 0.8031
Score menunjukan model memiliki performa yang baik dalam membedakan antara kelas positif dan negatif.

AUC sebesar 0.8031 mengindikasikan bahwa: Model memiliki kemungkinan 80.31% untuk memberikan skor prediksi lebih tinggi kepada pelanggan yang benar-benar akan berlangganan dibandingkan pelanggan yang tidak.

ROC CURVE & PR CURVE

PR Curve



Model ini memberikan trade-off yang baik antara precision dan recall.

- Threshold dapat disesuaikan dengan kebijakan marketing bank:
- Untuk meningkatkan F1 Score, threshold dapat dipertimbangkan diantara precision ~0.59 dan recall ~0.5, yang menunjukkan keseimbangan antara kedua metrik.

HASIL AKHIR

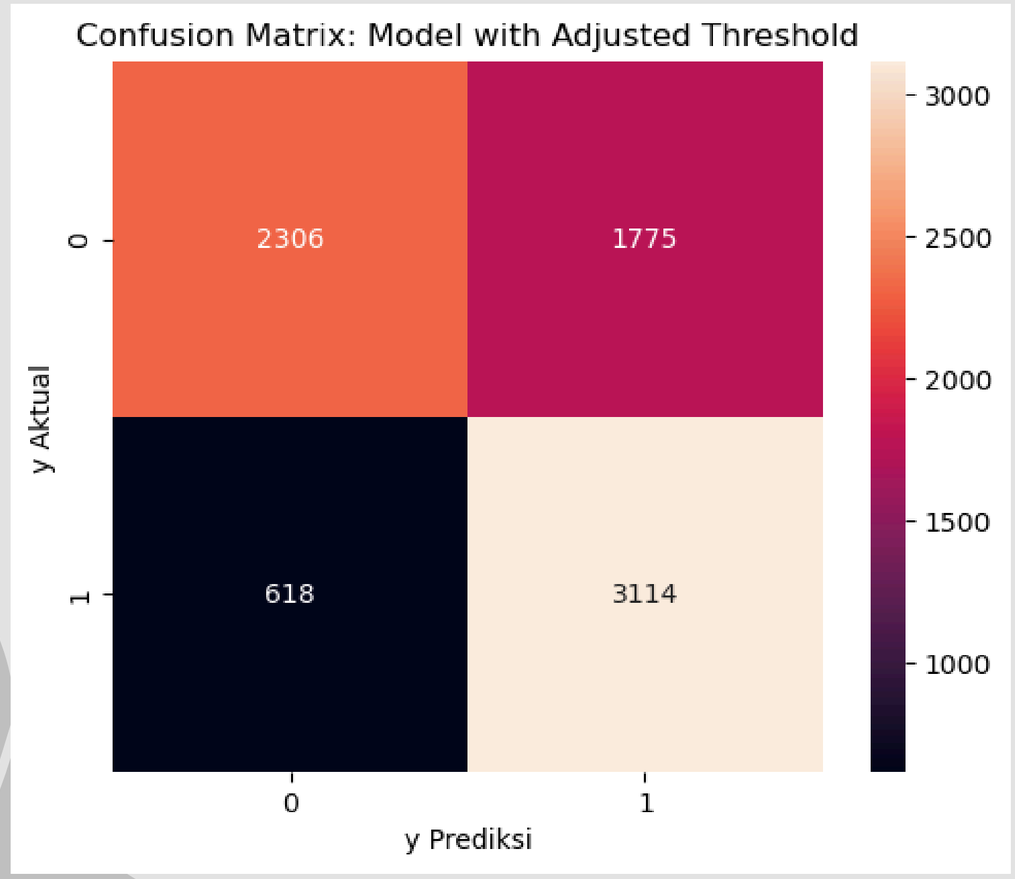
```
=== Classification Report without Threshold ===
              precision    recall  f1-score   support

           0       0.70      0.79      0.74       816
           1       0.74      0.63      0.68       747

    accuracy                           0.72      1563
   macro avg       0.72      0.71      0.71      1563
weighted avg       0.72      0.72      0.71      1563
```

```
=== Classification Report with Threshold 0.4 ===
              precision    recall  f1-score   support

           0      0.789     0.565     0.658      4081
           1      0.637     0.834     0.722      3732

    accuracy                          0.694      7813
   macro avg      0.713     0.700     0.690      7813
weighted avg      0.716     0.694     0.689      7813
```

- F1 Score sebelum threshold memberikan nilai 68%
- F1 Score dengan threshold 0.4 memberikan nilai 72%, namun akan mengorbankan nilai precision hingga 63%

HASIL AKHIR

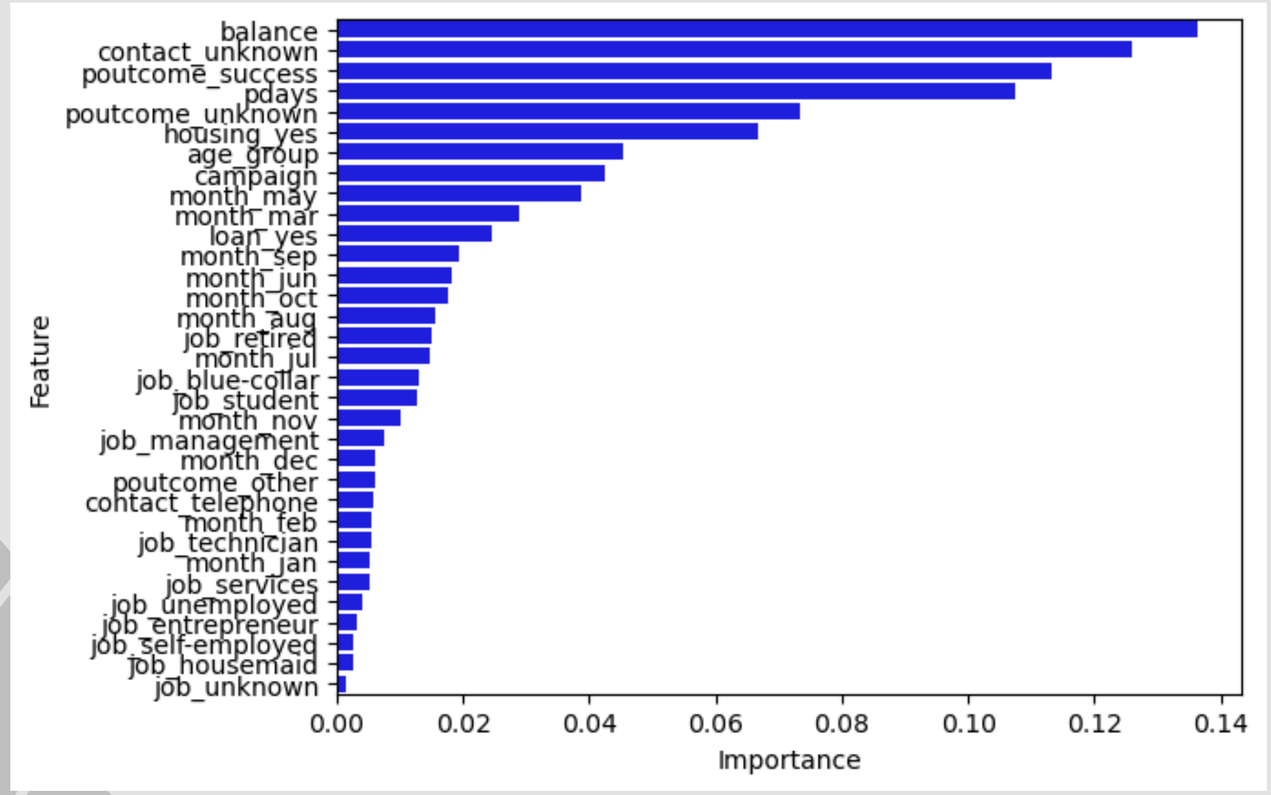
Confusion Matrix: Model dengan Threshold 0.4



- FP :1775. Biaya kampanye terbuang
- FN : 618. Gagal mendeteksi calon nasabah dan kehilangan peluang penjualan

MODEL INTERPRETATION

Feature Importances



- balance: Fitur paling berpengaruh dalam model, semakin tinggi saldo kemungkinan untuk term keputusannya lebih besar
- contact_unknown: Metode kontak yang tidak diketahui sangat memengaruhi hasil
- poutcome_success: Nasabah yang sebelumnya sukses dalam kampanye memiliki kecenderungan lebih tinggi untuk merespons positif

KESIMPULAN

Performa Model ML

Model ML yang telah dibuat sudah dapat memprediksi nasabah akan berlangganan term deposit/tidak berlangganan. Model yang dibuat adalah model RandomForestClassifier dengan NearMiss untuk resamplingnya.

Performa model ML:

- F1 Score setelah hyperparameter tuning sebesar 68%
- F1 Score setelah threshold menjadi 0.4 sebesar 72%.

Kesimpulan

- Model cukup seimbang dan sesuai dengan kebutuhan bisnis.
- **F1 Score** tetap menjadi metrik ideal karena:
  - Menyeimbangkan antara tidak menyia-nyiakan peluang (Recall) dan menghindari biaya promosi sia-sia (Precision)
  - Dengan threshold yang disesuaikan, bank berpotensi meningkatkan ROI kampanye secara signifikan.
- Informasi yang didapatkan dari model ML dapat digunakan oleh tim marketing Bank X

KESIMPULAN

Estimasi Kerugian Finansial dengan model ML

Asumsi:

- Kerugian 1 FN = Rp 5.000.000 (potensi pendapatan dari satu nasabah potensial tp tidak dilanjuti)
- Biaya 1 FP = Rp 500.000 (biaya kampanye per nasabah yang tidak potensial)

Hitung Total Kerugian:

- Total_FN_Loss = 618 × 5.000.000 = Rp 3.090.000.000
- Total_FP_Loss = 1.775 × 500.000 = Rp 887.500.000

Total Kerugian = Rp 887.500.000 + Rp 3.090.000.000

Total = Rp 3.977.500.000

Estimasi Kerugian Finansial tanpa model ML

- Semua nasabah yang deposit (label = 1) akan False Negatif (FN)
- Tidak ada False Positive (FP) karena tidak terdapat tindakan retensi.
- Jumlah nasabah deposit (label = 1) = 3732

Maka;

- FN = 3732
- FP = 0

Hitung Kerugian:

- FN Loss = 3732 x Rp 5.000.000 = Rp 18.660.000.000
- FP Loss = 0 x Rp 500.000 = 0

Total Kerugian = Rp 18.660.000.000

Terdapat efisiensi biaya : Rp 18.660.000.000 - Rp 3.977.500.000 = Rp 14.682.500.000

REKOMENDASI

Rekomendasi bisnis

- Optimalisasi fitur-fitur penting dari Feature Importance
  - Kampanye berfokus pada nasabah dengan balance tinggi yang berpotensi untuk langganan deposito
- Meminimalkan biaya promosi pada nasabah tidak potensial
  - melakukan strategi multi-tier marketing: dengan nasabah potensi tinggi dihubungi langsung, sisanya dapat melalui promosi biaya rendah

THANK YOU