

Prediksi Nasabah Produk Deposito Berjangka

Kampanye Pemasaran Bank

Mengoptimalkan pemasaran
menggunakan **Machine Learning**

Outline

Machine Learning dalam
memprediksi nasabah yang
potensial untuk membuka
deposito berjangka

- 01 Permasalahan Bisnis
- 02 Pemahaman Data
- 03 Model Machine Learning
- 04 Evaluasi Model
- 05 Kesimpulan
- 06 Rekomendasi



1. Permasalahan Bisnis

Latar Belakang

Bank tetap harus bersaing untuk **menarik nasabah baru** dan **mempertahankan nasabah yang sudah ada**.

Pemasaran yang efektif sangat dibutuhkan, tidak semua nasabah tertarik membuka deposito dan tidak semua hasil dari kampanye pemasaran pasti akan mendapatkan nasabah.

1. Permasalahan Bisnis

Memprediksi calon nasabah akan membuka deposito atau tidak setelah dilakukan kampanye pemasaran



Pernyataan Masalah	Tujuan	Pendekatan Analitik
Tidak semua nasabah tertarik untuk membuka deposito baru, dan tidak semua respons terhadap kampanye pemasaran berujung pada pembukaan produk	<ul style="list-style-type: none">• Identifikasi nasabah yang memiliki kemungkinan tinggi (potensi)• Menyusun strategi pemasaran yang efektif (meningkatkan konversi dari calon nasabah menjadi deposan)• Meningkatkan pendapatan bank melalui peningkatan jumlah deposito	<ul style="list-style-type: none">• Pengumpulan dan Pembersihan Data• Feature Engineering• Pemodelan• Evaluasi dan Tuning• Deployment

Evaluasi Metrik

False Positives (FP): Jumlah nasabah yang tidak membuka deposito, tetapi diprediksi akan membuka deposito.

False Negatives (FN): Jumlah nasabah yang membuka deposito, tetapi diprediksi tidak akan membuka deposito.



1. Permasalahan Bisnis

- Peningkatan biaya pemasaran **berbanding lurus** dengan yang dihasilkan.
- Berfokus pada meminimalisir potensi pendapatan yang hilang akibat **False Negative**.
- **Recall** menjadi metrik yang paling sesuai, mempertimbangkan keuntungan bisnis (semakin banyak nasabah yang membuka deposito akan **semakin profit**).

2. Pemahaman Data



Data berkualitas akan menghasilkan machine learning yang lebih baik daripada data yang kurang berkualitas.

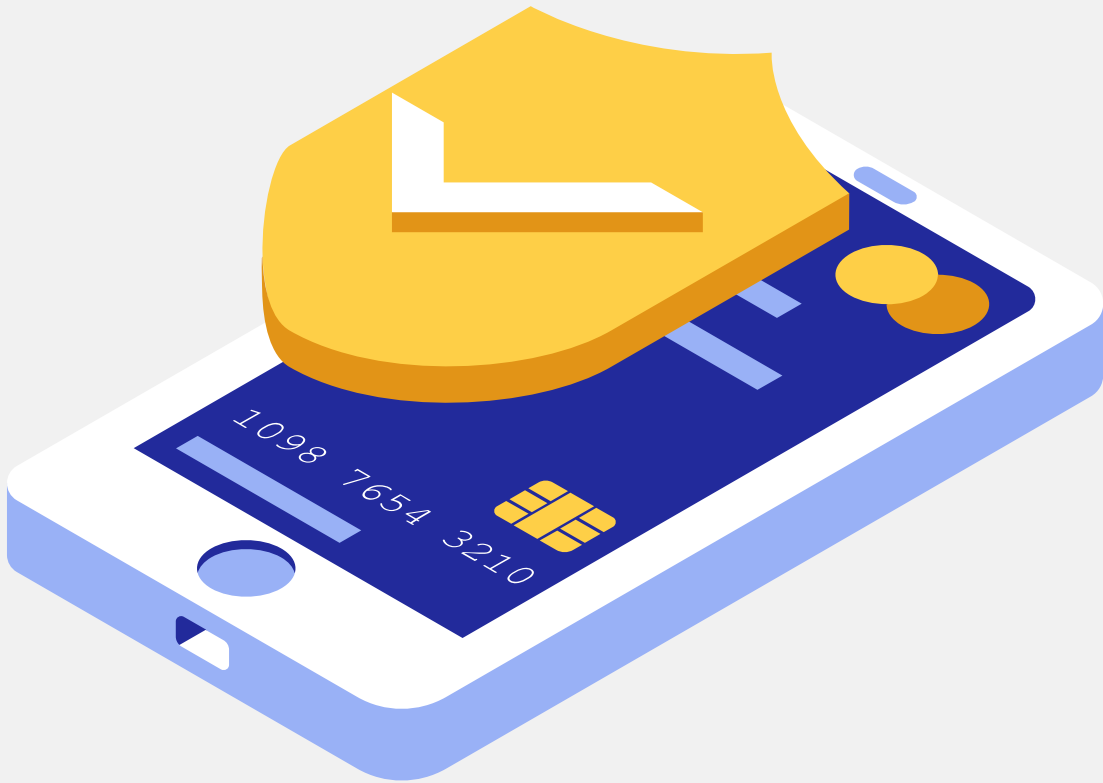
✓	EDA	Customer Profile (age, job, balance, housing, loan)
		Marketing Data (contact, month, campaign, pdays, poutcome, deposit)
✓	Hubungan Antar Variable	Adakah multikolinearitas?
		Apakah Data Seimbang?
✓	Pra-Pemrosesan Data	Pembersihan Data, Transformasi Data, Penanganan Ouliers
		Encoding, Duplikasi, Fitur (X) dan target (y), Pemisahan Data (train, test dan valdation set)

2. Pemahaman Data

Fitur pada Data

Terdapat **7813** baris data, dengan **10** kolom, dan **1** kolom untuk target

Terdapat kolom-kolom berisi **Profil Nasabah** dan kolom-kolom berisi **Data Pemasaran**.



Profil Nasabah

Fitur	Deskripsi	Tipe Data	Jumlah Nilai Unik
age	Usia nasabah	numerik	75
job	Jenis Pekerjaan Nasabah	kategori	11
balance	saldo rekening nasabah	numerik	3153
housing	apakah nasabah memiliki rumah	kategori	2
loan	apakah nasabah memiliki pinjaman	kategori	2

Data Pemasaran

Fitur	Deskripsi	Tipe Data	Jumlah Nilai Unik
contact	Jenis saluran komunikasi yang digunakan	kategori	3
month	bulan terakhir nasabah dihubungi	kategori	422
campaign	berapa kali nasbah ditawarkan	numerik	32
pdays	jumlah hari sejak nasabah terakhir dihubungi	numerik	422
poutcome	hasil dari kampanye pemasaran sebelumnya	kategori	4

Target

Fitur	Deskripsi	Tipe Data	Jumlah Nilai Unik
deposit	apakah nasabah membuka deposito	kategori	2

2. Pemahaman Data

Adakah Multikolinearitas pada Data?

	Variable	VIF
1	const	14.864787
2	age	1.011552
3	balance	1.018034
4	campiagn	1.025711
5	pdays	1.033152
6	deposit_numeric	1.050617

VIF < 5 untuk semua variabel menunjukkan bahwa **tidak ada** masalah serius dengan multikolinearitas

Apakah Data Seimbang?

- deposit no 4081
- deposit yes 3732
- **Rasio imbalance:**
1.0935155412647375

distribusi kelas dalam data **relatif seimbang**, dengan perbedaan jumlah sekitar 9%

2. Pemahaman Data

Pembersihan Data

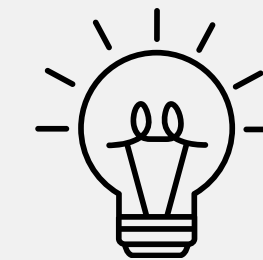
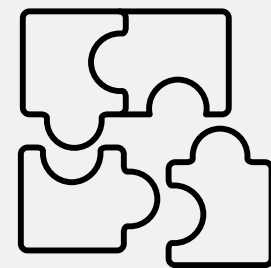
tidak terdapat missing values
namun terdapat beberapa nilai
'unknown' di beberapa kolom



value 'unknown'

✓	value 'unknown'	kolom job terdapat value 'unknown' sebanyak 54 baris
		digabungkan menjadi 1 kategori dg simple imputer dengan nilai most frequent
✓	value 'unknown'	kolom contact dengan nilai 'unknown' sebanyak 1639 (21%)
		akan diganti dengan others agar tidak miss-presepsi
✓	value 'unknown'	kolom poutcome dengan nilai unknown sebanyak 5819 (74%)
		disederhanakan menjadi kategori other (sudah ada sebelumnya)

Pembersihan Data



Outliers	Duplikat	Encoding & Scaling
<p>seluruh outliers dianggap sebagai natural outliers sehingga tetap dibiarkan:</p> <ul style="list-style-type: none">• kolom age memiliki 121 outliers• kolom balance memiliki 736 outliers• kolom campaign memiliki 429 nilai outliers• kolom pdays memiliki 1432 nilai outliers, dan 5809 berisi nilai negatif akan dilakukan imputasi menjadi 0 yang artinya belum di hubungi	<p>Data duplikat sebanyak 8 baris data dan dilakukan drop (penghapusan) untuk menghindari redundansi, agar model lebih memahami distribusi data yang sesungguhnya.</p>	<ul style="list-style-type: none">• One-hot encoding untuk seluruh kolom kategorikal (job, housing, loan, contact, month, poutcome)• binary encoding untuk target (deposit)• RobustScaler untuk seluruh numerikal (menghindari distorsi karena outliers)



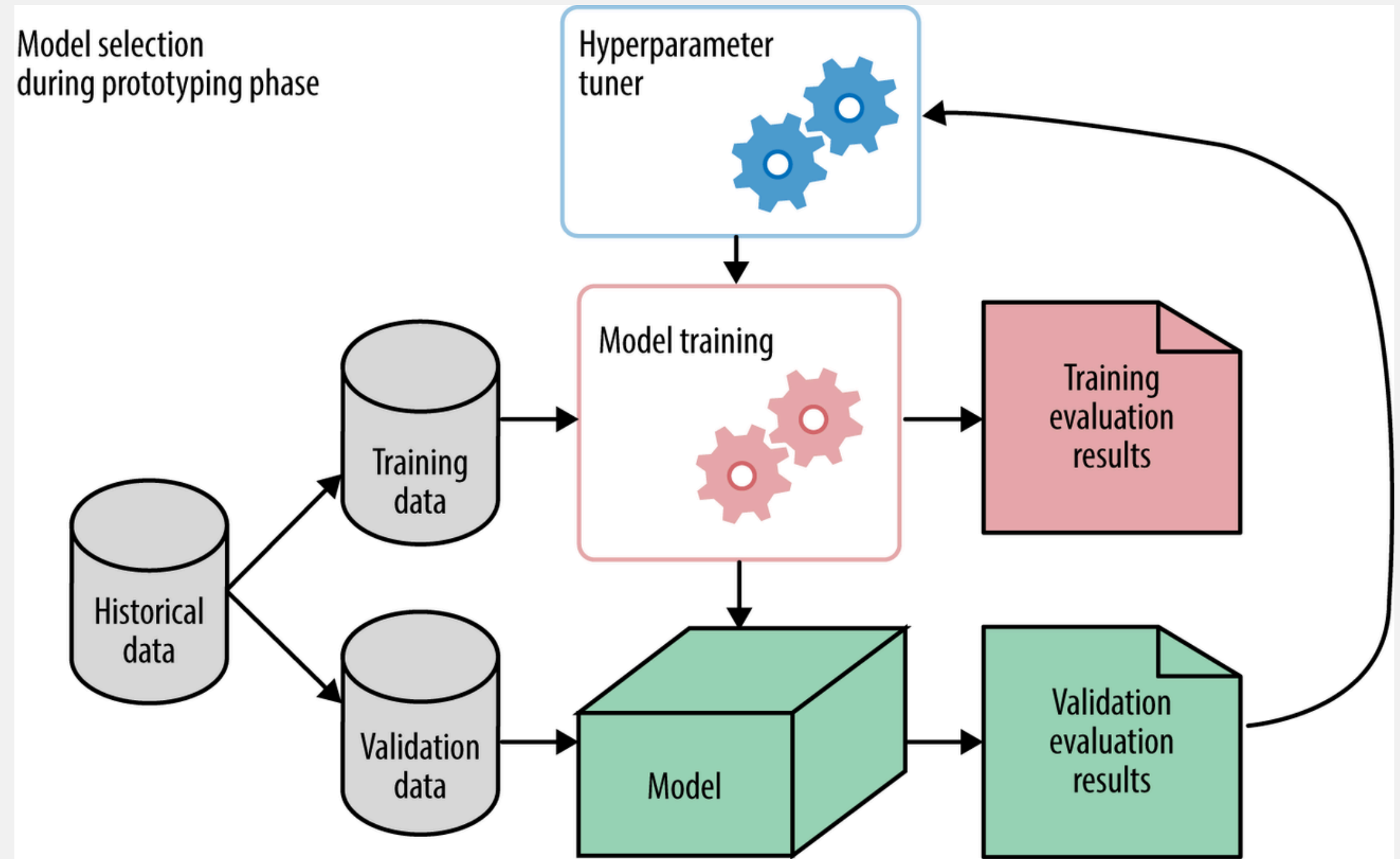
3. Model *Machine Learning*

- Pencocokan (*benchmarking*) model dengan Recall Score tertinggi pada data train maupun data test.
- Tidak *underfitting* atau pun *overfitting*
- Tune model untuk mendapat score yang lebih baik

3. Model *Machine Learning*

Mencari model dengan **Recall Score tertinggi**.

Validasi silang (*cross validation*) untuk mendapatkan **nilai rata-rata (mean) terbaik** dan **kestabilan** model.

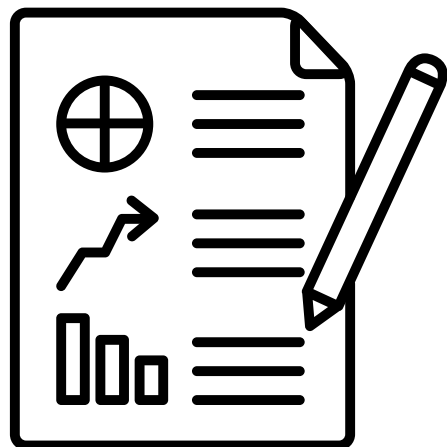
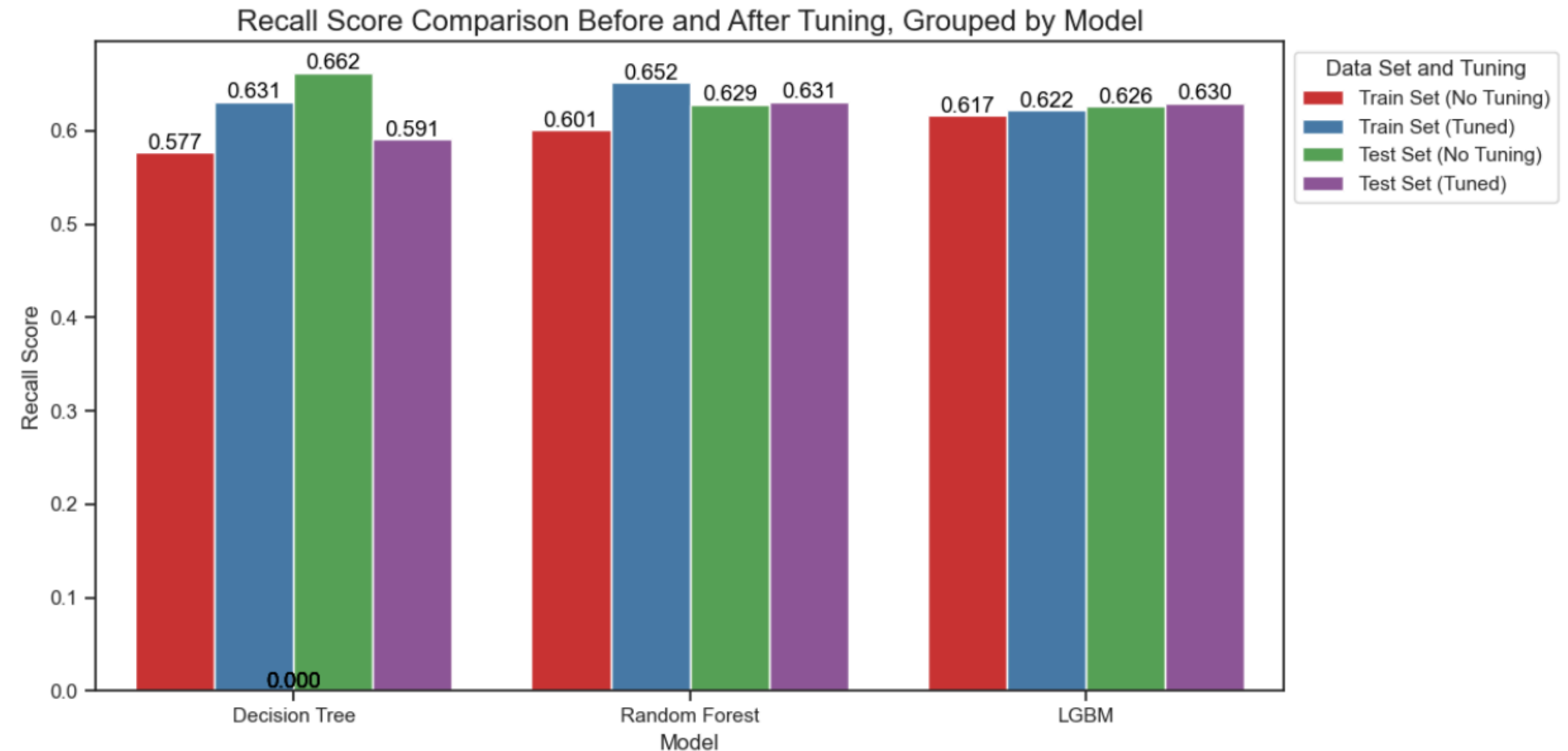


3. Model Machine Learning

Model Terbaik

LGBM Classifier

LGBM		
Train Set	No Tune	61.7%
	Tuned	62.2%
Test Set	No Tune	62.6%
	Tuned	63.0%

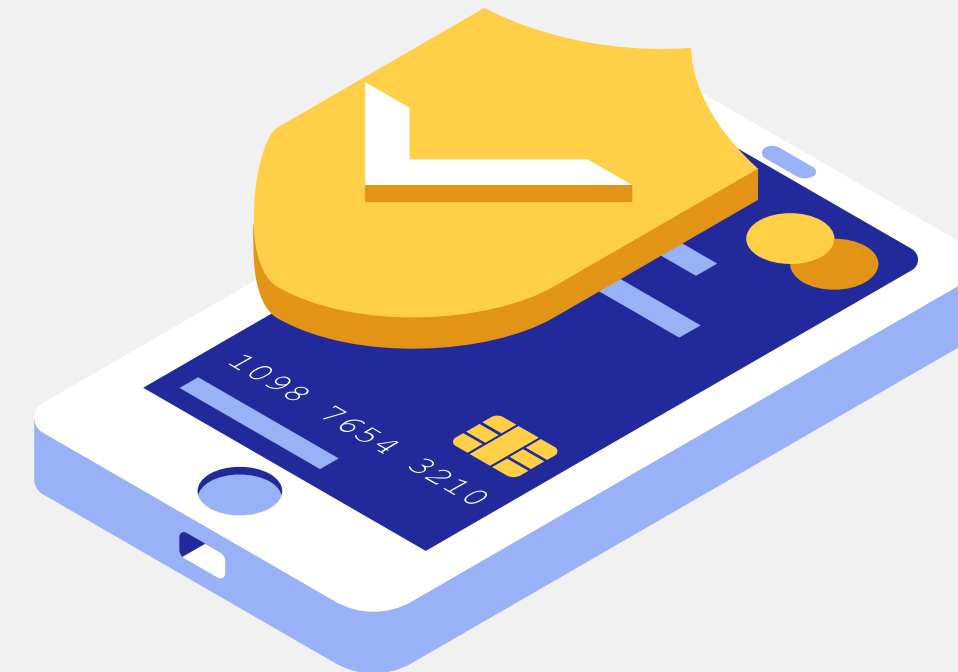
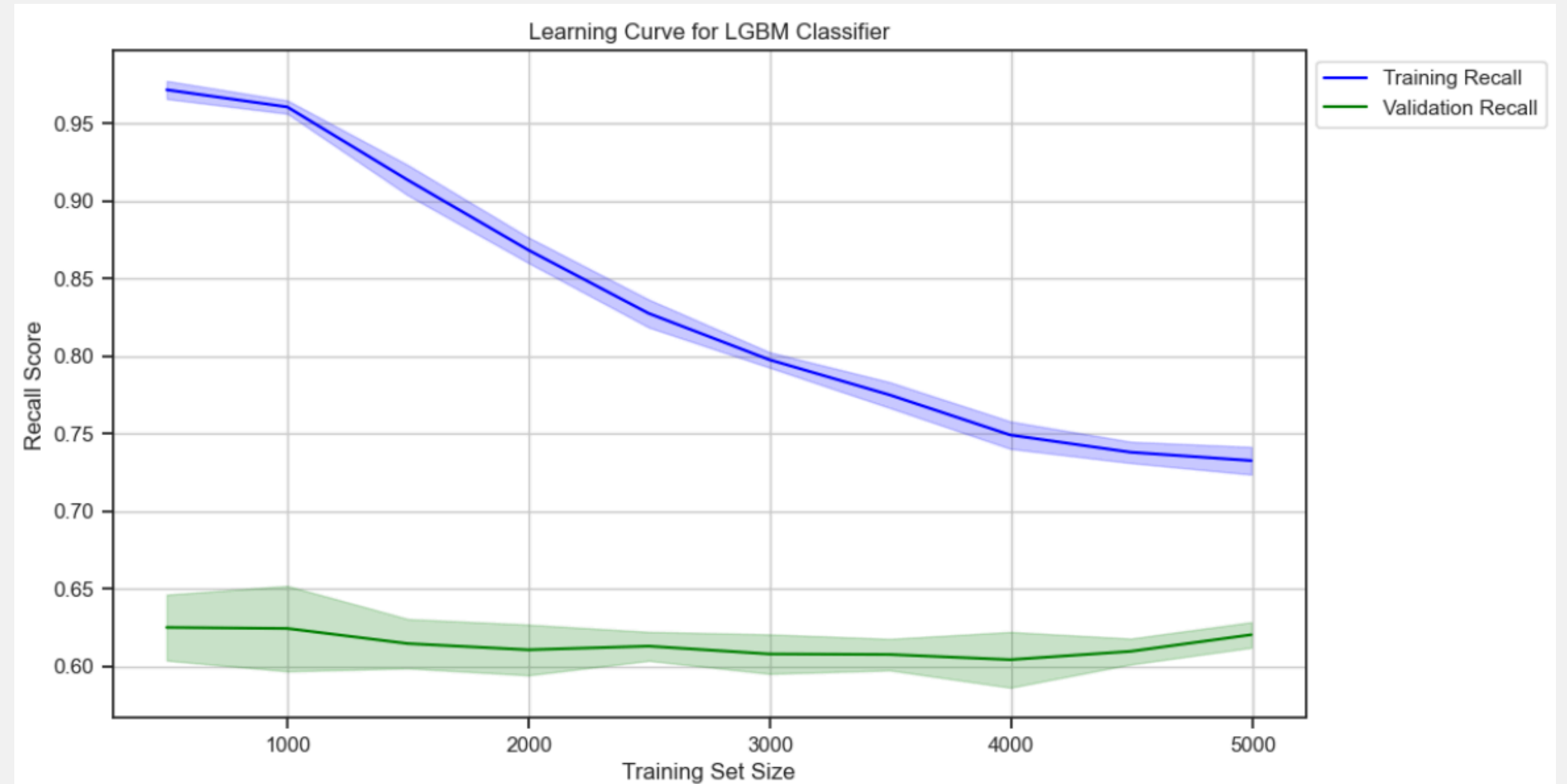


1. **LGBM paling stabil** dengan recall di kedua test set (sebelum dan setelah dilakukan *tuning*. Model Random Forest cukup kompetitif, namun sedikit di bawah pada test set. Performa Decision Tree masih lebih rendah dibandingkan 2 model lainnya.
2. LGBM menjadi model terbaik dengan recall score tertinggi 63% dan **hyperparameter** {'model__learning_rate': 0.2, 'model__max_depth': 5, 'model__min_child_samples': 40, 'model__n_estimators': 200, 'model__num_leaves': 31}.

4. Evaluasi Model

Learning Curve

- Tidak ada perbedaan yang jauh antara *training* dan *validation recall score*, model tidak *overfitting*.
- Model semakin baik dalam bekerja dengan data yang sebelumnya tidak terlihat (data validasi)
- Model sangat baik dalam hal *generalization*, penting karena di dunia nyata, data baru akan datang secara terus menerus.



4. Evaluasi Model

Feature Importance

1. Fitur paling penting:

- **pdays** / jumlah hari sejak nasabah terakhir dihubungi, berpengaruh pada model secara signifikan
- **poutcome** / keberhasilan sebelumnya memiliki pengaruh besar terhadap prediksi
- **campaign** / jumlah berapa kali ditawarkan juga berpengaruh

2. Fitur Kategorikal

- nasabah yang sudah memiliki rumah dan memiliki pinjaman serta dihubungi melalui telepon genggam cukup berpengaruh pada prediksi.

3. Fitur Numerikal

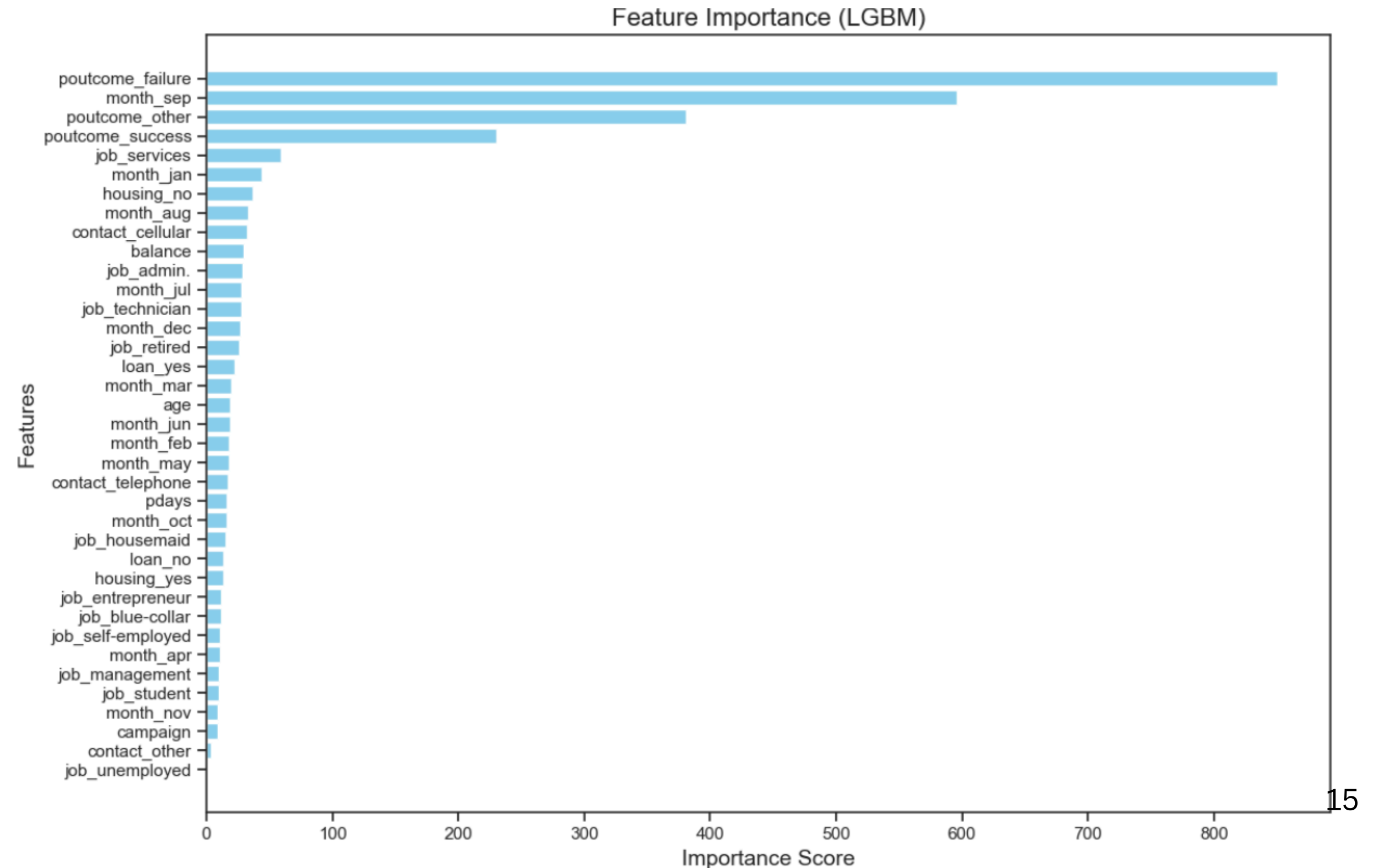
- fitur seperti age, campaign dan balance berkontribusi sedang namun tetap relevan untuk model.

4. Fitur dengan kontribusi rendah

- fitur month, job memiliki skor yang rendah, kemungkinan kurang relevan untuk memprediksi target.

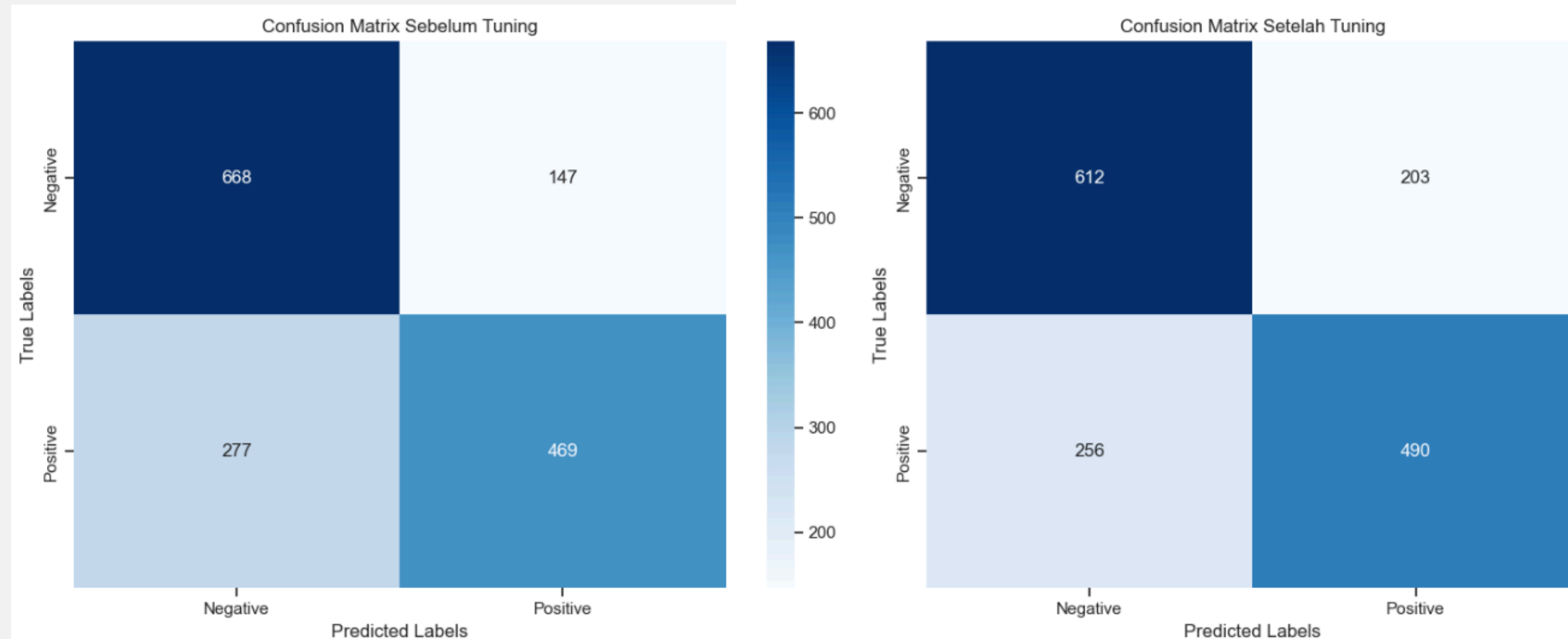
5. Keseimbangan antar fitur

- fitur numerik yang sudah dilakukan standarisasi (*RobustScaler*) memungkinkan model fokus pada variabel yang benar-benar signifikan. Fitur kategorikal yang sudah di *encode* menunjukkan perbedaan kelas yang penting.



4. Evaluasi Model

Confusion Matrix



Setelah dilakukan tuning, jumlah False Negative berkurang. Ini penting karena perusahaan tidak menginginkan potensi yang seharusnya didapatkan (profit) dari nasabah yang membuka deposito. Juga kelas positive harus semakin meningkat agar membantu tim marketing menyusun strategi yang lebih tepat sasaran (efektif).

5. Kesimpulan



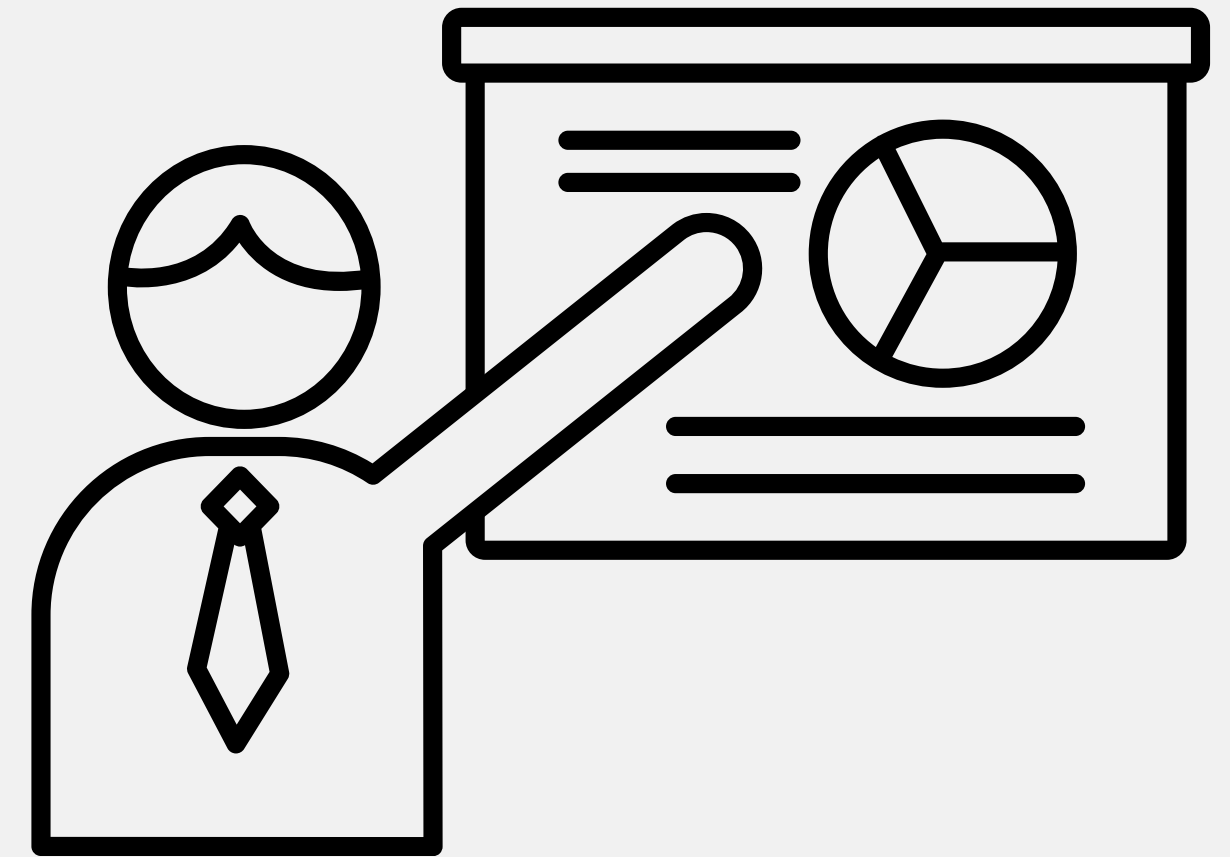
MODEL

- Metrik utama adalah ***Recall***, fokus pada potensi profit (mendapatkan sebanyak mungkin nasabah yang menempatkan deposito)
- Parameter terbaik dari model **LGBM Classifier** (model learning rate: 0.2, max depth: 5, min child sample: 40, n_estimator: 200, num leaves 31).
- Model berhasil meminimalisir kesalahan prediksi terhadap nasabah yang diprediksi tidak deposit, padahal aktualnya deposit (***False Negative***).

5. Kesimpulan

Interpretasi Fitur

- **pdays**, nasabah yang belum pernah dihubungi / sudah lama tidak dihubungi lebih mungkin melakukan deposit, efektivitas targeting nasabah baru.
- **poutcome**, nasabah yang berhasil dalam campaign sebelumnya (*success*) lebih mungkin melakukan deposit lagi, pentingnya data historis.
- **campaign**, nasabah yang sering ditawarkan lebih memungkinkan untuk melakukan deposit
- **contact**, metode other (selain telepon dan selular) mengurangi kemungkinan deposit, pentingnya metode kontak yang efektif.
- **balance**, nasabah dengan saldo tinggi cenderung tidak melakukan deposit, mungkin karena dana yang ada digunakan dengan instrumen investasi lain.



5. Kesimpulan Perhitungan Biaya

Type I Error
False Positive

Skala=
Total Data Confusion Matrix X Total Data Asli
1561 X 7813 ≈ **5.005**

[TN FN] = [3063 1016]
[FP TP] = [1281 2452]

Type II Error
False Negative

	Tanpa Model	Dengan Model	Penghematan
	seluruh nasabah (7,813)	prediksi positif (3,733)	4,080 nasabah
Biaya campaign @ Rp 132,000,-	Rp 1,031,316,000	Rp 492,756,000	Rp 538,560,000

	Tanpa Model	Dengan Model (tidak mencanangkan keuntungan)	Ril profit yang hilang
	Nasabah yang tidak deposit (4,081)	prediksi negatif (3,063)	False Negative (1,016)
Potensi kehilangan profit @ Rp 452,000,-	Rp 1,844,612,000	Rp 1,384,476,000	Rp 459,232,000



6. Rekomendasi Model

untuk mengembangkan Machine Learning agar lebih baik lagi

Pembaharuan

Lakukan pembaharuan / *update* karena suku bunga simpanan dan kurs selalu berubah-ubah seiring berkembangnya zaman

Eksplorasi algoritma

Mengembangkan model dengan menggunakan algoritma yang berbeda

Penambahan Fitur

Menambahkan fitur seperti segmentasi nasabah (misalnya segmentasi usia, status pernikahan, jenis kelamin, dan lain-lain)

6. Rekomendasi Bank / Perusahaan

agar perusahaan dapat
mengoptimalkan pencapaian profit

✓	Data Historis	Data hasil campaign sebelumnya
		Nasabah yang 'success' adalah kunci dalam memprediksi dan meningkatkan respons positif.
✓	Metode Kontak	Memilih metode kontak yang efektif sangat penting
		Menghindari metode kontak yang kurang efektif, dapat meningkatkan kemungkinan keberhasilan.
✓	Profil Keuangan Nasabah	Memahami beban finansial nasabah
		Apabila nasabah memiliki pinjaman pribadi, serta saldo rekening dapat membantu dalam menargetkan nasabah yang lebih mungkin membuka deposito.
✓	Usia	Usia tidak terlalu memiliki pengaruh
		Campaign dapat disesuaikan dengan preferensi keuangan dan perilaku nasabah dari pada usia



Terima kasih

capstone module 3
oleh Nuraini Septiana
JCDS 2604

