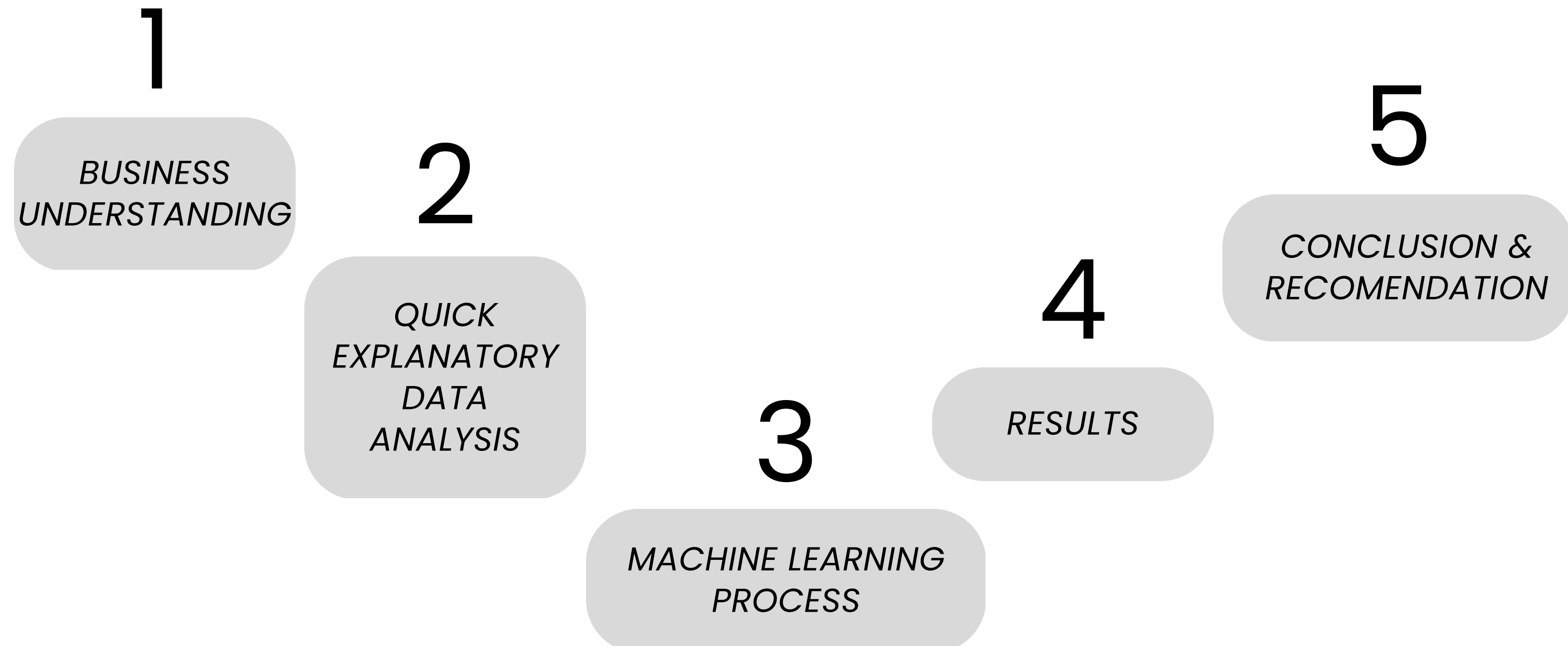


HOTEL BOOKING CANCELLATION PREDICTION

MACHINE LEARNING

CAN WE PREDICT?

OUTLINE



BUSINESS UNDERSTANDING

LATAR BELAKANG

Dewasa ini, teknologi semakin berkembang dan online booking semakin populer di industri perhotelan. Online booking memudahkan pengguna dalam memesan kamar dan juga dalam proses pembatalan. Namun, perilaku impulsive dari pengguna dan adanya biaya pembatalan gratis menyebabkan tingkat pembatalan menjadi tinggi. Ini adalah dampak negatif dari adanya fitur pembatalan gratis tersebut. Hal ini menjadi masalah bagi bisnis perhotelan karena dapat mengakibatkan kerugian finansial.

ANALYTICAL APPROACH

Jadi pada proyek ini akan membangun sebuah model klasifikasi yang mampu membedakan pelanggan yang berpotensi melakukan pembatalan dan tidak dalam reservasi sebuah kamar hotel.

OBJEKTIF

1. Perusahaan ingin menganalisa karakteristik seseorang yang melakukan reservasi hotel untuk memprediksi apakah pelanggan berpotensi untuk melakukan pembatalan reservasi atau tidak.
2. Model prediksi ini akan membantu pihak hotel untuk mengantisipasi kerugian finansial dan reputasi hotel yang disebabkan oleh potensi pembatalan pelanggan.
3. Dengan menggunakan model ini, pihak hotel akan memfokuskan pada pelanggan yang tidak berpotensi untuk melakukan pembatalan, sehingga dapat meningkatkan pendapatan dan reputasi bisnis hotel.
4. Model ini akan mempertimbangkan karakteristik pelanggan, seperti riwayat pemesanan sebelumnya, waktu pemesanan, metode pembayaran, dan faktor lain yang dapat mempengaruhi keputusan pembatalan.
5. Dengan menggunakan data historis, model ini akan mempelajari pola perilaku pelanggan dan menggunakan algoritma machine learning untuk membuat prediksi yang akurat.

BUSINESS UNDERSTANDING

METRIK EVALUASI

FALSE NEGATIVE

PELANGGAN YANG MELAKUKAN PEMBATALAN RESERVASI HOTEL NAMUN DI PREDIKSI OLEH MODEL TIDAK MELAKUKAN PEMBATALAN

Konsekuensi:

- PIHAK HOTEL TELAH MENYIAPKAN ROOM DAN PERSEDIAAN YANG TIDAK PENTING SERTA PENGERAHAN SUMBER DAYA MANUSIA YANG SIA SIA YANG AKAN BERUJUNG KEPADA MENINGKATNYA REVENUE LOSS PADA PIHAK PERUSAHAAN.
- BANYAKNYA KAMAR DENGAN STATUS TERISI NAMUN KENYATAANYA KOSONG, HAL INI AKAN MENYEBABKAN BERKURANGNYA KEUNTUNGAN PADA PIHAK PERUSAHAAN.
- BERPOTENSI TERJADINYA OVERBOOKING SERTA HILANGNYA PELUANG DALAM MENJUAL KAMAR HOTEL KEPADA PELANGGAN LAINNYA.

FALSE POSITIVE

PELANGGAN TIDAK MELAKUKAN PEMBATALAN RESERVASI HOTEL NAMUN DIPREDIKSI MELAKUKAN PEMBATALAN RESERVASI HOTEL.

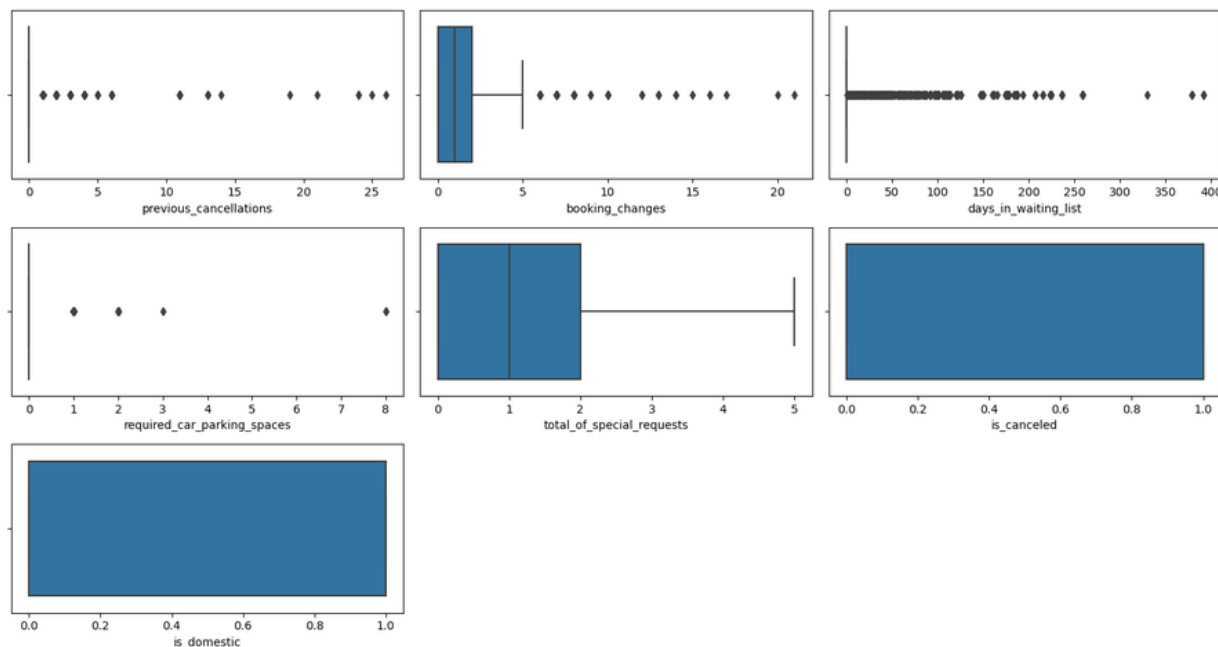
Konsekuensi:

- Tidak adanya persiapan dari pihak hotel dalam mendistribusikan persediaan serta pengerahan sumber daya yang berpotensi terjadinya miss management.
- Berpotensi hilangnya reputasi baik hotel yang diakibatkan kesalahan hotel karena memberikan kamar hotel kepada pelanggan lain.

QUICK EDA

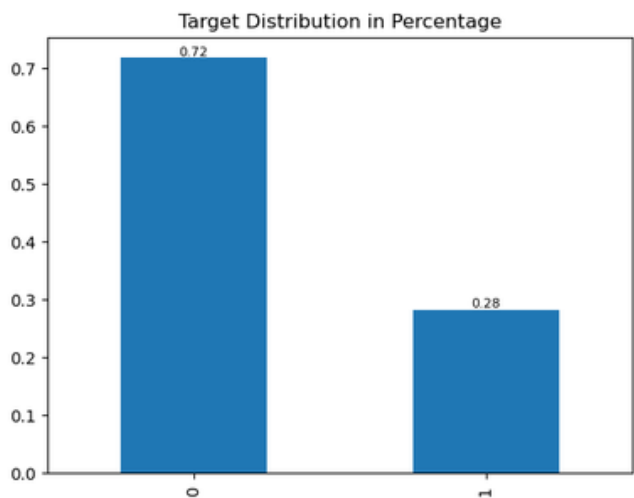
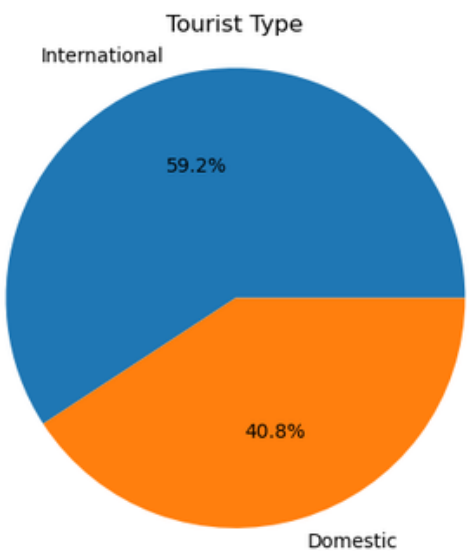
BEFORE

#	Column	Non-Null Count	Dtype
0	country	83222 non-null	object
1	market_segment	83573 non-null	object
2	previous_cancellations	83573 non-null	int64
3	booking_changes	83573 non-null	int64
4	deposit_type	83573 non-null	object
5	days_in_waiting_list	83573 non-null	int64
6	customer_type	83573 non-null	object
7	reserved_room_type	83573 non-null	object
8	required_car_parking_spaces	83573 non-null	int64
9	total_of_special_requests	83573 non-null	int64
10	is_canceled	83573 non-null	int64



AFTER

#	Column	Non-Null Count	Dtype
0	market_segment	83572 non-null	object
1	previous_cancellations	83572 non-null	int64
2	booking_changes	83572 non-null	int64
3	deposit_type	83572 non-null	object
4	days_in_waiting_list	83572 non-null	int64
5	customer_type	83572 non-null	object
6	reserved_room_type	83572 non-null	object
7	required_car_parking_spaces	83572 non-null	int64
8	total_of_special_requests	83572 non-null	int64
9	is_canceled	83572 non-null	int64
10	booking_type	83572 non-null	object
11	is_domestic	83572 non-null	int32



MACHINE LEARNING

TRAIN & VAL SET

	mean roc_auc	mean accuracy	mean recall	mean precision	mean f1 score	mean balanced accuracy
model						
GBoost	0.862553	0.792961	0.436042	0.728661	0.542450	0.684800
XGBoost	0.841881	0.748186	0.490526	0.564070	0.523419	0.670095
AdaBoost	0.840397	0.787668	0.453078	0.690940	0.546495	0.686271
Logistic Regression	0.839950	0.779969	0.427597	0.674151	0.523196	0.673194
LGBM	0.837297	0.748182	0.488831	0.565114	0.522106	0.669583
Random Forest	0.766042	0.700529	0.459887	0.470622	0.464046	0.627595
Decision Tree	0.621007	0.685602	0.422440	0.440779	0.430363	0.605851

Tanpa random sampling

	mean roc_auc	mean accuracy	mean recall	mean precision	mean f1 score	mean balanced accuracy
model						
GBoost	0.859040	0.753491	0.768434	0.546646	0.638170	0.758043
AdaBoost	0.840044	0.752525	0.723917	0.548052	0.623327	0.743838
XGBoost	0.839717	0.739052	0.679820	0.532761	0.596243	0.721118
LGBM	0.838881	0.734708	0.686542	0.524887	0.594021	0.720117
Logistic Regression	0.837881	0.735193	0.764827	0.521443	0.619873	0.744158
Random Forest	0.756171	0.682706	0.528017	0.447814	0.483648	0.635820
Decision Tree	0.619347	0.672592	0.476923	0.428725	0.451058	0.613294

Dengan random oversampling

	mean roc_auc	mean accuracy	mean recall	mean precision	mean f1 score	mean balanced accuracy
model						
GBoost	0.861011	0.742407	0.816051	0.529083	0.641774	0.764737
Logistic Regression	0.837252	0.734233	0.771708	0.520198	0.621261	0.745586
LGBM	0.836616	0.719781	0.754614	0.503720	0.603550	0.730327
XGBoost	0.835439	0.719298	0.761495	0.503055	0.605456	0.732090
AdaBoost	0.834064	0.738082	0.718832	0.527193	0.608036	0.732235
Random Forest	0.785090	0.688961	0.739200	0.467974	0.572883	0.704164
Decision Tree	0.683899	0.674993	0.648906	0.448556	0.529995	0.667071

Dengan random undersampling

MACHINE LEARNING

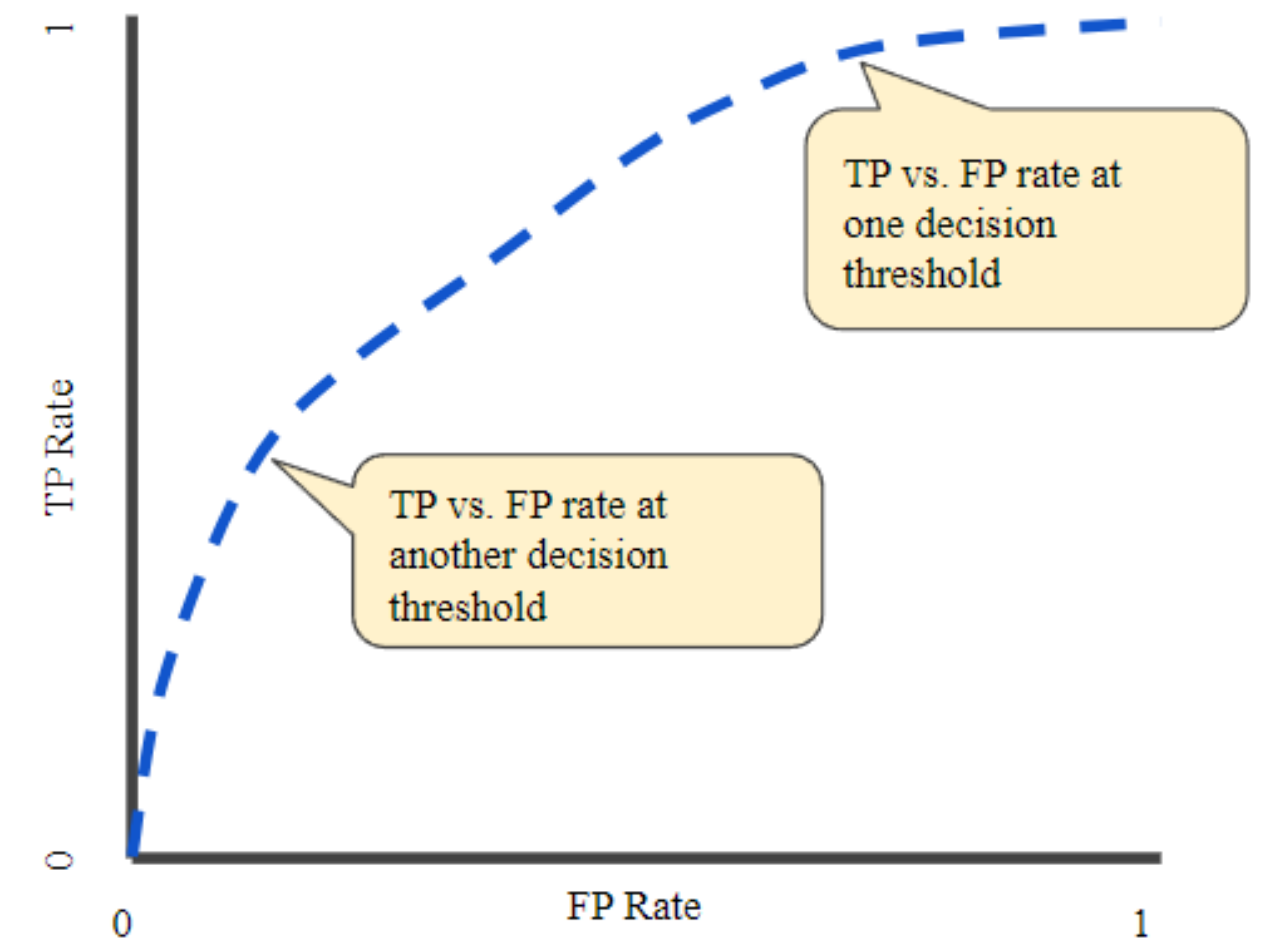
**APA ITU ROC-
AUC DAN
BALANCED
ACCURACY?**

$$\text{balanced-accuracy} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

predicted→ real↓	Class_pos	Class_neg
Class_pos	TP	FN
Class_neg	FP	TN

$$\text{TPR (sensitivity)} = \frac{TP}{TP + FN}$$

$$\text{FPR (1-specificity)} = \frac{FP}{TN + FP}$$



MACHINE LEARNING

TEST SET

Model	ROC AUC	Balanced Acc
GBoost	0.881855154930605	0.7105651912239427
XGBoost	0.8369079535299374	0.671071109408911
AdaBoost	0.8758001860261531	0.7236508544436542

Tanpa random sampling

Model	ROC AUC	Balanced Acc
GBoost	0.8855209644179387	0.7894804034214222
AdaBoost	0.8731283398077729	0.7783553099523992
XGBoost	0.8355218763108461	0.7351589429337418

Dengan random oversampling

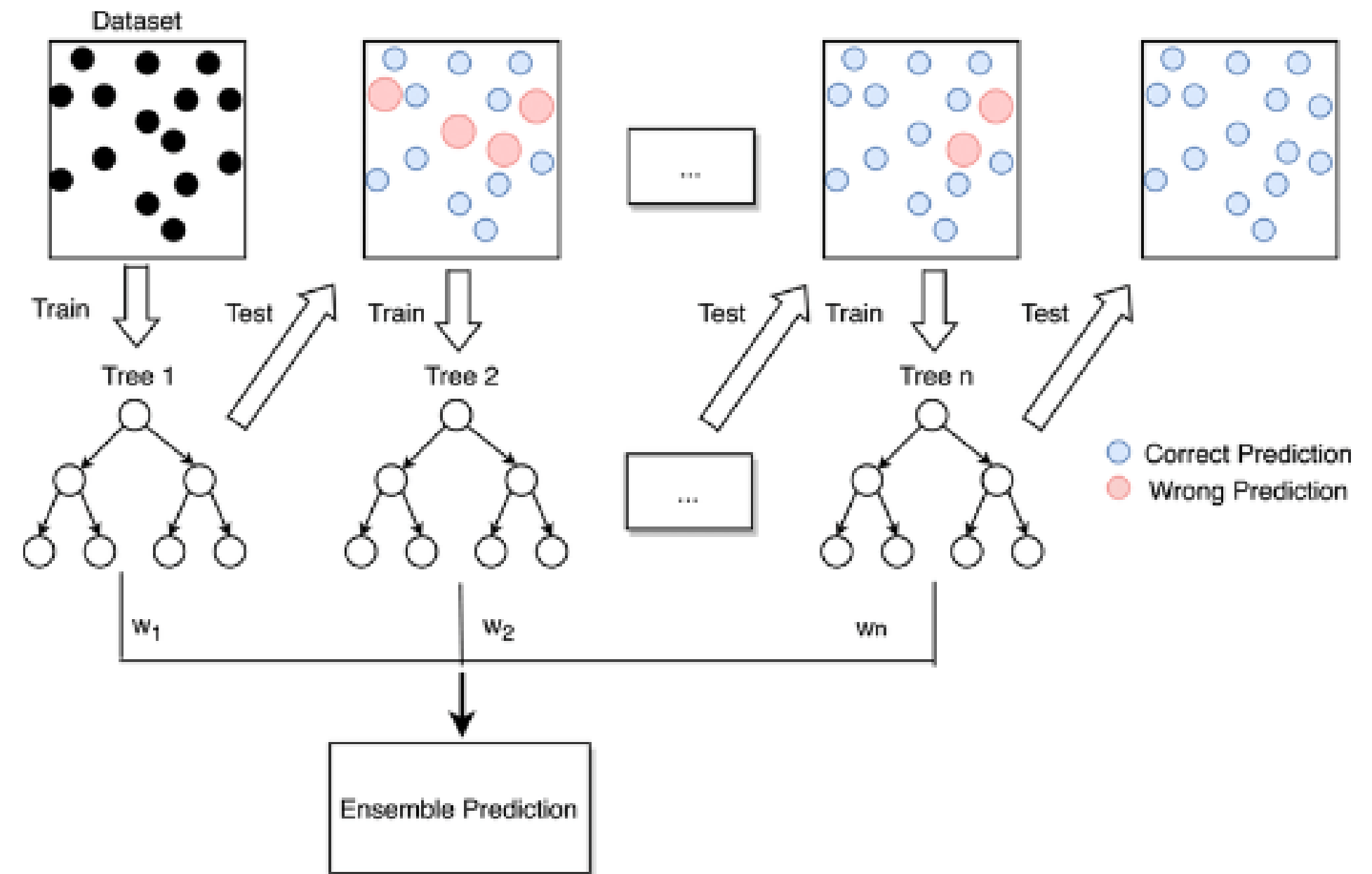
Model	ROC AUC	Balanced Acc
GBoost	0.8722985172621327	0.770613339169448
Logreg	0.8639546971603655	0.754947019022086
LGBM	0.8253998650398497	0.7308183326950083

Dengan random undersampling

MACHINE LEARNING

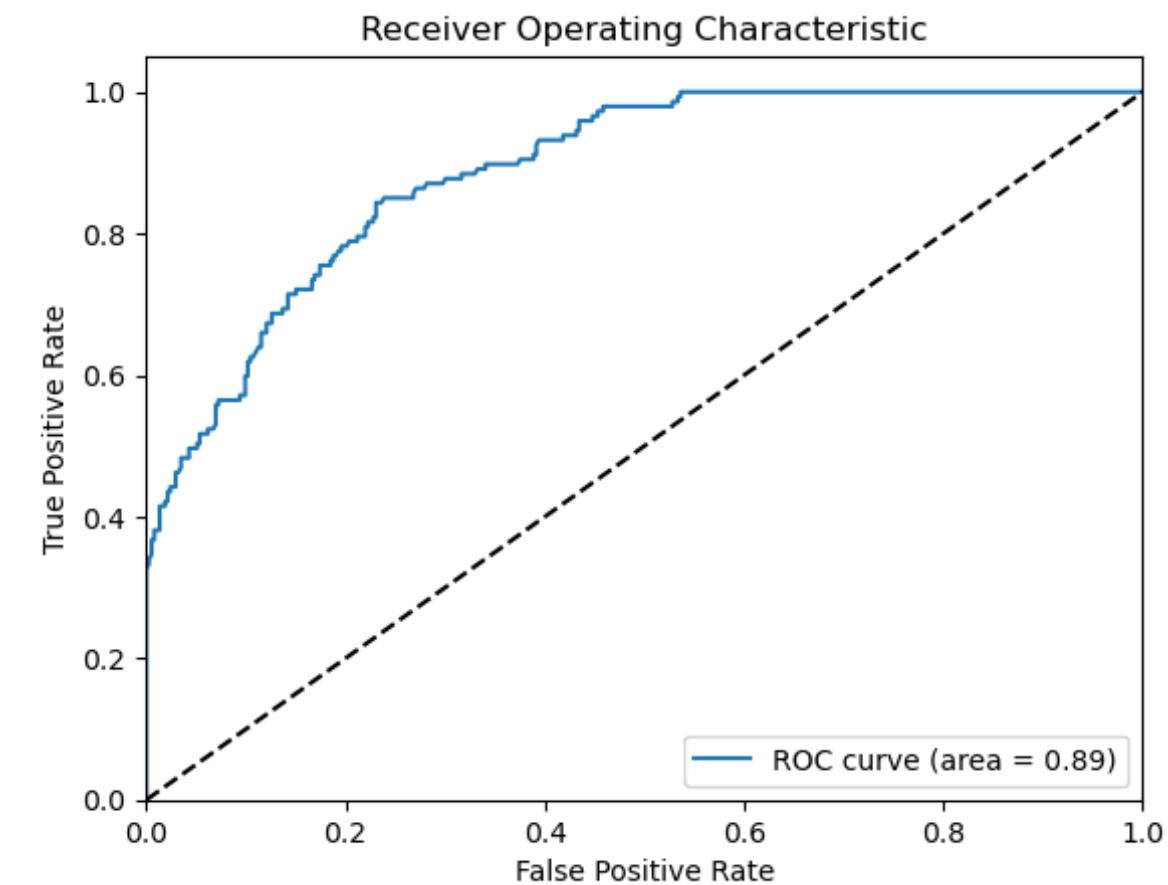
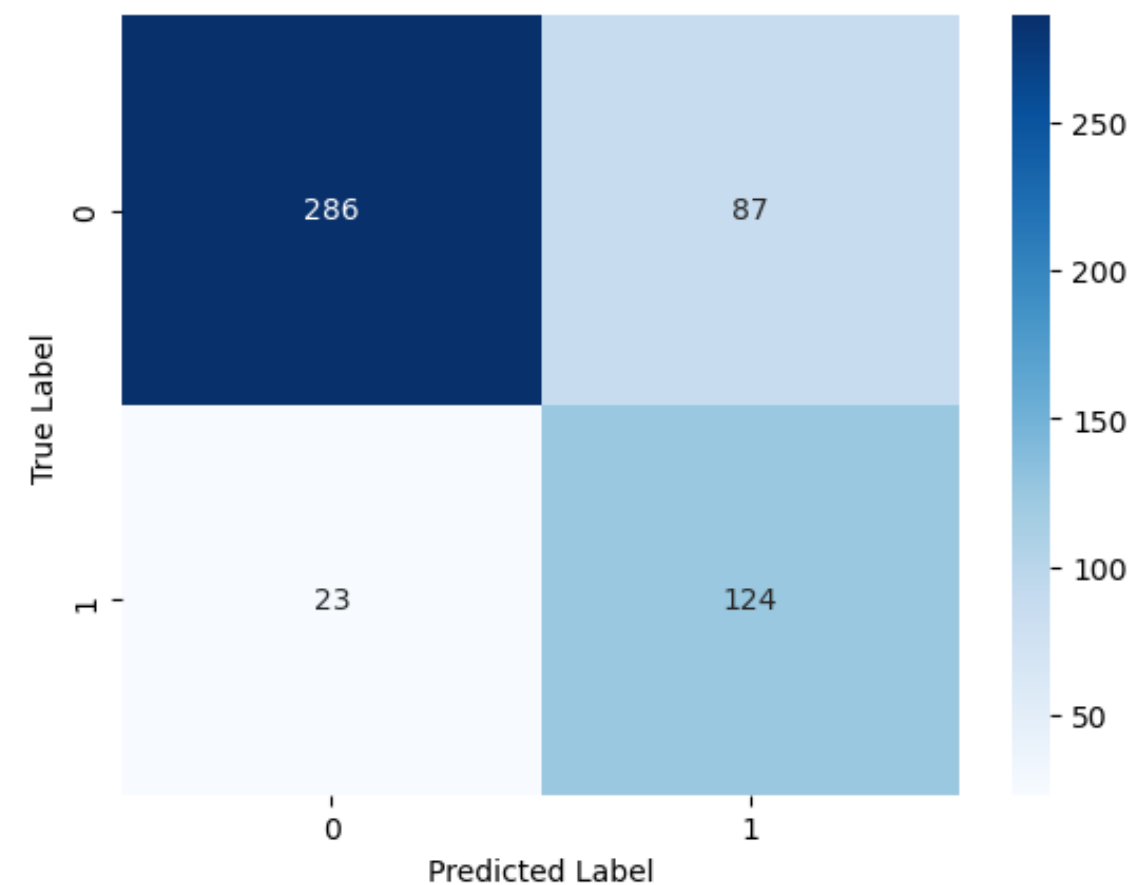
APA ITU GRADIENT BOOSTING?

GRADIENT BOOSTING CLASSIFIER ADALAH ALGORITMA MACHINE LEARNING YANG TERMASUK DALAM KATEGORI ENSEMBLE LEARNING. ENSEMBLE LEARNING ADALAH TEKNIK YANG MEMADUKAN BEBERAPA MODEL MACHINE LEARNING SEDERHANA UNTUK MENGHASILKAN MODEL YANG LEBIH KUAT.



RESULTS

APA ARTINYA?

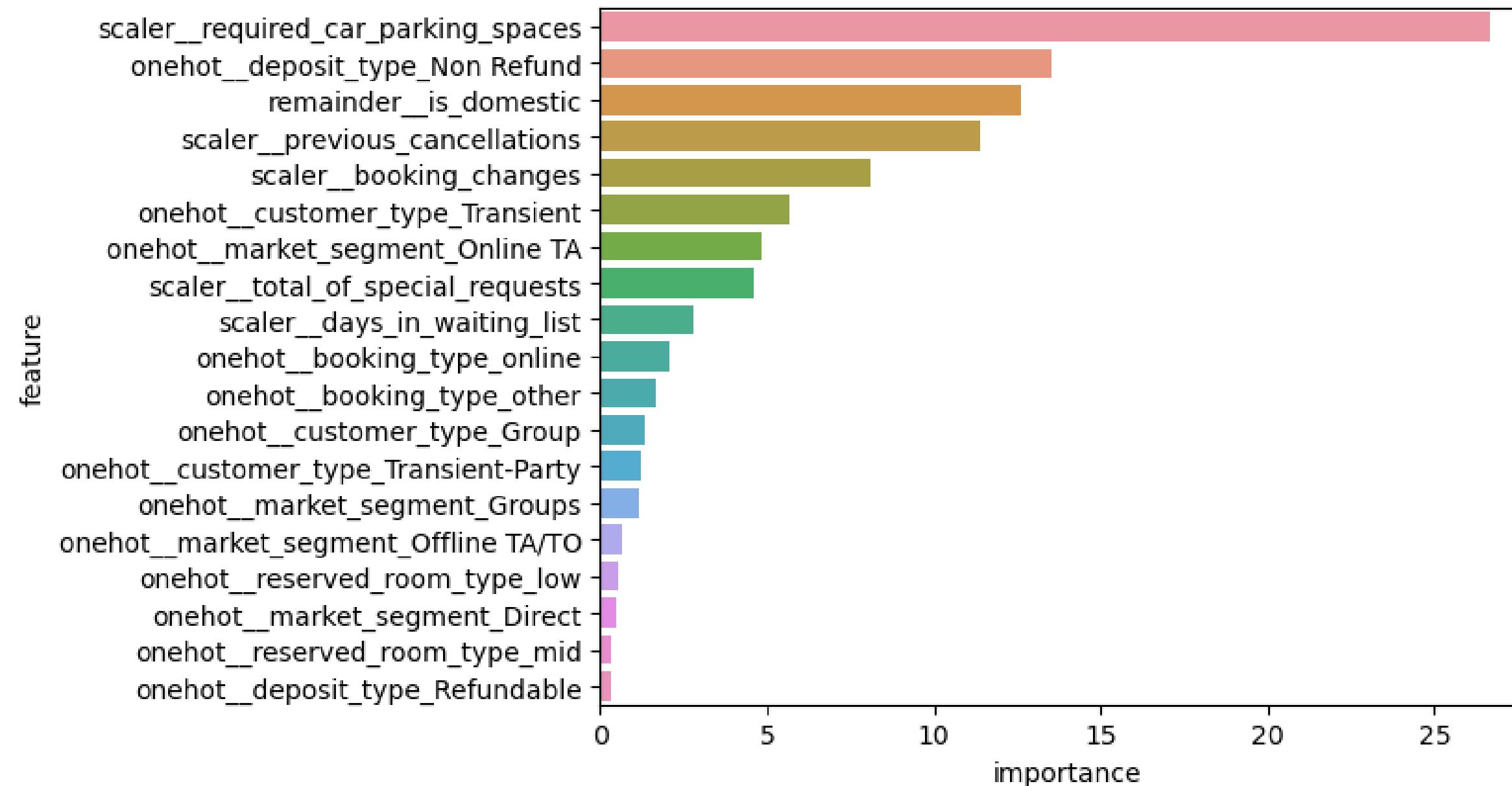


ROC AUC Score: 0.8888128978132808

Balanced Score: 0.8051467235687841

APA IMPACT NYA TERHADAP PERUSAHAAN?

RESULTS



RESULTS

CONTOH KASUS

Bila seandainya biaya untuk persiapan kamar hotel untuk calon pelanggan yang melakukan booking hotel adalah sebesar 25 EUR, dan andaikan jumlah calon pelanggan yang dimiliki untuk suatu kurun waktu sebanyak 500 orang (dimana andaikan 250 orang membatalkan booking, dan 250 tidak membatalkan booking. Maka perhitungannya kurang lebih akan seperti ini:

Total Biaya = (Jumlah Pelanggan Tidak Akan Membatalkan x Biaya Persiapan) + (Jumlah Pelanggan Akan Membatalkan x Biaya Persiapan)

Tanpa Model (Semua Kamar yang di booking akan dilakukan persiapan):

- Total biaya = $(250 \times 25 \text{ EUR}) + (250 \times 25 \text{ EUR}) = 12500 \text{ EUR}$
- Total calon pelanggan yang menginap = 250
- Total calon pelanggan yang membatalkan = 250
- Karena terdapat total calon pelanggan yang membatalkan sebesar 250, maka biaya persiapan akan terbuang sia sia.
- Total Kerugian = 6250 EUR

Dengan Model (Hanya kamar yang akan di tempati oleh calon pelanggan yang di persiapkan):

- Total biaya = $(210 \times 25 \text{ EUR}) + (58 \times 25 \text{ EUR}) = 5250 \text{ EUR} + 1450 \text{ EUR} = 6700 \text{ EUR}$
- Total calon pelanggan yang menginap = 210 (karena recall 1/yg tertarik itu $\rightarrow 250 \times 0.84$)
- Total calon pelanggan yang membatalkan = 40 (karena recall 1/yg tertarik itu 250×0.84)
- Biaya yang terbuang $\Rightarrow 58 \times 25 \text{ EUR} = 1450 \text{ EUR}$ (berdasarkan recall 0/yg tidak tertarik (58 orang $\rightarrow 0.23 \times 250$))
- Jumlah penghematan $\Rightarrow 193 \times 25 \text{ EUR} = 4825 \text{ EUR}$ (yang dihitung hanya yang memang membatalkan saja (193 orang $\rightarrow 0.77 \times 250$))

Dari perhitungan tersebut, terlihat bahwa dengan menggunakan model, perusahaan akan menghemat biaya persiapan kamar yang cukup besar. Dengan demikian, model tersebut dapat memberikan keuntungan bagi pihak hotel dalam mengelola bisnis mereka dengan lebih efektif dan efisien, sehingga dapat mengoptimalkan keuntungan dan meningkatkan kepuasan pelanggan.

KESIMPULAN DAN REKOMENDASI

BISNIS

- MEMASTIKAN KETERSEDIAAN TEMPAT PARKIR
- KEBIJAKAN DEPOSIT YANG JELAS
- MENINGKATKAN PENGALAMAN PELANGGAN
- KEBIJAKAN PEMBATALAN YANG FLEKSIBEL
- MENGURANGI PERUBAHAN RESERVASI
- DISKON KHUSUS
- UPGRADE KAMAR GRATIS
- POIN LOYALTY PROGRAM
- LAYANAN KHUSUS

MODEL

BERDASARKAN HASIL EVALUASI MODEL, ROC AUC SCORE YANG DIPEROLEH ADALAH 0.89, YANG MENUNJUKKAN BAHWA MODEL MEMILIKI KEMAMPUAN YANG BAIK UNTUK MEMBEDAKAN ANTARA KELAS POSITIF DAN NEGATIF. SEMENTARA ITU, BALANCED SCORE YANG DIPEROLEH ADALAH 0.805, MENUNJUKKAN BAHWA MODEL MAMPU MENGLASIFIKASIKAN DENGAN BAIK KELAS YANG TIDAK SEIMBANG.

SARAN YANG DAPAT DIBERIKAN ADALAH MELAKUKAN TUNING PADA MODEL UNTUK MENINGKATKAN PERFORMA PADA KELAS POSITIF DENGAN MEMPERTIMBANGKAN KEMBALI PEMILIHAN FITUR YANG DIGUNAKAN. SELAIN ITU, SEBAIKNYA DILAKUKAN PENGUMPULAN DATA LEBIH BANYAK UNTUK MENINGKATKAN AKURASI DAN PERFORMA MODEL.

TERIMA KASIH