



# MINIMIZING REVENUE LOSS: STRATEGIC PREDICTION OF HOTEL BOOKING CANCELLATIONS

P R E S E N T E D   B Y   A L I F S Y A   S A L A M



# PRESERVATION FLOW

1

## BACKGROUND

Gambaran umum konteks bisnis perhotelan, dampak finansial dari tingginya angka pembatalan, dan tujuan prediksi status pemesanan untuk meminimalkan kerugian pendapatan.

2

## EXPLORATORY DATA ANALYSIS

Pemahaman data, penanganan nilai yang hilang, dan analisis pola utama yang memengaruhi perilaku pembatalan seperti waktu tunggu, tipe deposit, dan segmen pasar.

3

## MODELING

Pengembangan model machine learning (Bagging Classifier) dengan Optimasi Threshold untuk memprioritaskan maksimalisasi profit dan mendeteksi potensi pembatalan secara akurat.

4

## CONCLUSION

Ringkasan kinerja model yang menonjolkan peningkatan metrik F2-score/Profit dan identifikasi faktor pendorong pembatalan paling berpengaruh menggunakan analisis SHAP.

5

## RECOMMENDATION

Tindakan strategis untuk memitigasi pembatalan, seperti penyesuaian kebijakan deposit untuk segmen berisiko tinggi dan penerapan strategi overbooking yang dinamis.



# BUSINESS BACKGROUND

objek bisnis dalam studi kasus ini merupakan sebuah hotel berbintang yang berlokasi di kawasan perkotaan di Portugal. Dalam menjalankan operasional reservasi, hotel menerapkan prinsip first-come, first-served, di mana kamar dialokasikan terlebih dahulu kepada tamu yang lebih dulu melakukan pemesanan dan konfirmasi.





# ANALYTICAL BUSINESS FRAMEWORK

## WHY MACHINE LEARNING?

Aturan-aturan tradisional tidak dapat menangkap pola non-linear yang kompleks (misalnya, interaksi antara waiting list, Deposit Type, dan Market Segment) dalam dataset yang sangat besar (~119 ribu baris).

Method: Supervised Learning (Classification)

### Project Goal & Target

Objective: Predict the booking status (`is_canceled`).

- Target Definition: 1 (Positive Class): Canceled or No-Show.
- 0 (Negative Class): Check-Out (Successful Stay).

Model Strategy:

- Ensemble Learning: Using Bagging Classifier to improve stability and reduce variance.
- Threshold Tuning: Adjusting the decision threshold (instead of default 0.5) to optimize sensitivity towards detecting cancellations.

### Evaluation Metric: F2-Score & Profit Analysis.

- Business Reasoning: Focus on Recall: We prioritize minimizing False Negatives (predicting a guest will come, but they cancel).
- Cost Impact: Kegagalan mendeteksi pembatalan mengakibatkan kamar kosong (Biaya Peluang/Opportunity Cost Tinggi), sedangkan Alarm Palsu (False Positive) hanya membutuhkan upaya verifikasi (Biaya Rendah).



# STAKEHOLDERS

1

## Front Office (Execution Level)

Goal: Seamless Guest Experience.

Value: Resource Planning & Mitigation.

2

## Revenue Manager (Strategic Level)

Goal: Maximize Revenue & Yield.

Value: Enables Dynamic Overbooking Strategy. Knowing which bookings are likely to cancel allows aggressive reselling without fear of empty rooms.

3

## Reservation Team (Operational Level)

Goal: Efficient Inventory Management.

Value: Prioritized booking control and Validation.



# BUSINESS PROBLEMS

## Inventory Spoilage (Spoiled Empty Room)

Loss of room revenue

Loss of potential additional revenue (F&B, hotel services)

Irrecoverable revenue leakage

## Inventory Spill (Opportunity Loss)

The hotel loses the opportunity to sell rooms to other guests

Potential higher prices (yield management) are not achieved

Market demand is not optimally utilized



# DATA UNDERSTANDING

## General Customer Data

**Country**

The country of origin of guests who make reservations.

**market\_segment**

Market segmentation/ordering channels

**previous\_cancellations**

Number of cancellations made by guests in the past

**booking\_changes**

Number of changes made to the order

**deposit\_type**

Deposit type (No Deposit, Non-Refundable, Refundable)

**days\_in\_waiting\_list**

The number of days for which reservations are on the waiting list

**customer\_type**

Customer type (Transient, Transient-Party, Contract, Group)

**reserved\_room\_type**

Room type code booked by the guest (e.g., A, D, E)

**required\_car\_parking\_spaces**

Number of parking spaces requested by guests

## additional General Customer Data

**total\_of\_special\_requests**

Number of special requests (extra pillows, \*view\*, etc.)

**is\_canceled**

Showing the final status of the order  
`0': Not canceled (Check-in) or `1': Canceled (Cancel)



# END-TO-END DATA PROCESSING

## 2 DATA PREPROCESSING

Data Profiling & Cleaning

Handling Duplicates

Feature Engineering

Cardinality Reduction

Scaling & Encoding

Fixed Data

- Column country: 351 NaN → Diisi "Other".
- Column market\_segment: Undefined → Diubah ke "Other".
- Mempertahankan semua baris (No dropping rows).

- Ditemukan: 73.371 data duplikat. Action: KEPT (Dipertahankan).
- Alasan: Tidak ada Unique ID; transaksi identik dianggap valid (tamu berbeda/rombongan).

- commitment\_score (Skor niat tamu).
- is\_repeat\_canceled (Flag pernah batal).
- is\_high\_risk (Deposit hangus + bukan korporat).
- commitment\_score (Skor Komitmen)
- booking\_stability (Stabilitas Booking)

Column country (162 negara) disederhanakan menjadi 3 grup: PRT (Portugal), Top International, & Other.

Kategori: One-Hot Encoding. Numerik: RobustScaler (karena banyak outliers pada days\_in\_waiting\_list).

Data siap dilatih menggunakan Bagging Classifier.

## 3 DATA ANALYSIS TOOLS



1



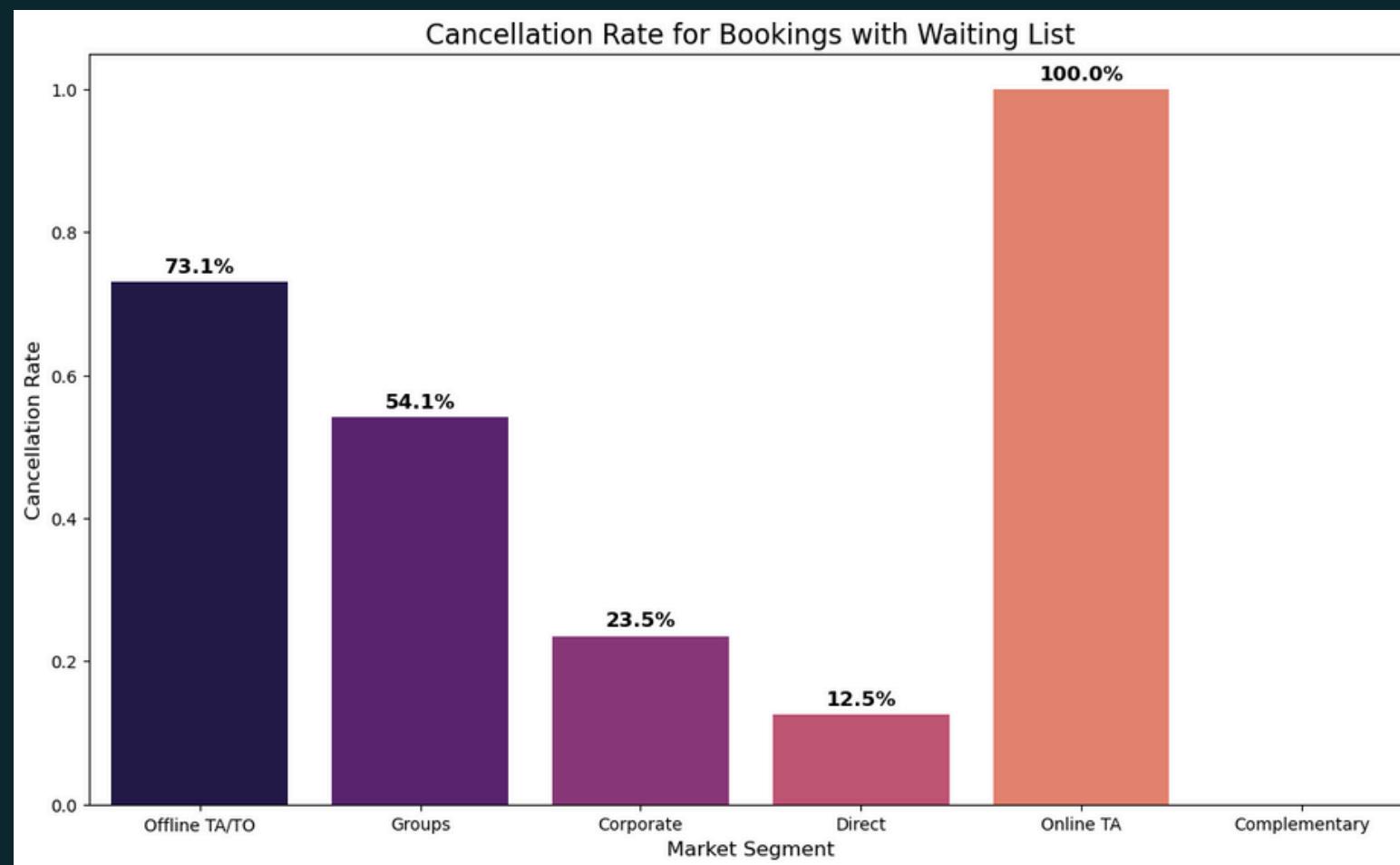
data\_hotel\_booking\_demand.csv  
Main Dataset

DATA COLLECTION  
(83.573 rows and 11 columns)



## WAITING LIST EFFICIENCY VARIES EXTREMELY BY CHANNEL

## ONLINE TA HAS THE HIGHEST CANCELLATION RATE ON WAITING LIST



## CRITICAL INSIGHT

100% tamu dari Online Travel Agents (OTA) yang masuk waiting list berakhir dengan pembatalan.

Sebaliknya ,segmen Direct (Pesan Langsung) adalah yang paling loyal dengan tingkat pembatalan terendah (12.5%).

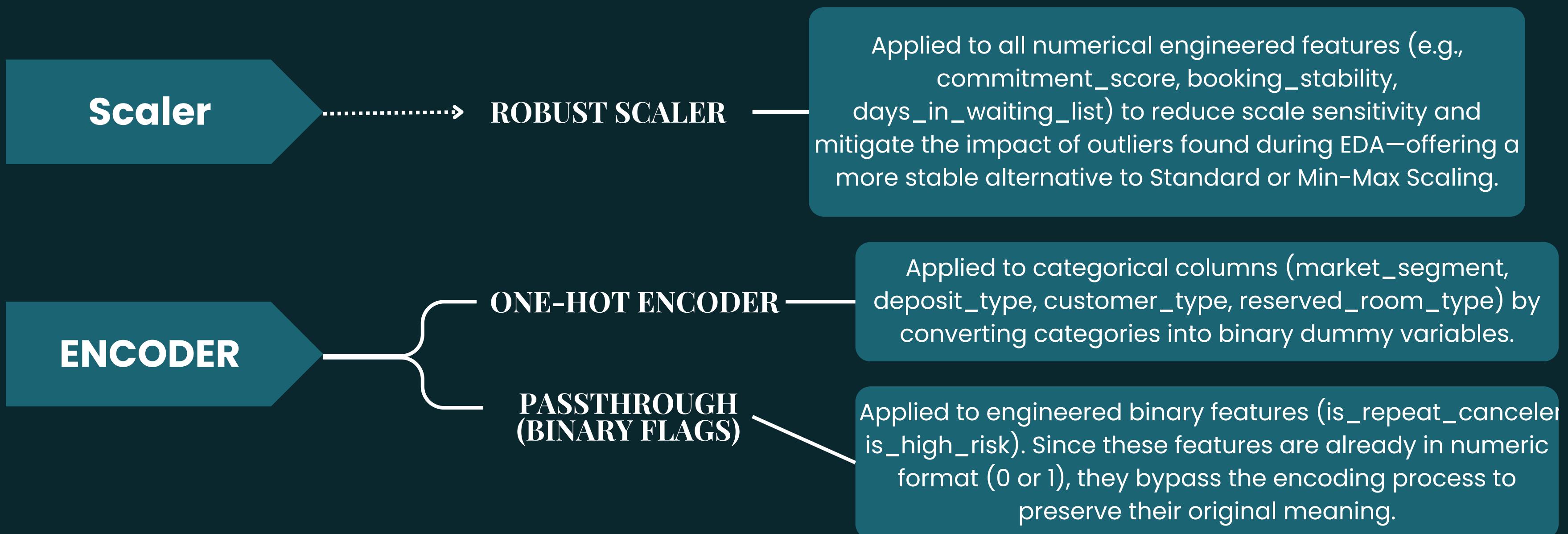
## STRATEGIC RECOMMENDATION

Hilangkan opsi waiting list untuk channel Online TA untuk mengurangi noise. Alihkan prioritas antrean kamar kosong kepada tamu Direct & Corporate karena mereka memiliki potensi konversi/jadi menginap paling tinggi.



# FEATURES TRANSFORMATION

Scaling Numerical Data & Encoding Categorical Variables





# BASE MODEL SUMMARY

RANK	F2 Mean Test Score	F2 Std Test Score	Model Category	Classification Method
1	0.711	0.005	Ensemble (Bagging)	Bagging Classifier
2	0.710	0.006	Ensemble (Boosting)	XGBClassifier
3	0.707	0.010	Ensemble (Bagging)	RandomForest Classifier
4	0.706	0.010	Tree Based	Decision Tree Classifier
5	0.681	0.004	Ensemble (Boosting)	Gradient Boosting



# HOW BAGGING CLASSIFIER WORKS

Bagging (Bootstrap Aggregating) adalah algoritma ensemble yang meningkatkan stabilitas dan akurasi algoritma machine learning (pembelajaran mesin) yang digunakan dalam klasifikasi dan regresi statistik. Algoritma ini mengurangi varians dan membantu menghindari overfitting.

Key Mechanics (The Process):

## 1. Bootstrapping (Pengambilan Sampel):

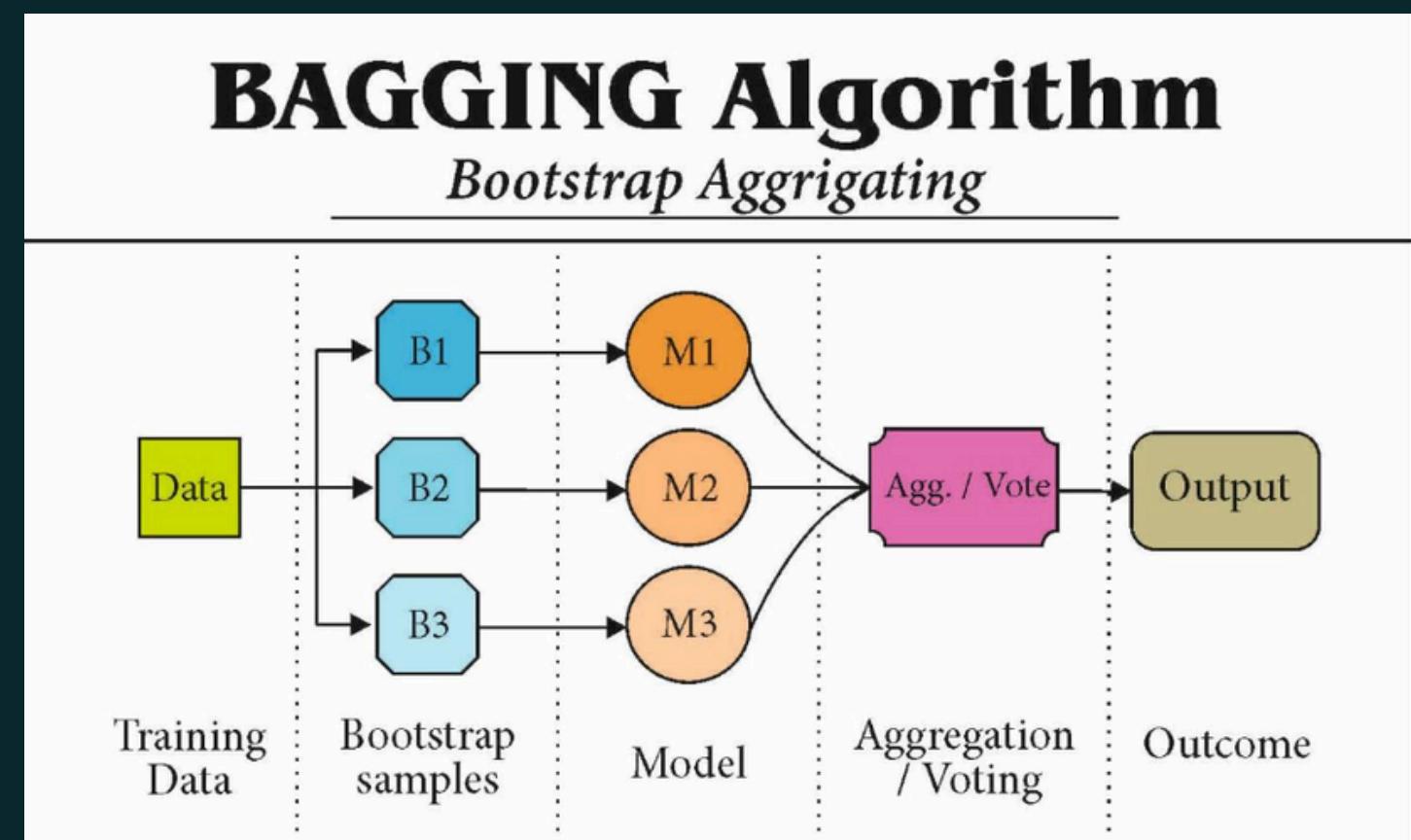
- Algoritma ini membuat beberapa subset (bagian data) dari data Hotel Booking asli dengan cara mengambil sampel secara acak dengan pengembalian (sampling with replacement).

## 2. Parallel Training (Pelatihan Paralel):

- Sebuah Decision Tree (Estimator Dasar) yang terpisah dilatih secara independen pada setiap subset acak tersebut. Tidak seperti pohon tunggal pada umumnya, pohon-pohon ini dibiarkan tumbuh sedalam mungkin.

## 3. Aggregation (Agregasi/Voting):

- Prediksi akhir dibuat dengan menggabungkan prediksi dari seluruh pohon individu. Untuk masalah klasifikasi seperti ini, algoritma menggunakan Majority Voting (Suara Terbanyak).





# MODEL LIMITATIONS OF BAGGING CLASSIFIER

## 1. Ketiadaan Variabel Eksternal

- Dataset saat ini hanya memuat catatan pemesanan internal (data CRM).
- Keterbatasan: Faktor-faktor eksternal penting yang menyebabkan pembatalan tidak tersedia, seperti Harga Pesaing, Kondisi Cuaca, Pembatalan Penerbangan, dan Acara Lokal.

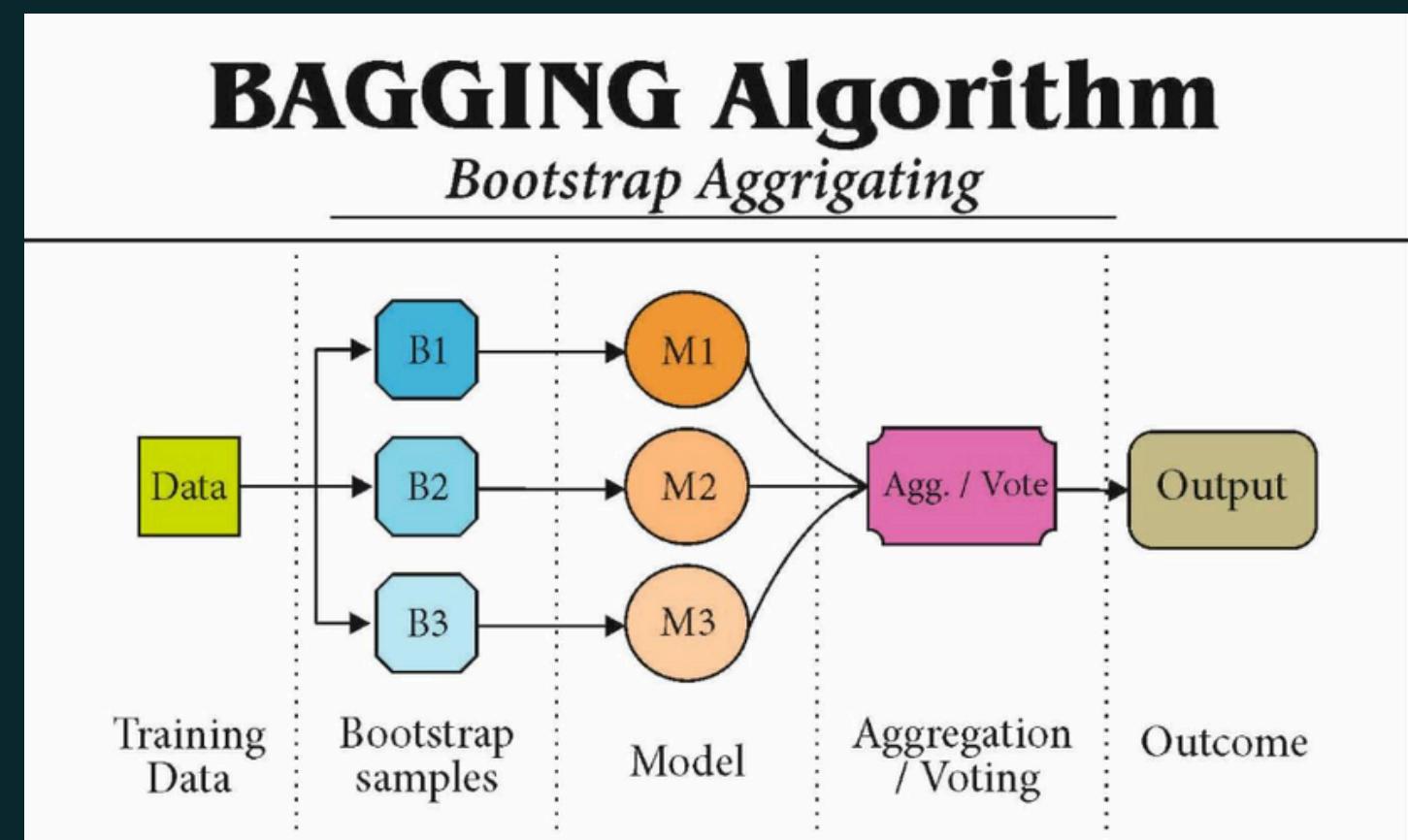
## 2. Kompleksitas Komputasi (Black Box)

Bagging Classifier melibatkan pelatihan banyak Decision Tree (Pohon Keputusan) yang dalam secara paralel.

- Keterbatasan: Secara komputasi, model ini lebih berat dibandingkan model sederhana (seperti Regresi Logistik) untuk inferensi real-time dan membutuhkan alat yang kompleks (seperti SHAP) untuk diinterpretasikan, tidak seperti model "White Box".

## 3. Batas Kinerja (F2-Score)

- Meskipun mencapai F2-Score yang baik (0.71), model ini masih menghasilkan False Negatives (Negatif Palsu).
- Keterbatasan: Beberapa pelanggan "Berisiko Tinggi" mungkin masih terlewat (gagal dideteksi), yang berarti hotel masih bisa menghadapi pembatalan tak terduga meskipun sudah ada prediksi dari model.





# Konfigurasi Final Model (HYPERPARAMETER TUNING)

## Efisiensi Pencarian (RandomizedSearchCV):

- Menguji 50 kombinasi acak (`n_iter=50`) dari ratusan kemungkinan untuk menemukan konfigurasi optimal secara cepat tanpa mengorbankan kualitas.

## Penanganan Imbalance Data (`class_weight='balanced'`):

- Secara eksplisit menguji Base Learner (Decision Tree) dengan bobot Balanced. Ini memaksa model untuk "memberi perhatian lebih" pada data minoritas (Tamu Batal).

## Optimasi Struktur Ensemble (Bagging Parameters):

- Mengatur kompleksitas model (`max_depth`, `n_estimators`, `max_samples`) untuk mencegah overfitting sekaligus menjaga stabilitas prediksi.

## Metrik Bisnis Kritis (`scoring=f2_scorer`):

- Poin Terpenting: Model tidak dinilai berdasarkan Akurasi, tapi F2 Score. F2 Score memberi bobot lebih besar pada Recall, memastikan prioritas model adalah menangkap tamu yang batal (Meminimalkan False Negative).

```
param_space = {  
    "modeling_n_estimators": [50, 100, 150, 200],      # 4 values  
    "modeling_max_samples": [0.5, 0.7, 0.8, 1.0],      # 4 values  
    "modeling_max_features": [0.5, 0.7, 0.8, 1.0],     # 4 values  
    "modeling_bootstrap": [True],                      # 1 value (fix)  
    "modeling_bootstrap_features": [True, False],        # 2 values  
    # Ganti nested estimator params dengan estimator langsung  
    "modeling_estimator": [  
        DecisionTreeClassifier(max_depth=5, class_weight='balanced', random_state=0),  
        DecisionTreeClassifier(max_depth=10, class_weight='balanced', random_state=0),  
        DecisionTreeClassifier(max_depth=15, class_weight='balanced', random_state=0),  
        DecisionTreeClassifier(max_depth=10, class_weight=None, random_state=0),  
    ]  
  
randomsearch = RandomizedSearchCV(  
    estimator=pipe_base,  
    param_distributions=param_space,  
    n_iter=50,  
    cv=3,  
    scoring=f2_scorer,  
    n_jobs=-1,  
    verbose=2,  
    random_state=0  
)  
  
randomsearch
```



# Konfigurasi Final Model (THRESHOLD OPTIMIZATION)

- Masalah Dasar (The Problem):

Secara default, model Machine Learning menggunakan batas 0.50 (50%) untuk memprediksi "Batal". Ini terlalu pasif/konservatif untuk bisnis hotel kita di mana biaya Kamar Kosong (200) jauh lebih mahal daripada biaya \*Overbooking\* (\75).

- Strategi Solusi (The Method):

Kami menggunakan teknik TunedThresholdClassifierCV untuk mencari titik potong probabilitas terbaik secara otomatis. Optimasi ini tidak mengejar Akurasi, melainkan F2-Score (yang memberi bobot 2x pada Recall) agar model fokus menangkap pembatalan.

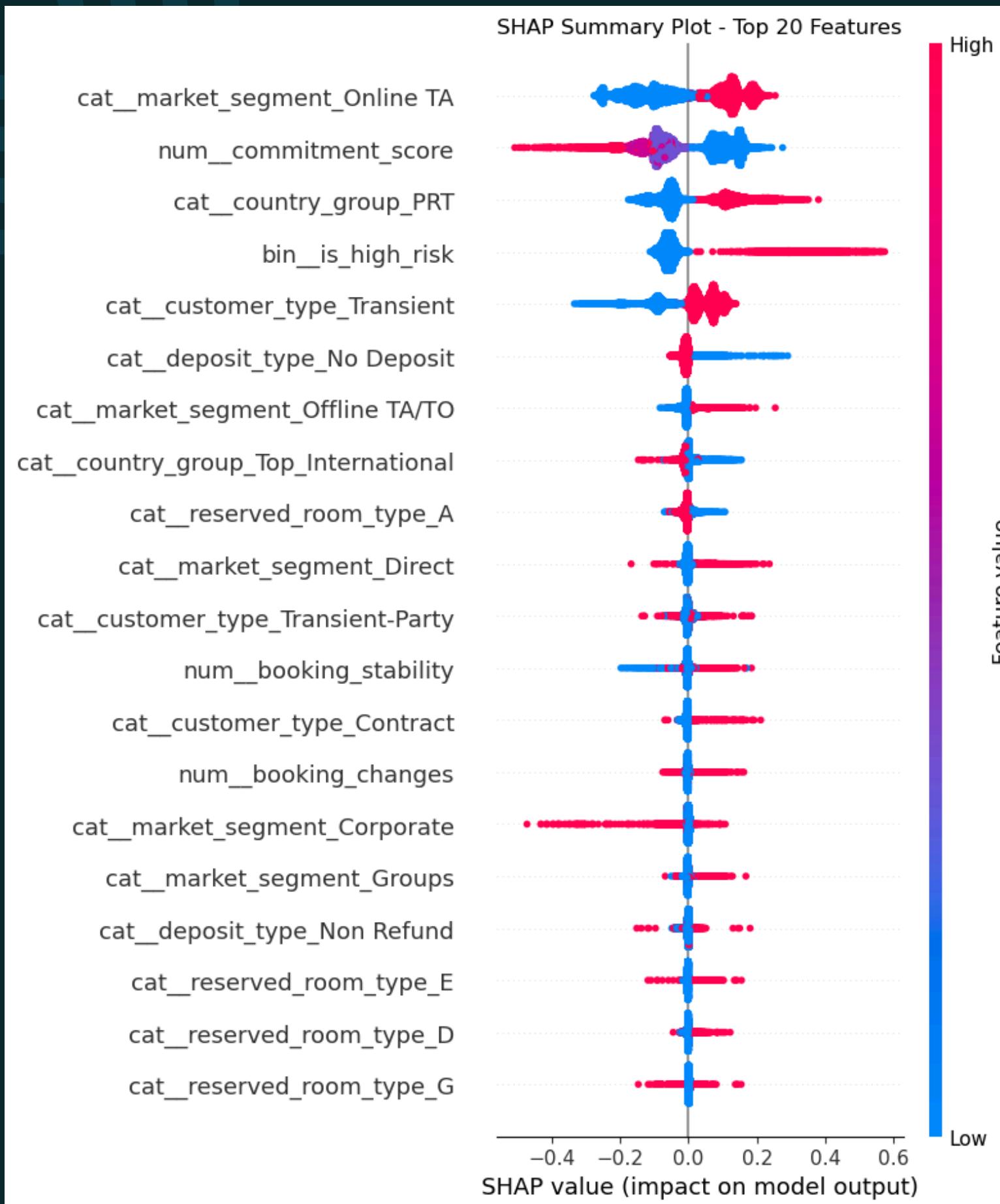
- Hasil Penyesuaian (The Result):

Ambang batas keputusan digeser drastis dari 0.50 menjadi 0.27. Artinya: Jika model mendeteksi probabilitas batal hanya 27% saja, sistem sudah akan menandainya sebagai "Prediksi Batal".

```
threshold_optimized_model = TunedThresholdClassifierCV(  
    estimator=pipe_tuned,  
    scoring=f2_scorer,  
    cv=5  
)  
# best estimator from RandomizedSearchCV  
# F2 emphasizes Recall > Precision  
# robust threshold via cross-validation
```



# MODEL EXPLAINABILITY: KEY DRIVERS OF CANCELLATION (SHAP ANALYSIS)



## SHAP (SHapley Additive exPlanations)

- Since the Bagging Classifier is a black-box model (non-interpretable), SHAP is used to explain the predictions
- SHAP provides the contribution value of each feature to individual predictions

### 1. Market Segment: Online TA (Peringkat 1)

- Observasi: Titik Merah (Booking via Online Travel Agent) menyebar di sebelah kanan (SHAP positif).

### 2. Commitment Score (Peringkat 2 - Engineered Feature)

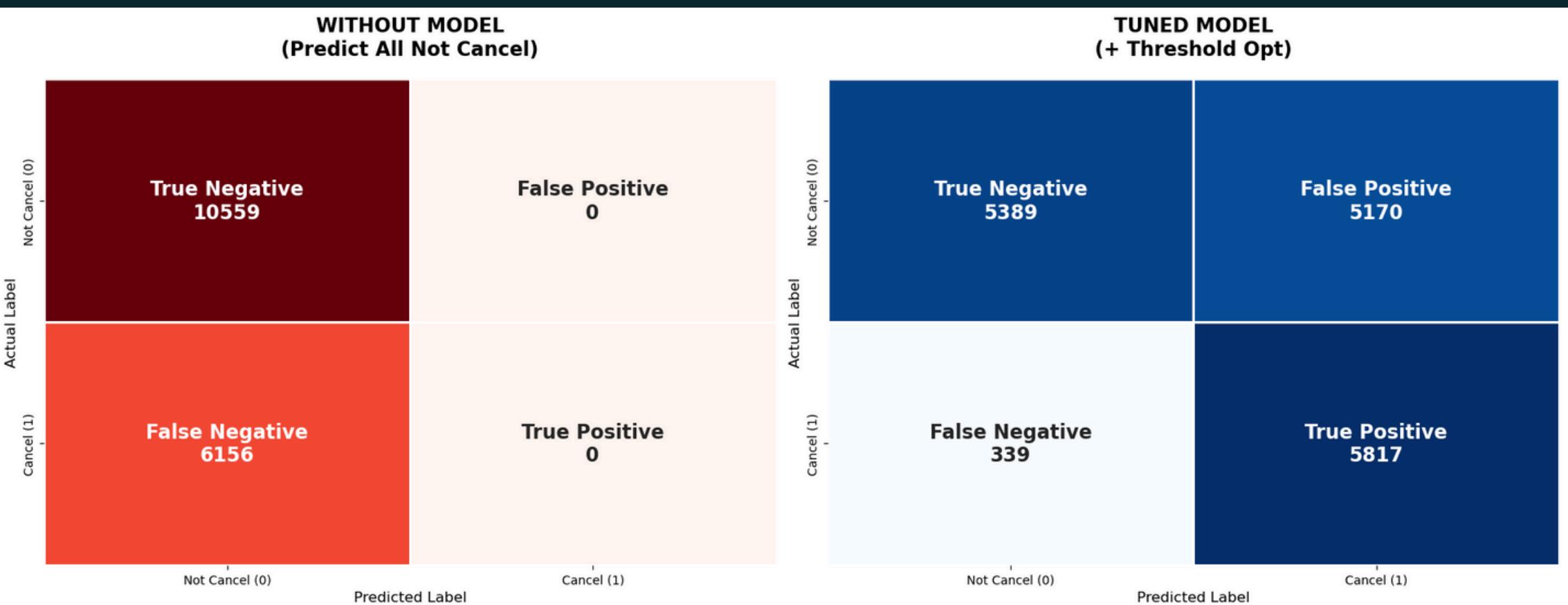
- Observasi: Titik Merah (Skor Komitmen Tinggi) berkumpul di Kiri Jauh. Titik Biru (Skor Komitmen Rendah) ada di Kanan.

### 3. Country Group: PRT / Portugal (Peringkat 3)

- Observasi: Titik Merah (Tamu dari Portugal) menumpuk di sebelah Kanan.



# BUSINESS OUTCOME EVALUATION



## Masalah Utama (Grafik Kiri):

Tanpa model, hotel "buta" terhadap pembatalan. Akibatnya, terdapat 6.156 kasus False Negative (tamu batal tapi kamar dibiarkan kosong), yang menciptakan kerugian pendapatan (Revenue Leakage) sebesar \$1.2 Juta.

## Solusi Model (Grafik Kanan):

Model difokuskan untuk menangkap pembatalan (High Recall). Hasilnya, False Negative turun drastis dari 6.156 menjadi 339. Sebanyak 5.817 kamar (True Positive) berhasil diselamatkan dan bisa dijual kembali ke pelanggan lain.

$$\begin{aligned} \text{Loss} &= (\text{False Negative} \times \$200) + (\text{False Positive} \times \$75) \\ \text{Loss} &= (6,156 \times \$200) + (0 \times \$75) \\ \text{Total Loss} &= \$1,231,200 \end{aligned}$$

$$\begin{aligned} \text{Loss} &= (339 \times \$200) + (5,170 \times \$75) \\ \text{Loss} &= \$67,800 + \$387,750 \\ \text{Total Loss} &= \$455,550 \end{aligned}$$

$$\begin{aligned} &\text{(Total Saving)} \\ &\text{Saving} = \$1,231,200 - \$455,550 = \$775,650 \\ &\text{(Model menghemat } \sim 63\% \text{ dari potensi kerugian)} \end{aligned}$$



# CONCLUSION

## Konteks Masalah (Problem Statement)

Tingkat pembatalan pemesanan (Cancellation Rate) sangat tinggi mencapai 37%, yang menjadi sinyal bahaya bagi manajemen karena menyebabkan hilangnya potensi pendapatan (revenue leakage) akibat kamar kosong yang terlambat dijual kembali.

## Pengaruh Fitur (Feature Importance)

Faktor paling dominan yang mempengaruhi keputusan pembatalan tamu adalah Commitment Score (skor interaksi permintaan khusus & parkir), Tipe Deposit (Non-Refund vs No Deposit), dan Riwayat Pembatalan (High Risk Flag).

## Performa Model (Model Performance)

Model terbaik yang digunakan adalah Bagging Classifier dengan Threshold Tuning.

- F2 Score (Test Set): 0.7300
- Recall (Test Set): 0.7315
- (Model difokuskan untuk meminimalisir False Negative agar tidak kehilangan peluang jual kamar).

## Dampak Bisnis (Business Impact)

Implementasi Machine Learning berhasil menurunkan risiko kerugian finansial secara signifikan, dengan estimasi penghematan sebesar ~\$771,825 (mengurangi sekitar 63% dari total potensi kerugian) dibandingkan strategi tanpa model.



# RECOMMENDATION

## Terapkan Aggressive Overbooking di High Season

Saat permintaan tinggi (harga kamar >\$150), gunakan prediksi model untuk melakukan overbooking secara agresif. Kerugian membiarkan kamar kosong (False Negative) terhitung 2,7x lebih mahal dibandingkan biaya kompensasi memindahkan tamu (False Positive).

## Integrasi "Soft Confirmation" via CRM

Gunakan hasil prediksi risiko tinggi (High Risk) untuk memicu pengiriman pesan otomatis (WhatsApp/Email) kepada tamu 3-7 hari sebelum kedatangan. Konfirmasi ulang rencana mereka sebelum memutuskan untuk menjual kembali kamar tersebut.

## Strategi Konservatif di Low Season

Saat harga kamar rendah (<\$80), naikkan ambang batas prediksi (threshold) agar model lebih berhati-hati. Hindari risiko membayar biaya kompensasi tamu (walk cost) yang sia-sia saat margin keuntungan hotel sedang tipis.

## Proteksi Segmen Korporat & VIP (Whitelist)

Kecualikan tamu dari segmen Korporat dan VIP dari sistem overbooking otomatis. Risiko kerusakan reputasi dan potensi hilangnya kontrak jangka panjang (Customer Lifetime Value) terlalu besar untuk dikorbankan demi optimasi jangka pendek.



SHODWE  
HOTEL

Home

About

Contact



# THANK YOU

W H E R E   C O M F O R T   M E E T S   L U X U R Y