A MACHINE LEARNING APPROACH TO CANCELLATION PREDICTION

Optimizing Hotel Revenue

Melin Ayu Safitri

January 13, 2025

AGENDA

01
BUSINESS UNDERSTANDING
Understand the business context, problems, and goals to predict hotel booking cancellations.

02
DATA CLEANING
Clean the data by addressing issues such as missing values, duplicates, or anomalies to ensure data quality.

03
EXPLORATORY DATA ANALYSIS
Identify patterns, relationships, and key factors affecting booking cancellations.

04
DATA PREPARATION
Prepare the data for modeling, including transformations, encoding, and handling data imbalance.

05
MODEL DEVELOPMENT
Build and evaluate predictive models to determine the best-performing one based on evaluation criteria.

06
MODEL IMPLEMENTATION
Use the model to support business decision-making and improve efficiency and revenue.

07
CONCLUSION & RECOMMENDATION
Summarize the analysis results and provide strategic recommendations to reduce cancellations and boost revenue.

01

BUSINESS UNDERSTANDING

CONTEXT : The success of a hotel depends on its ability to maximize room occupancy and operational efficiency.

High reservation cancellations pose a challenge !!

WHAT IS THE PROBLEM?

- **Financial Loss:** Last-minute cancellations result in empty rooms that could have been sold to other customers.
- **Operational Uncertainty:** It is difficult to plan for customer needs, inventory, and other resources when cancellation rates are unpredictable.

Goals

- Hotels predict cancellations to minimize risks with strategies like overbooking or last-minute deals.
- Hotels identify cancellation factors to adjust policies and offer promotions for at-risk customers.

Analytical Approach

- Analyze data to identify patterns and relationships that distinguish canceled bookings from non-canceled ones.
- Build a predictive model to forecast cancellations, helping hotels implement proactive strategies and optimize operations.

WHAT EVALUATION METRIC ARE USED?

Balancing False Positif Rate and True Positif Rate with ROC-AUC to Minimize Financial Losses and Improve Cancellation Rate Management.

BUSINESS METRICS

- **Cancellation Rate:** Measures the overall frequency of cancellations.
- **Revenue Recovery Rate:** Indicates how effectively the hotel mitigates losses from cancellations.

MACHINE LEARNING EVALUATION METRICS

- **False Positive (Type I Error):** Predicted to be canceled but is actually not.
- **False Negative (Type II Error):** Predicted not to be canceled but actually gets canceled.
- **Consequence:** Lost revenue due to rooms that remain empty and cannot be rebooked, resulting in financial loss for the hotel.

So, we need to reduce the financial loss caused by false negatives **(cancellations that the model fails to detect)**, while avoiding unnecessary actions arising from false positives **(overbooking)**.

Why use Machine Learning?

More Accurate For Prediction

- Predict customer cancellation probabilities with precision.
- Detect hidden patterns in booking behavior.
- Provide real-time prediction updates for better decision-making.

Identification of Risk Factors

- Analyze numerous variables, such as booking and customer details.
- Rank factors influencing cancellations by importance.
- Offer actionable insights to adjust overbooking strategies.

BUSINESS BENEFITS

- **Cost** ↓ Reduce revenue loss from undetected cancellations.
- **Efficiency** ↑ Optimize resource planning and operations.
- **Profit** ↑ Increase revenue recovery through smarter overbooking strategies.

Data Overview

Feature	Impact to Business
country	Optimizes regional marketing
market segment	Helps with pricing and cost control
previous cancellation	Highlights need for better policies
booking canges	Leads to inefficiencies and higher costs
deposit type	Secures revenue or builds trust
day in waiting list	Can lead to lost bookings
Customer Type	Enables tailored pricing and strategies
Reserved Room Type	Prevents overbooking and mismatched expectations
Required Parking Space	Ensures adequate parking for guests
Total Spesial Request	Increases satisfaction, adds complexity
Is Cancelled	Aids in forecasting and policy adjustment

- Total Entries: 83,573 rows
- Country just 83,222 rows (351 missing values).

- Most customers are from Portugal, showing the domestic market's dominance.
- The OTA market is key, emphasizing the importance of online bookings for the hotel.

02

DATA CLEANING

The data has 0.42% missing values, 87.51% duplicates, and 0.07% anomalies.

HOW TO HANDLING?

Missing Values:
With only **0.42%** missing values, drop the rows ensures data consistency without significant impact.

Anomalies:
Has 7 anomaly represent only **0.07%** of the data, making their removal a minor adjustment to improve model accuracy.

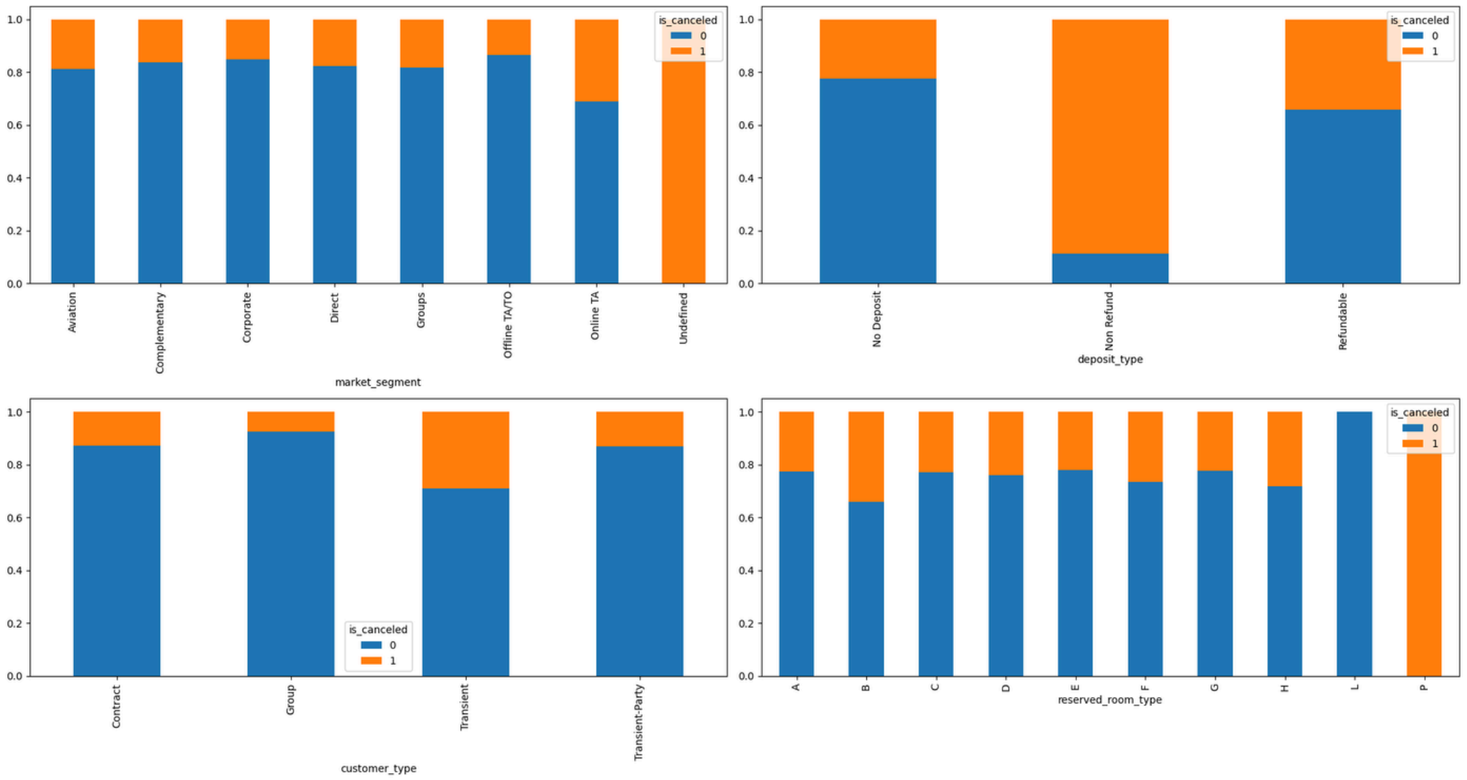
Duplicates:
Dropping **87.51%** duplicates is essential to avoid redundancy, ensure data accuracy, and prevent biased analysis.

Spelling Error:
The fact that no spelling errors were found indicates that the dataset is well-maintained, with clean and reliable text entries.

03

EXPLORATORY DATA ANALYSIS

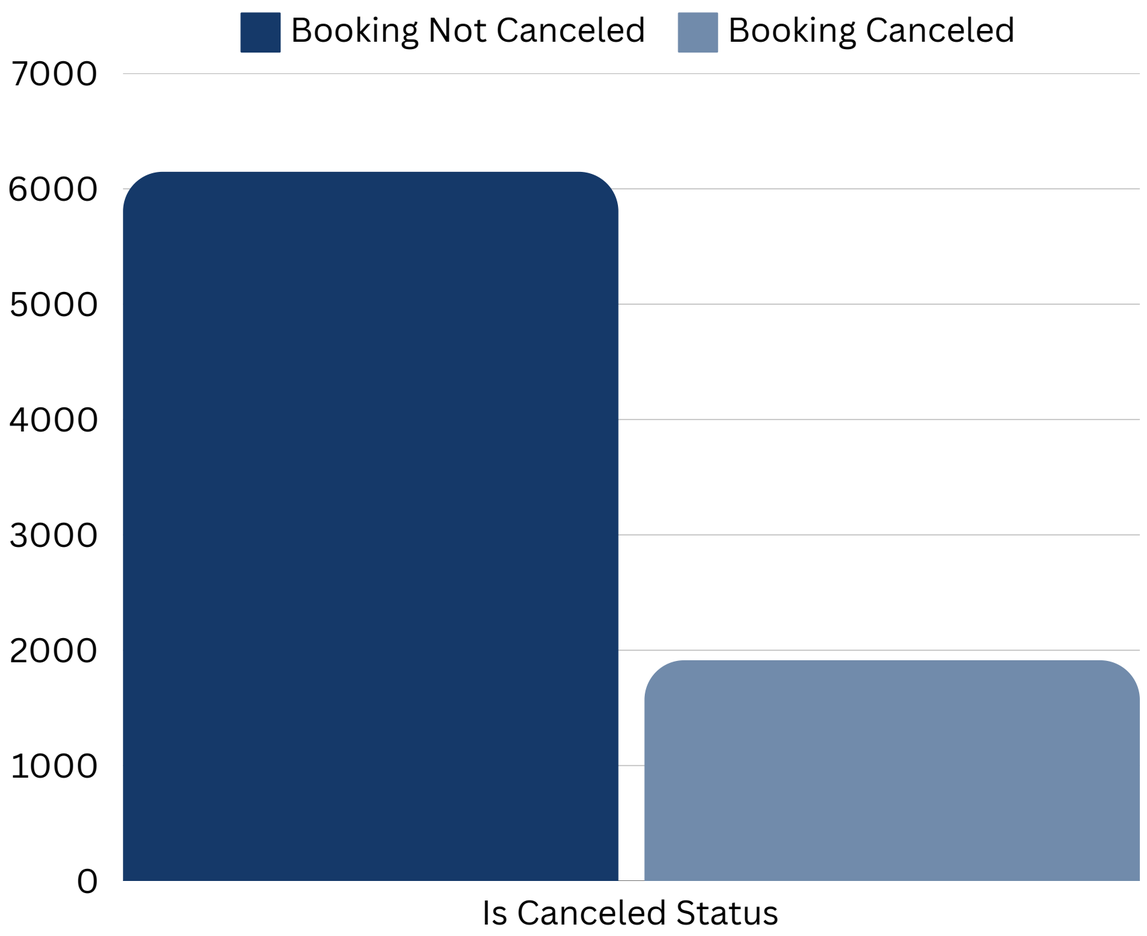
Visualization of Booking Cancellations Based on Various Factors



- Focus can be given to retention efforts for high-cancellation segments.

- Certain segments, deposits, customer types, and room types have higher cancellations.

Understanding Cancellation Patterns in Hotel Bookings



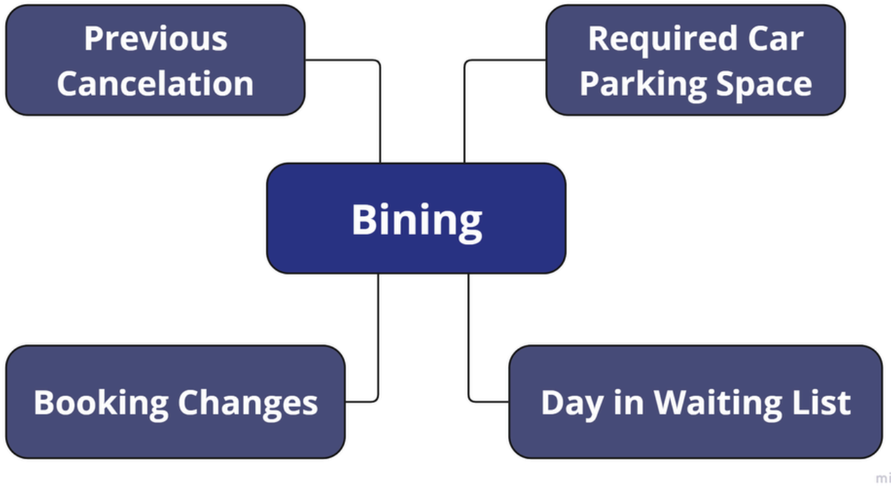
Higher count of Booking Not Canceled suggests strong guest retention.

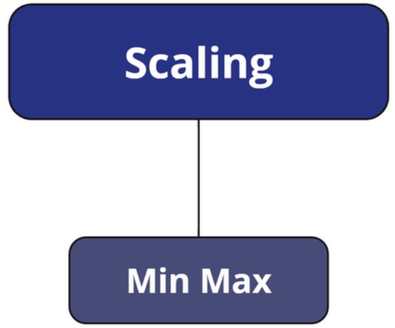
Analyzing contributing factors can help optimize booking policies, customer service, and operations.

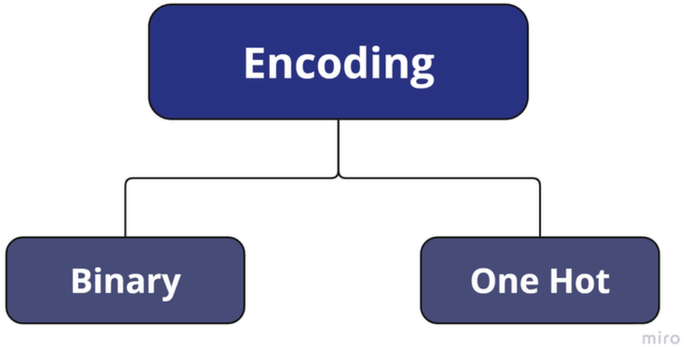
04

DATA PREPARATION

Data Transformation for Analysis







01 Bining

Splits each column into two categories to address the highly skewed

02 Scaling

Adjusts data to a distribution for consistency in machine learning models.

03 Encoding

Converting categorical data into numerical format

05

MODEL DEVELOPMENT

Benchmarking

No	List Model
1	**Logistic Regression**
2	**KNeighbors Classifier**
3	**Desicion Tree Classifier**
4	**Random Forest Classifier**
5	**XGB Classifier**
6	**LGBM Classifier**
7	**Gradient Boosting Classifier**
8	**Ada Boosting Classifier**
9	**Support Vector Classifier**

Resampling

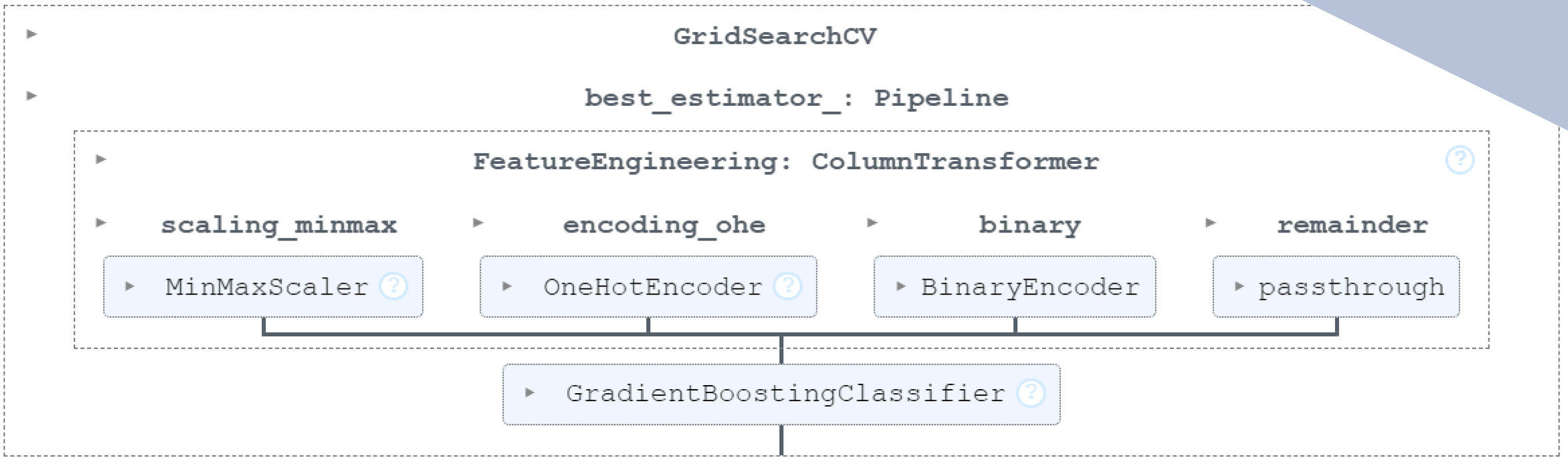
Oversampling

- SMOTE
- Random Over Sampler

Undersampling

- NEARMISS
- Random Under Sampler

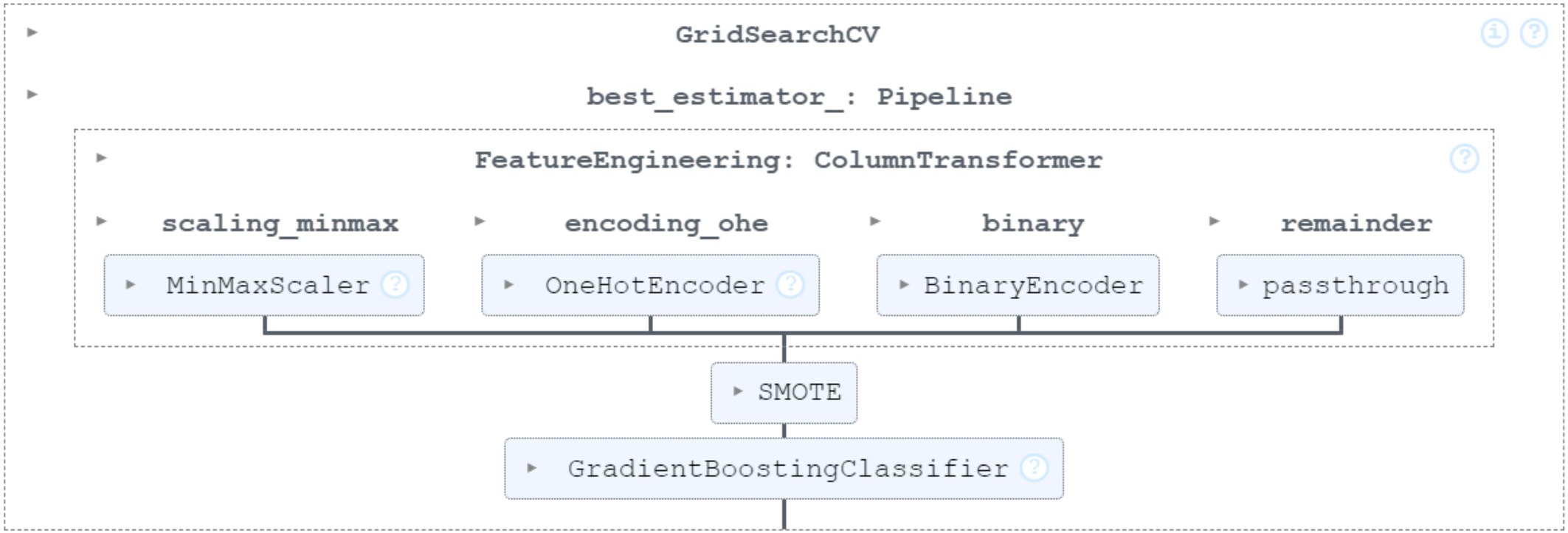
Best Model Without Resampling



- The **GradientBoostingClassifier** (GBC) has emerged as the **best-performing** model in this benchmarking process.

- **ROC-AUC Score:** 0.8289
- **Cross-validation** using **StratifiedKFold**.

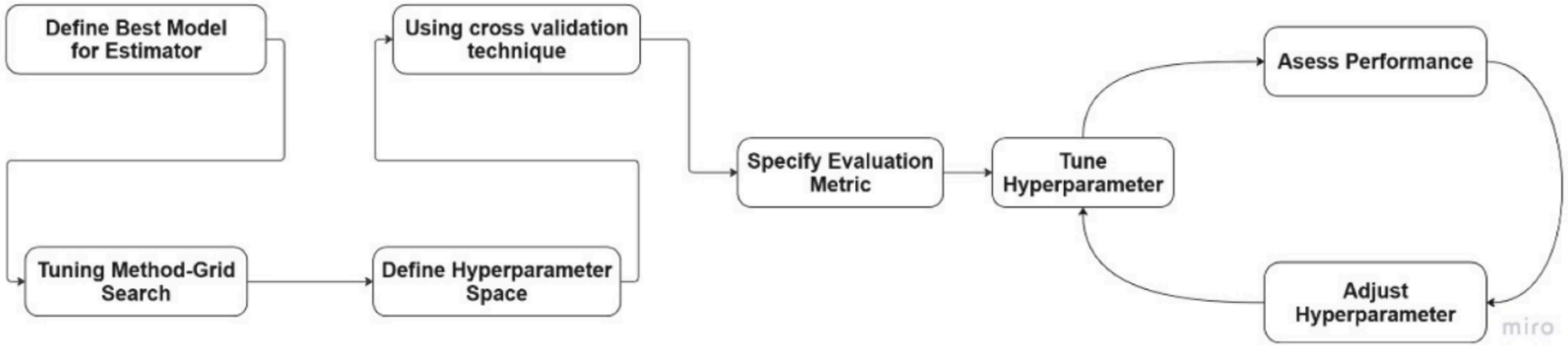
Best Model With Resampling



- The **GradientBoostingClassifier** continues to perform as the best model based on the results of fitting and evaluation.

- **ROC-AUC Score:** 0.829
- **Best Resampling :** SMOTE

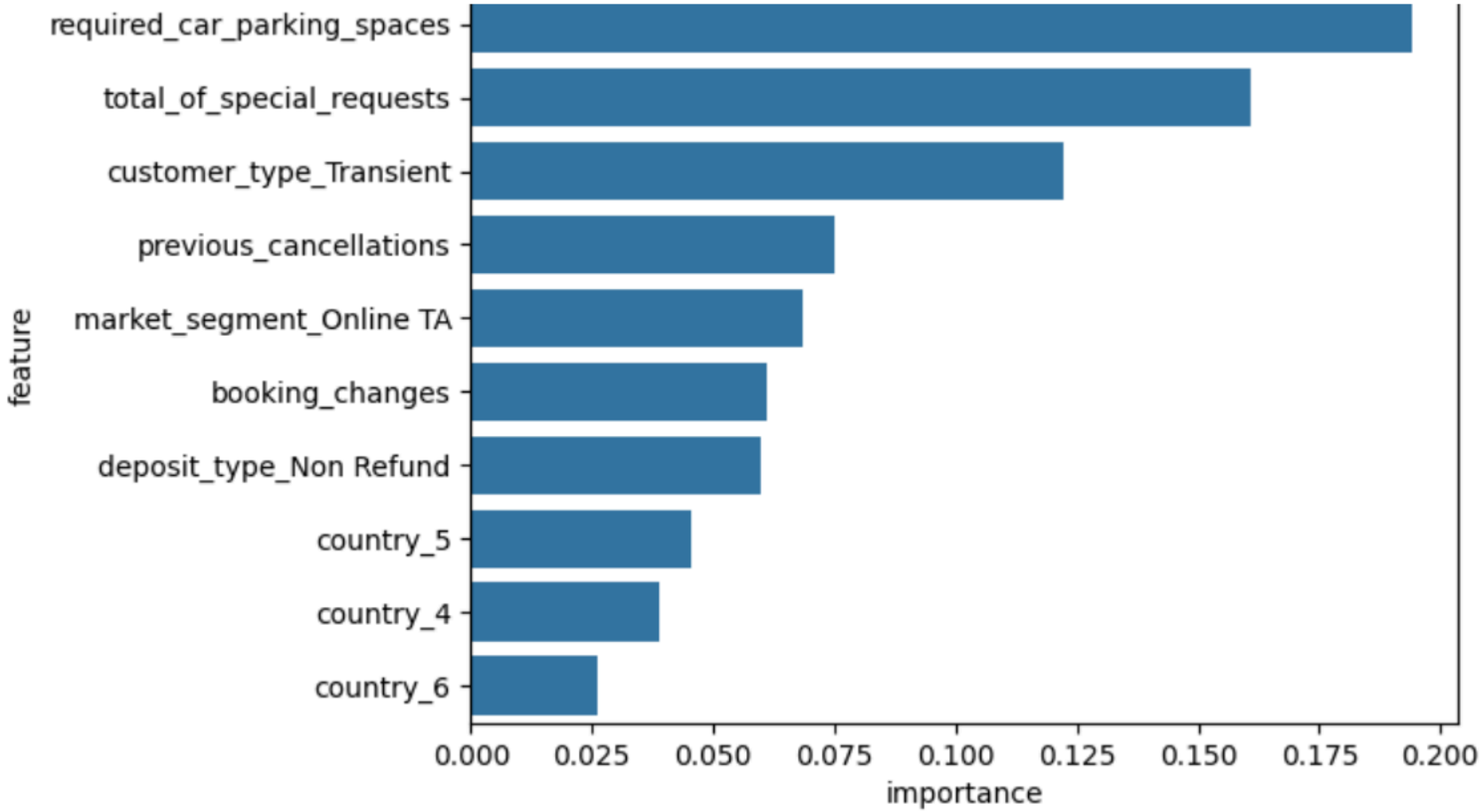
Hyperparameter Tuning



Tuning Result

	Training Dataset	Testing Dataset
Before Tuning	**0.85**	**0.831**
After Tuning	**0.86**	**0.835**

Feature Importances

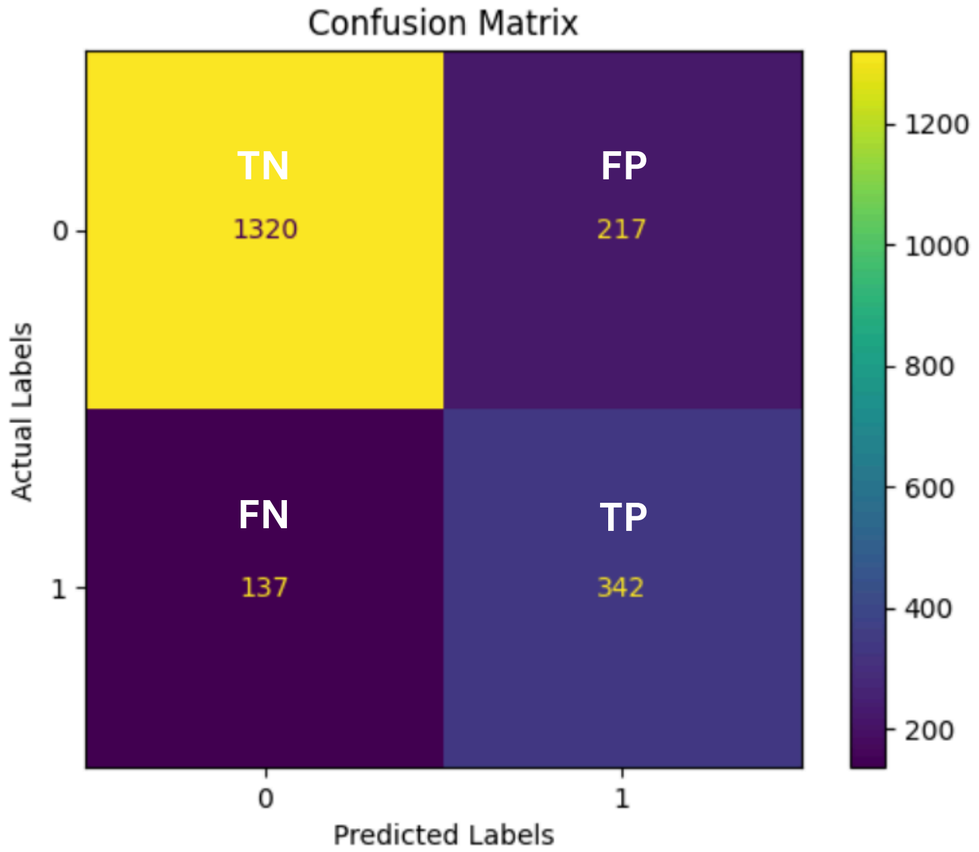


required_car_parking_spaces appears to be **the most important feature**, suggesting that the need for car parking significantly impacts the model's predictions.

06

MODEL IMPLEMENTATION

Business Calculation with Simulation



Assumptions

average_room_revenue = 1500000

cost_of_overbooking = 750000

Results :

1. Loss with Model: IDR 60M (overbooking cost offset by recovery).
2. Loss Without Model: IDR 205M (full revenue loss from cancellations).
3. Savings with Model: IDR 145M (reduced losses via overbooking strategies).

07

CONCLUSIONS & RECOMMENDATION

Conclusions

Machine Learning Model

- Revenue Recovery Rate (RRR) reached 12.24%, exceeding the 10% target.
- Losses from undetected cancellations (FN) dropped significantly from IDR 205M to IDR 60M.
- Total savings of IDR 145M, showcasing the model's effectiveness in managing cancellations and overbooking.

Business

- Financial gains include IDR 145M saved from reduced cancellations.
- RRR exceeding the 10% target confirms improved revenue recovery.
- Model-driven strategies enhance operational efficiency and balance cancellation risks with overbooking.

Recommendation

Machine Learning Model

- Reassess the 0.5 probability threshold to better balance overbooking risks and optimize predictions.

Business

- **Enhance Overbooking Strategy:** Use the model to refine strategies, factoring in seasonal trends and events.
- **Flexible Cancellation Policies:** Introduce reminders or incentives for high-risk customers to reduce cancellations.
- **Boost Revenue Recovery:** Offer last-minute deals or promotions to fill canceled rooms effectively.

THANK YOU