

# Pestana CR7 Hotel Madeira Portugal



DATA SET INTRODUCTION

*Portugal*





# Overview



**01**

Bussiness Understanding

**02**

Data Understanding

**03**

Data Cleaning

**04**

Modelling

**05**

Success Metric

**06**

Conclusion and  
Recommendation

• • •

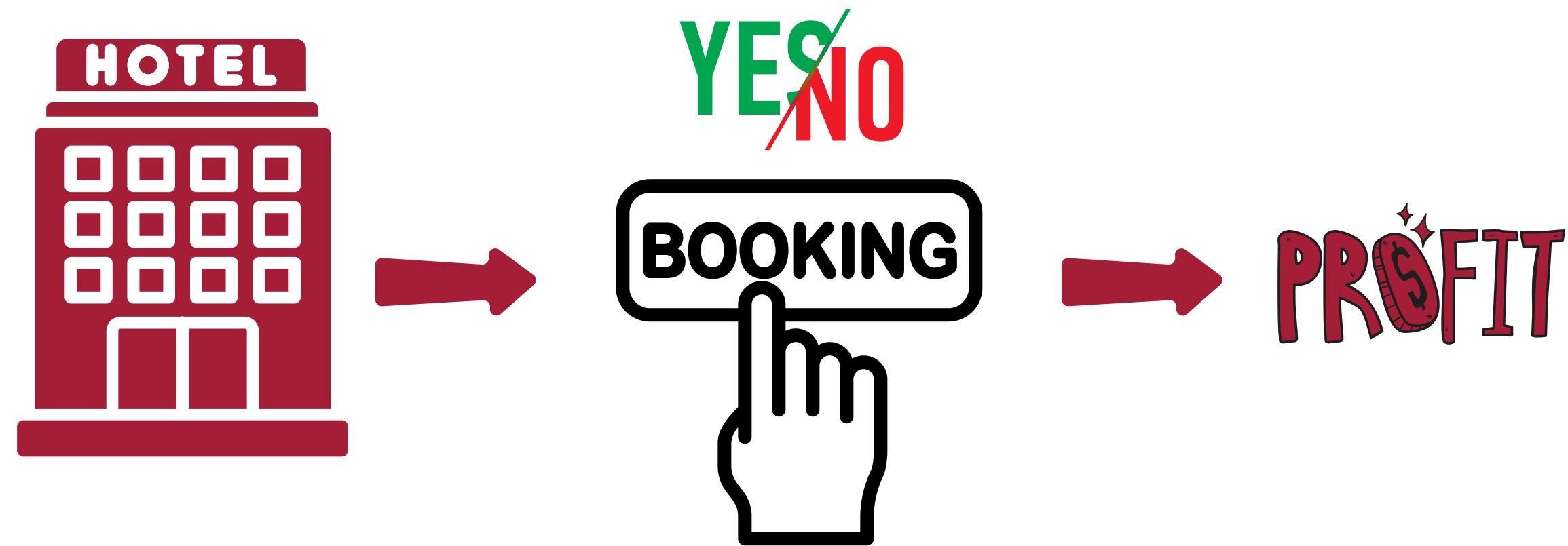


# 1. Business Understanding



- 
- 
-

# Background



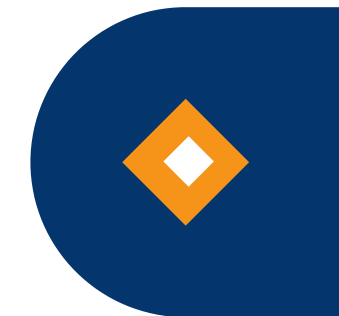
According to the Hotel Bed **Occupancy Rate** data in Portugal for 2023, the hotel occupancy rate in Portugal reached only **48%** throughout the year 2023.

# False Prediction

Aktual/Prediksi	Not Canceled(0)	Canceled(1)
Not Canceled(0)	True Negative Prediction: Not Canceled Actual: Not Canceled	False Positive Prediction: Canceled Actual: Not Canceled
Canceled(1)	False Negative Prediction: Not Canceled Actual: Canceled	True Positif Prediksi: Cancel Aktual: Cancel

**Type 1 error:** False Positive (FP): The model predicts a hotel booking will be canceled, but it turns out the booking is not canceled.

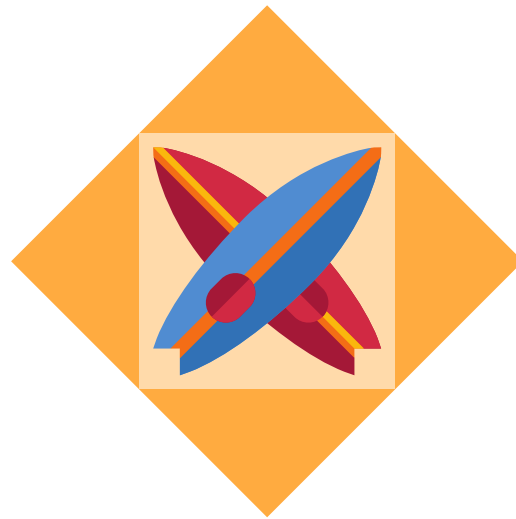
**Type 2 error:** False Negative (FN): The model predicts that a hotel booking will not be canceled, but it turns out the booking is canceled.



## Recall

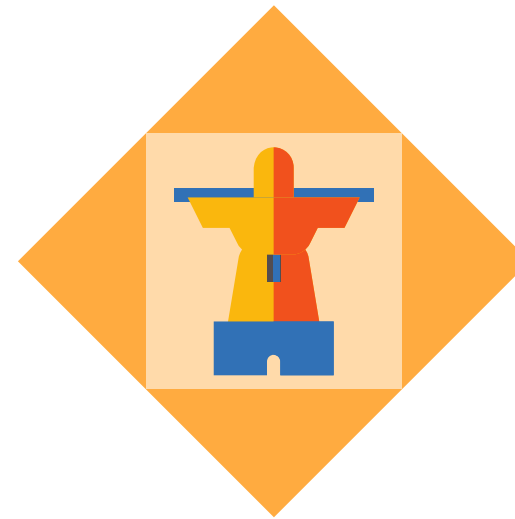
From the false predictions, it is known that for this booking case, False Negatives (FN) cause **more direct losses** than False Positives (FP). This is because when the model fails to predict a cancellation (FN), it leads **to lost revenue or missed opportunities**.





## Problem Statement

Booking cancellations have a direct impact on hotel profitability because canceled rooms cannot be resold in a short period of time.



## Goals

Optimizing Reservation Management and Maximizing Profitability



## Analytical Approach

Using the recall metric to minimize the number of people who will cancel their bookings.

# • 2.Data • Understanding





## Data Cleaning

- Handling Missing Value
- Handling Outlier
- Handling Duplicate Data





## Data Preprocessing

- Encoding
- Scaling
- Balancing



# Data Features



1. **country**: Country of origin. (Kategori)
  2. **market\_segment**: Market segment designation. (Kategori)
  3. **previous\_cancellations**: Number of previous bookings that were cancelled by the customer prior to the current booking. (Numerik)
  4. **booking\_changes**: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation. (Numerik)
  5. **deposit\_type**: Indication on if the customer made a deposit to guarantee the booking. (Kategori)
  6. **days\_in\_waiting\_list**: Number of days the booking was in the waiting list before it was confirmed to the customer. (Numerik)
  7. **customer\_type**: Type of booking. (Kategori)
  8. **reserved\_room\_type**: Code of room type reserved. Code is presented instead of designation for anonymity reasons. (Kategori)
  9. **required\_car\_parking\_space**: Number of car parking spaces required by the customer. (Numerik)
  10. **total\_of\_special\_request**: Number of special requests made by the customer (e.g. twin bed or high floor). (Numerik)
- 
- 



# Data Target



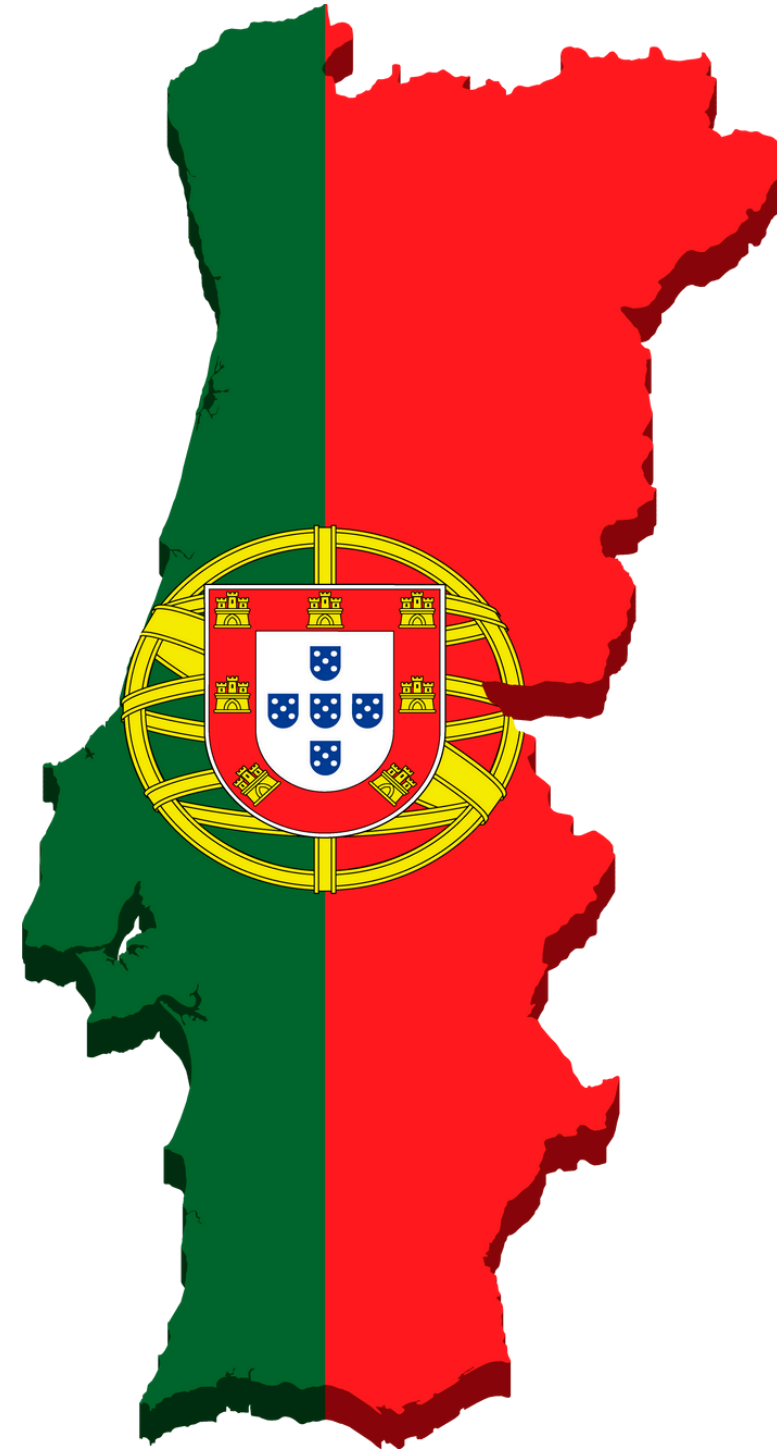
`is_canceled`: Value indicating if the booking was canceled (1) or not (0).

**There are 83,573 rows with 10 feature columns and 1 target column**





# 3.Data Cleaning





# Missing Value

Missing values are only present in the **country** column.

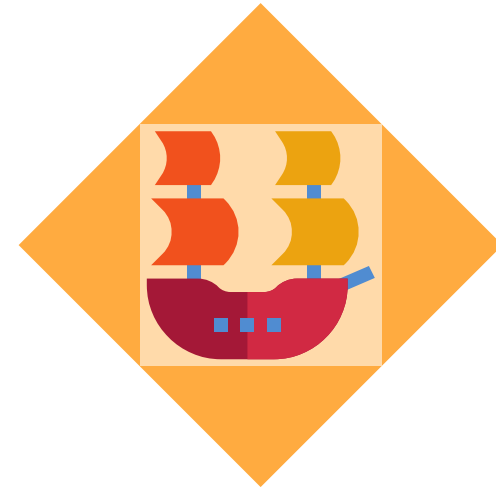


## Drop

Missing values were dropped because the amount of missing data was minimal, and thus, **it did not have a significant impact.**

country	118
market_segment	0
previous_cancellations	0
booking_changes	0
deposit_type	0
days_in_waiting_list	0
customer_type	0
reserved_room_type	0
required_car_parking_spaces	0
total_of_special_requests	0
is_canceled	0
dtype:	int64





## Outlier

It is still considered reasonable and relevant in the context of the analysis



## Duplicate Data

**73371** duplicates  
(**87.79%** of the data)



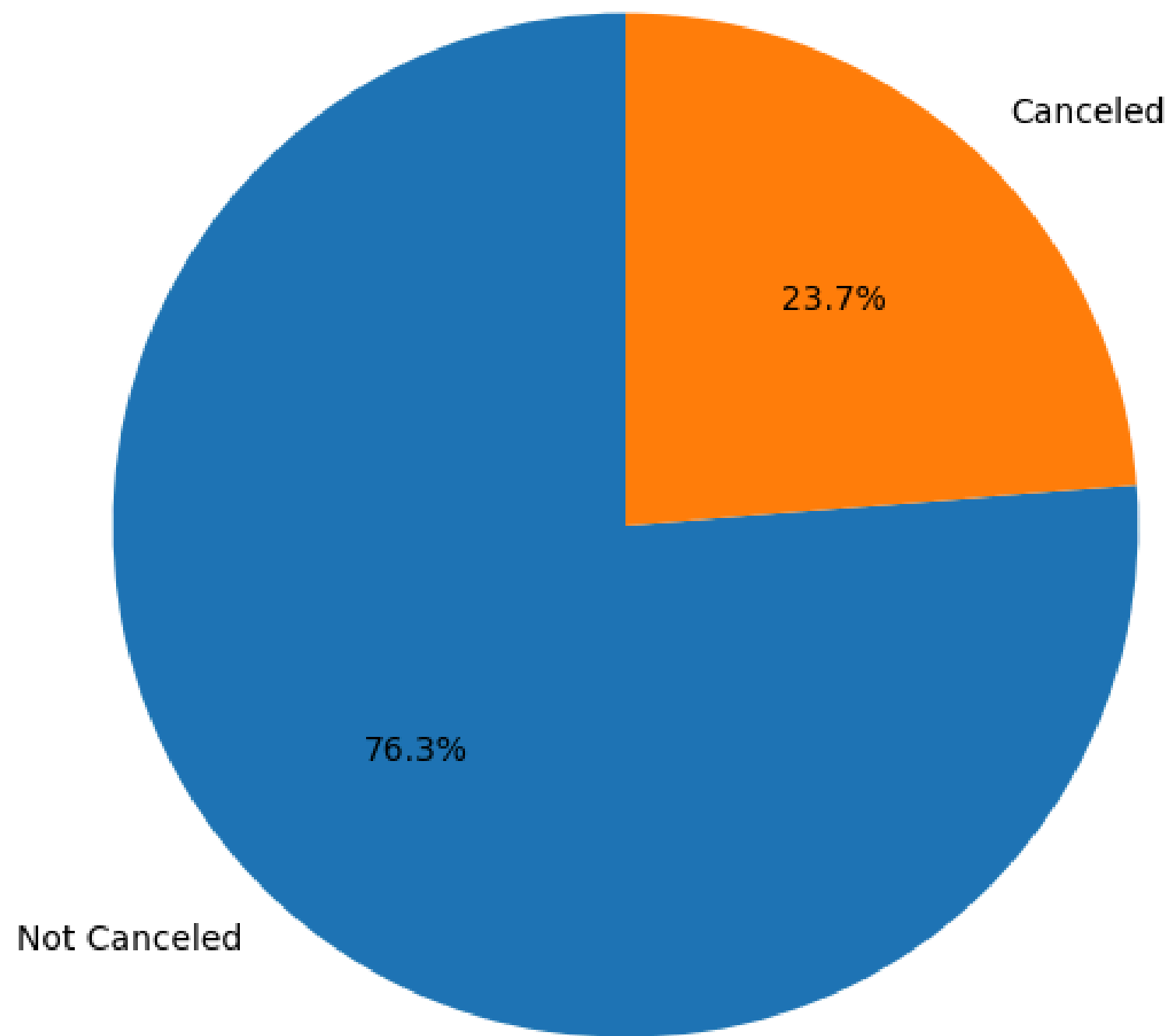
**Drop**







Percentage of Booking Cancellations

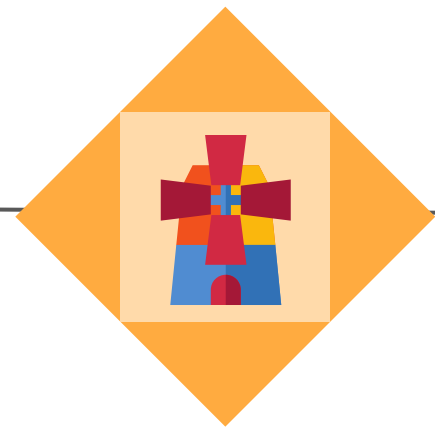


**Imbalanced Target Data**



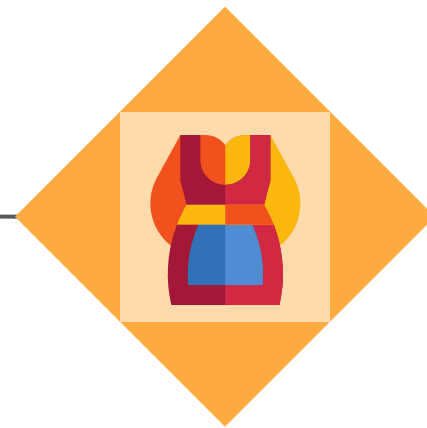
# Data Preprocessing

Encoding



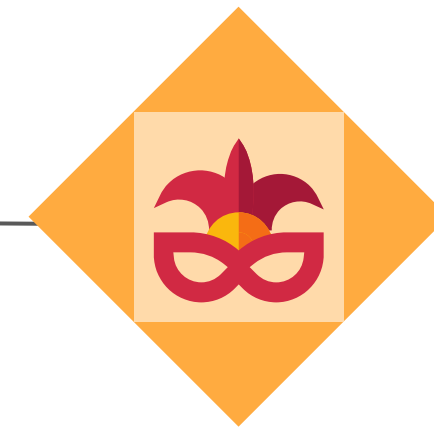
This dataset uses two types of encoding: one-hot encoding and binary encoding.

Scaling



Scaling using robust scaler

Balancing



Using resampler to balancing data

# 4. Modeling



# Model Benchmark and Evaluation

## Benchmark

The model achieved the **highest recall value** on both the training and test sets.

The best model was obtained using **LightGBM** and the **Random Under Sampling resampler**.

The higher the **recall score**, the better the prediction model.

Performing **hyperparameter** tuning to get best model.

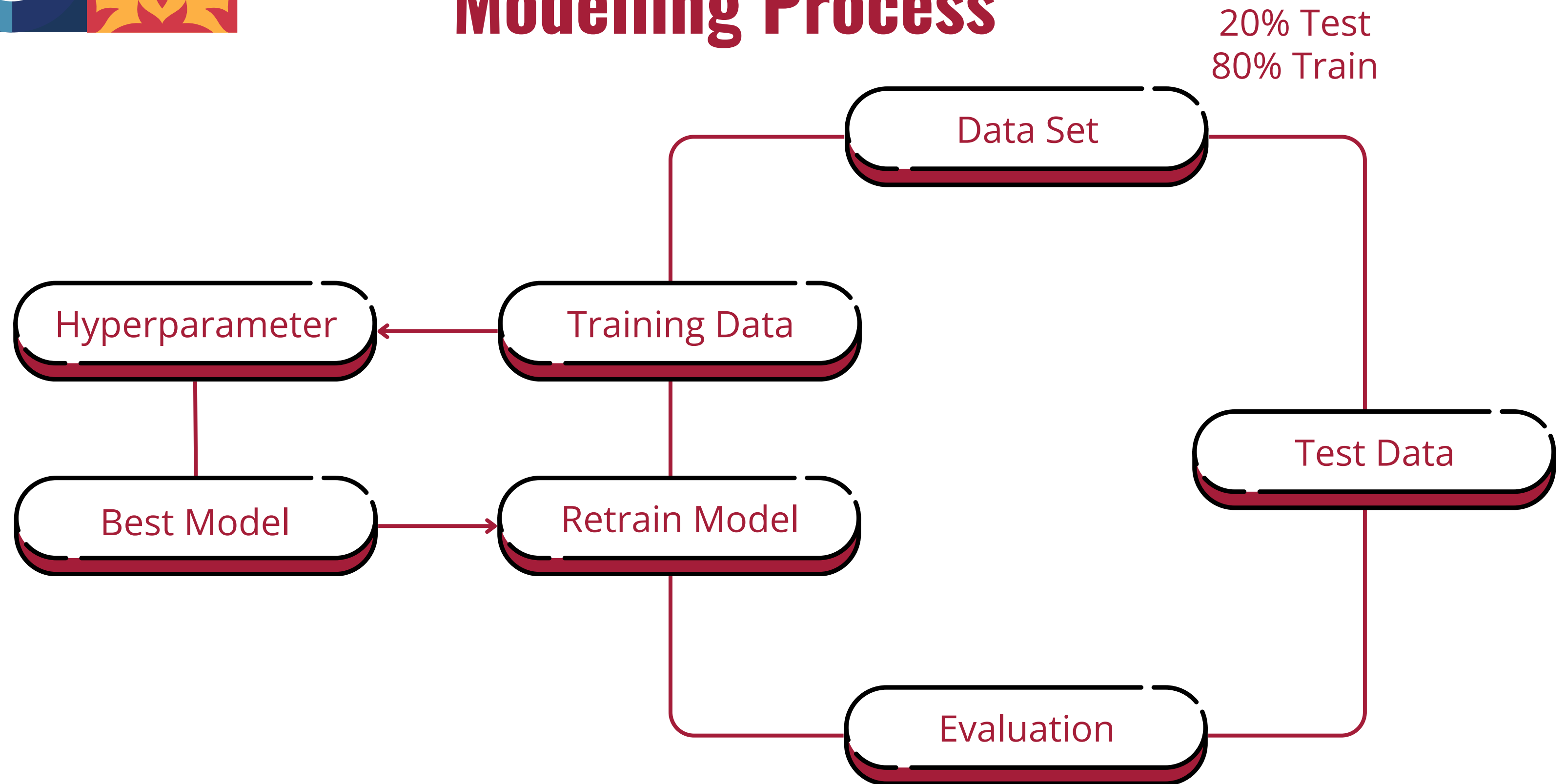


# Hyperparameter

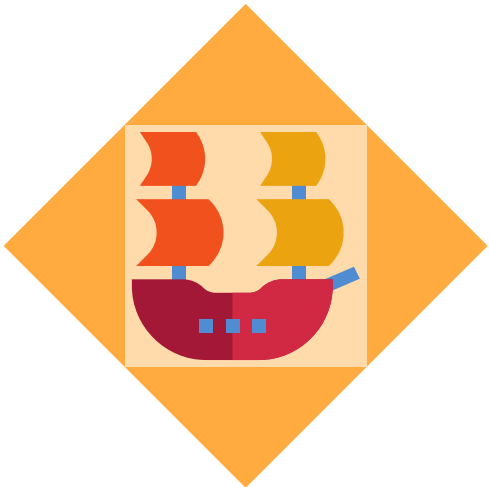
- Looking best combination with Cross Validation
- Using Random Search Sampling



# Modelling Process



# Best Model: LightGBM



Training Score

82.46 %



Test Score

87,89

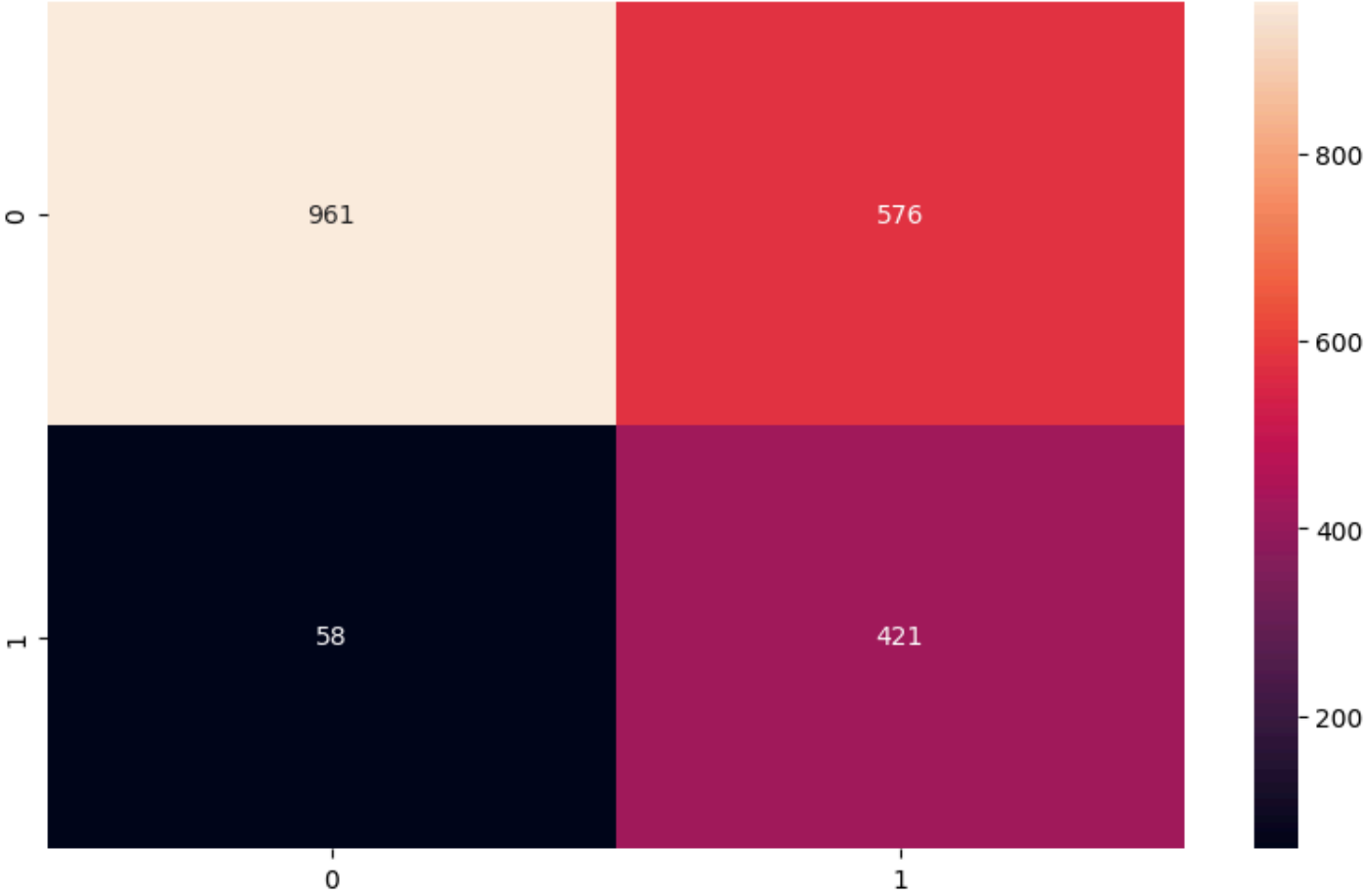
Before Tuned

	precision	recall	f1-score	support
0	0.92	0.66	0.77	1537
1	0.43	0.82	0.57	479
accuracy			0.70	2016
macro avg	0.68	0.74	0.67	2016
weighted avg	0.81	0.70	0.72	2016

After Tuned

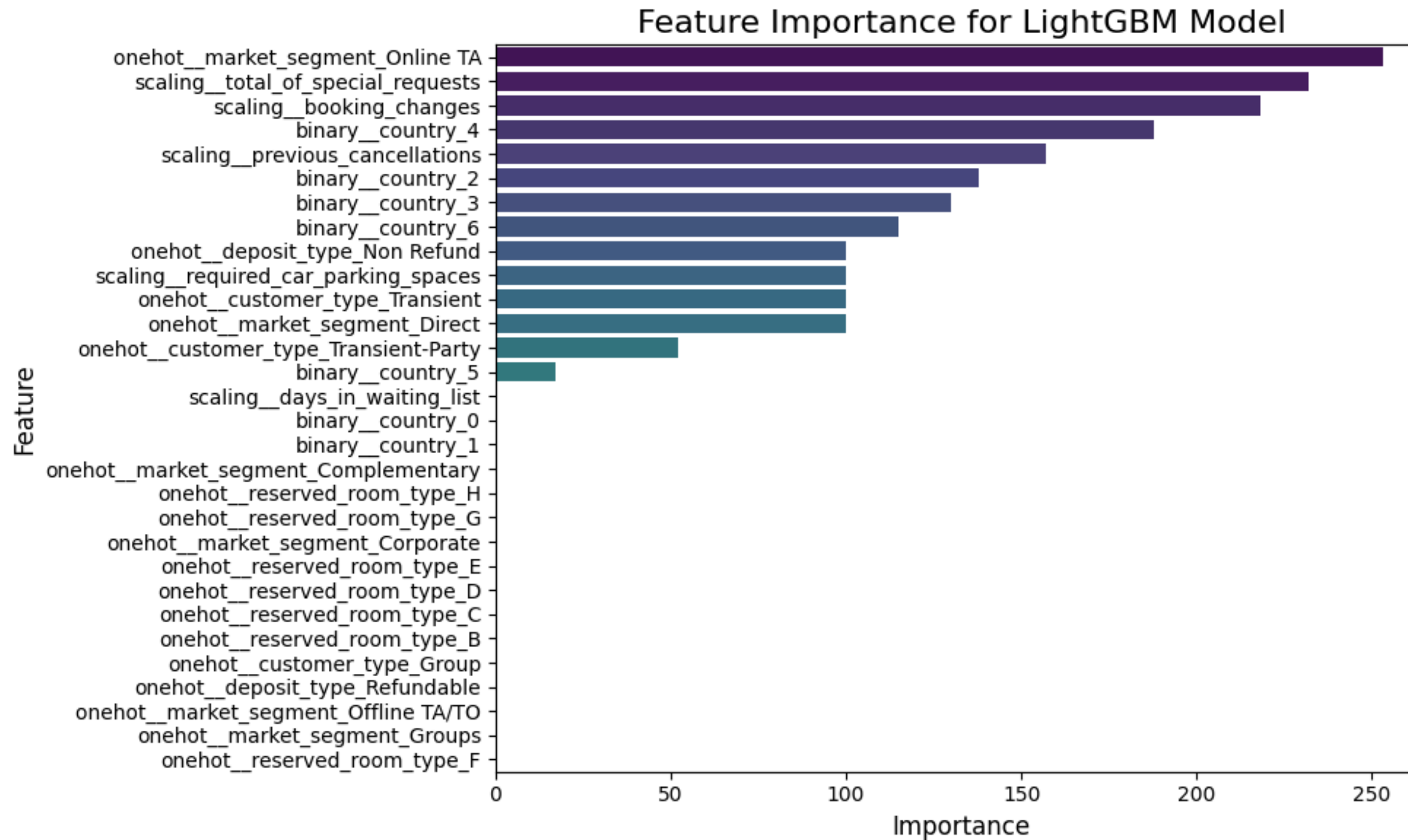
	precision	recall	f1-score	support
0	0.94	0.63	0.75	1537
1	0.42	0.88	0.57	479
accuracy			0.69	2016
macro avg	0.68	0.75	0.66	2016
weighted avg	0.82	0.69	0.71	2016

Confussion Matrix

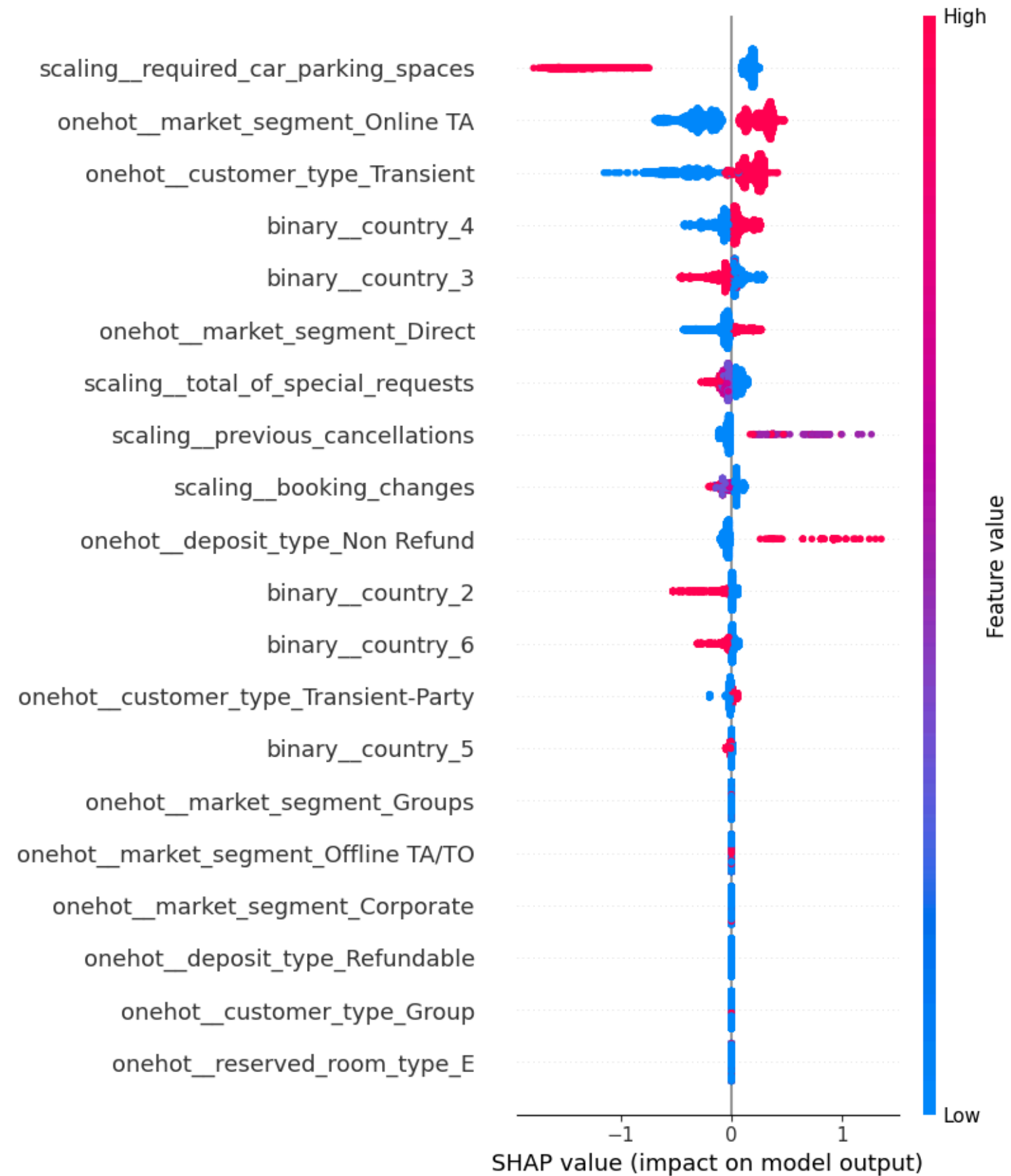




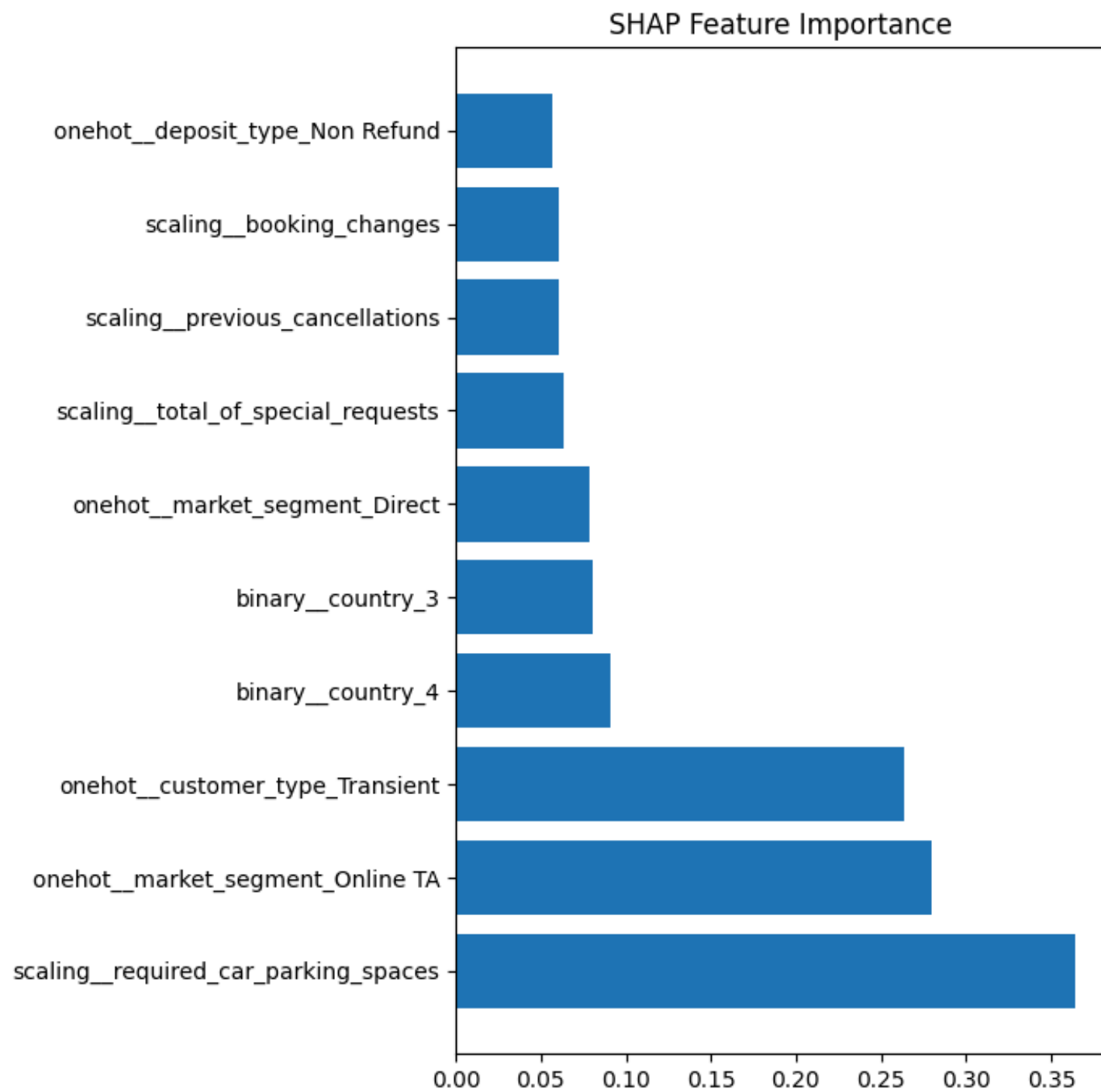
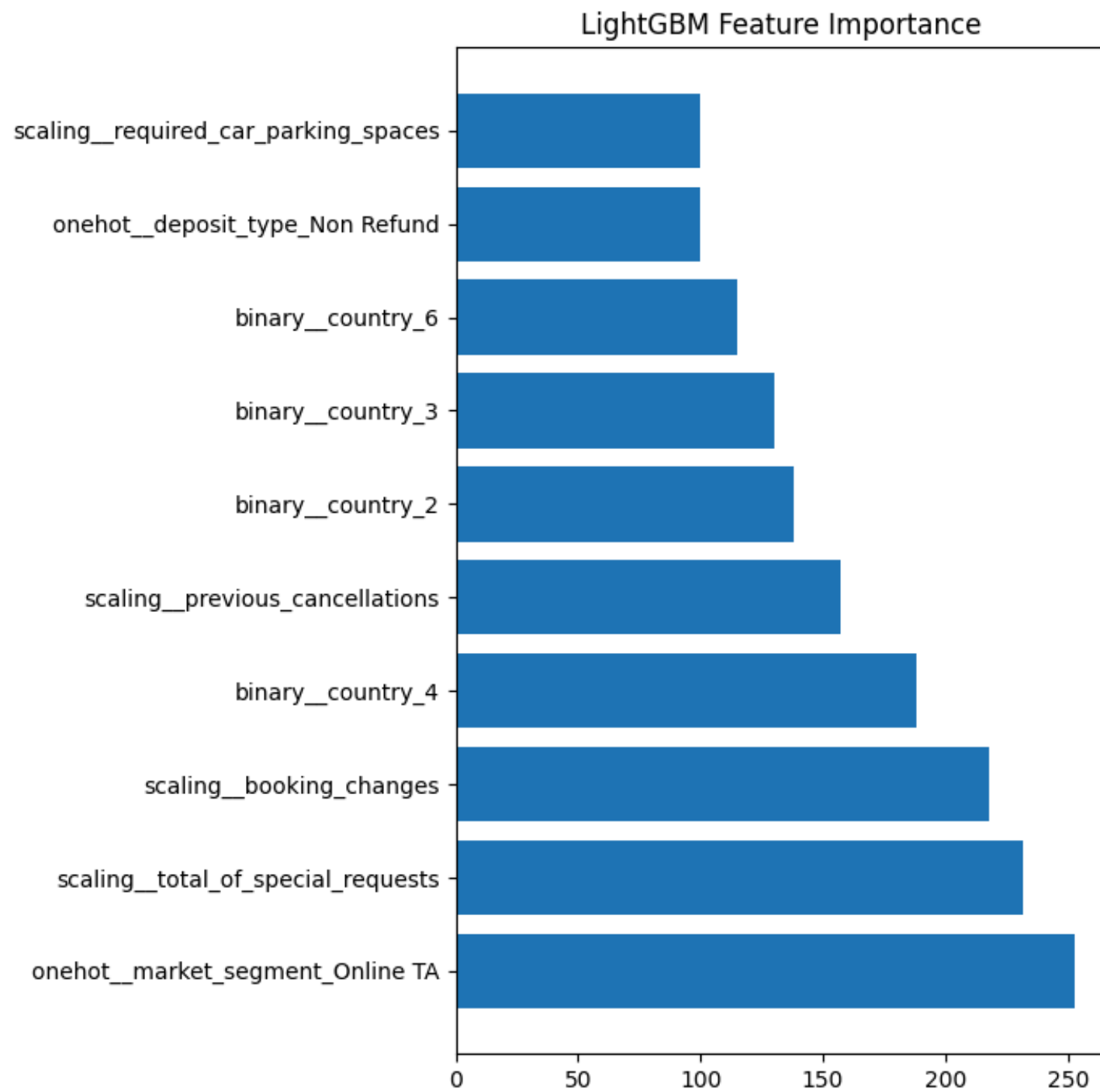
# Feature Importance LightGBM



# Feature Importance Shap



# Feature Importance Model



# 5.Success Metric



# Calculation With and Without Model



Explanation	Without Model	With Model
Total Customer	2016	2016
Price Each Room	50 USD	50 USD
Total Price	100,800 USD	70,800 USD
Total Served Customers Without Cancellation	1537	995
Total Unserved Customers Without Cancellation	0	542
Total Served Customers with Cancellation	0	421
Total Unserved Customers with Cancellation	0	58
Loss	23,950 USD	27,100 USD
Savings	0 USD	30,000 USD

# Conclusion

## 1. Best Model

- Model: LightGBM
- Resampler: RandomUnderSampling

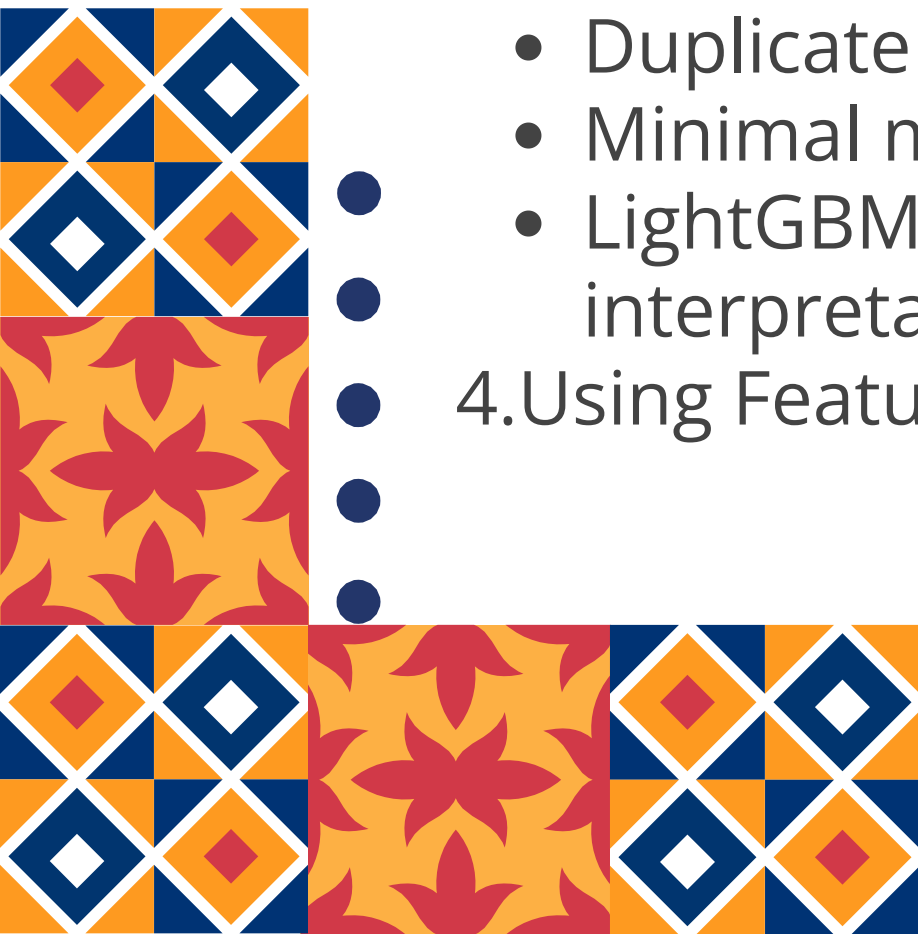
## 2. Model Interpretation

- Scaling: Robust Scaling
- Encoders: Binary and One-Hot Encoding

## 3. Model Limitations

- Duplicate data removed to prevent overfitting
- Minimal missing values dropped to avoid bias
- LightGBM is a "black-box" model, making it less interpretable

## 4. Using Feature Selection



# Bussiness Reccommendation



1. Stricter Cancellation Policies
  - Reconfirmation Before a Certain Date
  - No Cancellation for Certain Categories
  - Remove Refundable Option
2. Offer Special Deals or Discounts
  - Special Offers for Substitutes
  - Incentive Offers for Rescheduling
3. Implement Pre-Authorization Policies
  - Credit Card Pre-Authorization
  - Non-Refundable Deposit
4. Enhancing Customer Experience
  - Improved User Experience
  - Increase Customer Satisfaction

