Predicting Cancellations for Hotel Bookings

Matt Mascarelli

Context

➔ Cancellations = Revenue Loss

- **In the dataset there is approximately €34,446,903 worth of potential revenue if all of the bookings are successful. The cancellations account for about 33% of that, which is a significant loss.**

€11,478,718 LOSS

Plan

How can we mitigate this loss?

- Cancellation policy ❎
- Overbooking ❎
- Incentives ✅

How can we determine which bookings to offer incentives to?

- Machine Learning
- Predict high risk bookings

Problem Statement

- How can a hotel reduce their revenue loss by 20% for next year by <u>targeting customers that are likely to cancel</u> with incentives to retain their reservation?

The Data

Two Hotels in Portugal:

1) Hotel 1(Resort): 40,060 rows
2) Hotel 2(City): 79,330 rows
3) Bookings are between 2015-2017

After Cleaning:

- 86,207 rows
- 19 features

Training and Testing:

- Train Set: 68,965 rows
- Test Set: 17,242 rows

[Reference: https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib4]

Results

Of the 4,784 canceled bookings in the test data,

76%

are correctly predicted to cancel.

Features:

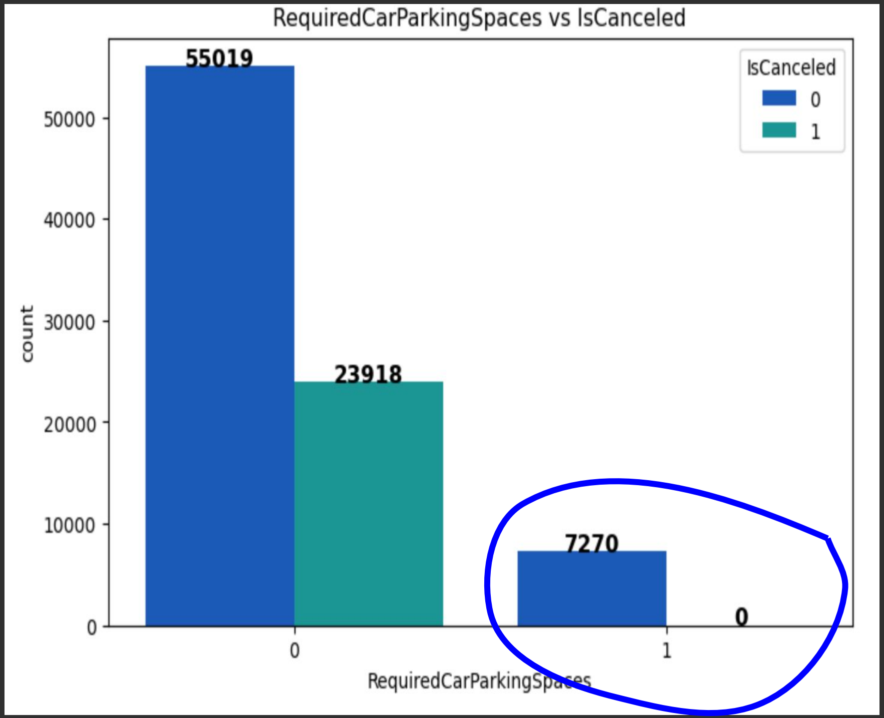
Car Spaces

Binary:

0 = No car space

1 = Car space

- 8% of bookings required a space
  - 0 cancellations

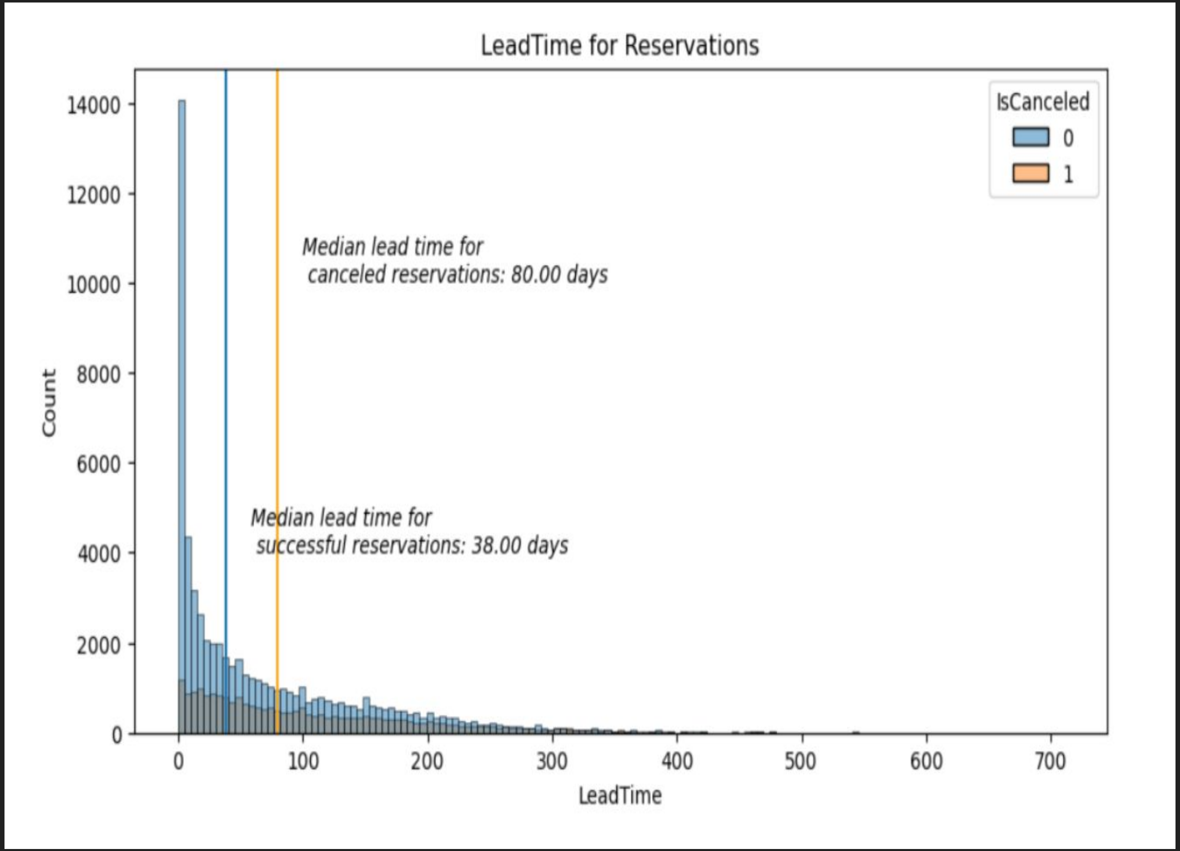


Lead Time

Number of days between reservation and arrival

<u>Median Lead Time:</u>

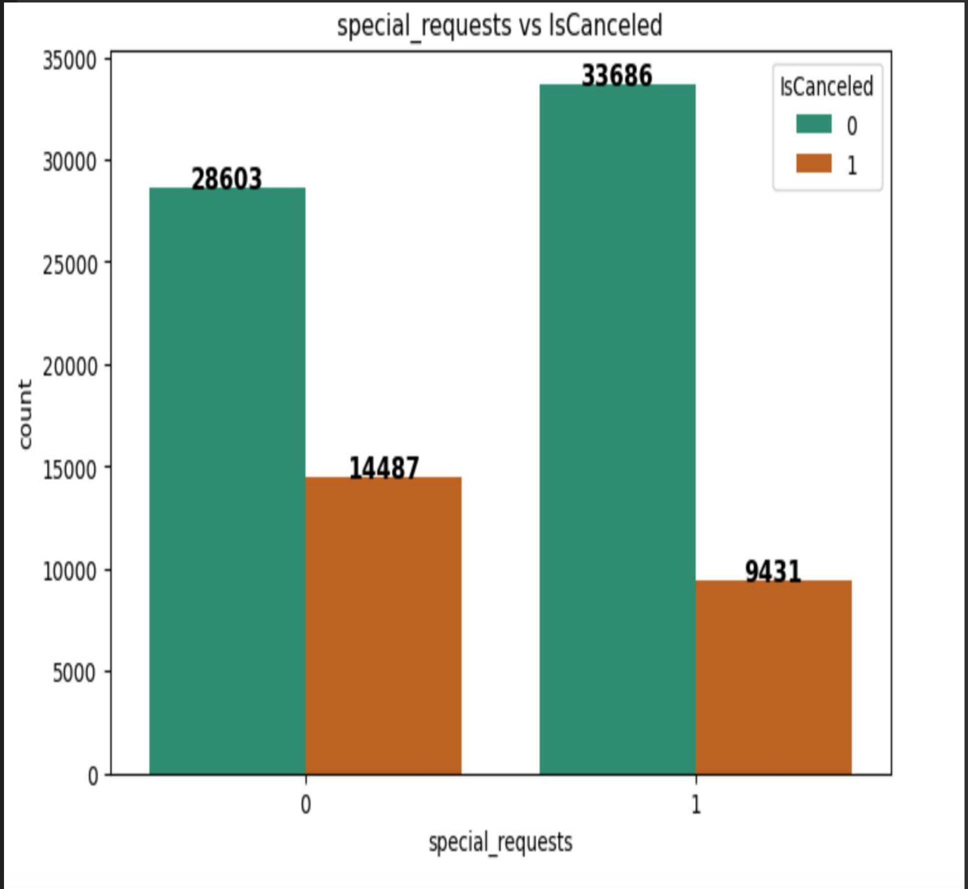
- **Canceled: 80 days**
- **Successful: 38 days**



Special Requests

Binary:
    0 = No requests
    1 = Requests

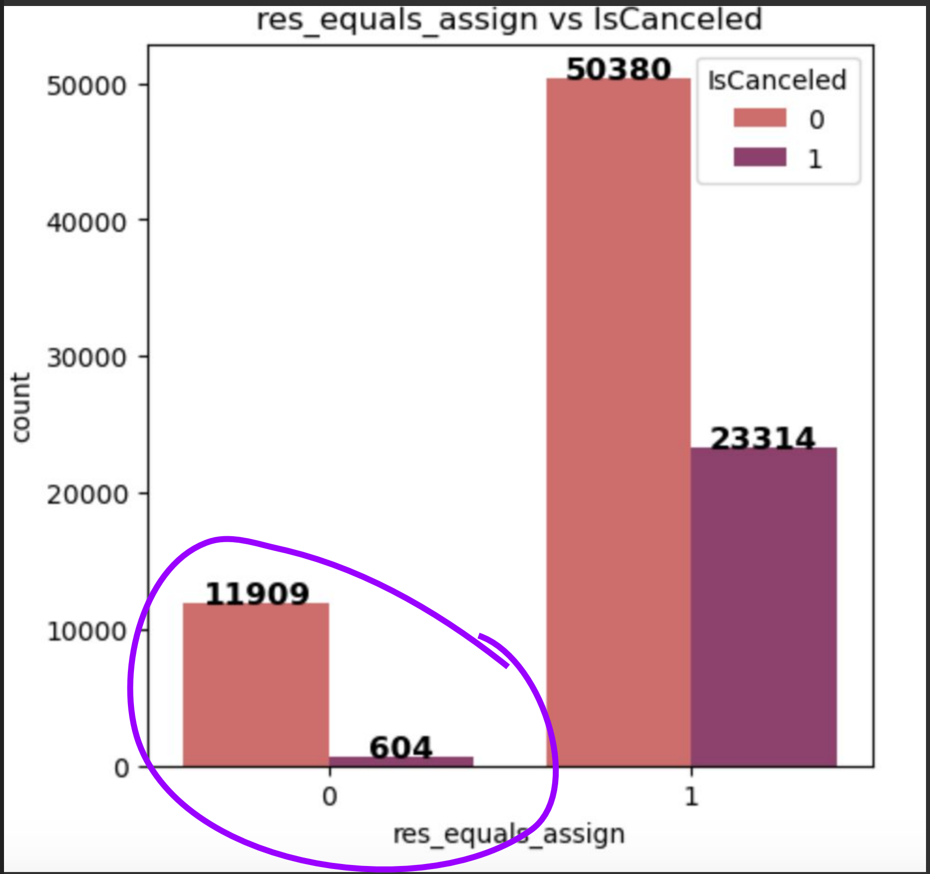
- Approximately 50/50 split between groups
- No Requests: 34% canceled
- Requests: 22% canceled



Reserved Room Type vs Assigned Room Type

Binary:
- 0 = Different Room Type
- 1 = Same Room type

➔ 14.5% of bookings received a different room type
  - Only 5% cancelled



Market Segment

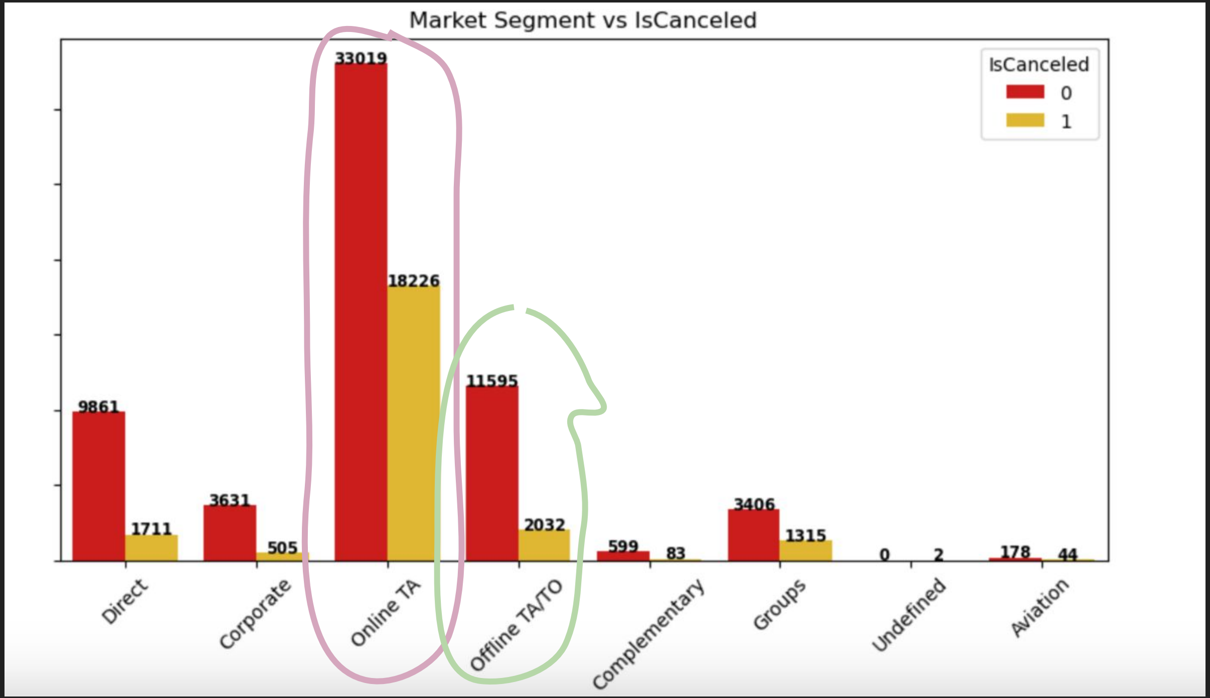
Categorical Feature

Online TA

➔ 59% of bookings
  ◆ 36% canceled

Offline TA/TO

➔ 16% of bookings
  ◆ 15% canceled



Agent Type

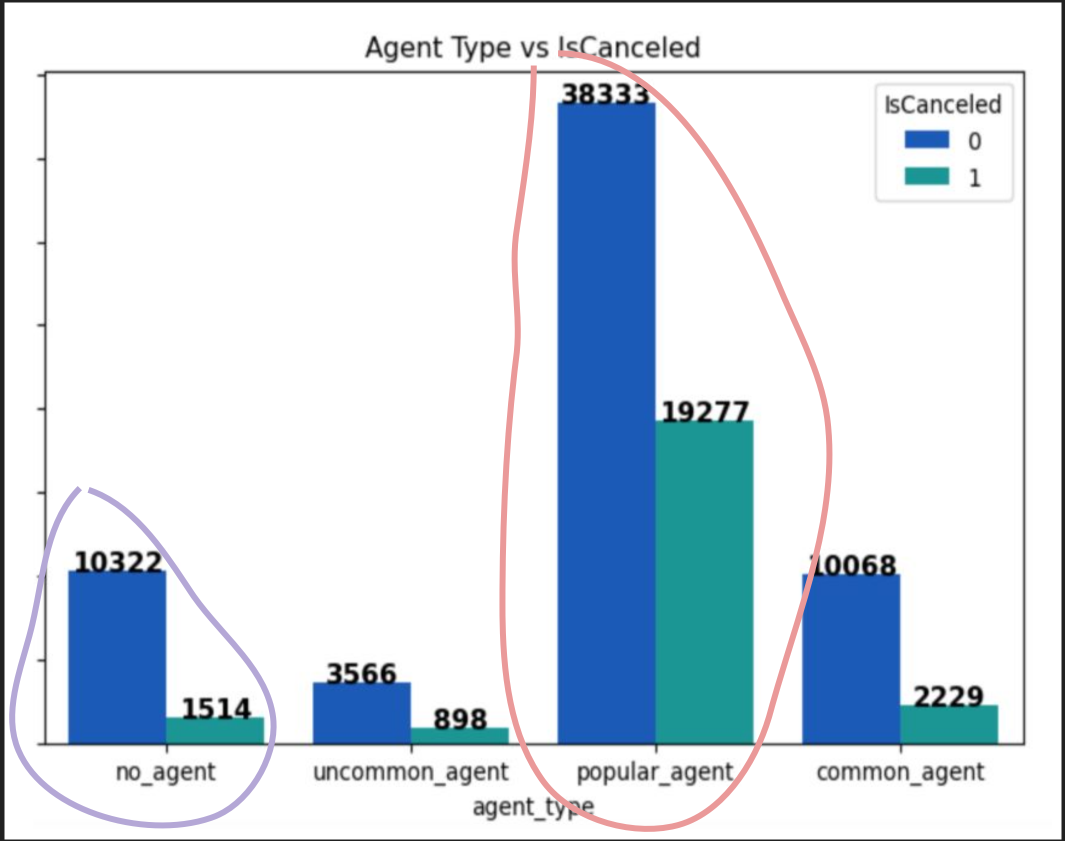
Categorical Feature:

Popular Agent:

- ➔ **66.8% of bookings**
  - ◆ **33.5% canceled**

No Agent:

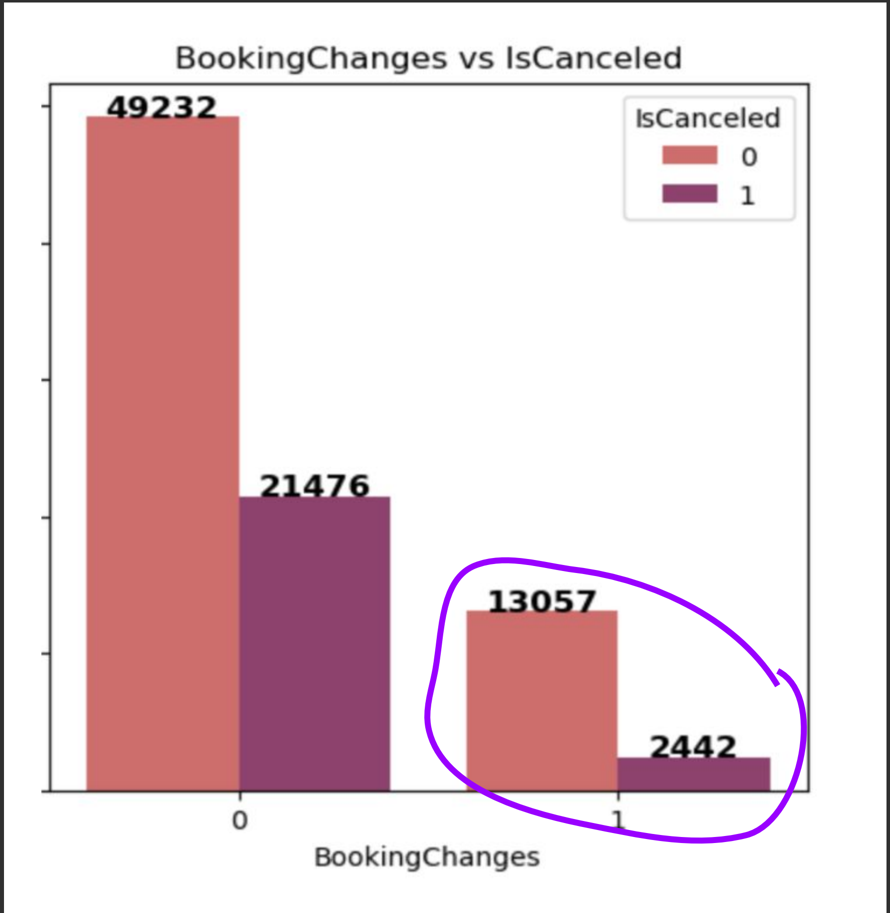
- ➔ **13.7% of bookings**
  - ◆ **12.8% canceled**



Booking Changes

Binary:
0 = No changes
1 = Changes

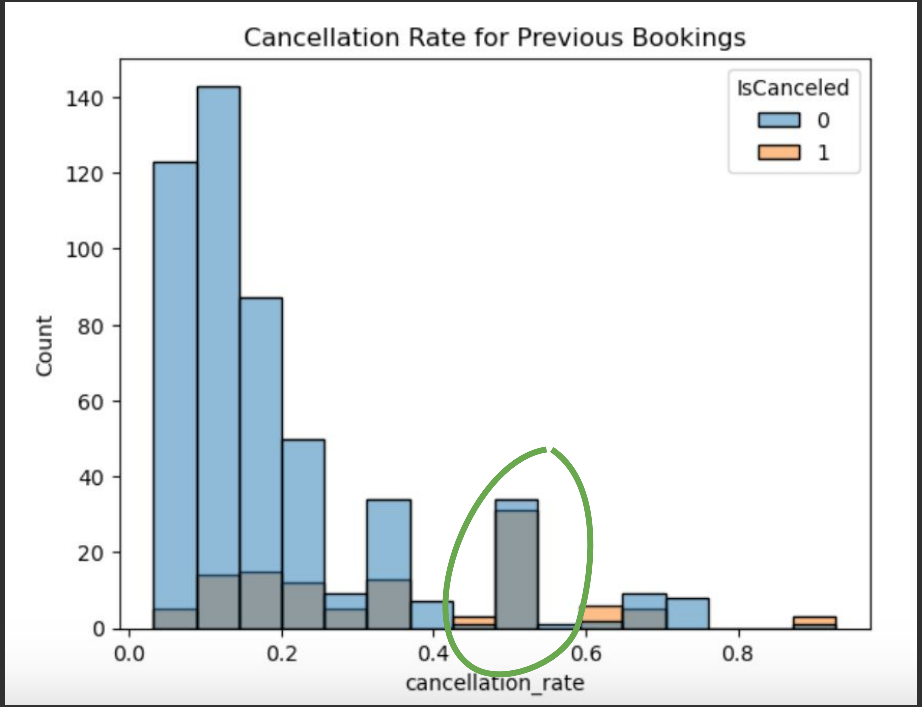
- ➔ No changes: 30% canceled
- ➔ Changes: 16% canceled



Previous cancellation rate

Group A: 0% PCR
- 98% are in this group
- 26% canceled

Group B: 100% PCR
- 1.2% are in this group
- 98% canceled AGAIN



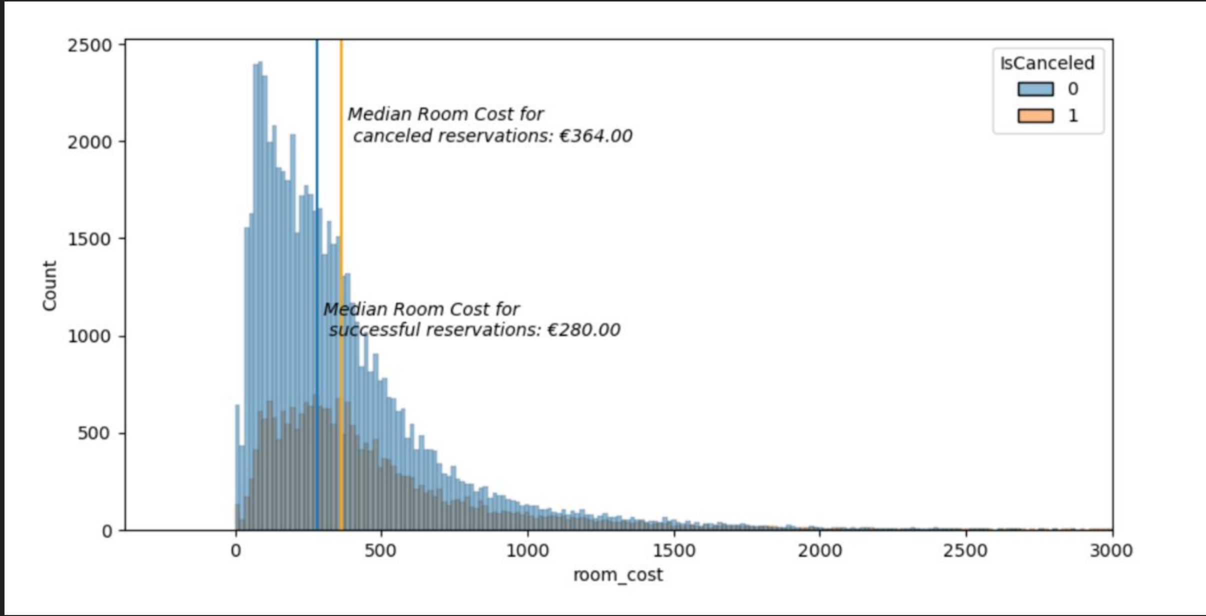
Group C: Everyone else
- 0.8% are in this group
- People who cancel previously, tend to cancel again

Room Cost

Total cost of the room

<u>Median Room Cost:</u>

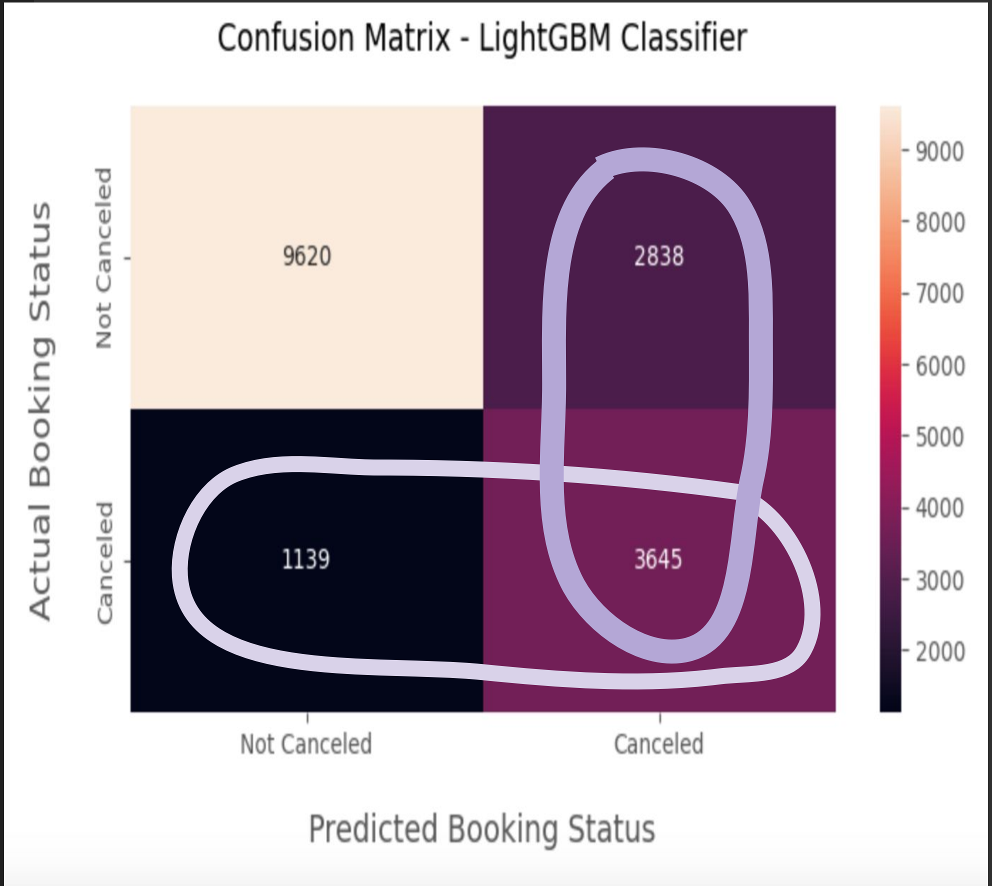
- **Canceled: €364**
- **Successful: €280**



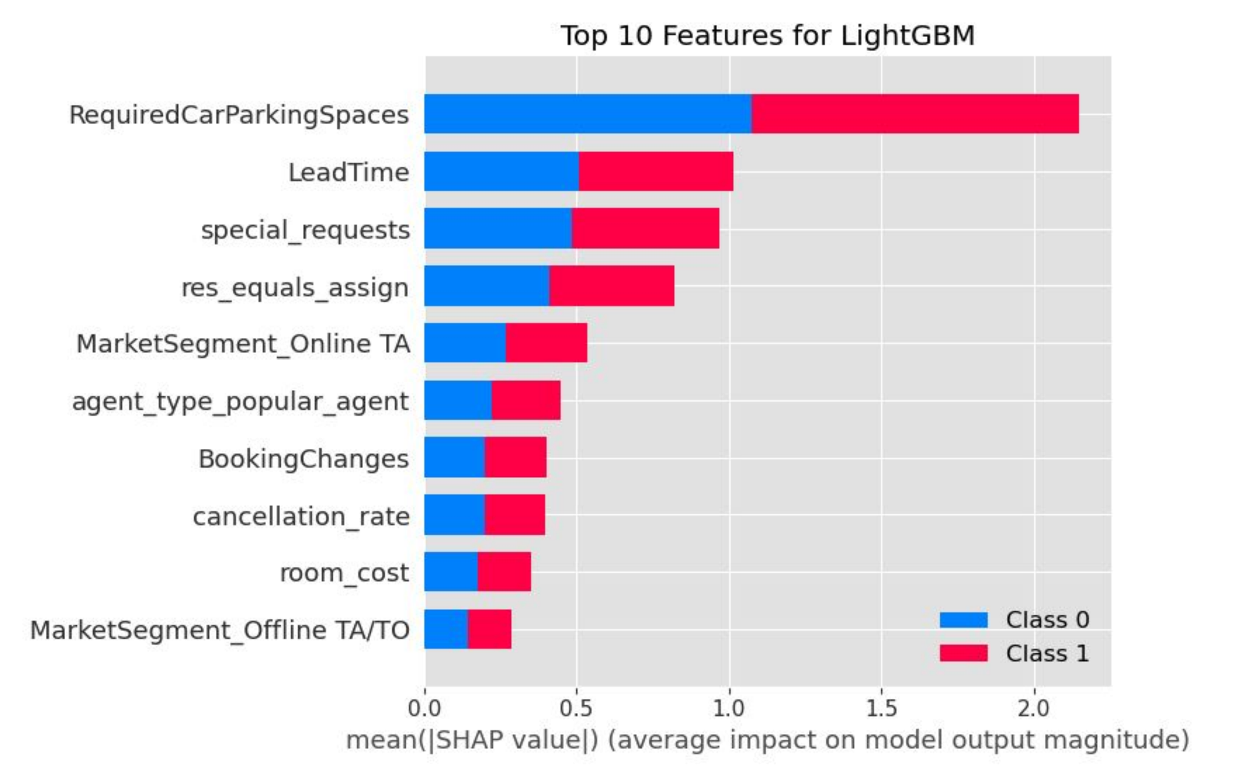
The Model

- **Models used: logistic regression, decision tree, random forest, Xgboost, LightGBM**
- **Best Model: LightGBM**

- **Recall: 76% (True positive Rate)**
- **Precision: 56%**

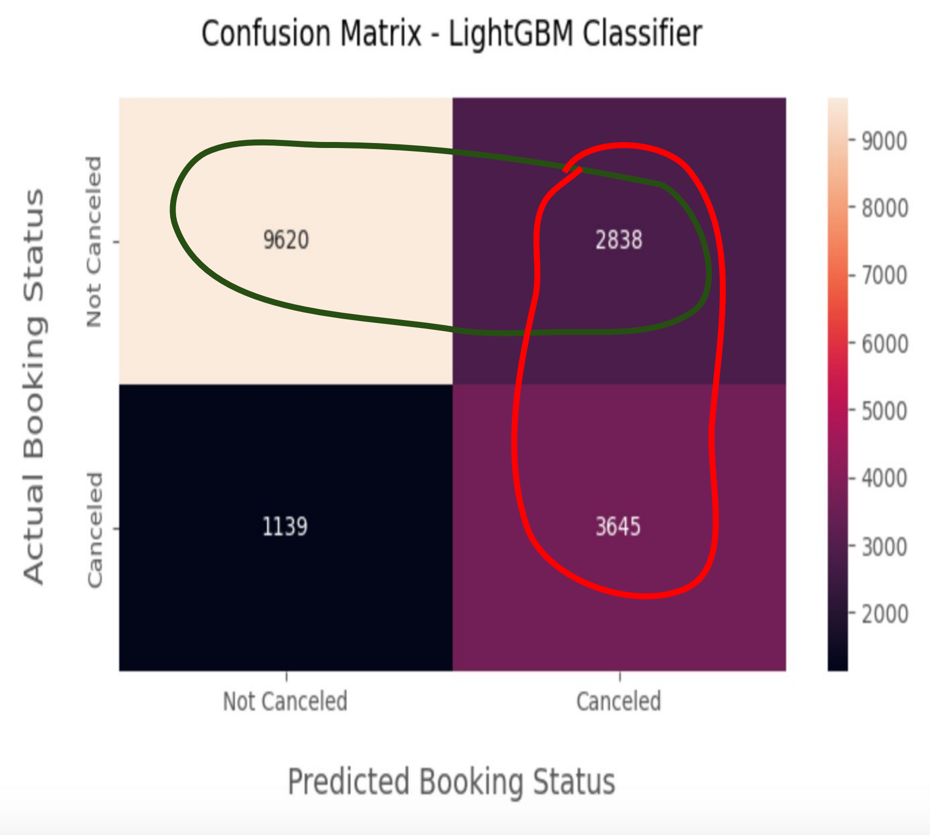


Important Features



Are the False Positives an Issue? (No, but maybe)

- 12,458 Successful bookings
  - 2,838 are false positives (22.7%)
- Successful booking = revenue gain✅
- Booking retention could be costly❌
  - 6,483 Bookings predicted to cancel
  - Complementary Rooms, meals, other services
  - Note: financial data not readily available to calculate profit.



Decision Threshold

	IsCanceled	predictions	probability_sucessful	probability_canceled
❎	0	1	0.412427	0.587573
✅	1	1	0.118317	0.881683
✅	0	0	0.949426	0.050574
❎	1	0	0.635925	0.364075
✅	0	0	0.701680	0.298320

Predictions in Practice

Bookings Predicted to Cancel:

1. **At least 70% probability (High Risk)**:
    a. Email/phone call to customer
    b. Offer Complimentary nights, meals, etc. (**if costs allow**)

2. **Less than 70% probability (Low Risk):**
    a. Email/phone call to customer

Further Work

1. Financial Data
   a. How many incentives can the hotel actually offer? (nights, meals, etc.)

2. Weather Forecast Data
   a. Could improve model accuracy