



INN Hotels

Predicting Cancellations

By Emilie Helen Wolf



Content

1. Problem Statement and Solution Approach
2. Data Overview
3. Exploratory Data Analysis
4. Model Performance
5. Recommendations



Problem Statement and Solution Approach

What profitable policies for cancellations and refunds can the hotel adopt?

1. Collect data on hotel cancellations for classification machine learning
2. Build and compare Logistic Regression and Decision Tree models
3. Determine the strongest influencing cancellation factors



Data Overview:

How the Data was Collected

Data Dictionary

Booking_ID: the unique identifier of each booking

no_of_adults: Number of adults

no_of_children: Number of Children

no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel

type_of_meal_plan: Type of meal plan booked by the customer:

required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)

room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group

lead_time: Number of days between the date of booking and the arrival date

arrival_year: Year of arrival date

arrival_month: Month of arrival date

arrival_date: Date of the month

market_segment_type: Market segment designation.

repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)

no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking

no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking

avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

booking_status: Flag indicating if the booking was canceled or not.



Data Overview: Initial Look

- 36275 rows and 19 columns
- One-third of bookings are cancelled
- No missing values
- No duplicates
- No outliers
- 72.5% of the dataset has questionable dates that don't line up. We suggest re-validating the dataset, but decided to assume the dates are accurate for model-building.

Data is tidy for the most part and can be used in full for model building

Exploratory Data Analysis

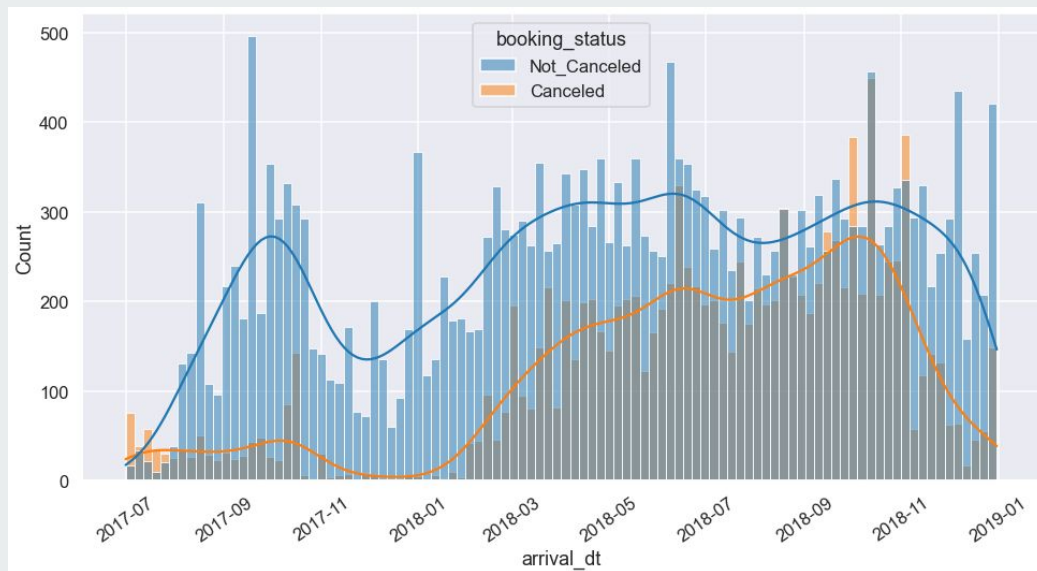
Univariate Plots

Histogram of All Bookings over Time

It looks like we have the second half of 2017 and a full year of 2018

There were more bookings in 2018 than 2017, showing growth

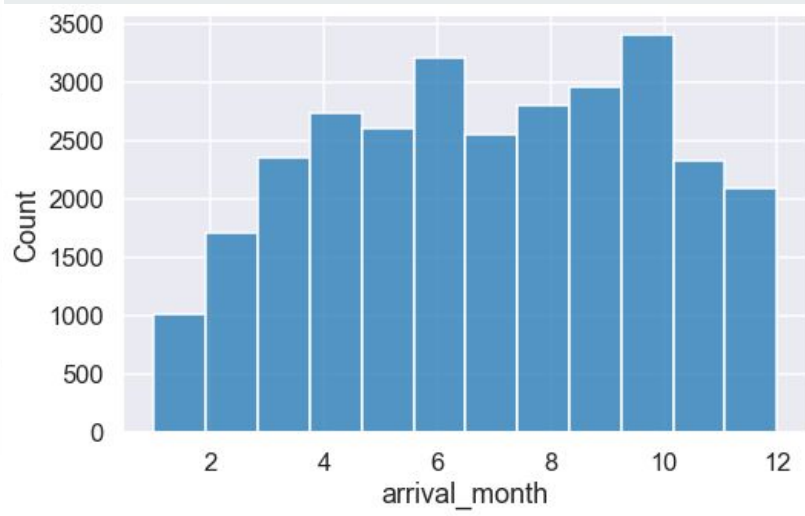
In September 2018, the number of cancellations almost met the number of non-cancellations



Histogram of 2018 Bookings by Month

This is from 2018 only

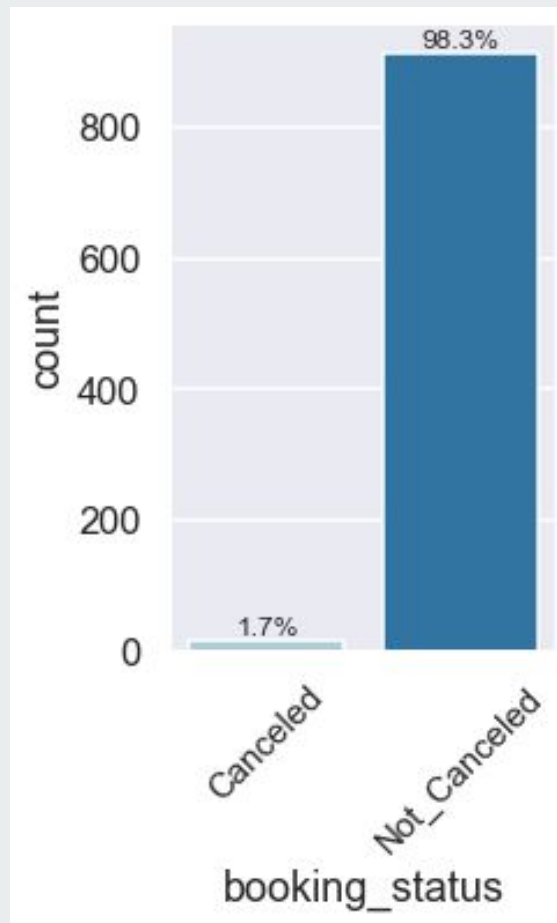
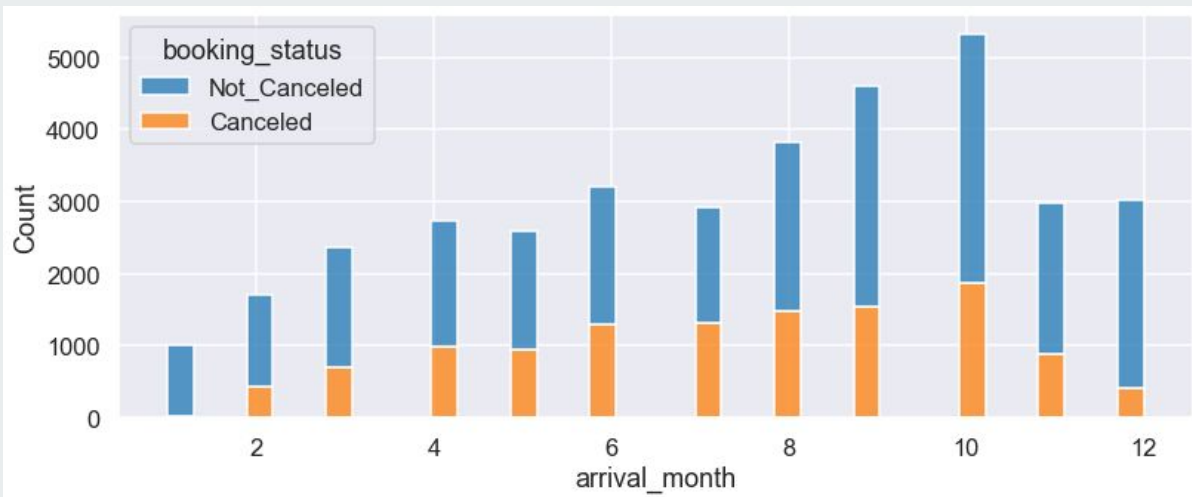
The busiest months are October, June, and September. The slowest month is January.



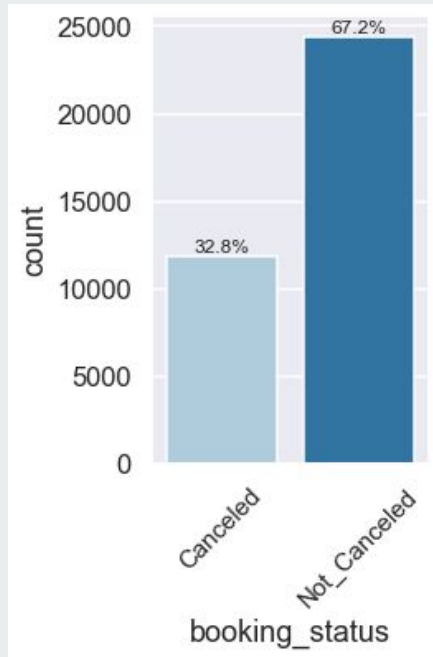
1.7% of Repeat Guests Cancel

Histogram of 2017 and 2018 Bookings by Month

Cancellations dip in December and January



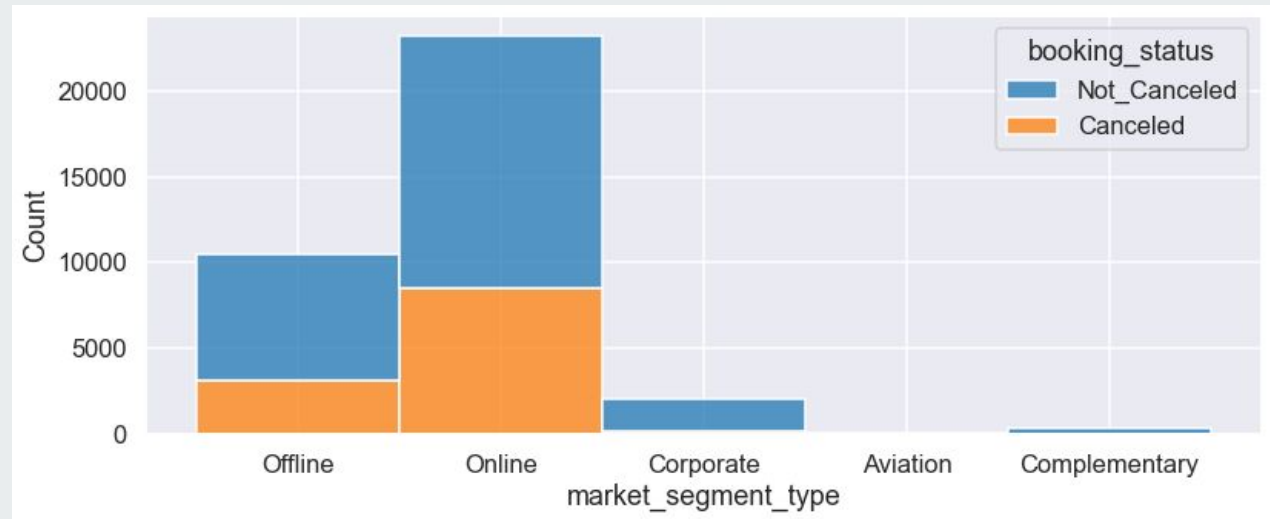
32.8% of bookings are cancelled in this dataset



Histogram of All Market Segments

Most customers come from the "Online" market segment

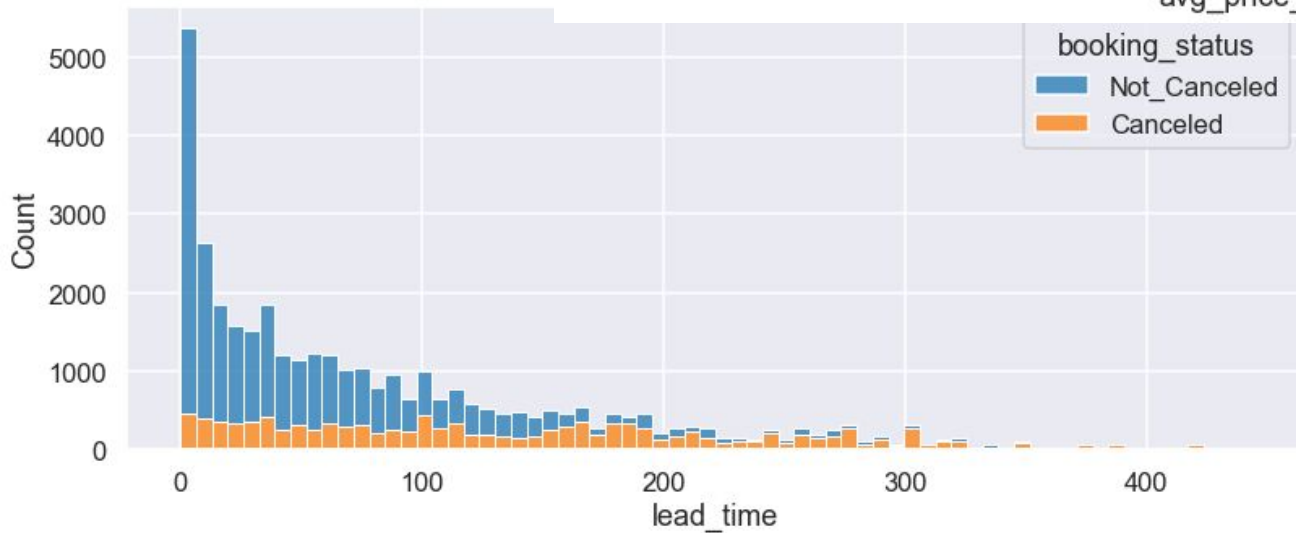
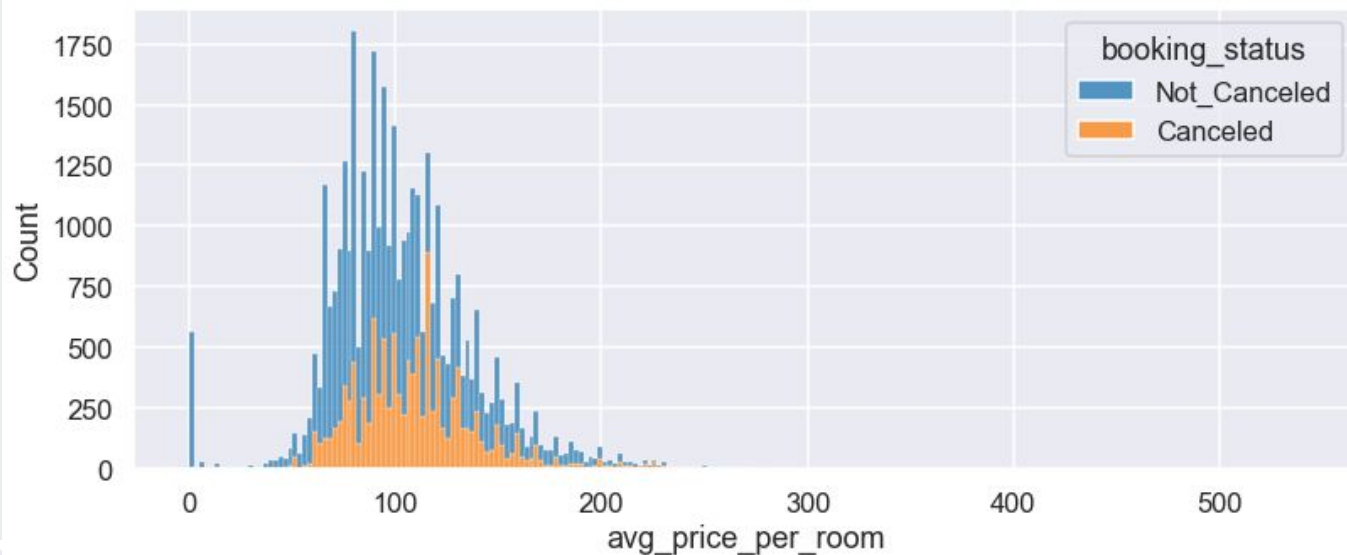
The smallest market segment is "Aviation"



Histogram of Average Price Per Room

The prices have a single peak and are right-skewed.

There are over 500 bookings that are free.



Histogram of Lead Time

Cancellations become more likely as lead time increases.

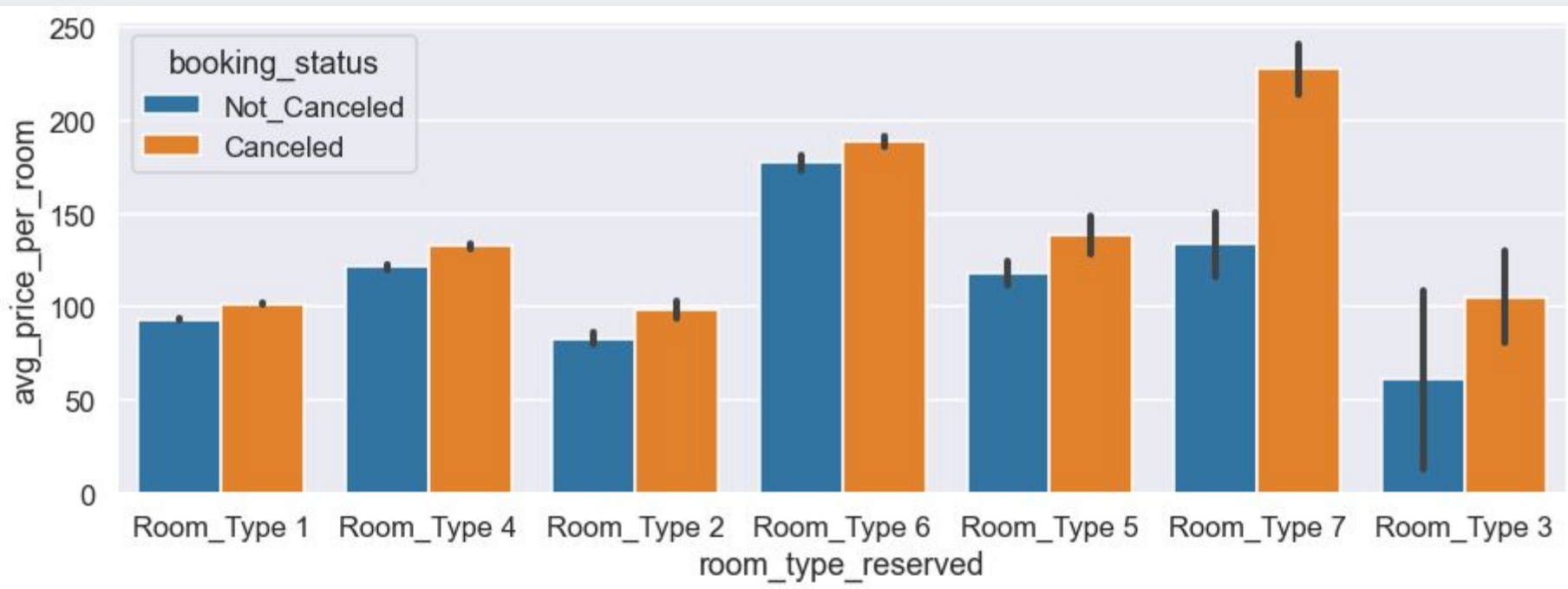
Exploratory Data Analysis

Multivariate Plots

Average Price per Room by Room Type

Room Type 6 is the most expensive

Cancellations have a higher average price per room than non-cancellations

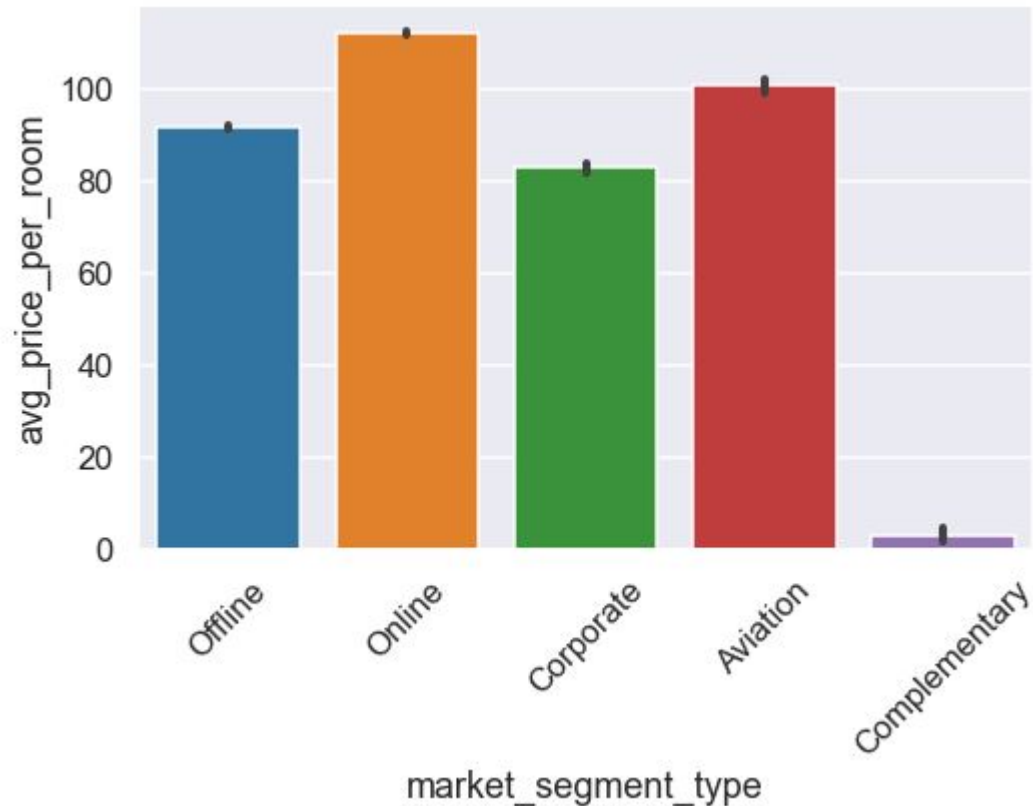


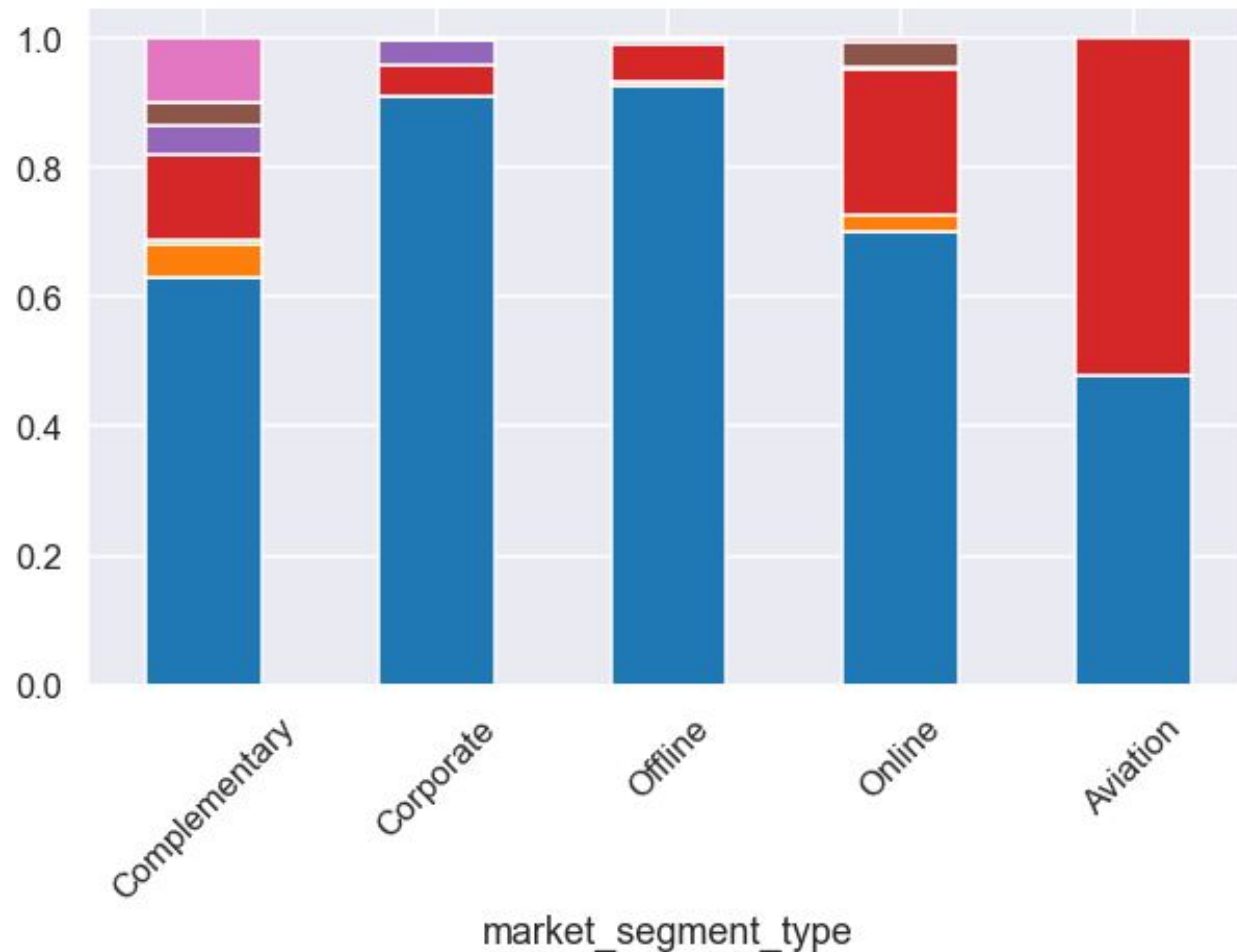
Average Price per Room by Market Segment

On average, the Online segment has the highest average price per room

Next is Aviation, then Offline, then Corporate

The average price of the Complementary segment is close to zero, indicating lots of free rooms

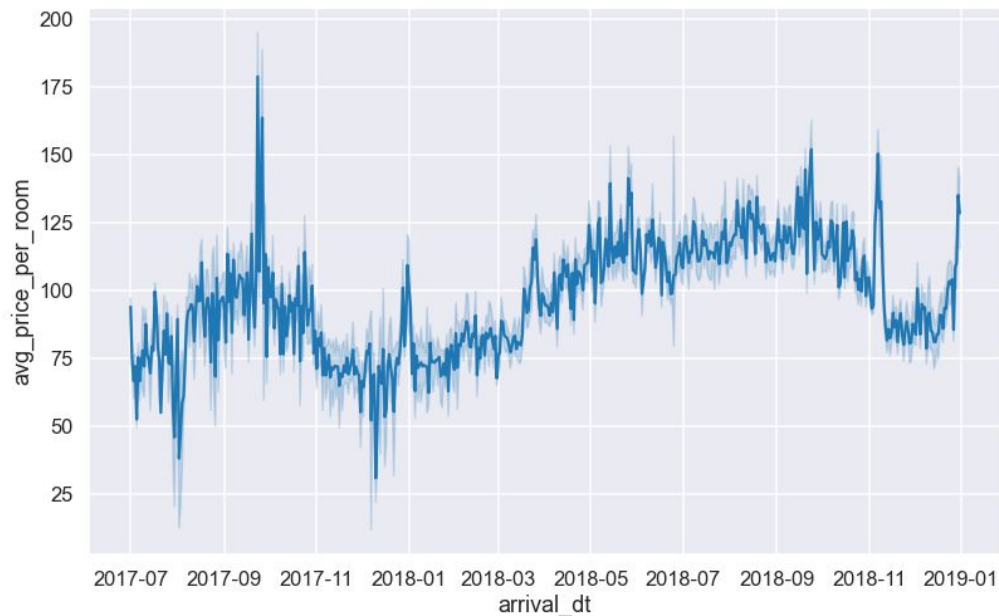




Distribution of Room Types by Market Segment

Most market segments are dominated by Room Type 1

Average Price per Room over Time



It looks like prices go down in November and December.

We saw earlier that bookings and cancellations are usually down during this time as well.

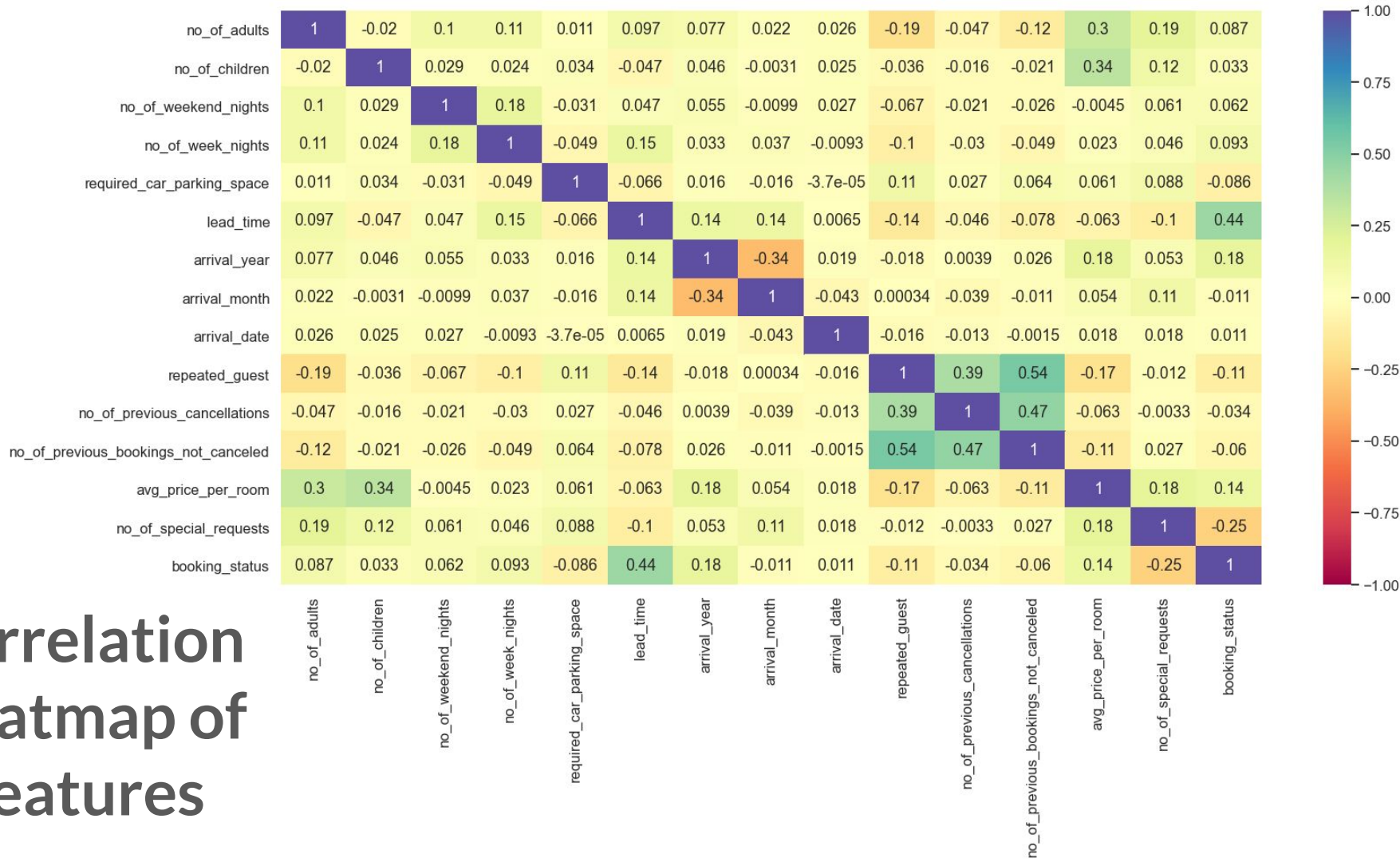
Perhaps cancellations are low during this time because the hotels have lowered prices to try and secure more guests.

Average Number of Special Requests by Booking Status

Less special requests are observed by bookings that are cancelled.



Correlation Heatmap of Features



Model Performance



We want a model that will predict whether or not a booking will be canceled.

- If the hotel predicts a booking will cancel (1) and they actually don't (0) (Type I error), the hotel might be understaffed and overbooked and customer service might suffer.
- If the hotel predicts a booking will not cancel (0) and they actually do (1) (Type II error), the hotel loses revenue on the room and is subject to costly, last-minute changes.

Type II Errors are more costly, so we want a model that has a **high recall score as well as a **high accuracy score**.**



We created 3 Logistic Regression Models and 3 Decision Trees
Here is a comparison of their scores on the testing sets:

	Accuracy	Recall	Precision	F1
Logistic 0.5	0.809152	0.647076	0.732091	0.686963
Logistic ROC 0.34	0.79932	0.747871	0.670229	0.706924
Logistic 0.44	0.804374	0.697615	0.697813	0.697714
Decision Tree 0	0.861343	0.795571	0.780284	0.787853
Tree Pre-pruned	0.610493	0.951732	0.451691	0.612629
Tree Post-pruned	0.866673	0.802669	0.789004	0.795778

The rows in green are the best models for each model type.
Notice the 95.1% Recall Score on the Pre-pruned Decision Tree Model.

Our Best Logistic Regression Model Has a Testing Recall Score of 74.8%

Accuracy Score: 79.9%

Model Assumptions Passed ✓



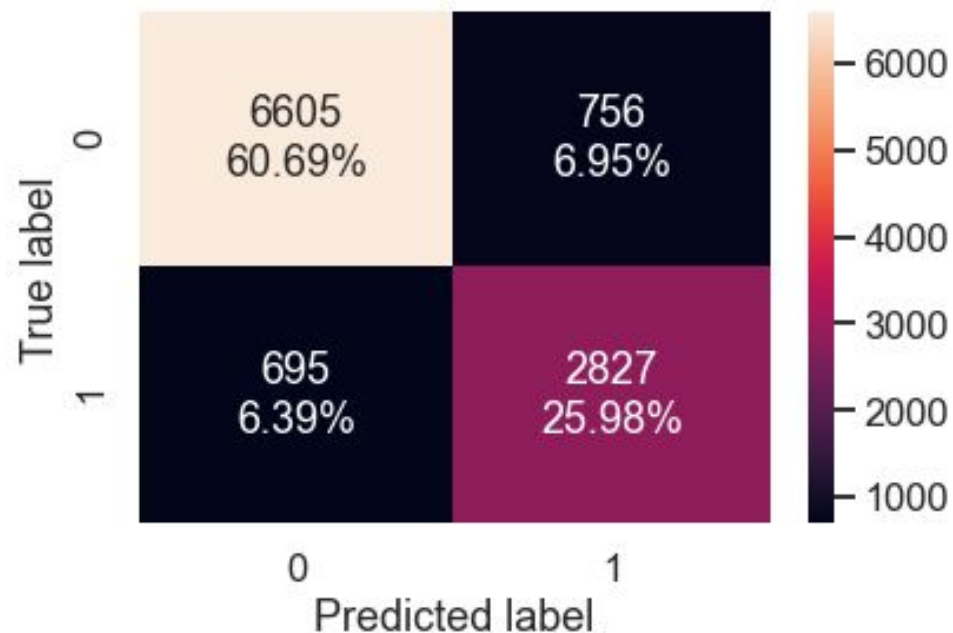
False Negatives account for 8.16%

We found the optimal threshold using the ROC curve:

0.34371724449936036

Our Best Decision
Tree Model Has a
Testing **Recall Score**
of **80.3%**

Accuracy Score: 86.7%



False Negatives account for 6.39%


This decision tree model was
optimized by using Cost Complexity
Pruning or post-pruning

Recommendations



We think the best model is the **post-pruned Decision Tree** because it has a high Recall and the highest Accuracy and F1 score, meaning it did the **best at reducing both Type I and Type II errors**, a great balance between no-shows and overbookings.

- With this model, we can predict whether or not a booking will cancel 87% of the time.
- The remaining 13% error is split as such: 7% Type I error and 6% Type II error.
- The variables with the most impact on booking status are **Lead Time, Average Room Price, the Online Market Segment, and Number of Special Requests.**

- 
- Any bookings with lead times greater than 6 months should be flagged as a possible cancellation, and the likelihood of cancellation increases if the booking price is higher than average, placed Online, or had zero special requests.
 - For bookings that are flagged as possible cancels, the hotel can try offering a free upgrade to a nicer, but less popular room. The better value might convince the guest not to cancel, and then the originally booked room (which is probably more popular) can be re-allocated to guests less likely to cancel.
 - We do not suggest refusing refunds, but we do suggest charging a cancellation fee if the guest cancels close to the arrival date.

Having more certainty over which bookings are likely to cancel will help hotel staff allocate resources and make better decisions.



Thank You