Hotel Booking

By Scarlett Scileppi, Campbell Isherwood, Logan Doan, Trevor Krall, and Shirui Zhou

Agenda

Problem Statement

Goal:

Hotel cancellations lead to lost revenue, unused rooms, and operation inefficiencies. Understanding what drives a customer to cancel a booking can help hotels improve forecasting, allocate resources more effectively, and design better pricing or communication strategies.

Problem:

Using the Hotel Booking dataset, we aim to predict the probability that a reservation will be cancelled based on a multitude of characteristics.

Problem Statement cont.

Why this matters:

Accurate cancellation predictions allow hotels to:

- Reduce overbooking risks
- Optimize staffing and inventory
- Improve revenue management
- Create targeted retention strategies

Our Task:

Build and evaluate predictive models to identify the most important drivers of cancellations and estimate cancellation likelihood for future bookings.

Describe the Data

Source:
- Hotel Booking Demand (Kaggle)

Scope:
- ~119k reservations
- from city and resort hotels
- (2015-2017)

Contains:
- Guest info: arrival date, lead time, length of stay
- Reservation activity: special requests, changes made
- Target: whether booking was cancelled
- Collected by researchers from University of Lisbon

Why Useful: Predict cancellation probability + understand what drives cancellations

Exploratory Data Analysis

HOTEL ML KEY CARD

BOOKING AVERAGES AND DISTRIBUTION

ROOM 576

ADR
$103
(0, 510)

Guests
~2
(1, 12)

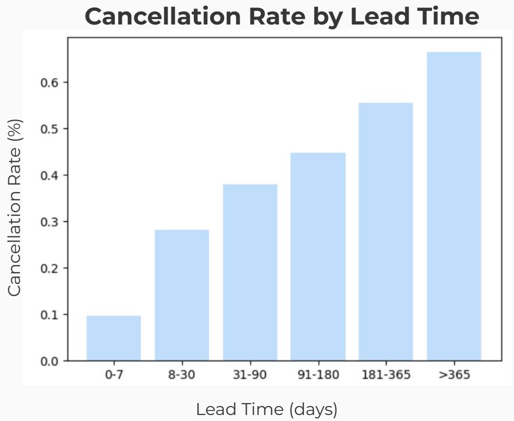
Nights Stayed
3.45
(1, 45)

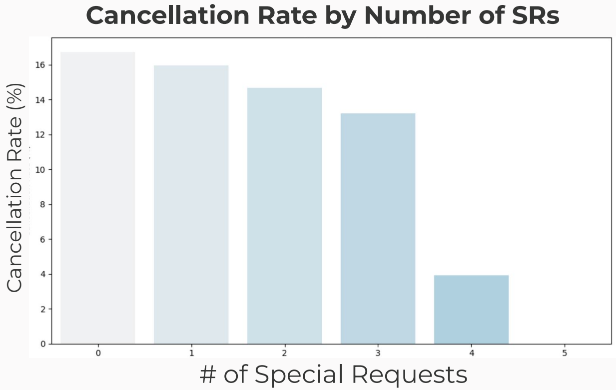
Lead Time
104.46
(0, 594)

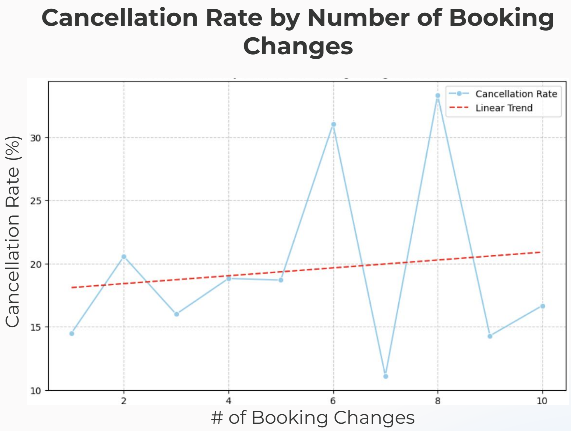
Special Requests
0.57
(0, 5)

Popular Month
August
(11,019)

Exploratory Data Analysis







Longer lead times are associated with **higher rates** of cancellation, especially for those booked over a year

Guests arranging **accommodations** (special requests) are **less likely** to cancel

Cancellation rate slightly **increases** with the number of **booking changes**

Exploratory Data Analysis - Overview

Key Observations

Right-Skewed Distributions

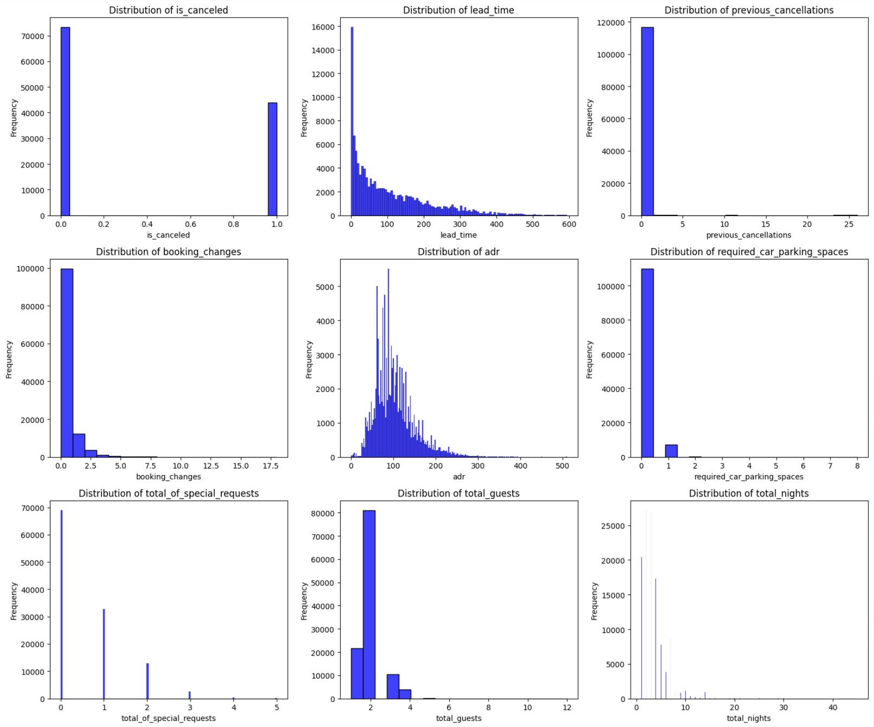
- Distributions are largely right-skewed, indicating the presence of outliers
- Outliers addressed through source paper and hotel industry standards

Two Hotel Types

- City Hotel: 61%
- Resort Hotel: 49%

Geographic Context

- Both hotels located in Portugal
- 90% of guests are European



Model Overview

Baseline: 62.5%

Forward Selection

- Forward Selection Accuracy Rate: ~73%
- 22/42 features selected

Backward Selection

- Backward Selection Accuracy Rate: ~64%
- 10/42 features selected

Lasso & Ridge

- Lasso Accuracy Rate: ~73%
- Ridge Accuracy Rate: ~73%
- Optimal lambda: 10

Random Forest

- Train Accuracy: ~80%
- Test Accuracy: ~79%

Boosting Tree

- Train Accuracy Rate: ~74%
- Test Accuracy Rate:~73%

Decision Tree

- Train Accuracy Rate: ~77%
- Test Accuracy Rate: ~76%

KNN

- Accuracy Rate: ~78%
- Test Rate: ~77%

Logistic Regression

- Train Accuracy: ~75%
- Test Accuracy: ~73.4%

Neural Network

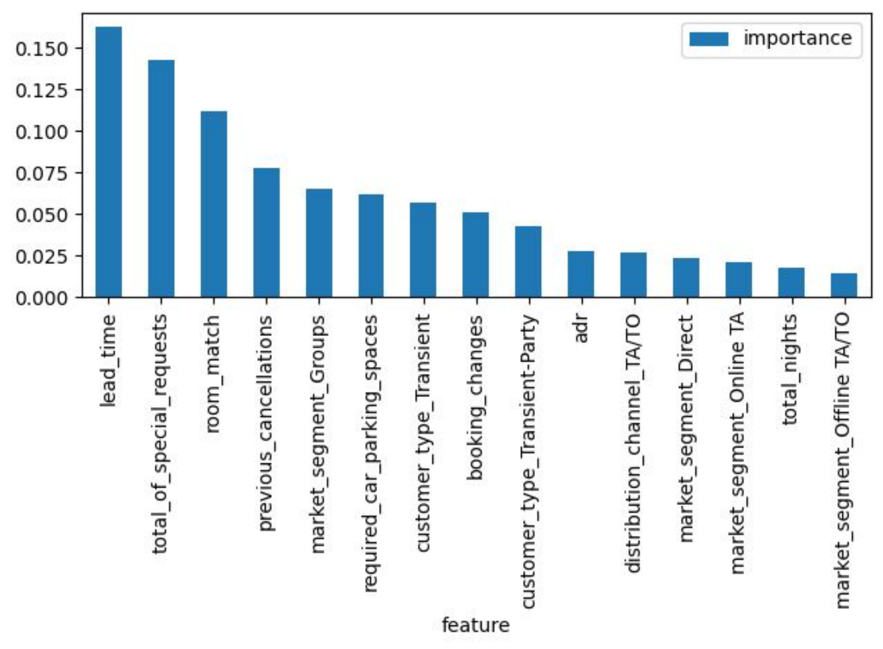
- Train Accuracy Rate: ~82%
- Test Accuracy Rate: ~80%

Main results
Random Forest

Train Accuracy Rate: ~80%

Test Accuracy Rate: ~79%

Important Features Of Random Forest



Main results

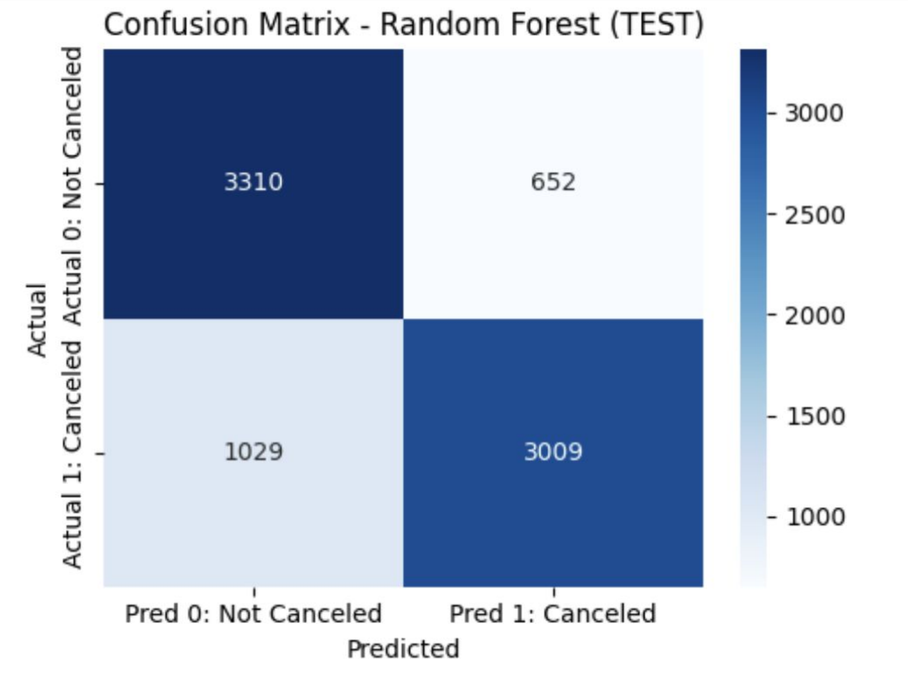
Random Forest

Class0 (Not Canceled)

Precision: 0.81
Recall: 0.79
F1 Score: 0.80

Class1 (Canceled)

Precision: 0.80
Recall: 0.81
F1 Score: 0.81

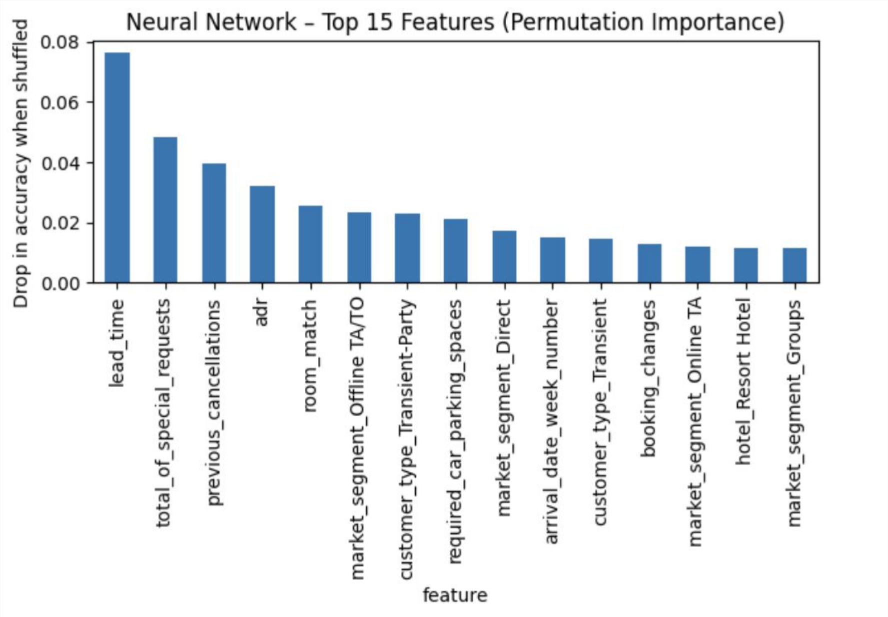


Main results
Neural Network

Train Accuracy Rate: ~81%

Test Accuracy Rate: ~80%

Important Features Of Neural Networks



Main results

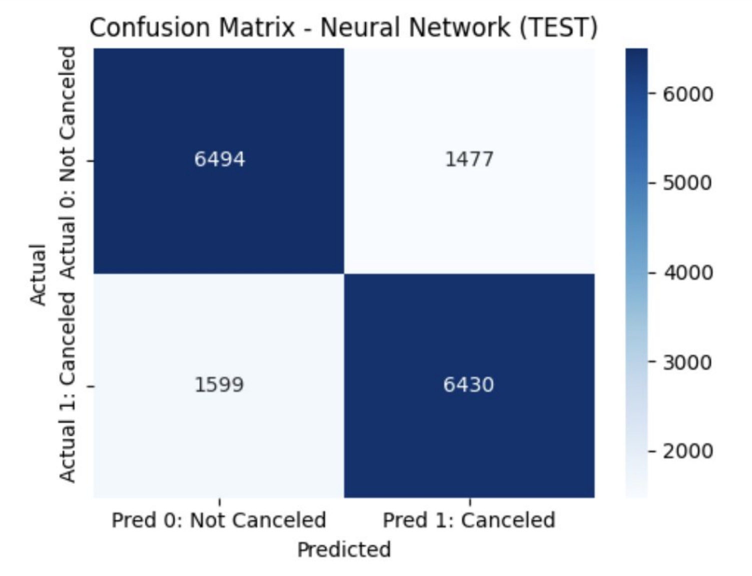
Neural network

Class0 (Not Canceled)

Precision: 0.80
Recall: 0.81
F1 Score: 0.81

Class1 (Canceled)

Precision: 0.81
Recall: 0.80
F1 Score: 0.81

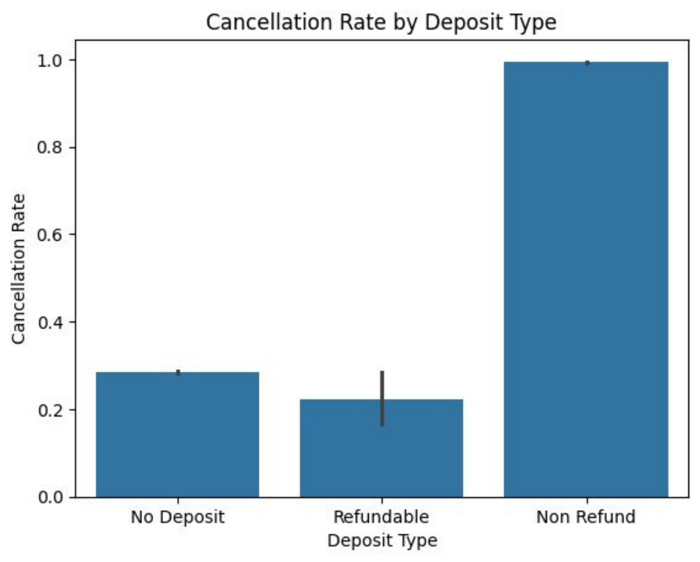


Challenge 1: Counter intuitive result

Counter intuitive result

Extremely High Cancellation for Non-Refundable Bookings:

- ~99.6% cancellation rate of non-refundable type
- ~ 67% cancellation cases are Non-refundable type
- in real-world hotel operations, non-refundable bookings typically have the lowest cancellation rates, and refundable/no-deposit bookings cancel more frequently.
- Industry research (Gould et al.) reports hotel no-show/cancellation rates around 5%–15%,



Impact on the Model

"Deposit type" becomes an artificially dominant predictor

- Disrupts the model's ability to learn meaningful patterns

Method we have tried

- Remove cancelled Non-refundable reservation
- Industry research (Gould et al.) reports hotel no-show/cancellation rates around 5%–15%,

Challenge 2: Data Geography

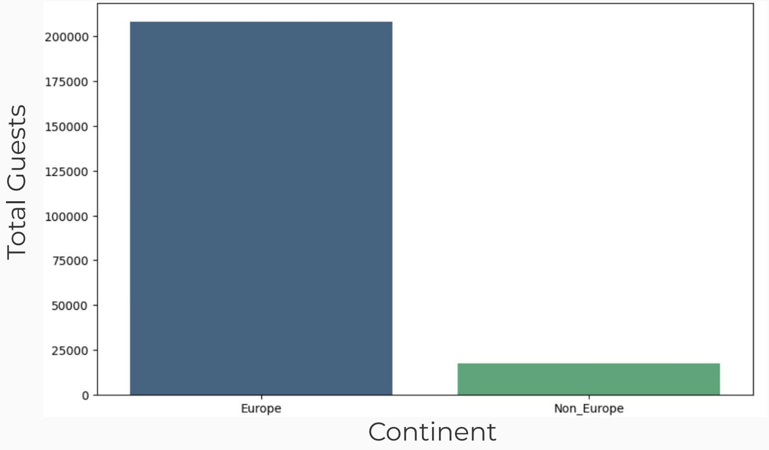
Data Collection

- **Limited Geographic Scope**
  Data for both hotels were collected in Portugal to study consumer behavior
- **Homogenous Guest Population**
  90% of hotel guests were travelling from Europe

Limited Application

- **Behavior Bias**
  Consumer insights are limited to Portuguese/European contexts and cultural factors
- **Overfitting**
  Predictive power could be overfit to local patterns

Guest Distribution by Continent



Top 3 Countries by Total Guests	
Portugal	48,590
Great Britain	12,129
France	10,415

Conclusion

Model Performance

Random Forests and Neural Nets had the highest predictive power
Capture complex, non-linear relationships with large number of predictors

Important Features and Business Application

Lead Time	Longer lead times associated with higher cancellation suggesting targeted strategies for re-engagement / earlier bookings
Special Requests	Special requests correlate to lower cancellations, creating an opportunity to offer customizations for higher commitment
Previous Cancellations	Guests with a history of cancellation or more likely to do so, increasing the importance of collecting booking data

Opportunities for Improvement

Incorporate diverse hotel data and test ensemble methods
Increasing generalizability and capturing consumer complexity

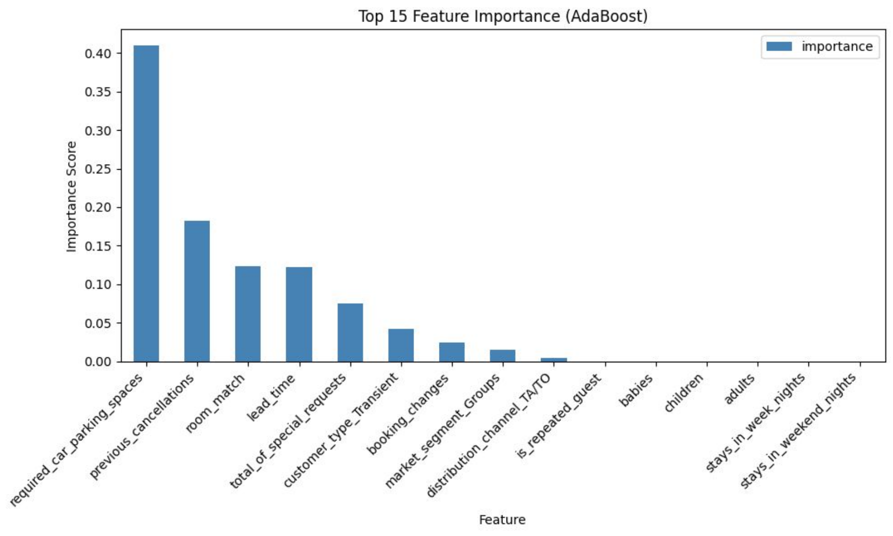
Thank you

Appendix

Boosting Tree

Train Accuracy Rate: ~74.5%

Test Accuracy Rate: ~74.23%



Forward Selection

Train Accuracy Rate: ~73.5%

Test Accuracy Rate: ~73.5%

Number of Features: 20

Top 5 Features

Feature	Type
lead_time	Numeric
total_of_special_requests	Numeric
customer_type_Transient	Categorical
room_match	Binary
market_segment_Groups	Categorical

Backward Selection

Train Accuracy Rate: ~73.4%

Test Accuracy Rate: ~73.1%

Number of Features: 20

Top 5 Features

Feature	Type
arrival_date_week_number	Numeric
stays_in_week_nights	Numeric
is_repeated_guest	Binary
previous_cancellations	Binary
previous_bookings_not_canceled	Binary

Logistic Regression

Train Accuracy Rate: 75%

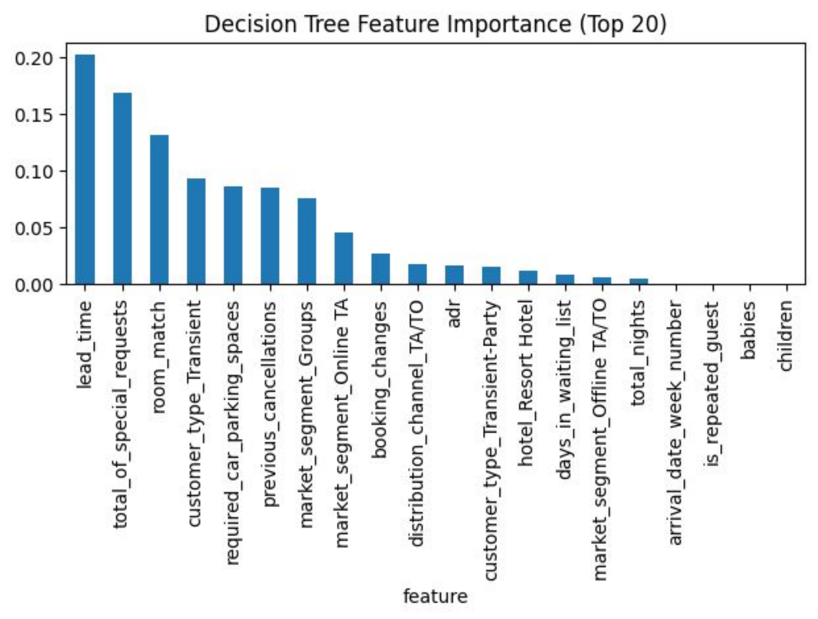
Test Accuracy Rate: 74.3%

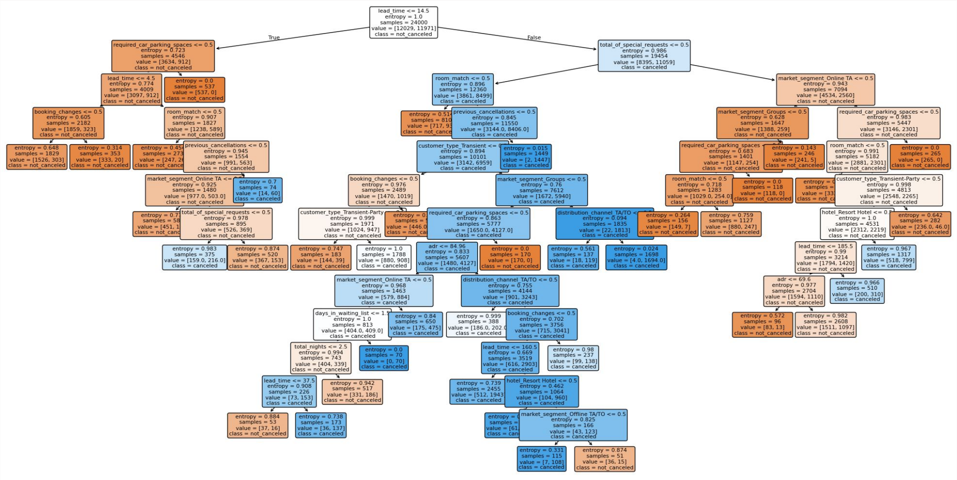
Number of Features: 20

Top 5 Features

Feature	Type
lead_time	Numeric
market_segment_Groups	Categorical
total_of_special_requests	Numeric
previous_cancellations	Binary
required_car_parking_spaces	Numeric

Decision Tree





Train Accuracy Rate: ~77.28%

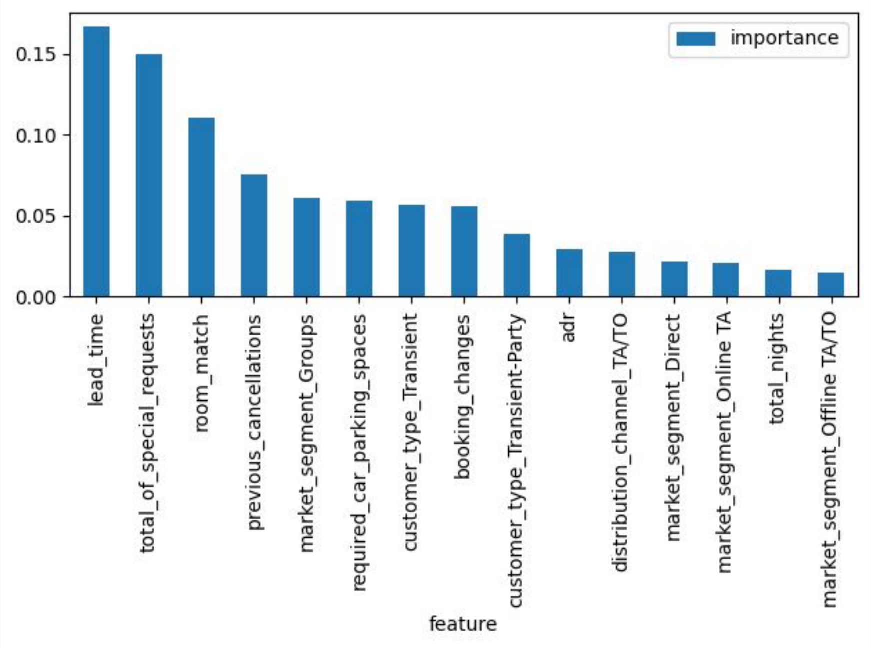
Test Accuracy Rate: ~77.5%

Main results
Random Forest

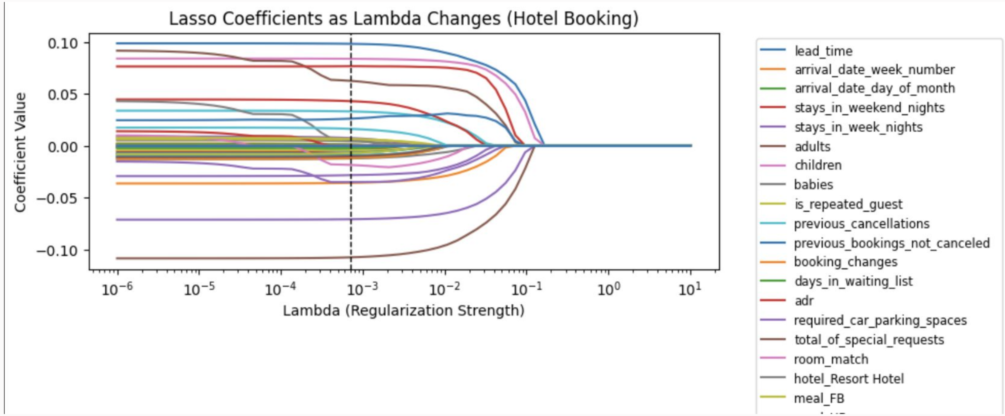
Train Accuracy Rate: ~80%

Test Accuracy Rate: ~79.2%

Important Features Of Random Forest



Lasso Regression



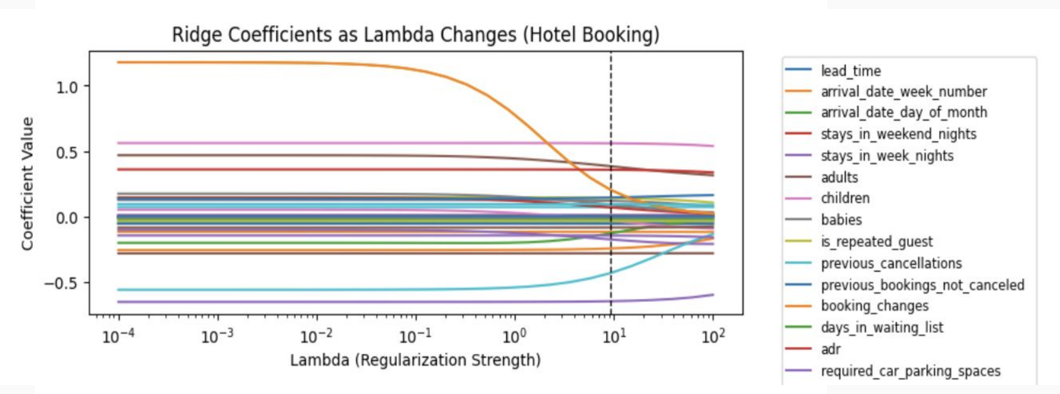
Train Accuracy Rate: ~73%

Test Accuracy Rate: ~73%

Top 5 Features

Feature	Type
deposit_type_non_refund	Binary
total_of_special_requests	Numeric
lead_time	Numeric
market_segment__online	Binary
assigned_room_type	Categorical

Ridge Regression



Train Accuracy Rate: ~74.3%

Test Accuracy Rate: ~74%

Top 5 Features

Feature	Type
required_car_parking_spaces	Numeric
room_match	Binary
distribution_channel_GDS	Numeric
market_segment_Groups	Categorical
customer_type_Transient	Categorical

KNN

Train Accuracy Rate: 77.74%

Test Accuracy Rate: 77.75%

Optimal Neighbors: 5
Grid Search CV

Data correlation

Key Drivers of the Outcome Variable

Positive Features

- Lead Time
- Room match
- Market Segment Groups

Negative Features

- Required car parking space
- Total Special Request