Final Project Report
Data Analytics

Project Title:
Analyzing Guest Behavior and Cancellations in Hotels: Insights to Optimize Performance

Prepared by:
Devi Sri Swetha Tanuku(N01623362)

AIGC 5000 – Fall 2024
Humber College

Problem Statement

The analysis aims to find out the behavior, preference, and pattern of hotel guests in order to optimize operations, reduce cancellations, and increase revenue.

We identify the best-performing country, in terms of guest bookings, and present its key drivers-stay duration, preferences, and trends.

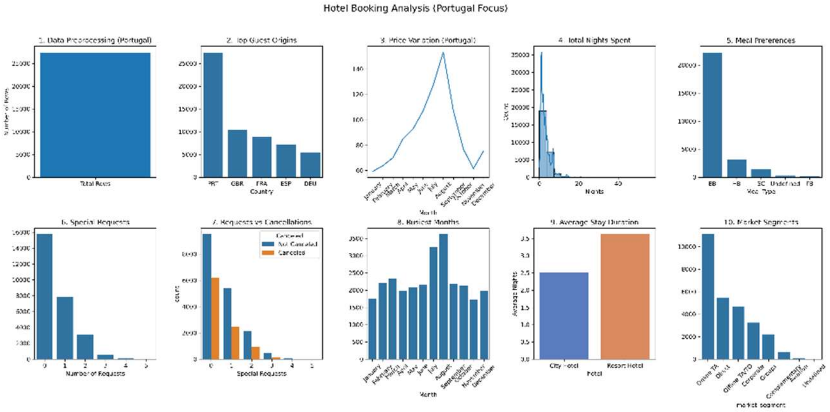
By fully understanding what makes the best-performing country the biggest market, we are able to offer concrete ways for other countries to achieve similar performance and attract more bookings.

Dataset Description

The Hotel Booking Dataset contains basic information about the guest booking at two hotel types, City Hotel and Resort Hotel. The dataset is made up of columns depicting guest booking behavior, stay information, preferences, and cancellations; hence, helping in establishing key trends and patterns.

Key Variables:hotel: Hotel type, arrival_date_month, country, total_nights, meal, adr, market_segment, is_canceled, total_of_special_requests

Dataset Analysis and Observations



Guest Origins:Portugal heads the bookings, providing the highest number of guests.

Price Variation:Prices are higher in August and July, thus summer seasons.

Nights Spent:1–4 nights spent by guests in both types of hotels.

Meal Preferences:Bed & Breakfast is the most chosen meal preference by Portuguese guests.

Special Requests:Most guests make 0 or 1 special request.

Requests vs Cancellations:Higher cancellations occur with 0 special requests, indicating less engagement.

Busiest Months:August and July are the busiest months for Portuguese guests.

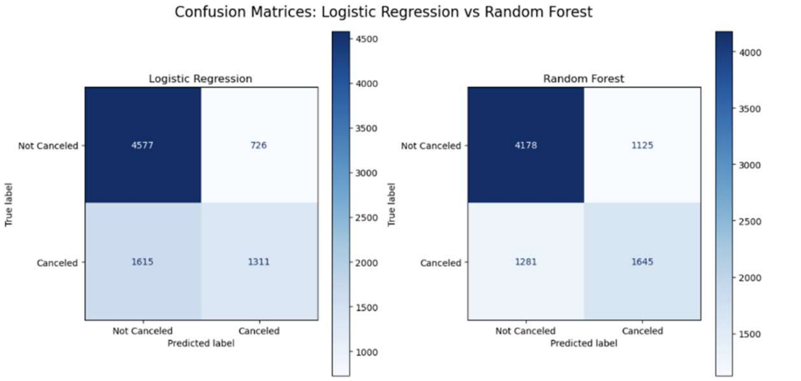
Average Stay Duration:Longer stays are slightly more common in Resort Hotels compared to City Hotels.

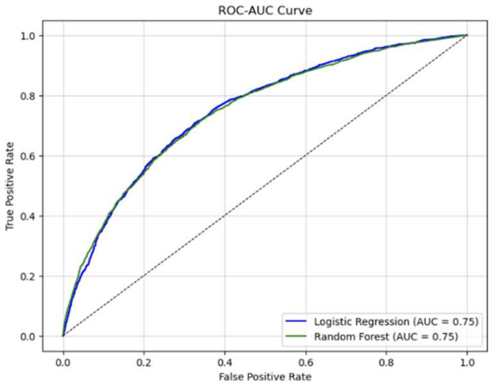
Market Segments:Most bookings come through Direct and Online Travel Agents (TA).

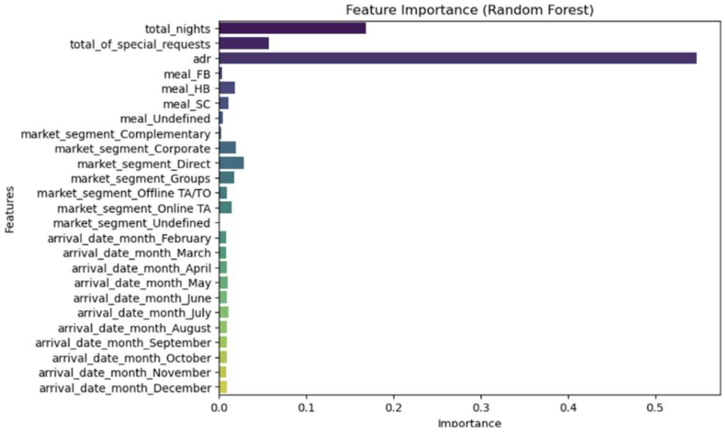
Proposed Analytical/Prediction Model

The major challenges, according to the analysis, would be cancellations, which affect the operations and revenues.That is why I've trained Logistic Regression and Random Forest Classifier to predict booking cancellation based on the following list of features:Total Nights Stayed, Special Requests, ADR: Average Daily Rate, Market Segment, Meal Preferences,Arrival Month.

For both models, I have trained and tested the dataset, evaluated their performance, and further compared metrics such as accuracy and ROC-AUC score for a proper understanding of predictive capabilities.







The Logistic Regression model outperforms the Random Forest Classifier, with:

Accuracy: 71.55%

ROC-AUC: 74.84%

Results and Discussions

Confusion Matrix for Logistic Regression and Random Forest models to show how the model has classified the bookings as canceled and non-canceled. From the matrix, I gather that Logistic Regression has slightly better accuracy and fewer misclassifications compared to Random Forest.

I show the ROC-AUC Curve for both models, comparing their ability to classify cancellations (1) and non-cancellations (0). Logistic Regression enjoys a score of 74.84%, and Random Forest, a score of 74.65%, indicating that the performance of Logistic Regression stands out, although only just.

I illustrate Feature Importance from the Random Forest model.

Key takeaways from the plot for feature importance: Fewer requests are strongly associated with higher cancellations. Higher prices influence cancellations. Shorter stays lead to more cancellations.

Give attention to increasing special requests to reduce cancellations. Consider a pricing strategy for guests with short stays to improve retention. Target high-cancellation market segments with better engagement strategies.