Prediction of Hotel Booking Cancellations for Operational Optimization

Andrés Padrón Quintana, Juan Guillermo Palestino Martínez, Cuahtemoc Maya Maldonado, Gabriel Ángel Laurel Membrillo

What factors influence booking cancellations, and how can the hotel reduce its cancellation rate?

Objectives

- Identify the factors that predict booking cancellations.
- Propose data-driven strategies to reduce cancellations
- Design a predictive model that helps anticipate cancellations to allow the hotel to optimize its occupancy and increase revenues

Problem Description

Holiday Inn, a hotel chain with global presence, is experiencing a growing problem with booking cancellations, which has led to significant financial losses. While there is a cancellation penalty for the customer, and part of the reservation amount is recovered, the potential revenue from the guests' stay at the hotel, including additional services offered by the chain, is not recovered.

For this reason, they provided us with a database that includes 120,000 observations of two types of hotels in the chain: City Hotel and Resort Hotel.

Feature	City Hotel	Resort Hotel
Location	Urban centers, business districts	Tourist destinations, beaches, countryside
Main Purpose	Business trips, short-term stays	Leisure, vacations, family or couple getaways
Booking Lead Time	Shorter, often last-minute	Planned in advance
Customer Profile	Business travelers, solo travelers	Families, couples, groups

The reservations are delimited between July 1, 2015, and August 31, 2017. It contains 36 variables that describe each of the reservations:

- *lead_time:* Number of days between the booking date and the arrival date.
- *total_stays:* Total number of nights booked.
- *booking_changes:* Number of changes made to the reservation.
- *previous_cancellations:* Number of previous cancellations by the customer.
- *hotel:* Type of hotel (resort or city).
- *deposit_type:* Type of deposit (none, refundable, or non-refundable).
- *is_repeated_guest:* Indicates whether the customer has visited the hotel previously.
- *customer_type:* Classification of the customer type (contract, group, transient, or transient-party).
- *market_segment:* Channel through which the reservation was made (direct, corporate, online, etc.).
- *arrival_date_year, arrival_date_month, arrival_date_day_of_month:* Customer's arrival date.

Methodology

Exploratory Data Analysis (EDA)

The data exploration revealed that approximately 37% of the reservations were canceled (*is_canceled* = 1). This confirms that cancellations are a significant problem for the hotel chain. At the same time, City Hotels have a higher proportion of cancellations compared to Resort Hotels.

One hypothesis is that urban hotels might have more impulsive or less planned reservations, while resorts might have reservations made further in advance. At the same time, a greater lead_time was observed in canceled reservations compared to those that were not canceled, suggesting that reservations made well in advance have a higher probability of cancellation.

Data cleaning

Before running any models, we focused on transforming the raw dataset into a clean and structured one. This included removing incomplete or irrelevant columns, imputing missing values in a statistically sound way, and creating aggregated variables like *total_stays* and *total_guests*, which improved the model's ability to capture key behavioral patterns. By filtering

out outliers and encoding categorical variables, we ensured that all algorithms would perform optimally without being misled by noise.

During data cleaning, we initially removed variables like company, agent, credit_card, phone_number, and name due to high proportions of missing g values or lack of predictive power. However, after conducting exploratory analysis and correlation checks, we decided to retain and use key variables such as deposit_type, lead_time, and hotel because they demonstrated strong explanatory value for cancellations.

New Variable	Description	Reason for Creation
total_stays	Sum of stays_in_week_nights and stays_in_weekend_nights	Captures the total length of stay in a single variable
total_guests	Sum of adults, children, and babies	Reflects total group size, which affects cancellation behavior
is_canceled (factor)	Recoded from 0/1 to "Canceled" and "Not Canceled" for interpretability	Improves visualization and model output clarity
lead_time (numeric)	Converted to numeric type (if factor) to allow calculations and binning	Enables quantitative filtering and plotting

Models

Logit, Probit, Lasso, and Random Forest models were used to predict the factors that most influence booking cancellations:

Model	Accuracy	Sensitivity	Specificity	Key Takeaway
Logit	0.762	0.979	0.393	Good at detecting non-cancellations, poor specificity
Probit	0.758	0.979	0.383	Similar to Logit, slightly worse performance
LASSO	~0.76	Balanced	Balanced	Penalizes irrelevant features, more interpretable
Random Forest	0.771	0.389	0.996	Excellent at confirming bookings, misses cancellations

Results Analysis

Logit, Lasso, and Probit showed very similar performance. All these models have high sensitivity, indicating that they are effective at correctly identifying cancellations; however, their specificities are low, meaning they have difficulty correctly identifying reservations that were not canceled. In contrast, Random Forest showed higher accuracy compared to the other models. Although its sensitivity is very low, its specificity is extremely high, indicating that this model is very effective at correctly predicting non-canceled reservations, but has difficulty correctly identifying cancellations..

If the primary goal is to correctly identify cancellations, Logit, Probit, or Lasso are preferable due to their high sensitivity. However, if the goal is to minimize false positives in the predictions, the Random Forest model stands out due to its high specificity.

Based on the Random Forest model we can see that:

Variable	Interpretation
deposit_type	The most important variable: the type of deposit strongly influences cancellation likelihood.
lead_time	Bookings made well in advance are more likely to be canceled.
previous_cancellations	Customers who have canceled before are more likely to cancel again.
hotel	The hotel type also plays a role, but its influence is significantly lower.

Factors Influencing Booking Cancellations

Based on the results obtained and the analysis of variable importance using the Random Forest model, the factors that have the greatest influence on booking cancellations are::

- deposit_type (Type of Deposit)

Reservations with a "Non Refund" deposit have a significantly lower probability of cancellation, as customers have a greater financial commitment from the beginning.

Recommended action: the hotel can incentivize the use of non-refundable deposits through discounts or attractive offers to encourage this type of reservation.

- previous_cancellations (Previous Cancellations)

Customers with a history of cancellations are more likely to cancel again.

Recommended action: the hotel could implement restrictions or additional conditions for customers with a high number of previous cancellations, such as requiring an advance payment or limiting flexibility in changes.

- lead_time (Booking Lead Time)

Reservations made far in advance have a higher risk of cancellation.

Recommended action: implementing tiered cancellation policies, where penalties increase as the check-in date approaches, could reduce the number of late cancellations.

- total_stays (Total Length of Stay)

Longer reservations have a lower probability of cancellation.

Recommended action: offering incentives for longer stays, such as progressive discounts, can help reduce cancellations.

- booking_changes (Changes to the Reservation)

A higher number of modifications to the reservation is related to a higher probability of cancellation.

Recommended action: establishing reasonable limits on the number of changes allowed or charging a fee for excessive modifications could discourage this behavior.

Economic Approach

Assuming the hotel processes 5,000 bookings per month and the average revenue per booking is $150. Based on the results obtained in the analysis, these may be some of the effects that we can predict:

1) A 1% reduction in cancellations (50 recovered bookings), then $7,500 in additional revenue/month.
2) A 5% reduction ($37,500 per month), then over $450,000/year in retained revenue.

Why Use Machine Learning for a "Simple" Cancellation Problem?

While it's true that some cancellation patterns (like long lead times) seem intuitive, human intuition can't evaluate dozens of variables simultaneously, quantify risk probabilities, or adapt automatically to new data. Machine Learning allows us to turn business intuition into quantifiable, scalable, and automated decision-making. For example:

- With Logit, we can state that a customer with lead_time > 90 days and deposit_type = None has an 85% probability of canceling, allowing proactive response.
- With Random Forest, we can build a real-time alert system to flag reservations most likely to cancel and prioritize them for reconfirmation or overbooking strategy.

This is why Holiday Inn should hire us: our model doesn't just predict cancellation, it delivers actionable, data-driven insights to optimize operations and revenue, something no manual review or simple dashboard can achieve at this scale.

References

- Mojtaba. (2022). *Hotel Booking Demand Dataset* [Dataset]. Kaggle. https://www.kaggle.com/datasets/mojtaba142/hotel-booking
- Tudón Maldonado, J. (2025). Slides, Readings, and Data Lists. Dropbox. . (2025). Slides, Readings, and Data Lists. Dropbox. https://www.dropbox.com/scl/fi/c0doxdvcnclk46fh8n5us/slides_readings_and_data_lists.pdf?rlkey=hrtgbkvyzodqh3fav1xc86m03
- Statista. (2023). *Holiday Inn hotels gross revenue worldwide from 2009 to 2023*. Retrieved from https://www.statista.com/statistics/223345/holiday-inn-hotels-revenue/
- OpenAI. (2025). *ChatGPT* (versión GPT-4o) [Modelo de lenguaje]. https://chat.openai.com/