Prediction of Cancellations in Hotel Reservations for Operational Optimization

Data-Driven Business Decisions

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


Holiday Inn®
TM

Why cancellations are a problem?

Cancellations cause:

- Daily actualized income to drop below projections when rooms cannot be re-booked.
- Forced price reduction to fill vacancies.
- Increased distribution costs to sell those vacant rooms.

Predicting cancellations to decrease uncertainty and increase revenue

Objectives:

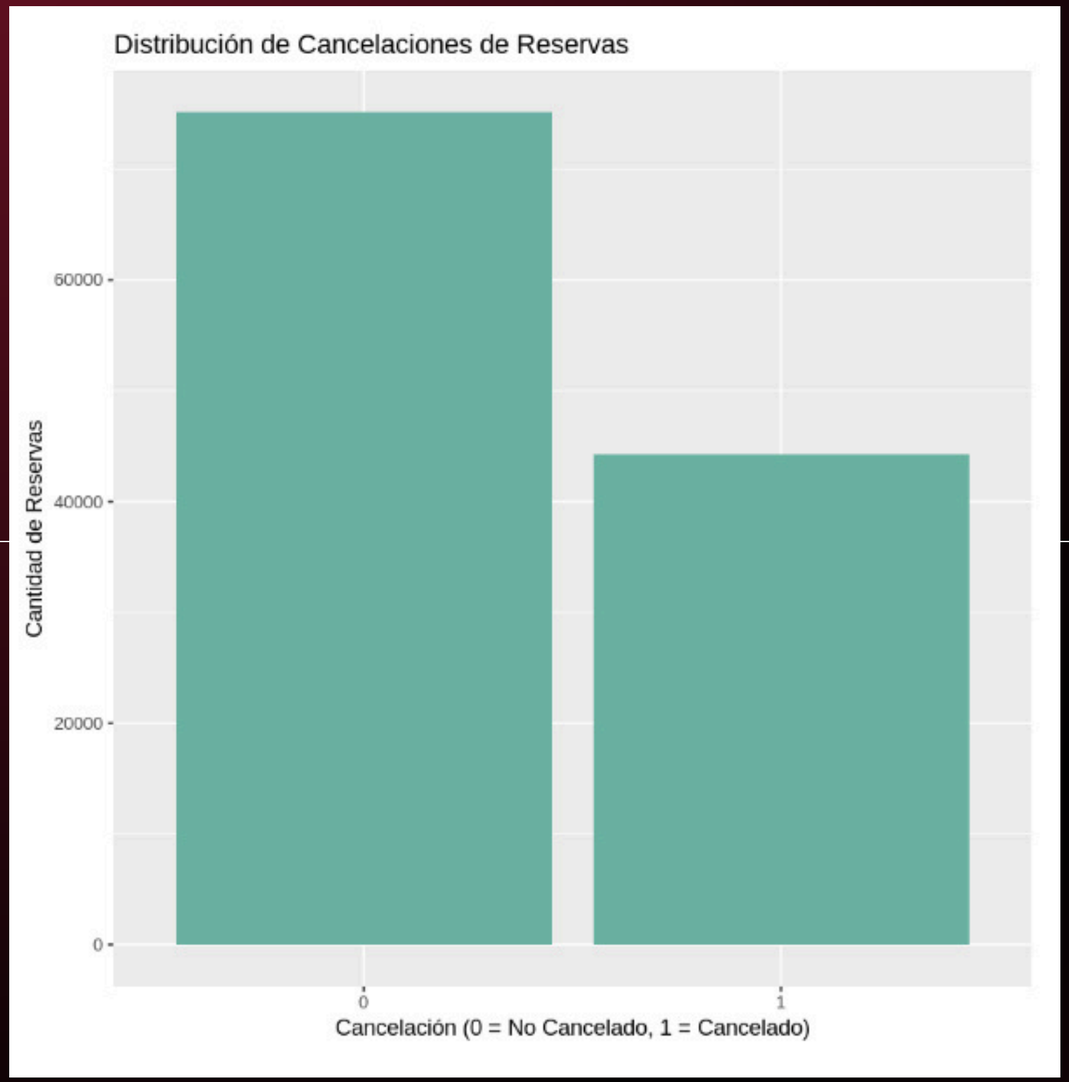
- Identify the factors that predict booking cancellations.
- Propose data-driven strategies to reduce cancellations.
- Design a predictive model to anticipate cancellations, enabling the hotel to optimize occupancy and increase revenue.

Data and Variables

- Observations: 120 000 reservations (2015–2017)
- Hotel Type: City and Resort
- Key Variables:
  - *lead_time*: Days between booking date and arrival date
  - *total_stays*: Total nights booked
  - *booking_changes*: Number of changes made to the booking
  - *previous_cancellations*: Previous cancellations by the customer
  - *hotel*: Type of hotel (Resort or City)
  - *deposit_type*: Type of deposit (None, Refundable, Non-Refundable)
  - *is_repeated_guest*: Whether the customer has previously visited the hotel
  - *customer_type*: Customer classification (Contract, Group, Transient, or Transient-Party)
  - *market_segment*: Booking channel (Direct, Corporate, Online, etc.)
  - *arrival_date*: Arrival date (year, month, and day)

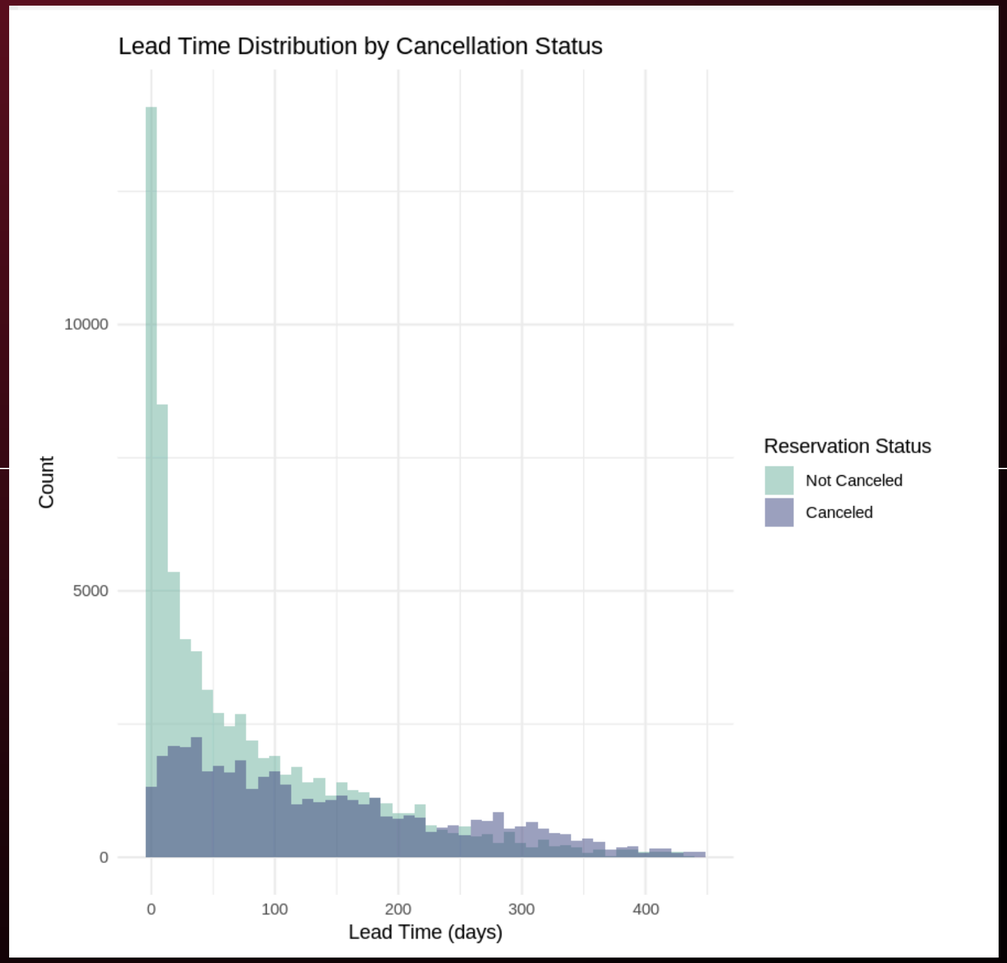
City vs Resort

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



37% of the reservations were canceled

Cancellations are a significant problem for the hotel. This proportion suggests that there is enough information to identify patterns that can help predict and prevent future cancellations.

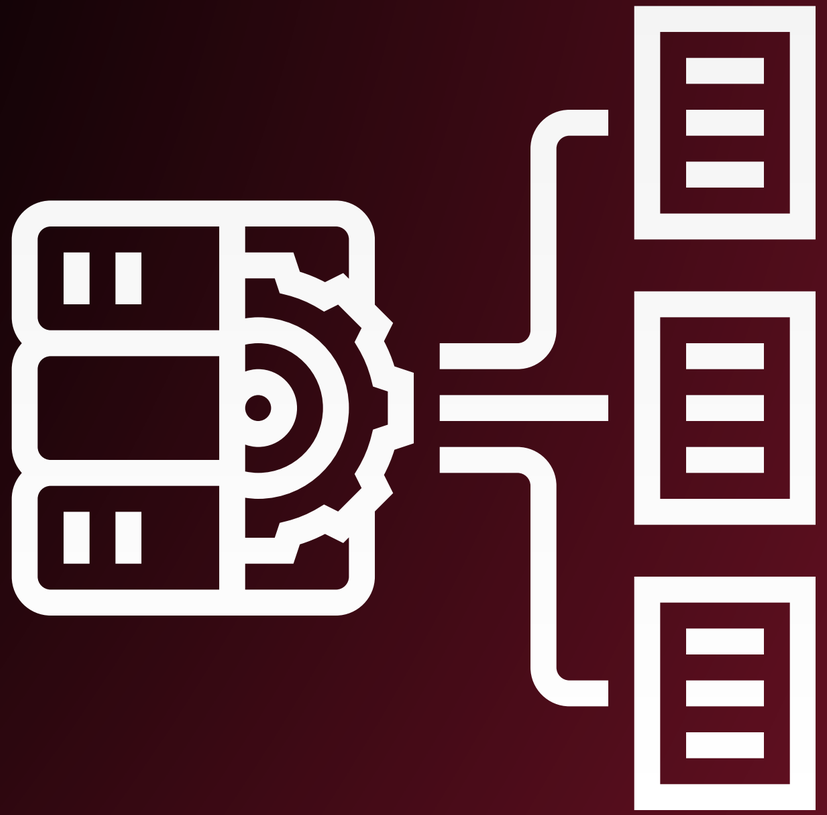


Cancellations tend to have a significantly higher lead time compared to non-canceled bookings.

- Customers who booked closer to the arrival date were much more likely to show up.
- Bookings made well in advance had a higher probability of cancellation.
- Overlap exists, but the distribution differs clearly:
- Long lead time → Higher cancellation risk.
- Short lead time → Lower cancellation risk.

Data Cleaning

- **Elimination of Null Values**:
  - *company* and *agent* columns were removed due to too many null values.
- **Elimination of Irrelevant Data**:
  - name, phone_number, and credit_card were removed due to a lack of predictive value.
- **Conversion of Categorical Data**:
  - Categorical variables were converted into factors for compatibility with R models..
- **Creation of New Variables:**
  - total_stays and total_guests were created to improve the predictive capability of the model.



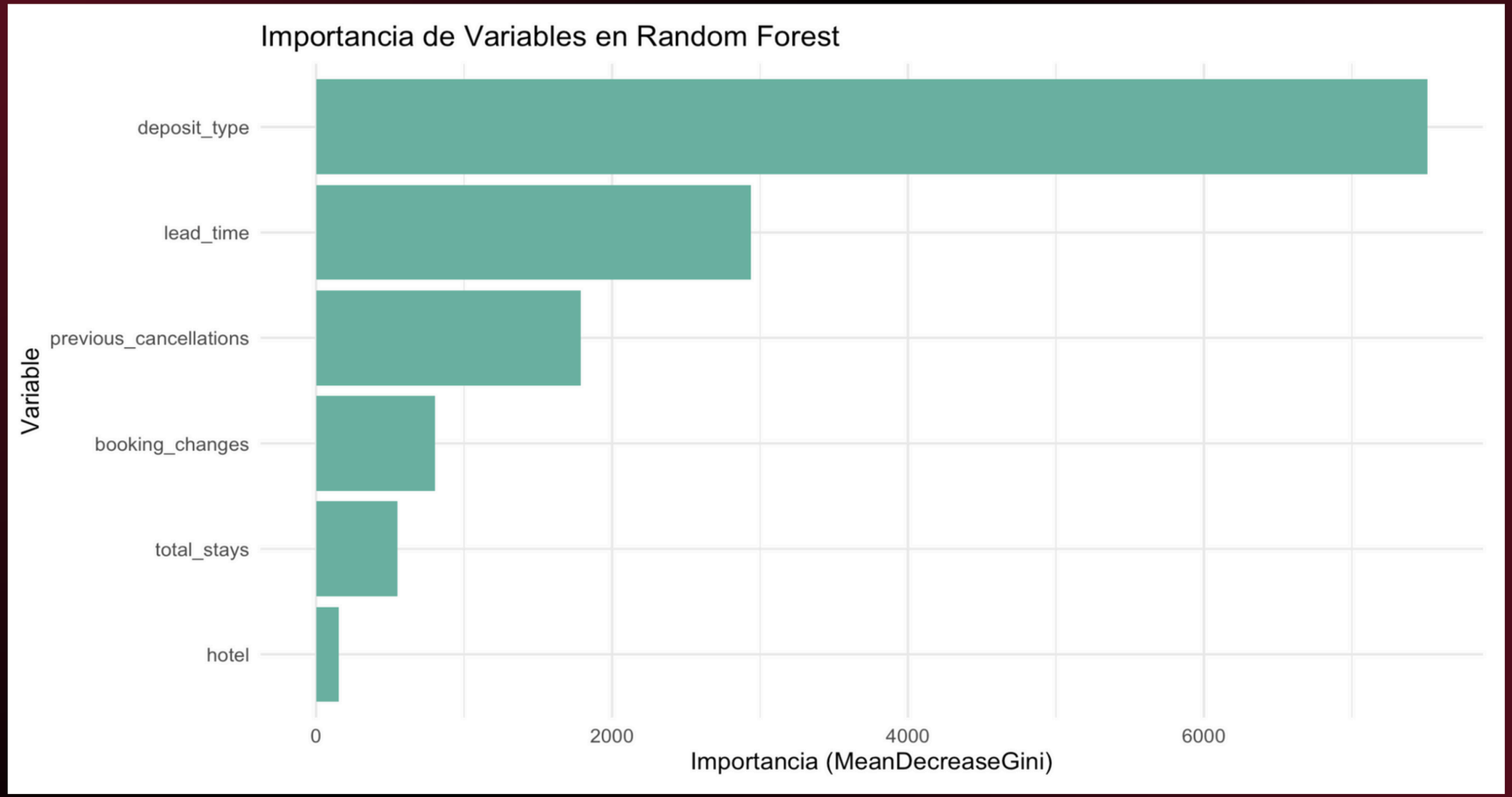
New Variables

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

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
Random Forest	0.771	0.389	0.996	Excellent at confirming bookings, misses few cancellations

Predictive Models

- Random Forrest is the best one:
  - RF is the most accurate in predicting for cancelled bookings
  - RF has a higher specificity, so it has a lower type I error rate.


Importancia de Variables en Random Forest
deposit_type
lead_time
previous_cancellations
booking_changes
total_stays
hotel
Variable
0
2000
4000
6000
Importancia (MeanDecreaseGini)

Key Factors

- **Deposit Type**: Non-refundable deposits reduce cancellations.
- **Previous Cancellations**: Customers with a history of cancellations are more likely to cancel again.
- **Lead Time**: Reservations made well in advance are more likely to be canceled.
- **Length of Stay:** Longer stays have a lower probability of cancellation.
- **Booking Changes**: A higher number of booking changes increases the risk of cancellation.

Economic Approach

- Assuming the hotel processes 5,000 bookings per month and the average revenue per booking is $150
- A 1% reduction in cancellations = 50 recovered bookings → $7,500 in additional revenue/month.
- A 5% reduction = $37,500/month → Over $450,000/year in retained revenue.

Extra revenue from reservations that will not be canceled

- Food and Beverage: Each reservation typically generates $30 to $50 per night in dining revenue.
- Transportation and Tours: Those who book tend to spend $40 to $100 on transportation and local experiences.
- Spa and Wellness: Non-canceled reservations can bring in $100 to $200 through spa services.
- Room Upgrades: Early detection of likely cancellations allows hotels to reallocate premium rooms, potentially increasing revenue by $100 to $300 per reservation.

Scenario	EDA Only (Descriptive)	Random Forest (Predictive)
Current Cancellation Rate	37%	32% - 27% (expected 5-10% reduction)
Monthly Bookings	5,000	5,000
Average Revenue per Booking	$150 per booking	$150 per booking
Monthly Revenue	$472,500	547,500
	Extra revenues (commodities)	62,500
	Total monthly difference	127,500 (1.27)

Why This Model Adds Business Value

- Automated decision support: real-time cancellation predictions for operational planning.
- Customers: personalize incentives for high-risk profiles.
- Scalable: can be deployed across multiple hotel brands and locations.
- Cost-saving potential: optimizes room inventory and staffing.
- Actionable insights: identifies why customers cancel, not just who cancels.

Why Use Machine Learning for a “Simple” Cancellation Problem?

Traditional Approach	Machine Learning Approach
Manual review of bookings	Automated prediction at scale
Based on intuition or rules of thumb	Based on statistical patterns across 30+ variables
Difficult to adapt to changing trends	Continuously adaptable to new data
No quantification of cancellation risk	Outputs precise cancellation probabilities (e.g., 85%)
Limited to basic reports	Enables real-time alerts and strategic overbooking


RECEPTION

References

- 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&e=2&dl=0

- Mojtaba. (2022). Hotel Booking Demand Dataset [Dataset]. Kaggle. https://www.kaggle.com/datasets/mojtaba142/hotel-booking

- Statista. (2023). Holiday Inn hotels gross revenue worldwide from 2009 to 2023. Retrieved from https://www.statista.com/statistics/223345/holiday-inn-hotels-revenue/