

Hotel Booking Cancellation Predictive Modeling: A Comprehensive Report on Variable Selection and Data Acquisition

Table of Contents

- 1. Introduction
- 2. Scope and Objectives
- 3. Background and Importance
- 4. Description of Selected Variables
 - Lead Time
 - Arrival Date Week Number
 - Arrival Date Day of Month
 - Stays in Weekend Nights
 - Stays in Week Nights
 - Adults
 - Children
 - Babies
 - Is Repeated Guest
 - Previous Cancellations
 - Previous Bookings Not Canceled
 - Required Car Parking Spaces
 - Total of Special Requests
 - Average Daily Room Rate
 - Hotel
 - Arrival Date Month
 - Meal
 - Market Segment
 - Distribution Channel
 - Reserved Room Type
 - Deposit Type
 - Customer Type
 - Is Canceled
- 5. Data Sources and Acquisition
- 6. Hypothetical Team Collaboration and Data Collection Process
 - Kickoff Meeting
 - Data Inspection and Cleaning Session
 - Final Data Readiness Meeting
- 7. Data Quality Checks and Validation
- 8. Data Cleaning Methodology
- 9. Documented Challenges and Resolutions
- 10. Regulatory and Compliance Considerations
- 11. Conclusion and Next Steps
- 12. Appendices
- 13. References

1. Introduction

Hotel bookings are a major component of the global travel and hospitality industry, and **predicting whether a booking will be canceled** has become critically important for revenue management, customer service planning, and operational efficiency. Overbooking strategies, staffing requirements, and marketing efforts are all influenced by the likelihood of cancellations.

This formal report centers on the **key variables** used in a machine learning (ML) model that predicts hotel booking cancellations. By elucidating the meaning, range, and potential data sources for each variable, it provides a reference framework for data professionals, data scientists, product owners, and hotel operations teams. Ultimately, a deeper understanding of these variables can enhance modeling accuracy, enabling more profitable and customer-friendly decision-making.

2. Scope and Objectives

- 1. **Scope**
 - Offers a comprehensive overview of the data fields—both **numerical** (e.g., lead_time, number of adults) and **categorical** (e.g., hotel, meal, deposit_type).
 - Discusses the **significance** of each variable in the context of predicting cancellations.
- 2. **Objectives**
 - Clarify the **purpose** and **meaning** of each feature.
 - Outline the plausible **range** or **list of allowable values** for each field.
 - Provide domain context relevant to data scientists and hotel management personnel.

- Present an **official record** of the data sources, typical data collection methods, and hypothetical discussions between team members throughout the data acquisition and preparation phase.

3. Background and Importance

With the growth of **online travel agencies (OTAs)**, direct booking platforms, and evolving consumer habits, hoteliers face rising cancellation rates. Understanding **which factors** make a booking likely to result in cancellation or retention is central to:

- **Revenue Management:** Overbooking rooms can safeguard revenue but also risks negative customer experiences if it leads to forced relocations.
- **Operational Planning:** Cancellations impact staffing, inventory management, and service availability. Predictive modeling can help streamline day-to-day operations.
- **Marketing and Promotions:** Knowing the customer segments most prone to cancel allows targeted promotions or flexible booking policies to **mitigate** no-shows.

This set of variables captures dimensions such as booking timelines (lead_time), guest composition (adults, children, babies), booking history (is_repeated_guest, previous_cancellations), and financial aspects (average_daily_room_rate). Each dimension adds invaluable context to the puzzle of **forecasting cancellations**.

4. Description of Selected Variables

Below, each variable is presented with its type, recommended range or category set, and relevance for modeling. Understanding these fields is crucial for data validation, feature engineering, and algorithm training.

4.1. Lead Time (INT, Range [0, 730])

- **Description:** The number of days between the booking date and the arrival date.
- **Business Rationale:** Longer lead times can increase the chance of cancellation as guests' plans are more likely to change over a longer period. Conversely, last-minute bookings might be more certain but can be susceptible to sudden changes.
- **Modeling Considerations:** Often correlated with seasonal trends, promotions, and booking behaviors. Can be bucketed into short lead times (0–7 days), medium (8–30), or long (30+) for feature engineering.

4.2. Arrival Date Week Number (INT, Range [1, 53])

- **Description:** Identifies which week of the year the guest arrives.
- **Business Rationale:** Useful for capturing seasonal peaks or lows. Weeks closer to holidays (e.g., weeks 51–52 near Christmas) or high-travel seasons can see higher or lower cancellation rates.
- **Modeling Considerations:** The model can detect patterns tied to specific weeks, possibly combined with year-level data or month to account for holiday shifts.

4.3. Arrival Date Day of Month (INT, Range [1, 31])

- **Description:** The specific day of the month on which the guest arrives.
- **Business Rationale:** Some travelers may habitually book certain dates (e.g., end of the month, payday cycles). Payment schedules or personal travel patterns might influence cancellations.
- **Modeling Considerations:** May be combined with the month feature to form a complete temporal context.

4.4. Stays in Weekend Nights (INT, Range [0, 26])

- **Description:** Number of weekend nights (Saturday or Sunday) included in the stay.
- **Business Rationale:** Weekend stays can be leisure-related; leisure bookings may have different cancellation behaviors than business-oriented weekday stays.
- **Modeling Considerations:** Large weekend stays could correspond to special events or holiday packages. May be combined with total nights to see if extended stays occur over weekends.

4.5. Stays in Week Nights (INT, Range [0, 65])

- **Description:** Number of weekday nights (Monday to Friday) in the booking.
- **Business Rationale:** Frequent among corporate or business travelers. Business travelers can be subject to last-minute schedule changes or corporate policies, influencing cancellations.
- **Modeling Considerations:** Sum of weekday and weekend nights often gives total length of stay. Could be a key feature for revenue analysis as well.

4.6. Adults (INT, Range [1, 5])

- **Description:** Total number of adult guests in the booking.
- **Business Rationale:** More adults might indicate family trips or group travel, each with different cancellation risks. Solo travelers might exhibit distinct patterns as well.
- **Modeling Considerations:** Used to gauge occupancy. If "adults" is zero (which shouldn't occur in practice), it might indicate data errors or special cases.

4.7. Children (INT, Range [0, 10])

- **Description:** Number of children in the booking.
- **Business Rationale:** Family vacations can be more sensitive to external factors (e.g., health or school schedules) that lead to cancellations.
- **Modeling Considerations:** Combined with adults and babies to approximate total occupancy, which can be cross-checked with room types or rates.

4.8. Babies (INT, Range [0, 10])

- **Description:** Number of babies included in the booking.

- **Business Rationale:** Traveling with infants may involve additional uncertainties, such as health considerations, potentially influencing cancellation decisions.
- **Modeling Considerations:** Could be grouped with “children” or considered separately, depending on the modeling approach.

4.9. Is Repeated Guest (INT, Range [0, 1])

- **Description:** Indicates whether the booking is from a returning guest (1) or a new guest (0).
- **Business Rationale:** Repeat guests often have established loyalty or familiarity with the property, potentially reducing cancellation likelihood. Conversely, a first-time guest may be more likely to change plans.
- **Modeling Considerations:** Typically encoded as a binary variable. Strong indicator of loyalty and reliability.

4.10. Previous Cancellations (INT, Range [0, 30])

- **Description:** Total number of canceled bookings by this guest prior to the current booking.
- **Business Rationale:** Past behavior is often predictive of future behavior. A guest with multiple prior cancellations could be more likely to cancel again.
- **Modeling Considerations:** May influence risk-scoring systems, allowing hotels to overbook or secure additional deposits if necessary.

4.11. Previous Bookings Not Canceled (INT, Range [0, 80])

- **Description:** Total number of successful stays (non-canceled bookings) by this guest.
- **Business Rationale:** A history of bookings honored to completion is a sign of reliability. High values suggest loyalty and trustworthiness, typically lowering the cancellation risk.
- **Modeling Considerations:** Helps differentiate guests who have a history of stable bookings from those with uncertain patterns.

4.12. Required Car Parking Spaces (INT, Range [0, 8])

- **Description:** Number of car parking spaces requested.
- **Business Rationale:** Guests who require parking (especially multiple spaces) may be traveling by car and thus have set plans, possibly indicating lower cancellation risk. However, certain large-event travelers could also frequently revise or cancel if plans change.
- **Modeling Considerations:** Typically indicative of local or regional guests. Could be a minor predictor but still relevant for operational logistics.

4.13. Total of Special Requests (INT, Range [0, 5])

- **Description:** Number of special requests made (e.g., extra bed, early check-in, late checkout).
- **Business Rationale:** Guests who articulate many preferences may be **more committed** to the trip, or in some cases, they might be more **finicky** and prone to dissatisfaction.
- **Modeling Considerations:** Often correlated with certain guest profiles (families, corporate VIPs). Impact on cancellation can be nuanced.

4.14. Average Daily Room Rate (FLOAT, Range [1, 540])

- **Description:** Mean cost per room per night, possibly measured in local currency units.
- **Business Rationale:** Higher average rates might indicate premium stays (special packages) or surge pricing in peak seasons. Premium bookings might see lower or higher cancellation rates depending on guest profiles and the flexibility of those bookings.
- **Modeling Considerations:** Price sensitivity is a significant factor—some guests might cancel if they find a cheaper deal elsewhere. Rate fluctuations can also be seasonal.

4.15. Hotel (CAT, Values: 'City Hotel', 'Resort Hotel')

- **Description:** The category/type of hotel.
- **Business Rationale:** City hotels often cater to business travelers, while resort hotels may attract holiday or leisure travelers. These two segments can have drastically different booking behaviors and cancellation rates.
- **Modeling Considerations:** Often used as a **segmentation** variable. Models may train separate sub-models per hotel type or use this as a feature for a single unified model.

4.16. Arrival Date Month (CAT, Values: Various Months)

- **Description:** Categorical variable indicating the month of arrival.
- **Business Rationale:** Directly ties into **seasonality**. High-demand months for tourism or business events could see different cancellation patterns compared to off-peak periods.
- **Modeling Considerations:** One-hot encoding or cyclical encodings may help the model capture month-related patterns.

4.17. Meal (CAT, Values: 'SC', 'BB', 'HB', 'FB')

- **Description:** Type of meal plan chosen: SC (no meal), BB (bed and breakfast), HB (half board), FB (full board).
- **Business Rationale:** Guests choosing certain meal options might be either cost-conscious or convenience-focused. Meal plan selection can reflect the length of stay or trip purpose, thus influencing likelihood to cancel.
- **Modeling Considerations:** May correlate with total booking value and guest profiles (families vs. business travelers).

4.18. Market Segment (CAT, Values: e.g., 'Online TA', 'Direct', etc.)

- **Description:** The market segment classification describing how the booking was made.
- **Business Rationale:** Different segments (corporate, groups, online travel agency) have distinct cancellation behaviors. Group bookings may be less likely to cancel individually but can result in block cancellations; online TA bookings sometimes face higher cancellation rates due to flexible policies.
- **Modeling Considerations:** Highly relevant feature for e-commerce-based cancellation analysis. Often combined with `distribution_channel`.

4.19. Distribution Channel (CAT, Values: e.g., 'Corporate', 'Direct', 'TA/TO', etc.)

- **Description:** The channel used for the booking, such as direct hotel website, corporate agreements, travel agents, or tour operators.
 - **Business Rationale:** Similar to market segment but more specific to *how* the booking is processed. Different channels have different booking conditions (cancellation fees, deadlines).
 - **Modeling Considerations:** Supports segment-level forecasting. Channels with flexible cancellation policies might see higher cancellation volumes.
- 4.20. Reserved Room Type (CAT, Values: e.g., 'A', 'B', 'C', etc.)**
- **Description:** The assigned internal code for the room type.
 - **Business Rationale:** Certain room types (e.g., suites or premium categories) might have different cancellation propensities compared to standard rooms. Rate and deposit requirements can also vary by room category.
 - **Modeling Considerations:** Ensures the model captures relationships between room choice and cancellation risk.
- 4.21. Deposit Type (CAT, Values: 'Non Refund', 'Refundable', 'No Deposit')**
- **Description:** The deposit policy linked to the booking (full non-refundable, partially refundable, or no deposit at all).
 - **Business Rationale:** A **non-refundable** deposit typically reduces cancellations since guests risk losing their payment. Conversely, no-deposit policies can invite more speculative bookings.
 - **Modeling Considerations:** Strong predictor, as deposit policies directly influence the financial implications of canceling.
- 4.22. Customer Type (CAT, Values: 'Contract', 'Group', 'Transient', 'Transient-Party')**
- **Description:** Classifies the customer's booking context:
 - **Contract:** Bookings with stable deals between businesses or institutions and the hotel.
 - **Group:** Often part of a larger reservation block for events, tours, etc.
 - **Transient:** Individuals or small parties booking independently for short stays.
 - **Transient-Party:** Similar to transient but includes additional guests.
 - **Business Rationale:** Contract or group travelers might be less prone to cancel due to corporate or group coordination. Transient guests often have more freedom to cancel.
 - **Modeling Considerations:** Distinct patterns in each type can significantly affect cancellation likelihood.
- 4.23. Is Canceled (INT, Range [0, 1])**
- **Description:** The target variable—whether the booking was canceled (1) or kept (0).
 - **Business Rationale:** Central measure for the model. This binary outcome helps hotels forecast potential losses or occupancy shortfalls.
 - **Modeling Considerations:** All other variables aim to predict this outcome. Balanced or imbalanced class issues may arise if the cancellation rate is significantly lower or higher than 50%.
-

5. Data Sources and Acquisition

Typical data pipelines for these variables include:

1. **Property Management Systems (PMS):** Real-time databases that track reservations, modifications, arrivals, departures, and payments.
 2. **Channel Manager Software:** Aggregates data from various distribution channels (OTAs, direct websites, GDS) and updates availability and pricing.
 3. **CRM Systems:** Stores guest profiles, loyalty status, and historical booking/cancellation behavior.
 4. **Manual Logs or Additional Spreadsheets:** In smaller hotels or for special events, staff might update records by hand, later integrating them into the primary database.
-

6. Hypothetical Team Collaboration and Data Collection Process

Below is a fictional yet representative account of how various teams might collaborate to assemble and validate data for this predictive model.

6.1. Kickoff Meeting

Attendees:

- Alice Kim (Lead Data Scientist)
- Jorge Martinez (Hotel Operations Manager)
- Linda Green (Data Engineer)
- Dr. Miriam Ross (Revenue Management Specialist)

Key Discussion Points:

- Alice Kim introduced the project scope, emphasizing that accurately forecasting cancellations could drive improvements in overbooking strategies and resource planning.
- Jorge Martinez highlighted data availability from the property management system. However, he flagged potential inconsistencies in manual logs for older or smaller properties.
- Dr. Miriam Ross stressed the importance of deposit policy data, as she observed strong correlations between deposit types and no-show rates.
- Linda Green discussed integration challenges and the necessity of a well-defined data schema.

Outcomes:

- Agreement to create a standardized data dictionary (covering lead_time, is_repeated_guest, etc.).
- Timeline for extracting historical records, focusing on the past three years of data.

6.2. Data Inspection and Cleaning Session

Attendees:

- Linda Green (Data Engineer)
- Eva Sanders (Junior Data Scientist)
- Sarah Wei (Business Analyst)

Key Discussion Points:

- Linda presented an initial merged dataset from the PMS and channel manager logs, highlighting missing values for children in certain older records.
- Eva identified outliers, such as negative lead_time (likely data entry errors) and bizarre average_daily_room_rate values (e.g., zero or extremely high, beyond the known rate range).
- The group discussed possible approaches to handle ambiguous booking records with unusual deposit_type codes not in the standard set.

Outcomes:

- A structured plan to **impute** or **remove** erroneous data.
- Confirmed the final structure of each variable (e.g., separate fields for arrival_date_week_number and arrival_date_month).
- Documented assumptions for missing or inconsistent records (e.g., children defaulted to zero if not specified).

6.3. Final Data Readiness Meeting

Attendees:

- Alice Kim (Lead Data Scientist)
- Jorge Martinez (Hotel Operations Manager)
- Linda Green (Data Engineer)
- Dr. Miriam Ross (Revenue Management Specialist)

Key Discussion Points:

- Linda shared a cleansed, consolidated dataset now containing consistent values for each variable.
- Dr. Ross verified the ranges for monetary fields like average_daily_room_rate.
- Jorge ensured that the final dataset aligned with the hotel chain’s privacy standards.
- Alice Kim proposed an immediate start on exploratory data analysis, focusing on cancellation correlations with lead_time, deposit_type, and market_segment.

Outcome:

- Official sign-off on the dataset for model development, with explicit acknowledgment of permissible usage and next steps.

7. Data Quality Checks and Validation

Typical **quality checks** include:

1. **Range Verification:** For numerical fields (e.g., lead_time must be between 0 and 730).
2. **Category Matching:** Ensuring categorical fields (e.g., deposit_type) only contain valid labels.
3. **Statistical Outlier Detection:** Identifying abnormal values in average_daily_room_rate (e.g., negative or extremely large daily rates) or unrealistic numbers of children.
4. **Consistency Checks:** For instance, if stays_in_week_nights is 5, but arrival_date_day_of_month suggests a weekend overlap, investigate anomalies.

8. Data Cleaning Methodology

A standardized approach was implemented to produce a reliable dataset:

- **Unit and Format Conversions:** While not typically an issue with these booking variables, ensure consistency in currency or date formats.
- **Missing Data Handling:** Impute children or babies as 0 if not recorded, but label them as “missing” if the context is ambiguous.
- **Outlier Treatment:** Apply percentile-based capping for average_daily_room_rate if extreme rates are deemed unrealistic.
- **Deduplication:** Remove or merge duplicate reservations that might appear across multiple channel sources.

9. Documented Challenges and Resolutions

1. **Inconsistent Field Definitions:** Different hotel branches sometimes labeled the same feature differently (e.g., “baby” vs. “infant”). The team standardized terms under the “babies” variable.
2. **Data Entry Errors:** Negative lead times were identified and corrected after verifying booking logs. Zero rates were also flagged.
3. **Seasonal Overlaps:** Some data had partial overlap in the year boundary, requiring a careful approach to arrival_date_month and arrival_date_week_number alignment.
4. **Mixed Data Sources:** Integrating data from multiple PMS systems introduced duplicates and conflicting records, which were systematically reconciled through robust matching rules.

10. Regulatory and Compliance Considerations

- **Personal Data Protection:** Although these fields do not typically include sensitive PII beyond occupant counts, privacy regulations (like GDPR in Europe) require safeguards on any dataset that can be traced back to individual guests.
- **Data Usage Policy:** Cancellation predictions often feed **revenue management** or marketing strategies. The dataset should be handled in line with the hotel's data retention and usage policies, ensuring no misuse.
- **Anonymization:** Certain variables (e.g., `is_repeated_guest`) could link to loyalty profiles. Proper anonymization or encryption is essential if data is shared externally.

11. Conclusion and Next Steps

This formal report underscores the **critical nature** of each variable, illustrating how they collectively shape a predictive model for **hotel booking cancellations**. By systematically defining the numeric ranges, categorical values, and domain relevance:

- **Data Scientists** gain a well-documented feature set, aiding in model selection and feature engineering.
- **Hotel Management** obtains clarity on which data points are most relevant for revenue strategies and operational adjustments.
- **Stakeholders** at all levels can reference these definitions to maintain consistent data standards and promote reliable analytics.

Next Steps:

1. **Exploratory Data Analysis (EDA):** Examine correlations, distribution plots, and missing-data patterns to refine feature engineering.
2. **Model Development:** Train, validate, and compare machine learning algorithms (e.g., logistic regression, random forest, gradient boosting) to predict `is_canceled`.
3. **Pilot Deployment:** Use real-time data integration to forecast daily or weekly cancellation probabilities, enabling dynamic overbooking thresholds.
4. **Continuous Improvement:** Update the dataset as new variables or external factors (e.g., pandemic-related changes) arise, ensuring the model remains robust and up to date.

12. Appendices

Appendix A: Example Booking Record (Fictitious)

Booking_ID	lead_time	arrival_date_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations
00012345	45	July	2	3	2	1	0	0	0

Appendix B: Common Abbreviations

- **OTA:** Online Travel Agency
- **PMS:** Property Management System
- **TA/TO:** Travel Agent/Tour Operator
- **CRM:** Customer Relationship Management

13. References

1. **Hotel Revenue Management Journals** – Studies linking advanced cancellation predictions to improved overbooking strategies.
2. **Hospitality Technology Forum Proceedings** – Papers discussing integration of PMS and Channel Manager data.
3. **Sample Datasets** – For instance, open-source hotel booking datasets from academic or professional resources (e.g., UCI Machine Learning Repository).

Document Control

- **Version:** 1.0
- **Prepared By:** Alice Kim (Lead Data Scientist)
- **Reviewed By:** Jorge Martinez (Hotel Operations Manager)
- **Approved By:** Dr. Miriam Ross (Revenue Management Specialist)
- **Effective Date:** March 1, 2026