Prediction of hotel booking cancellation

Aya Bekhtiar, Fatma Braham, Apolline Hadjal

December 14, 2025

Github Repository : `https://github.com/FatmaZBh/Hotel-booking-cancellation-prediction.git`

Aim of study

This report revolves around studying hotel booking cancellations. The aim of our work is to attempt answering the following question: *can we predict hotel booking cancellations in advance in order to reduce revenue loss and optimize overbooking and pricing strategies?*

Dataset documentation

The main dataset is sourced from `https://www.sciencedirect.com/science/article/pii/S2352340918315191` and contains two collections of hotel reservation records. The first property (H1) is a resort hotel, while the second (H2) is a city hotel. Although the locations differ, both datasets share an identical structure with 31 features: H1 includes 40,060 entries and H2 contains 79,330. The records cover bookings scheduled between July 1, 2015 and August 31, 2017, including both completed and canceled reservations.
The dataset features describe when and how bookings are made, customer characteristics, and indicators of commitment to support cancellation analysis. They capture booking anticipation and seasonality, customer profile and past behavior, booking channels, financial conditions such as price and deposit, and signals of engagement or instability during the booking process.

Data preprocessing for ML

Handling numerical and categorical variables

For numerical columns, to improve model performance, we created new features or did some changes in the columns. We added a new column `room_type_mismatch` if the assigned room `reserved_room_type` and `assigned_room_type` differ.
Since the `agent` column represents the agent ID and `corporate` tells us if the booking is for business purposes, we replaced them respectively with `is_agent_booking` and `is_corporate`. When the value is not empty, we assign 1; otherwise, we assign 0. For the agent column, 1 means the booking was made through an agency, while 0 indicates a direct booking.
Our dataset contains categorical and numerical columns, so before doing any machine learning algorithm, we need to convert these columns to numerical ones using categorical encodings. For `hotel`, we used label encoding thanks to the built-in `LabelEncoder` from `scikit-learn` (convert City Hotel and Resort Hotel). For `arrival_date_month`, we used manual mapping to convert months written in text to numbers.
For all the other columns that have many multi-unique values, namely `arrival_date_month`, `meal`, `market_segment`, `distribution_channel`, `reserved_room_type`, `deposit_type`, and `customer_type`, we used one-hot encoding to convert them into unique numerical binary features, and it works well for all models (Logistic Regression, KNN, Decision Tree). To avoid multicollinearity, the first category was dropped for each encoded variable.

Feature Scaling

We applied `StandardScaler` to all features (excluding the target variable `is_canceled`). This is crucial since we are going to use distance-based algorithms for classification like Logistic Regression and KNN.

Handling Class Imbalance

We observe the class distribution of `is_canceled`: 72.5% not canceled versus 27.5% canceled. The imbalance is moderate (we consider classes imbalanced when they are below 30%), but it is still significant. Therefore, when splitting the data into training and testing sets, we added the `stratify=y` parameter to maintain class proportions. Also, for Logistic Regression, we used the `class_weight='balanced'` parameter.

Machine Learning

Supervised Learning

Supervised learning was used to predict whether a booking would be canceled. This task directly supports the project's objective of anticipating cancellation risk in order to reduce revenue loss and improve operational planning. Several classification algorithms were tested to balance predictive performance, interpretability, and robustness to class imbalance.

Logistic Regression

Logistic Regression was used as a baseline classification model due to its simplicity and interpretability. It provides a clear understanding of how booking characteristics influence cancellation probability, which is valuable for managerial interpretation. The model was particularly useful for identifying high-risk bookings early, thanks to its strong recall performance.

K-Nearest Neighbors

K-Nearest Neighbors classifies bookings based on similarity with past observations. This approach was used to capture local behavioral patterns, such as bookings with similar lead times, prices, and customer profiles. Although intuitive, its performance highlights the limits of purely distance-based methods for complex cancellation behavior.

Decision Tree

Decision Trees were applied to model non-linear relationships between booking features and cancellation outcomes. Their rule-based structure makes them easy to interpret and suitable for understanding conditional cancellation patterns, such as the combined effect of lead time and deposit type.

Random Forest

Random Forest extends Decision Trees by aggregating multiple trees, improving generalization and reducing overfitting. This model was well suited for the project as it captures diverse cancellation patterns across customer types and booking conditions, resulting in strong predictive performance.

Gradient Boosting

Gradient Boosting was used to model complex interactions between booking features through an ensemble of sequential learners. Its strong performance indicates that cancellation behavior is driven by multiple weak signals rather than a single dominant factor, making it the most effective model for this prediction task.

Unsupervised Learning

In addition to prediction, unsupervised learning was applied to identify distinct customer segments based on booking behavior. This approach supports the business objective of tailoring cancellation prevention strategies to different customer profiles.

K-Means Clustering

K-Means clustering was used to segment bookings into groups with similar characteristics, such as lead time, price level, and repeat guest behavior. These segments revealed meaningful differences in cancellation rates, enabling targeted managerial actions rather than uniform policies.

Principal Component Analysis

Principal Component Analysis was applied to reduce dimensionality and visualize customer segments. While not used for prediction, PCA helped assess the structure of the data and supported the interpretation of clustering results by highlighting underlying behavioral patterns.

Model Evaluation

Metrics and Model Comparaison

We did an evaluation function called `evaluate_model` that we call for each supervised model we used. The function computes 5 metrics :

- **Accuracy**, which measures the overall proportion of correct predictions. However, in our case, this metric can be misleading due to class imbalance.
- **Precision** measures, among all bookings predicted to be cancelled, the proportion that were actually cancelled. A high precision indicates fewer false alarms
- **Recall** measures the proportion of actual cancellations that were correctly identified by the model.A High recall means catch more cancellations
- **F1-score** balances both precision and recall
- **ROC-AUC** measures the model's ability to distinguish between classes across all decision thresholds. A value of 0.5 corresponds to random guessing, while a value of 1.0 indicates perfect classification.

The function outputs a detailed analysis of the confusion matrix,where we look specifically at False positives and False Negatives.
In addition, after we plotted ROC curves for all five models to visually compare their performance across different decision thresholds.

Results

Predictive Modeling Performance

Our analysis evaluated five machine learning models, with Gradient Boosting emerging as optimal (80.9% accuracy, 0.846 ROC-AUC). While Logistic Regression showed highest recall (78.3%), its low precision (47.6%) would generate excessive false positives. Gradient Boosting achieved balanced performance (73.3% precision, 48.0% recall, F1: 0.580), making it most practical for deployment.
The model correctly identifies 2,308 cancellations while producing only 840 false alarms—an acceptable trade-off enabling targeted interventions without overwhelming operations.

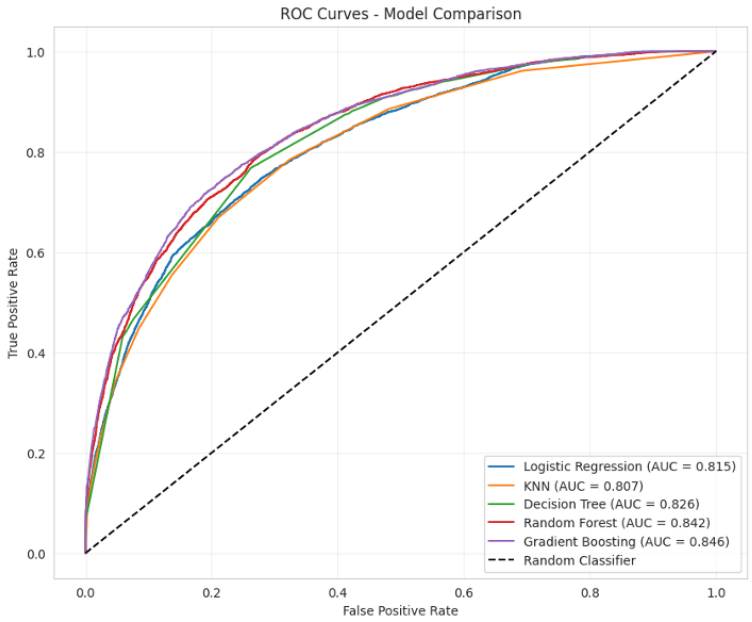

Figure 1: ROC Curves for All Models

Key Cancellation Risk Factors

Lead time emerged as the strongest predictor: bookings 181+ days in advance show dramatically higher cancellation rates versus those within 30 days.

Customer type segmentation revealed Transient customers (71,986 bookings) exhibit 30.1% cancellation rate—highest among all types—while Group bookings show 9.9% cancellation rate. Most critically, non-refundable deposits correlated with 94.7% cancellation rate versus 26.7% for no-deposit bookings—suggesting current policies create perverse sunk-cost incentives demanding immediate revision.

Business Insights

Strategic Recommendations

We recommend five core strategies:

1. ML-Powered Risk Scoring: Integrate Gradient Boosting to assign real-time cancellation probabilities. High-risk bookings (¿60%) receive immediate engagement; medium-risk (30-60%) receive standard monitoring; low-risk (¡30%) receive minimal intervention.

2. Differentiated Overbooking: Apply risk-adjusted rates: 15-20% for high-risk segments (long lead times, transient customers), 10-12% for medium-risk, 5% for low-risk (groups, short lead times). This can reduce revenue loss by 30-40%.

3. Restructure Deposit Policies: Eliminate non-refundable options creating sunk-cost psychology. Implement moderate refundable deposits (10-15%) for long lead-time bookings; maintain flexible policies for short lead-time bookings exhibiting lower risk.

4. Proactive Engagement: High-risk bookings receive automated reminders at 60, 30, and 14 days with flexible modification options. Medium-risk bookings receive reminders at 30 and 7 days with upselling opportunities.

5. Dynamic Pricing: Incorporate cancellation risk into revenue management, adjusting rates and terms based on risk profiles while applying premium pricing for low-risk segments.

Expected Impact

With baseline 27.5% cancellation rate (24,025 cancellations), interventions reducing cancellations by 25-35% translate to 6,000-8,400 additional confirmed bookings annually. At average ADR of

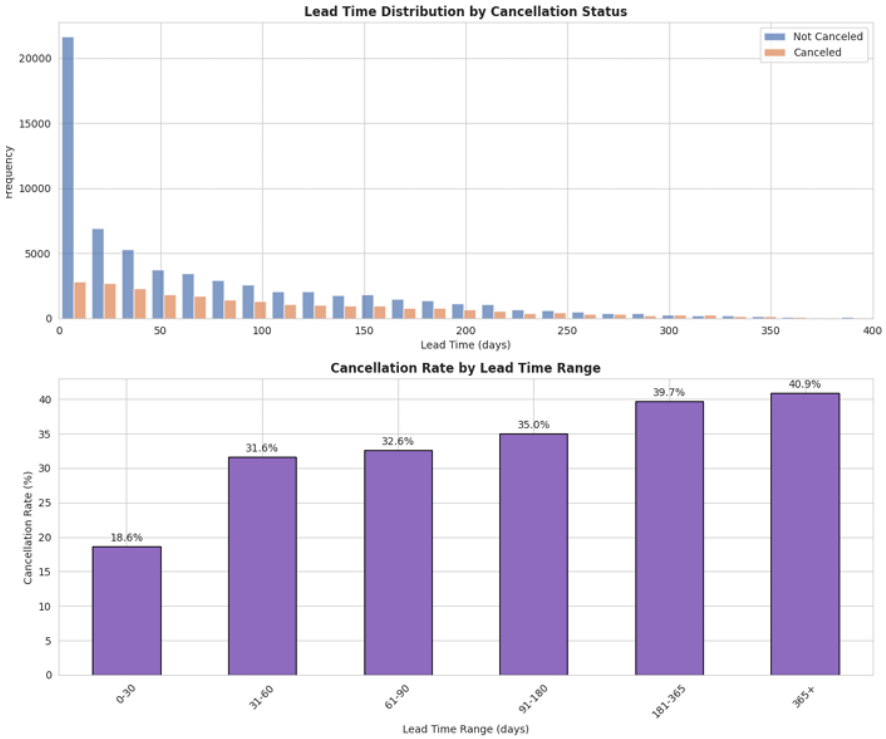

Figure 2: Lead Time Distribution and Cancellation Rate

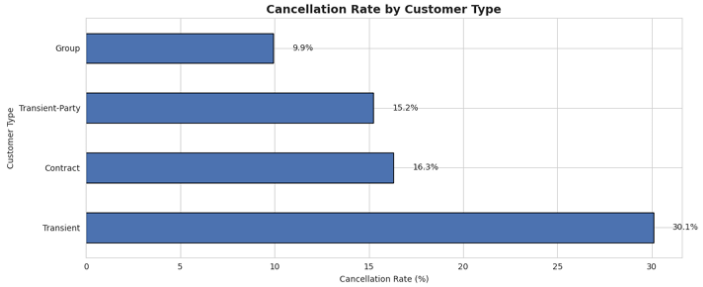

Figure 3: Cancellation Rate by Customer Type

$102, this represents **potential revenue recovery of $612,000-$856,000 per year**. Implementation follows three phases: Phase 1 (0-3 months) focuses on deposit policy revision and risk-based reminders; Phase 2 (3-6 months) involves full ML deployment with dashboards; Phase 3 (6-12 months) integrates dynamic pricing and quarterly retraining. Success metrics should track cancellation rate reduction from 27.5% to 20-22%, improved occupancy, and increased revenue per available room.

Conclusion

This study demonstrates hotel booking cancellations can be predicted with 80.9% accuracy using Gradient Boosting. Three critical insights emerged: lead time is the strongest predictor, customer type significantly influences stability, and non-refundable deposits paradoxically drive cancellations.

Implementing ML-powered risk scoring, restructured deposit policies, and differentiated overbooking can potentially recover $612,000-$856,000 annually—a 25-35% reduction in cancellation losses. The framework is scalable and adaptable, transforming cancellation management from reactive cost center into strategic revenue optimization opportunity.