

Executive Summary

ABC Hotels has been losing a substantial portion of its projected revenue due to high booking cancellation rates. Nearly one-third of all reservations end up being canceled, which complicates occupancy forecasting and inventory management. The goal of this analysis was to build a predictive model that estimates the likelihood of a booking being canceled so that hotel management can take proactive action. Using a dataset of 36,238 historical reservations, two dense neural network models were developed and compared. The best-performing model achieved an area under the ROC curve (AUC) of 0.921 on the validation set, demonstrating excellent discriminatory power.

Key insights emerged from the analysis. Bookings made far in advance had a dramatically higher probability of cancellation, suggesting uncertainty in long-term travel planning. Guests who made special requests or were repeat customers were far less likely to cancel, indicating that engagement and loyalty predict reliability. The final model not only delivers strong predictive performance but also produces actionable probabilities that can be integrated into the booking system to identify at-risk reservations in real time. By leveraging this model, ABC Hotels can reduce last-minute cancellations, optimize overbooking strategies, and stabilize revenue forecasting.

Approach and Data

The analytic goal was to convert the hotel's historical booking data into a tool that could forecast the probability of cancellation at the moment a reservation is made. The project began with the Analytic Plan, where several candidate modeling approaches were considered. After reviewing the preliminary results, the final phase focused on deep learning models due to their ability to capture nonlinear interactions between variables like pricing, lead time, and customer history. The dataset consisted of 36,238 records containing details about arrival dates, room types, prices, special requests, and previous customer behavior. The target variable, `booking_status`, indicated whether a booking was canceled. Data preparation involved removing the non-informative `Booking_ID` column, converting the target variable to binary format, and parsing arrival dates into separate month and season features. Categorical variables such as `type_of_meal_plan`, `room_type_reserved`, and `market_segment_type` were one-hot encoded to make them suitable for machine learning. Missing numeric values were replaced with the median, and categorical ones with the most frequent label. Numerical variables were standardized to improve convergence during neural network training.

The dataset was stratified into 80% training and 20% validation subsets to ensure that the class imbalance (32.8% canceled vs 67.2% not canceled) was preserved. This setup allowed consistent evaluation of model generalization while avoiding information leakage. Exploratory analysis from earlier stages revealed strong trends that informed feature selection: longer lead times, higher prices, and fewer special requests were associated with a greater likelihood of cancellation. These insights provided the foundation for model development.

Detailed Findings and Model Evaluation

Two feedforward neural network models were designed to test different levels of complexity. The first, referred to as Model A, contained two hidden layers with 64 and 32 neurons respectively, using ReLU activation and Adam optimization. The second, Model B, was a smaller architecture with 32 and 16 neurons and slightly higher regularization to test whether reduced complexity could yield faster, more generalizable performance. Both models were trained with early stopping and adaptive learning rates to prevent overfitting.

Model A achieved a training AUC of 0.933 and a validation AUC of 0.921, while Model B achieved 0.917 and 0.906 respectively. The modest difference between training and validation AUC for Model A indicated excellent generalization. Model A also produced a validation accuracy of 86%, precision of 0.84, recall of 0.72, and F1-score of 0.77. These metrics mean that when the model predicts a booking will be canceled, it is correct roughly 84% of the time, and it captures about 72% of actual cancellations. In contrast, Model B was slightly less precise and missed more cancellations.

ROC curve analysis confirmed that Model A's performance remained stable across probability thresholds, maintaining a strong balance between sensitivity and specificity. The learning curve suggested that both training and validation AUCs plateaued together, showing the model had learned as much as possible from the available data without overfitting. Calibration analysis demonstrated that the model's predicted probabilities aligned closely with observed cancellation rates. For example, bookings assigned a 70% cancellation probability actually canceled about 70% of the time. This reliability is essential for operational use, where decision thresholds can be tuned to the organization's tolerance for risk.

Feature importance analysis, using permutation-based techniques, highlighted several critical predictors. Lead time was by far the most influential factor: the longer the time between booking and arrival, the more likely a cancellation. Average room price and the number of special requests also played major roles; customers who paid higher prices but made few requests were the most likely to cancel. Repeat guests and those with prior non-canceled bookings were the most dependable customers. These findings reinforce the intuitive link between engagement, commitment, and follow-through.

Recommendations and Business Application

Based on performance and interpretability, Model A was selected as the final model. Its predictive probabilities can be directly integrated into the hotel's booking management system. Each new reservation can be assigned a real-time cancellation probability, allowing staff to identify high-risk bookings immediately. Bookings exceeding a 0.75 probability threshold could trigger a set of proactive measures such as personalized reminder emails, flexible rebooking offers, or partial prepayment options to reduce the likelihood of last-minute cancellations. From a revenue management perspective, the model's probabilities also support strategic overbooking decisions. By estimating the expected number of cancellations for upcoming dates,

managers can confidently sell slightly more rooms without risking customer dissatisfaction from overselling. Over time, the model can be retrained quarterly as new booking data accumulates, ensuring it adapts to changing traveler behaviors and economic conditions.

Operationally, the model's outputs should be incorporated into a simple dashboard that classifies bookings as low-, medium-, or high-risk. Customer service teams can prioritize outreach accordingly, improving guest relationships while simultaneously reducing revenue volatility.

Conclusion

This project demonstrates how machine learning can transform historical data into a practical decision-making tool. The final dense neural network model provides ABC Hotels with a reliable, data-driven method for anticipating booking cancellations, offering both predictive accuracy and actionable insights. Beyond its technical performance, the analysis revealed human patterns in guest behavior like commitment, engagement, and loyalty that have clear strategic implications. By implementing this model and continuing to refine it with new data, ABC Hotels can move from reactive crisis management to proactive forecasting, strengthening both financial outcomes and customer trust.