Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Hotel Booking Cancellations with Explainable Machine Learning: Implications for Revenue Management

NUDRAT NUDRAT

nfariha@umass.edu

Unversity Of Massachusetts https://orcid.org/0009-0004-4825-1500

Kazi Moshiur Rahaman

240378802@aston.ac.uk

Aston University https://orcid.org/0009-0002-5468-1297

Research Article

Keywords: hotel booking cancellations, revenue management, XGBoost, SHAP, explainable AI, overbooking

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11585608/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Booking cancellations are a routine and costly feature of hotel demand: 37.0 percent of the bookings in the dataset used here were cancelled before arrival. Hotels respond with overbooking and with deposit and cancellation policies, but both responses are usually set in aggregate, while the decisions a revenue manager faces concern individual bookings on specific arrival dates. This study uses the Hotel Booking Demand dataset of Antonio, de Almeida and Nunes (2019c), covering 119,390 bookings at a Portuguese resort hotel and a city hotel, to predict individual booking cancellations under a protocol designed for decisions rather than for headline accuracy. Post-outcome variables are excluded, models are evaluated out of time, and a recording audit tests three fields whose values accumulate over a booking's life. A logistic regression baseline is compared with random forest and XGBoost on discrimination, calibration, and classification at cost-based thresholds. The best-performing model is explained with SHAP values at global, local and segment level, separating drivers a manager can act on, such as deposit type, distribution channel terms and offers targeted by lead time, from drivers that can only be priced for, such as a guest's own cancellation history. The predicted probabilities are then translated into expected cancellations and net demand by arrival date, an uncertainty range around expected cancellations, and a threshold-based rule for retention intervention, with the costs of false positives and false negatives stated explicitly. The study's contribution is a transparent benchmark, a decision-oriented explanation, and a quantified translation on a widely used public dataset; it proposes no new algorithm. Out of time, XGBoost performed best at both properties. On the twenty predictors as recorded, ROC-AUC was 0.840 at the city hotel and 0.849 at the resort hotel (logistic regression: 0.817 and 0.833); in an audited specification that excludes three fields whose recording is entangled with the outcome, it was 0.794 and 0.784 (logistic regression: 0.778 and 0.761). Deposit type was the strongest recorded driver at the city property, subject to a data caveat set out in Section 4.3.

1. Introduction

Every cancelled booking leaves a hotel with the same problem. The room was held for a guest who will not arrive, a room night cannot be stored, and if the room is not resold, the revenue is lost permanently. The scale of the problem is not small. Cancellation rates of up to about 20 percent of bookings are reported generally, and up to 60 percent for airport and roadside hotels (Romero Morales & Wang, 2010). In the dataset used in this study, 44,224 of 119,390 bookings, 37.0 percent, were cancelled before arrival (Antonio, de Almeida & Nunes, 2019c). Cancellations are therefore not occasional accidents in this industry. They are a routine feature of demand, and revenue management has to be organised around them.

The two standard responses pull in opposite directions. Overbooking accepts more reservations than the hotel has rooms, on the expectation that some guests will not arrive. It protects occupancy, but when fewer guests cancel than expected, the hotel denies service to guests holding valid reservations, with the damage to satisfaction and future demand that denied service causes (Noone & Lee, 2011). Deposits and cancellation policies protect the individual booking instead, but restrictive terms are judged by guests on fairness and can reduce bookings and patronage (Smith et al., 2015), and the element of a policy that changes guest behaviour is its deadline rather than the size of its fee (Chen, Schwartz & Vargas, 2011). Both responses are usually set in aggregate: an overbooking level for a date, a policy for a segment. The decision a revenue manager actually faces each day is narrower. It concerns the specific bookings on the book for a

specific arrival date: which of them are likely to disappear, and what should be done about those bookings while there is still time to act.

Booking-level prediction is the natural answer to that narrower question, and machine learning models can find patterns in reservation data that aggregate forecasts and simple rules miss. Yet prediction by itself has not settled the manager's problem, for three reasons that the review in Section 2 documents. First, the reported performance of cancellation models is hard to trust. On materially the same data, reported accuracy ranges from about 70 percent to above 95 percent, and the differences owe less to the algorithms than to design choices: random rather than time-based test splits, resampling applied before splitting, and the inclusion of variables that are only known after the outcome, such as the reservation status fields in the public data. Under a stricter protocol that splits by time and removes leakage-prone variables, conventional models score AUC values around 0.78 (Yang & Miao, 2024). A manager cannot set protection levels from numbers whose size depends on how the test was designed. Second, where models are explained, the explanation usually stops at a global ranking of important variables. A ranking does not say which of the important drivers the hotel can act on and which it can only price for, and it does not convert a booking's risk into the quantities revenue management works with: expected cancellations on a date, net demand, and a defensible overbooking level. Third, a score that a manager cannot interrogate is unlikely to be used at all. One visible mistake is often enough for users to turn away from an algorithm for good, even one they have watched perform better than a person (Dietvorst, Simmons & Massey, 2015), and people who forecast for a living are among the least likely to defer to algorithmic advice (Logg, Minson & Moore, 2019). Revenue managers are forecasting experts, judging bookings daily against their own experience.

This study addresses the three problems together, using the public Hotel Booking Demand dataset (Antonio, de Almeida & Nunes, 2019c): 119,390 bookings from a resort hotel and a city hotel in Portugal, for arrivals between July 2015 and August 2017. The design has three parts. First, we benchmark an interpretable logistic regression baseline against random forest and XGBoost under a leakage-audited, out-of-time evaluation, judging the models on calibration and on classification at cost-based thresholds, not on accuracy alone. Second, we explain the best-performing model with SHAP values at global, local and segment level, separating the drivers a revenue manager can act on, such as deposit type, distribution channel terms and offers targeted by lead time, from those the manager can only price for, such as the guest's own cancellation history and the booking's stay pattern. Third, we translate the predicted probabilities into decisions: expected cancellations and net demand by arrival date, an uncertainty range around expected cancellations, and a threshold-based rule for retention intervention whose false-positive and false-negative costs are made explicit. Two limits on the claim should be stated at the outset. This study proposes no new algorithm, and it claims no performance record. Its contribution is discipline in evaluation, explanation organised by decision, and quantified translation.

The study is organised around three research questions. RQ1: Under an out-of-time, leakage-audited evaluation, how do random forest and XGBoost compare with a logistic regression baseline in discrimination and calibration when predicting individual booking cancellations? RQ2: Which booking characteristics drive cancellation risk at the global, local and segment level, and which of these drivers can a revenue manager act on directly, as distinct from those that can only be priced for? RQ3: How can predicted cancellation probabilities be translated into expected cancellations and net demand by arrival date, an uncertainty range

around expected cancellations, and a retention rule, once the costs of false positives and false negatives are stated explicitly?

The contributions follow the same order. First, the study provides a transparent and reproducible benchmark on a widely used public dataset, with an evaluation protocol strict enough for the comparison between a statistical baseline and two standard machine learning models to be believed. Second, it provides an explanation of cancellation risk that is organised by the decisions it is meant to support, distinguishing actionable from price-for-only drivers within segments rather than only in aggregate. Third, it provides a worked translation from predicted probabilities to the quantities a revenue manager consumes, with the cost trade-offs stated rather than assumed, so that a hotel facing different prices for empty rooms and denied service can set its own operating point.

The paper proceeds as follows. Section 2 reviews the literature on cancellations and revenue management, on machine learning prediction, and on explainable AI and managerial trust. Section 3 describes the data, the leakage controls, the models, and the evaluation and translation procedures. Section 4 reports the results. Section 5 discusses the findings, Section 6 states the implications for revenue managers, Section 7 sets out the limitations and directions for future research, and Section 8 concludes.

2. Literature Review

2.1 Booking cancellations and revenue management

Cancellations are not a side issue in hotel revenue management; they are one of its central uncertainties. A room night cannot be stored, so a booking that cancels late leaves capacity that may never be resold, while a hotel that expects cancellations and sells beyond capacity risks turning guests away. The standard reference on revenue management treats cancellations and no-shows as a core problem of the field, and notes that a cancellable reservation serves the customer twice over: it holds the room, and it leaves the customer free to abandon it (Talluri & van Ryzin, 2004). The losses are large. Studies report cancellations of up to about 20 percent of bookings in general, and up to 60 percent for airport and roadside hotels (Romero Morales & Wang, 2010).

The classical response to this uncertainty is overbooking. In an early hotel-specific model, Liberman and Yechiali (1978) allowed customers to cancel at any time without penalty and showed that the profit-maximising policy has a three-region structure, with upper and lower bounds on the inventory the hotel should hold. Overbooking, however, is not free. When a hotel denies service to a guest with a valid reservation, satisfaction falls and complaining rises; cash compensation repairs satisfaction better than vouchers, but even generous compensation does not restore the guest's intention to return (Noone & Lee, 2011). Protecting against cancellations therefore has a cost on both sides: too little protection leaves rooms empty, and too much protection creates denied service and damages future demand.

A second response is to forecast. Traditional hotel forecasting works at an aggregate level: it predicts how many rooms will be occupied on a future date, not what will happen to a specific booking. In a comparison on Choice Hotels and Marriott data, Weatherford and Kimes (2003) found that exponential smoothing, pickup

and moving-average methods were the most robust of the methods tested. These methods remain the background against which booking-level prediction must justify itself, because they are simple, stable and already embedded in practice.

A third response is policy: deposits, deadlines and cancellation fees. The evidence here is more subtle than hotels often assume. In a week-long booking experiment with deal-seeking customers, Chen, Schwartz and Vargas (2011) found that the cancellation deadline shaped whether customers booked, booked and kept searching, or only searched, while the size of the cancellation fee did not significantly change behaviour. Chen and Xie (2013) showed that cancellation policies in the U.S. hotel industry are differentiated by hotel type, so there is no single policy that fits all properties. Smith, Parsa, Bujisic and van der Rest (2015) add a fairness warning: restrictive, non-refundable policies can reduce bookings and revenue, because guests judge such policies on fairness and adjust their patronage. Riasi, Schwartz and Chen (2018) argue that research on cancellation practices needs proposition-based theory rather than purely descriptive work, because the same policy can have opposite effects in different settings.

Finally, a behavioural literature studies who cancels, rather than how to respond. Using 233,000 bookings from nine Finnish hotels, Falk and Vieru (2018) estimated probit models separately by channel. The overall cancellation rate was 8 percent, but it was 17 percent for online bookings, 12 percent for offline bookings and 4 percent for travel-agency bookings; two characteristics stood out above the rest (how far in advance the booking was made, and the guest's country of residence), and both mattered most in the online channel. Romero Morales and Wang (2010), reviewing passenger-name-record data mining for cancellation and no-show forecasting, add a point that matters for everything that follows: what predicts a cancellation early in the booking window is not the same as what predicts it near arrival. Cancellation is therefore not one static problem. The risk attached to a booking changes as the arrival date approaches, and the information that predicts it changes as well.

This literature leaves one level of analysis underdeveloped. Overbooking models, aggregate forecasts and policy studies all work with rates: the share of bookings that cancel in a period or a segment. The decision a revenue manager faces each day, however, is about individual bookings: which reservations on the book for a given arrival date are likely to disappear, and what should be done about those reservations now. That is the level at which prediction enters, in Section 2.2.

2.2 Machine learning for cancellation prediction and revenue management

Machine learning research on hotel demand has moved in the same direction as the industry: from aggregate series toward the individual booking. Song and Li (2008), reviewing 121 tourism demand studies, organise the field into time-series, econometric and artificial-intelligence models. Within hotels, Pereira (2016) connects the traditional revenue management forecasting toolkit with newer statistical methods, and Pereira and Cerqueira (2022) show that machine learning regression models outperform seasonal-naive and exponential-smoothing benchmarks for daily hotel demand. Viverit, Heo, Pereira and Tiana (2023) take a design that managers can follow: they cluster historical booking curves and then apply the additive pickup model within each cluster, beating traditional approaches at all forecasting horizons while keeping a method

revenue managers already use. The general pattern is that machine learning improves accuracy. The question of what a manager should do differently because of a prediction is usually left unanswered.

Cancellation prediction on the dataset used in this study has a dense history, and honesty about that history is necessary. The dataset's creators built the first models themselves. Antonio, de Almeida and Nunes (2017b) estimated a machine learning classification model on the two hotels' property management data, the extraction later published as the public dataset, and showed that the cancellation of an individual booking could be classified with useful accuracy. In related work on four Algarve resort hotels over 2013–2015, the same authors reported test accuracies between roughly 0.90 and 0.99 across hotels (Antonio, de Almeida & Nunes, 2017a). They then deployed an automated machine learning decision support system in the two hotels themselves; in live operation it reached an AUC of 0.88, with precision above 82 percent and accuracy above 84 percent, and the authors report that acting on the bookings it flagged as high-risk brought cancellations down by 37 percentage points (Antonio, de Almeida & Nunes, 2019a). A companion study combined eight hotels' data with weather, holidays, events, reputation and price data and used interpretable machine learning to expose the drivers of cancellation, so that policies and overbooking tactics could be adjusted (Antonio, de Almeida & Nunes, 2019b). Booking-level cancellation prediction on this data, including a deployed system and an interpretable-model study in a leading journal, therefore already exists. This study does not claim otherwise.

Later work on the same file has pushed in four directions. First, class imbalance: Adil, Ansari, Alahmadi, Wu and Chakrabortty (2021) addressed the uneven class distribution by combining SMOTE-ENN resampling with five classifiers - logistic regression, decision tree, AdaBoost, gradient boosting and random forest - and reported the highest accuracies for random forest with SMOTE-ENN, above 95 percent in both hotels. Second, model architecture: Herrera, Arroyo, Jiménez and Herrero (2024) compared neural networks, decision trees, random forest, AdaBoost and XGBoost on a Portuguese resort and city hotel, and Yang and Miao (2024) proposed a tree-based neural network that reached AUC values of 0.855 (resort) and 0.837 (city). Yang and Miao's study is the most important methodological comparator for this paper, because they split the data by time, the first 75 percent of bookings by date for training and the last 25 percent for testing, and they removed variables that leak the outcome, including country, a field that in these extracts was only settled when the guest checked in. Under that stricter protocol, standard gradient boosting, XGBoost and random forest models on the city hotel produced AUC values between 0.776 and 0.784. The contrast between figures of this size and the much higher numbers reported under more permissive designs is the central measurement problem of this literature, and Section 3 returns to it. Third, interpretability inside the prediction model: Chen, Ngai, Ku, Xu, Gou and Zhang (2023) integrated machine learning with a probability model built on interpretable feature interaction, and Liu, De Bock and Zhang (2025) trained a heterogeneous stacking ensemble toward profitability, with global post-hoc interpretation, reporting greater profit than other profit-oriented models. Fourth, timing: Kundu, Roy, Shukla and Mateen (2025) modelled the city hotel in two stages, whether a booking cancels and when, using partial dependence plots to interpret the timing stage, with average daily rate, arrival month, arrival day and lead time as significant predictors; Duong and Phan (2026) combined TabNet classification (AUC 0.862 for the resort hotel and 0.846 for the city hotel) with DeepHit survival analysis (concordance index 0.812), and presented the output as cancellation risk at 7-, 14- and 30-day horizons, intended to guide price changes, retention offers and adjustments to overbooking.

Two findings from the wider literature shape this study's design. Sánchez-Medina and C-Sánchez (2020), using Canary Islands booking data, showed that a small set of 13 variables, all available at booking time, can carry most of the predictive signal; parsimony is a practical virtue when a hotel must collect the inputs for every reservation. Liu, Jiang, Wang, Du, Niu and Zhang (2023) optimised XGBoost for profit rather than accuracy and earned more than accuracy-oriented models, and Ampountolas (2025), on 209,545 reservations from a four-star chain, reported precision of 97.65 percent for XGBoost alongside low recall: a clear demonstration that headline accuracy or precision is the wrong target when the cost of a missed cancellation falls on the hotel. The most recent work also warns that performance is not stable over time. Silvestre, Antonio and Carrasco (2026), on 670,343 bookings from four Portuguese hotels including the pandemic period, found static-model AUC values between 0.70 and 0.93 depending on the hotel, with sliding-window retraining adding up to five percentage points; notably, their models were not deployed, a limitation they state openly. Ampountolas and Legg (2026) similarly move the target from the individual booking to short-term cancellation-rate forecasting for pricing.

Read together, this literature is strong on prediction and weaker on two things: evaluation discipline and decision translation. Reported performance ranges from about 70 percent to above 95 percent accuracy on materially the same data, and the differences come less from the algorithms than from how the data were split, whether resampling was applied before splitting, and whether post-outcome variables were left in the model. And in nearly every study, the managerial use of the prediction is asserted - better pricing, better inventory control - rather than worked out in quantities a manager can apply. Those two weaknesses define the space for this study, together with the explanation problem taken up next.

2.3 Explainable AI and managerial trust

An accurate model that a manager cannot read will not be used. Explainable AI addresses the reading problem. The reference method in this study is SHAP (SHapley Additive exPlanations). Lundberg and Lee (2017) showed that additive feature-attribution methods, which explain a single prediction by assigning each feature a contribution, have a unique solution that satisfies local accuracy, missingness and consistency, and that this solution unifies several earlier explanation methods. SHAP therefore gives this study a principled way to say, for one booking, how much each of its characteristics pushed its cancellation risk up or down. Two cautions follow from the method itself. SHAP explains the model, not the world: an attribution is a statement about how the model uses a variable, not proof that the variable causes cancellation. And whether an output explains anything depends on who is meant to understand it. Barredo Arrieta et al. (2020), in the standard survey of the field, put the audience inside the definition of explainable machine learning: an output counts as an explanation only if its intended reader can make sense of it. A feature ranking produced for a data scientist is not automatically an explanation a revenue manager can act on.

Applied work on cancellation has begun to use these tools, but thinly. Hermawan, Amalia, Rafif, Azzahra and Ragasa (2025) applied SHAP to an XGBoost model fitted on a reduced subset of this same dataset (83,222 bookings and 10 predictors), finding deposit type, required car parking spaces, previous cancellations and online travel agency segment as the leading predictors; we note this prior use openly, because it means SHAP on this dataset is not new. Silvestre, Antonio and Carrasco (2026) used SHAP differently and more deeply: not as a one-off ranking, but to track how predictor importance rotates as conditions change. Lead

time stayed dominant, while previous cancellations fell from fourth to thirteenth in importance in one city hotel during the pandemic, and they remark on how rarely SHAP has been applied to hotel cancellation problems at all. Where hospitality research uses explainability more broadly, it is mostly to interpret guest-generated text and sentiment, not revenue decisions. In short, explanation in this literature stops at global feature rankings, followed by general implications. The rankings tell a manager which variables matter on average. They do not tell the manager what to do about the arrivals of a specific date, which of the important variables the hotel can actually change, or what an intervention is worth when it is applied to a booking that would not have cancelled anyway.

Whether managers will use such a system at all is a behavioural question, and the evidence cuts both ways. Dietvorst, Simmons and Massey (2015) showed that people who saw an algorithm err avoided it afterwards and chose an inferior human forecaster instead, even when they had seen the algorithm outperform the human, because confidence in an algorithm falls faster after a mistake than confidence in a person. Dietvorst, Simmons and Massey (2018) found that the remedy is partly in the design: giving users the freedom to adjust an algorithm's forecasts, even within narrow limits, made them readier to rely on it, happier with it, and readier to forgive the mistakes it still made. A cancellation tool whose threshold a manager can set, and whose flags a manager can override, is therefore more likely to survive contact with practice than a fixed black box. The counterweight is Logg, Minson and Moore (2019): in their experiments people usually took algorithmic advice over a person's, but the preference shrank in two situations: when the competing advice was the person's own judgement, and when the person forecast as part of their job. Revenue managers are forecasting experts who judge bookings daily. Neither distrust nor acceptance can be assumed for them; it has to be designed for.

The gap this study addresses follows from Sections 2.1 to 2.3 together. That a single booking's cancellation can be predicted from this dataset is no longer in question: the dataset's creators showed it (Antonio, de Almeida & Nunes, 2019a), and later work has carried the prediction toward profit (Liu, De Bock & Zhang, 2025), toward the timing of cancellation (Kundu et al., 2025; Duong & Phan, 2026), and toward keeping the model valid as conditions drift (Silvestre, Antonio & Carrasco, 2026). What is still missing is a benchmark on this dataset that a reader can fully check and a manager can fully use, in three respects. First, no study has put a simple logistic regression baseline and the two standard tree ensembles through the same strict test on this file: post-outcome variables removed, three fields whose recording is entangled with the outcome audited separately, models judged on later bookings than the ones they were trained on, and judgement passed on calibration and discrimination together rather than on accuracy by itself; second, no explanation of a model on this data has been organised the way a revenue manager's choices are organised, with SHAP reported for the whole book, for single bookings and for segments, and a clear line drawn between what the hotel can still change about a booking (its deposit terms, its channel conditions, an offer made while there is still lead time) and what it can only charge for, such as the guest's record of past cancellations and the shape of the stay itself; and third, predicted probabilities have not been carried through into the numbers a revenue manager works with each day: how many cancellations to expect on a given arrival date, what demand remains after them, which bookings justify a retention intervention once the cost of intervening unnecessarily is set against the cost of missing a cancellation that then happens. The gap, in short, is methodological and

practical rather than algorithmic, and it is that gap this study attempts to fill. It offers no new algorithm, and no claim to the best score on this data.

3. Data and Methodology

3.1 Data

We use the Hotel Booking Demand dataset published by Antonio, de Almeida and Nunes (2019c) in Data in Brief. The data were extracted from the property management systems of two Portuguese hotels, one resort hotel and one city hotel, and contain 119,390 bookings in total: 40,060 at the resort hotel and 79,330 at the city hotel. Each observation is a single booking with an arrival date between 1 July 2015 and 31 August 2017. Its authors built it for cancellation prediction, and they offer it as a common test bed: new models can be run on the same file and set against the published results. That purpose suits this study, which is a benchmark and translation exercise rather than a search for a new algorithm.

Each booking carries the information a hotel records around a reservation: lead time; arrival date and length of stay, split into weekend and week nights; party composition (adults, children, babies); meal plan; country of origin; market segment and distribution channel; deposit type; whether the guest is a repeat guest, and the guest's record of previous cancellations and previous completed bookings; reserved room type; booking changes; days in the waiting list; customer type; average daily rate (ADR); required car parking spaces; and the number of special requests. The outcome variable is IsCanceled, which equals 1 if the booking was cancelled before arrival and 0 otherwise. In total, 44,224 of the 119,390 bookings (37.0 percent) were cancelled. The rate is higher at the city hotel (about 42 percent) than at the resort hotel (about 28 percent), a difference that runs through the analysis: models are estimated separately for each hotel and reported side by side, so that a result driven by one property is not presented as a result for both. One definition needs to be explicit. Some studies in this literature treat no-shows as cancellations. In this dataset, IsCanceled follows the property management system's cancellation flag, and a no-show is not coded as a cancellation in that variable; we keep the dataset's own definition and model cancellations only.

3.2 Data preparation and leakage control

We restrict the predictors to information that is known at the time of booking, or that accumulates before arrival in the normal course of managing the reservation, because the model is meant to be used in advance, while the hotel can still act. Three accumulating fields are audited separately in Section 3.3. Four fields are excluded before any modelling, because they are recorded only during or after the stay, or because they reveal the outcome. ReservationStatus and ReservationStatusDate are post-outcome fields: they state whether the guest checked out, cancelled or did not show, and when that status was set, so including them would hand the answer to the model. AssignedRoomType is excluded because the room actually assigned is decided by the hotel at check-in, not chosen at booking; the guest's own choice, ReservedRoomType, is retained. Following Yang and Miao (2024), we also exclude Country from the main specification: in these extracts a booking's country entry was not final until check-in, so it cannot be treated as information known when the booking was made. We report a sensitivity check with country included, so that the size of its effect is visible rather than assumed. Booking changes are retained, because a change is an event in the life of a

live reservation that a manager can observe before arrival; we note, however, that changes recorded very close to arrival carry information a manager may not yet have acted on, and we treat this variable's importance with corresponding care at the explanation stage.

Missing values are handled within the training data only. Where a field is missing for a small share of bookings, we impute the median for numeric fields and the mode for categorical fields, with imputation values computed on the training set and applied unchanged to the test set; fields that are structurally absent for most bookings, such as company and agent identifiers, are not used as predictors. Categorical variables are one-hot encoded. No resampling is applied to the test set under any circumstances, and the class imbalance, 37 percent cancellations overall, but about 42 percent at the city hotel and 28 percent at the resort hotel, is addressed through threshold choice and cost weighting (Section 3.4) rather than by synthesising bookings, so that the evaluated data remain real reservations with their natural prevalence.

The evaluation design is out of time. Bookings are ordered by arrival date, and the earliest 75 percent form the training set and the most recent 25 percent the test set, separately within each hotel, following the protocol of Yang and Miao (2024). This is stricter than the random splits used in much of the earlier work on this dataset, and it is the honest design for the question a hotel actually faces: a model trained on the past is used on bookings that arrive later. All preprocessing, tuning and threshold selection are performed inside the training period only. Performance under this design is expected to be lower than under random splits (Section 2.2); we regard that as the cost of an honest benchmark, not as a defect to be engineered away.

3.3 Models

We estimate three classifiers for each hotel. The baseline is logistic regression. It is the standard statistical model for a binary outcome, its coefficients can be read directly as changes in log-odds, and it is the model a revenue manager is most likely to have seen; any machine learning model in this study must justify itself against it, not against chance. The two machine learning models are random forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016). Both are tree ensembles that capture non-linear effects and interactions, for example between lead time and market segment, without the analyst specifying them in advance, and both have been among the strongest conventional performers in the earlier work on this dataset (Section 2.2). We deliberately do not add a new architecture: the study's claim is about evaluation, explanation and translation, and adding a novel algorithm would blur it. The two ensembles use fixed, conventional hyperparameter settings, reported with the implementation details below, rather than tuned ones, so that no part of the evaluation is spent on model selection; the logistic regression baseline is likewise estimated without tuning. All three models output a predicted probability of cancellation for each booking, because the decision analysis in Section 3.6 works with probabilities, not with hard class labels.

Implementation was deliberately plain. Logistic regression used class weights inversely proportional to class frequency. The random forest grew 400 trees. XGBoost used 500 rounds with maximum depth 6, a learning rate of 0.05, and row and column subsampling of 0.9. All random seeds were fixed at 42, and the full pipeline was re-run end to end to confirm that the reported figures reproduce exactly. The 20 predictors are: lead time; weekend and week nights stayed; adults, children and babies; meal; market segment; distribution channel; the repeated-guest indicator; previous cancellations; previous bookings not cancelled; reserved room type;

booking changes; deposit type; days in the waiting list; customer type; average daily rate; required car parking spaces; and total of special requests.

A recording audit qualifies the predictor set. Three of the twenty fields — required car parking spaces, total special requests and booking changes — accumulate over a booking's life, and their recording in this dataset is entangled with the outcome: a parking space is recorded for none of the cancelled bookings at either property, against 4.2 percent of arriving bookings at the city hotel and 19.0 percent at the resort hotel, the pattern expected if the field is completed at check-in rather than at booking; special requests and booking changes show the same asymmetry in weaker form. A model can read such fields as partial sight of the outcome, so headline results are reported for two specifications: the twenty predictors as recorded, and an audited specification that excludes the three accrual-contingent fields. A further variant also excludes deposit type, given the recording pattern reported in Section 4.3.

3.4 Evaluation

Evaluation has three layers, because a revenue management tool can fail in three different ways. The first layer is discrimination: can the model rank a booking that will cancel above one that will not? We report the area under the ROC curve (AUC-ROC) and, because cancellations are the minority class, the area under the precision-recall curve (AUC-PR) alongside it. The second layer is classification at a working threshold: accuracy, precision, recall and the F1 score. These are reported to keep the study comparable with earlier work, but we do not select the study's operating point by accuracy, for the reason Ampountolas (2025) makes concrete: a model can post precision above 97 percent while missing most of the cancellations that matter. The third layer is calibration: whether bookings with a score of 0.7 actually cancel about 70 percent of the time. Calibration is assessed with calibration curves and the Brier score, and it is not optional in this study, because Section 3.6 sums predicted probabilities across bookings to produce expected cancellations; sums of badly calibrated probabilities are badly wrong in a way that good ranking cannot rescue.

The operating threshold is chosen on cost grounds, inside the training data. The two errors have different prices. A false positive, a booking flagged as high-risk that would not have cancelled, triggers an intervention: a deposit request or a confirmation contact. Its cost is friction and, at the limit, a lost booking, since restrictive treatment can itself drive guests away (Chen, Schwartz & Vargas, 2011; Smith et al., 2015). A false negative, a cancellation the model missed, leaves the room unprotected, and its cost is the expected resale value of the room night, net of any chance of a late resale. We select the threshold that minimises expected total cost on the training period under a stated cost ratio, and we report results across a range of cost ratios rather than a single assumed value, so that a hotel facing different prices can read off its own operating point. All evaluation measures are reported for the test period only, separately for the resort and the city hotel. In this study the thresholds were selected on out-of-fold predictions from three-fold cross-validation within the training period, not on the model's fitted values, so the selection is not tuned to cases the model has memorised; even so, a threshold chosen this way is an estimate, and differences from the plug-in values implied by the cost ratios are expected.

3.5 Explanation with SHAP

Explanation uses SHAP values (Lundberg & Lee, 2017), computed for the best-performing model of each hotel as judged by the evaluation in Section 3.4, and computed on the test period, so that the explanations describe the model as it performs out of time, not as it fits its own training data. The explanation is built at three levels. At the global level, mean absolute SHAP values and summary plots show which booking characteristics move cancellation risk across all bookings, and in which direction. At the local level, we decompose the prediction for individual bookings, including correctly flagged cancellations, missed cancellations and false alarms, because the local explanation is what a manager sees when deciding what to do about one reservation, and because the failure cases show most clearly where the model's reasoning differs from a manager's. At the segment level, SHAP values are aggregated within segments defined by distribution channel, deposit type and lead-time band, separately for each hotel, so that the study can say whether the same driver matters equally for a direct booking and an online travel agency booking, or for a refundable and a non-refundable rate.

Segment-level explanation is also where the design separates two kinds of driver, following the distinction set out in Section 2.3. Actionable drivers are those the hotel can change for a booking, or for the flow of future bookings: deposit type, distribution channel terms, and offers targeted by lead time. Price-for-only drivers, the guest's own cancellation history is the clearest case, cannot be changed for the booking in hand; they can only be priced into the rate, the deposit requirement or the protection level. Keeping these two groups separate in the results is what prevents a true but useless recommendation of the form 'guests with past cancellations cancel more', and turns it instead into a rule about what to charge or require from such bookings. SHAP values are used throughout as explanations of the model's behaviour. They are not treated as causal estimates, and no policy claim in this paper rests on a SHAP value alone.

3.6 From prediction to revenue management decisions

The final analysis translates probabilities into the quantities a revenue manager consumes, in three steps. First, for each arrival date in the test period, the predicted probabilities of all bookings for that date are summed to give expected cancellations, and net demand is computed as bookings on hand minus expected cancellations. This is the number that protection and pricing decisions actually need, and it is only as good as the calibration established in Section 3.4, which is why calibration is treated as a first-class criterion in this study. Second, expected cancellations are reported with an uncertainty interval, treating each booking's cancellation as an independent draw, because a protection level set from a point forecast alone is over-precise. The dataset records no room capacity, so converting that interval into a property-specific overbooking limit — where the cost of crossing it is a denied guest, with the consequences documented by Noone and Lee (2011) — is left to the property; Section 7 states this limit of the study plainly. Third, the cost-based threshold of Section 3.4 becomes a retention rule: bookings whose predicted probability exceeds the threshold are flagged for intervention, a confirmation contact, or a deposit request timed before the free-cancellation deadline passes, the timing that Chen, Schwartz and Vargas (2011) show is the element of policy that actually changes behaviour. For each candidate rule, Section 4.2 reports the flag share, precision and recall it produces on the test period at the stated cost ratios, so that the rule is evaluated as a decision, not only as a classification. Section 6 states the resulting rules in the form a manager would apply them.

4. Results

4.1 Model performance out of time

Table 1 reports the test-period results for the two hotels, estimated separately as set out in Section 3. The city hotel contributes 59,497 bookings to the training period (41.1 percent cancelled) and 19,833 to the test period (43.5 percent cancelled); the resort hotel contributes 30,045 (25.7 percent) and 10,015 (34.0 percent). Each model uses the same 20 predictors. Thresholds in the last three columns were chosen inside the training period only, under the three cost ratios of Section 3.4.

Table 1
Out-of-time performance on the test period, by hotel (as-recorded specification). City Hotel (test n = 19,833)

Model	ROC-AUC	PR-AUC	Brier	Accuracy	Precision	Recall	F1	Thr (2x)	Thr (3x)	Thr (5x)
XGBoost	0.840	0.831	0.162	0.761	0.770	0.643	0.701	0.45	0.20	0.09
Random forest	0.818	0.810	0.176	0.741	0.729	0.642	0.683	0.38	0.26	0.12
Logistic regression	0.817	0.818	0.172	0.740	0.696	0.714	0.705	0.50	0.45	0.14

Resort Hotel (test n = 10,015)

Model	ROC-AUC	PR-AUC	Brier	Accuracy	Precision	Recall	F1	Thr (2x)	Thr (3x)	Thr (5x)
XGBoost	0.849	0.757	0.151	0.769	0.764	0.464	0.577	0.26	0.25	0.13
Logistic regression	0.833	0.738	0.177	0.708	0.546	0.828	0.658	0.62	0.53	0.36
Random forest	0.826	0.720	0.161	0.761	0.743	0.455	0.564	0.35	0.25	0.14

XGBoost is the strongest model at both hotels on discrimination and on the Brier score, and it is the model carried into the explanation and translation steps below. Its margin over logistic regression is real but modest: 0.023 AUC points at the city hotel (0.840 against 0.817) and 0.016 at the resort hotel (0.849 against 0.833). Random forest adds almost nothing over the baseline at the city hotel and falls behind it at the resort hotel. The pattern matters for the study's claim: under an honest out-of-time test, most of what is predictable about a cancellation is already visible to a well-specified statistical model, and the ensembles earn a small, consistent improvement rather than a transformation. At the resort hotel the ensembles also trade recall for precision at the 0.5 threshold: XGBoost catches only 46.4 percent of the cancellations that occur, against 82.8 percent for logistic regression, which is one reason the operating point is set on cost grounds rather than left at 0.5. One qualification on the Brier comparison: the logistic regression's class weighting shifts its probabilities upward by design, which penalises its Brier score while leaving its ranking intact, so the Brier gap overstates the ensembles' calibration advantage. Bootstrap confidence intervals (300 resamples of the test period) put the XGBoost AUC at 0.840 [0.835, 0.846] for the city hotel and 0.849 [0.841, 0.857] for the resort hotel, against 0.817 [0.811, 0.822] and 0.833 [0.825, 0.842] for logistic regression; the interval for the

difference excludes zero at both properties ([0.021, 0.027] and [0.011, 0.021]), so XGBoost's advantage, though small, is not sampling noise. Figure 4 shows the ROC and precision–recall curves for the audited specification.

The sensitivity check promised in Section 3.2 can be read off the same design. Adding country of origin to the predictor set raises test AUC by 0.037 to 0.040 at the city hotel and by 0.029 to 0.042 at the resort hotel, depending on the model: for XGBoost, from 0.840 to 0.877 and from 0.849 to 0.884. The gain is not trivial, and we report it openly; the main specification still excludes the variable, because in these extracts the country entry was not final until check-in and a benchmark that a hotel can reproduce at booking time cannot rest on it.

Calibration is good in the middle of the range and imperfect at the ends (Fig. 1). At the city hotel the deciles track the diagonal closely up to a predicted probability of about 0.42, then overshoot: bookings scored near 0.78 cancelled 63 percent of the time, and those scored near 0.92 cancelled 83 percent of the time, before the top decile — the non-refundable cluster of Section 4.3 — cancels in full. At the resort hotel the middle deciles under-predict throughout (a decile scored 0.25 cancelled 37 percent of the time), which is the rate shift of Section 4.5 visible at booking level. A property that uses these scores for pricing should recalibrate on recent data before trusting the levels, even where the ranking is sound.

The recording audit of Section 3.3 is material to performance. Excluding the three accrual-contingent fields lowers the XGBoost AUC to 0.794 at the city hotel and 0.784 at the resort hotel (logistic regression: 0.778 and 0.761); excluding deposit type as well brings both properties to 0.777, the level of the conventional baselines in Yang and Miao (2024). Excluding deposit type alone costs almost nothing (0.832 and 0.848). Two conclusions follow. XGBoost leads in every specification, by a margin of 0.016 to 0.034; random forest does not — it falls below logistic regression in both reduced specifications (0.769 against 0.778 with the accrual-contingent fields removed and 0.791 against 0.800 with deposit type removed at the city hotel; 0.755 against 0.761 and 0.819 against 0.830 at the resort hotel), so the ensemble advantage in this study is XGBoost's alone. And part of what this dataset makes predictable lives in fields whose recording is entangled with the outcome, so the honest range for out-of-time discrimination on it is roughly 0.78 to 0.85, depending on how much of the recording process the analyst is willing to trust. The explanations that follow are reported for the as-recorded specification, with the audited specification's ranking stated wherever the two differ. The size of deposit type's SHAP values and the small AUC cost of removing it are consistent with one another. The signal is concentrated in a small block of bookings recorded as non-refundable, and other fields partly proxy for it, so the model loses little when the variable is dropped even though it relies on it heavily when present. This is what a recording pattern would produce, and it is a further reason not to read the SHAP magnitude literally.

Table 2
The recording audit at a glance: XGBoost ROC-AUC by specification (logistic regression in parentheses, random forest in brackets).

Specification	City XGB	City LR	City RF	Resort XGB	Resort LR	Resort RF
As recorded (20 predictors)	0.840	0.817	0.818	0.849	0.833	0.826
Audited (17 predictors)	0.794	0.778	0.769	0.784	0.761	0.755
Audited, deposit type also excluded (16 predictors)	0.777	0.743	0.735	0.777	0.754	0.738
Deposit type alone excluded (19 predictors)	0.832	0.800	0.791	0.848	0.830	0.819

4.2 Cost-based thresholds

The thresholds in Table 1 fall as the cost of a missed cancellation rises relative to a false alarm, as they should, and they fall much further at the city hotel, where the higher base rate makes an aggressive flagging rule cheaper to run. At the middle ratio (a missed cancellation priced at three false alarms), the operating threshold is 0.20 for the city hotel's XGBoost model and 0.25 for the resort hotel's. Applied to the test period, the city rule flags 14,198 of 19,833 bookings (71.6 percent) for attention, and the resort rule flags 4,471 of 10,015 (44.6 percent). These are large shares, and they are the honest consequence of the stated costs: when a missed cancellation costs three times a needless intervention, the model should speak up often. A hotel that prices the two errors equally would work near the 2x column instead, and the table lets it read off that point without re-estimating anything.

Table 3
Performance at the cost-based thresholds on the test period (XGBoost, as recorded).

Hotel	Ratio	Threshold	Flagged	Precision	Recall	F1
City	2x	0.45	38.1%	0.757	0.663	0.707.
City	3x	0.20	71.6%	0.566	0.931	0.704.
City	5x	0.09	86.4%	0.497	0.987	0.661.
Resort	2x	0.26	43.6%	0.606	0.778	0.681.
Resort	3x	0.25	44.6%	0.603	0.792	0.685.
Resort	5x	0.13	59.1%	0.522	0.908	0.663.

Table 3 reports what those thresholds do on the test period for the model carried forward. At the 2x threshold the city rule is genuinely targeted — it flags 38 percent of bookings at a precision of 0.76 — while at 3x it becomes a broad watchlist, flagging 72 percent to catch 93 percent of cancellations.

On the audited specification the thresholds move in both directions, but at the middle ratio both hotels flag more bookings than as recorded: 73.9 percent at the city hotel (against 71.6) and 48.6 percent at the resort

hotel (against 44.6). XGBoost's thresholds are 0.33, 0.21 and 0.12 at the city hotel and 0.31, 0.23 and 0.12 at the resort hotel; at the middle ratio the city rule flags 73.9 percent of test bookings at precision 0.538 and recall 0.915, and the resort rule flags 48.6 percent at precision 0.542 and recall 0.775. The operating rules in Section 6 are read from this specification (Table 3b).

Table 3
b. Performance at the cost-based thresholds on the audited specification (XGBoost).

Hotel	Ratio	Threshold	Flagged	Precision	Recall	F1
City	2x	0.33	63.3%	0.581	0.846	0.689.
City	3x	0.21	73.9%	0.538	0.915	0.678.
City	5x	0.12	86.9%	0.487	0.973	0.649.
Resort	2x	0.31	40.5%	0.584	0.696	0.635.
Resort	3x	0.23	48.6%	0.542	0.775	0.638.
Resort	5x	0.12	64.9%	0.461	0.880	0.605.

4.3 What drives the predictions

SHAP values for the XGBoost models, computed on the test period, differ sharply between the two properties. At the city hotel the leading driver is deposit type (mean absolute SHAP 1.47), well ahead of the next features: total special requests (0.68), market segment (0.41), lead time (0.36), average daily rate (0.26) and customer type (0.25). Deposit type is a variable the hotel itself sets, and it leads the audited specification as well (1.52). Under the strict definition of Section 3.5 — deposit type and distribution channel terms — the actionable share of the city model's importance is 35.0 percent as recorded and 48.0 percent audited, and in both cases almost all of it is deposit type: distribution channel terms carry about 1 percent. Guest-driven fields such as special requests are not counted as actionable here, however much weight they carry, because the definition of Section 3.5 does not allow it. Figure 2 shows the global SHAP importance for both properties as recorded.

One caution applies to the headline driver. In these extracts, bookings recorded as non-refundable cancel at rates no real rate plan produces: 99.8 percent of the city hotel's 12,868 non-refundable bookings are recorded as cancelled, and 96.0 percent of the resort hotel's 1,719. A guest who chooses a non-refundable rate is normally more certain of travelling, not less, so a cancellation rate near one hundred percent points to a recording pattern in the data — most plausibly rate-plan changes and rebookings written into the extract as cancellations. Deposit type must therefore be read as the strongest recorded driver, with its size inflated by this pattern. What survives the caveat is the comparison inside the city hotel: deposit terms separate speculative from committed bookings far more strongly in group and offline-agent segments than elsewhere (Table 5), and the resort hotel's weaker deposit signal shows the pattern is not purely mechanical.

At the resort hotel the as-recorded order is headed by required car parking spaces (1.03) — and that lead is the recording artifact of Section 3.3: the field is empty for every cancelled booking in the data, so the model is partly reading the property's paperwork rather than the guest. In the audited specification the resort ranking is

lead time (0.61), market segment (0.42), deposit type (0.36), customer type (0.27) and average daily rate (0.24): timing and booking terms, not guest requests. The strictly actionable share at the resort hotel is 8.8 percent as recorded and 16.4 percent in the audited specification. The explanation therefore supports different conversations at the two properties: at the city hotel, a deposit-policy conversation; at the resort hotel, where most of the signal describes who books and when, the realistic responses are pricing and protection levels rather than intervention.

Local explanations make the same point at the level of a single booking (Table 4; two are drawn as waterfall plots in Fig. 5). The city hotel's most confident correct flag was a group booking taken 200 days out on a non-refundable rate through a travel agent: the model returned a probability of effectively 1.00, driven almost entirely by the deposit term (+ 5.90 in log-odds), and the booking did cancel. Its clearest miss was also a group booking, taken 409 days out with no deposit; one special request and the customer type pulled the score down to 0.004, and the booking cancelled anyway — a reminder that a long lead time cuts both ways and that the model cannot see a group's internal plans. At the resort hotel, the loudest false alarm was a direct booking whose guest had cancelled twice before: previous cancellations alone contributed + 5.40, the model returned 0.995, and the guest arrived. That is the price-for-only category behaving as Section 2.3 predicts — the score is information for pricing and protection, and a poor reason, on its own, to contact the guest.

Table 4
Three local explanation examples from the test period (XGBoost).

Case	Actual	Predicted	Main drivers (SHAP, log-odds)
City Hotel, flagged and cancelled	cancelled	1.00	deposit type (+ 5.90), lead time (+ 1.05).
City Hotel, missed cancellation	cancelled	0.004	special requests (-1.22), deposit type (-0.60), customer type (-0.55).
Resort Hotel, false alarm	arrived	0.995	previous cancellations (+ 5.40), daily rate (+ 1.04).

4.4 Segment differences

Aggregating SHAP values within market segments (Table 5) sharpens the city result. Deposit type is far more decisive for group bookings (4.96) and for offline travel-agent and tour-operator bookings (2.79) than for any other segment (0.46 to 1.55), which is consistent with block bookings being held speculatively until a deposit is requested. A guest's record of previous completed bookings matters in the Aviation, Complementary and Corporate segments (0.63 to 0.82) and almost nowhere else (0.02 to 0.06 in Direct and the travel-agent channels), so history-based pricing would bite only on a narrow slice of the book. At the resort hotel, deposit type again spikes for group bookings (1.44) against 0.20 to 0.42 elsewhere. One caution on reading Table 5: the segment gradient in deposit importance tracks the segment gradient in non-refundable prevalence — 55 percent of group bookings and 29 percent of offline TA/TO bookings at the city hotel are recorded as non-refundable, against almost none in the direct and online segments — so the table shows where the recording pattern of Section 4.3 concentrates, as much as where behaviour concentrates. Within segments the pattern is the same: among city-hotel group bookings, no-deposit bookings cancel at 31.2 percent while non-

refundable bookings cancel at 99.9 percent — a gap no real rate plan produces, in the segment where behaviour should vary most.

Table 5
Mean absolute SHAP value of deposit type, by market segment.

Market segment	City Hotel	Resort Hotel
Groups	4.96	1.44
Offline TA/TO	2.79	0.26
Corporate	1.55	0.42
Direct	0.56	0.22
Aviation	0.47	—
Complementary	0.47	0.23
Online TA	0.46	0.20

The channel and lead-time cuts add one more pattern. Within the city hotel, the non-refundable deposit term matters most inside the TA/TO channel (1.29) and for corporate bookings (0.98), and its weight grows with lead time, from 0.56 for bookings made in the final week to 1.58 for bookings made three to six months out — speculative holding is a long-lead phenomenon. At the resort hotel, required parking dominates the direct channel (1.81) and the shortest lead band (1.62), which fits the drive-to reading of Section 4.3: the guests a resort can least afford to lose are the ones who booked a car space at short notice. (The parking cuts in this paragraph describe the as-recorded specification and inherit the recording caveat of Sections 3.3 and 4.3.)

4.5 From probabilities to net demand

Summing the audited XGBoost probabilities by arrival date gives the translation of Section 3.6 (Fig. 3). At the city hotel, over 161 arrival days in the test period, predicted cancellations total 8,602 against 8,627 actual, an understatement of 0.3 percent: expected net demand averages 70 arrivals a day against 123 bookings a day on hand. The city model is close to calibrated in aggregate, so a buffer read from it would be roughly right on average over the season. At the resort hotel, over 180 arrival days, predicted cancellations total 2,830 against 3,403 actual, an understatement of 16.8 percent. (On the as-recorded specification the two totals are 8,839 and 2,850.) The cause is visible in Table 1: the resort test period cancelled at 34.0 percent against 25.7 percent in training, and a model trained in the calmer period carries the calmer rate forward. The practical lesson is the one Section 3.6 anticipates: net demand and the overbooking buffer are only as current as the cancellation rate behind them, and the buffer must be re-estimated as that rate drifts, not set once from a fitted model. One framing caution: this translation is a retrospective accounting on final records. The bookings summed for an arrival date are all the bookings recorded for that date, including some that had already cancelled and some made only days before; a live system would work from the bookings on hand at a snapshot taken days or weeks earlier, which this dataset cannot reconstruct, because cancellation dates are not recorded. The comparison with the naive rate is like-for-like, since both are computed on the same final records, but the errors reported here are not the errors of a live forecast.

Benchmarked at day level, the translation earns its keep. Against a naive forecast that applies the training-period cancellation rate to each day's bookings on hand, the model's mean absolute error in net demand falls from 12.6 to 6.1 rooms per day at the city hotel (bias + 0.2 rooms against + 2.9) and from 6.6 to 4.4 rooms per day at the resort hotel (bias + 3.2 rooms against + 4.6). The probabilities also carry their own uncertainty: treating each booking's cancellation as an independent draw, the 95 percent interval around a day's expected cancellations averages plus or minus 8.7 rooms at the city hotel (7.9 percent of the day's bookings) and plus or minus 5.5 rooms at the resort hotel (10.5 percent). A protection level set from the point forecast alone is therefore over-precise; the interval, not the point, is the quantity a manager should protect to.

5. Discussion

The first thing the results clarify is what an honest test does to the claims made for machine learning in this area. Under an out-of-time, leakage-audited protocol, XGBoost beat a logistic regression baseline by 0.016 to 0.023 AUC points at both hotels, in both the as-recorded and the audited specification. The absolute level depends on how much of the recording process is trusted: 0.84 to 0.85 as recorded, 0.78 to 0.79 once the three accrual-contingent fields are excluded, and 0.777 without deposit type as well. The audited levels match the temporal-split baselines of Yang and Miao (2024); the as-recorded levels should be read as an upper bound. All of them sit well below the accuracies above 95 percent that appear when post-outcome variables or random splits inflate the estimate. That is not a disappointing result; it is the study's first finding. A benchmark that a reader can reproduce sets the level at which prediction can honestly support decisions, and it gives managers a defence against vendor claims built on leaked data. The comparison with Yang and Miao (2024) needs one qualification: their standard-model baselines sit near 0.78 and their temporal network reaches 0.84 to 0.86 under a pooled two-hotel architecture, while this study estimates each property separately with a different baseline set, so the figures are indicative rather than directly comparable. What both studies agree on is the ordering: honest temporal evaluation places this problem between the high 0.70s and the mid 0.80s, not in the high 0.90s.

The second finding is that the strongest actionable driver is a policy variable. Deposit type leading the city hotel's explanations, and carrying far more weight in group and offline-agent segments than elsewhere, extends the cancellation-policy literature (Chen, Schwartz & Vargas, 2011; Chen & Xie, 2013; Riasi, Schwartz & Chen, 2018) in a specific direction: policy does not merely correlate with cancellation in aggregate, it identifies where a deposit requirement does its work. A hotel does not need to tighten terms for everyone, which the fairness evidence warns against (Smith et al., 2015); it needs deposit discipline on the segments where speculative holding is concentrated. The explanation layer is what turns a general policy lever into a targeted one. The size of the deposit signal in these data is inflated by the recording pattern set out in Section 4.3, and its segment gradient partly tracks where non-refundable bookings are prevalent (Section 4.4); the argument here rests on deposit terms being the largest block of signal at the city hotel in both specifications, not on the signal's literal size.

The contrast between the two properties is the third finding. At the city hotel the largest single block of signal is a policy variable — deposit type — even after the recording caveats, so the model hands the manager one real lever and a short intervention list. At the resort hotel the audited ranking is led by lead time and market segment, the strictly actionable share is under a fifth, and the honest output is a protection level: risk that

cannot be talked out of a booking has to be priced into overbooking and deposit requirements instead. The same prediction system therefore supports different management conversations in different properties. Where actionable drivers dominate, the output is an intervention list. Where they do not, the honest output is a protection level. Treating the two situations as one problem is how prediction projects lose credibility with the managers who run them.

Fourth, the resort hotel's translation error carries a methodological lesson. Its test-period cancellation rate rose from 25.7 to 34.0 percent, and the model, trained in the calmer period, understated cancellations by about a sixth. The test window runs from March to August, so part of that shift may be seasonal rather than drift in the strict sense; the operational point survives either way. This is the validity problem that Silvestre, Antonio and Carrasco (2026) address, observed here in the quantity a manager actually consumes. A cancellation model is not an installation but a maintained instrument: the rate behind the buffer has to be monitored and the model re-estimated on a schedule, and a widening gap between expected and realised cancellations is itself a signal worth reporting to management.

Finally, the design choices in this study were made with adoption in mind. Managers abandon systems that err visibly and cannot be adjusted (Dietvorst, Simmons & Massey, 2015), and they use systems more when they can adjust the output and see the reasoning (Dietvorst, Simmons & Massey, 2018; Logg, Minson & Moore, 2019). A local explanation attached to each flagged booking, and a threshold presented as a menu across cost ratios rather than as a fixed rule, are not cosmetic features; on the behavioural evidence, they are the conditions under which a model like this one survives in practice.

6. Implications for Revenue Managers

The practical output of this study is a short set of rules a revenue manager can apply: which bookings to score at reservation time, which high-risk bookings to ask for a deposit or a confirmation, and how risk scores should feed into overbooking decisions on high-demand dates.

Rule 1: Score every booking when it is made, and score it again at each review point as its details change. A booking change is itself information (it carries weight in both hotels' explanations), so a risk score computed only at reservation quietly goes stale. One caution from this study's design: the scores here are computed from the values finally recorded for each booking, so a live system should expect weaker performance than the figures in Table 1, which are an upper bound: excluding the three accrual-contingent fields alone lowers AUC by 0.05 to 0.07.

Rule 2: Set the flagging threshold from your own cost ratio, using the menu in Table 3b, and work the flagged list in score order. At the middle ratio the city rule flags about three bookings in four; if the team cannot contact that many guests, the answer is to start from the top of the list, not to raise the threshold without saying so, because an unstated threshold change is an unstated change in the hotel's risk policy.

Rule 3: Match the response to the driver. For flagged group and offline-agent bookings, where deposit type carries the risk, the response is a deposit request or firmer cancellation terms — piloted and measured, not assumed to work, because the deposit pattern in these data is partly a recording artifact (Section 4.3) and this study does not estimate the causal effect of asking for a deposit. For flagged direct bookings, it is a

confirmation contact with a flexible rebooking offer. Where the score is high only because of the guest's history, do not intervene at all: price it, through the deposit required at booking, and leave the guest's experience alone.

Rule 4: Feed net demand, not bookings on hand, into protection and overbooking decisions, and refresh the cancellation rate behind the buffer at least monthly. In the resort hotel's test period, a season-old rate would have understated cancellations by about a sixth, which is the difference between a buffer and a guess.

Rule 5: Keep the override human, and keep the explanation attached to the flag. A manager who can see why a booking was flagged can dismiss the flags that are wrong for reasons the data cannot hold; a pattern in those dismissals is diagnostic information about the model, not a discipline problem.

7. Limitations and Future Research

The data come from two hotels in one country, so the findings may not transfer directly to other markets. The data also end in 2017, before the pandemic changed booking behaviour; the same approach should be tested on recent data. Future research can add text data, such as guest messages and reviews, and can test whether acting on the risk scores actually reduces cancellation losses in a live setting. The cost ratios used to choose the thresholds are illustrative rather than estimated from hotel accounts; a property applying the framework should replace them with its own intervention costs and resale values.

Six further limits should be kept in view. First, the test period covers March to August 2017 only, so the rate shift reported in Section 4.5 mixes any genuine drift with seasonality, and the study cannot separate the two. Second, all predictors are the values finally recorded for each booking; a system scoring bookings as they arrive would see earlier, less complete values, so the reported performance is an upper bound for live use. Third, the deposit-type pattern is partly an artifact of how these extracts record non-refundable bookings (Section 4.3); the direction of the deposit finding is consistent with the policy literature, but its size in this dataset should not be read literally, and nothing in the study shows that changing a deposit term changes a guest's behaviour. Fourth, scoring guests partly on their own cancellation history raises a fairness question the study does not resolve: the framework uses such scores for pricing and protection only, never for refusing service, but a property adopting it should state that boundary explicitly. Fifth, the dataset records no room capacity, so the study cannot compute a property-level overbooking limit or a denied-service tolerance; it reports expected cancellations with an uncertainty interval, and converting that interval into rooms above capacity remains the property's own calculation.

Sixth, three predictor fields are recorded in a way that is entangled with the outcome (Section 3.3); the audited specification bounds their effect, but only a dataset that timestamps requests, changes and parking at the moment they occur would remove the ambiguity entirely.

8. Conclusion

This study set out to give revenue managers a cancellation model they can check and use. On a public dataset of 119,390 bookings, under an out-of-time, leakage-audited protocol, XGBoost predicted cancellations with an AUC of 0.840 at a city hotel and 0.849 at a resort hotel as recorded, and 0.794 and

0.784 once three accrual-contingent fields are excluded, only modestly ahead of logistic regression in both cases — a result that argues for judging such systems by their evaluation design rather than by headline accuracy. Explanation separated what a hotel can change from what it can only price: deposit policy dominates the actionable drivers at the city property, while the resort property's risk is carried more by who books than by what the hotel does. Translated into arrival-date terms, the probabilities produced net-demand figures close to realised demand where booking behaviour was stable, and visibly biased where it had shifted, which is the honest boundary of the method. The contribution is deliberately practical: a reproducible benchmark, a decision-oriented explanation, and a quantified translation that a revenue manager can adopt with ordinary reservation data.

Declarations

Declaration of AI use

An AI assistant was used under the authors' direction to carry out the analyses reported here and to assist with drafting, review and editing of the text. All design decisions, interpretations and the final text are the authors' own, and the authors take full responsibility for the content.

Data and code availability

The dataset is public (Antonio, de Almeida & Nunes, 2019c). The analysis notebooks behind this study are available from the authors on request.

References

1. Adil M, Ansari MF, Alahmadi A, Wu JZ, Chakrabortty RK (2021) Solving the problem of class imbalance in the prediction of hotel cancelations: A hybridized machine learning approach. Processes 9(10):1713. https://doi.org/10.3390/pr9101713
2. Ampountolas A (2025) Predicting hotel booking cancellations: A comprehensive machine learning approach. J Revenue Pricing Manag 24(6):539–550. https://doi.org/10.1057/s41272-025-00532-x
3. Ampountolas A, Legg M (2026) A comprehensive approach to enhancing short-term hotel cancellation forecasts through dynamic machine learning models. Tour Econ 32(2):321–341. https://doi.org/10.1177/13548166251318768
4. Antonio N, de Almeida A, Nunes L (2017a) Predicting hotel booking cancellations to decrease uncertainty and increase revenue. Tourism Manage Stud 13(2):25–39. https://doi.org/10.18089/tms.2017.13203
5. Antonio N, de Almeida A, Nunes L (2017b) Predicting hotel bookings cancellation with a machine learning classification model. In Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1049–1054). IEEE. https://doi.org/10.1109/ICMLA.2017.00-11
6. Antonio N, de Almeida A, Nunes L (2019a) An automated machine learning based decision support system to predict hotel booking cancellations. Data Sci J 18(1):32. https://doi.org/10.5334/dsj-2019-032

7. Antonio N, de Almeida A, Nunes L (2019b) Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior. Cornell Hospitality Q 60(4):298–319. https://doi.org/10.1177/1938965519851466
8. Antonio N, de Almeida A, Nunes L (2019c) Hotel booking demand datasets. Data Brief 22:41–49. https://doi.org/10.1016/j.dib.2018.11.126
9. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
10. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
11. Chen C-C, Schwartz Z, Vargas P (2011) The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. Int J Hospitality Manage 30(1):129–135. https://doi.org/10.1016/j.ijhm.2010.03.010
12. Chen C-C, Xie KL (2013) Differentiation of cancellation policies in the U.S. hotel industry. Int J Hospitality Manage 34:66–72
13. Chen S, Ngai EWT, Ku Y, Xu Z, Gou X, Zhang C (2023) Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. Decis Support Syst 170:113959. https://doi.org/10.1016/j.dss.2023.113959
14. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794)
15. Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114–126. https://doi.org/10.1037/xge0000033
16. Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Manage Sci 64(3):1155–1170. https://doi.org/10.1287/mnsc.2016.2643
17. Duong TN, Phan DT (2026) AI-driven prediction of hotel booking cancellations and time-to-cancellation using a two-stage deep learning framework. Inform Technol Tourism. https://doi.org/10.1007/s40558-026-00394-y
18. Falk M, Vieru M (2018) Modelling the cancellation behaviour of hotel guests. Int J Contemp Hospitality Manage 30(10):3100–3116. https://doi.org/10.1108/IJCHM-08-2017-0509
19. Hermawan A, Amalia I, Rafif M, Azzahra NA, Ragasa R (2025) Optimizing machine learning models for predicting and mitigating hotel booking cancellations. Jurnal Publikasi Teknik Informatika 4(2):21–41. https://doi.org/10.55606/jupti.v4i2.4055
20. Herrera A, Arroyo A, Jiménez A, Herrero A (2024) Forecasting hotel cancellations through machine learning. Expert Syst 41(9):e13608. https://doi.org/10.1111/exsy.13608
21. Kundu S, Roy S, Shukla A, Mateen A (2025) Unveiling cancellation dynamics: A two-stage model for predictive analytics. Data Knowl Eng 160:102467. https://doi.org/10.1016/j.datak.2025.102467
22. Liberman U, Yechiali U (1978) On the hotel overbooking problem: An inventory system with stochastic cancellations. Manage Sci 24(11):1117–1126

23. Liu Z, Jiang P, Wang J, Du Z, Niu X, Zhang L (2023) Hospitality order cancellation prediction from a profit-driven perspective. Int J Contemp Hospitality Manage 35(6):2084–2112. https://doi.org/10.1108/IJCHM-06-2022-0737
24. Liu Z, De Bock KW, Zhang L (2025) Explainable profit-driven hotel booking cancellation prediction based on heterogeneous stacking-based ensemble classification. Eur J Oper Res 321(1):284–301. https://doi.org/10.1016/j.ejor.2024.08.026
25. Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. Organ Behav Hum Decis Process 151:90–103. https://doi.org/10.1016/j.obhdp.2018.12.005
26. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774
27. Noone BM, Lee CH (2011) Hotel overbooking: The effect of overcompensation on customers' reactions to denied service. J Hospitality Tourism Res 35(3):334–357. https://doi.org/10.1177/1096348010382238
28. Pereira LN (2016) An introduction to helpful forecasting methods for hotel revenue management. Int J Hospitality Manage 58:13–23
29. Pereira LN, Cerqueira V (2022) Forecasting hotel demand for revenue management using machine learning regression methods. Curr Issues Tourism 25(17):2733–2750. https://doi.org/10.1080/13683500.2021.1999397
30. Riasi A, Schwartz Z, Chen C-C (2018) A proposition-based theorizing approach to hotel cancellation practices research. Int J Contemp Hospitality Manage 30(11):3211–3228
31. Romero Morales D, Wang J (2010) Forecasting cancellation rates for services booking revenue management using data mining. Eur J Oper Res 202:554–562
32. Sánchez-Medina AJ, C-Sánchez E (2020) Using machine learning and big data for efficient forecasting of hotel booking cancellations. Int J Hospitality Manage 89:102546. https://doi.org/10.1016/j.ijhm.2020.102546
33. Silvestre P, Antonio N, Carrasco P (2026) Navigating uncertainty: Enhancing hotel cancellation predictions with adaptive machine learning. Inform Technol Tourism 28(1):1–28. https://doi.org/10.1007/s40558-025-00349-9
34. Smith SJ, Parsa HG, Bujisic M, van der Rest J-P (2015) Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry. J Travel Tourism Mark 32(7):886–906. https://doi.org/10.1080/10548408.2015.1063864
35. Song H, Li G (2008) Tourism demand modelling and forecasting: A review of recent research. Tour Manag 29(2):203–220
36. Talluri KT, van Ryzin GJ (2004) The theory and practice of revenue management. Springer
37. Viverit L, Heo CY, Pereira LN, Tiana G (2023) Application of machine learning to cluster hotel booking curves for hotel demand forecasting. Int J Hospitality Manage 111:103455. https://doi.org/10.1016/j.ijhm.2023.103455
38. Weatherford LR, Kimes SE (2003) A comparison of forecasting methods for hotel revenue management. Int J Forecast 19(3):401–415

39. Yang D, Miao X (2024) Predicting hotel booking cancellations using tree-based neural network. PeerJ Comput Sci 10:e2473. https://doi.org/10.7717/peerj-cs.2473

Figures

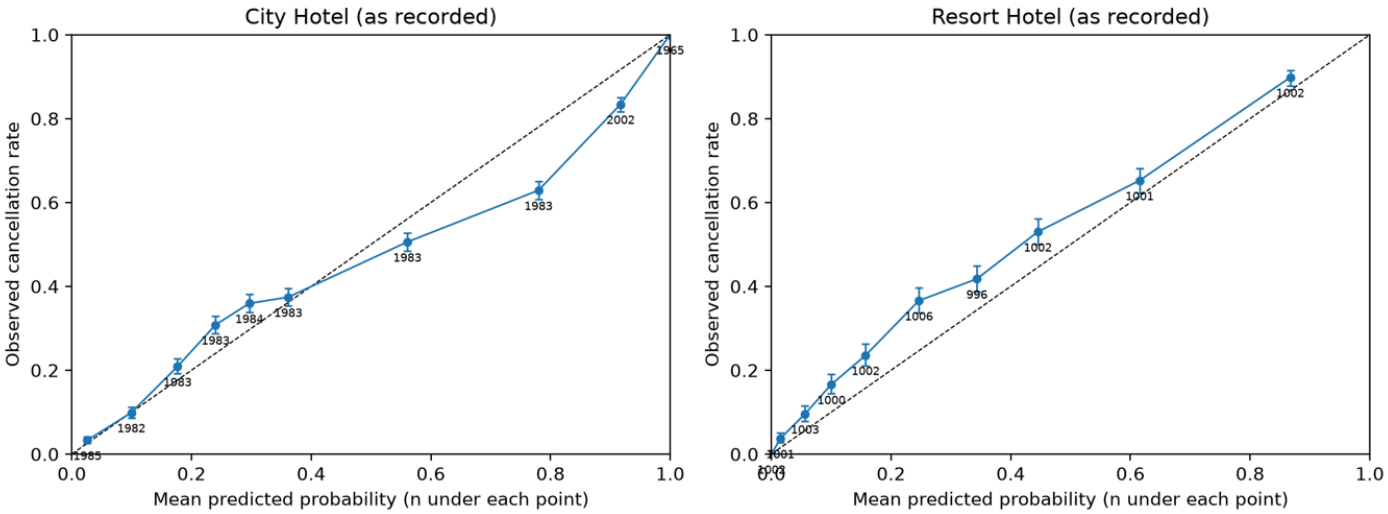


Figure 1

Calibration by predicted-probability decile (XGBoost, test period). (a) City Hotel; (b) Resort Hotel (as recorded). Whiskers are 95 percent binomial confidence intervals, and the number of bookings in each decile is printed under its point. [The chart is embedded in the submission file.]

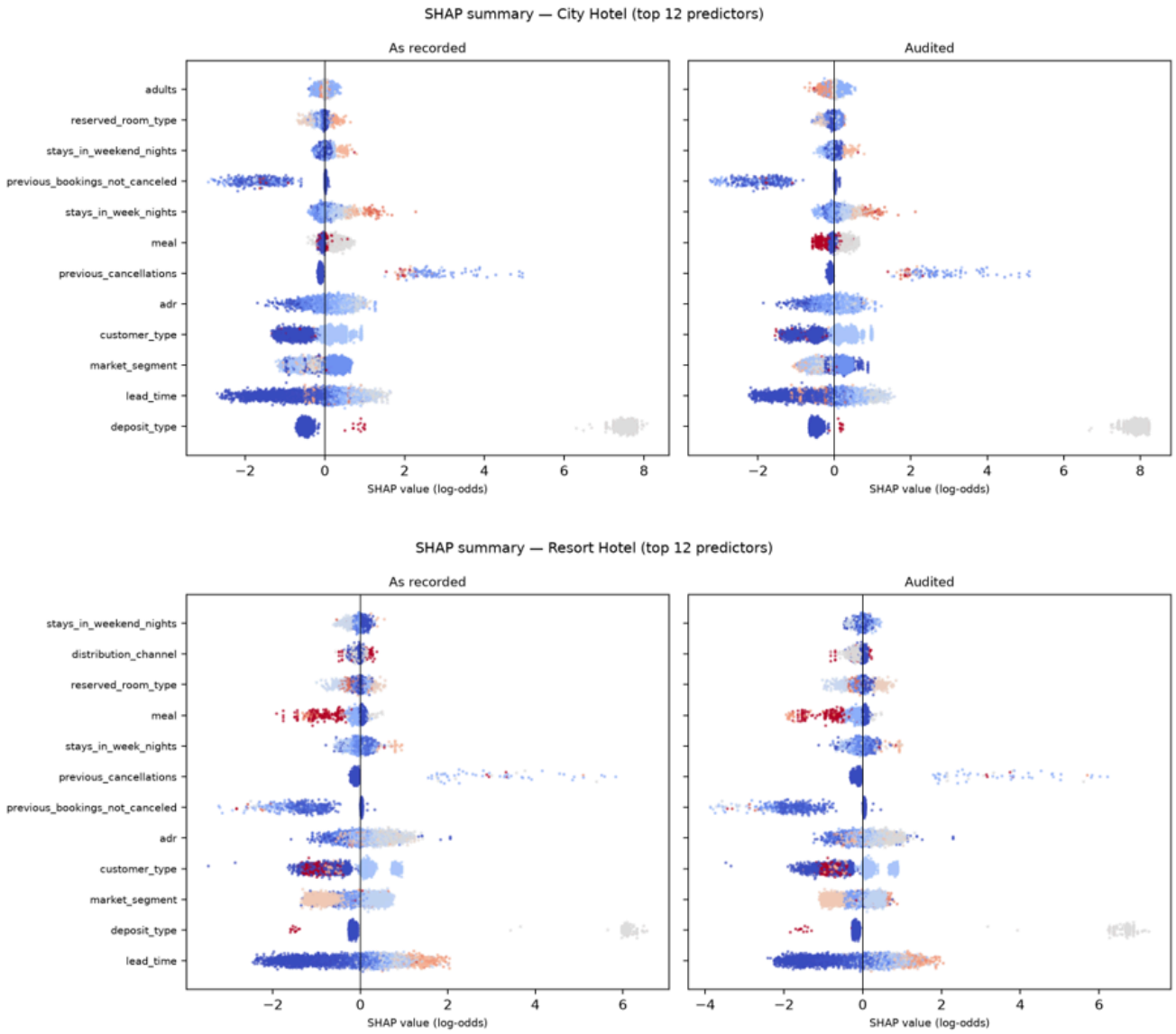


Figure 2

SHAP summary (beeswarm) plots for the XGBoost models on the test period, as recorded (left) and on the audited specification (right). (a) City Hotel; (b) Resort Hotel. The twelve highest-ranked predictors are shown, of twenty in the as-recorded model and seventeen in the audited model; each dot is one booking, its colour is the feature value, and its horizontal position is the SHAP value. [The charts are embedded in the submission file.]ed probability 0.995).

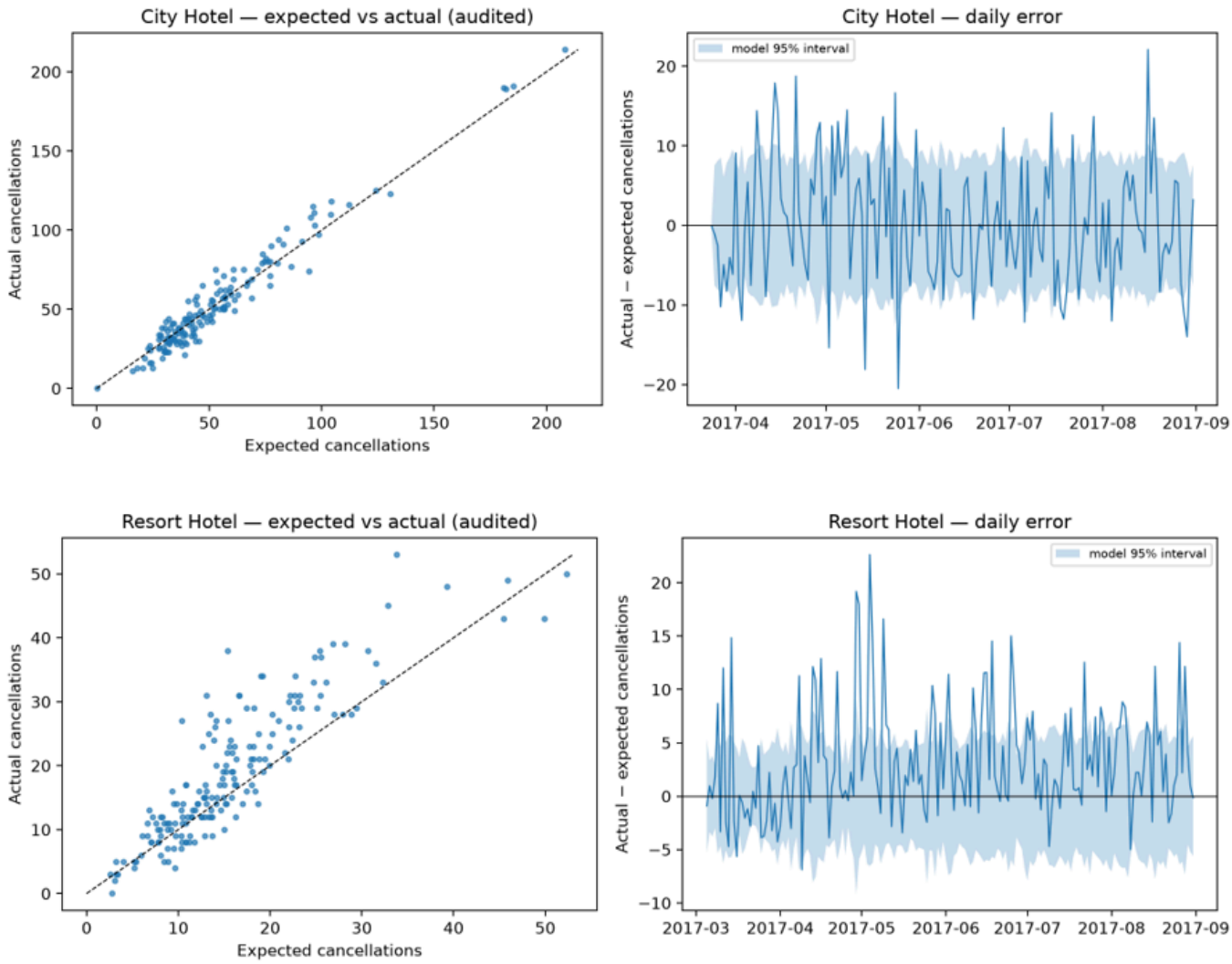


Figure 3

Daily cancellation translation on the audited specification. Left: expected against actual cancellations by arrival date, with the 45-degree line. Right: daily error (actual minus expected) with the model's own 95 percent interval. (a) City Hotel; (b) Resort Hotel.

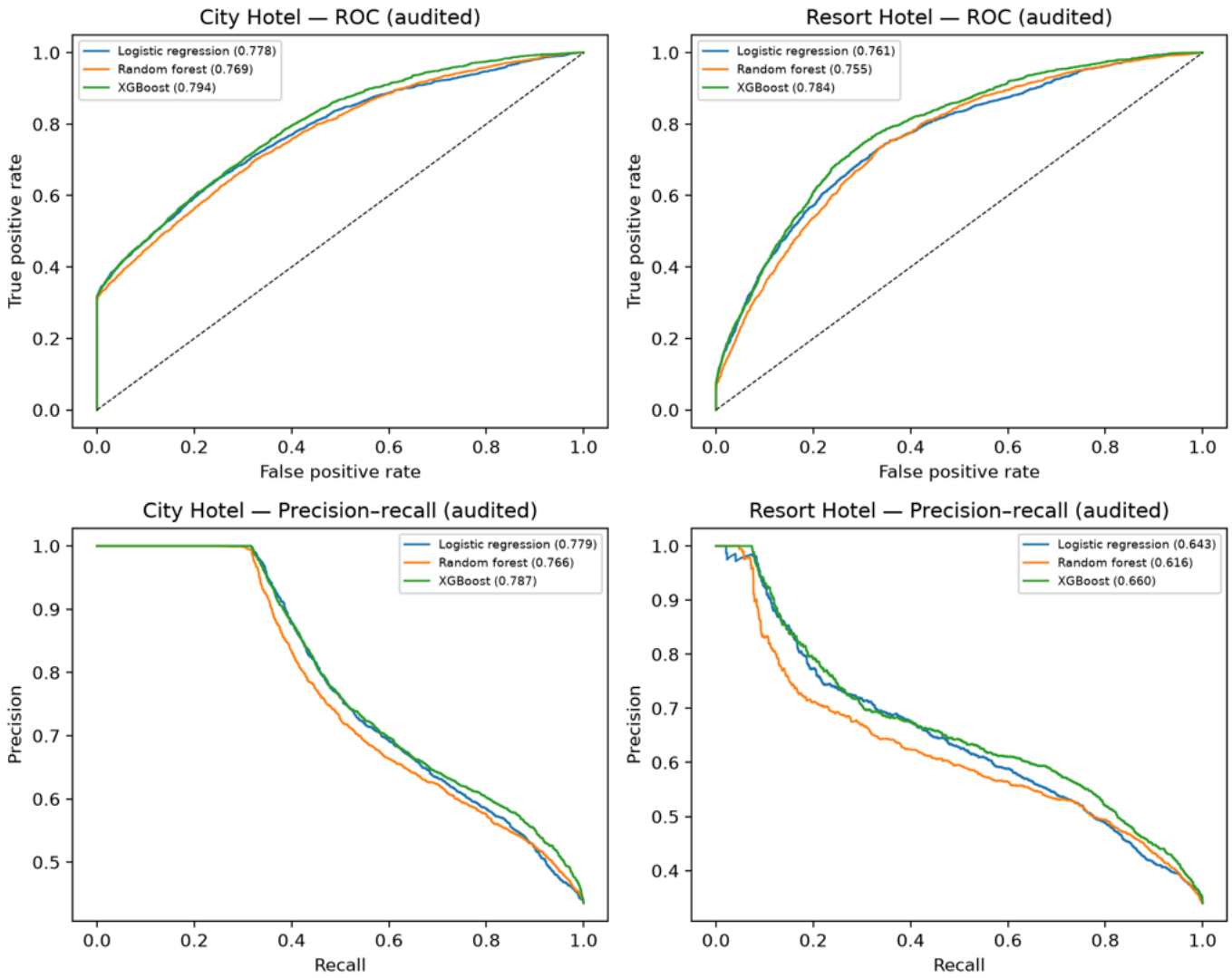


Figure 4

ROC and precision–recall curves on the audited specification (test period). (a) City Hotel, ROC; (b) City Hotel, precision–recall; (c) Resort Hotel, ROC; (d) Resort Hotel, precision–recall.

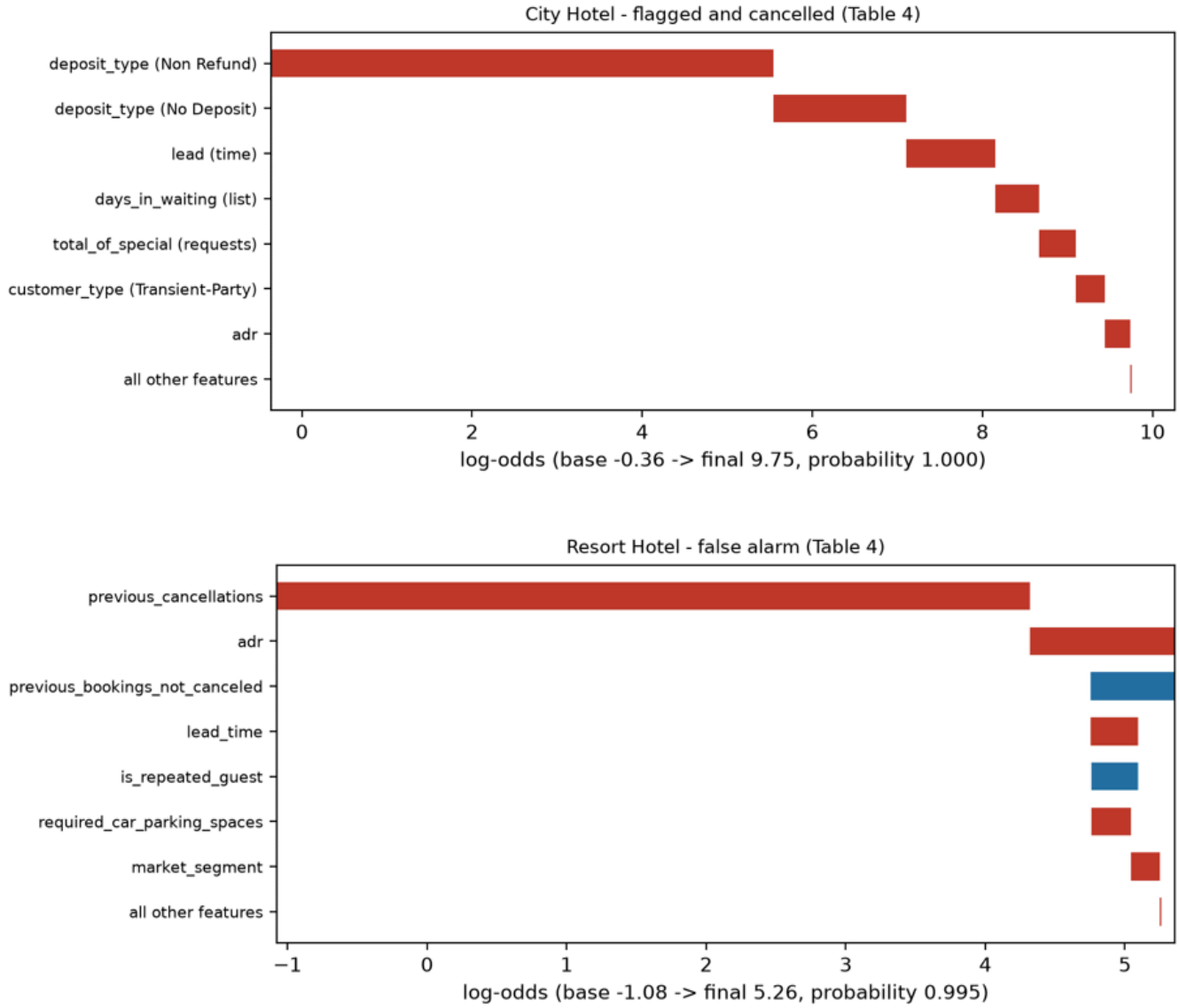


Figure 5

Waterfall plots for two bookings from Table 4. (a) City Hotel, flagged and cancelled (predicted probability 1.00). (b) Resort Hotel, false alarm (predicted probability 0.995).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HotelCancellationModellingNotebookv2.ipynb
- HotelCancellationModellingNotebookv2b.ipynb
- HotelCancellationModellingNotebookv2c.ipynb
- HotelCancellationModellingNotebookv2g.ipynb
- HotelCancellationModellingNotebookv2f.ipynb
- HotelCancellationModellingNotebookv2e.ipynb

- HotelCancellationModellingNotebookv2d.ipynb