PROJECT ABSTRACT

Hotel Booking Cancellation Prediction using Machine Learning

A gradient-boosted classifier predicting which hotel bookings will cancel on the 119,390-booking Hotel Booking Demand dataset, with a strict leakage audit of post-outcome columns, an interactive booking demo, notebook and viva kit.

01. Title

Hotel Booking Cancellation Prediction using Machine Learning

02. Introduction / Background

More than a third of hotel bookings never materialize — guests cancel, rooms sit empty, revenue evaporates. But cancellations are not random: long lead times, no deposit, online-travel-agency bookings and guests with cancellation histories all raise the odds, and the hotel knows all of this at booking time. The Hotel Booking Demand dataset (Antonio, Almeida & Nunes, 2019): 119,390 real reservations from a Portuguese resort and city hotel with 31 booking attributes. The central engineering lesson is the leakage audit — columns like reservation status are only known after the outcome and must be excluded, or the model "predicts" the past.

03. Problem Statement

Revenue-management teams need a cancellation-risk score at booking time so they can intervene (personal outreach, deposit incentives) while there is still time. A defensible project must train only on booking-time features, document every excluded leaky column with justification, and evaluate with AUC plus calibration — the hotel must be able to trust the probabilities.

04. Objectives

- Train a gradient-boosted classifier reaching AUC ≥ 0.85 (design target).
- Perform and document a strict leakage audit excluding all post-outcome columns.
- Evaluate with AUC, precision/recall at the operating threshold and a calibration curve.
- Analyse and visualize what drives cancellations (lead time, deposit type, segment, guest history).
- Build an interactive demo scoring any booking with risk-tiered intervention recommendations.
- Deliver the complete student kit: notebook, trained model, demo, report, PPT and viva Q&A.;

05. Proposed Methodology

The 119,390 bookings (H1 resort: 40,060; H2 city: 79,330) are loaded; the leakage audit removes reservation_status, assigned_room_type, booking_changes and other post-outcome columns with written justification. Remaining features (lead_time, hotel, market_segment, deposit_type, is_repeated_guest, previous_cancellations, adr, special requests) are encoded; a gradient-boosted classifier trains with stratified cross-validation and class weighting. Evaluation reports AUC, precision/recall and calibration; feature-importance and partial-dependence analysis reveal the cancellation drivers for the report.

06. System Architecture / Working

Input (~20 booking-time features) → encoding → gradient-boosted trees → P(cancel) with a risk tier. In the web demo, three sample bookings (OTA long-lead, loyal direct guest, corporate group) preload; every field is editable, and prediction shows the risk gauge, the exact factors pushing risk up or down, a lead-time-band cancellation-rate chart from the dataset, and a tiered action (personal outreach / automated reminder / standard flow). The footer states the demo runs a compact in-browser model while the full boosted model ships separately.

07. Hardware / Software Requirements

- Python 3.10+, scikit-learn, Pandas, NumPy, Matplotlib, Jupyter
- Hotel Booking Demand dataset (public, 119,390 bookings)
- Any modern laptop; training takes ~5–15 min on CPU
- Web demo: any current browser, runs offline after download

08. Expected Outcomes

- AUC ≥ 0.85 (design target), precision/recall, calibration curve
- Leakage-audit table and cancellation-driver analysis documented in the report
- Interactive booking-risk demo with intervention recommendations
- Measured values are reported after the built-to-order training run — the target is not presented as a measured result

09. Applications

Demonstrates applied ML for hospitality revenue management; the leakage-audit pattern applies to churn, no-show and default prediction. The model predicts risk — whether an intervention saves the booking needs A/B testing the project does not run, stated plainly.

10. Future Scope

Intervention A/B-test design with uplift modeling; net-demand forecasting combining cancellation risk with pickup curves; overbooking optimization on top of risk scores.