2025

MLG382
CYO - Porject

Machine Learning

Anele Nkayi 577168
Rourke Veller 601052
Kealeboga Molefe 577482
Willem Booysen 600613

Introduction

In an increasingly competitive hospitality landscape, hotels must balance occupancy maximisation with dynamic pricing while accounting for volatile customer behaviour. Cancellations erode forecast accuracy, and sub-optimal room rates directly suppress revenue. Traditional rule-based revenue-management systems are ill-equipped to model the complex, non-linear relationships that now govern booking behaviour.

Problem Statement

Hotels struggle with two intertwined questions; Will this booking actually materialise?, At what nightly rate should we sell a room to maximise revenue? Rising online-booking volumes have amplified two long-standing revenue-management problems for hotels:

- Demand uncertainty: Last-minute cancellations leave rooms unsold or trigger costly overbooking counter-measures.
- Price optimization: Static rate tables fail to capture seasonality, room mix, and guest behaviour, leaving revenue on the table.

Proposed Solution

This project therefore applies contemporary machine-learning techniques to historical hotel-booking data in order to deliver real-time, reservation-level predictions of (i) cancellation probability and (ii) optimal Average Daily Rate (ADR). The resulting insights are surfaced through an interactive Dash web application intended for day-to-day use by revenue-management staff.

Dataset Overview

Layer	Approach
Data	Use the Hotel Booking Demand dataset 95 000 cleaned records.
Models	- Random Forest Classifier: predict is_canceled. - XGBoost Regressor (+ log target): predict nightly adr.
Feature Engineering	Lead-time, stay length, guest mix, room/meal type, distribution channel, seasonality flags.
Web Interface	Dash app for staff to input booking details, view: cancellation % , nightly ADR, total ADR, plus rule-based recommendations.
Deployment	Containerised via gunicorn; one-click deploy on Render; models & encoders loaded from artifacts/.

Cleaning steps: dropped sparse ‘company’, filled ‘country’ and ‘children’, trimmed ADR outliers 0–5000.

Machine Learning Pipelines

Step	Cancellation Model	ADR Model
Pre-processing	One-Hot encode 14 categorical cols, impute numeric means	Label-encode 8 categorical cols, log-transform adr
Algorithm	RandomForestClassifier (100 trees)	XGBRegressor (200 trees)
Metrics	Accuracy	R^2 & MAE
Scores	87 % accuracy	0.46 R^2, R 23.4 MAE
Artifacts	model_cancel.pkl	model_adr.pkl, adr_features.pkl, encoders.pkl

Feature importance exported to artifacts/feature_importance.csv.

Web Development (Dash)

Feature	Implementation
Form inputs	Check-in/out pickers with date-logic, room/meal dropdowns, guest counters
Predictions	Live call to both pickled models on button click
Total ADR	Calculates nightly_ADR × nights
Recommendations	Rule-based engine (flags high cancel risk, low ADR, etc.)
Code file	Src/app.py

Deployment Steps (Render)

1. Repo pushed to GitHub
2. Created a new Render Web Service
3. Environment spin-up installs Dash, scikit-learn, XGBoost.
4. On first boot, if models are absent train_models.py runs, then the Dash server starts.

Key Findings

Cancellation:

- Lead-time and previous cancellations are the strongest predictors.
- Resort-hotel bookings cancel 12 % more than city-hotel bookings during summer.

Pricing:

- Room-type and stay length explain > 40 % of ADR variance.
- Premium suites (code P) average 3× base ADR.

Challenges and Lessons Learned

Challenge	Lesson
Strict feature-order requirement in XGBoost	Solution: saved adr_features.pkl and enforced ordering in app
Encoding drift between train & inference	Persisted all LabelEncoders to encoders.pkl
Unbalanced room-type labels	Created user-friendly mapping and rare-type handling
Outliers skewing ADR	Log-transform + outlier trim improved R^2 from 0.20 → 0.46

Conclusion

The Hotelligence platform delivers a practical decision-support tool for hotel revenue managers. Cancellation risk is predicted with an accuracy of 87 percent, enabling proactive measures such as deposit enforcement or targeted reminders. ADR forecasting achieves a mean absolute error of approximately R 23, providing reliable nightly and total-stay revenue estimates.

The Dash interface integrates both predictions with actionable recommendations, thereby translating analytics into immediate operational value. Planned future work includes hyper-parameter optimisation, addition of holiday and event features, time-series cross-validation to respect booking chronology, and automated re-training pipelines to monitor and mitigate model drift in production.

Link to dash: https://hotelligence.onrender.com