HOTEL BOOKING CANCELLATION PREDICTION REPORT

Trendalytix - Team 5

Date of Submission: 30 June, 2025

S.no	Name	Role
1	Varshini Chilakala	Intern (TL)
2	Konda Sai Sreekar Reddy	Intern
3	Diti Solanki	Intern

1. Objective

To build a machine learning model that predicts hotel booking cancellations using structured booking data. The aim is to help hotel managers improve room allocation, reduce overbooking losses, and enhance customer satisfaction by identifying cancellation-prone reservations.

2. Dataset

- **Source:** Kaggle – Hotel Booking Demand Dataset
- **Size:** 119,390 rows, 32 columns
- **Important Features:** Hotel type, lead time, booking dates, number of guests, market segment, deposit type, special requests, assigned vs. reserved room type, cancellation status.

3. Tools & Technologies

- **Programming:** Python (Pandas, NumPy, Scikit-learn, XGBoost)
- **Visualization:** Power BI, Seaborn, Matplotlib
- **Notebook:** Jupyter, Google Colab
- **Version Control:** Git & GitHub

4. Process Overview

Problem Understanding

Cancellations disrupt hotel operations. We analyzed customer behavior to:

- Identify key factors causing cancellations
- Understand trends by customer type, time, and region
- Improve room planning and marketing

Data Preprocessing

- Handled missing values (e.g., filled children with mode, country as 'Unknown')
- Combined date fields into `arrival_date`
- Feature engineering:
    - `total_stay`, `total_guests`, `is_family`, `room_mismatch`, `is_repeated_customer`
- Handled outliers using capping

5. Exploratory Data Analysis (EDA)

Univariate Insights

- Most bookings came from **City Hotels**
- **Lead time** was highly skewed; long lead time correlated with cancellations
- "No Deposit" and "Online TA" bookings dominated

Bivariate Insights

- **City Hotels** and **Transient customers** showed higher cancellation rates
- **Room mismatch** increased chances of cancellation
- Seasonal trend: More cancellations in **summer** (Jul–Aug)
- Bookings with **children** had a higher cancellation rate

6. Model Development

Models Tested:

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	84%	0.76	0.65	0.70
Logistic Regression	74%	0.53	0.79	0.63
XGBoost (Best)	84%	0.74	0.67	0.71
Decision Tree	70%	0.48	0.89	0.62

Key Features (XGBoost):

1. Required car parking spaces
2. Room mismatch
3. Market segment – Online TA
4. Country – PRT
5. Deposit type – Non Refund
6. Number of special requests

7. Power BI Dashboard Insights

A dynamic Power BI dashboard was created for visual trend monitoring.

Key Visuals:

- **Booking Cancellations by Hotel Type:** City hotels faced more cancellations
- **Market Segment:** Online TA had the highest cancellations
- **Seasonality:** Summer months (Jul-Aug) had peak ADR & cancellations
- **Demographics:** Bookings with children were more prone to cancellations
- **KPIs:**
    - Total cancellations: ~22,000
    - Bookings from Portugal highest
    - Room mismatch and no-parking requests = higher risk

8. Results & Benefits

- **XGBoost model** achieved 84% accuracy with solid balance of precision and recall
- **Power BI dashboard** allowed real-time stakeholder insights

Business Impact:

- Revenue forecasting improvement
- Cancellation reduction through pre-emptive action
- Optimized room allocation & marketing strategy

9. Challenges & Future Work

Challenges:

- Imbalanced dataset (many more confirmed bookings than cancellations)
- Feature leakage from date fields
- Nonlinear dependencies required tree-based models

Future Work:

- Deploy model as a web app
- Integrate real-time bookings for live prediction
- Add customer reviews sentiment analysis

10. Deliverables

- Final Report (this document)
- Power BI Dashboard
- GitHub Repo: hotel_booking_cancellation_prediction_project
- Model & Code in Jupyter Notebook

End of Report

Team 5 – Trendalytix Internship, June 2025