HOTEL BOOKING CANCELLATION PREDICTION

Team Members:

- Malak Mohamed Abdelmoniem
- Farah Yehia Ismail Awaad
- Mariam Mohamed Ali El Nakeeb
- Malak Khaled Abdelaziz Hamed Ali
- Jana Mostafa Mohamed Sabry
- Youssef Mohamed Nabil

Project Summary

The Hotel Booking Cancellation Prediction project addresses one of the key challenges in the hospitality industry, the high rate of booking cancellations that impact revenue and operational planning. The main objective of this project is to develop a system capable of predicting the likelihood of a reservation being canceled before arrival, allowing hotels to make informed, proactive decisions. By understanding the factors that lead to cancellations, hotels can optimize room allocation, improve financial forecasting, and enhance customer satisfaction. This project contributes to building a smarter, data-informed hospitality environment where hotels can minimize losses, plan resources efficiently, and deliver a more reliable booking experience for their guests.

1. Project Planning

The Hotel Booking Cancellation Prediction Project follows a complete data science and machine learning lifecycle, designed to transform raw hotel booking data into actionable predictive insights. The workflow includes six main phases: data collection, exploratory data analysis (EDA), data preparation and preprocessing, modeling, evaluation, and maintenance.

1. Data Collection

The dataset was obtained from Kaggle using the kagglehub API. The data file — hotel_booking.csv — contains 119,390 records and 36 features, representing real-world hotel booking data from both City Hotels and Resort Hotels. Data attributes include guest demographics, booking details, stay duration, deposit type, and special requests. The dataset serves as the foundation for understanding cancellation patterns and training predictive models.

2. Exploratory Data Analysis (EDA)

The EDA phase was performed using **pandas**, **matplotlib**, and **seaborn** to uncover trends, correlations, and potential data issues. Numerical and categorical summaries were generated to understand the data distribution and detect anomalies. Key visualizations included:

- **Histograms** for lead time, ADR, and special requests distributions.
- **Count plots** for categorical variables such as hotel type, market segment, deposit type, and customer type.
- **Correlation heatmap** to analyze relationships among numerical variables.
- **Pie chart** showing class balance between canceled and non-canceled bookings.
  The EDA revealed that approximately **37% of bookings were canceled**, with higher cancellations linked to **No Deposit bookings**, **longer lead times**, and **Online Travel Agent segments**.

3. Data Preparation and Preprocessing

Data preprocessing ensured data quality and consistency prior to modeling:

- **Handling Missing Values:** Missing entries in `children`, `country`, `agent`, and `company` were treated using targeted imputation methods — mode, median, or replacing with zero when absence indicated "no record."
- **Removing Duplicates:** Over **8,000 duplicate rows** were detected and removed to maintain unique records.
- **Outlier Treatment:** Using the IQR method and logical thresholds, extreme values in lead_time, stays_in_week_nights, and adr were capped (winsorization) to prevent skewness.
- **Data Type Fixes:** The reservation_status_date column was converted to datetime for proper time-based analysis.
- **Feature Reduction:** Irrelevant and directly revealing columns such as name, email, phone-number, credit_card, and reservation_status were dropped to avoid data leakage.
- **Feature Engineering:** New features were created to enhance model performance:
  - total_guests = adults + children + babies
  - total_nights = week + weekend nights
  - is_family = 1 if children or babies > 0
  - stay_duration_category = Short / Medium / Long
  - season derived from arrival_date_month
- **Encoding:** Label Encoding was used for arrival_date_month, while **One-Hot Encoding** transformed nominal variables like hotel, meal, market_segment, deposit_type, customer_type, and season into numeric dummy variables.
- **Feature Scaling:** Numeric features were standardized using **StandardScaler** to ensure uniform range and improve model training stability.

4. Modeling

The processed dataset was divided into **training and testing sets (80:20 split)** using train_test_split with stratification to maintain class balance.
A **Random Forest Classifier** served as the baseline model due to its robustness against noise and ability to handle mixed feature types. The pipeline included:

- Model initialization with reproducible random state.
- Training on the scaled feature set (X_train, y_train).
- Automatic feature importance extraction for interpretability.

The Random Forest achieved **perfect performance (100%) on training data** and **~89% accuracy on test data**, indicating strong performance with slight overfitting.

5. Evaluation

Model performance was rigorously assessed using multiple metrics:

- **Accuracy, Precision, Recall, F1-score, and ROC-AUC** via sklearn.metrics.
- **Confusion Matrix** and **ROC Curve** visualizations provided interpretability of classification performance.
- **Classification reports** were generated for both training and test datasets.
  The evaluation confirmed that the model performed well on unseen data, achieving an F1-score of **85%** and ROC-AUC of **~0.90**, showing reliable discrimination between canceled and non-canceled bookings.

6. Maintenance and Future Improvements

Post-deployment, the model requires **continuous monitoring and retraining** as hotel booking trends evolve. Maintenance activities include:

- **Data Drift Detection:** Monitoring feature distributions (e.g., lead_time or deposit_type) to detect behavioral changes in customer booking habits.
- **Periodic Retraining:** Incorporating new booking records into the dataset to keep the model up-to-date.
- **Hyperparameter Tuning:** Experimenting with optimized Random Forest or advanced models like **XGBoost** and **LightGBM**.
- **Integration:** Deploying the model as a REST API or Streamlit dashboard for real-time hotel management use.

Timeline:

Phase	Tasks	Duration
1. Data Collection & Understanding	Import dataset, explore features	Week 1
2. Data Cleaning & Preprocessing	Handle missing values, outliers, encoding	Week 2
3. EDA & Feature Engineering	Visual analysis, create new features	Week 3
4. Model Training & Evaluation	Train Random Forest, evaluate	Week 4
5. Documentation & Presentation	Prepare final report & visual dashboard	Week 5

2. Stakeholder Analysis

The success of the project depends on clear identification of stakeholders and understanding their roles. Below is an analysis of each stakeholder group and their contribution to the project.

Key Stakeholder Need:

Hotel management needs a **real-time, interpretable prediction tool** to identify high-risk bookings and plan resource allocation accordingly.

Stakeholder	Role/Interest	Contribution	Impact
Hotel Management	Decision-makers seeking cancellation risk insights	Provides operational feedback	High
Revenue Managers	Optimize pricing and overbooking	Use model outputs for forecasting	High
Data Science Team (You)	Develop predictive model	Design and evaluate ML pipeline	High
IT / Database Admins	Maintain data infrastructure	Ensure data security and access	Medium
Guests / Customers	Indirect beneficiaries	Experience improved booking reliability	Low

3. Database Design

Data Source:

- Dataset: *Hotel Booking Demand Dataset* (Kaggle)
- Size: 119,390 records × 36 features
- Entities include **Hotels**, **Bookings**, **Guests**, and **Rooms**.

4. UI/UX Design

User Interface Concept

The UI focuses on **simplicity and interpretability** for hotel managers.
It allows users to input booking details and instantly view the predicted cancellation risk.

Core Screens:

1. **Dashboard Home** – Displays cancellation statistics, class balance, and key metrics.
2. **Prediction Form** – Input fields for lead time, deposit type, market segment, etc.
3. **Results Page** – Shows prediction ("Likely to Cancel" / "Not Likely"), confidence score, and key influencing factors.
4. **Insights Tab** – Interactive visualizations (histograms, feature importance plots, trend charts).

Design Principles

- **Minimalistic Layout:** Focus on core KPIs like cancellation rate, ADR, and lead time.
- **Color Coding:**
  - Green → Not Canceled
  - Red → High Cancellation Risk
- **Visualization:** Seaborn/Matplotlib charts integrated into a Streamlit dashboard.
- **Responsiveness:** Optimized for desktop and tablet use by hotel managers.

Tools and Technology:

Category	Tool / Library
Programming Language	Python
Development Environment	Google Colab
Data Manipulation	Pandas, NumPy
Data Visualization	Matplotlib, Seaborn
Machine Learning	Scikit-learn
Model Used	Random Forest Classifier
Dataset Source	KaggleHub