Hotel Booking Cancellation Prediction

This project focuses on predicting hotel booking cancellations using machine learning. Hotel cancellations create significant operational and financial challenges for hotels, including revenue loss, overbooking issues, staffing inefficiencies, and inaccurate demand forecasting. By developing an accurate prediction model, hotels can proactively manage bookings, apply smart overbooking strategies, and allocate resources more effectively. The project follows a complete end-to-end machine learning workflow, from data collection and cleaning to advanced preprocessing, feature engineering, model training, evaluation, and interpretation.

Dataset Overview The dataset contains 36 features and more than 119,000 booking records from a City Hotel and a Resort Hotel. These features describe booking details, guest characteristics, hotel operations, and historical behaviour.

Key Features Include:

Booking details: hotel type, lead time, arrival dates

Guest information: adults, children, babies, country

Stay characteristics: number of weekend and week nights

Booking behaviour: previous cancellations, booking changes, deposit type

Room features: reserved vs assigned room type

Market & distribution: channel, meal type, market segment

Target variable: is_canceled

These features allow us to study seasonality, customer behaviour, room availability, and operational conditions that influence cancellation likelihood.

Exploratory Data Analysis (EDA):

EDA was performed to understand the structure, patterns, and potential issues in the data.

Key Insights:

1. Higher lead times showed significantly higher cancellation rates.
2. City hotels had more cancellations than resort hotels.
3. Certain months and seasons (summer, winter holidays) had peaks in cancellation activity.
4. Bookings from certain countries had unusually high cancellation patterns.
5. Longer stays and bookings with children showed unique trends in cancellation likelihood.

EDA helped guide feature engineering, data cleaning, and modelling decisions.

Data Cleaning & Preprocessing:

Thorough data cleaning phase ensured data consistency and reliability.

Main Cleaning Actions:

1. Removed duplicate rows to prevent biased learning.
2. Fixed invalid or impossible values (e.g., negative guests or nights).
3. Standardized date formats, converting reservation_status_date to datetime.
4. Handled missing values:
5. Numerical values → replaced with median
6. Categorical values → replaced with mode

Removed leaking features and irrelevant personal information (name, email, phone number, credit card, and features directly revealing cancellation outcomes).

This ensured the dataset was clean, accurate, and safe for modelling.

Outlier Detection Outliers

The chosen method is **capping**, where it was selected because it reduces the influence of extreme values without removing observations, preserving data size for modelling.

1. **Lead Time:** Capped at **365** → removed unrealistic booking times beyond a year.
2. **Stays in Weekend Nights:** Capped at **10** → prevented extreme rare weekend stays.
3. **Stays in Week Nights:** Capped at **30** → avoided very long stays that distort averages.
4. **ADR (Average Daily Rate):** Cleaned → negative values set to 0, extreme values capped at 1000.

5. **Adults:** Column fixed → bookings with 0 adults replaced with 1 (ensuring realistic records).

The dataset is now free from **unrealistic extreme values**, making it **more reliable for analysis and modelling**

Feature Engineering

Feature engineering was a crucial part of enhancing model performance.

1. The following meaningful features were created:
2. Total Guests = adults + children + babies
3. Total Nights = week nights + weekend nights
4. Is Family → 1 if children or babies > 0
5. Stay Duration Category → Short / Medium / Long
6. Season → derived from arrival month (Winter, Spring, Summer, Autumn)

These new features capture patterns related to guest type, travel purpose, seasonality, and booking behaviour, adding significant value beyond the raw dataset.

Encoding Categorical Variables

All categorical variables were transformed into numerical values so the model could understand them.

Encoding Steps:

1. Ordinal Encoding for ordered variables (arrival month → 1–12)
2. One-Hot Encoding for nominal variables such as: hotel, meal, market segment, distribution channel, room type, deposit type, customer type, season, stay duration category, and country
3. Boolean features were converted into 0/1 integers

Encoding ensured all features were machine-readable while preserving their meaning.

Handling Class Imbalance

The target variable (is_canceled) was imbalanced. To solve this, **SMOTE (Synthetic Minority Oversampling Technique)** was used to generate synthetic examples for the minority class. This improved model fairness and prevented bias toward the non-cancelled category.

Although SMOTE was applied and tested, the best-performing model (XGBoost) worked optimally on the original unbalanced dataset, so sampling was not used in the final version.

Train–Test Split: The data was split into an 80% training set and 20% testing set

This allowed for robust evaluation of model generalization.

Model Training

Multiple machine learning models were trained and compared, including:

Logistic Regression, Decision Tree, Random Forest, Gradient Boosting / XGBoost ,Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)

Hyperparameters were tuned, and performance differences were analysed. The objective was to find the model with the best balance of accuracy, precision, recall, F1-score, and AUC.

Model Evaluation

Evaluation metrics included: Accuracy, Precision, Recall, F1-score, ROC-AAUC, and Confusion Matrix

These metrics offered a complete understanding of how well the model predicted cancellations, especially the ability to detect cancelled bookings without too many false alarms.

Final Model Selection: XGBoost

After evaluating multiple machine learning models under different sampling strategies, **XGBoost trained on the original (unbalanced) dataset** was selected as the **final and optimal model** for deployment as it provides:

1. **Highest Overall Performance (Best F1-Score)**
    a. The model achieved the **top F1-Score (81.75%)**, indicating the best balance between Precision and Recall.
    b. This makes it the most reliable model for maintaining both correctness and sensitivity.
2. **Strong Generalization Ability**

   a. XGBoost consistently delivered high **Accuracy (85.84%)** and **AUC (94.32%)** without requiring any resampling.
   b. This reflects strong learning capability even under class imbalance.

3. **No Need for Sampling**
   a. Tree-based models like XGBoost naturally handle class imbalance due to their splitting strategy.
   b. Using the original dataset avoids the potential noise introduced by oversampling or information loss from undersampling.
4. **Most Stable and Interpretable Results**
   a. Results were more stable compared to SMOTE, Undersampling, or SMOTE+Tomek.
   b. This stability makes XGBoost more reliable for real production settings.

Conclusion

Overall, the project demonstrates a complete end-to-end machine learning pipeline, from raw data to final model selection. The final XGBoost model provides a highly reliable and practical solution for predicting hotel booking cancellations.