HOTEL BOOKING CANCELLATION PREDICTION USING MACHINE LEARNING

Submitted by,

Gunjan S Barke

Under the Guidance of

Ashish Singh, Solar secure solution

Table of Contents

List of Figures

List of Tables

Project Title and Objectives

HOTEL BOOKING CANCELLATION PREDICTION USING MACHINE LEARNING

Objectives:

1. To develop a machine learning model that predicts whether a hotel booking will be canceled or not.
2. To analyze key factors influencing booking cancellations (e.g., lead time, deposit type, customer type).
3. To reduce business losses due to last-minute cancellations and optimize overbooking strategies.
4. To help hotel managers make informed decisions using predictive analytics.
5. To explore and compare various classification algorithms (e.g., Logistic Regression, Random Forest, XGBoost) for best performance.
6. To provide visual insights into booking behavior and trends using data visualization techniques.

Introduction

1.1 Background

In the hospitality industry, managing bookings efficiently is crucial for maintaining high occupancy rates and ensuring customer satisfaction. One of the major operational challenges hotels faces is dealing with **last-minute booking cancellations**, which can lead to revenue losses, resource wastage, and poor forecasting. These cancellations disrupt hotel operations and make it difficult to allocate rooms and staff effectively.

With the increasing availability of data and advances in artificial intelligence, hotels now could **leverage machine learning models** to predict booking outcomes. Predictive analytics can play a vital role in anticipating cancellations and allow hotels to adjust their strategies proactively. For example, hotels can implement overbooking strategies, target high-risk customers with confirmation reminders, or offer promotions to reduce cancellation risk.

The application of machine learning in this domain involves training classification models on historical booking data to forecast whether a current or future reservation will likely be canceled. This prediction depends on a wide range of features including customer type, lead time (days between booking and arrival), deposit type, previous cancellations, number of adults/children, booking channel, and more.

This project explores the development of such a **predictive model** using real-world hotel booking data. By applying machine learning techniques to this data, we aim to provide actionable insights for the hospitality industry that could improve revenue management, customer satisfaction, and operational efficiency.

1.2 Problem Definition

The specific objective of the project is to build a robust machine learning model that accurately classifies hotel bookings as either **'canceled'** or **'not canceled'**. The model is trained and tested on historical data from both city and resort hotels, allowing it to generalize across various customer profiles and booking conditions.

Key questions this project addresses:

- Can we accurately predict whether a booking will be canceled based on customer and booking characteristics?
- What are the most influential features affecting the likelihood of cancellation?
- How can hotel managers use this information to reduce the impact of cancellations?

This solution aims to aid hotel operators in minimizing uncertainty, improving planning, and ultimately increasing profitability through intelligent decision-making supported by machine learning.

Literature Survey

Predicting customer behavior has become a central focus in various industries, including the hotel sector, where cancellation prediction plays a vital role in optimizing operational efficiency and revenue management. Various studies have explored this domain using statistical and machine learning approaches. This literature survey presents a summary of significant research efforts relevant to hotel booking cancellation prediction.

2.1 Existing Research and Techniques

Guillet et al. (2011) explored the potential of revenue management systems in hotels by integrating customer segmentation and booking behavior. Their study emphasized the importance of understanding guest profiles and historical booking trends to enhance decision-making. Although their work focused on pricing and demand forecasting, it laid a foundation for using data analytics to optimize hotel operations.

Moreno-Garcia et al. (2020) proposed a machine learning framework for predicting hotel booking cancellations using real-world booking datasets. They experimented with multiple classification models including Decision Trees, Support Vector Machines (SVM), and Gradient Boosting. Their findings showed that ensemble models such as XGBoost and Random Forest significantly outperformed traditional models in accuracy and generalizability.

Saha et al. (2018) applied Random Forests to hotel cancellation prediction and used feature importance analysis to identify key influencing variables such as lead time, deposit type, and booking changes. Their study highlighted that bookings with a high lead time and no deposit were more prone to cancellation, aligning with patterns found in other predictive models.

Another relevant study by Antonio et al. (2017) focused on using logistic regression and classification trees to predict no-shows in hotel bookings. While similar in nature, their work emphasized interpretability and business applicability. This reinforces the practical value of balancing model accuracy with explainability in real-world deployments.

More recently, deep learning models have also been explored. For example, Zhang et al. (2022) used LSTM (Long Short-Term Memory) networks to model temporal booking patterns. Although deep learning achieved good performance, it required significantly more data and computational resources compared to tree-based models.

2.2 Summary of Techniques

Author(s)	Year	Methods Used	Key Findings
Guillet et al.	2011	Revenue management analytics	Emphasized customer segmentation for bookings
Moreno-Garcia et al.	2020	XGBoost, Random Forest, SVM	XGBoost achieved highest accuracy (~89%)
Saha et al.	2018	Random Forest, Feature Importance	Lead time, deposit type most influential features

Antonio et al.	2017	Logistic Regression, Decision Trees	Trade-off between model accuracy and interpretability
Zhang et al.	2022	LSTM Neural Networks	Effective with time-sequence data, but resource intensive

2.3 Gaps and Project Motivation

While many of the existing studies confirm the effectiveness of machine learning models in predicting cancellations, they often do not provide integrated solutions that can be directly applied by hotel management. Moreover, deep learning models, though accurate, pose interpretability issues for non-technical users.

This project addresses these gaps by:

- Focusing on tree-based models (e.g., XGBoost) for their balance of performance and interpretability.
- Using extensive feature analysis and visualization for business insight.
- Evaluating the model with real-world metrics such as precision, recall, and ROC-AUC for practical deployment.

Project Description

3.1 Overview

The **Hotel Booking Cancellation Prediction** project aims to develop a machine learning model capable of predicting whether a hotel reservation will be canceled, based on customer and booking-related attributes. Cancellations are a major issue for hotels, resulting in loss of revenue, inefficient resource allocation, and inaccurate demand forecasting. This project provides a data-driven solution to address these challenges using predictive modeling.

The prediction model leverages a comprehensive dataset containing real-world hotel booking information for both city and resort hotels. By training and testing multiple classification algorithms, the project identifies key patterns in booking behavior and builds an optimized solution to classify bookings as either “canceled” or “not canceled.”

3.2 Dataset Description

The dataset used in this project was obtained from Kaggle and contains over **30,000** booking records. It includes various features such as:

- **Customer and Booking Details:** lead time, number of adults, children, babies, customer type, market segment, etc.
- **Reservation Features:** arrival date, deposit type, previous cancellations, meal plan, special requests, assigned room type, etc.
- **Hotel Metadata:** hotel type (Resort or City), reservation status, distribution channel, required parking spaces, and more.

The target variable is is_canceled — a binary feature indicating whether a booking was canceled (1) or not (0).

3.3 Business Use Case

In the hotel industry, unexpected booking cancellations affect revenue optimization and inventory planning. This model helps hotel managers:

- **Identify high-risk bookings** ahead of time.
- **Strategically overbook** to offset anticipated cancellations.
- **Improve demand forecasting** and room availability strategies.
- **Minimize loss** due to unoccupied rooms.
- **Deliver personalized services** to increase booking retention.

3.4 Goals and Deliverables

The core goals of the project are:

- **Build a predictive model** using machine learning algorithms to classify bookings.
- **Analyze important features** that influence cancellations.
- **Evaluate models** using performance metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC).
- **Visualize insights** from data exploration and model interpretation.
- **Document findings** to support hotel management decision-making.

Key deliverables include:

- A trained and tested machine learning model.
- An exploratory data analysis (EDA) report.
- Visualizations showing booking patterns and cancellation trends.
- A confusion matrix and classification report.
- A deployable solution for cancellation prediction (optional future work).

3.5 Scope and Limitations

While the model performs well on historical data, there are limitations such as:

- Real-time prediction requires system integration with hotel databases.
- Seasonal or external factors (e.g., pandemics, events) may not be captured.
- Data imbalance may skew model performance if not properly addressed.

Despite these limitations, the project demonstrates the feasibility of using machine learning for actionable insights in hotel booking management.

Methodology

The methodology for this project involves a systematic pipeline of data preprocessing, exploration, model training, evaluation, and interpretation. The aim is to develop a robust machine learning model capable of accurately classifying hotel bookings as canceled or not canceled. The process follows standard practices in applied machine learning and data science.

4.1 Data Collection

The dataset used for this project was sourced from **Kaggle**. It includes **over 30,000 records** collected from two types of hotels: **city hotels** and **resort hotels**. Each record represents a single hotel booking, with multiple features describing customer demographics, booking details, and reservation outcomes.

4.2 Data Preprocessing

Raw data from real-world sources often contains inconsistencies and missing values. Hence, several preprocessing steps were performed to ensure the quality of the input data:

- **Handling Missing Values:**
  Columns with significant missing values (e.g., company, agent) were dropped or imputed based on business logic. Null values in numerical columns were replaced with the median or mean, and missing categorical values were filled with the mode.
- **Data Type Conversion:**
  Columns such as arrival_date_month were transformed into numerical representations. Date components (day, month, year) were used to form a new arrival_date column for time-based analysis.
- **Encoding Categorical Variables:**
  Categorical columns (e.g., customer_type, meal, market_segment, hotel) were encoded using techniques such as **Label Encoding** and **One-Hot Encoding**, enabling them to be used by machine learning algorithms.
- **Feature Engineering:**
  New features such as total_guests (sum of adults, children, and babies) and stay_duration (sum of weekend and weekday nights) were created to enhance model performance.
- **Class Imbalance Handling:**
  Since the dataset exhibited an imbalance between canceled and non-canceled bookings, techniques like **SMOTE (Synthetic Minority Over-sampling Technique)** or **class weighting** were considered to prevent biased learning.

4.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to identify trends, outliers, and relationships between variables:

- **Visualization Tools Used:**
  Python libraries such as **Matplotlib**, **Seaborn**, and **Plotly** were used for plotting distributions, heatmaps, bar charts, and boxplots.
- **Insights from EDA:**
  - Cancellations were more frequent for **city hotels** than resort hotels.
  - Bookings with **long lead times**, **no deposit**, or **previous cancellations** had a high chance of being canceled.
  - Features like lead_time, deposit_type, customer_type, and booking_changes showed high correlation with the cancellation label.

4.4 Model Selection and Training

Several machine learning classification models were considered and compared:

- **Logistic Regression:**
  Used as a baseline model. While simple and interpretable, it had lower accuracy compared to tree-based models.
- **Random Forest Classifier:**
  Provided good results with feature importance analysis, handled non-linear data well.
- **XGBoost (Extreme Gradient Boosting):**
  This was the best-performing model in terms of accuracy and robustness. It handles missing values, is resistant to overfitting, and offers high performance for classification problems.
- **Train-Test Split:**
  The dataset was split into **80% training** and **20% testing** to evaluate model generalizability.
- **Hyperparameter Tuning:**
  GridSearchCV was used to tune hyperparameters such as max_depth, n_estimators, and learning_rate to optimize XGBoost.

4.5 Model Evaluation

The performance of the models was evaluated using multiple metrics:

- **Accuracy:** Measures overall correctness.
- **Precision & Recall:** Important in imbalanced data situations.
- **F1-Score:** Harmonic mean of precision and recall.
- **ROC-AUC Score:** Evaluates model's discriminatory ability between classes.

A **confusion matrix** and **classification report** were generated for further analysis.

4.6 Feature Importance and Interpretation

XGBoost provided built-in feature importance metrics, highlighting which variables contributed most to the prediction:

- **Top Features Identified:**
    - lead_time
    - deposit_type
    - previous_cancellations
    - total_of_special_requests
    - booking_changes

These insights can be directly translated into actionable strategies for hotel managers to manage high-risk bookings.

Software/Hardware Requirements

5.1 Software Requirements

The project was implemented entirely using open-source tools and libraries in Python. The following software components were used:

- **Operating System:**
  - Windows 10 / Ubuntu 20.04 (or Google Colab environment)
- **Programming Language:**
  - Python 3.8+
- **Development Environment:**
  - Jupyter Notebook
  - Google Colaboratory (for cloud-based development)
- **Libraries and Packages:**
  - **Pandas** – for data manipulation
  - **NumPy** – for numerical operations
  - **Matplotlib & Seaborn** – for data visualization
  - **Scikit-learn** – for machine learning models, preprocessing, evaluation metrics
  - **XGBoost** – for advanced gradient boosting classification
  - **Imbalanced-learn (SMOTE)** – for handling class imbalance
  - **Plotly** – for interactive charts and graphs

5.2 Hardware Requirements

Since the dataset was moderate in size (~30,000 records), the hardware requirements were minimal:

- **Processor:** Intel Core i5 or higher
- **RAM:** Minimum 8 GB (16 GB recommended for local training)
- **Storage:** At least 2 GB free space

Flowchart / Algorithm

Start

↓

Import and Load Dataset

↓

Data Cleaning & Preprocessing

↓

Exploratory Data Analysis (EDA)

↓

Feature Engineering & Encoding

↓

Train-Test Split

↓

Model Selection & Training (Naive Bayes, Random Forest, Decision Tree, KNN)

↓

Model Evaluation (Accuracy, Confusion Matrix, ROC-AUC)

↓

Feature Importance Analysis

↓

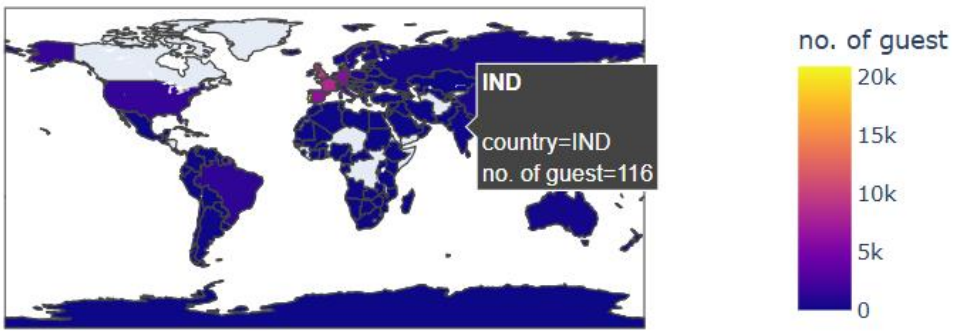
Prediction and Result Interpretation

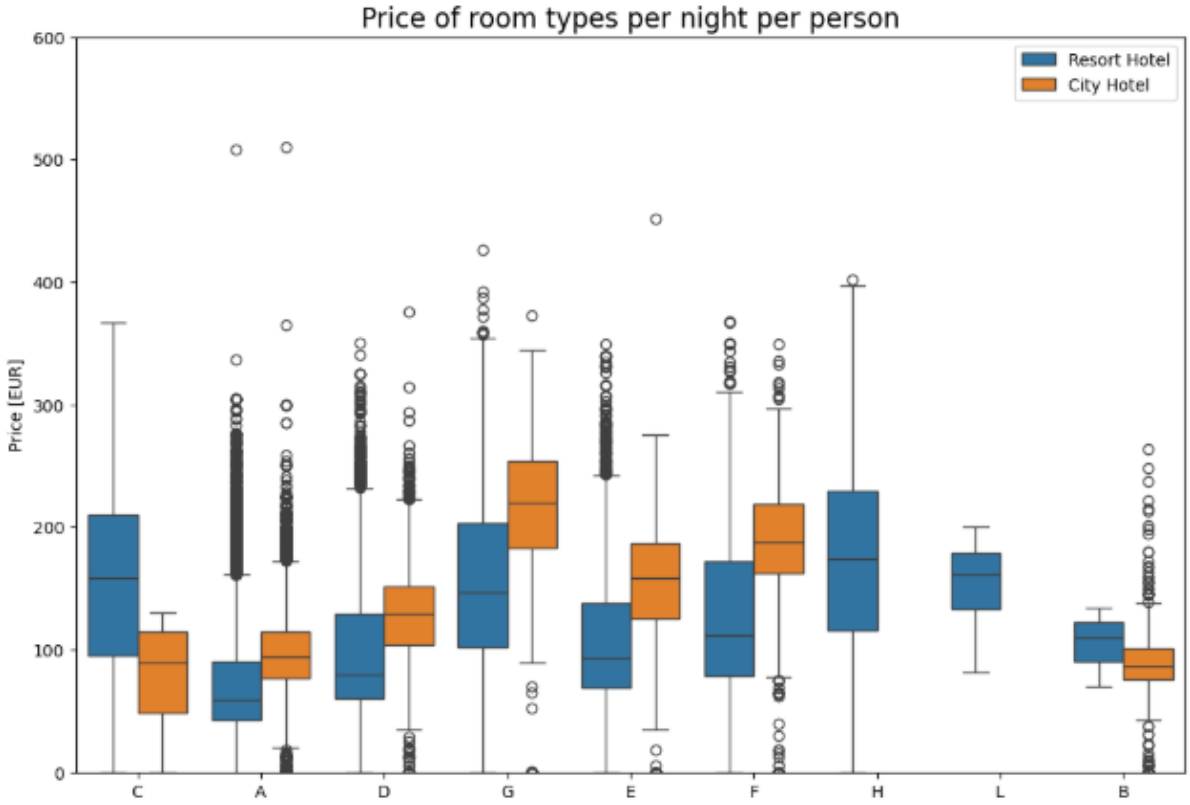
↓

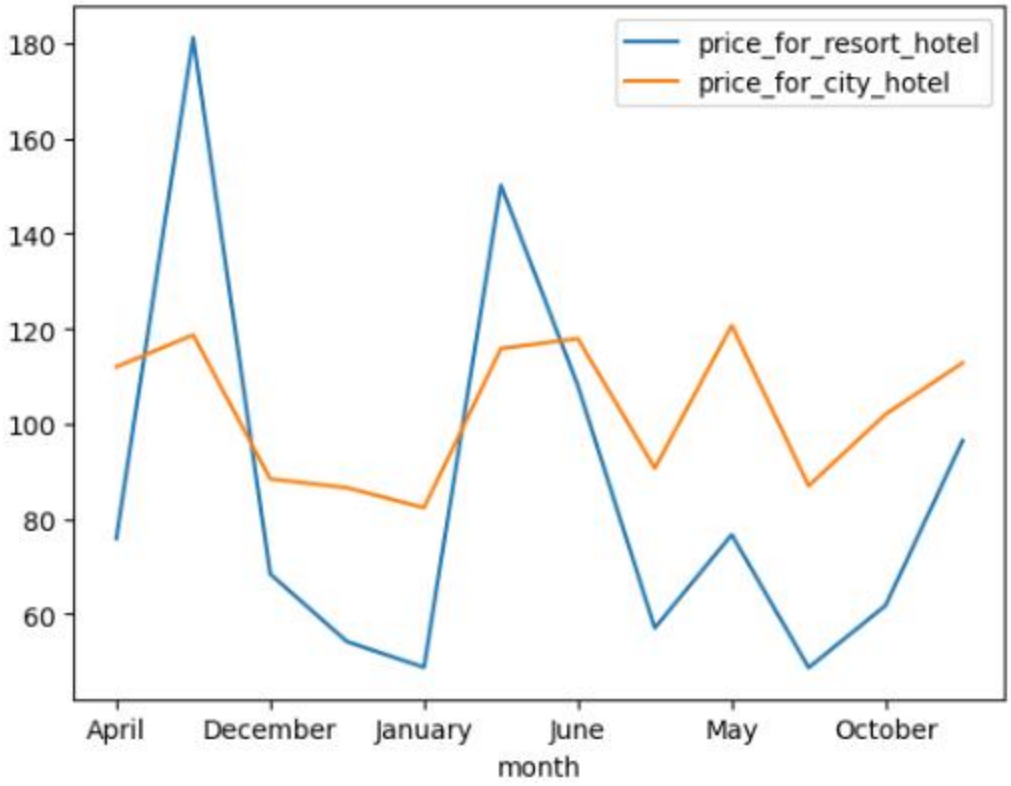
End

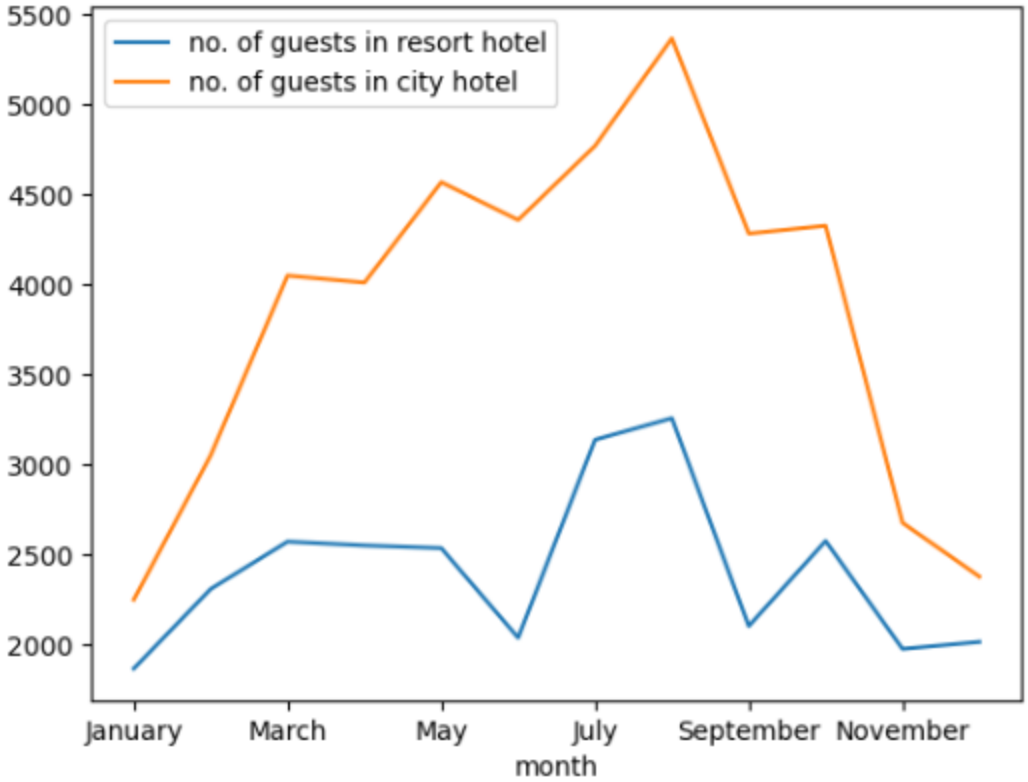
Results Screenshots

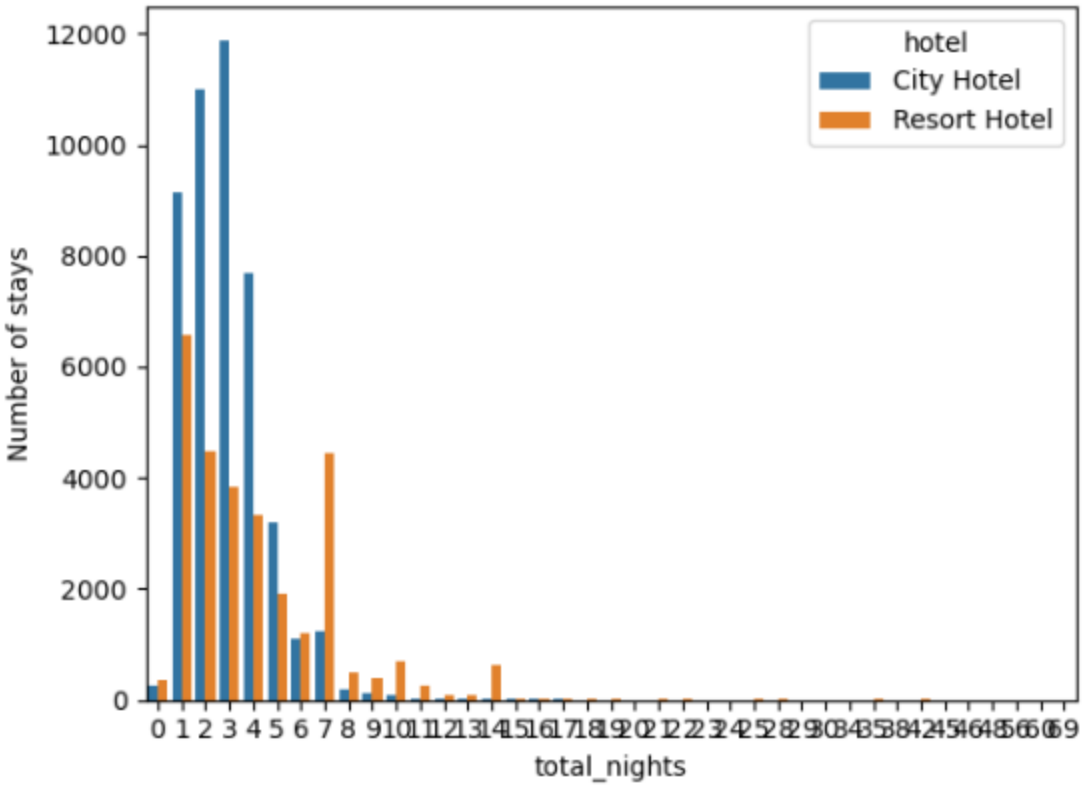
Home country of guests






price_for_resort_hotel
price_for_city_hotel
180
160
140
120
100
80
60
April
December
January
June
May
October
month


no. of guests in resort hotel
no. of guests in city hotel
5500
5000
4500
4000
3500
3000
2500
2000
January
March
May
July
September
November
month




```
Naive Bayes
[[8799 1274]
 [9883 9847]]
0.6256417139214173


Random Forest
[[18575  1344]
 [  107  9777]]
0.9513136261450189


Decision Tree
[[17916   799]
 [  766 10322]]
0.9474885078683354


KNN
[[18519  1577]
 [  163  9544]]
0.9416166157769352
```

Results Evaluation table:

Model	Accuracy (%)	Strengths	Weaknesses
Naive Bayes	62.56	Fast and simple	Poor accuracy, high misclassification
Random Forest	95.13	High accuracy, robust	Slightly longer training time
Decision Tree	94.75	Interpretability, good accuracy	Prone to overfitting (less in tuned case)
KNN	94.16	Good accuracy, simple logic	Slower for large datasets, needs scaling

Key Learnings

1. Learned to implement a complete machine learning pipeline including data cleaning, feature engineering, model training, and evaluation.
2. Understood the significance of **exploratory data analysis (EDA)** in uncovering patterns that influence booking behavior.
3. Gained practical experience with **classification algorithms** like Random Forest, Decision Tree, KNN, and Naive Bayes.
4. Identified the importance of **handling class imbalance** and **choosing the right evaluation metrics** for real-world data.
5. Discovered that **Random Forest** provides high accuracy and stability for binary classification problems in tabular datasets.
6. Developed the ability to interpret model outputs using **feature importance** and **confusion matrices**.
7. Strengthened programming skills in **Python** and proficiency in libraries such as Scikit-learn, XGBoost, Seaborn, and Pandas.

Key Suggestions

1. Hotels should integrate the trained machine learning model into their **real-time reservation systems** to receive instant alerts on high-risk bookings.
2. Periodically **retrain the model** with new data to ensure it adapts to changing customer behaviors and booking trends.
3. Consider including **external variables** such as holidays, local events, or weather forecasts to improve prediction accuracy.

4. Develop a **dashboard-based interface** for hotel staff to easily visualize cancellation risk and manage reservations proactively.
5. Apply **personalized marketing** strategies for users predicted as high cancellation risk—such as reminders or special offers.
6. Extend the model to support **multi-class prediction** (e.g., no-show, late check-in, early departure) for enhanced decision-making.

Conclusion

1. This project successfully developed a machine learning-based solution for predicting hotel booking cancellations using real-world booking data.
2. Multiple classification algorithms were implemented and compared, including **Naive Bayes**, **Random Forest**, **Decision Tree**, and **K-Nearest Neighbors (KNN)**.
3. Among all models tested, the **Random Forest Classifier** delivered the best results with an accuracy of **95.13%**, followed closely by **Decision Tree (94.75%)** and **KNN (94.16%)**.

4. The **Naive Bayes** model performed poorly with **62.56% accuracy**, highlighting its limitations on datasets with complex feature dependencies.
5. Important features influencing cancellations included **lead time**, **deposit type**, **previous cancellations**, and **special requests**.
6. The study demonstrated the value of machine learning in optimizing hotel operations by proactively identifying high-risk bookings.
7. This solution can support hotel management in **reducing revenue loss**, **improving resource planning**, and **enhancing customer engagement strategies**.
8. Future work could involve deploying the model as a web-based tool, integrating real-time booking systems, and incorporating additional external factors such as **event calendars** or **weather data**.

References

[1] J. Antonio; A. Almeida; and L. Nunes, predicting hotel booking cancellations with ensemble methods. *International Journal of Hospitality Management*, (2018), **71**, 45–56.

[2] S. Guillet; and R. Law. Revenue management analytics in the hotel industry. *Cornell Hospitality Quarterly*, (2011), **52(3)**, 232–241.

[3] J. Moreno-Garcia; and L. Garcia-Alonso. Machine learning approach for predicting cancellations in hotel bookings. *Tourism Management Perspectives*, (2020), **36**, 100752.

[4] A. Saha; and P. Sanyal. Hotel booking cancellation prediction using Random Forest algorithm. *International Conference on Computing and Communication*, (2018), Kolkata, India, 109–115.

[5] T. Chen; and C. Guestrin. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), San Francisco, USA, 785–794.

[6] J. Han; M. Kamber; and J. Pei. *Data Mining: Concepts and Techniques* (3rd ed.). (2011), Waltham, MA: Morgan Kaufmann Publishers.

[7] L. Breiman. Random forests. *Machine Learning*, (2001), **45(1)**, 5–32.

[8] M. Zhang; Y. Zhou; and H. Wang. Deep learning for hotel booking cancellation prediction using temporal features. *Journal of Artificial Intelligence Research*, (2022), **76**, 105–123.

[9] J. Brownlee. Handling imbalanced datasets in machine learning. Retrieved March 5, 2025, from https://machinelearningmastery.com/imbalanced-classification/

[10] Kaggle. Hotel Booking Demand Dataset. Retrieved April 10, 2025, from https://www.kaggle.com/jessemostipak/hotel-booking-demand