Predicting Hotel Reservation Cancellations: Turning Data into Revenue

This presentation explores how predictive analytics can help hotels minimize cancellations and boost revenue. We will delve into the cancellation landscape, key data insights, and practical strategies to optimize hotel operations.

by Rishita Shah

Introduction :

1 Objectives :

The primary objective of this project is to develop robust machine learning models that can accurately predict hotel reservation cancellations.

- Enhance revenue management
- Improve resource Allocation
- Understand customer Behavior

2 Dataset :

The dataset contains various features that provide insights into booking details, guest behavior, and factors influencing hotel reservation cancellations.

- Total rows : 36275
- Total columns : 19
- Target variable : `booking_status` (0 = Not Cancelled, 1 = Cancelled)

3 Importance :

Accurately predicting hotel reservation cancellations is crucial for improving operational efficiency, maximizing revenue, and enhancing customer satisfaction. This project holds significant importance in several key areas .

- Revenue Optimization
- Resource management
- Customer insights

Made with Gamma

Here is a data dictionary summarizing the key columns in the dataset :

Column Name	Description
Booking_ID	Unique identifier for each booking
no_of_adults	Number of adults
no_of_children	Number of children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday)
no_of_week_nights	Number of weeknights (Monday to Friday)
meal_type	Meal type booked by the customer
required_car_parking_spaces	Does the customer require a car parking space? (0 - No, 1 - Yes)
lead_time	Number of days between the booking date and arrival date
arrival_year	Year of arrival
arrival_month	Month of arrival
arrival_date	Date of arrival
market_segment	Market segment designation
repeated_guest	Is the customer a repeated guest? (0 - No, 1 - Yes)
no_previous_cancellations	Number of previous bookings canceled by the customer prior to the current booking
previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
avg_price_per_room	Average price per day of the reservation (in euros)
no_of_special_requests	Total number of special requests made by the customer (e.g., high floor, view from the room, etc)
booking_status	Flag indicating if the booking was canceled or not

Made with Gamma

Steps :

Made with Gamma

EDA (Exploratory Data Analysis)

Descriptive Statistics

This step involves calculating basic statistics like mean , median , standard deviation of key features , to understand data characteristics .

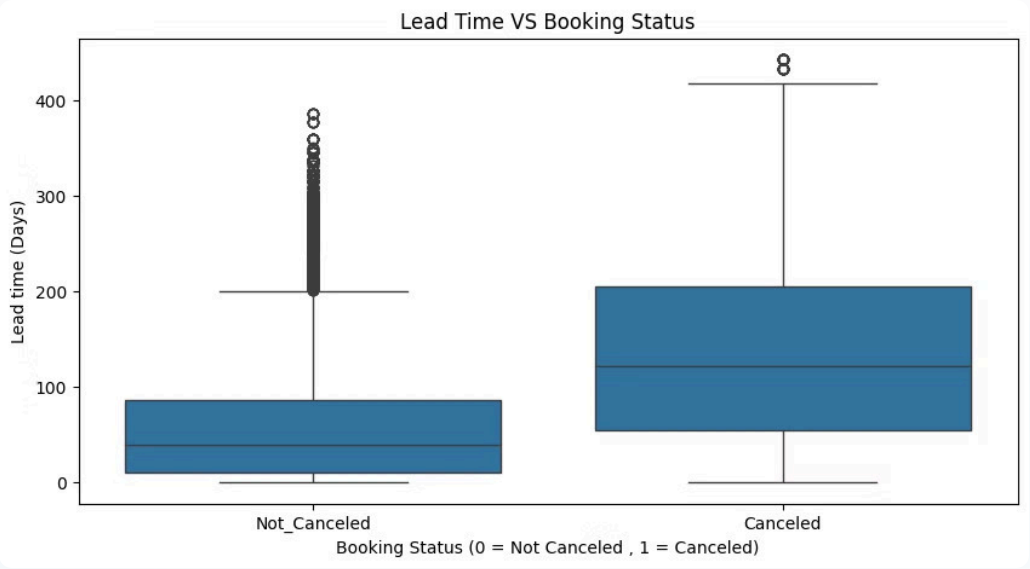
Correlation Analysis

Examining correlations between features reveals potential relationships & helps identify which features are most relevant for segmentation .

Visualizations

Creating visualizations like Heatmaps , Pair-plots , Scatterplots & Boxplots helps understand data patterns & outliers in the data .

Made with Gamma



Boxplot of Lead Time vs Booking Status :

1️⃣ Not-Canceled (0)

2️⃣ Canceled (1)

- **Non-canceled Bookings** : Most bookings have lead times below **100 days**, with some extending to around **200 days**.
- Outliers above **200 days** indicate rare cases with unusually long lead times.

- **Canceled Bookings** : A wider spread in lead times, ranging up to **400+ days**, suggests cancellations are more common with distant bookings.
- Outliers are present but less concentrated than in the "Not-Canceled" group.

Made with Gamma

Count-plot of Booking Timing Impact :

1️⃣ Not - Canceled (Blue)

2️⃣ Canceled (Orange)

- **Seasonal Trends -**

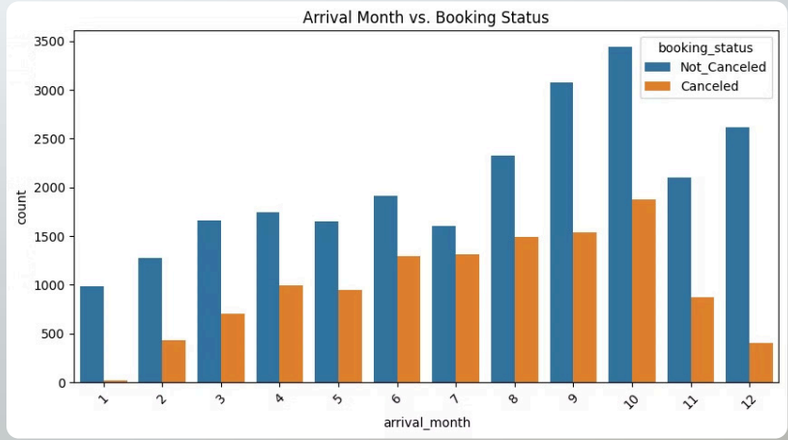
Peak Months : Months **9 (September)** and **10 (October)** show the highest number of total bookings.

Low Months : Months like **1 (January)** and **2 (February)** have the fewest bookings, indicating a possible off-season period.

- **Cancellation Patterns -**

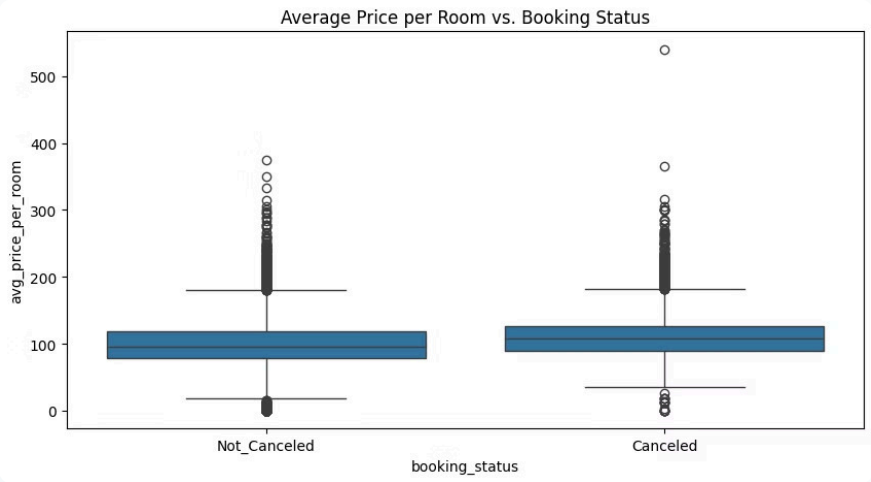
Higher Cancellations in Peak Seasons : Noticeable increase in cancellations during the busiest months (**September** and **October**). This trend may be influenced by overbookings, bulk reservations, or holiday rush.

Lower Cancellations in Early Months : Months like **January** and **February** have fewer cancellations, possibly reflecting lower booking volumes or more committed travelers.



Made with Gamma

Boxplot of Price & Cancellation Relationship



1️⃣ Price Distribution :

- Both groups show a **similar median price** range of approximately **100-120** units.
- This suggests that **price alone may not be a strong differentiator** between canceled and non-canceled bookings.

2️⃣ Outliers :

- Both groups exhibit significant outliers, particularly at the **higher price range**.
- The **Canceled** category shows a few extreme values exceeding **500 units**, indicating that **high-priced bookings are more prone to cancellations**.

Made with Gamma

Data Cleaning & Outlier Handling :

- Effective data cleaning and outlier handling are crucial steps to ensure your dataset is accurate, consistent, and optimized for machine learning models .

Outlier Detection & Handling :

✅ **Visual Detection :** Use **boxplots** or **Histograms** to identify extreme values.

✅ **Statistical Detection :** Use the **IQR (Interquartile Range)** method.

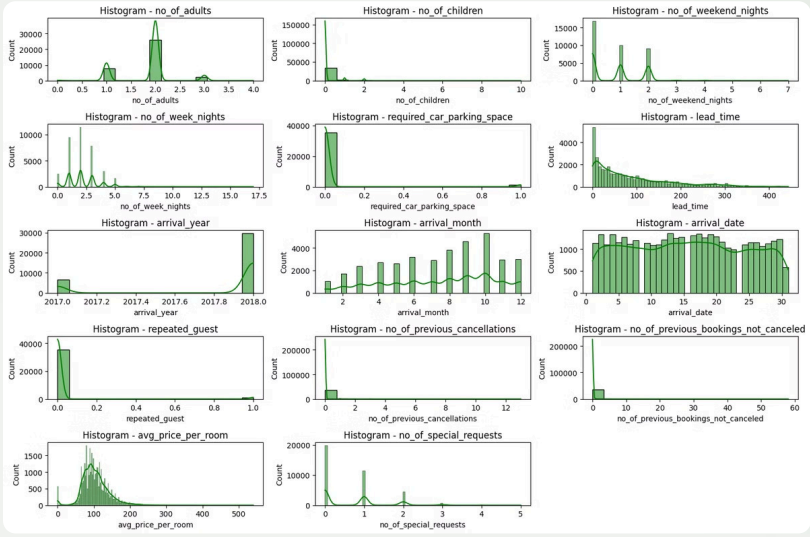
Made with Gamma

Outlier Detection :

1 Histogram for Distribution Analysis :

This image presents multiple histograms visualizing the distribution of various features from a hotel reservation dataset

- Outliers : columns like Lead_time , no_of_previous_bookimgs_not_canceled , & avg_price_per_room show potential outliers .
- Most bookings have zero or one special request , with a few having multiple .
- Majority have no previous non-canceled bookings , with rare occurrences of higher values .



Made with Gamma

Introduction to Model Building :

Objective :

To develop predictive models that can accurately classify customer segments.

Process Overview :

- Data Preparation
- Model Evaluation
- Training
- Evaluation
- Comparison

Model Implemented For Prediction :

Decision Tree Classifier .

- A **Decision Tree Classifier** is a supervised machine learning algorithm used for classification tasks. It is based on a tree-like structure where data is split into branches based on feature conditions until a decision (or class label) is reached.

Random Forest Classifier .

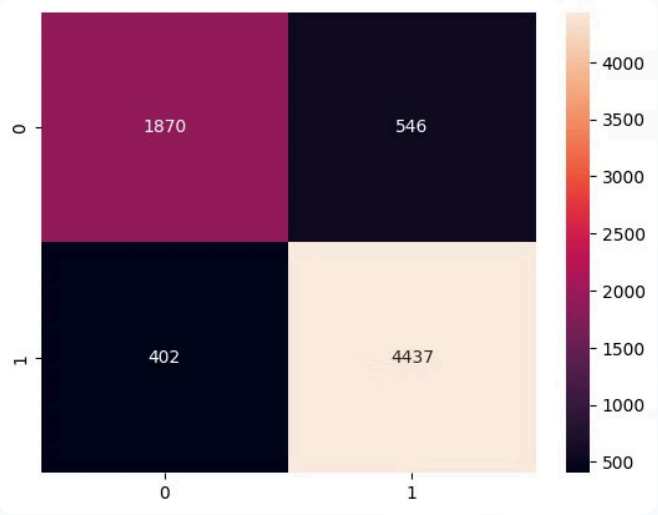
- The **Random Forest Classifier** is an ensemble machine learning algorithm that combines multiple decision trees to improve accuracy and reduce overfitting. It is widely used for classification tasks due to its robustness and versatility.

Logistic Regression .

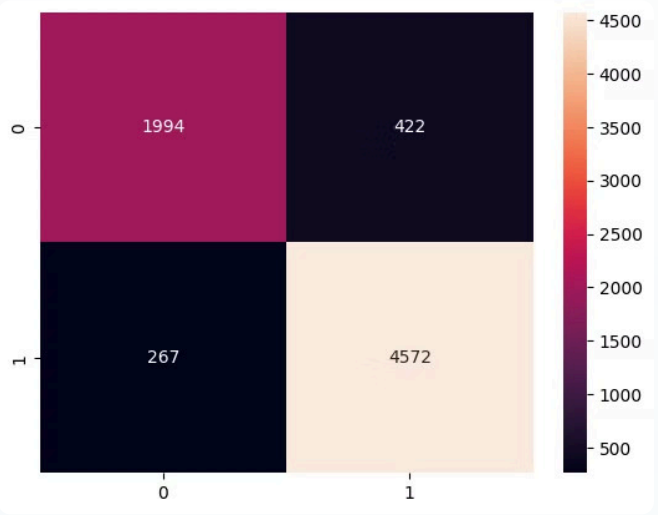
- **Logistic Regression** is a supervised machine learning algorithm used for **binary** and **multiclass classification** tasks. It predicts the probability that a given input belongs to a particular class.

Made with Gamma

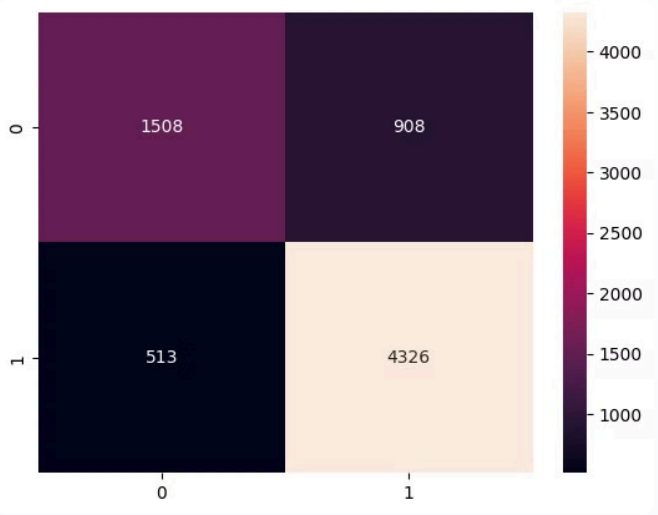
Comparative Analysis for Model Evaluation :



Decision Tree Classifier



Random Forest Classifier



Logistic Regression

Made with Gamma

Model Performance Summary :

Model : DecisionTreeClassifier

Accuracy : 0.86

Model : RandomForestClassifier

Accuracy : 0.90

Model : LogisticRegression

Accuracy : 0.80

Made with Gamma

Conclusion: The Future of Hotel Revenue Management

By leveraging data-driven decision-making, hotels can optimize operations, maximize revenue, and deliver exceptional customer experiences. The future holds exciting possibilities as AI and machine learning continue to advance, further enhancing cancellation prediction and revenue management.

Made with Gamma

Thank You !!

Made with Gamma