### Project Title : Hotel Revenue Management System – Data Science

### Name : Balram Shah

### Affiliation : Internship at HCL Tech

### Date of Submission : 04/29/2025

**1. Introduction to the Problem Statement**

In the intensely competitive hotel sector, revenue management becomes crucial to attain optimal profitability and sustainable business development. Hotels are required to streamline pricing, occupancy and demand forecasting to be able to compete and meet customers' needs. However traditional pricing models pay no attention to shifting demand, market and competitor behavior resulting in loss of revenue and wasteful usage of resources.



This project focuses on the development of a Machine Learning and Data Analytics-Driven Hotel Revenue Management System to address these requirements. Based on historical booking data, patron consumption behavior and market trends. The system will enhance decision-making in pricing, demand forecasting, and occupancy management.

With interactive visualization and predictive analytics, hotel management is able to make informed decisions that maximize revenue, improve operational effectiveness, and maintain customer satisfaction.

**The overall objectives of this project are\***

**• Room Rate Optimization:**

Applying dynamic rate strategies based on demand, seasonality, and competitor information.

**• Enhancing Occupancy Management:**

Forecasting demand variations to achieve the best room mix.

**• Enhancing Decision-Making:**

Providing actionable intelligence through data visualization and predictive analytics.

**• Enhancing Competitiveness:**

Leverage real-time intelligence to stay competitive in the market.

This paper gives a comprehensive overview of the approaches used in the construction of the Hotel Revenue Management System, from data preprocessing and feature engineering to machine learning modeling and system integration with interactive Streamlit UI. The results and outputs to be produced will allow hotel companies to improve revenue planning and overall business performance in a data-driven manner.

**2. Data Collection & Preprocessing**

**- Dataset Description**

The data set used in this project is a hotel revenue data set that contains the most impactful variables required for revenue management.

The dataset used in this project was initially noisy unstructured, named “hotel\_revenue\_data\_noisy.csv”. It contained over 70,000+ records with information on booking dates, room types, customer segments, lead times, pricing, night stayed, cancellations and more. After data cleaning, the refined dataset “cleaned\_hoteldata\_before\_encoding.csv.” was used for all further analysis and modelling.

The data provides a general indication of how the various variables affect revenue production in the hotel industry.

**Key Variables:**

**• Room Rate** – Daily rates for various categories of rooms.

**• Occupancy Rate** – Proportion of rooms occupied at any one time.

**• Demand Levels** – Seasonal and market fluctuation in room bookings.

**• Competitor Pricing** – The price of the competitors.

**• Customer Segmentation** – Segmentation of visitors according to habits and likes of booking.

**3. Data Cleaning**

Data cleaning was a critical phase to ensure quality input for analysis and modeling. The following steps were taken:

* Removed duplicates and irrelevant columns
* Handled missing values (imputation and removal)
* Standardized date formats and categorical values
* Engineered features such as occupancy rate, final price, and lead time buckets

**Outlier Detection:**

* Outliers in avg\_daily\_rate were detected across different room types using the Interquartile Range (IQR) method.
* However, actual outlier removal was not performed to preserve real-world booking price variations.
* Boxplots were used to visualize the distribution and potential impact of extreme values.

**4. Exploratory Data Analysis (EDA)**

EDA was performed to uncover patterns in:

* Monthly revenue trends
* Room type and customer segment behavior
* Booking lead time and cancellation likelihood
* Seasonal patterns affecting occupancy and pricing and more.

Key insights were visualized using libraries like **Seaborn, Matplotlib** and **Plotly.**

**5. Visual Insights**

The Streamlit dashboard includes:

* Monthly Hotel Revenue Trends
* Room price vs. occupancy analysis
* Booking cancellation by customer segment
* Revenue contribution by room type
* Average daily rate across months
* Lead time distribution

These visuals help stakeholders understand data-driven decisions clearly.

**6. Feature Engineering & Transformation**

**Feature Scaling**

To ensure uniformity in numerical features, the following techniques were applied:

* **Standardization:** Applied to features like room rates and revenue to scale them to a mean of **0** and standard deviation of **1**.
* **Normalization:** Used for occupancy rates and demand patterns to bring values within a **0-1** range.

**Encoding Methods Applied:**

* **Label Encoding:** Used for ordinal and some nominal features where categories were represented as integers. For example, customer types or occupancy status were encoded using simple integer labels.
* **Map Encoding:** Applied for converting string categories to numeric format by directly mapping values (e.g., room types like 'Deluxe' to 1, 'Standard' to 0).
* **Binary Flags:** Boolean features such as cancellation status were encoded as binary values (0 or 1).

These transformations ensured that all categorical data could be effectively used in model training without loss of meaning or model performance.

**7. Machine Learning Model & Analysis**

A machine learning model using XGBoost was trained to predict expected revenue.

**- Model Selection**

* **Algorithms tried:** Linear Regression, Random Forest, Dicision Tree, XGBoost Regressor, Gradient Regressor,etc.
* **Model comparison :**  Performance metrics (R², RMSE, MAE)
* **Hyperparameter tuning :**  GridSearchCV, RandomizedSearchCV

**Best Model :**

* **Chosen Model:** XGBoost performed the best with **R² Score = 0.9983 (test data), MAE = 31.04** And **RMSE = 54.68**
* **Feature Importance:** List top contributing factors to revenue

**Example-Insight:**  
-- "Room price, occupancy rate, night stayed and competitor pricing were the most influential features."

**8. Dashboard Implementation**

The final system was deployed using **Streamlit**. Key features:

* Interactive filters (room type, month, lead time, etc.)
* Real-time predictions using form inputs
* Dynamic visual charts
* Downloadable filtered data

**9. Key Findings & Business Impact**

* Revenue varies significantly by season and customer type
* Cancellations are common for short-lead bookings
* Room pricing has a direct correlation with occupancy
* Smart pricing improves profit without compromising customer experience

### **10. Conclusion**

The Hotel Revenue Management System project demonstrates how data science can transform hotel operations. From noisy data to interactive analytics and accurate forecasting, the system covers all aspects of smart revenue management. It empowers hotel managers to make data-backed decisions confidently.

**11. Future Enhancements**

* Add real-time competitor scraping
* Expand to multiple property support
* Include guest review sentiment analysis
* Integrate with external booking APIs
* Future work can include advanced outlier handling techniques to further improve model performance.

**• Github\_Link :** [**https://github.com/balramshah01/hotel\_webapp**](https://github.com/balramshah01/hotel_webapp)

**• App\_Link : https://balramshah-hotel-webapp.streamlit.app**