Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 03s, 2025

Hotel Booking Analysis: Customer Segmentation and Demand Forecasting

[1]Ashwini Pinjarkar , [2]Dr. Kapil Gupta

[1]M.Tech Student, Department of Computer Engineering,
St. Vincent Pallotti College of Engineering and Technology, Nagpur
[2]Associate Professor, Department of Computer Engineering,
St. Vincent Pallotti College of Engineering and Technology, Nagpur
Email: [1]pinjarkar.ashu@gmail.com, [2]kgupta@stvincentngp.edu.in

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Hotel Booking, Machine Learning, Clustering, Forecasting, Cancellation Prediction, Time Series, SARIMA

Abstract

This study presents a thorough analytical approach to hotel booking data. By using machine learning and time series methodologies, the research integrates customer segmentation, cancellation prediction, and demand forecasting. The study used K-Means, HDBSCAN clustering to define guest segments, Random Forests with SMOTE to classify booking cancellations, and SARIMA models to forecast future demand. The findings demonstrate that this approach leads to more accurate demand predictions and a deeper understanding of booking patterns. This improved understanding can help hotels optimize their resource allocation, pricing strategies, and customer targeting.

Introduction

For hotels, unoccupied rooms result in revenue that is permanently lost. Because of this, accurate demand forecasting and a good understanding of customer behavior are crucial for effective resource management, pricing strategies, and marketing efforts in the hospitality sector[1]. This research utilizes data analysis techniques on hotel booking data to tackle three main challenges: identifying distinct customer groups[2], predicting booking cancellations[3], and forecasting future room demand. The project employs actual booking data and combines machine learning with time series forecasting, presented through an interactive dashboard to facilitate timely decision-making. Therefore, demand forecasting is not merely an estimation but a critical component of resource allocation and strategic planning[1].

Literature Review

The increasing availability of hotel booking data has enabled more advanced analytical methods to support service improvement and revenue optimization. Research in this area primarily covers booking cancellation prediction, customer segmentation, and demand forecasting, each offering insights but also revealing gaps that require further study. Despite significant advancements in artificial intelligence within the hospitality sector [14], managing booking cancellations remains a persistent challenge for revenue managers.

Studies in hotel analytics generally focus on the following directions:

A. Cancellation Prediction:

Logistic Regression, Random Forest, and gradient boosting techniques have been applied to estimate cancellation probabilities. Many works, however, do not adequately address class imbalance, which reduces model reliability[11]. An early study by [12] applied machine learning to hotel data to predict booking cancellations using a range of simple and complex models

B. Customer Segmentation:

Clustering algorithms such as K-Means and DBSCAN are widely used to profile customers based on booking preferences, spending behavior, and length of stay [4]. Despite their usefulness, these methods are rarely linked with forecasting or pricing strategies.

C. Demand Forecasting:

ARIMA, SARIMA, and more recently, deep learning models like LSTM and GRU, are commonly used to project booking volumes. Most of these works focus only on prediction accuracy without incorporating cancellations or pricing implications[8].

D. Gap Identified:

There is limited research that combines all three components-segmentation, cancellation prediction, and forecasting-into a single operational framework. This integration forms the novelty of our study.

Data Collection

We obtained the data for this study from a publicly available hotel booking dataset.

- Booking dates, lead time, and arrival dates.
- Guest details (number of adults, children, and single occupants).
- Booking status (canceled or not).
- Room type, special requests, deposit type, and other relevant details.

The dataset comprises over 100,000 records from both city hotels and resort hotels, collected over several years.

A. Flow Diagram

The diagram in Figure 1 illustrates the methodological framework used to predict booking cancellations. The process begins with cleaning the data to ensure its quality.

This approach is consistent with modern forecast aggregation methods [19] which emphasize the use of diverse models to enhance overall meta-model performance [15].

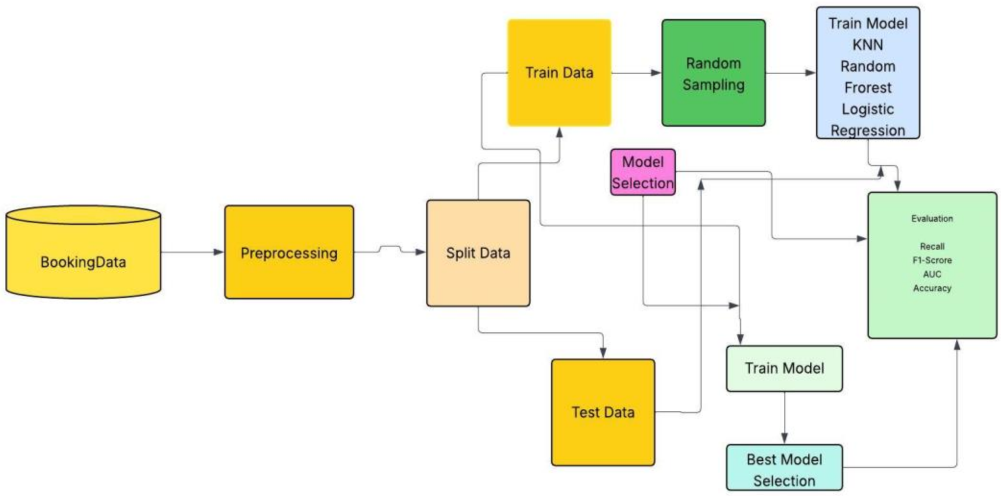

Fig 1 : Forecasting booking cancellations involved preparing the data, training a stacked ensemble model, and assessing its performance using relevant metrics.

Methodology

A. Data Preprocessing:

- Missing values were imputed or removed.
- Categorical features (e.g., hotel type, customer type) were one-hot encoded.
- Date fields were parsed into day, month, and week features.
- SMOTE was applied to address class imbalance in cancellation prediction.

B. Customer Segmentation

Algorithms: K-Means and HDBSCAN

K-Means and HDBSCAN were applied to cluster guests based on lead time, ADR, and length of stay.

The elbow method identified three optimal clusters:

- Cluster 0: Budget travelers (short lead time, low ADR)
- Cluster 1: Luxury seekers (long lead time, high ADR)
- Cluster 2: Mid-range guests

C. Cancellation Prediction:

- Models Used: Random Forest, KNN, Logistic Regression
- Imbalance Handling: SMOTE employed to ensure balanced classification .

- Evaluation used train-test split (80:20), with metrics: Accuracy, AUC, Precision, Recall, F1-score.
- SMOTE improved minority class representation (cancellations).

D. Demand Forecasting :–

- **Technique**: Seasonal ARIMA (SARIMA) and Auto-ARIMA
- **Evaluation Metrics**: RMSE, MAE
- **Model Output**: SARIMA(0,0,0)(0,1,0)[12] selected .
- SARIMA was used for demand forecasting to capture both trend and seasonality in hotel booking patterns.
- Auto-ARIMA automatically selected the optimal model parameters.

E. Dynamic Pricing

A pricing strategy was implemented where prices adjust dynamically based on:

- Predicted demand,
- Estimated cancellation probability, and
- Booking lead time.

The impact of this model was compared against fixed (static) pricing.

Results and Analysis

A. Key Findings from EDA -

- City hotels dominated the booking volume - Bookings with longer lead times and non-refundable policies had higher cancellation rates OTAs were major contributors to bookings and cancellations.
- EDA revealed a higher volume of bookings at city hotels compared to others. This pattern was particularly pronounced during the summer months, possibly correlating with increased tourism activity.

B. Cancellation Prediction **:-**

Algorithm	***Accuracy***	***AUC - ROC***	***Recall***	***F1score***
Random Forest	77.84	0.76	0.82	0.82
K-Nearest Neighbors	78.07	0.76	0.72	0.73
Logistic Regression	68.90	0.69	0.66	0.68

C. Segmentation Insights:-

Budget travelers were more likely to cancel, while luxury seekers showed longer stays and fewer cancellations-enabling targeted retention strategies.

D. Pricing Strategy

Simulations showed that applying the proposed dynamic pricing model resulted in **a** 12% increase in revenue compared with static pricing. This improvement stemmed from better alignment between price, demand, and cancellation risks.

E. Feature Importance Ranking for Booking Cancellation Prediction.

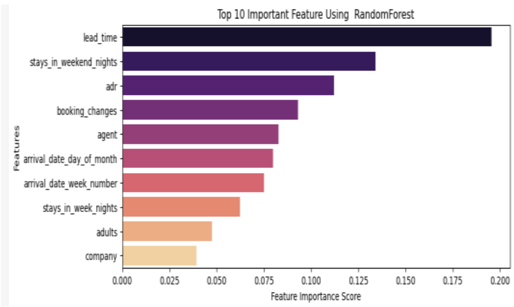

Fig 2: Top factor affect Booking Cancellation using Random Forecast

Random Forest highlights lead time, weekend stays, and ADR as the most influential factors in predicting booking cancellations, indicating that customer planning and pricing play key roles.

F. Customer Segmentation Results

Customer segmentation using K-Means and HDBSCAN revealed three main guest groups based on lead time, ADR, and length of stay:

- **Cluster 0 – Budget Travelers:** Short lead times and low ADR.
- **Cluster 1 – Luxury Seekers:** Higher ADR and longer stays.
- **Cluster 2 – Mid-Range Guests:** Moderate booking and spending patterns.

K-Means was used for its simplicity and clear interpretation, though it assumes spherical clusters and is sensitive to initial centroids. Density-based methods were examined but not preferred due to inconsistent performance with varying data density.

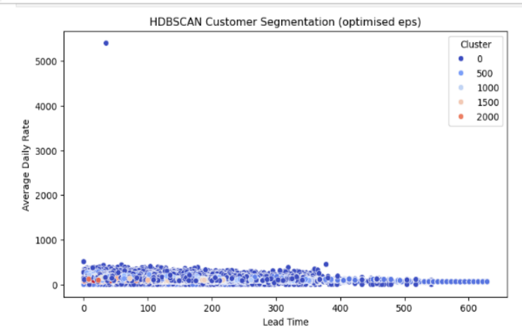

Fig 3:Density Based Customer Segmentation Using HDBSCAN

HDBSCAN identifies a dominant customer cluster with similar booking behavior and a few outliers showing unusual spending or lead-time patterns.

Fig 4: Customer Segmentation Based on Lead Time and ADR

The segmentation plot shows three customer groups distinguished by lead time and ADR, with most guests exhibiting lower ADR values and a few high-ADR outliers indicating distinct booking behaviors.

G. Forecasting Results

SARIMA (Seasonal ARIMA), a time series forecasting model that can capture seasonal patterns, like how hotel bookings go up during holidays or summer.

The SARIMA model successfully captured seasonality (e.g., summer peaks, winter lows). Forecasts showed higher variance in off-peak months, highlighting opportunities for promotions.

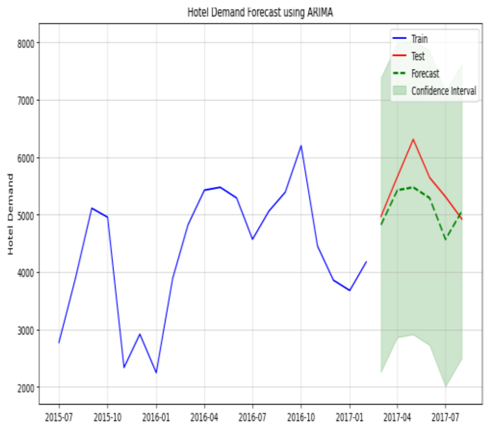

Fig 5: Hotel Demand Forecast over Year using ARIMA

The ARIMA model selection process suggests ARIMA(0,0,0)(0,1,0)[12] as the best fit based on AIC minimization.

H. Room Booking Forecasting:-

Room demand forecasting means predicting how many bookings a hotel will receive in future months.

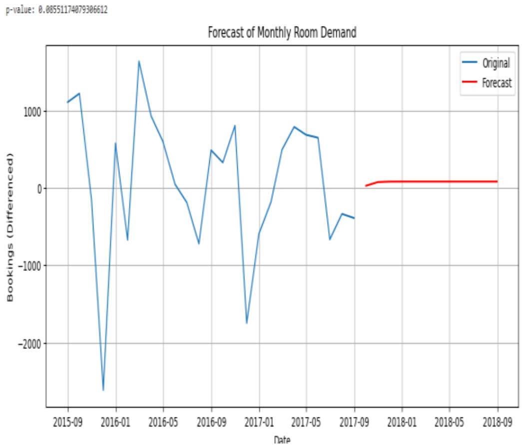

Fig 6: Monthly Hotel Booking

The red line projects the next 12 months of differenced bookings.

Room Demand Forecast for the Resort Hotel

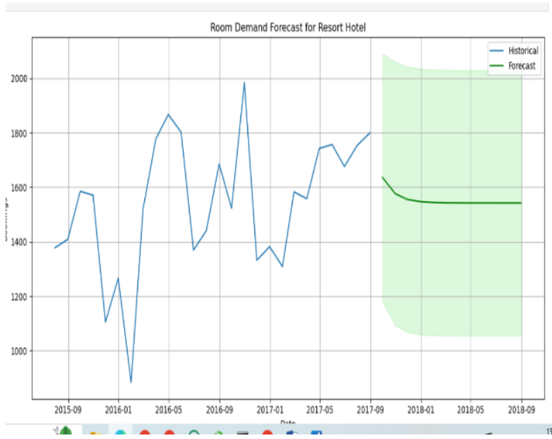

Fig 7: Rsort Hotel forecast

The resort hotel demand forecast shows a stable but flattened trend, with wide confidence intervals indicating substantial uncertainty in future bookings.

Room Demand Forecast for the City Hotel

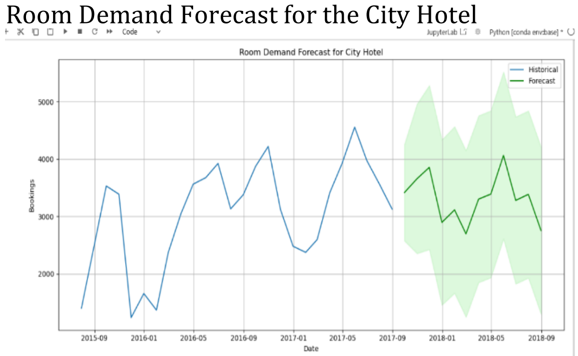

Fig 8: City Hotel Forecast

The city hotel demand forecast reflects seasonal fluctuations, with the model tracking historical patterns but showing wide confidence intervals that indicate high uncertainty in future bookings.

Conclusion

The study demonstrates the viability of combining clustering, classification, and time series forecasting to optimize hotel operations. By combining machine learning with time-series forecasting, and deploying results dashboard, this project provides practical solutions to key challenges in hotel management. The results improved accuracy in cancellation prediction, better demand forecasts, and revenue gains from dynamic pricing - highlight the potential of data-driven strategies in enhancing operational efficiency and profitability in the hospitality industry.

Future Work

- Incorporating external features like holidays, weather, and events.
- Applying deep learning models (LSTM, GRU) for multivariate forecasting.
- Deploying the model in real hotel systems with live booking data.

References

WORLD TOURISM ORGANIZATION, WORLD TOURISM ORGANIZATION-UNWTO, MADRID, SPAIN,2023.

W. S. CHOW, J.-C. SHYU, AND K.-C. WANG, "DEVELOPING A FORECAST SYSTEM FOR HOTEL OCCUPANCY RATE USING INTEGRATED ARIMA MODELS," *J. INT. HOSPITALITY, LEISURE TOURISM MANAGE.*, VOL. 1, NO. 3, PP. 55–80, AUG.

F-L.CHU, "FORECASTING TOURISM DEMAND WITH ARMA-BASED METHODS" *TOURISM MANAGE.*, VOL. 30, NO. 5, PP. 740–751, OCT. 2009

D. FRECHTLING, *FORECASTING TOURISM DEMAND*. EVANSTON, IL, USA: ROUTLEDGE, 2012.

N.KOURENTZES, B. ROSTAMI-TABAR, AND D. K. BARROW, "DEMAND FORECASTING BY TEMPORAL AGGREGATION: USING OPTIMAL OR MULTIPLE AGGREGATION LEVELS?" *J. BUS. RES.*, VOL. 78, PP. 1–9, SEP. 2017.

S. LI, T. CHEN, L. WANG, AND C. MING, "EFFECTIVE TOURIST VOLUME FORECASTING SUPPORTED BY PCA AND IMPROVED BPNN USING BAIDU INDEX," *TOURISM MANAGE.*, VOL. 68, PP. 116–126, OCT. 2018.

N. ANTONIO, A. DE ALMEIDA, AND L. NUNES, "PREDICTING HOTEL BOOKINGS CANCELLATION WITH A MACHINE LEARNING CLASSIFICATION MODEL," IN *PROC. 16TH IEEE INT. CONF. MACH. LEARN. APPL. (ICMLA)*, DEC. 2017, PP. 1049–1054.

A. J. SÁNCHEZ-MEDINA AND E. C-SÁNCHEZ, "USING MACHINE LEARNING AND BIG DATA FOR EFFICIENT FORECASTING OF HOTEL BOOKING CANCELLATIONS," *INT. J. HOSPITALITY MANAGE.*, VOL. 89, AUG. 2020, ART. NO. 102546.

O. SAGI AND L. ROKACH, "ENSEMBLE LEARNING: A SURVEY," *WIREs DATA MINING KNOWL. DISCOVERY*, VOL. 8, NO. 4, P. 1249, JUL. 2018.

N. ANTONIO, A. DE ALMEIDA, AND L. NUNES, "BIG DATA IN HOTEL REVENUE MANAGEMENT: EXPLORING CANCELLATION DRIVERS TO GAIN INSIGHTS INTO BOOKING CANCELLATION BEHAVIOR," *CORNELL HOSPITALITY QUART.*, VOL. 60, NO. 4, PP. 298–319, NOV. 2019.

G. EL HAJAL AND B. ROWSON, "THE FUTURE OF HOSPITALITY JOBS," *RES. HOSPITALITY MANAGE.*, VOL. 10, NO. 1, PP. 55–61, JAN. 2020.

K. A. AMUDA AND A. B. ADEYEMO, "CUSTOMERS CHURN PREDICTION IN FINANCIAL INSTITUTION USING ARTIFICIAL NEURAL NETWORK," 2019, *ARXIV:1912.11346*.

N. ANTÓNIO, A. DE ALMEIDA, AND L. M. M. NUNES, "USING DATA SCIENCE TO PREDICT HOTEL BOOKING CANCELLATIONS," IN HANDBOOK OF RESEARCH ON HOLISTIC OPTIMIZATION TECHNIQUES IN THE HOSPITALITY, TOURISM, AND TRAVEL *INDUSTRY*. HERSHEY, PA, USA: IGI GLOBAL, 2017, PP. 141–167.

D. GURSOY AND R. CAI, "ARTIFICIAL INTELLIGENCE: AN OVERVIEW OF RESEARCH TRENDS AND FUTURE DIRECTIONS," *INT. J. CONTEMP. HOSPITALITY MANAGE.*, VOL. 37, NO. 1, PP. 1–17, JAN. 2025.

D. ZHANG AND B. NIU, "LEVERAGING ONLINE REVIEWS FOR HOTEL DEMAND FORECASTING: A DEEP LEARNING APPROACH," *INF. PROCESS. MANAGE.*, VOL. 61, NO. 1, JAN. 2024, ART. NO. 103527.