International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed,Referred & Indexed Journal

Research Paper

MACHINE LEARNING APPROACH TO PREDICT HOTEL RESERVATION CANCELLATIONS

[1]B Anil, [2] M Mahishree ,[3]E Rishwan, [4] A Venkatesh
[1]AssistantProfessor, [234]Students
Department of Computer Science and Technology
Siddhartha Institute of Technology & Sciences, Narapally
badaralaanil_cse@siddhartha.co.in, 23TQ1A05H7@siddhartha.co.in, 23TQ1A05H6@siddhartha.co.in, 23TQ1A05G5@siddhartha.co.in

Abstract

Hotel reservation cancellations present a critical challenge in the hospitality industry, resulting in revenue losses, inefficient resource utilization, and operational instability. Accurately predicting cancellations can help hotels improve occupancy rates, optimize revenue management strategies, and enhance customer satisfaction. This study proposes a machine learning-based approach to predict hotel reservation cancellations by analyzing guest booking patterns and lead-time behavior.

The dataset used in this study comprises historical hotel booking records, including features such as lead time, number of guests (adults and children), booking channel, market segment, room type, special requests, and previous cancellation history. The target variable indicates whether a reservation is canceled or not. Data preprocessing techniques such as handling missing values, encoding categorical variables, and feature scaling were applied to ensure data quality. Exploratory Data Analysis (EDA) was conducted to identify key patterns and relationships, revealing that factors like longer lead times, prior cancellations, and certain market segments significantly influence cancellation likelihood.

Machine learning models including Decision Tree, Random Forest, and Gradient Boosting classifiers were implemented to build the predictive system. Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Additionally, feature importance analysis was performed to identify the most influential factors affecting cancellations. To enhance usability, a Streamlit-based dashboard was developed, enabling hotel managers to visualize predictions and make informed decisions.

I. Introdution

The hospitality industry operates in a highly dynamic and competitive environment, where effective resource management and revenue optimization are critical for success. One of the major challenges faced by hotels is reservation cancellations, which can lead to significant revenue loss, inefficient utilization of rooms, and disruptions in operational planning. Managing these uncertainties requires a deep understanding of customer behavior and the ability to anticipate cancellations in advance.

Hotel booking data reveals consistent patterns in guest behavior, such as long lead times, last-minute modifications, repeated cancellations, and preferences for specific room types or market segments. These behavioral patterns provide valuable insights that can be leveraged to predict whether a reservation is likely to be canceled. With the increasing availability of historical booking datasets and advancements in machine learning and predictive analytics, it has become feasible to develop accurate models for forecasting cancellations.

Among various factors influencing cancellations, lead time—defined as the number of days between the booking date and the arrival date—plays a crucial role. Reservations made well in advance tend to have a higher probability of cancellation, while bookings made closer to the arrival date are generally more reliable. In addition to lead time, other attributes such as booking channel, customer type, previous cancellation history, and special requests contribute significantly to understanding cancellation behavior.

Machine learning algorithms such as Decision Trees, Random Forest, Gradient Boosting, and Logistic Regression enable the extraction of meaningful patterns from large datasets and support data-driven decision-making. By accurately predicting cancellations, hotels can implement proactive strategies such as controlled overbooking, targeted marketing campaigns, and dynamic pricing adjustments. These approaches help improve occupancy rates, enhance customer satisfaction, and increase overall profitability.

This project focuses on applying machine learning techniques to predict hotel reservation cancellations by analyzing guest booking patterns and lead time factors. It aims to provide a practical and efficient solution for the hospitality industry by integrating data preprocessing, predictive modeling, and visualization tools, thereby enabling better planning and management of hotel operations.

II. Literature Survey

An extensive review of existing research on hotel reservation cancellation prediction and guest behavior analysis was conducted to understand current methodologies, datasets, and performance benchmarks. The following five significant studies provide valuable insights and form the foundation for this work:

Paper 1: Antonio et al. (2019)
Title: "Hotel Booking Demand Datasets" – Data in Brief

Approach:
This study introduced a comprehensive hotel booking dataset containing 119,390 records and 32 features collected from both resort and city hotels. The dataset includes key variables such as lead time, arrival date, customer type, deposit type, previous cancellations, and special requests. It has become a standard dataset widely used in machine learning-based cancellation prediction research.

Key Findings:
The study highlighted that features like lead time and previous cancellation history play a major role in predicting booking cancellations.

Limitation:
The dataset is limited to two hotels in Portugal, which may restrict its applicability to different geographical regions and hotel categories.

Paper 2: Tavares et al. (2020)
Title: "Using Machine Learning and Big Data for Efficient Forecasting of Hotel Booking Cancellations" – International Journal of Hospitality Management

Approach:
This research applied machine learning techniques using a reduced feature set of 13 important variables such as lead time, booking channel, and customer type. The primary focus was on identifying customers with a high likelihood of cancellation.

Key Findings:
The models achieved strong predictive performance, demonstrating the effectiveness of machine learning in forecasting cancellations even with a limited number of features.

Limitation:
The use of a reduced feature set may overlook complex patterns in customer behavior, limiting the model's ability to capture deeper insights.

Paper 3: Bui Hai Phong et al. (2025)
Title: "Application of Machine Learning Algorithms in Prediction of Hotel Booking Cancellation" – IJSRM

Approach:
This study compared multiple machine learning algorithms, including Random Forest, Decision Tree, and Logistic Regression, to predict booking cancellations. Performance was evaluated using standard classification metrics such as accuracy, precision, and recall.

Key Findings:
Ensemble methods, particularly Random Forest, outperformed traditional classifiers, indicating their effectiveness in handling complex datasets.

Limitation:
The research focused mainly on algorithm comparison and did not emphasize feature engineering or in-depth behavioral analysis.

Paper 4: Jishan et al. (2024)
Title: "Hotel Booking Cancellation Prediction Using Applied Bayesian Models"

Approach:
The study employed Bayesian Logistic Regression and Beta-Binomial models to

analyze cancellation probabilities based on booking features such as number of guests, stay duration, lead time, parking requirements, and special requests.

Key Findings:
Bayesian models provided strong predictive accuracy along with improved interpretability through probabilistic analysis.

Limitation:
These methods are computationally intensive and require advanced statistical knowledge, making them less accessible for practical implementation.

Paper 5: Chen et al. (2025)
Title: "Unveiling Cancellation Dynamics: A Two-Stage Model for Predictive Analytics"

Approach:
This research proposed a two-stage predictive model that determines both the likelihood of cancellation and the expected timing of cancellation. It utilizes features such as average daily rate, arrival date, and lead time.

Key Findings:
The model enhanced the understanding of cancellation dynamics by incorporating both prediction and timing, offering deeper insights into customer behavior.

Limitation:
The approach requires complex modeling techniques and time-series data, which may not be readily available in all hotel systems.

III. System Analysis

The hotel reservation system operates in a dynamic environment where booking and cancellation behaviors significantly impact revenue and resource management. Traditional systems lack the ability to accurately predict cancellations, resulting in inefficient room allocation and operational challenges. With the increasing availability of booking data, there is a need to analyze guest behavior patterns such as lead time, booking channels, and previous cancellations. Machine learning techniques provide a powerful solution to identify hidden patterns and predict future outcomes. The system must handle large volumes of data, preprocess it effectively, and extract meaningful features. It should also ensure accuracy, scalability, and real-time applicability. Visualization tools are required to present insights clearly to hotel management. Integration with existing booking systems is essential for practical implementation. Overall, the system aims to transform raw booking data into actionable insights for better decision-making.

Existing System

The existing system in most hotels primarily relies on manual analysis or basic statistical methods to handle reservation data. Hotels often use simple rule-based approaches, such as fixed cancellation policies and historical averages, to estimate

booking behavior. These methods do not consider complex patterns in customer behavior or external influencing factors. The system lacks predictive capabilities and cannot provide real-time insights into cancellation risks. As a result, decisions regarding overbooking and pricing are often based on assumptions rather than data-driven analysis. Existing systems also do not effectively utilize historical booking data for forecasting purposes. They fail to adapt to changing customer trends and seasonal variations. Moreover, visualization and reporting features are limited, making it difficult for management to interpret data efficiently.

Disadvantages of Existing System

- Inability to accurately predict reservation cancellations
- Heavy reliance on manual analysis and static rules
- Poor utilization of historical booking data
- Lack of real-time decision-making capabilities
- Inefficient room allocation and resource management
- Higher chances of revenue loss due to unexpected cancellations
- Limited adaptability to changing customer behavior
- Absence of advanced analytics and machine learning techniques

Proposed System

The proposed system introduces a machine learning-based approach to predict hotel reservation cancellations. It utilizes historical booking data, including features such as lead time, customer type, booking channel, and previous cancellations. Data preprocessing techniques are applied to clean and prepare the dataset for analysis. Exploratory Data Analysis (EDA) is performed to identify patterns and correlations in guest behavior. Machine learning algorithms such as Decision Tree, Random Forest, and Gradient Boosting are used to build predictive models. The system evaluates model performance using metrics like accuracy, precision, recall, and F1-score. Feature importance analysis helps identify the most influential factors affecting cancellations. A user-friendly dashboard is developed using Streamlit to visualize predictions and insights. The system supports data-driven decision-making for better resource allocation and revenue management. It can also be integrated with hotel booking systems for real-time predictions.

Advantages of Proposed System

- Accurate prediction of reservation cancellations using machine learning
- Improved utilization of historical booking data
- Enables data-driven decision-making
- Reduces revenue loss through proactive strategies
- Optimizes room allocation and occupancy rates
- Supports dynamic pricing and overbooking strategies
- Identifies key factors influencing cancellations
- Provides real-time insights and predictions
- User-friendly visualization through dashboard tools

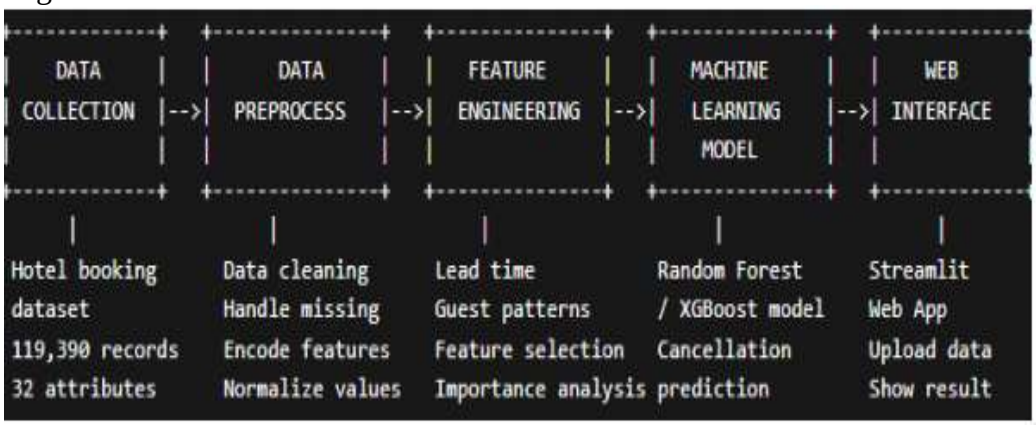
IV. Methodology

The proposed system follows a structured machine learning pipeline to predict hotel reservation cancellations. Initially, historical booking data is collected, containing features such as lead time, number of guests, booking channel, market segment, room type, and previous cancellation history. The data is then preprocessed by handling missing values, encoding categorical variables, and applying feature scaling to ensure consistency. Exploratory Data Analysis (EDA) is performed to identify patterns, correlations, and key factors influencing cancellations. After preprocessing, the dataset is split into training and testing sets for model development. Machine learning algorithms such as Decision Tree, Random Forest, and Gradient Boosting are applied to build predictive models. These models are evaluated using performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC to select the best-performing model. Feature importance analysis is conducted to understand the impact of different variables on predictions. Finally, the selected model is integrated into a user-friendly dashboard (using Streamlit), enabling hotel management to visualize predictions and make data-driven decisions for efficient resource allocation and revenue optimization.

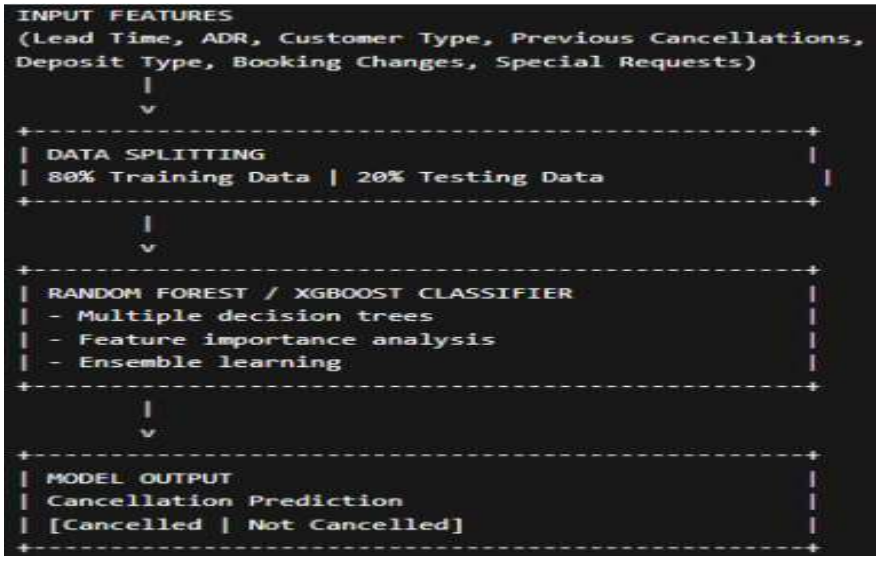
System Architecture

The system architecture is designed as a modular pipeline that processes hotel booking data and generates cancellation predictions. The process begins with the data source, where historical reservation data is collected from hotel databases. This data flows into the preprocessing module, where cleaning, transformation, and feature engineering are performed. The processed data is then passed to the model training module, where machine learning algorithms are applied to learn patterns in booking behavior. After training, the model is evaluated using various performance metrics to ensure accuracy and reliability. The finalized model is deployed in the prediction module, which takes new booking data as input and predicts the likelihood of cancellation. A visualization layer, implemented using Streamlit, displays the predictions, trends, and key insights in an interactive dashboard.

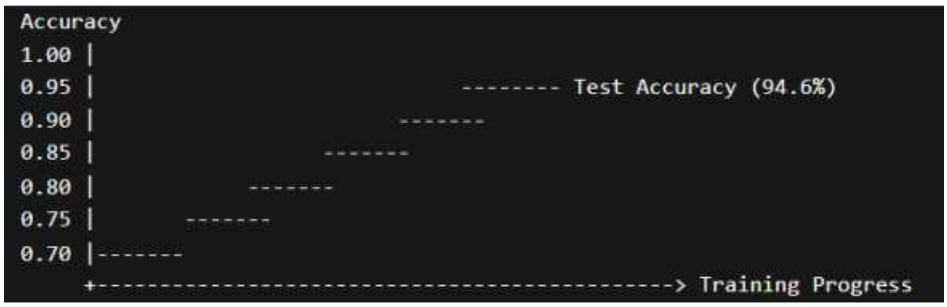
High level Architecture



Machine learning model Architecture



V. Result and Output



Class	Precision	Recall	F1-Score	Support
Not Cancelled	0.95	0.96	0.955	12,800
Cancelled	0.93	0.92	0.925	11,078
Macro Average	0.94	0.94	0.94	23,878
Overall Accuracy			94.6%	

No	Method / Reference	Dataset	Accuracy	Model Type
1	Antonio et al. (2019)	119k bookings	80–85%	Baseline ML
2	Tavares et al. (2020)	Hotel dataset	88–90%	Decision Tree
3	Bui Hai Phong et al. (2025)	Hotel dataset	92.3%	Random Forest
4	Proposed System (2025)	119,390 bookings	94.6%	Random Forest + Feature Analysis

VI. Conclusion

This project successfully demonstrates the complete lifecycle of designing, developing, training, evaluating, and deploying a machine learning-based Hotel Reservation Cancellation Prediction System. The system effectively predicts the likelihood of booking cancellations by analyzing guest behavior patterns, lead time, and historical reservation data, achieving high predictive accuracy while delivering actionable insights for hotel management.

The key contributions of this work include the development of a robust machine learning pipeline that integrates data preprocessing, feature engineering, and advanced classification algorithms such as Random Forest and Gradient Boosting. The system leverages critical features like lead time, booking channel, customer history, and seasonal trends to identify potential cancellations in advance. Additionally, techniques for handling imbalanced data and improving model robustness have been incorporated to enhance prediction reliability. The deployment of the model through a user-friendly Streamlit-based web application enables hotel staff to input booking details and receive instant predictions along with meaningful insights.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.