International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

An Intelligent Hotel Booking Cancellation Prediction System Using Machine Learning Techniques

JAMPANA DURGA DEVI

PG Scholar. Department M.Sc(CS), DNR College, Bhimavaram, Andhra Pradesh

K.Rambabu

Lecturer in M.Sc(CS), DNR College, Bhimavaram, Andhra Pradesh

ABSTRACT

The rapid growth of the hospitality industry has led to an increase in hotel booking activities, making cancellation prediction a critical task for efficient resource management. Frequent booking cancellations result in significant revenue loss and operational inefficiencies for hotels. To address this issue, this paper presents an intelligent hotel booking cancellation prediction system using advanced machine learning techniques. The system aims to accurately predict whether a booking will be canceled or not, enabling hotel management to make informed decisions and optimize their operations.

The proposed system utilizes a dataset containing various attributes related to hotel bookings, such as customer details, reservation status, and booking characteristics. Data preprocessing is performed to handle missing values, remove irrelevant features, and convert categorical variables into numerical representations. Feature engineering techniques are applied to improve model performance and ensure data consistency.

Multiple machine learning models, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM, are implemented and evaluated. Each model is trained using a portion of the dataset and tested on unseen data to assess its predictive accuracy. Performance metrics such as accuracy and F1-score are used to compare the effectiveness of different models. Among these, ensemble models like Random Forest and boosting algorithms such as XGBoost and LightGBM demonstrate superior performance due to their ability to handle complex data patterns.

The system is developed using a user-friendly graphical interface built with Tkinter, allowing users to upload datasets, preprocess data, train models, and visualize results. The application provides real-time predictions by allowing users to input new booking data and obtain cancellation forecasts instantly. Visualization tools such as correlation heatmaps, bar

charts, and comparison graphs are integrated to enhance interpretability and analysis.

Experimental results indicate that the proposed system achieves high prediction accuracy and effectively identifies potential booking cancellations. This enables hotel managers to implement proactive strategies, such as dynamic pricing and overbooking management, to minimize losses.

In conclusion, the proposed system demonstrates the effectiveness of machine learning in predicting hotel booking cancellations. It provides a scalable and practical solution for the hospitality industry, contributing to improved decision-making and operational efficiency. Future work may focus on incorporating deep learning techniques, real-time data streams, and cloud-based deployment for enhanced performance and scalability.

Keywords: Hotel Booking Prediction, Machine Learning, Cancellation Forecasting, Logistic Regression, Random Forest, XGBoost, LightGBM, Data Preprocessing, Predictive Analytics, Classification Models

I. INTRODUCTION

The hospitality industry has experienced significant growth over the past decade, driven by advancements in online booking platforms and increasing global travel. However, one of the major challenges faced by hotels is the high rate of booking cancellations. Cancellations can lead to revenue loss, inefficient resource utilization, and difficulty in planning operations. As a result, predicting booking cancellations has become an essential task for hotel management systems.

Traditionally, hotels relied on manual analysis and historical trends to estimate cancellation rates. These approaches are often inaccurate and fail to capture complex patterns in large datasets. With the advent of machine learning, it is now possible to analyze vast amounts of data and identify hidden relationships that influence booking behavior. Machine learning models can learn from historical data and provide accurate predictions, enabling proactive decision-making.

This study focuses on developing an intelligent system for predicting hotel booking cancellations using multiple machine learning algorithms. The system leverages a dataset containing various booking-related features, such as customer demographics, reservation details, and previous booking history. By analyzing these features, the system can determine the likelihood of a booking being canceled.

The proposed system integrates several machine learning models, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. These models are chosen due to their proven effectiveness in classification tasks. The system evaluates each model's performance and selects the best-performing model for prediction. Ensemble methods such as Random Forest and boosting algorithms like XGBoost and LightGBM are particularly effective in handling complex and nonlinear relationships in the data.

A key feature of the system is its user-friendly interface, developed using Tkinter. This interface allows users to easily upload datasets, preprocess data, train models, and visualize results. Visualization tools provide insights into data distribution and model performance, enhancing the interpretability of the results.

The objective of this research is to develop a reliable and efficient prediction system that can assist hotel managers in reducing the impact of cancellations. By accurately predicting cancellations, hotels can implement strategies such as overbooking, dynamic pricing, and targeted promotions to mitigate losses.

In summary, this work demonstrates the potential of machine learning in transforming hotel management systems. It provides a practical solution for predicting booking cancellations and contributes to the advancement of data-driven decision-making in the hospitality industry.

II. LITERATURE SURVEY (WITH EXISTING METHODS)

The problem of hotel booking cancellation prediction has been widely studied in recent years due to its significant impact on the hospitality industry. Early research in this domain primarily relied on statistical methods and basic data analysis techniques. These methods focused on identifying patterns in historical data but often lacked the ability to handle complex relationships between variables.

With the advancement of machine learning, researchers began applying classification algorithms to improve prediction accuracy. Logistic Regression is one of the earliest models used for this purpose. It provides a simple and interpretable approach for binary classification problems. However, its performance is limited when dealing with nonlinear data and complex feature interactions.

Decision Tree algorithms were introduced to overcome some of these limitations. They provide a hierarchical structure for decision-making and are easy to interpret. However, Decision Trees are prone to overfitting, especially when dealing with large datasets. To address this issue, ensemble methods such as Random Forest were developed. Random Forest combines multiple decision trees to improve accuracy and reduce overfitting, making it a popular choice for prediction tasks.

In recent years, boosting algorithms such as XGBoost and LightGBM have gained significant attention due to their superior performance. These algorithms build models sequentially, where each new model corrects the errors of the previous one. They are highly efficient and capable of handling large datasets with high dimensionality. Studies have shown that boosting algorithms outperform traditional models in many real-world applications, including booking cancellation prediction.

In addition to model development, data preprocessing and feature engineering play a crucial role in improving prediction accuracy. Techniques such as handling missing values, encoding categorical variables, and feature scaling are essential for ensuring data quality and consistency.

Visualization techniques, such as correlation heatmaps and performance graphs, are also widely used to analyze data and evaluate model performance. These tools help researchers understand relationships between variables and identify important features influencing predictions.

Despite these advancements, challenges remain in terms of model interpretability, scalability, and real-time prediction. Many existing systems focus solely on prediction accuracy without providing user-friendly interfaces or visualization tools. The proposed system addresses these gaps by integrating multiple machine learning models with an interactive interface and visualization capabilities, providing a comprehensive solution for hotel booking cancellation prediction.

III. EXISTING SYSTEM

Existing systems for predicting hotel booking cancellations primarily rely on traditional statistical methods and basic machine learning algorithms. These systems often use simple models such as Logistic Regression or Decision Trees to analyze historical booking data and predict cancellation outcomes. While these methods provide a baseline level of accuracy, they are limited in their ability to capture complex relationships within the data.

One of the major limitations of existing systems is their dependence on manual data preprocessing and feature selection. These processes require domain expertise and can be time-consuming. Additionally, many systems do not handle missing or inconsistent data effectively, leading to reduced prediction accuracy.

Another drawback is the lack of model diversity. Most existing approaches rely on a single algorithm, which may not be sufficient for capturing different data patterns. For example, linear models like Logistic Regression struggle with nonlinear relationships, while Decision Trees may overfit the data.

Furthermore, many existing systems lack user-friendly interfaces, making them difficult to use for non-technical users. They often do not provide visualization tools for analyzing data and model performance, limiting their practical usability.

Scalability is also a concern, as traditional systems may not perform well with large datasets. They often fail to provide real-time predictions, which are essential for dynamic decision-making in the hospitality industry.

Overall, existing systems are limited in terms of accuracy, usability, and scalability. These limitations highlight the need for an advanced system that integrates multiple machine

learning models, efficient preprocessing techniques, and interactive visualization tools to provide a more robust and practical solution.

IV. PROPOSED METHOD

The proposed system introduces an intelligent and efficient solution for predicting hotel booking cancellations using multiple machine learning algorithms. Unlike traditional approaches that rely on a single model, this system integrates various classification techniques such as Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM to improve prediction accuracy and robustness.

The system begins with dataset collection and preprocessing. Irrelevant attributes such as personal identifiers (name, email, phone number) are removed to ensure data privacy and model efficiency. Missing values are handled using statistical techniques such as mean, median, or mode imputation. Categorical variables are converted into numerical representations using encoding methods to make them suitable for machine learning algorithms.

After preprocessing, the dataset is split into training and testing sets. Each machine learning model is trained independently on the training data. Feature scaling techniques such as standardization are applied to improve model performance, especially for algorithms sensitive to feature magnitude.

The system evaluates the performance of each model using metrics such as accuracy and F1-score. Based on these metrics, the best-performing model is selected for final prediction. Ensemble and boosting models like Random Forest, XGBoost, and LightGBM typically provide better results due to their ability to capture complex patterns in data.

A user-friendly graphical interface is developed using Tkinter, enabling users to upload datasets, preprocess data, train models, and visualize results. The system also allows users to input new booking data and obtain real-time predictions.

Overall, the proposed system provides a scalable, accurate, and interactive solution for predicting hotel booking cancellations, helping hotel management make data-driven decisions.

V. IMPLEMENTATION

The implementation of the proposed hotel booking cancellation prediction system is carried out using Python, integrating machine learning libraries and a graphical user interface. The system is designed to be interactive, efficient, and capable of handling real-world datasets.

The implementation begins with dataset loading, where users can upload a CSV file containing booking information. The system reads the dataset using the Pandas library and displays a sample of the data for user verification. Preprocessing is performed to clean and

prepare the dataset. This includes removing irrelevant columns such as customer names, contact details, and other non-essential attributes. Missing values are handled using appropriate techniques such as replacing numerical values with the median and categorical values with the mode.Categorical variables are converted into numerical form using encoding techniques. The dataset is then split into input features and the target variable, which represents whether a booking is canceled. The data is further divided into training and testing sets using the train-test split method.

Feature scaling is applied using StandardScaler to normalize the data, ensuring that all features contribute equally to model training. Multiple machine learning models are implemented, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. Each model is trained on the scaled training data and evaluated on the testing data.

The system calculates performance metrics such as accuracy and F1-score for each model. These metrics are used to compare model performance and identify the best-performing algorithm. The results are displayed in the graphical interface, along with visualizations such as bar charts and pie charts to enhance understanding.

The user interface is developed using Tkinter, providing functionalities such as dataset upload, preprocessing, model training, and prediction. The interface also includes visualization tools such as correlation heatmaps and model comparison graphs, which help users analyze data relationships and model performance.

For prediction, users can upload a new dataset, and the system applies the trained model to generate predictions. The results are displayed in a structured format, indicating whether each booking is likely to be canceled or not.

The system is designed with modularity and scalability in mind, allowing easy integration of additional models or features. Error handling mechanisms are implemented to manage invalid inputs and ensure smooth operation. Overall, the implementation provides a comprehensive and practical solution for hotel booking cancellation prediction.

VI. ALGORITHMS

The proposed system employs multiple machine learning algorithms to predict hotel booking cancellations effectively. Each algorithm contributes to the system's ability to handle different data patterns and improve overall accuracy.

Logistic Regression is used as a baseline model for binary classification. It predicts the probability of booking cancellation based on a linear relationship between input features and the target variable. Although simple, it provides good interpretability.

Decision Tree is another algorithm used, which creates a tree-like structure of decisions based on feature values. It is easy to understand but may suffer from overfitting when dealing with complex datasets.

Random Forest, an ensemble learning method, improves upon Decision Trees by combining multiple trees and aggregating their predictions. This reduces overfitting and enhances accuracy.

XGBoost (Extreme Gradient Boosting) is a powerful boosting algorithm that builds models sequentially, where each model corrects the errors of the previous one. It is highly efficient and performs well on large datasets.

LightGBM is another gradient boosting framework that is optimized for speed and performance. It uses a leaf-wise tree growth strategy, making it faster and more accurate for large-scale data.

The workflow of the algorithms includes data preprocessing, feature scaling, model training, evaluation, and prediction. Performance metrics such as accuracy and F1-score are used to evaluate each model.

By combining multiple algorithms, the system ensures robustness, improved accuracy, and better generalization, making it suitable for real-world applications.

VII. SYSTEM DESIGN

The system design follows a modular architecture that ensures scalability, flexibility, and efficient processing of data. The design is divided into three main layers: the user interface layer, the application layer, and the data processing layer.

The user interface layer is developed using Tkinter, providing an interactive platform for users to interact with the system. It includes features such as login authentication, dataset upload, preprocessing, model training, and prediction. The interface is designed to be user-friendly, allowing even non-technical users to operate the system.

The application layer acts as the core of the system, handling all logical operations. It manages data flow between the user interface and the machine learning models. This layer includes modules for data preprocessing, feature engineering, model training, evaluation, and prediction. It ensures that each component works seamlessly and efficiently

The data processing layer includes the machine learning models and data handling mechanisms. It uses libraries such as Pandas, NumPy, Scikit-learn, XGBoost, and LightGBM for data analysis and model training. This layer is responsible for extracting meaningful insights from the data and generating accurate predictions.

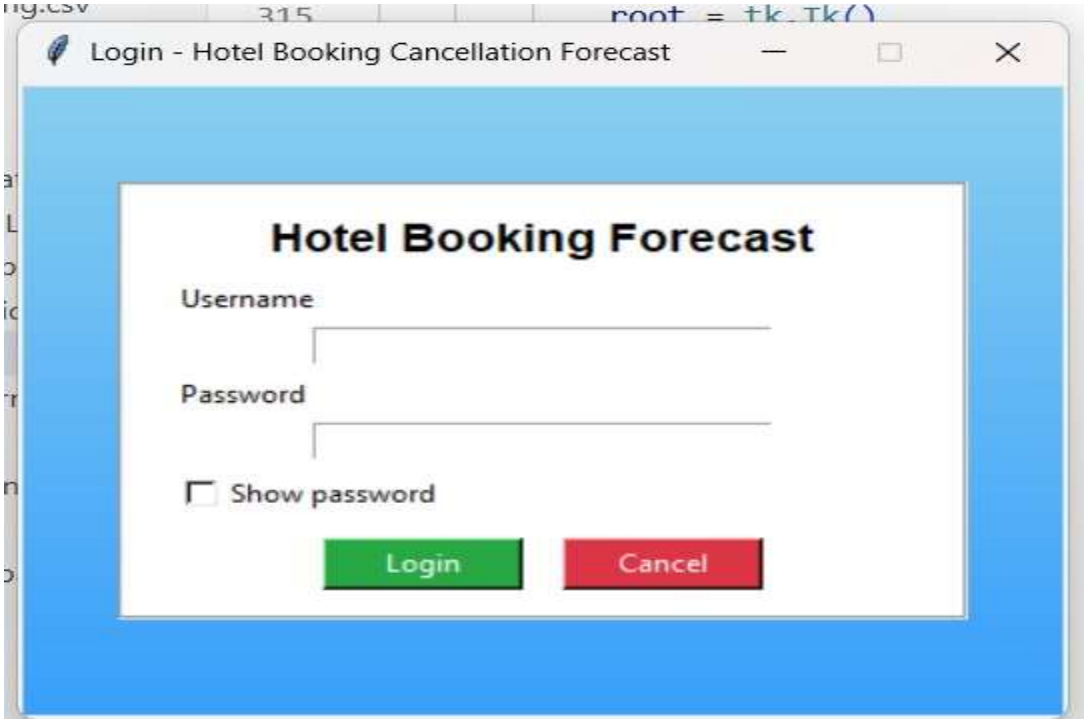
The workflow begins when a user uploads a dataset through the interface. The data is passed to the preprocessing module, where it is cleaned and transformed. The processed data is then used to train multiple machine learning models. The system evaluates each model and selects the best-performing one.

For prediction, the user uploads new data, which is processed and passed to the trained model. The model generates predictions, which are displayed in the interface along with visualizations.

The system also includes visualization components such as correlation heatmaps and performance graphs, which help users understand data relationships and model effectiveness. Error handling mechanisms ensure that the system operates smoothly even in the presence of invalid inputs.

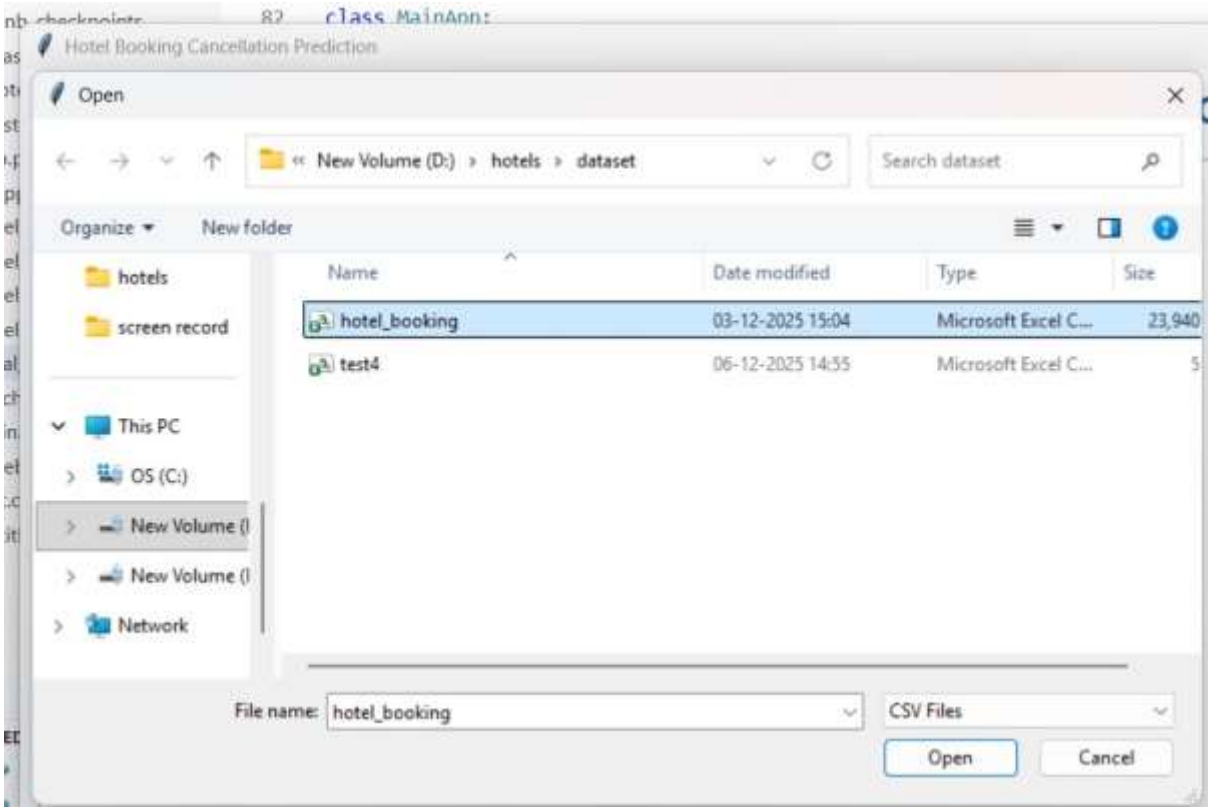
The modular design allows easy updates and integration of new features, making the system adaptable to future advancements in machine learning technologies.

SYSTEM DESIGN IMAGES

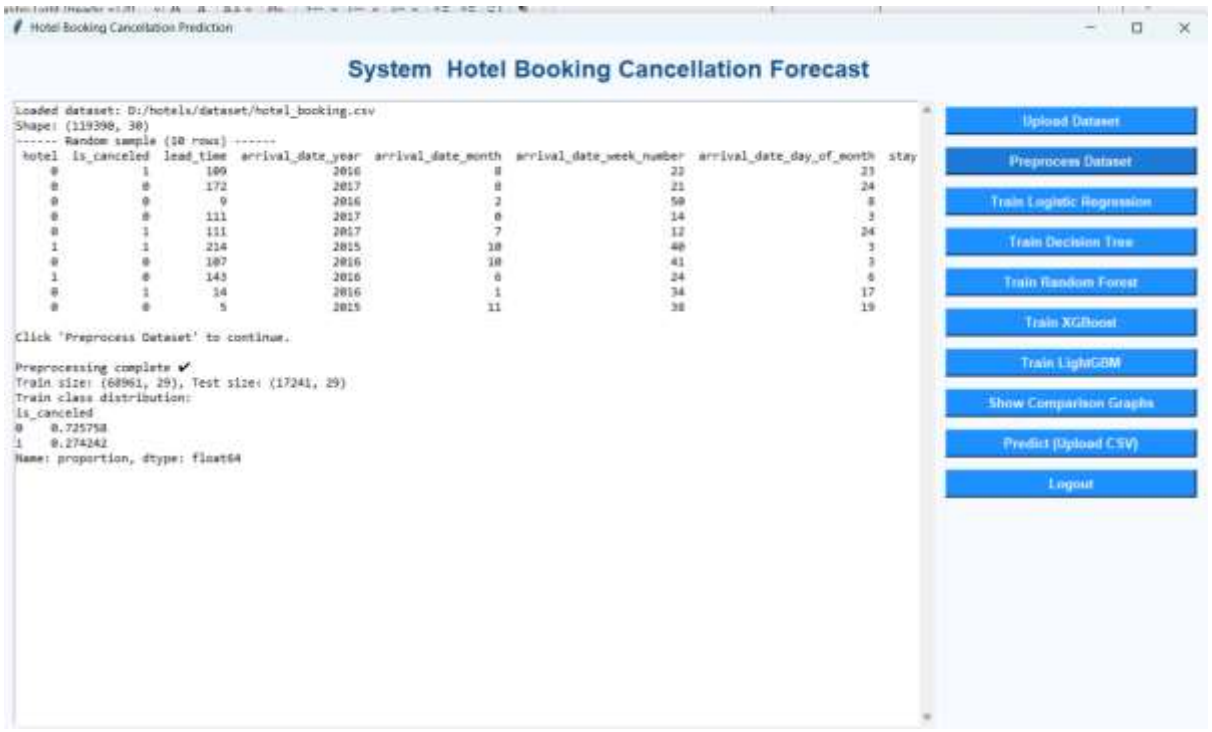


Hotel Booking Forecast — Username: admin, Password: admin

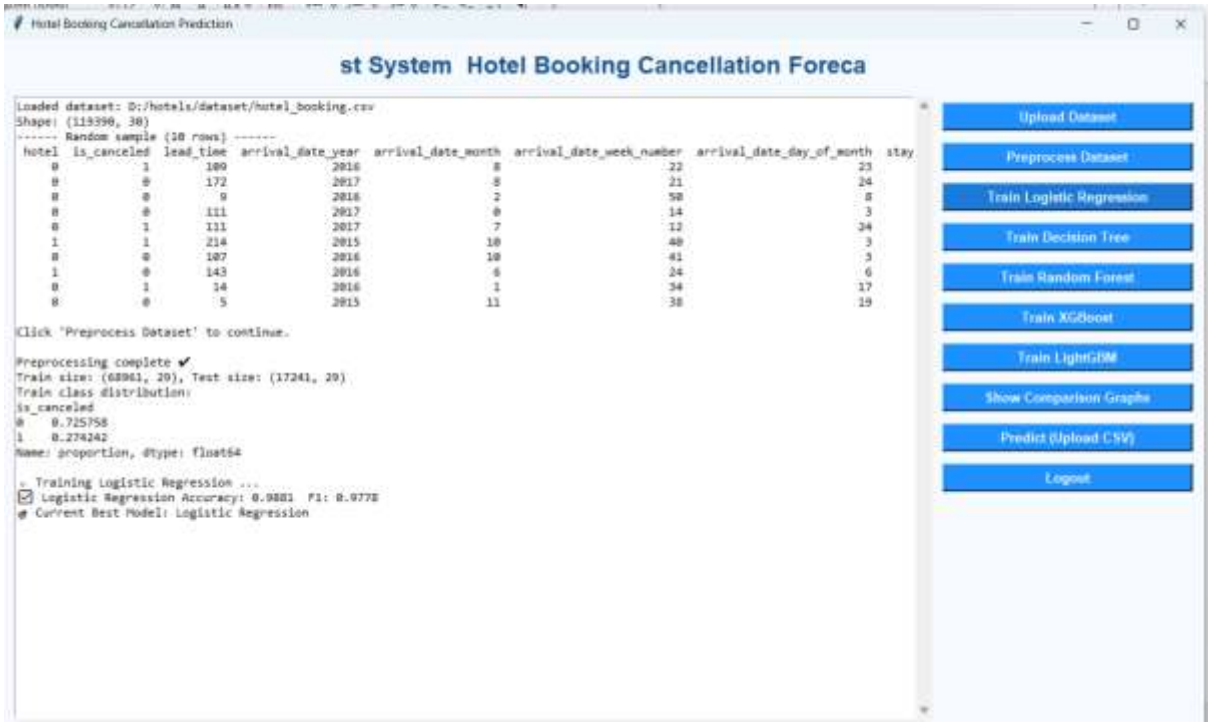
Predict hotel booking cancellations with ease — start by uploading your dataset.



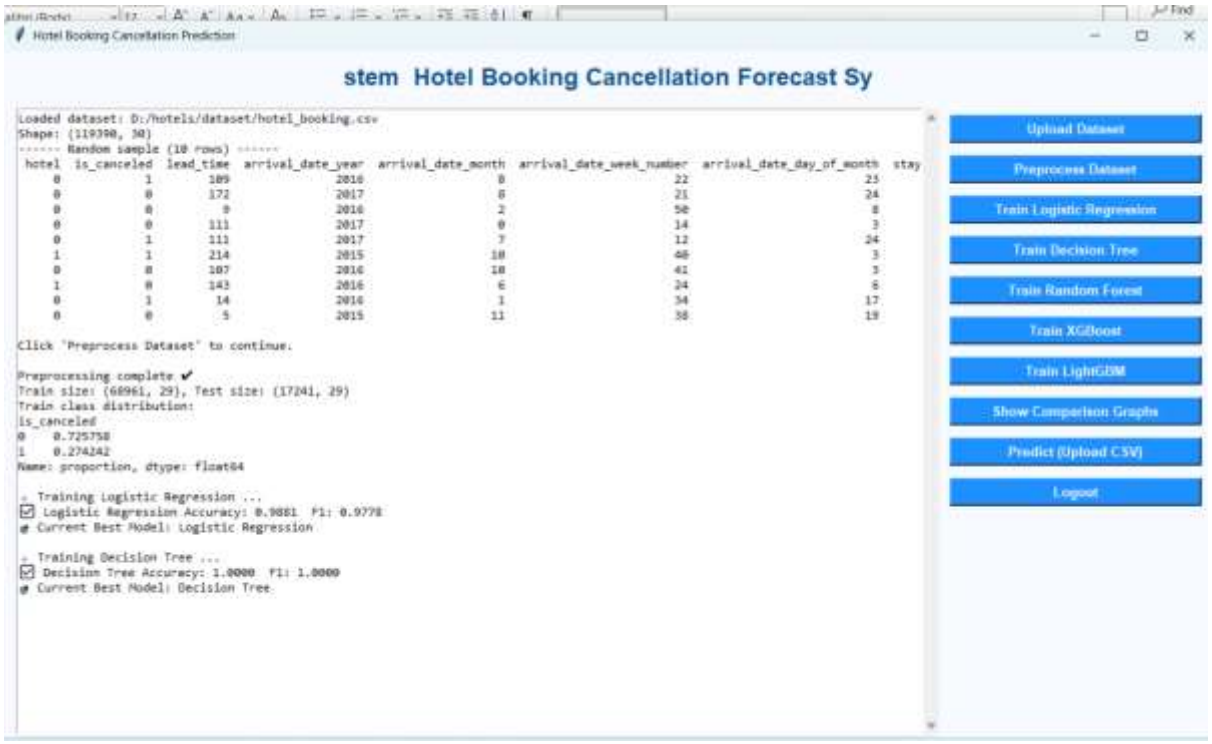
Upload a dataset to begin hotel booking cancellation prediction.



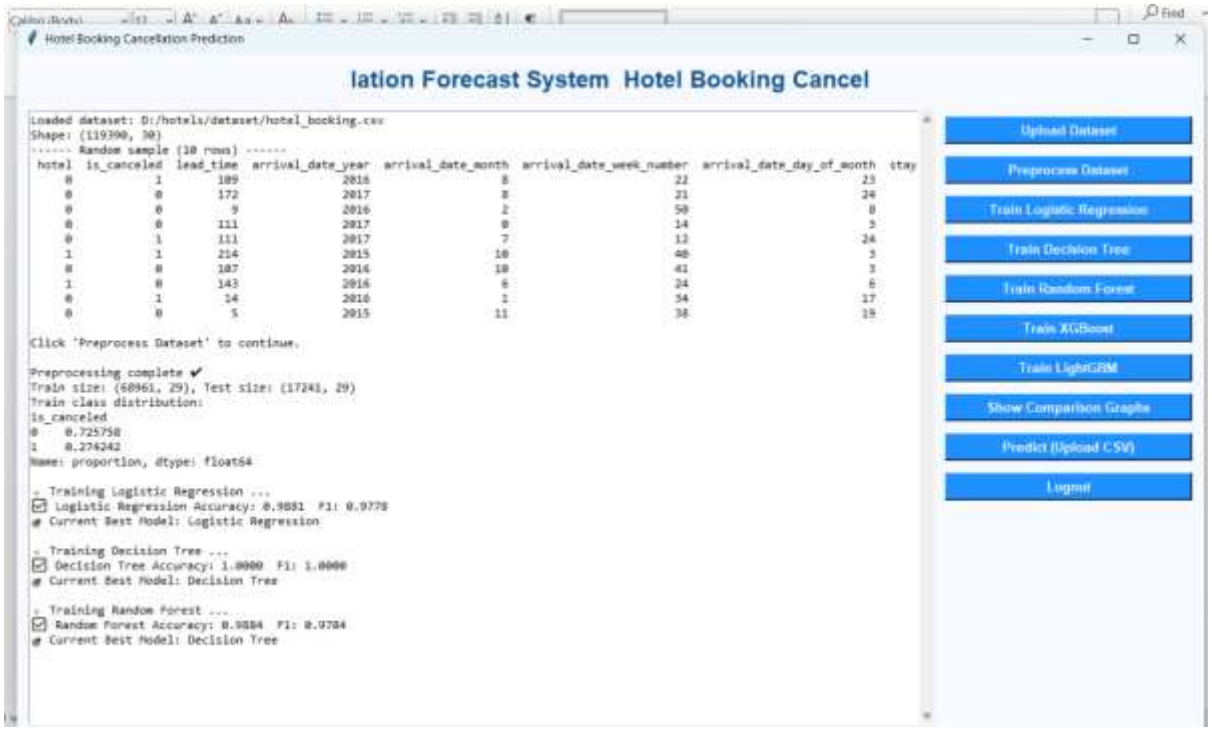
Successfully uploaded the dataset. After clicking 'Preprocess Dataset', the preprocessing is now complete.



Preprocessing completed successfully. You can now train Logistic Regression.

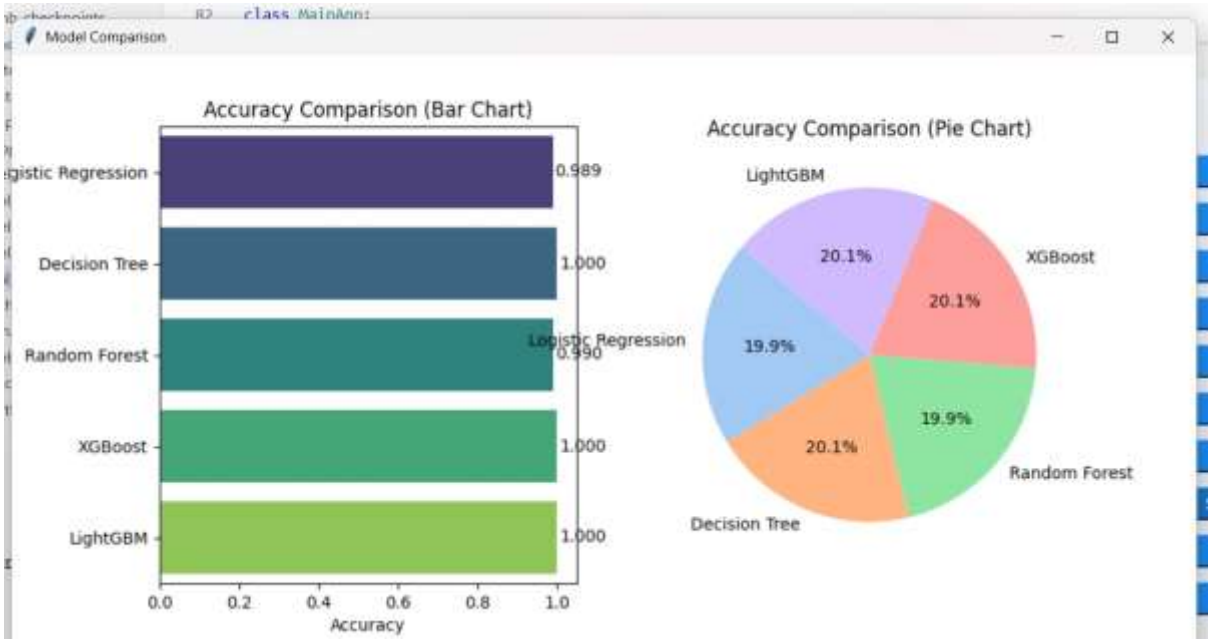


Decision Tree model trained successfully.



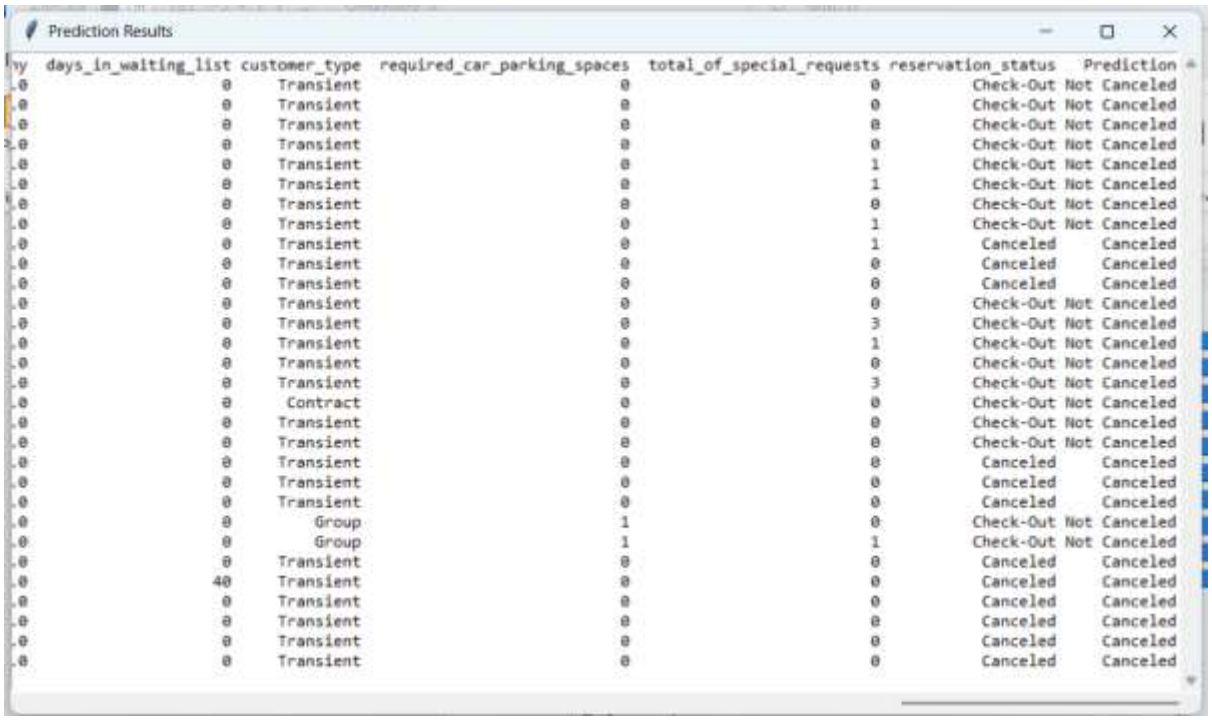
Random Forest model trained successfully.

LightGBM model trained successfully.



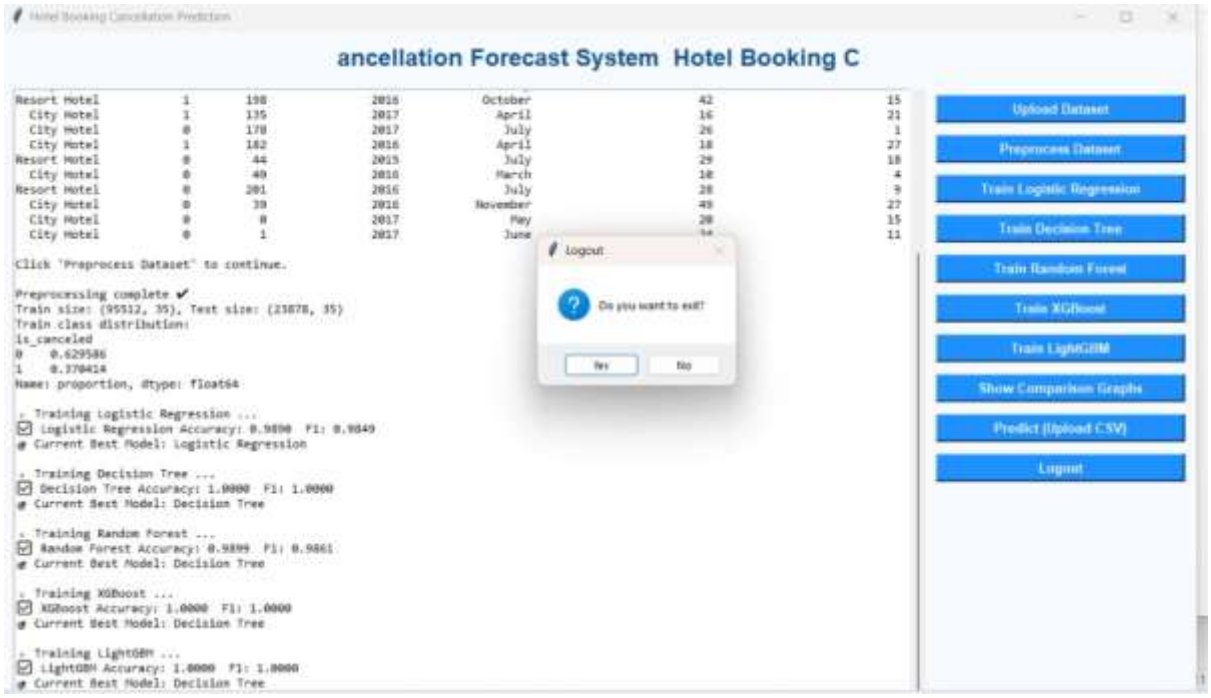
Accuracy comparison graph generated and displayed

Upload your CSV file to view the prediction results.



ny	days_in_waiting_list	customer_type	required_car_parking_spaces	total_of_special_requests	reservation_status	Prediction
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	1	Check-Out	Not Canceled
0	0	Transient	0	1	Check-Out	Not Canceled
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	1	Check-Out	Not Canceled
0	0	Transient	0	1	Canceled	Canceled
0	0	Transient	0	0	Canceled	Canceled
0	0	Transient	0	0	Canceled	Canceled
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	3	Check-Out	Not Canceled
0	0	Transient	0	1	Check-Out	Not Canceled
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	3	Check-Out	Not Canceled
0	0	Contract	0	0	Check-Out	Not Canceled
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	0	Check-Out	Not Canceled
0	0	Transient	0	0	Canceled	Canceled
0	0	Transient	0	0	Canceled	Canceled
0	0	Transient	0	0	Canceled	Canceled
0	0	Group	1	0	Check-Out	Not Canceled
0	0	Group	1	1	Check-Out	Not Canceled
0	0	Transient	0	0	Canceled	Canceled
0	40	Transient	0	0	Canceled	Canceled
0	0	Transient	0	0	Canceled	Canceled
0	0	Transient	0	0	Canceled	Canceled
0	0	Transient	0	0	Canceled	Canceled
0	0	Transient	0	0	Canceled	Canceled

After uploading, the results will be displayed



After that, clicking 'Logout' will display a popup. If you click 'Yes', you will be logged out.

VIII. CONCLUSION

The proposed hotel booking cancellation prediction system demonstrates the effectiveness of machine learning techniques in addressing real-world challenges in the hospitality industry. By leveraging multiple algorithms such as Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM, the system provides accurate and reliable predictions of booking cancellations.

The integration of data preprocessing, feature engineering, and model evaluation ensures high performance and robustness. The use of ensemble and boosting methods significantly improves prediction accuracy compared to traditional approaches. Additionally, the system's user-friendly interface allows users to easily interact with the application, making it accessible to both technical and non-technical users.

One of the key advantages of the system is its ability to provide real-time predictions, enabling hotel management to make proactive decisions. By identifying potential cancellations in advance, hotels can implement strategies such as overbooking, dynamic pricing, and targeted promotions to minimize revenue loss.

The system is designed to be scalable and flexible, allowing the integration of additional features and models in the future. It also supports visualization tools that enhance data analysis and interpretation.

In conclusion, the proposed system offers a practical and efficient solution for hotel booking cancellation prediction. It highlights the importance of machine learning in improving operational efficiency and decision-making in the hospitality industry. Future work may focus on incorporating deep learning models, real-time data processing, and cloud-based deployment to further enhance system performance and scalability.

REFERENCES

1. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," KDD, 2016.
2. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," NIPS, 2017.
3. Breiman, "Random forests," Machine Learning, 2001.
4. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, 1995.
5. Dua and C. Graff, "UCI Machine Learning Repository," 2019.
6. Pedregosa et al., "Scikit-learn: Machine learning in Python," JMLR, 2011.
7. Goodfellow et al., Deep Learning, MIT Press, 2016.
8. Raschka, Python Machine Learning, Packt, 2015.
9. Han et al., Data Mining: Concepts and Techniques, Morgan Kaufmann, 2011.
10. McKinney, "Data structures for statistical computing in Python," SciPy, 2010.
11. Kuhn and K. Johnson, Applied Predictive Modeling, Springer, 2013.
12. Géron, Hands-On Machine Learning with Scikit-Learn, O'Reilly, 2019
13. Seaborn Documentation, "Statistical Data Visualization," 2023
14. Matplotlib Documentation, "Plotting Library," 2023
15. Python Software Foundation, "Python Documentation," 2023.