Hotel Booking Cancellations Prediction Using Machine Learning Techniques for Optimized Management

R R Shantha Spandana[1], M S Arumugam[2], Syed Jeelan[3]

[1]Assistant Professor, Department of MCA, Sri Venkatesa Perumal College of Engineering & Technology, Puttur,
E-mail: shanthaspandana@gmail.com, ORC-ID: https://orcid.org/0009-0003-4236-1250

[2]P.G Scholar, Department of MCA, Sri Venkatesa Perumal College of Engineering & Technology, Puttur,
E-mail: msarumugam.ms@gmail.com, ORC-ID: https://orcid.org/0009-0002-0860-4223

[3]Assistant Professor, Department of CSE, Sri Venkatesa Perumal College of Engineering & Technology, Puttur,
E-mail: syedjeelan1971@gmail.com, ORC-ID: https://orcid.org/0009-0005-8042-9521

To Cite this Article

R R Shantha Spandana, M S Arumugam, Syed Jeelan na, "Hotel Booking Cancellations Prediction Using Machine Learning Techniques for Optimized Management", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 277-286, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp277-286

Submitted: 28-02-2026 Accepted: 01-04-2026 Published: 08-04-2026

Abstract: Reservation cancellations are a massive issue in the hospitality sector as they disrupt the demand forecasts and cost businesses away 20 percent of their revenues. You must be capable of making proper predictions regarding cancellations in order to optimize the operations, price and resource management of your hotel. The dataset that we used was Hotel Booking Demand, with 32 characteristics of the hotel of a resort and city. These characteristics consist of categorical and numerical data regarding customer booking and their type of customer and the specifics of their booking. Preprocessing involved a process of dealing with missing values, eliminating duplicate values and the process of evening out the distribution of classes using the RandomUnderSampler. Some of the classification techniques employed are LR, DT, RF, XGBoost, Gradient Boosting, LGBMClassifier, SVM, and MLP. The hyperparameters were fine-tuned with the help of gridSearchCV. The precision of the forecasts was enhanced through Stacking Classifier where the major model consisted of XGBoost, Random Forest, Gradient Boosting, and Logistic Regression. Categorical features were further encoded as labels to make them even more accurate with the help of Label Encoding and RFE was utilized to identify the most significant factors that influenced cancellations. The Voting Classifier which used a combination of Decision Tree and MLP models increased the accuracy to 98.7%. The importance of each feature was also determined using XAI tools such as LIME and SHAP. Flask-based user interface enabled real time and live prediction to aid good decisions by the hotel management systems.

"Index Terms: *Cancellation forecasting, hotel booking, artificial intelligence, machine learning, revenue management, explainable artificial intelligence".*

1. INTRODUCTION

Tourism has today been one of the main stimuli to the development of the economies in most parts of the world and has led to employment, development of infrastructure, and UNWTO, in 2023, the amount of foreign tourists was approximately 1.3 billion, and it is 88 percent of the pre-pandemic levels. The levels are projected to be exceeded by 2024 [1]. Tourism does not only inject money through tourist expenditure, it also supports numerous other sectors and sustainability of the economy in the long run through maintenance of high demand in hospitality services [2]. Hotel industry is a critical component of this environment as it interrelates the supply and demand with customer experience. However, since hotel rooms are sensitive resources in that they become available soon, and the hotel incurs expenses in cases where the rooms are not used, demand planning is a significant yet challenging undertaking by hotels seeking to remain in business [3].

Although demand forecasting is significant, as a business point of view, it remains difficult to estimate the behavior of customers in the hotel industry. This is compounded by the fact that external elements such as volatile governments, environmental issues and shifting tourist preferences are difficult to forecast hence making accurate predictions difficult [4]. This has been complicated by the process of digitalizing the booking process particularly through online travel sites thus making it easy to change or cancel reservations at short notice. Studies show that

between 10 and 30 percent of hotel reservations are dropped in the end [5]. This has led to a drastic increase in the cancellation rates. Such cancellations may disrupt operational planning, demand change signals and cause up to 20% declines in earnings [6]. Although past studies have focused on general demand forecast, little has been conducted on the individual booking cancellation prediction or the determination of the behavior determinants that make people do such things [7]. Due to this study gap, hotel managers lack the sufficient predictive tools to reduce the operational and financial risks associated with having customers who behave in unpredictable manners. In order to address these issues, this research paper gives a prediction method that equally considers accuracy and practicality. New developments in AI and ML have changed many parts of the hotel industry. The majority of the applications however have been to enhance customer experiences, automate tasks and make service delivery more efficient [8]. Nevertheless, the explainable prediction models have not been well understood yet, particularly when one is trying to predict the cancellation behavior. It is equally important to understand why a model predicts that a booking will be canceled as to make the prediction. This allows for data-driven, clear decisions that are in line with operational goals [9]. This research attempts to bridge that gap by developing a model that does not only predict cancellations with a high degree of accuracy but also provides some valuable information regarding the reasons behind cancellations.

This research contributes three items into the field. First, it contributes to the body of knowledge of hotel demand forecasts by filling the gap in the literature on the necessity to have straightforward ways to predict cancellations on the customer level. Second, it introduces a very good prediction modeling method that uses few features to make predictions. This renders it to be adaptable and scalable to a broad scope of operation conditions. Third, it allows managers to make decisions that are more effective because it provides them with results that are easy to read and demonstrate the factors that lead to cancellations. All these attempts justify the application of AI-based analytics to additional variables of hospitality revenue management. This will keep operations more predictable, transparent and factual, which will keep businesses afloat in a global tourism market that is rapidly evolving [10].

2. LITERATURE REVIEW

The data-driven analytics have become much better recently, which allows predicting the cancellation of hotel bookings much easier. Nevertheless, the questions of the extent to which the predictions could be comprehended and applied in other circumstances are still quite issues. The system was a ML-based predictive hotel cancellation system developed by Herrera et al. [11]. They demonstrated that they were highly accurate in their predictions using ensemble models. Their approach was good to capture non-linear relationships in the booking data but the feature-level explanations were not clear hence making it more difficult to understand by managers. Similarly, Hikmawati et al. [12] applied data mining to search the trends in the number of rooms per night (ADR), cancellations, and the manner in which individuals reserved their rooms. Their findings revealed the role of price trends in influencing cancellations, however, due to the lack of one prediction framework, it could not be applied appropriately in all types of hotels.

The application of AI in the hospitality business is also of interest to more and more researchers. Limna [13] provided an extensive overview of the current application of AI with emphasis on how it can transform the process of automation of services and engage with customers. However, the review was largely on high-level applications and did not discuss some of the issues with prediction modeling that arise with cancellation management. In another similar study, Sekhon and Ahuja [14] examined machine learning approaches to cancellations and discovered that ensemble algorithms such as the Random Forest and Gradient Boosting were significantly better. Although they discussed the effectiveness of models, their studies revealed that AI-based predictive systems remain difficult to comprehend and, thus, it is difficult to make hotel managers understand why predictions are the way they are. Cai et al. [15] investigated the influence of AI on the cognition of customers regarding it and discovered that trust and brand association is a significant factor in the acceptability of AI-enabled systems. This implies that explainable and accurate models are highly significant in influencing people to believe in smart forecasting devices.

Similarly, other works have also examined other related fields of hotel management that employ big data to enhance performance. Based on the analysis of a large amount of transaction data, Gomez-Talal et al. [16] offered a data management tool that companies could utilize to waste fewer foods. Their effectiveness in improving processes at the level indicates how predictive analytics can be used to assist in sustainability and reduction of costs. A model that can be comprehended combining machine learning and statistical feature association in the hotel cancellations area was proposed by Chen et al. [17]. Their approach caused predictions to be more

understandable, yet it was based on much feature engineering detailed, which might complicate scaling and adaptation to booking trends that shift rapidly. Similarly, Nababan et al. [18] examined the potential of using the K-nearest neighbors algorithm and the Synthetic Minority Over-sampling Technique (SMOTE) to correct the case of class imbalance in cancellation data. Their technique had higher recall rates but lower precision, which indicated that accuracy and computational speed do not necessarily equal each other.

Ensemble and meta-learning techniques have proved to be promising to more accurate predictions in other fields than hospitality. Barton and Lennox [19] employed model stacking to enhance the strength of the variable importance in the industrial soft sensor contexts. This demonstrated the usefulness of multi-layer learning systems. In addition, the results by Sahu et al. [20] demonstrated that stacking classifiers are superior to single models in medical diagnostics. This implies that the same techniques may be applied to hotel data to predict the accurate cancellation more effectively. All these studies indicate that ensemble learning is more accurate when making predictions and its application in hotel forecasting that is easy to comprehend remains poorly understood.

Two issues with the already existing research are: (1) there are no frameworks that render the rationale behind prediction of cancellations easy to comprehend, and (2) not many researchers have been conducted on lightweight, adaptive models that can achieve high accuracy without being overly dependent on features. In order to bridge these gaps, this paper proposes a simplified and easy to understand machine learning model that seeks to strike a good balance between the ability to predict and the ability to explain. This paper facilitates the development of intuitive and practical AI-based predictive systems to enable the hotel managers to make proactive decisions using the data by synthesizing the concepts of new ensemble and interpretability-oriented strategies.

3. MATERIALS AND METHODS

The proposed system will be based on highly advanced machine learning and ensemble-based models, which will effectively and effortlessly comprehend the occurrence of hotel reservations being canceled. Hotel Booking Demand data is used to predict the behavior of customers and their reservation patterns based on the 32 categorical and numerical features of the city and resort hotels that comprise it. The approach will have structured pre-processing such as label encoding, data cleaning, and class balancing of random under-sampler to ensure that the training of the model is not biased. The most important predictors of cancellations are found with the help of Recursive Feature Elimination (RFE) that makes features useful. A Stacking Classifier is employed to make it more flexible and reliable and a combination of XGBoost, Random Forest, and Gradient Boosting is taken as base learners with the meta-model being Logistic Regression. The predictions made by a Voting Classifier which is an amalgamation of Decision Tree and Multi-layer Perceptron (MLP) are also more accurate. SHAP and LIME are used to obtain model transparency to interpret AI-based explanations of models. The flask-based interface can be deployed in real-time and hence enable the operations of the hotel to make improved decisions and to manage their income more efficiently.

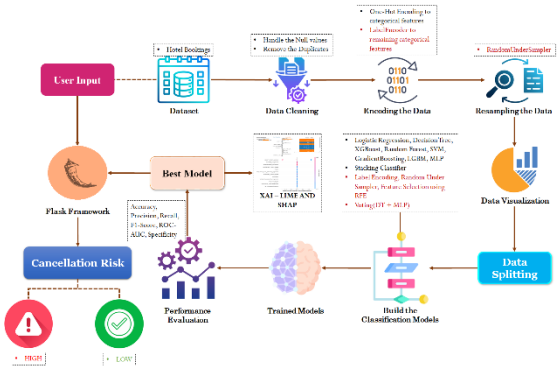


Fig.1 Proposed Architecture

The system architecture displays a machine learning model of making predictions on when hotel bookings are going to be canceled. It takes preprocessed ticket data that has been split, and run through a 5-fold cross-validation with random undersampling to ensure that all the classes are well represented. Most of the models, e.g. Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, and MLP are trained and optimized with hyperparameter tuning. The most effective of the models are assembled into a stacking ensemble where there is meta-model selection. Performance is measured using accuracy, recall, precision, F1-score, sensitivity and AUC and SHAP ensures that the output is understandable and all is transparent.

a) Dataset Collection:

The Hotel Booking Demand data that can be obtained in the Kaggle public repository consists of 119,390 samples with 32 numerical and categorical variables that provide details on the demographics of the customers, their reservations, and booking preferences on both resort and city hotels. The target variable, is cancelled, lets you know whether a ticket was cancelled or not. The information within the dataset is of numerous varieties such as lead time, booking channels, room types, and customer types that demonstrate the difference in the hotel processes in the real world. The dataset is even more complex and helpful to analyze due to small disparities in the data and the absence of values in such characteristics as agent, company, and country. It is ideal to create strong and generalizable machine learning models used to predict cancellations and demand in the hotel industry because of its large size, diversity and richness in time.

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults
0	Resort Hotel	0	342	2015	July	27	1	0	0	2
1	Resort Hotel	0	737	2015	July	27	1	0	0	2
2	Resort Hotel	0	7	2015	July	27	1	0	1	1
3	Resort Hotel	0	13	2015	July	27	1	0	1	1
4	Resort Hotel	0	14	2015	July	27	1	0	2	2

5 rows × 32 columns

Fig.2 Hotel Booking Demand Dataset

b) Pre-Processing:

Preprocessing step ensures that the data is sound and the model is prepared to be used. This is done through cleaning, encoding, and resampling, feature selection, visualization and partitioning of the data in a systematic manner. This renders the subsequent predictive modeling to be more reliable and effective.

i. Data Cleaning: The dataset was also properly cleaned and cleaned using a rigorous data cleaning technique. The all-important attributes such as country, agent, company, and children were filled with appropriate values when they were absent. Duplication of records was discovered and removed to eliminate wastages. This was a critical step that ensured that the data was kept clean and prevented errors during model learning. The subsequent stages of feature extraction, resampling and model training were better when the data was complete and clean.

ii. Data Encoding: The encoding technique was a hybrid one, which transformed categorical variables into a number format that could be used by machine learning systems. Label encoding was used for ordinal or binary categorical variables and one-hot encoding was used for multi-class categorical characteristics. This transformation ensured that every trait was represented in a quantitative manner and not ordinarily. Encoding enabled algorithms to comprehend categorical data with greater ease which enhanced model convergence, readability and accuracy with a broad range of data properties.

iii. Data Resampling: An even sampling method was applied randomly to the dataset to balance the dataset such that the canceled and non-cancellation bookings had no imbalance. This approach reduced the size of the majority group to have an equal representation of both classes. The model was less biased towards the ruling classes as the classes were balanced and this made the model more fair and useful in general. The resampling was a requirement that was used to enhance the classification accuracy and maintain the cancellation prediction constant.

iv. Feature Selection: The feature selection was done in a recursive elimination based technique to identify the most significant predictors of booking cancellation. The twelve main attributes that were reviewed and were retained in the process were critical in determining the effectiveness of the model in predicting the future. The dimensionality of the data was decreased by eliminating the already existing traits or ones that were insignificant. This simplified the calculations and the model. This move ensured that the final learning model was founded on the aspects that produced the greatest distinction.

v. Data Visualization: It was made easier to visualize data using data visualization techniques and determine the balance of the dataset and the cancellation of bookings. Pie and bar charts were created before and after resampling to demonstrate the percentage of canceled plans to those that were not canceled. These visual hints were used to locate cases of imbalance in the classes and ensure that the sampling technique was effective. Visualization made people better comprehend the way the data was spread, make more informed preprocessing decisions, and the results of the subsequent analytical modeling easier to interpret.

c) Training and Testing:

The processed data was divided into training and test data in 80: 20 proportions. The stratified sampling was employed to maintain the same distribution of classes. In this way, the number of canceled and non-canceled bookings in each group stayed the same. The training set assisted the model in learning and the testing set provided a decent preview of the ability of the model to guess. This data partitioning was highly sought after to test model generalization, overfitting and ensure the system performed well with new data that it had never encountered.

d) Algorithms:

A Logistic Regression is a kind of statistical classification algorithm that employs logistic function to approximate the probability of a two-option outcome. It transforms raw features into a probability between 0 and 1 which allows it to classify the difference between classes using a straight-line decision boundary. The interpretability, efficiency, and ability to counter overfitting in the algorithm make it appropriate in the evaluation of the baseline performance. Its capability to approximate the importance of features increases the model transparency and it plays a role in having a clear picture of the decision making processes in predictive analytics.

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(W^T x + b)}} \quad (1)$$

Decision Tree algorithm classifies data into sets by dividing the space of features into similar subsets repeatedly according to the most significant characteristics. Leaf nodes are used to display the outcomes of a class and every internal node represents a decision process. It is easy to comprehend and is effective in both numeric and categorical data using this hierarchical framework. Its non-parametric feature allows it to be able to capture nonlinear relationship, and thus it is much more accurate in classification, yet it also leaves things clear and easy to see in occupations where there is need to make decisions.

$$I\,(i) = 1 - \sum_{i=1}^{k} p_i^2 \quad (2)$$

Extreme Gradient Boosting (XGBoost) is a form of ensemble learning which creates a series of decision trees in such a manner that corrects the errors of the previous trees. It is a gradient-based optimization that eliminates loss functions in the fastest possible time, which ensures enhanced prediction performance. XGBoost takes advantage of parallel processing to make its computations faster and regularization to ensure that models do not fit too well. This enables it to be highly reliable and accurate with large and multidimensional datasets.

RF is a variant of ensemble classifier that involves a combination of more than one decision tree to make the predictions more secure and more precise. It reduces overfitting and enhances generalization when using data that the model has not encountered yet through introducing randomness in feature selection and data sampling. The end classification is determined by each tree and it is resilient to noise and outliers. It is common in challenging classification tasks since it is able to make use of varying kinds of data and determine the most valuable features.

$$Gini = 1 - \sum_{i=1}^{C} (Pi)^2 \quad (3)$$

SVM is a powerful supervised learning algorithm which forms the optimal hyperplane to classify classes with the maximum space. It transforms raw data into high-dimensional spaces that are able to classify nonlinear patterns using the help of kernel functions. The SVM margin-based algorithm reduces the generalization error hence it is effective in both equal and unequal data sets. It is suitable in the case of a high dimensional classification task since it is stable and can be applied on complicated feature distributions.

$$minimize\ \frac{1}{2}\ ||W||^2 + C \sum_{i=1}^{n} \xi_i \quad (4)$$

Gradient Boosting is an ensemble algorithm which creates one model after another. The models that follow the current one minimise the errors of the previous ones by gradient descent optimization. It picks up weak learners such as decision trees and combines them to create a strong forecast model. Gradient Boosting is extremely precise and adaptable as it concentrates on the circumstances that are difficult to forecast. The fact that it is iteratively revised is what makes models more stable and ensures improved performance on both structured and non-structured data.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (5)$$

LGBM Classifier is a high-tech gradient boosting system, which is supposed to be efficient and scalable. It employs a tree growth technique involving searching leaves in a tree where the depth is limited and, as such, more precise and converges faster. LGBM is effective when dealing with large amounts of data with many features, and it is inherently suitable with categorical variables. Since it is small, does not consume much memory, and capable of making predictions in a short time, it is ideal in real-time classification activities that require it to be accurate and consume less computer power.

MLP is a feedforward neural network that employs layers of neurons linked to one another in order to map the input data to the predictions. It learns the use of complicated features though the use of backpropagation-based optimization of weights and nonlinear activation functions. MLP does a good job at discovering the complex relationships in data and is quite effective at generalization and adaptation. It is capable of making a prediction with high accuracy due to its deep learning architecture that enables it to model nonlinear and multidimensional classification problems.

$$\hat{y} = f\left(W^L f\left(W^{L-1} \cdots f(W^1 X + b^1) + b^{(L-1)}\right) + b^L\right) \quad (6)$$

Stacking Classifier is a form of ensemble learning, which involves the use of a meta-model to interpolate predictions of a number of base classifiers to create the general predictions more precise. Estimates in this method are made by different base estimators such as XGBoost, RF and Gradient Boosting. These forecasts are then pooled together using a Logistic Regression meta-model. This hierarchy exploits the strengths of each model and fills the weaknesses. This renders the system more general and dependable as far as classification jobs are concerned.

$$\hat{y} = g(Y_{base}) = g\left(f_1(x), f_2(x), \ldots, f_m(x)\right) \quad (7)$$

Voting Classifier is an ensemble learning algorithm which involves the use of predictions of multiple base models i.e. the Decision Tree and the Multilayer Perceptron and combines the two to come up with improved classifications. In order to make this algorithm more precise and applicable in any situation, a number of preprocessing extensions were incorporated. The categorical features were converted to numbers through the use of Label Encoding, it ensured that the distribution of classes was fair through the use of Random Under-Sampling and it was used to select the most important characteristics through the use of Recursive Feature Elimination (RFE). All these modifications enhanced the efficacy, accuracy and simplicity of the Voting Classifier in many data circumstances.

$$\hat{y} = argmax_c \left(\sum_{i=1}^{n} II\ (\hat{y}_i = c) \right) \quad (8)$$

e) Integration of XAI & Flask Framework:

XAI approaches were introduced to ensure that the model of predicting the cancellation of hotel bookings was more accessible and transparent. LIME was used to come up with local explanations by looking at how each feature affects the model's output by studying feature contributions for each prediction. The value of global features was also figured out using SHAP. This assisted us to know more about the model behavior with respect to various predictions. These approaches make sure that the process of making decisions is readable and credible to the management at the hotel.

The predictive system was configured in a Flask-based web interface. This enabled individuals to have a live interaction with the trained model. With such integration, entering data becomes simple, it is possible to guess quickly whether a reservation is going to be canceled or not and instantly look at the insights that LIME and SHAP are going to produce that clarifies things. Flask is a lightweight, small and fast framework to deploy, which links the machine learning service to end users and allows them to make practical decisions.

4. EXPERIMENTAL RESULTS

Accuracy: The ability of a test to distinguish between the sick and healthy individuals is referred to as its accuracy. To obtain a notion of the accuracy of a test, we need to determine the percentage of true positives and false negative cases. Mathematically this can be expressed as.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} (9)$$

Precision: The ability of a test to distinguish between the sick and healthy individuals is referred to as its accuracy. To obtain a notion of the accuracy of a test, we need to determine the percentage of true positives and false negative cases. Mathematically this can be expressed as.

$$Precision = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} (10)$$

Recall: In ML, the measure of recall is used to indicate the capability of a model to achieve all the significant cases of a particular category. It demonstrates the ability of a model to capture the instances of some class. It is computed as the number of correct positive predictions that are made divided by the number of actual positives.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} (11)$$

F1-Score: F1 score is a method of measuring the accuracy of ML model. It sums up the accuracy and recall scores of a model. Accuracy measure is a measure that counts the number of times when a model made a correct guess over the entire data set.

$$F1\ Score = 2 * \frac{Recall\ \mathrm{X}\ Precision}{Recall + Precision} * 100 (12)$$

AUC-ROC Curve: The AUC-ROC Curve demonstrates the effectiveness of a classification problem at the various benchmark levels. The True Positive Rate is plotted against the False Positive Rate by ROC. auc: The AUC is used to measure the ability of the model to distinguish among the classes; the larger the AUC the better the model.

$$AUC = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2} \quad (13)$$

Specificity: It is determined by the number of people who are found negative to a disease divided by the number of all people who are not affected by the disease. This is the people who were tested negative as well as the people who were tested positive but did not have the disease.

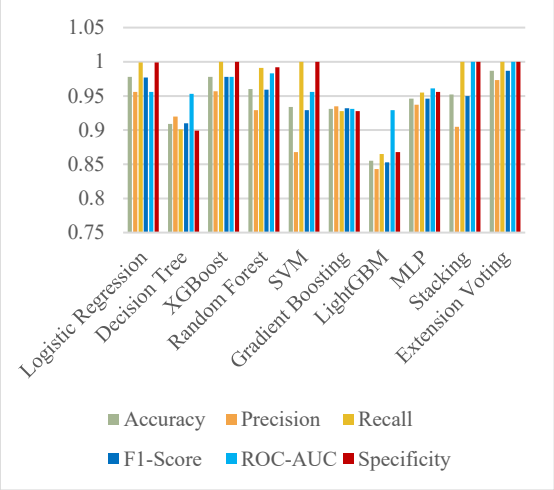
$$Specificity = \frac{TN}{(TN + FP)} \quad (14)$$

Table.1 Performance Evaluation Table

ML Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Specificity
Logistic Regression	0.978	0.956	0.999	0.977	0.956	0.999
Decision Tree	0.909	0.920	0.901	0.910	0.953	0.899
XGBoost	0.978	0.957	1.000	0.978	0.978	1.000
Random Forest	0.960	0.929	0.991	0.959	0.983	0.992
SVM	0.934	0.868	1.000	0.929	0.956	1.000
Gradient Boosting	0.931	0.935	0.928	0.932	0.931	0.928
LightGBM	0.855	0.843	0.865	0.853	0.929	0.868
MLP	0.946	0.937	0.955	0.946	0.961	0.956
Stacking	0.952	0.905	1.000	0.950	1.000	1.000
Extension Voting	**0.987**	**0.973**	**1.000**	**0.987**	**1.000**	**1.000**

Table.1 indicates the performance of various classification models. The longest Voting Classifier was the most accurate and strong in terms of prediction.

Fig.3 Comparison Graph



The graph that compares ability.1 shows that the extended Voting model regularly does better than all other machine learning classifiers in terms of accuracy, recall, and F1-score.

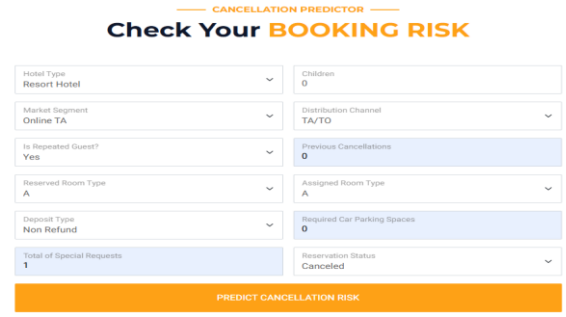


Fig.4 Enter Input Data – 1

Fig.4 is used to insert details of bookings using a web interface so that the user can have an idea of the number of hotel reservations that could be canceled.

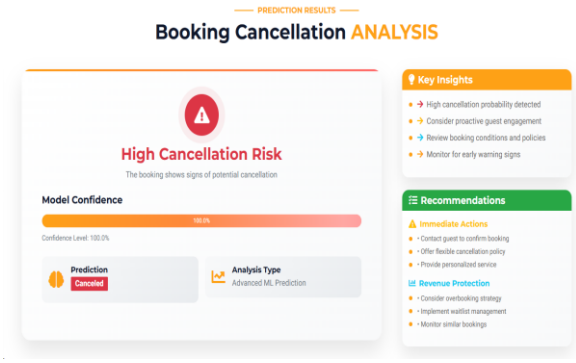


Fig.5 Booking Cancellation Analysis – 1

The result of the estimate of the given input values is given in figure 5 which indicates a "High Cancellation Risk."

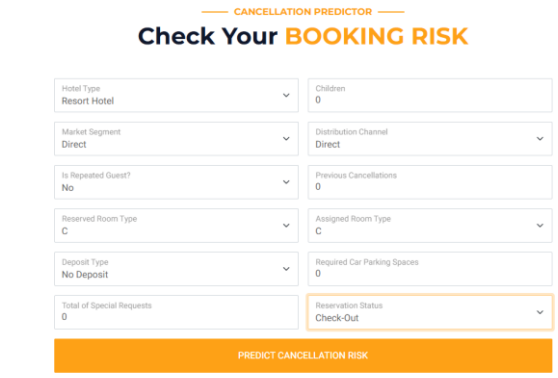


Fig.6 Enter Input Data – 2

The result of the estimate of the given input values is given in figure 5 which indicates a "High Cancellation Risk."

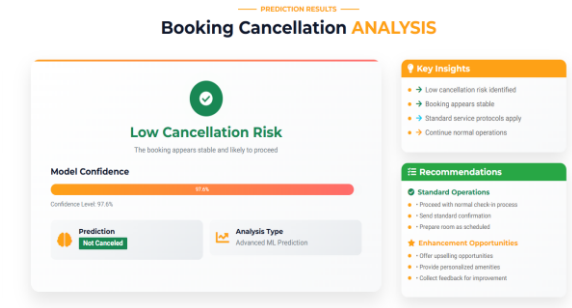


Fig.7 Booking Cancellation Analysis – 2

In Fig.7, the system provides a projection of the information that was typed in to depict a Low Cancellation Risk.

5. CONCLUSION

Lastly, the designed system is good at addressing the issue of hotel reservations canceled with the help of Hotel Booking Demand dataset. In this dataset, there are 32 numerical and categorical variables of both tourist and city hotels. The quality and balance of the data was also made possible through careful preparation which involved working with missing values, eliminating duplication, and the RandomUnderSampler. To apply and optimize several machine learning models, such as Logistic Regression, Decision Tree, Random Forest, XGBoost, Gradient Boosting, LGBMClassifier, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP), gridSearchCV was applied. This was in order to identify optimal hyperparameters. The Stacking Classifier that was a combination of XGBoost, Random Forest, Gradient Boosting, and Logistic Regression as the lead model was quite good at prediction. Categorical features were encoded using Label Encoding to enhance the performance further and Recursive Feature Elimination (RFE) was employed to locate the most significant factors that influenced cancellations. Ensemble learning could be quite useful as the Voting Classifier that combined Decision Tree and MLP was highly accurate (98.7%). Moreover, Explainable AI (XAI) software such as LIME and SHAP simplified the data since it indicated the significance of each feature. The Flask-based interface allowed the users to make predictions and interact with the system in real time and is therefore convenient in making quick and data-driven decisions in hotel management.

This project should be continued by including real-time booking systems and dynamic price platforms in the predictive framework to enable the framework to make more automatic decisions in the hotel management. It is possible to add more behavioral, financial and seasonal factors so that the model becomes even more efficient in adjusting to the market changes. Researchers can consider hybrid deep learning systems in the future to prevent strategic cancellation and temporal pattern recognition through reinforcement learning. By deploying the model into the cloud, it can also be possible to enable the model to continuously learn on a large scale based on live streams of data. It will become even more helpful in data-driven operations of hospitality by adding supplementary visualization tools that will simplify the understanding process and provide users with the opportunity to work with it.

REFERENCES

[1] Lakshmi, J. M., Prasad, K. K., & Viswanath, G. (2025). Proactive Security in Multi-Cloud Environments: A Blockchain Integrated Real-Time Anomaly Detection and Mitigation Framework. Cuestiones De Fisioterapia, 54(2), 392–417.

[2] Luo, Z. (2025, August). Hotel Cancellation Rate Prediction: A Machine Learning Based Prediction Model. In 2025 3rd International Conference on Image, Algorithms, and Artificial Intelligence (ICIAAI 2025) (pp. 318-327). Atlantis Press.

[3] Ganesh, B. R., B M, P., Prasad K, K., Swapna, G., & G, Viswanath. (2025). Data Mining-Driven Multi-Feature Selection for Chronic Disease Forecasting. Journal of Neonatal Surgery, 14(5S), 108–124. https://doi.org/10.52783/jns.v14.1993

[4] Alkan, T. A. H. A. (2025). A Comparative Study of Machine Learning and Deep Learning Approaches For Hotel Booking Cancellation Prediction. International Journal Of Scientific Research In Engineering & Technology, 5(03).

[5] Bhardwaj, A., Yadav, T., & Chaudhary, R. (2024, June). Predicting Hotel Booking Cancellations using Machine Learning Techniques. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[6] 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.

[7] Y. Y. Febrian, D. R. Wijaya, and E. Ervina, "Hotel reservation cancellation prediction using boosting model," Proc. 2nd Int. Conf. Softw. Eng. Inf. Technol. (ICoSEIT), Feb. 2024, pp. 138–143.

[8] M. Yoo, A. K. Singh, and N. Loewy, "Predicting hotel booking cancelation with machine learning techniques," J. Hospitality Tourism Technol., vol. 15, no. 1, pp. 54–69, Jan. 2024.

[9] H. Lv, S. Shi, and D. Gursoy, "A look back and a leap forward: A review and synthesis of big data and AI literature in hospitality and tourism," J. Hospitality Marketing Manage., vol. 31, no. 2, pp. 145–175, Feb. 2022.

[10] Z. Liu, L. Zhang, W. Wang, and S. Chen, "Hospitality order cancellation prediction from a profit-driven perspective," Int. J. Contemp. Hospitality Manage., vol. 35, no. 6, pp. 2084–2112, May 2023.

[11] A. Herrera, P. Morales, and L. Torres, "Forecasting hotel cancellations through machine learning," Expert Syst., vol. 41, no. 9, p. 13608, Sep. 2024.

[12] N. K. Hikmawati, Y. Ramdhani, and W. Wartika, "Exploring ADR trends: A data mining approach to hotel room pricing, cancellations, and EDA," J. Appl. Data Sci., vol. 5, no. 1, pp. 189–202, 2024.
[13] P. Limna, "Artificial intelligence (AI) in the hospitality industry: A review article," Int. J. Comput. Sci. Res., vol. 7, pp. 1306–1317, Jan. 2023.
[14] Korapati, Dolasankar., Viswanath, G., G, Prathyusha. (2023). A Real-Time Video Based Vehicle Classification, Detection And Counting System. Industrial Engineering Journal, 52(9), 474–480.
[15] R. Cai, L. N. Cain, and H. Jeon, "Customers' perceptions of hotel AI-enabled voice assistants: Does brand matter?," Int. J. Contemp. Hospitality Manage., vol. 34, no. 8, pp. 2807–2831, Jul. 2022.
[16] J. Gómez-Talal, M. Alvarez, and C. Lopez, "Avoiding food waste from restaurant tickets: A big data management tool," J. Hospitality Tourism Technol., vol. 15, no. 2, pp. 232–253, Mar. 2024.
[17] S. Chen, X. Liu, J. Ma, and H. Zhang, "Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction," Decis. Support Syst., vol. 170, Jul. 2023, Art. no. 113959.
[18] A. A. Nababan, M. Jannah, and A. H. Nababan, "Prediction of hotel booking cancellation using K-nearest neighbors (K-NN) and SMOTE," INFOKUM, vol. 10, no. 3, pp. 50–56, 2022.
[19] G Loge, T Sunil Kumar Reddy, G Swapna, & G Viswanath. (2025). Interpretable AI for Precision Brain Tumor Prognosis: A Transparent Machine Learning Approach. In International Journal of Health Sciences and Pharmacy (IJHSP) (Vol. 9, Number 1, pp. 180–195). Zenodo. https://doi.org/10.5281/zenodo.15523628
[20] B. Sahu, R. Pradhan, and A. Singh, "Performance analysis of state-of-the-art classifiers and stack ensemble model for liver disease diagnosis," Springer, 2022.
[21] A. Chatzimparmpas, R. Martins, and A. Kerren, "StackGenVis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics," IEEE Trans. Vis. Comput. Graphics, vol. 27, no. 2, pp. 1547–1557, 2020.
[22] H. Duan, "Computational methods for estimating global sensitivity indices Shapley values," Ph.D. dissertation, Florida State Univ., 2023.
[23] S. M. Lundberg, G. Erion, and S. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Mach. Intell., vol. 2, no. 1, pp. 56–67, Jan. 2020.
[24] A. Rai, "Explainable AI: From black box to glass box," J. Acad. Marketing Sci., vol. 48, no. 1, pp. 137–141, Jan. 2020.
[25] Z. A. Andriawan, T. W. Purboyo, and M. Z. Arifin, "Prediction of hotel booking cancellation using CRISP-DM," Proc. 4th Int. Conf. Informat. Comput. Sci. (ICICoS), Nov. 2020, pp. 1–6.
[26] G. Chen, X. Liu, and D. Zhang, "Attending to customer attention: A novel deep learning method for leveraging multimodal online reviews to enhance sales prediction," Inf. Syst. Res., vol. 35, no. 2, pp. 829–849, Jun. 2024.
[27] 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.
[28] P. Limna, "Artificial intelligence (AI) in the hospitality industry," Int. J. Comput. Sci. Res., vol. 7, pp. 1306–1317, 2023.
[29] World Tourism Organization (UNWTO), World Tourism Organization–UNWTO, Madrid, Spain, 2023.
[30] M. Li, X. Zhou, and Y. Wang, "A systematic review of AI technology-based service encounters: Implications for hospitality and tourism operations," Int. J. Hospitality Manage., vol. 95, May 2021, Art. no. 102930.