APPLICATION OF MACHINE LEARNING MODEL IN PREDICTING HOTEL ROOM CANCELLATIONS

Nguyen Thi Thu Thuy[1,*], Dang Thi Hong Ha[1]

DOI: https://doi.org/10.57001/huih5804.2025.400

ABSTRACT

This study investigates the application of machine learning models to predict booking cancellations in the hospitality industry. Driven by the urgent need to enhance operational efficiency and optimize revenue - especially given that cancellation rates range from 20% to 40%-the study focuses on developing models to forecast cancellation behavior to support managerial decision-making. Three machine learning models are implemented: Decision Tree (DT), Random Forest (RF), and Gradient Boosting (XGBoost). The data processing steps include cleaning, encoding categorical variables, balancing the dataset using SMOTE (Synthetic Minority Over-sampling Technique), and evaluating model performance through metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC, combined with k-fold cross-validation. The dataset was collected from two hotels in Portugal during the period 2015 - 2017, reflecting a wide range of customer behaviors from both a city hotel and a resort hotel. The results indicate that factors such as lead time, deposit type, and the number of special requests significantly impact the likelihood of cancellation. Customers who book through Online Travel Agencies (OTAs), with long waiting periods and flexible policies, are identified as the most likely to cancel. The study contributes an effective tool for analyzing customer behavior, thereby proposing solutions such as dynamic pricing, prioritizing low-risk customers, and optimizing the booking process to improve management effectiveness in the hotel industry.

Keywords: *Booking cancellation, machine learning, XGBoost, hotel industry, prediction.*

[1]School of Economics, Hanoi University of Industry, Vietnam

[*]Email: nguyenthithuthuy@haui.edu.vn

Received: 05/8/2025

Revised: 12/11/2025

Accepted: 28/11/2025

1. INTRODUCTION

As the tourism and hospitality industry recovers and grows rapidly after the pandemic, hotels are increasingly facing the need to optimize revenue and manage operations effectively. One of the biggest challenges that accommodation businesses face is customer cancellations - a factor that can lead to wasted resources, reduced occupancy rates and severely impact revenue. Globally, cancellation rates reached a high of up to 40% in 2018 [1], mainly due to flexible cancellation policies of OTA (Online Travel Agency) channels such as Booking.com or Expedia. Although this rate has gradually decreased in recent years thanks to more stable post-pandemic customer sentiment and hotel adjustments, fluctuations still exist, especially in urban hotels or hotels that use OTAs as their main distribution channels [2]. Cancellations are not only different by geographic region but are also influenced by booking patterns, hotel types, and new consumer behavior in the digital age.

In the academic field, many international studies have demonstrated the effectiveness of applying machine learning models such as Decision Tree (DT), Random Forest (RF), or Artificial Neural Network (ANN) in predicting the possibility of cancellations, supporting hotels to proactively adjust pricing policies, room allocation, and revenue strategy planning [3-5]. However, in Vietnam, the application of machine learning in hotel management - especially in predicting cancellation behavior - is still relatively new, mainly limited to traditional methods such as linear regression or basic statistical analysis, not taking advantage of the power of big data and modern algorithms.

Based on that reality, the article is developed with the goal of building a highly accurate prediction model, helping hotels to identify early cancellation risks, thereby making appropriate adjustment decisions to optimize operations, improve management efficiency and improve customer experience. Not only contributing to filling the research gap in the domestic market, the topic also expands the potential for practical application in revenue management and hotel business strategy in

Vietnam during the post-pandemic recovery period and adapting to digital transformation.

2. RESEARCH METHODOLOGY

2.1. Research model

In this study, the authors apply three advanced machine learning models: Decision Tree (DT), Random Forest (RF), and XGBoost to predict the probability of hotel cancellation. Each model is selected because of its own strengths in the ability to analyze, generalize, and evaluate the relationship between multiple factors with behavioral decisions.

Decision Tree Model

Decision Tree is one of the most intuitive and simple algorithms used in classification and regression [6]. Decision Tree builds a tree with a structure consisting of nodes representing conditions, branches representing the results of the conditions, and leaf nodes representing the final conclusions. The tree construction algorithm is based on criteria such as Entropy, Information Gain, Gain Ratio or Gini Index to split the data set in the direction of maximizing purity in the subgroups [6] .

In this study, the decision tree helps analyze the influence of factors such as advance booking time, booking channel, deposit type, number of days of stay, etc. on the decision to cancel a room. This model helps to clearly define the decision rules based on the input features.

Random Forest Model

Random Forest is a machine learning technique in the ensemble learning group, proposed by Leo Breiman. This model builds many decision trees on random subsets from the original data set (bootstrap sampling), and predicts the results based on the majority vote (for classification) or the predicted average (for regression) [7].

In this study, RF is applied to find out the contribution of each factor to the probability of cancellation. RF has better overfitting resistance than single DT because of the diversity in the tree set. In addition, RF allows to evaluate the importance of input variables, support decision making and optimize pricing policies.

XGBoost Model

XGBoost (Extreme Gradient Boosting) is an optimized version of Gradient Boosting that maximizes speed and performance, developed by Tianqi Chen and Carlos Guestrin [8]. Unlike Random Forest, the trees in XGBoost are built sequentially, each tree corrects the errors of the previous tree, focusing on the wrongly predicted data points [9].

XGBoost was chosen in the study due to its outstanding strengths such as: the ability to handle large data, flexible model adjustment, high accuracy and limited overfitting. In the study, XGBoost was trained on a processed and simplified dataset, with hyperparameters such as learning rate, max_depth, subsample... fine-tuned through grid search.

Combining the above three models, the study compared the performance through the Accuracy, Precision, Recall, F1-Score, and AUC-ROC indexes. This result helps to propose optimal models and decision support strategies for hotel businesses.

2.2. Research data

In this study, the authors used a dataset of 119,390 records, recording booking information at two main types of hotels: City Hotels (66%) and Resort Hotels (34%). The data was collected from the hotel management system, including diverse information about customers, booking history, transaction types and consumer behaviors related to the stay. The target variable in the study is the cancellation status, represented in binary form with the value 0 being no cancellation and 1 being a cancellation. The initial cancellation rate was about 37%, indicating an imbalance between the two data layers and increasing the complexity of the forecast.

In Table 1, the original dataset has 36 columns. Through the review process, columns that have no analytical value or are likely to cause information leakage were removed to increase accuracy and protect user privacy. Columns containing information after the booking time such as booking status and booking status date were removed to avoid data leakage. Personal identification fields such as full name, email, phone number and credit card number were also removed to secure data. The booking company code column was removed because it had up to 94% missing values. After this step, there were 29 input variables that could predict cancellation behavior.

Missing values in the dataset were handled by filling in the average or common value. Specifically, the number of children was filled in with an average value of about 0.1; the booking agency code was filled in with 0, representing non-agent bookings; The customer country is filled with the most common value of Portugal (PRT),

Table 1. Research data

No	Variable name	Variable type	Data type	Meaning
1	Hotel	Categorical	Categorical	Distinguish City Hotel and Resort Hotel
2	is_canceled (Objective)	Dependent	Binary (0/1)	Indicates whether the customer canceled the booking or not
3	advance booking time	Continuous	Real	Time from booking to check-in date
4	arrival_date_year	Categorical	Integer	Year in which the guest stayed
5	arrival_date_month	Categorical	Categorical	Customer arrival month (January - December)
6	arrival_date_week_number	Categorical	Integer	Week number of the year when the guest arrives
7	arrival_date_day_of_month	Categorical	Integer	Specific date of arrival
8	stays_in_weekend_nights	Continuous	Integer	Number of nights guests stay on weekends (Saturday, Sunday)
9	stays_in_week_nights	Continuous	Integer	Number of nights guests stay on weekdays (Monday - Friday)
10	Adults	Continuous	Integer	Number of adults travelling in the booking
11	Children	Continuous	Integer	Number of children in booking
12	Babies	Continuous	Integer	Number of children under 2 years old
13	meal	Categorical	Categorical	Food service packages included (BB, HB, FB, SC...)
14	country	Categorical	Categorical	Customer nationality
15	market_segment	Categorical	Categorical	Where do customers come from (Online, Offline, Corporate...)
16	distribution_channel	Categorical	Categorical	Booking channels such as direct, through agents, OTA, GDS...
17	is_repeated_guest	Binary	0/1	Indicates whether this guest is a regular or not.
18	previous_cancellations	Continuous	Integer	Past Cancellation History
19	previous_bookings_not_canceled	Continuous	Integer	Number of previous bookings without cancellation
20	reserved_room_type	Categorical	Categorical	Original room type code (A, B, C...)
21	assigned_room_type	Categorical	Categorical	Actual room type code received by the guest
22	booking_changes	Continuous	Integer	Number of times the guest changed information during the booking process
23	deposit_type	Categorical	Categorical	No deposit, refundable deposit, non-refundable deposit
24	agent,company	Categorical	Categorical	Agent or service provider code
25	days_in_waiting_list	Continuous	Integer	Number of days the booking was on the waiting list for confirmation
26	customer_type	Categorical	Categorical	Individual, contract, group, walk-in guests
27	Average price per day (ADR)	Continuous	Real	Amount paid by the guest per night of stay
28	required_car_parking_spaces	Continuous	Integer	Number of parking spaces requested by the guest
29	total_of_special_requests	Continuous	Integer	Number of special requests such as baby beds, low floors, non-smoking...

accounting for about 40% [9]. After processing, the dataset has no missing values. Outliers in numeric columns such as booking time, average daily price, and number of nights are handled using the IQR (Interquartile Range) method. For example, the average daily price can initially be more than 5,000 USD, which is limited to 300 USD to reduce the impact of outliers. Unreasonable negative values in columns such as number of adults, number of children, number of babies, average daily price are also replaced with 0 to ensure reasonableness.

To prepare the machine learning model, the data is processed by one-hot encoding for unordered categorical variables such as hotel type, meal type, deposit type, and customer type. These variables are split into multiple binary columns, increasing the number of variables to about 50 columns. Label encoding is applied to ordered or multi-valued categorical variables such as guest arrival month, booked room type, actual room type assigned, and guest country [3]. Numeric variables such as advance booking time, average daily price, days on the

waiting list, and weeknight and weekend nights are normalized using StandardScaler, bringing the values to a mean of 0 and a standard deviation of 1, which improves training performance and uniformizes the scale.

Since the non-cancellation rate is 63% and the cancellation rate is only 37%, the authors apply the SMOTE (Synthetic Minority Over-sampling Technique) technique to generate additional records belonging to the minority class by interpolating between neighboring points. After applying SMOTE, the total number of records increases to about 150,332, and the ratio between the two classes is balanced at 50% - 50%. This balancing helps the model avoid bias and improve classification accuracy.

The post-processed data was divided into a training set of 80% (about 120,266 records) and a test set of 20% (about 30,066 records). The data division process used the stratify parameter to maintain the ratio between the two classes (cancelled and non-cancelled) in both sets, and used random_state = 42 to ensure that the results could be reproduced when retraining the model.

Thus, the study carried out a thorough data processing process from cleaning, handling missing and outliers, encoding variables, normalization, to class balancing and data division to best prepare for building models to predict hotel cancellations.

3. RESEARCH RESULTS AND DISCUSSION

3.1. Decision Tree model results

Table 2. Decision Tree model training results

Class	Precision	Recall	F1-score	Support
Do not cancel (0.0)	0.87	0.88	0.88	15.034
Cancel (1.0)	0.88	0.87	0.88	15.033

Population	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-score	Total number of samples
	0.88	0.88	0.88	0.88	30.067

The decision tree model was trained with optimal parameters from GridSearchCV including a maximum depth of 15 and a minimum number of samples for branching of 10. After training, the model was tested on a dataset of 30,067 records. The results of decision tree mode (Table 2) showed that the model achieved 88% accuracy, Precision and Recall both reached 88% with an average F1-score of 0.88. The model also achieved an AUC-ROC score of 0.95, reflecting the ability to discriminate well between the two groups of canceled and non-cancelled bookings. Although the overall performance was very positive, the model still showed a slight overfitting phenomenon due to the difference between the accuracy on the training set (92%) and the test set (88%).

The model identified several variables that had a significant impact on cancellation behavior, including long advance booking times (over 150 days), non-refundable policies, and exceptionally low request volumes. Clear cutoffs such as advance booking times greater than 300 days increased the likelihood of cancellation by 90%, or if guests chose a non-refundable policy, the cancellation rate increased to around 85%.

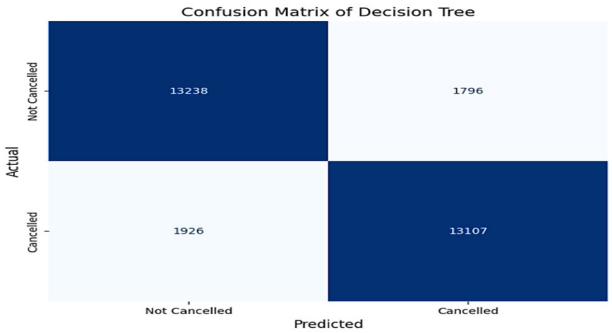

Figure 1. Confusion Matrix of Decision Tree

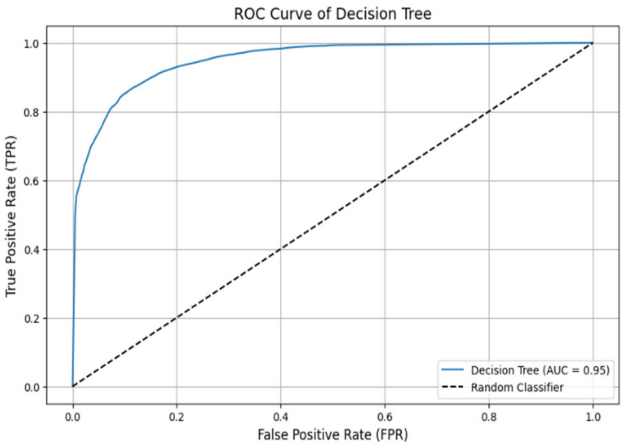

Figure 2. ROC curve of Decision Tree

However, some errors are still notable. According to Figure 1, there are about 1,796 records that are misclassified as cancelled (False Positive), accounting for about 11.9%, and 1,926 records that are misclassified as not cancelled (False Negative), accounting for about 12.8%. These errors often fall into intermediate cases, such as bookings with a 100 - 150 day lead time or no special requirements. The model has the advantage of being easy to interpret and fast to train (~25 seconds on

120,265 records), making it suitable for experimental deployment.

From an application perspective and ROC curve of Decision tree (Figure 2), hotels can use the model results to come up with policies such as limiting advance booking times, considering more flexible refund policies, and encouraging guests to make special requests to reduce cancellation rates.

3.2. Random Forest Model Results

The random forest model was trained with optimal parameters of 200 trees, a maximum depth of 15, and the feature selection at each node was the square root of the total number of fields. The data was class-balanced using the SMOTE technique before training. On the test set of 30,067 records, the model achieved Accuracy 89%, Precision and Recall both reached 89%, the average F1-score was 0.89, and the AUC-ROC was 0.96 (Table 3).

Table 3. XGBoost model training results

Class	Precision	Recall	F1-score	Support
Do not cancel (0.0)	0.88	0.91	0.89	15.034
Cancel (1.0)	0.91	0.87	0.89	15.033

Population	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-score	Total number of samples
	0.89	0.89	0.89	0.89	30.067

Compared to the decision tree, the model significantly reduces Type I errors (to 1,346 records) and maintains Type II errors at 1,926 records. The top important fields identified are booking time (~30%), non-refundable policy (~25%), and number of special requests (~15%). The model also shows confidence in its classification, with predicted probabilities strongly concentrated near 0 and 1 (Figure 4).

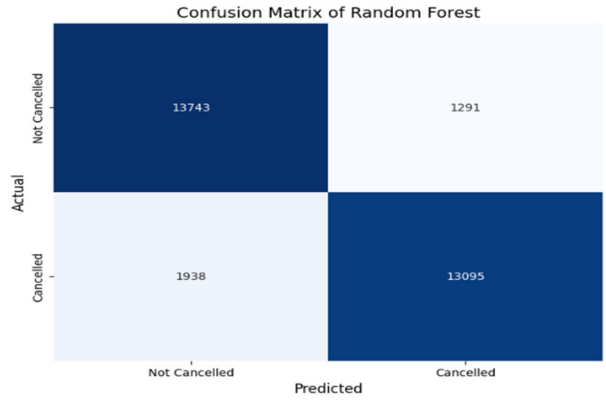

Figure 3. Confusion matrix of Random Forest

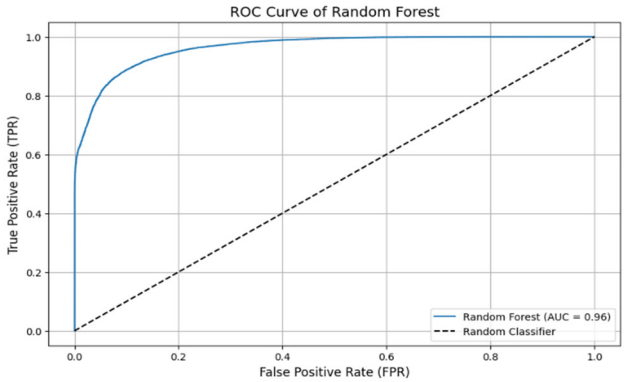

Figure 4. ROC curve of Random Forest

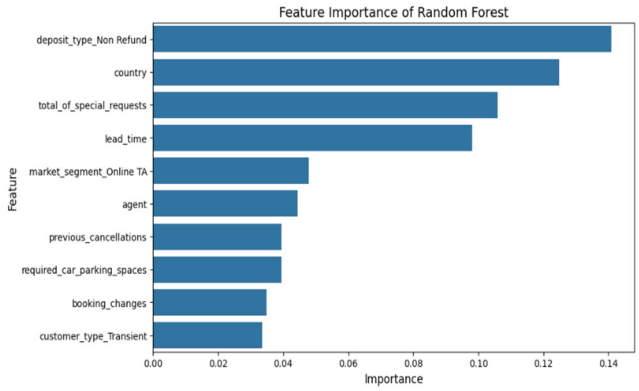

Figure 5. Random Forest feature importance graph

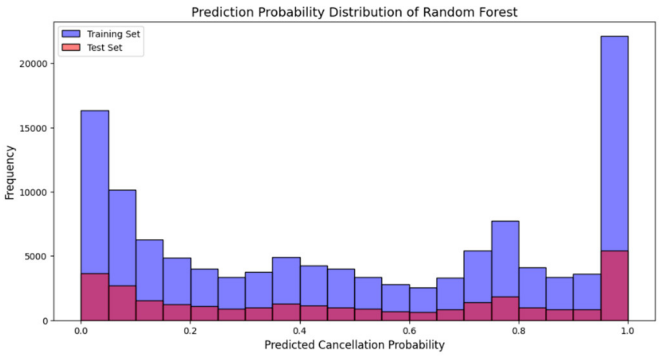

Figure 6. Prediction probability distribution plot of Random Forest

Despite the longer training time (~150 seconds), the model has better generalization ability, suitable for practical deployment. The error cases still occur frequently in the medium booking period group (100 - 150 days) and there are no special requirements, indicating that further improvement is needed in handling these cases (Figure 5, 6).

3.3. XGBoost model results

The XGBoost model was trained with optimal parameters including learning rate 0.1, maximum depth 7 and 200 trees. The data was also processed for class balance using SMOTE. On the test set of 30,067 records, the model achieved Accuracy 90%, Precision 90%, Recall

90%, average F1-score of 0.90 and AUC-ROC score up to 0.97, the highest among the three models (Table 4).

Table 4. Training results table of XGBoost model

Class	Precision	Recall	F1-score	Support
Do not cancel (0.0)	0.90	0.91	0.91	15.034
cancel (1.0)	0.91	0.90	0.90	15.033

Population	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-score	Number of samples
	0.90	0.90	0.90	0.90	30.067

The model correctly predicted 26,987 records, resulting in only 1,311 Type I errors and 1,569 Type II errors, a reduction of ~27% and ~18%, respectively, compared to the decision tree model. The fields with the highest importance were the non-refundable policy (~50%), the number of parking spaces required (~15%), and the online booking channel (~10%). Similar to Random Forest, XGBoost also demonstrated high confidence in classification with a probability distribution concentrated at both ends (Figure 7, 8).

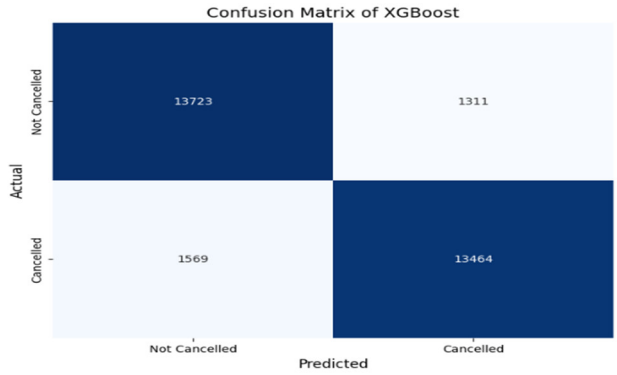

Figure 7. Confusion Matrix of XGBoost

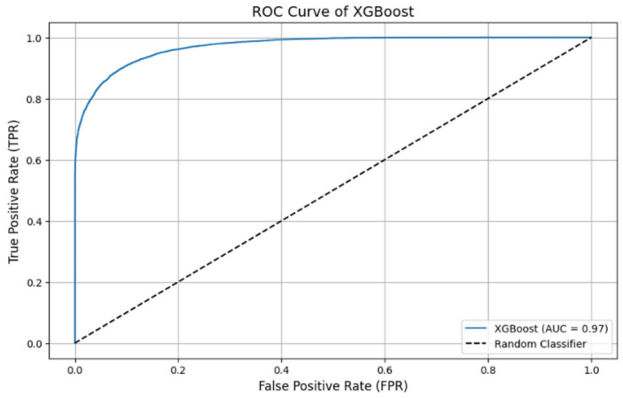

Figure 8. ROC curve of XGBoost

Despite the high performance, the model has the longest training time (~200 seconds) and still struggles with intermediate cases like 100 - 150 days in advance and no special requirements. Type II errors are reduced but still account for 10.4%, which can impact revenue if left unchecked according to Figure 9 and Figure 10.

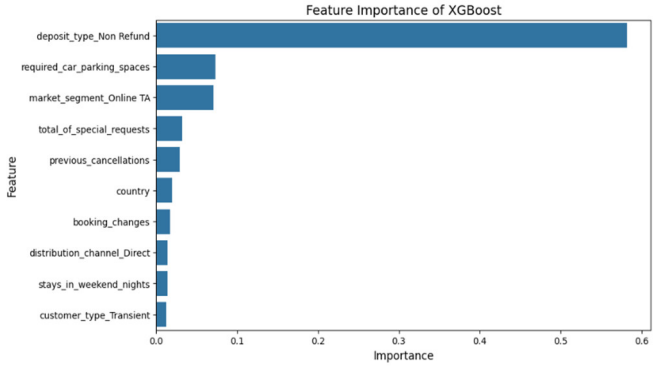

Figure 9. XGBoost Prediction Probability Distribution Histogram

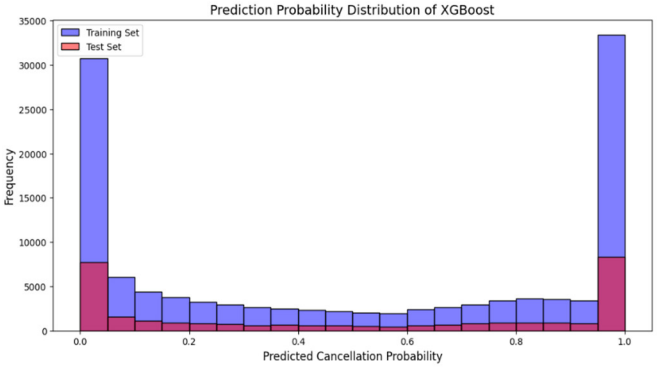

Figure 10. XGBoost Feature Importance Histogram

Overall, XGBoost is the best performing model, suitable for deployment in real-time cancellation prediction systems. Information such as deposit policy, advance booking time, and customer behavior can be exploited to optimize revenue and manage risk more effectively.

4. CONCLUSION AND RECOMMENDATIONS

The application of machine learning models, especially XGBoost, in hotel reservation management has proven to be highly effective in accurately classifying reservations with a high risk of cancellation. With an AUC-ROC of 0.97, XGBoost showed a clear ability to differentiate between two classes of cancelled and non-cancelled reservations, while significantly reducing type I and type II errors compared to other models. This opens up the potential for integrating the model into the reservation management system (PMS) to provide early warning of risky reservations, thereby providing timely responses such as holding offers or contacting customers for confirmation. In addition, the model also supports optimizing room allocation during peak seasons by

identifying more stable reservations, helping to minimize the room vacancy rate due to unexpected cancellations.

Additionally, the results from XGBoost suggest that it is necessary to improve booking policies to reduce cancellation rates [13]. Specifically, the current non-refundable policy leads to cancellation rates of up to 85%. Testing a flexible refund policy, such as offering 50 - 70% refunds for early cancellations, could significantly reduce cancellation rates and increase customer satisfaction and loyalty. Similarly, guests who book too far in advance (more than 150 days) are more likely to cancel; therefore, hotels should design incentives for bookings within 90 days to encourage early commitment and reduce cancellations [10].

Another important factor influencing cancellation rates is the customer experience. The model shows that guests with special requests or parking needs are less likely to cancel. Therefore, hotels should increase service personalization, expand care programs and packages for guests with special requests, especially during peak seasons [7]. At the same time, upgrading the online booking experience and optimizing the website with exclusive offers also helps increase direct bookings, reducing dependence on online agents that have higher cancellation rates.

Although the current models have achieved good performance, the article also proposes expanding the application of other machine learning algorithms such as LightGBM, CatBoost and artificial neural networks to find a more optimal model, especially when deployed on a large data scale or in complex conditions [11]. In parallel, expanding the data range from multiple hotels and integrating additional external data fields such as weather conditions, competitive room rates or real-time customer behavior data will help the prediction model be more flexible and increase the ability to generalize.

From an application perspective, hotels should focus on improving data collection and processing capabilities, integrating the prediction system into the current PMS, and training staff to understand and effectively exploit the warnings given by the model [12, 14]. Combining real-time data and intelligent management policies based on predictive models will be the key to helping hotels optimize operations, reduce cancellation rates and increase sustainable profits in the context of an increasingly competitive tourism market.

REFERENCES

[1]. Ampountolas A., "Predicting hotel booking cancellations: A comprehensive machine learning approach," *Journal of Revenue and Pricing Management*, 1-12, 2025.

[2]. Antonio N., De Almeida A., Nunes L., "Predicting hotel booking cancellations to decrease uncertainty and increase revenue," *Tourism & Management Studies*, 13(2), 25-39, 2017.

[3]. Breiman L., Jerome F., Richard A. O., Charles J. S., *Classification and regression trees*. Chapman and Hall/CRC, 2017.

[4]. Brownlee, J., *Deep learning with Python: Develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery, 2016.

[5]. Chen T., Guestrin C., XGBoost: a scalable tree boosting system supplementary material, International Conference on Knowledge Discovery and Data Mining," in *Proceedings of the 22nd ACM SIGKDD*, New York, 785-794, 2016.

[6]. Esther Hertzfeld, *Study: Canceellation rate at 40% as OTAs push free change policy*, 2019. https://www.hotelmanagement.net/tech/curator-hotel-resort-collection-partners-hovr-video-storytelling

[7]. Garreta, R., Guillermo M., *Learning scikit-learn: machine learning in python*. Birmingham: Packt Publishing, 2013.

[8] Herrera A., Arroyo A., Jiménez A., Herrero A., "Forecasting hotel cancellations through machine learning," *Expert Systems*, 41(9), e13608, 2024.

[9]. Kevin M., PhocusWire*, Hotel cancelation rate at 40% as online travel agencies push free change policy*. D-Edge Hospitality Solutions. https://www.phocuswire.com/Hotel-distribution-market-share-distribution-analysis

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

[11]. Nekouei F., *Hotel booking cancellation prediction*, 2023.

[12]. Pham Dinh Khanh (KhanhBlog), *Introduction to the forest model (Random Forest), 2021*. https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_RandomForest.html

[13]. Quinlan J. R., "Induction of decision trees," *Machine learning*, 1(1), 81-106, 1986.

[14]. Yoo M., Singh A. K., Loewy N., "Predicting hotel booking cancelation with machine learning techniques," *Journal of Hospitality and Tourism Technology*, 15(1), 54-69, 2024.