International Journal of Scientific Research and Management (IJSRM)
||Volume||13||Issue||09||Pages||2615-2619||2025||
Website: https://ijsrm.net ISSN (e): 2321-3418
DOI: 10.18535/ijsrm/v13i09.ec01

Application of Machine Learning Algorithms in Prediction of Hotel Booking Cancellation

Bui Hai Phong[1], Nguyen Le Phuong Vy[2], Trieu Ngoc Van[2], Nguyen Viet Nhat Minh[3], Nguyen Ngoc Thanh[4], Trinh Nam Khanh[2] and Ngo Nam Khanh[2]

[1]Hanoi Architectural University, Hanoi, Vietnam
[2]Hanoi-Amsterdam High School for the Gifted, Hanoi, Vietnam.
[3]HUS High School for Gifted Students, Hanoi, Vietnam.
[4]Tan Hiep High School, Dong Thap, Vietnam.

Abstract:
Machine learning algorithms have been applied in various fields. In hotel booking, machine learning algorithms have shown promising results. The applications of machine learning algorithms in prediction of hotel booking cancellation have attracted much attention. The capability of machine learning plays an important role to predict hotel booking cancellation early and accurately. The paper presents the applications of various machine learning algorithms of prediction of hotel booking cancellation. The performance evaluation on a large public dataset has shown the applicable results of the machine learning applications.

Keywords: Hotel booking cancellation prediction; Machine learning; Feature extraction

1. Introduction

In recent years, applications of online hotel booking have rapidly developed. Users can search for and book comfortable hotel rooms using mobile and web applications. Moreover, users can book hotel rooms using email or phone applications. The booking applications are convenient for users and hotel management activities. In hotel room booking applications, the cancellation of hotel booking is very important. The function is crucial for hotel booking management and resource allocation [1]. However, there are different reasons for canceling hotel reservations. The factors that caused the cancellation of hotel reservations can be: flight changes, itinerary changes, working schedule changes. To check if a hotel booking is canceled or not, traditional methods require much time and human effort to contact customers by email or phone. With the advances in machine learning researches, the prediction of cancellation of hotel reservations can be performed accurately and early [2]. The application of machine learning helps hotel managers to predict the percentage of cancellation of hotel reservations. Using machine learning algorithms, the cancellation of hotel reservations can be predicted based on collected data from customers. The prediction is performed by machine automatically without using human checking. Based on obtained results, hotel managers can make precise decisions and support customers better.

Hotel managers can save much time and resources for hotel room management. The paper proposes a method to predict the cancellation of hotel reservations using various machine learning algorithms such as k nearest neighbor, Random forest. The proposed method is evaluated on a large dataset of hotel reservations. Moreover, we compare performance of proposed with existing methods to analyze the strength and weakness of the proposed method.

2. Literature Review

The section reviews significant approaches related to hotel room booking and cancellation. In hotel booking, it is more important to accurately predict hotel room occupancy in an entire hotel.

This study explores the use of customer ratings and review texts to forecast monthly hotel occupancy using Long Short-Term Memory (LSTM) networks. Sentiment analysis was conducted on hotel reviews from Taiwan, with LSTM outperforming five other models: Back Propagation Neural Network (BPNN), General Regression Neural Network (GRNN), Least Squares Support Vector Regression (LSSVR), Random Forest

(RF), and Gaussian Process Regression (GPR). The integration of sentiment scores with ratings significantly improved prediction accuracy [3].

Another study aimed to reduce booking cancellations and optimize revenue by implementing a prepaid deposit strategy, leveraging a Random Forest model for rate adjustment and cancellation prediction [4]. Similarly, artificial intelligence techniques were employed to predict hotel cancellations using Passenger Name Record (PNR) data, achieving 80% accuracy for forecasting cancellations within a 7-day window [5]. This research compared four machine learning models—Logistic Regression, Decision Tree, K-Nearest Neighbors, and Random Forest Classifier—with Random Forest demonstrating superior performance in revenue management [6].

Additionally, a separate study proposed a cancellation prediction method using 13 independent variables and achieved a prediction accuracy of up to 98% [7]. Using a Kaggle dataset, another investigation assessed various machine learning models to evaluate their effectiveness in forecasting cancellations, identifying XGBoost as the most accurate among individual classifiers [8].

Accurate demand forecasting remains a significant challenge in the hospitality industry, largely due to booking cancellations. Enhancing forecast precision could reduce cancellations and strengthen demand management strategies [9].

Furthermore, revenue management practices such as addressing no-shows can substantially increase profitability. Data mining of Passenger Name Records has helped uncover dynamic patterns in cancellation behavior and assess the performance of advanced predictive models [10]. Although, the researches on the prediction of cancellation of hotel reservations has obtained great advances, the issue needs to be further investigated to improve the high accuracy.

3. Proposed method

The steps of our proposed method for the prediction of cancellation of hotel reservation is described in Figure 1. Firstly, we collect data of hotel booking from Kaggle repository [11]. The collected data is divided into training and testing datasets. After that, various machine learning algorithms are applied to predict the cancellation of hotel reservation. In the work, we investigate the performance of the prediction using kNN, SVM and Random forest algorithms. Finally, we compare the performance of proposed and existing methods.

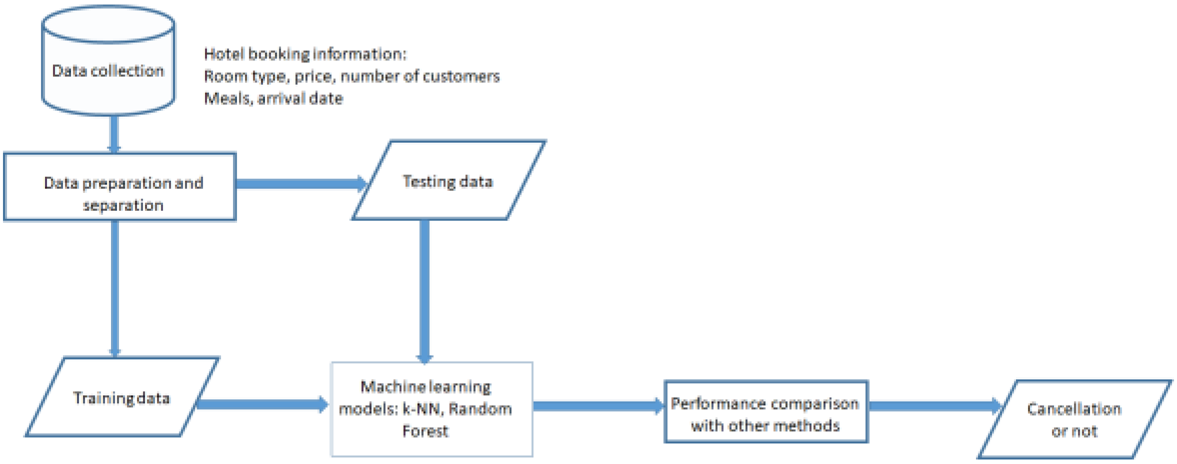

Data collection
Hotel booking information:
Room type, price, number of customers
Meals, arrival date
Data preparation and separation
Testing data
Training data
Machine learning models: k-NN, Random Forest
Performance comparison with other methods
Cancellation or not

Figure 1. Framework of the proposed method for predicting cancellation of hotel reservations using machine learning algorithms

3.1. Data description

Data in hotel reservations has been collected from Kaggle repository [11]. The dataset contains 13 features of hotel reservations of customers. In the dataset, there are two types of a hotel reservation: *cancel and not cancel*. Machine learning algorithms are designed to predict the correct label of a hotel reservation. Table 1 describes the features of hotel reservations of customers.

Table 1. Feature description in hotel reservation dataset

Feature	Description
no-adults	Number of adults booking hotel
no-children	Number of Children booking hotel
no-nights	Number of week nights booking hotel
no-weekend-nights	Number of weekend nights booking hotel
type-meal-plan	Type of meal requested by customers
car-parking-space	Is a car parking space required or not
room-type-reserved	Type of room reserved by the customer
lead-time	Number of days between the date of booking and the arrival date
no-cancellations	Number of previous bookings that were canceled
no-not-canceled	Number of previous bookings that were not canceled
avg-price-room	Check if the booking was canceled or not
booking-status	Average price per day of the reservation

3.2. Machine learning algorithms

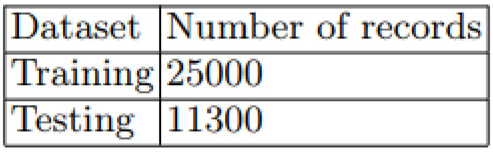
k-Nearest Neighbors (kNN) [12] is a simple, intuitive, and widely used supervised learning algorithm used for classification and regression tasks. It is a nonparametric and instance-based learning method, meaning it doesn’t make strong assumptions about the underlying data distribution and stores all available cases for decision-making.

Support Vector Machine (SVM) [13] is a powerful supervised learning algorithm used for classification and regression tasks. It is especially effective in high-dimensional spaces and with datasets where the number of features exceeds the number of samples. The main goal of SVM is to find the optimal hyperplane that best separates the data into classes. A hyperplane is a decision boundary that divides the space; in two dimensions, it is simply a line. For binary classification, SVM finds the hyperplane that maximizes the margin between the two classes.

A Decision Tree [14] is a popular and easy-to-understand supervised learning algorithm used for both classification and regression tasks. It models decisions in a tree-like structure, where internal nodes represent features, branches represent decision rules, and leaf nodes represent outcomes (e.g., class labels or predicted values).

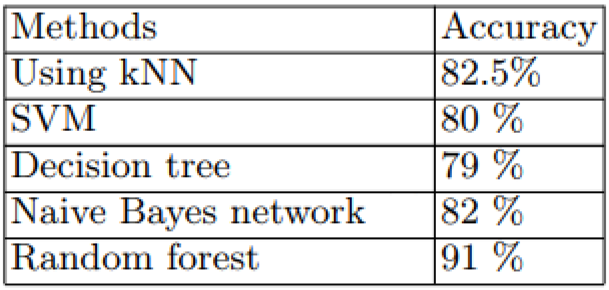
Random Forest [15] is a popular and powerful ensemble learning algorithm used for both classification and regression tasks. It works by building multiple decision trees during training and combining their outputs to improve accuracy and control overfitting. A Random Forest is an ensemble of many individual Decision Trees. Instead of relying on a single decision tree (which can be prone to overfitting), random forest aggregates the results from multiple trees to produce more robust and accurate predictions.

Table 2. Description of training and testing dataset of hotel reservation



Dataset	Number of records
Training	25000
Testing	11300

Table 3. Performance comparison of proposed and existing methods for the prediction of cancellation of hotel reservation



Methods	Accuracy
Using kNN	82.5%
SVM	80 %
Decision tree	79 %
Naive Bayes network	82 %
Random forest	91 %

4. Experimental results

The proposed method is evaluated on the public dataset of hotel reservations. The data set consists of 36300 records of customer information. Table 2 describes the number of records in training and testing dataset. In the paper, we apply the accuracy metric [2] that is commonly used in machine learning algorithms. The metric is described as follows:

$$\text{Accuracy} = \frac{\text{Correct} - \text{predictions}}{\text{Total} - \text{predictions}} \quad (1)$$

Table 3 shows the performance of the prediction of cancellation of hotel reservations using various machine learning algorithms. The random forest obtains the highest score because the algorithm aggregates results of hundreds of Decision trees. The SVM algorithm obtains the lowest score. The Decision tree and Naive Bayes obtained better results compared to SVM and k-NN algorithms. The Random Forest algorithm gains better accuracy compared to the method using Naive Bayes in [1].

5. Conclusion

The paper investigates various machine learning algorithms for the prediction of cancellation of hotel reservations. Random forest obtains highest score compared to other methods. The proposed method is evaluated on a large dataset that consists of a wide range of features. Obtained results have shown the promising applications of the proposed method. In the future, the results can be developed in the hotel booking applications.

References

1. M.Jishan et al. Hotel Booking Cancellation Prediction Using Applied Bayesian Models. 3 2024 International Conference on Decision Aid Sciences and Applications (DASA), 2024, doi: 110.1109/DASA63652.2024.10836282.
2. N.A.Putro et al., PREDICTION OF HOTEL BOOKING CANCELLATION USING DEEP NEURAL NETWORK AND LOGISTIC REGRESSION ALGORITHM, Jurnal Techno Nusa Mandiri 18(1):1-8, 2021.
3. Y.-M. Chang, C.-H. Chen, J.-P. Lai, Y.-L. Lin, and P.-F. Pai, "Forecasting Hotel Room Occupancy Using Long Short-Term Memory Networks with Sentiment Analysis and Scores of Customer Online Reviews", Applied Sciences, vol. 11, no. 21, p. 10291, 2021.
4. E. C. Sánchez, A. J. Sánchez-Medina, and M. Pellejero, "Identifying critical hotel cancellations using artificial intelligence", *Tourism Management Perspectives*, vol. 35, p. 100718, 2020.
5. M. A. Afrianto and M. Wasesa, "Booking Prediction Models for Peerto-peer Accommodation Listings using Logistics Regression, Decision Tree, K-Nearest Neighbor, and Random Forest Classifiers", *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 2, pp. 123–132, 2020.
6. *J. Sánchez-Medina and E. C.-Sánchez, "Using machine learning and big data for efficient forecasting of hotel booking cancellations",* International Journal of Hospitality Management, *vol. 89, p. 102546, 2020.*

7. M. S. Satu, K. Ahammed, and M. Z. Abedin, “Performance Analysis of Machine Learning Techniques to Predict Hotel Booking Cancellations in Hospitality Industry”, in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6, 2020.
8. Antonio, A. de Almeida, and L. Nunes, “Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior”, *Cornell Hospitality Quarterly*, vol. 60, no. 4, pp. 298–319, 2019.
9. D. Romero Morales and J. Wang, “Forecasting cancellation rates for services booking revenue management using data mining”, European Journal of Operational Research, vol. 202, no. 2, pp. 554-562, 2010.
10. Herrera, A. Arroyo, A. Jimenez, and Á. Herrero, “Forecasting hotel cancellations through machine learning”, Expert Systems, 2024.
11. Rice Leaf Diseases Dataset, https://www.kaggle.com/datasets/dedeikhsandwisaputra/riceleafs-disease-dataset, Accessed: 2025-04-20.
12. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer. ISBN: 978-1-4614-7137-0.
13. Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297. DOI: 10.1007/BF00994018.
14. Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). Classification and Regression Trees. Wadsworth. ISBN: 978-0412048418.
15. Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32, DOI:10.1023/A:1010933404324.