

Hotel Booking Cancellation Prediction Using Machine Learning for Improved Reservation Management

Lamyaa Sabeeh Ashour¹ Sin Han Jinn² Lee Jeong Luke³

¹Department of Mathematics, College of Basic Education, University of Misan, Misan, Iraq

²School of Computer Science, University Sains Malaysia, 11800 Penang, Malaysia

³School of Computer Science, University Sains Malaysia, 11800 Penang, Malaysia

doi: <https://doi.org/10.37745/ejcsit.2013/vol14n65872>

Published October 03, 2026

Citation: Ashour L.S., Jinn S.H., Luke L.J. (2026) Hotel Booking Cancellation Prediction Using Machine Learning for Improved Reservation Management, *European Journal of Computer Science and Information Technology*, 14(6),58-72

Abstract: Hotel booking is a vital part of hotel operations that supports the information needed for better management of room inventory and availability, occupancy, and resources. However, cancellations can complicate expected occupancy levels and have a continuing impact on reservation planning and revenue management. The objective of this study is to improve and compare four machine learning (ML) models (XGBoost, Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN) to identify the most effective model for hotel booking cancellation prediction. The analysis was performed based on the Hotel Booking Demand dataset containing a total of 119,390 records with 31 attributes. This involved missing value treatment, feature engineering, and data type conversion for encoding categorical variables. Models were assessed using accuracy, recall, precision, F1-score, and confusion matrices. XGBoost obtained an accuracy of 98.39%, precision values of (0.99 and 0.97), recall values of (0.98 and 0.98), and F1-scores of (0.99 and 0.98) for Classes 0 and 1 respectively, outperforming the other models. It was followed by RF with 98.24%, DT with 97.95%, and KNN with 97.82%. These results show that XGBoost achieved the strongest predictive performance among the evaluated models and can be utilized for reservation planning, resource allocation, capacity planning, and revenue management.

Keywords: hotel booking cancellation, machine learning, CRISP-DM, hotel booking hotel booking demand, reservation management.

INTRODUCTION

The hospitality industry is a pivotal part of the tourism sector, and therefore, effective management of hotels relies heavily on accurate analysis of booking patterns and customer behavior. Hotel reservations are a challenging task where reservation factors include arrival date, departure date, stay duration, guest capacity per room, and reservation cancellations. However, with the increase in hotel booking data in recent years, it has opened the opportunities to utilize data-driven and machine learning (ML) approaches to assist hotel management efforts and enhance operational decision making-processes [1].

Canceling a booking is one of the biggest problems hotels face, as this uncertainty that creates or independent hotel rooms can really damage revenue management and resource allocation. For hotels, identifying bookings that are explicitly prone to be cancelled early can help them with capacity planning and implementing strategies for distribution of remaining rooms. The need to utilize ML -based cancellation forecasts has been emphasized in a recent study in order to support hotel revenue and capacity management [2,3].

ML offers a data-driven method for forecasting hotel booking cancellations by learning the patterns and relationships present in historical reservation records. When cancellations are predicted with accuracy, hotels can anticipate shifts in expected occupancy and establish the operational processes required to manage reservations. The predictive models, run by analyzing the characteristics related to booking data, provide the probability of cancellation and help in reservation planning, room capacity management, and resource allocation. By applying and comparing the enactment of various ML classification models on this dataset, we can determine a good predictor model for hotel booking cancellation [4].

The Hotel Booking Demand dataset, presented by Antonio et al. [5]. This dataset consists of 119,390 booking records and 31 variables pertaining to reservations at a resort hotel and a city hotel. The data contains a total of 119,390 booking records containing information about 31 variables relative to reservations at a city hotel and a resort hotel. The dataset records reservations with dates of arrival starting from (July 2015 until August 2017) and summarizes information on some booking related characteristics such as lead time, length of stay, number of guests, market segment and previous booking information.

The purpose of these experiments is to predict the possibility of cancellation, and accordingly, XGBoost, Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN) are used as algorithms. The dataset is processed with missing-value treatment, feature engineering, and categorical-variable encoding before model development. Evaluation of models is done using accuracy, recall, precision, and F1-score with confusion matrices. A more important consideration in cancellation prediction is the number of evaluation measures, as an overall high accuracy can simply mean that a model cannot predict cancelled bookings well. A recent set of experiments successfully demonstrated how a model with high precision could also yield false negatives and a quite low recall when given the task to predict hotel cancellations, meaning it's very important to evaluate various performance metrics [6].

The key objective of the present study involves developing and comparing different ML approaches for hotel booking cancellation prediction and determine which one delivers the best predictive performance. The results aim to illustrate how ML can provide useful predictive analytics that could help hotel managers with reservation forecasting, resource planning, capacity planning, and other revenue-related decision-making processes.

LITERATURE REVIEW

ML algorithms and their applications in forecasting hotel registration behavior and demand. Across the many studies evaluating ML techniques, researchers have been working to optimize operational efficiency and customer satisfaction.

ML algorithms have been used for many years to predict and analyze hotel booking demand and anticipate customer-specific needs. The algorithms in question include Naive Bayes, DT, RF, Deep Learning, and Generalized Linear Models. For example, F. Ijaz et al. [7] presented that the RF algorithm outperformed the other algorithms in classifying hotel booking data, achieving an accuracy of 83%.

Satu & colleagues [8] conducted another study assessing different ML algorithms to anticipate prospective reservation cancellations such as Logistic Regression (LR), XGBoost, and KNN. They found that the most appropriate methodology for improving prediction accuracy was XGBoost, and the Information Gain Attribute and Gain Ratio Attribute were used for feature selection.

Ali Shah [9] proposed a contemporary model for predicting resort bookings using ML algorithms, including LR, Naive Bayes, KNN, RF, and DT. The study integrated various strategies consisting of feature evaluation,

data preparation, feature engineering, and complete evaluation of algorithms. Among the models examined, the RF algorithm validated the pleasant overall performance, reaching an accuracy of 95.40%, a recall of 98.81%, a precision of 88.77%, and an F1-score of 93.47%.

Kumar & Sharma [10] presented ML techniques to forecast booking cancellations in hotels, a critical task with direct implications for revenue management in the industry. Various ML algorithms were used in the study, including LR, KNN, DT, RF, and XGBoost. RF was the best model with optimized hyperparameters after correction, achieving 85.36% accuracy. This model helps hotel managers by reducing loss of revenue and having more flexible cancellation policies and also taking actions on overbooking risks.

Li & Liu [11] performed a comparative evaluation of several ML algorithms, including LR, Naive Bayes, DT, RF, SVM, and Neural Network, using 10,000 hotel reviews from TripAdvisor with ratings between one and five stars. The models were assessed through two 10-fold cross-validation tests. The lowest classification accuracy (84%–87%) was achieved using only review titles as input (an average of 25.7 bytes) to train the models. However, when the actual reviews (full text) are combined and much of the boilerplate words are removed, SVM achieves an accuracy of about 92%.

Lee & Kim [12] investigated factors affecting hotel selection for extended stays using Expedia data and DT algorithm. Package reservations were the number one driver of longer stays across the regions, with prediction accuracies for the Asia (75.47%), Americas (72.16%), and Europe (63.08).

Saputro & Nanang [13] utilized ML algorithms (RF, DT, Extra Trees Classifier, and XGBoost) for predicting hotel booking cancellations. This method required to tackle the increasing problem of cancellations accompanying online reservations, an evolution in accommodation that brings with it operational difficulties. Accordingly, for this dataset RF and Extra Trees Classifier both had the highest precision rates (and outperformed in recall), predicting 79% of actual cancellations as correct.

Another study presented several ML algorithms to recommend hotels based on customer data [14]. Results showed that the complex configurations are more appropriate than default linear ones for optimal hotel recommendations and that a more advanced ML method is required to better grasp consumers' preferences in hotel selection.

B. Zhang et al. [15] introduced forecasting hotel accommodation demand using Long Short-Term Memory Networks (LSTM). The results indicated that LSTM can be used to integrate internet search indices into demand forecasts, which highlights the power of using deep learning methods in this area.

Putra & Azhar [16] used an artificial neural network (ANN) and LR and defined the best model through hyperparameter tuning in five experiments to predict canceled hotels. The five experiments were ANN, LR with L1 Penalty, and LR with L2 Penalty and used GridSearchCV for the LR. Among these, the LR with GridSearchCV (fifth experiment) gave the best result, which was accuracy of 79.77%, precision of 85.86%, and recall 55.07%.

Finally, Deep Neural Networks with L R were utilize for hotel booking cancellation prediction. It has stretched the effect of feature selection on prediction accuracy and features that highly influence the model.

Overall, the literature indicates that there is a strong trend towards the application of different ML techniques in solving class issues within hotel management. Together, these studies support the idea that ML has a lot of

potential when it comes to operational decision-making in hotels across several applications, ranging from cancellation predictions to exploiting recommendations.

METHODOLOGY

A. Research Framework

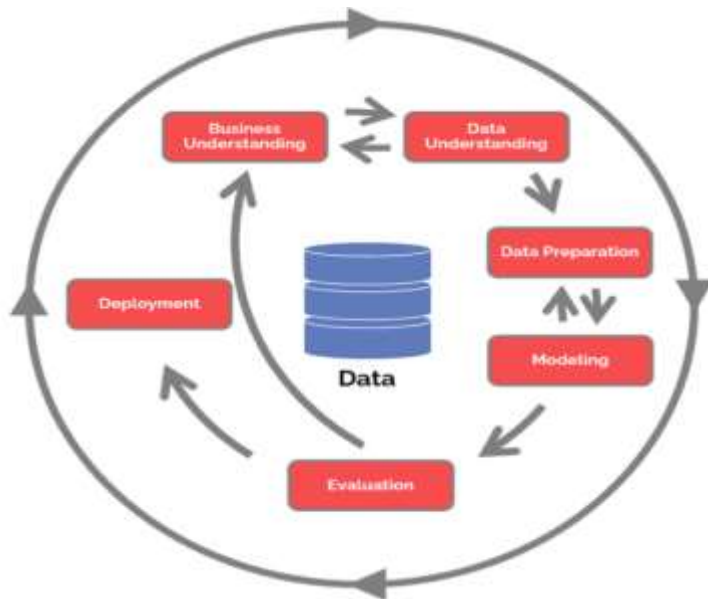


Fig. 1. Diagram of CRISP-DM Phases [11]

This study employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology as a framework for the study. CRISP-DM contains six phases that naturally describe the data science life cycle, as illustrated in Fig. 1 [17]. There are several reasons for adopting this methodology, as it is a proven and repeatable framework that requires relatively limited data mining expertise, provides a common structure, and, most importantly, offers flexibility throughout the research process. The initial five phases were applied in this study, while the deployment phase is beyond the scope of the present study.

1. **Business Understanding:** This stage defines the aims and needs of the study. The main objective of this research is to use classification methods and ML algorithms on hotel booking dataset to predict cancellation in a very precise way. It will help hotel management to optimize bookings and maximize revenue.
2. **Data Understanding:** Identifying relevant datasets, collecting data and analysing it. In this project will use the hotel booking demand dataset, which consists of booking data for city and resort hotels.
3. **Data Preparation:** The dataset undergoes a thorough preparation process that includes cleaning, transforming, and constructing new features to ensure it is suitable for modelling.
4. **Modelling:** Various ML models are developed and assessed using different techniques. This phase includes selecting appropriate algorithms, designing tests, and building models.
5. **Evaluation:** After building, the evaluate each model against business objectives defined upfront. The top model would be chosen through its predictive capabilities of cancellations.

Dataset and Preparation

The Hotel booking demand dataset on Kaggle is a combination of booking history collected by Nuno Antonio et al. [5]. This dataset has information about booking at a city hotel and resort hotel, when the booking was made, how long the stay was, adults number, children number or babies number etc. The target variable for this study is the column “cancelled” which indicates whether a hotel booking has been cancelled or not. The variable contains two classes: 0 for non-cancelled bookings and 1 for cancelled bookings. The dataset contains 31 attributes and 119,390 booking records, including both numerical and categorical variables.

For data preparation, the first step done was to find the total number of missing null values. There are four features with null values company, agent, country, and children. Highest missing values are present in the company column, followed by agent and country. Reading the description and explanation of the dataset, the author described the ‘agent’ column as the “ID of the travel agency that made the bookings”.

This means that those with ‘agent’ in the dataset are the only ones that made the booking via a travel agency and those with null values are those that did not. Based on this, the null values in this column were filled with 0 to represent no agents used. Same method was used for the ‘company’ column as the rows with null values are those who book the hotel on their own and not through a company. ‘Country’ column represents country codes of different countries and as it is a categorical feature, the null values were replaced with the mode (most common country appearing in dataset).

The ‘children’ column contains the number of children, and it is a numerical data, thus the null values were replaced with the mean value. There are many rows with zero guests as well (Zero adults, children, and babies), thus those rows were dropped after that, standardization of data types was done on the ‘children’, ‘company’, and ‘agent’ column. These three columns were converted from floating numerical data to integer type to fit in with the other data. how long the stay was, the number of adults, children, or babies, etc. The target variable for this study is the column “canceled,” which indicates whether a hotel booking. Next step, used dimensionality reduction by deriving new features from existing attributes to decrease the total number of features in use.

The Dataset has 2 features which are ‘reserved_room_type’ and ‘assigned_room_type’. A new feature which return ‘1’ if the guest is assigned with the same room number as reserved and ‘0’ otherwise. Based on columns ‘previous_cancellations’ and ‘previous_bookings_not_cancelled’ an additional new feature was created which contained ‘1’ if the guest had cancelled more bookings in the past than booked with no cancellations; otherwise is ‘0’.

The less important features that no longer contributes to the prediction such as dates and the features mentioned above are then dropped. The dataset containing object data types are converted into numerical integer form using Label Encoder from Sklearn to make them suitable for use in the modelling algorithm later.

A Heatmap is then plotted as shown in Fig.2 to observe the correlations among the attributes and the target variable, is_canceled. The Heatmap illustrates an overview of the correlations among the numerical attributes and the target variable.

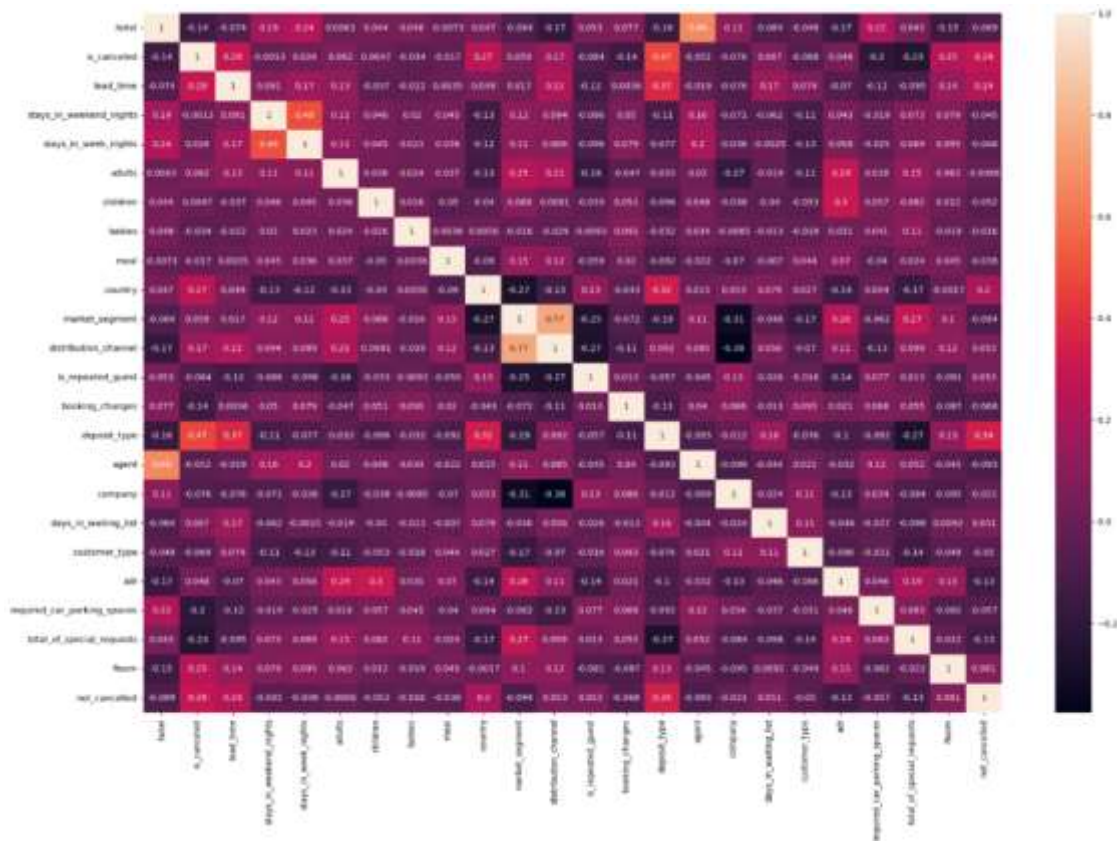


Fig. 2. Correlation Heatmap

Lastly, the dataset is splitted into train-test with ratio of 80:20 using train_test_split function from sklearn library.

PROPOSED METHOD

In this study, four ML cancellations: XGBoost, RF, DT, and KNN algorithms are employed to predict hotel booking. Each algorithm is chosen for its unique strengths in handling classification tasks, particularly in scenarios with imbalanced datasets and will be applied at predicting hotel booking cancellations. A flow diagram of the modeling approach proposed in this paper is provided in Fig. 3.

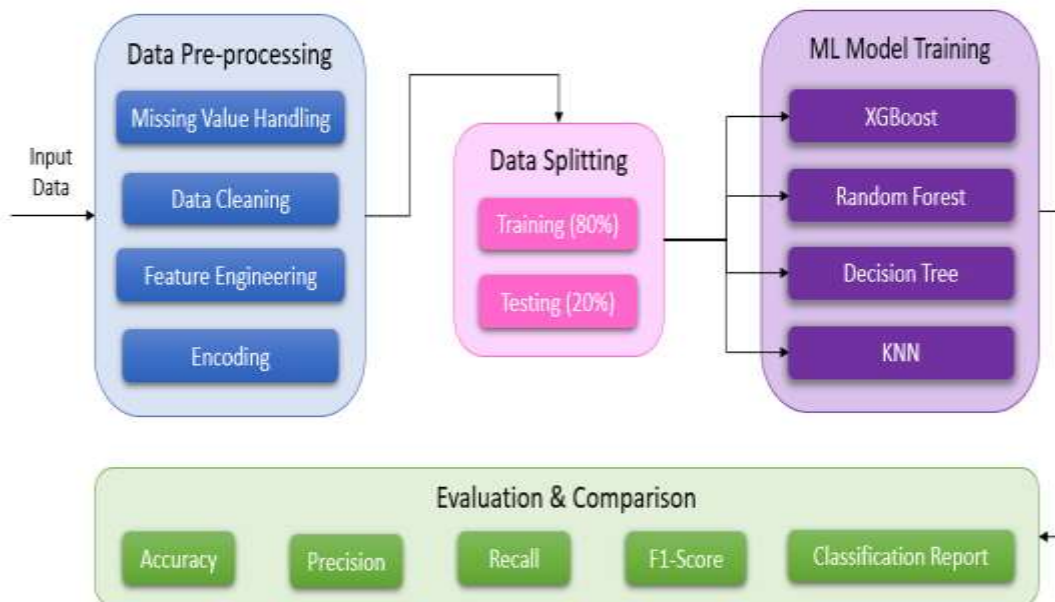


Fig. 3. Flow Diagram of Proposed Method

XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced, optimized tool for ensemble learning technique using gradient boosting. More commonly, you will encounter it in a context where multiple weak learners (often decision trees) are built sequentially in such a way that new trees correct errors made by prior ones. XGBoost is aimed at optimizing the computational complexity for finding the best split which in tree construction algorithms, is usually the most time-consuming part. [18]. XGBoost reduces the bias and variance of predictions by optimizing the model through regularization techniques and advanced tree pruning methods. Imbalance XGBoost offers two mechanisms to deal with imbalance: weighted-XGBoost and focal-XGBoost [19]. Additionally, more iterations can lead to a more accurate model but may also increase the risk of overfitting if not properly managed, which necessitates careful tuning of hyperparameters. In this study, XGBoost is an efficient and scalable implementation of gradient boosting framework. It is particularly effective for structured data and has gained popularity due to its performance and speed. In this implementation, the XGBoost model was initialized with the evaluation metric set to 'logloss', which measures the logistic loss between predicted probabilities and actual labels. To address the class imbalance in the dataset, the `scale_pos_weight` parameter was utilized. This parameter computes the ratio of negative instances to positive instances in the training set to help balance the classes during training. The model was fit to the training data and predictions were made on the test set.

Random Forest

RF is an ensemble method, which associates multiple randomized DT by averaging their predictions. In particular, RF mitigates the high variability observed in individual decision trees through the use of random training samples and random subsets of features to train small trees which are then aggregated [20]. With more trees you would have a stronger model with lower chance of overfitting but at the cost of requiring additional time to process and a more complicated model to interpret. It is appropriate in situations where the number of variables is greater than observations, and it can scale to large problems and adapt readily to many types of ad hoc learning tasks whilst returning measures of variable importance[21]. RF is an ensemble learning algorithm that creates numerous DT from training data and gives as output the mode of their predictions for classification tasks [22]. The RF model parameters were logically designed so as to achieve high accuracy. Notably, the model used 'balanced' class weights to compensate for the fact that the dataset is biased towards certain classes by nature. This strategy ensured that equal weight was placed on both classes when training the model, addressing the problem of class imbalance. The random state is defined as 42 for the reproducibility of the results in this case. The RF model was trained on all of the training set predictors and then used to predict the test set predictors, with predictions also serving as a very good measure of performance for the unseen data.

Decision Tree Classifier

DT is the most easily understood and commonly used supervised learning method for classification & regression. Unlike other techniques that work on a one-shot decision, the DT works hierarchically to traverse over explore and classify [23]. DT is also simple to use since it creates a more interpretable model, which can be visualized and broken down into parts. A DT starts from the root and goes down, typically represented left to right [24]. Every internal node of the tree is a decision to be made on a specific feature, and every leaf node is an outcome. DT is also effective for both numerical and categorical data types, thus making them applicable on a variety of datasets. They are quite simple, but they can be a building block for more complicated ensemble algorithms such as RF and Gradient Boosting. According to a tree-like model of decisions and their possible consequences, DT Classifier. However, it is simple and easy to understand but may easily overfitting. The DT model for this study has defined random state 0 and criterion= 'entropy' This 'entropy' criterion showed that the model would use information gain to determine which splits are optimal in training [25]. Setting these parameters, the DT model would give predictions on the test set, allowing sensitivity on its ability to generalize.

K-Nearest Neighbors (KNN)

KNN is a supervised learning algorithm that can be used for both classification and regression tasks but is mostly used as a non-parametric method. It is based on finding the 'k' nearest data points in the feature space and making predictions. Generally, the KNN classifies the dataset using the nearest training data points (the neighbors) to test the query [26]. KNN reduces the complexity of the model by relying on the proximity of data points rather than building a model explicitly. KNN incurs lower model complexity because it does not construct an explicit model however this method relies on the

distance between data points. The basic KNN algorithm is very sensitive to noise characteristics because it uses a simple Euclidean distance to compute the similarity information. KNN is very effective when the boundary that separates the three classes is arbitrary; K-nearest neighbors can work on different types of data and low parameters need to be optimized [26]. Additionally, it is often useful to take more than one into account, so the technique is more commonly referred to as KNN classification, where KNN is used in determining the class [27].

The KNN model was instantiated with a meticulously chosen number of neighbors ($k=5$). This helped the model in learning regional patterns while reducing overfitting. Then, the test set was used to generate predictions. The weighting scheme was set as distance so that nearer neighbors outweighed farther ones when making predictions across the three models studied to address the class imbalance effect effectively. This will reduce the impact of class imbalance and optimize the overall performance of a model [28]. The various algorithm will be compared based on key parameter like accuracy, precision, recall and F1-score to find out what kind of model is giving better prediction towards hotel booking cancellation. The best fit to these detected cases so the recommendations strategically recommendations will be developed.

EXPERIMENT AND ANALYSIS

Experimental Setup

The Hotel Booking Demand dataset was used in the experiment to predict hotel booking cancellations by implementing different ML algorithms. Therefore, the focus was limited to on four models (XGBoost, RF, DT, and KNN). The performance of each model was evaluated using accuracy, precision, recall, and F1-score. A defined evaluate model function was used to calculate all the metrics in one go and also print out the confusion matrix for smoother evaluation across models. It accepts true labels, predicted labels, and the model name as parameters. Here, inside this function, with the help of true and predicted labels, usual evaluation metrics and the number of mislabelled points are calculated. This is a confusion matrix to understand how many instances have been classified correctly and incorrectly. This matrix allows us to discover other issues on which the model is working poorly. This is called with the relevant parameters each time a model has to be evaluated, and the results are returned.

Results

The evaluation of the four machine learning models created from this standard hotel booking transactions dataset to predict the cancellation scenario is analyzed here. This evaluation is done using various metrics to judge the performance of the models, which include verifying some important parameters such as accuracy, precision, recall, and F1-score. Apart from classification reports, confusion matrices give a visualization of the respective results and show how the models perform in making predictions.

XGBoost

The model was trained using the training dataset, and it produced predictions on the testing dataset that can be evaluated. There were a total of 23,842 instances, and the model misclassified 383 of them, achieving an accuracy of 98.39%. A summary of the results is presented in Table 1. As a result, the

XGBoost model made a fantastic prediction about hotel booking cancellations. From the confusion matrix, we see that there are high precision and recall values for both classes (cancelled bookings and non-cancelled bookings), which also indicates that the model is doing very well in predicting both classes, as shown in Fig.4, with minimal misclassification.

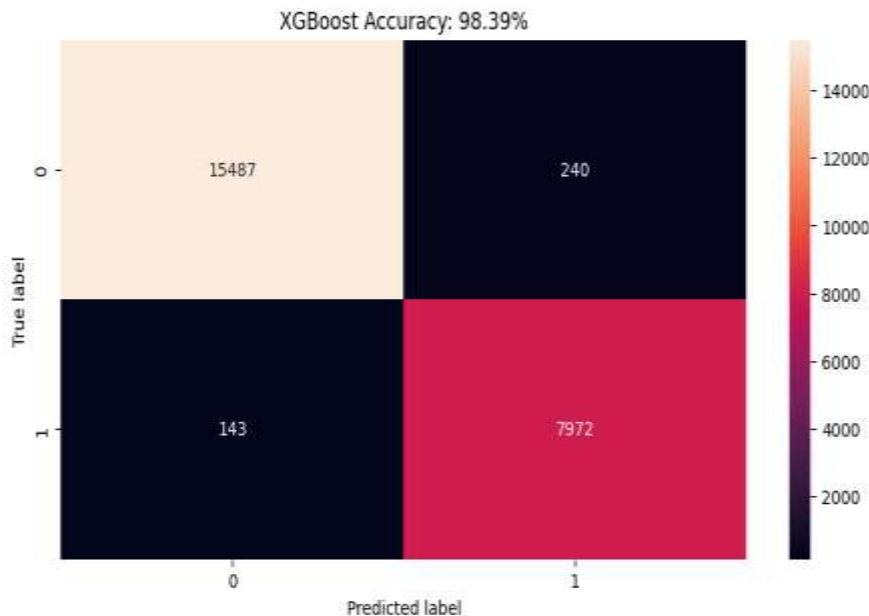


Fig. 4. Confusion Matrix for XGBoost Algorithm

Random Forest

The RF model is evaluated using the test dataset by comparing the predicted outcomes to the actual labels. It attained 98.24% accuracy, misclassifying 419 of 23,842 instances. The results are provided in Table 1. The Random Forest model performed well in terms of both overall accuracy and precision for predicting hotel booking cancellations, showing that it is a good classification method to use in this case. The confusion matrix and classification report help in the analysis of model performance across both classes (cancelled and not cancelled) within classification models, which easily identifies the bookings that are likely to be cancelled, as shown in Fig. 5.

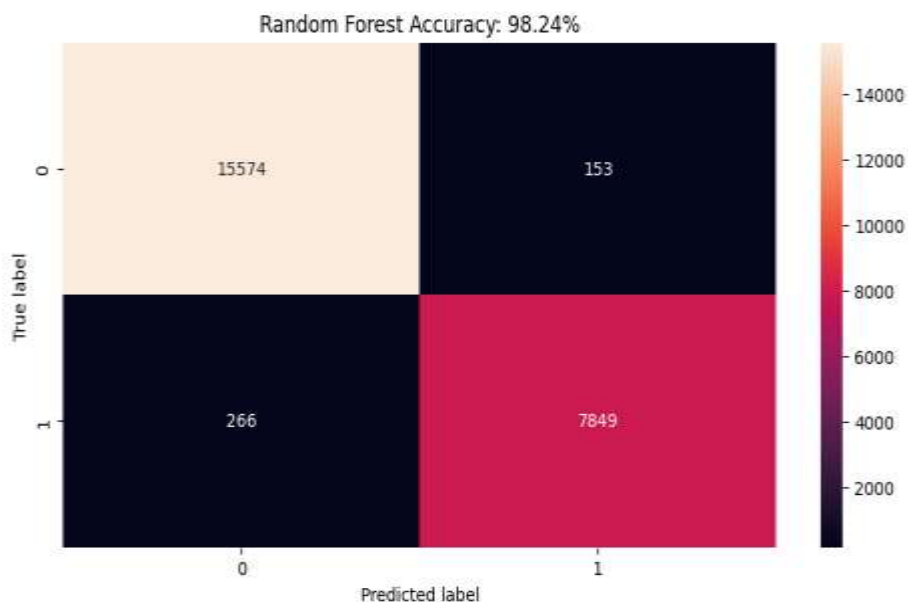


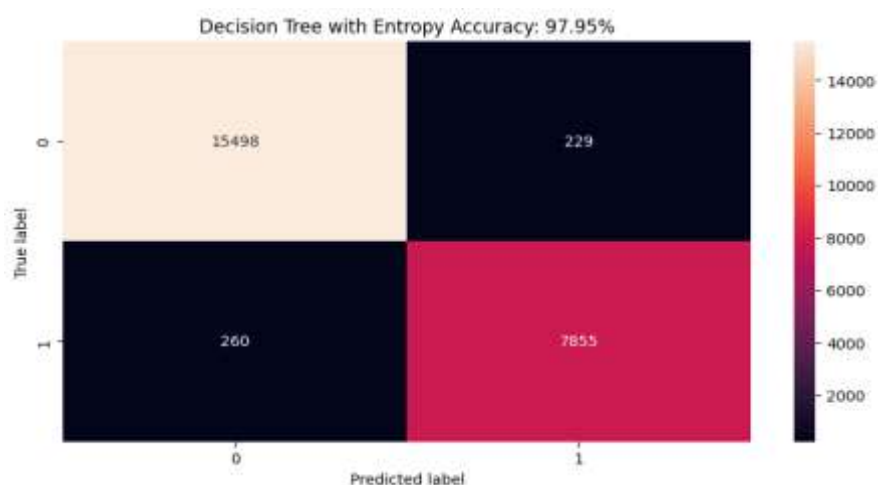
Fig.
for Random Forest Algorithm

5. Confusion Matrix

Decision Tree

This is an evaluation of the Decision Tree model with the setting of Criterion='entropy', allowing us to evaluate its performance in predicting hotel booking cancellations. The overall accuracy of classifying the test set is 97.95%, which means that there are just 489 misclassified instances. Table 1 provides an overview of the performance. The DT model with the 'entropy' criterion for feature importance performed well in predicting whether a hotel booking was cancelled. The high precision and recall values confirm the correctness of the model in identifying both classes (cancelled or not cancelled), as shown in Fig. 6.

Fig. 6. Confusion Matrix for Decision Tree (Entropy) Algorithm



KNN

To evaluate the K-Nearest Neighbors model by assessing its ability to classify hotel bookings cancellations. The KNN model achieved accuracy at 97.82% which translates into 520 out of 23842 instances being mislabelled. Summary of the results are provided in table 1. KNN Model for Hotel Booking Cancellations Both classes had high precision and recall, as is expected, leading to an effective model that correctly identified both the cancelled and non-cancelled bookings, as shown in Fig. 7.

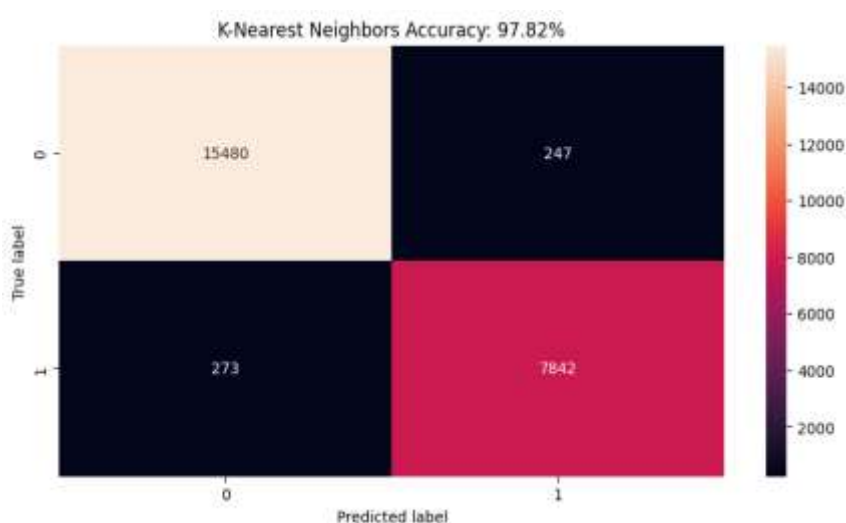


Fig. 7. Confusion Matrix for KNN Algorithm

TABLE1: PERFORMANCE COMPARISON OF THE MACHINE LEARNING MODELS

Model	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-score (Class 0)	F1-score (Class 1)	Accuracy
XGBoost	0.99	0.97	0.98	0.98	0.99	0.98	98.39%
Random Forest	0.98	0.98	0.99	0.97	0.99	0.97	98.24%
Decision Tree	0.98	0.97	0.99	0.97	0.98	0.97	97.95%
KNN	0.98	0.97	0.98	0.97	0.98	0.97	97.82%

Comparative Analysis

All models are compared together to recognize the most efficient model for predicting hotel booking cancellations. All models are evaluated based on the key performance metrics summarized in Table 1. The analysis has also revealed that all models are highly predictive of hotel booking cancellations, with XGBoost achieving an accuracy of 98.39%, outperforming RF, which achieved an accuracy rate of 98.245% and a similar mislabelling rate. The DT and KNN models also obtained high accuracies (97.95% and 97.82%, respectively), but lower accuracies and higher mislabelling rates than their ensemble versions. These results emphasize the importance of algorithm selection based on performance metrics for solving problems such as bookings cancelled in the hospitality sector, setting the basis for strategic insights to manage booking cancellation rates effectively.

CONCLUSION

This paper describes the generation and evaluation of ML algorithms that predict hotel booking cancellations, one of the main issues in hoteliers. Adopting a methodical approach reflecting CRISP-DM principles, we recorded every decision and its rationale throughout the book. First, the analysis was done on four different machine learning models DT, RF, XGBoost, and KNN. Every model was thoroughly tested and assessed to determine how well or poorly it predicted the relevant metrics. At 98.39%, the XGBoost classifier was the best performing, followed by RF at 98.24%. These insights can be used for recommendations to reduce cancellations in hotel management services, as confirmed using Principal Component Analysis. Through this process, hotels can identify risk-prone customer segments and their behaviours using booking information; allowing them to come up with proactive strategies like flexible booking policies or marketing efforts against such customers. This project not only shows how machine learning features can be designed to enhance productivity in hotel management but also sets the stage for later research. There are also other possibilities where we could work along with these predictive models combined with live booking or alert systems to flag and manage bookings in advance to maximize revenue.

REFERENCES

- [1] E. Rahmawati, G. S. Nurohim, C. Agustina, D. Irawan, and Z. Muttaqin, "Development of Machine Learning Model to Predict Hotel Room Reservation Cancellations," *Jurnal Teknologi Informasi dan Terapan*, vol. 11, no. 2, pp. 58–65, 2024, doi: 10.25047/jtit.v11i2.431.
- [2] A. Herrera, Á. Arroyo, A. Jiménez, and Á. Herrero, "Forecasting hotel cancellations through machine learning," *Expert Systems*, vol. 41, no. 9, e13608, 2024, doi: 10.1111/exsy.13608.
- [3] I. Gomez-Talal, M. Azizsoltani, P. Talón-Ballesteros, and A. Singh, "Machine Learning in Hospitality: Interpretable Forecasting of Booking Cancellations," *IEEE Access*, vol. 13, pp. 26622–26638, 2025, doi: 10.1109/ACCESS.2025.3536094.
- [4] S. Chen, E. W. T. Ngai, Y. Ku, Z. Xu, X. Gou, and C. Zhang, "Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction," *Decision Support Systems*, vol. 170, Art. no. 113959, 2023, doi: 10.1016/j.dss.2023.113959.
- [5] N. Antonio, A. de Almeida, and L. Nunes, "Hotel booking demand datasets," *Data in Brief*, vol. 22, pp. 41–49, 2019.
- [6] A. Ampountolas, "Predicting hotel booking cancellations: a comprehensive machine learning approach," *Journal of Revenue and Pricing Management*, vol. 24, no. 6, pp. 539–550, 2025, doi: 10.1057/s41272-025-00532-x.
- [7] F. Ijaz, T. Hussain, M. Iram, and M. Hussain, "an Intelligent Classification of Hotel Data Using Machine Learning," *GS Int. Conf. Comput. Sci. Eng.* 2020, vol. 2020, 2020.

Publication of the European Centre for Research Training and Development -UK

- [8] M. S. Satu, K. Ahammed, and M. Z. Abedin, "Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry," ICCIT 2020 - 23rd Int. Conf. Comput. Inf. Technol. Proc., pp. 19–21, 2020, doi: 10.1109/ICCIT51783.2020.9392648.
- [9] S. O. Ali Shah, "Optimizing Hotel Booking Prediction: A Comparative Study of Five Machine Learning Algorithms," International Journal of Trendy Research in Engineering and Technology, vol. 8, no. 4, pp. 31-41, 2024, doi: 10.54473/ijtret.2024.8406.
- [10] P. Kumar, S. Sharma, and S. O., "Hotel Booking Prediction using Machine Learning," Int. J. Res. Appl. Sci. Eng. Technol., vol. 10, no. 3, 2022, doi: 10.22214/ijraset.2022.43036.
- [11] X. Li and C. Liu, "Comparison of machine learning models for sentimental analysis of hotel reviews," in IOP Conf. Ser.: Mater. Sci. Eng., vol. 806, no. 1, p. 012029, Apr. 2020, doi: 10.1088/1757-899X/806/1/012029.
- [12] Y. Lee and D.-Y. Kim, "The decision tree for longer-stay hotel guest: The relationship between hotel booking determinants and geographical distance," Int. J. Contemp. Hosp. Manag., vol. 33, no. 6, pp. 2264-2282, 2021, doi: 10.1108/IJCHM-06-2020-0594.
- [13] P. H. Saputro and H. Nanang, "Exploratory data analysis & booking cancelation prediction on hotel booking demands datasets," J. Appl. Data Sci., vol. 2, no. 1, pp. 40-56, 2021, doi: 10.47738/jads.v2i1.20.
- [14] E. Thomas, A. G. Ferrer, B. Lardeux, M. Boudia, C. Haas-Frangii, and R. A. Agost, "Cascaded Machine Learning Model for Efficient Hotel Recommendations from Air Travel Bookings," in Proceedings of the 12th ACM Conference on Recommender Systems, ACM RecSys Workshop on Recommenders in Tourism (RecTour 2019 vol 2435), 2019, pp. 9-16.
- [15] B. Zhang, Y. Pu, Y. Wang, and J. Li, "Forecasting hotel accommodation demand based on LSTM model incorporating internet search index," Sustain., vol. 11, no. 17, pp. 1–14, 2019, doi: 10.3390/su11174708.
- [16] M. S. T. Putra and Y. Azhar, "Perbandingan model logistic regression dan artificial neural network pada prediksi pembatalan hotel," JISKA (Jurnal Informatika Sunan Kalijaga), vol. 6, no. 1, pp. 29-37, 2021, doi: 10.14421/jiska.2021.61-04.
- [17] Data Science Process Alliance, "CRISP-DM," Aug. 24, 2021. [Online]. Available: (link unavailable).
- [18] R. Genuer and J. M. Poggi, "Random forests," pp. 33-55, Springer International Publishing, 2020.
- [19] G. Biau and E. Scornet, "A random forest guided tour," TEST, vol. 25, no. 2, pp. 197-227, 2016/06/01 2016, doi: 10.1007/s11749-016-0481-7.
- [20] Y. Q. Zhang, "FORECASTING HOTEL DEMAND USING MACHINE LEARNING APPROACHES." [Online]. Available: https://ecommons.cornell.edu/bitstream/handle/1813/67733/Zhang_cornell_0058O_10695.pdf?sequence=1&isAllowed=y. [Accessed: 06-Jan-2022].
- [21] D.-W. Sun, Computer Vision Technology for Food Quality Evaluation. Amsterdam: Elsevier, 2016.
- [22] Random Forests and Decision Trees September 2012 LicenseCC BY-NC-ND 4.0 JCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012ISSN (Online): 1694-0814www.IJCSI.org
- [23] Artificial Intelligence, Big Data Analytics and Insight, "Machine learning models: Generative vs. Discriminative," Feb. 22, 2021. [Online]. Available: (link unavailable). Accessed: Jan. 6, 2022.
- [24] Scikit, "Sklearn.naive_bayes.multinomialnb." [Online]. Available: (link unavailable). Accessed: Jan. 6, 2022.
- [25] A Survey of Decision Trees: Concepts, Algorithms, and Applications Corresponding author: Ibomoiye Domor Mienye (imienye@wsu.ac.za) Digital Object Identifier 10.1109/ACCESS.2024.3416838
- [26] Uddin, S., Haque, I., Lu, H. et al. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep 12, 6256 (2022). <https://doi.org/10.1038/s41598-022-10358-x>
- [27] A Survey of kNN Algorithm May 2018Information Engineering and Applied Computing 1(1) DOI: 10.18063/ieac.v1i1.770 LicenseCC BY-NC 4.0 Jingwen SunWeixing DuNiancai Shi

- [28] k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples)
<https://doi.org/10.48550/arXiv.2004.04523> Focus to learn more Related DOI:
<https://doi.org/10.1145/3459665> Focus to learn more