cisdi Journal

Computing, Information Systems, Development Informatics and Allied Research Journal
Vol. 16 No. 3, September 2025 - www.cisdijournal.net

Development of a Machine Learning–Enabled Decision Support Framework for Hotel Booking Cancellation Prediction

[1]Afolorunso, A.A., [2]Aimuel, B.E., [3]Abiodun, A.O., [4]Adesina, A.O. & [5]Ajagbe, S.A.
[1,2,3]Department of Computer Science, National Open University of Nigeria, Abuja
[4]Department of Computer Science, Olabisi Onabanjo University, Ago-Iwoye, Nigeria
[5]Department of Computer Science, University of Zululand, Kwadlangezwa 3886, SouthAfrica
E-mail: [1]aafolorunsho@noun.edu.ng; [2]blessingaimuel@gmail.com; [3]aabiodun@noun.edu.ng; [4]ademola.adesina@oouagoiwoye.edu.ng; [5]bright80@gmail.com

ABSTRACT

Booking cancellations have a significant impact in demand-management decisions in the hospitality industry. These cancellations often limit the production of accurate forecast which is a significant tool in terms of revenue management performance. In order to curb these booking cancellation problems, hotels tend to employ rigid cancellation policies and overbooking strategies which could be detrimental to revenue generation and reputation of the hotel. This study is aimed at creating a machine learning model that could, with very high accuracy and precision, predict hotel booking cancellations. The dataset used is based on the individual bookings drawn from a hotel reservation system from a resort hotel in Portugal. The model was built using random forest algorithm with the dataset being split into 80% for the training set and 20% for the test dataset. By addressing booking cancellation prediction as a classification problem in data science context, the findings showed that it is possible to build models for predicting booking cancellations with an accuracy result of over 88%. This allows hotel managers accurately predict net demand and build better forecasts, improve cancellation policies, define better overbooking strategies and hence come up with more realistic and profitable pricing and resources allocation strategies.

Keyword: Machine Learning, Hotel Booking Cancellation, Random Forest Algorithm, Prediction.

CISDI Journal Reference Format
Afolorunso, A.A., Aimuel, B.E., Abiodun, A.O., Adesina, A.O. & Ajagbe, S.A. (2025): Development of a Machine Learning–Enabled Decision Support Framework for Hotel Booking Cancellation Prediction. Computing, Information Systems, Development Informatics and Allied Research Journal. Vol 16 No 3, Pp 23-36. Available online at www.isteams.net/cisdijournal.
dx.doi.org/10.22624/AIMS/CISDI/V16N3P2

1. INTRODUCTION

According to Kimes and Wirtz (2003), revenue management is “the application of information systems and pricing strategies to allocate the right capacity to the right customer at the right price at the right time”. This was originally developed in 1966 in the aviation industry (Chiang et al, 2007) and has been gradually implemented in other service industries, such as hotels, rental cars and golf courses. Hotel bookings used to be a hassle as one would have probably needed to do a walk-in upon arrival before making reservations. This led to a number of problems ranging from unavailability of rooms, getting a room that does not suit your taste and so on. Whereas today, such would be a matter of a few clicks, calls or emails away. For a guest, reservation increases the chances of a better deal for assured accommodation on arrival. For a hotel, reservations ensures a better management of guest experience during peak seasons and otherwise. There are various ways of making or cancelling a reservation, they are categorized as online and offline sources.

Mehrotra and Ruttley (2006) opined that good demand forecasting is an important aspect of revenue management. Morales and Wang (2010) and Ivanov and Zhechev (2012) also identified service demand as one of the aspects where forecasting is necessary. With this need for demand forecasting comes booking cancellations because with industries such as hospitality industry that require advanced bookings this do not represent the true demand for their services, since there are likely going to be a considerable number of cancellations (Liu, 2004; Morales and Wang, 2010; Schwartz et al., 2016; Gomez-Talal et al., 2024; Sun, 2025).

Cancelling a reservation poses a problem for the hotel as it can lead to loss of revenue and inability to reallocate the room as a result of probably short time interval between the cancellation and booking date. This problem can be curtailed by building a system that can predict new/future booking cancellations using machine learning as well as predict their net demand and help in making decisions about which bookings to accept and reject. Machine learning has been a very important tool in creating models that can predict these cancellations and other similar problems. Similar works has been done by several other researchers using various Machine Learning algorithms such as Support Vector Machines (SVM), Hidden Markov Model (HMM), etc. (Pereira, 2016; Liang et al., 2017; Sarica et al., 2017; Antonio et al, 2019a; Chen and Wang, 2020; Alavi and Khosravi, 2023; Jishan et al., 2024).

For this work, random forest model was deployed in building the cancellation prediction model. Random forest is a tree-based algorithm in machine learning that can be applied on both regression and classification problems. The algorithm was first created by Tin Kam Ho in 1995 and was further developed by Leo Breiman in 2001. With random forest, the classification trees can only use a predetermined number of randomly selected explanatory variable when performing splits (James et al., 2013; Yang and Miao, 2024). The rest of this paper is organised thus: Section 2 presents the related works, Section 3 discusses the methodology, Section 4 is the results and discussion while Section 5 concludes the paper.

2. LITERATURE REVIEW

Cole (2020) in his work defined Churning as the rate at which a commercial customer would leave the commercial platform where they are currently a paying customer. Using the logistic regression model, he predicted customers who churned. It has a recall and precision that was lower than expected and concluded that building models is an iterative process and suggested using other methods to improve the score. Zidek, et al. (2017) used various machine learning models. The Artificial Neural Network (ANN) method was found out to give the best result in comparison to the others. But the model did not have an impressive prediction accuracy.

Alotaibi (2020) worked on Application of Machine Learning in the Hotel Industry. In his research, he found out that machine learning is helpful in demand forecasting, price forecasting and work efficiency and it outperforms in the forecast accuracy against the statistical methods. He noted that machine learning provides simplicity but social influence makes it difficult to control the predictions.
Jishan et al. (2024) apply Bayesian modelling techniques to predict hotel booking cancellations using a dataset from Kaggle with 36,285 bookings and 17 features. They compare a Bayesian Logistic Regression model on 14 features and 6,000 randomly selected observations with a Beta-Binomial model. They find that the logistic variant yields superior predictive accuracy.

Key predictive factors identified include number of adults and children, stay duration, lead time before check-in, availability of car parking space, room type, and "special requests". The modelling approach is validated via Leave-One-Out Cross-Validation (LOO-CV), showing strong alignment between observed cancellations and predicted probabilities, suggesting robustness in the Bayesian framework. Generally, the study demonstrates the practicality and effectiveness of Bayesian approaches in booking-cancellation prediction, offering actionable insights for hotel revenue-management and operational decision-making. But while Bayesian models provide interpretability, they are computationally expensive and less scalable than ensemble methods like Random Forest or Gradient Boosting when applied to large real-time hotel booking systems. In an exploratory study from the United Arab Emirates (UAE) on Artificial intelligence (AI) towards hotels' competitive advantage written by Al-shami, et al. (2021), a model was developed that can determine how the interaction between the AI factors and uses influence hotel performance. But it was found out that majority of the hotels in UAE use AI in managing trip planning, reception service and room services. It was suggested that further studies should address this limitation by investigating technical and managerial factors.

3. METHODOLOGY

3.1 Model Architecture

The model architecture is as presented in Figure 1 below. The first stage of the model architecture is business understanding and its objective is to identify the key business factors that the analysis needs to predict. The variables are referred to as the model target variables, and the metrics associated with them are used to determine the success of the model. Defining the success metrics was also a key step in business understanding.

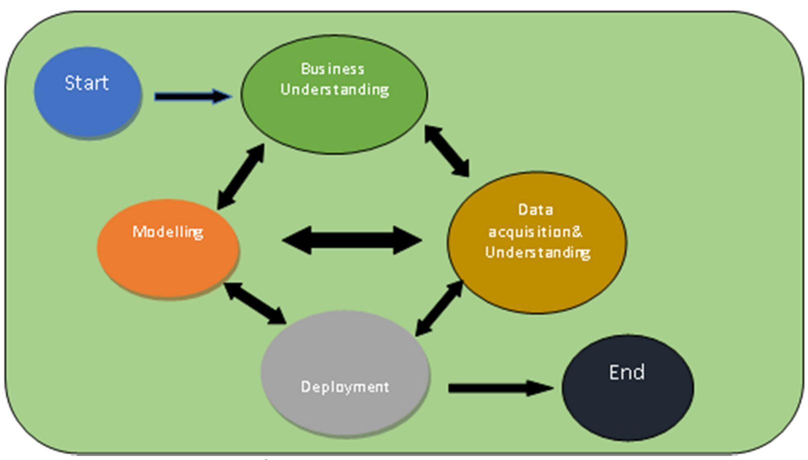

Figure 1: Business Model architecture

The next stage in building the model is the data acquisition and understanding. The dataset used in this study was collected from a real hotel whose identity would remain anonymous in order to protect the customer's identity. R programming language is used in writing the codes for this study. The dataset consists of hotel booking records between 1st July 2015 and 31st August 2017, including both utilized and cancelled bookings. It has a total of 119,390 records combined and 31 variables (attributes). The dataset and their description can be found and downloaded from an open data set (Antonio et al., 2019b) which is available at https://www.sciencedirect.com/science/article/pii/S2352340918315191.

The research system model architecture is as in Figure 2 below. The data was acquired, preprocessed to remove data imbalance and select appropriate features for the model. The model was then trained with and tested with 80% and 20% data respectively.

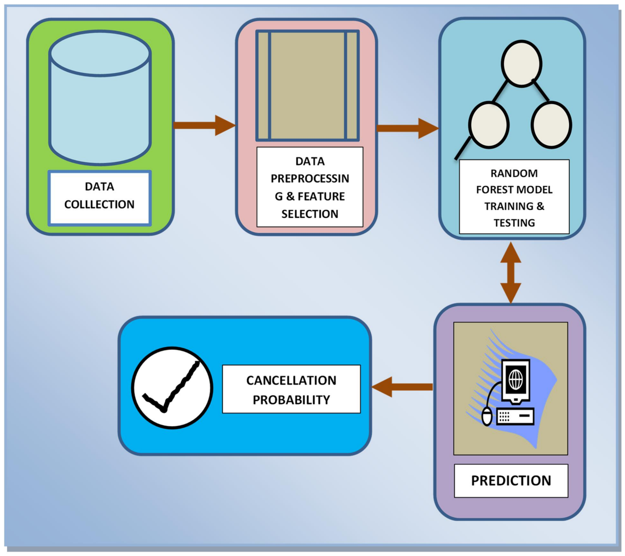

Figure 2: Random Forest Hotel Booking Cancellation Prediction Model

Table 3.1: A sample of the hotel dataset

IsCanceled	0	1	1	1	1
LeadTime	6	88	65	92	100
ArrivalDateYear	2015	2015	2015	2015	2015
ArrivalDateMonth	July	July	July	July	July
ArrivalDateWeekNumber	27	27	27	27	27
ArrivalDateDayOfMonth	1	1	1	1	2
StaysInWeekendNights	0	0	0	2	0
StaysInWeekNights	2	4	4	4	2
Adults	1	2	1	2	2
Children	0	0	0	0	0
Babies	0	0	0	0	0
Meal	HB	BB	BB	BB	BB
Country	PRT	PRT	PRT	PRT	PRT
MarketSegment	Offline TA/TO	Online TA	Online TA	Online TA	Online TA
DistributionChannel	TA/TO	TA/TO	TA/TO	TA/TO	TA/TO
IsRepeatedGuest	0	0	0	0	0
PreviousCancellations	0	0	0	0	0
PreviousBookingsNotCanceled	0	0	0	0	0

ReservedRoomType	A	A	A	A	A
AssignedRoomType	A	A	A	A	A
BookingChanges	0	0	0	0	0
DepositType	No Deposit	No Deposit	No Deposit	No Deposit	No Deposit
Agent	6	9	9	9	9
Company	NULL	NULL	NULL	NULL	NULL
DaysInWaitingList	0	0	0	0	0
CustomerType	Transient	Transient	Transient	Transient	Transient
ADR	0	76.5	68	76.5	76.5
RequiredCarParkingSpaces	0	0	0	0	0
TotalOfSpecialRequests	0	1	1	2	1
ReservationStatus	Check-Out	Canceled	Canceled	Canceled	Canceled
ReservationStatusDate	7/3/2015	7/1/2015	4/30/2015	6/23/2015	4/2/2015

Before training the model, there was a sound understanding and prepossessing of the data. This is because real-world datasets often have missing values or a host of other discrepancies (Anthonio et al., 2019). Using data summarisation and visualisation, the quality of the data was audited and from it the information needed to process the data to get it ready for modelling was extracted. This was an iterative process. After the data was cleaned, the patterns that are intrinsic in the data were studied and analysed. This aided in choosing and developing an appropriate predictive model for the target variables. In the model architecture feature importance was used as evidence to show how the predictor variables affects the target variables and then determine whether there is sufficient data to proceed to the next modelling step, which is model training and evaluation.

The process of model training involves the following steps:

i. Randomly splitting the input data for modeling into a training dataset and a test dataset in a bid to effectively predict new data.
ii. Building the model by using the training dataset.
iii. Evaluating the training and the test data set, using a random forest machine-learning algorithms along with the various associated tuning parameters (known as a parameter sweep) that are geared toward answering the question of interest with the current data.

3.2 Random Forest (RF)

RF is an ensemble learning method that builds multiple decision trees and aggregates their predictions through **voting (classification)** or **averaging (regression).** It is a classifier consisting of a collection of tree-structured classifiers denoted as:

$$\{h(\mathbf{x}, \Theta_k),\ k = 1, 2, \ldots, K\}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

where:

$h(\mathbf{x}, \Theta_k)$ is the k^{th} decision tree trained on random vector Θ_k,
$\{\Theta k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x** (Breiman, 2001), and K is the total number of trees

Then the **Random Forest predictor** is given by:

Classification case:

$\hat{y}$ = mode{$h_1(x)$, $h_2(x)$, ..., $h_K(x)$}, ... (2)
where the predicted $\hat{y}$ is the majority vote among all tree predictions.

Regression case:

$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} h_k(x)$ .... (3)

Which means the final prediction is the **average** of all the individual tree outputs

Error Estimation:
The RF model performance is evaluated using the Out-of-Bag (OOB) metric where a lower OOB error corresponds to a higher OOB score.

The OOB error is computed using equation (4) below:

OOB Error = $\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \neq \hat{y}_{i,OOB})$ .... (4)
Where:
N is the total number of training samples,
$\mathbb{I}$ is the indicator function (0 if condition is false, and 1 otherwise)
$\hat{y}_{i,OOB}$ is the aggregated prediction from trees that are exclusive of sample i.

Feature importance:

The **Mean Decrease in Impurity (MDI)** for a feature X_j is given by equation (5):
Imp(X_j) = $\frac{1}{K}\sum_{k=1}^{K} \sum_{t \in T_k} p(t)\Delta_i(s_t, t)$ .... (5)

Where:
T_k is the set of nodes in tree k,
$p(t)$ is proportion of samples reaching node t,
$\Delta_i(s_t, t)$ is the impurity decrease from splitting node t on feature X_j.

The Random forest was the choice for this research for its desirable characteristics such as:

- Good accuracy, which is many atimes better than Adaboost.
- Relatively robust in capturing outliers and noise.
- Faster than bagging or boosting.
- Renders useful internal estimates of error, strength, correlation and variable importance.
- Simple and easily parallelized.

The general algorithm for the RF is as in Figure 3 below:

General Algorithm for Random Forest

Step 1: Draw K bootstrap samples from the training data.

Step 2: For each sample, grow a decision tree:

- At each node, select a random subset of features $m \ll M$.
- Choose the best feature and split point based on impurity reduction such as Gini index or entropy.

Step 3: Aggregate predictions (majority vote or average).

Step 4: Estimate OOB error for generalization performance.

Figure 3: Random Forest Algorithm

The random forest model is run with the features that are classified as of high importance after feature engineering/importance.

LeadTime	DepositType	Country
4958.66009	4766.57493	4455.71970
ADR	MarketSegment	TotalOfSpecialRequests
2920.05125	2587.08859	2314.52659
ArrivalDateDayOfMonth	ArrivalDateWeekNumber	ArrivalDateMonth
1928.59411	1663.17959	1653.36507
PreviousCancellations	StaysInWeekNights	CustomerType
1351.15133	1291.75866	1167.37666
AssignedRoomType	ArrivalDateYear	RequiredCarParkingSpaces
1130.25733	942.03383	915.45409
BookingChanges	StaysInWeekendNights	ReservedRoomType
837.64541	817.13862	693.43217
Meal	DistributionChannel	Adults
550.90628	516.86112	491.16145
Children	PreviousBookingsNotCanceled	IsRepeatedGuest
237.70990	177.29102	84.08374
DaysInWaitingList	Babies	
77.23932	31.50452	

Figure 4: Feature Importance

```
Call:
 randomForest(formula = IsCanceled ~ ., data = trainbinded, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 11.22%
Confusion matrix:
      0     1 class.error
0 52628  3746  0.06644907
1  6301 26867  0.18997226
```

Figure 5: Accuracy Score

Some of the important parameters as shown in the figure 5 includes:

- The first parameter specifies the formula IsCanceled (It is the feature to be predicted using the remaining columns of data)
- Number of trees; ntree = 500 (ntree defines the number of trees to be generated.
- Number of variables tried at each split; mtry = 5
- The Out-of-bag (OOB) estimate of error rate is a measure of the level of accuracy. It is useful measure to differentiate between different random forest classifiers. For instance, after comparing the number of variables to be considered, the smallest value for this error rate was selected. The random forest model developed in this study has an 88% accuracy.
- The confusion matrix, which is a good way of looking at how well the classifier is performing when presented with new data, displays the result of the Random forest model evaluated on the test data. It also represents counts from the predicted and actual value. The True Negative (value = 52628) is the number of people predicted to cancel their bookings who actually cancelled it; False positive (value = 3746) is the number of people predicted not to cancel their bookings but actually cancelled it; False negative (value = 6301) is the number of people predicted to cancel their bookings who didn't cancel it; True positive (value = 26867) is the number of people predicted to cancel their bookings who actually cancelled it.

3.3 Performance Evaluation Metrics

The model evaluation was carried out using accuracy, precision, F1-score and recall metrics. After performing the classification predictions, we got four types of results, which constitute the confusion matrix and they are:

- True positives: predicting an observation as belonging to a class, which truly belongs to that class.
- True negatives: predicting an observation does not belong to a class, which actually does not belong to that class.
- False positives: occur when the outcom eis predicted to belong to a class when it actually does not.
- False negatives: occur when the outcome is predicted as not belonging to a class when it actually does.

Confusion matrix is often used to compute the accuracy of the machine learning algorithm in classifying the data into corresponding labels. It is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the values are known. In the field of machine learning, the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualisation of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix).

In the final stage, the model was deployed to a production or production-like environment for final user acceptance which in the case of this study was deployed using R Shiny app. Depending on the business requirements, predictions are made either in real-time or in batch. In order to deploy this model, it is exposed with an open API interface. The interface enables the model to be easily consumed from various applications, such as: Online websites, Spreadsheets, Dashboards, Line-of-business applications, Back-end applications.

4. RESULTS AND DISCUSSION

As earlier stated, the accuracy of the random forest model was tested using the confusion matrix, accuracy, precision, recall and F1-Score. The accuracy of the prediction test is 83% which is a very good score. This shows that the model could predict new values accurately. Sensitivity and Specificity mathematically describes the accuracy of a test which reports the presence or absence of a condition. Sensitivity is a measure of how well a test can identify true positives and Specificity is a measure of how well a test can identify the negatives. In all data tests, there is usually a trade-off between the both of them as higher sensitivities will mean low specificities and vice versa.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 17796  3886
         1   996  7170

               Accuracy : 0.8364
                 95% CI : (0.8322, 0.8406)
    No Information Rate : 0.6296
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6294

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9470
            Specificity : 0.6485
         Pos Pred Value : 0.8208
         Neg Pred Value : 0.8780
             Prevalence : 0.6296
         Detection Rate : 0.5962
   Detection Prevalence : 0.7264
      Balanced Accuracy : 0.7978

       'Positive' Class : 0
```

Figure 6: Model Sensitivity

The goal of this study is to identify everyone who would cancel the hotel booking, so this means the number of false negatives should be low which requires a high sensitivity. This means that the customers that are highly likely to cancel would be identified by the test. This is especially important as the consequences of failing to identify them in time would either lead to loss of revenue for the hotel management or inability to resell the room as a result of the short notice. As seen in figure 6, the sensitivity is 94% which is very satisfactory as it shows that the rate of predicted true positives is very high.

Accuracy: The accuracy represents the ratio of correctly classified hotel booking cancellations which is 88.5% (See figure 7).

```
accuracy_score <- (52484+26783)/(52484+3890+26783+6385)
print(accuracy_score)

[1] 0.8852494
```

Figure 7: Booking Cancellation Prediction Accuracy

Precision: The precision of an algorithm is the ratio of correctly classified booking cancellations (TP-True positive) to the total number of bookings predicted to be cancelled. The precision of the random forest model is 87.3% as shown in figure 8.

```
precision <- (26783)/(26783+3890)
print(precision)

[1] 0.8731784
```

Figure 8: Booking Cancellation Prediction Precision

Recall: The recall of the algorithm is the ratio of correctly classified booking cancellations to the total number of bookings that was actually cancelled. The recall is 80.7% for the model as depicted in figure 9.

```
recall <- (26783)/(26783+6385)
print(recall)

[1] 0.8074952
```

Figure 9: Booking Cancellation Prediction Recall

F1-Score, which is the equilibrium between the precision and the recall is given in Figure 10

```
f1_Score <- (2*((0.87*0.80)/(0.87+0.80)))
print(f1_Score)

[1] 0.8335329
```

Figure 10: F1-Score Booking Cancellation Prediction

The performance of the system across all metrics is as shown in figure 11 below:

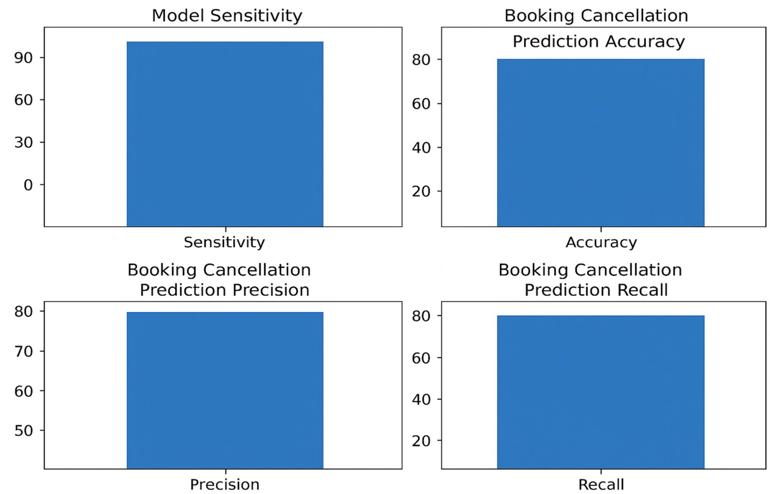

Figure 11: General Performance of the Model across metrics

5. CONCLUSION

This study confirms that bookings with high probability of being cancelled can be predicted. This would encourage hotel management to take measures to prevent possible cancellations, which could either be through discounts, incentives, or additional recreational activities that can be a source of attraction to customers. However, new daily records should be run against this model as it can make future predictions thereby making it possible for hotel operators to predict their net demand and also build stronger overbooking and cancellation strategies. This would lead to lower costs and reduced risks.

This paper greatly contributes to reduce the sparseness of studies in predictive analytics and shows how analytics as a service of decision support systems can be built and deployed. In addition, this study has also demonstrated how data-splitting method selection and having domain knowledge in feature engineering are of utmost importance in machine learning modelling and its influence in improving prediction models. This study provides new empirical evidence of the probability to cancel a hotel booking, based on unique data originating directly from a booking system. This research has shown that lead time, which is the number of days between the booking and the date of arrival, is the most important variable when predicting cancellations with the Random Forest algorithm. A distinctive point of this study is the use of open-source tools such as R to build a cloud-based service-oriented decision support system. The performance of the system and its results prove the usefulness and adequacy of these tools in tackling hotel booking cancellation problems.

The prediction model enables hotel managers to mitigate revenue loss derived from booking cancellations and to reduce the risks associated with overbooking (reallocation costs, cash or service compensations etc.). They would also provide hotel managers the ability to implement less rigid cancellation policies without increasing uncertainty. This would in turn translate to more sales and generate more bookings. Asides from providing hotel managers the ability to take action on bookings likely to be cancelled, it also allows them to produce precise demand forecasts.

It is necessary to understand that one model cannot fit all hotels as different hotels would have different weights on their feature importance and therefore every hotel should have its own model. As seen in figure 4, not all features have the same order of importance, nor do they contribute in the same way towards predicting whether a booking would be cancelled or not. These features may also vary with different hotels i.e., if there is a hotel A with feature “RequiredParkingSpaces”, it may be ranked as of high importance in comparison with hotel B that may have it ranked as the least important which may be as a result of hotel B having limited parking space.

Therefore, hotel revenue management and the business domain is not enough to undertake a good feature selection. It is necessary to also understand individual hotel’s operation and characteristics before applying the model or building a different model with new parameters that is befitting for the task. This would make a difference on the final model performance and adequacy. A lack of quality can affect overall model performance. For this reason, hotels that want to build prediction models must ensure that a data quality policy is in place.

In future studies, if a system can generate new features there is a high probability of improving the model’s performance. The bookings that are acted on are cancelled less frequently than bookings where no action was taken, hence a feature with the indication if and what category of action was taken should be added to the data to improve the models’ performance. Additionally, recording the actions or incentives made in each booking to avoid cancellation (e.g., offering a room upgrade, asking about the bed type preference, offering a spa day) would be useful in another machine learning model in order to recommend the actions that should be taken in the bookings that are predicted as likely to cancel.

This has the potential of prompting the development of a fully automated system. It would be a system that not only can predict a bookings cancellation outcome but can also select which customers should be contacted, whether to make initial contact, and engage in a discussion with the customer via a chatbot, where human intervention would only be required in the occasion the system is not able to provide an answer. Furthermore, future research should explore the development and implementation of machine learning-based systems that can predict overall demand, service delays, social reputation ratings, or slow responses to customers’ requests, and others.

REFERENCES

1. Alavi, M. T., and Khosravi, S. H. (2023). Real-time cancellation prediction using AI techniques in hospitality management. International Journal of Hospitality Management, 98, 102–113
2. Alotaibi, E. (2020). “Application of Machine Learning in the Hotel Industry: A Critical Review”. Journal of Association of Arab Universities for Tourism and Hospitality (JAAUTH), 18(3), 78-96.
3. Al-shami, S.A.H., Mamun, A.A., Ahmed, E.M. and Rashid, N. (2021). “Artificial intelligent towards hotels’ competitive advantage. An exploratory study from the UAE”. https://doi.org/10.1108/FS-01-2021-0014
4. Antonio, N., Almeida, A., and Nunes, L. (2019). “An automated machine learning based decision support system to predict hotel booking cancellations”. In: An automated machine learning based decision support system to predict hotel booking cancellations, 1, 1–20.
5. Antonio, N., de Almeida, A., and Nunes, L. (2019). Hotel booking demand datasets. Data in Brief, 22, 41–49. https://doi.org/10.1016/j.dib.2018.11.126

6. Breiman, L. (2001). “Random forests”. Machine learning. Kluwer Academic Publisher. 45(1), 5–32.
7. Chen, T. H., and Wang, Y. L. (2020). Big data analytics in hotel industry: Predicting cancellation rates. Tourism Management Perspectives, 35, 100–110.
8. Chiang, W.C., Chen, J. C. and Xu, X. (2007). “An overview of research on revenue management: current issues and future research," International Journal of Revenue Management, Inderscience Enterprises Ltd, 1(1), 97-128.
9. Cole, A. (2020). “Predicting customer churn with classification modelling”. https://towardsdatascience.com/customer-churn-analysis-eda-a688c8a166ed.
10. Gómez-Talal, I., Azizsoltani, M., Talón-Ballestero, P., and Singh, A. (2024). Machine learning in hospitality: Interpretable forecasting of booking cancellations. *IEEE Access*, *12*, 1-15. https://doi.org/10.1109/ACCESS.2024.0429000
11. Grogan, M. (2020). “Support vector machine and regression analysis”. Published in Towards Data Science. https://towardsdatascience.com/support-vector-machines-and-regression-analysis-ad5d94ac857f.
12. Ivanov, S. and Zhechev, V. (2012). Hotel marketing (in Bulgarian). Varna: Zangador
13. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). “An introduction to statistical learning with Application in R”, Springer Texts in Statistics (STS, volume 103).
14. Jishan, M. A., Singh, V., Ghosh, A. K., Alam, M. S., Mahmud, K. R., and Paul, B. (2024). Hotel booking cancellation prediction using applied Bayesian models. *arXiv*. https://arxiv.org/abs/2410.16406 arXiv
15. Kimes, S. E., and Wirtz, J. (2003). Has revenue management become acceptable? Findings from an International study on the perceived fairness of rate fences. Journal of Service Research, 6(2), 125–135. DOI:10.1177/1094670503257038
16. Liang, L., Kong, F., Martin, C., Pham, T., Wang, Q., Duncan, .J, and Sun, W. (2017). Machine learning-based 3-D geometry reconstruction and modeling of aortic valve deformation using 3-D computed tomography images. Int J Numer Method Biomed Eng. 33(5). doi: 10.1002/cnm.2827.
17. Liu, P. H. (2004). Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods. In: I. Yeoman and U. McMahon-Beattie (Eds.), Revenue management and pricing: Case studies and applications (pp. 91–101). Cengage Learning EME.
18. Mehrotra, R., and Ruttley, J. (2006). Revenue management (second edition.). Washington, DC, USA: American Hotel and Lodging Association (AHandLA).
19. Morales, D. R., and Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, 202(2), 554–562. https://doi.org/10.1016/j.ejor.2009.06.006
20. Pereira, L.N. (2016). "An introduction to helpful forecasting methods for hotel revenue management", *International Journal of Hospitality Management*, 58, 13-23. DOI:10.1016/j.ijhm.2016.07.003
21. Sarica, A., Ceraca, A. and Quattrone, A. (2017) “Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review". Frontiers in Aging Neuroscience. Gale OneFile: Health and Medicine, link.gale.com/apps/doc/A511188454/HRCA?u=anon~4921db01andsid=googleScholarand xid=6d9055e7.

22. Schwartz, Z., Uysal, M., Webb, T. and Atlin, M. (2016). "Hotel daily occupancy forecasting with competitive sets: a recursive algorithm", International Journal of Contemporary Hospitality Management, 28(2), 267-285. https://doi.org/10.1108/IJCHM-10-2014-0507
23. Sun, J. (2025). Hotel booking cancellation and machine learning. *Advances in Engineering Innovation*, *15*(1), 45-62. https://doi.org/10.54254/2977-3903/2025.20826
24. Ting, K. M. (2017). “Confusion matrix”. Encyclopedia of machine learning and data mining, 260. https://www.researchgate.net/publication/316116586_Confusion_Matrix.
25. Yang, D. and Miao, X. (2024). Predicting hotel booking cancellations using tree-based neural network. *PeerJ Computer Science*, *10*, Article e2473. https://doi.org/10.7717/peerj-cs.2473
26. Zidek, K., Pitel, J. and Hosovsky, A. (2017). “Machine learning algorithms implementation into embedded systems with web application user interface”. 2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES), 77 – 82. doi: 10.1109/INES.2017.8118532.