Jurnal Sistem Informasi Bisnis 02(2025)
Copyright ©2025 , JSINBIS, p-ISSN: 2502-2377, e-ISSN: 2088-3587
On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/71817

Optimization of Prediction for Cancellation of Hotel Room Reservation Using Decision Tree with Feature Selection and Resampling

Eka Rahmawati*, Galih Setiawan Nurohim

Information Systems, Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Jakarta, Indonesia 10450

Submitted: February 7th, 2025; Revised: March 15th, 2025; Accepted: April 13th, 2025; Available Online: May 31st, 2025
DOI : 10.14710/vol15iss2pp211-215

Abstract

The hotel industry is highly competitive and faces challenges, such as fluctuating demand, intense competition, and shifting consumer preferences. One critical issue that hotels frequently encounter is the cancellation of room reservations, which disrupts operational planning and resource management and leads to significant financial losses. Accurately predicting the likelihood of reservation cancellation is essential to mitigate these negative impacts and optimize revenue management strategies. This study focuses on the development of a predictive model for hotel room reservation cancellations using a decision-tree algorithm. The Decision Tree was selected for its ability to manage complex relationships between variables and ease of interpretation, making it accessible to hotel managers without technical expertise. To enhance the performance of the model, a forward selection technique was employed to identify the most relevant features, ensuring a balance between the model complexity and predictive accuracy. Additionally, resampling techniques were applied to address class imbalance in the dataset, which is common in cancellation cases where non-cancelled reservations outnumber cancelled reservations. This study explores the prediction of hotel room reservation cancellations using a decision tree algorithm enhanced by feature selection and resampling. The model achieved an accuracy improvement to 90%, with precision and recall each increasing by 5,5% after applying these techniques. These findings suggest practical applications for improving cancellation predictions and optimizing revenue management strategies for hotels. The study provides insights into how data-driven approaches can enhance decision-making processes within the competitive hospitality industry.

Keywords: Hotel Room Reservation Cancellation; Decision Tree; Predictive Modeling; Feature Selection; Resampling.

1. Introduction

The hospitality industry is one of the most dynamic and competitive sectors. In its operations, hotels face several challenges. One of the challenges that are usually faced is changes in consumer preferences. These challenges can affect the cancellation of room reservations by guests. Frequent cancellations can disrupt operational planning and resource management. Therefore, the ability to accurately predict the likelihood of cancellations is crucial to mitigate negative impacts and maximize revenue for the hotel.

Predicting cancellation of room reservations is an essential part of revenue management in the hotel industry (Yoo et al., 2024). One effective model prediction allows hotels to take preventive measures, such as implementing stricter cancellation policies or offering incentives for guests who may cancel to maintain their reservations(Antonio et al., 2019). Furthermore, accurate predictions enable hotels to implement more effective overbooking strategies, reducing the risk of unintended vacancies (Antonio et al., 2017).

This study focuses on the development of a cancellation prediction model for hotel room reservations using the Decision Tree algorithm. The Decision Tree algorithm was chosen due to its ability to handle complex relationships between input variables and output, especially when data has intricate connections between them (Mienye & Jere, 2024). The algorithm also offers ease of interpretation, allowing hotel managers without strong technical backgrounds to understand the prediction results.

One of the main challenges in developing prediction models is selecting the most relevant and significant features (Kappen et al., 2018). Using too many features can lead to overfitting, where the model becomes too specific to the training data and fails to generalize to new data. Conversely, using too few features can reduce model accuracy. To address this, the present study applies the forward selection technique, allowing for the iterative selection of features based on their contribution to improving prediction accuracy (Hamdard & Lodin, 2023).
Resampling techniques are applied to address the issue of data imbalance, which often occurs in cases of

*) Corresponding author: eka.eat@bsi.ac.id

cancellation reservations. In the dataset used, it is likely that the number of non-cancelled reservations is much higher than the number of cancelled reservations. This imbalance can cause the model to classify most reservations as non-cancelled, reducing prediction accuracy for cancellation cases. By applying resampling techniques, the distribution of data is balanced, allowing the model to be trained more effectively and resulting in more accurate predictions (Kim & Jung, 2023).

Accurately predicting hotel reservation cancellations is critical for optimizing occupancy rates and improving revenue management strategies in the hotel industry. Existing models, however, often fail to integrate feature selection and resampling techniques, leading to less precise predictions. This study addresses this gap by proposing a decision tree model that incorporates these advanced techniques to improve prediction accuracy. By focusing on feature selection and resampling, this research aims to provide a robust framework that enhances cancellation predictions and supports hotel management in decision-making processes.

2. Literature Review/ Related Works

Based on the survey conducted on the Traveloka.com website, information was obtained indicating that there are currently approximately 181 accommodations around Candi Borobudur. These accommodations include a variety of types, ranging from star-rated hotels to homestays. The presence of diverse accommodation options provides an opportunity for visitors to tailor their lodging experience to their individual preferences and needs. Star-rated hotels may offer luxurious facilities and comprehensive services, while homestays provide a more intimate and local experience. In fact, the actual number of accommodations around Candi Borobudur is greater than what is recorded on the website, as not all businesses utilize digital marketing services to manage their operations. In 2020, the government implemented the KSPN (National Strategic Tourism Area) program, which significantly increased the number of accommodations around Candi Borobudur. This program had a positive impact by promoting the growth of homestays as one of its outcomes. With the emergence of homestays, KSPN not only increased the number of accommodations overall, but also provided economic opportunities for local communities involved in the tourism industry. This creates more options for visitors, enriches their travel experiences, and contributes positively to the development of tourism in the area surrounding Candi Borobudur. The KSPN program has been proven to play a strategic role in developing the tourism sector and supporting local economic growth.

In addressing this issue, the manager requires an appropriate decision support system. One approach that can be taken is by utilizing information technology. With the adoption of information technology, the manager can optimize data analysis, gain deeper insights, and make more effective decisions in managing operations. In the study by Chen et al., predictions of hotel guest cancellations were made using a machine learning approach that combined Bayesian network algorithms and Regression Lasso. This approach provided significant results, offering deep insights into the probabilities of cancellations of reservations (Chen et al., 2023).

In the latest study, Yoo and his team utilized the XGBoost algorithm as the main foundation for predicting cancellations of hotel room reservations. Interestingly, they employed an ensemble method by combining three additional algorithms, namely Support Vector Machine (SVM), RandomForest, and XGBoost itself. Through this approach, the research results showed that the accuracy of predictions for hotel room cancellations reached a significant level (Yoo et al., 2024). In other research studies, Sarilidis et al. employed the decision tree algorithm due to its ability to provide clear interpretations of decisions (Sarailidis et al., 2023) This algorithm is effective in mapping complex patterns that influence hotel room cancellations, with an advantage in breaking down complex problems into simple decisions. The algorithm not only provides accurate results, but also easy-to-understand interpretations (Herrera et al., 2024).

Although accuracy may be achieved in some cases, it may not reach the desired level or be too low in other instances. Therefore, efforts should be made to improve it, and one strategy that can be applied is using the resample techniquekup (Kuptametee & Aunsri, 2022) This technique allows for adjustments to be made to the data distribution that may be unbalanced, enabling machine learning models to better handle variations and complexity in the data. By implementing the resample technique, it is hoped that the model's performance in predicting hotel room cancellations can be improved and more reliable results can be achieved (Aldoseri et al., 2023).

This study uses decision tree algorithm, feature selection, and resampling techniques to address class imbalance.

2.1. Decision Tree

A decision tree is a predictive modelling technique used in machine learning that recursively splits the dataset into branches to form a tree-like structure, enabling classification or regression based on decision rules derived from the data(Shahrizan et al., 2023). J48 is an implementation of the C4.5 algorithm, a type of decision tree used in machine learning for classification tasks, which generates a decision tree by splitting data into subsets based on attribute values, aiming to optimize the classification accuracy (Asim Shahid et al., 2024).

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/71817

2.2. Feature Selection

Feature selection is a process in machine learning that involves identifying and selecting the most relevant features from a dataset, with the goal of improving model performance by reducing dimensionality, enhancing accuracy, and minimizing overfitting (El-Hasnony et al., 2020). Forward selection is a stepwise feature selection method in machine learning where features are iteratively added to the model one at a time, starting with the most significant feature, to enhance model performance by identifying the most relevant subset of features.

2.3. Resampling

Resampling techniques, such as oversampling and undersampling, are crucial in machine learning for balancing class distributions within a dataset (Khushi et al., 2021). By increasing the frequency of the minority class or reducing the instances of the majority class, these methods aim to mitigate the bias introduced by imbalanced data. Consequently, the application of resampling enhances the model's capacity to generalize and accurately classify underrepresented classes.

3. Method

A dataset comprising 2,000 hotel reservations was obtained from Kaggle, including various attributes, such as the number of adults, number of children, number of weekend nights, number of weeknights, type of meal plan, required car parking space, room type reserved, lead time, arrival year, arrival month, arrival date, market segment type, repeated guest status, number of previous cancellations, number of previous bookings not canceled, average price per room, number of special requests, and booking status. This process involved utilizing data from Kaggle, following the example of researchers like Kuo and Chang, who used Emergency Medical Service Data from the platform as a sample for their study (Kuo et al., 2023).

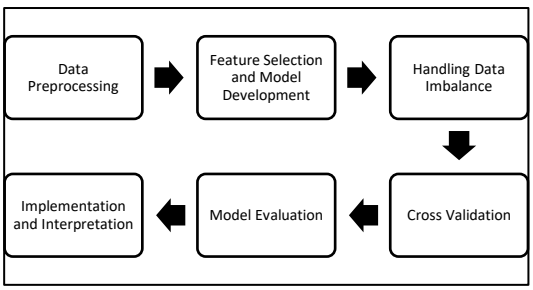

Figure 1. Proposed Method.

Figure 1 shows that this study employed a systematic approach to develop and optimize a prediction model for hotel room reservation cancellations using the Decision Tree algorithm. The methodology consists of key stages: data collection, data preprocessing, feature selection and model development, and evaluation.

1. Data Preprocessing
   Preprocessing involved data cleaning, feature encoding and normalization (Maharana et al., 2022).
2. Feature Selection and Model Development
   Feature selection and model development were integrated:
   a. Feature Selection: Forward selection was used to iteratively add features that improved prediction accuracy.
   b. Model Development: The Decision Tree algorithm was trained on an 80/20 training-validation split, with hyperparameter tuning to optimize performance.
3. Handling Data Imbalance
   Resampling techniques were applied Oversampling of cancelled reservations and under sampling of non-cancelled reservations were used to balance the dataset, improving model accuracy in predicting cancellations.
4. Cross-Validation
   Cross-validation was performed to validate the model's robustness, reducing the risk of overfitting by training on multiple data subsets.
5. Model Evaluation
   The model was evaluated using accuracy, precision, recall, and F-Measure. A confusion matrix was generated to further assess performance, ensuring the model effectively distinguished between cancelled and non-cancelled reservations.
6. Implementation and Interpretation
   The final model was applied to a test set, and the results were analysed to guide hotel revenue management strategies, focusing on optimizing booking practices and reducing the impact of cancellations.

4. Results and Discussion

The first step undertaken is data preprocessing. The steps involved in data preprocessing include:

1. Data Cleaning: Addressing missing values, correcting inconsistencies, and removing irrelevant or duplicate records.
2. Feature Encoding: Converting categorical features (e.g., room type, payment method) into numerical formats using one-hot encoding.
3. Normalization: Scaling numerical features to ensure consistency and improve algorithm performance.

Next, experiments were conducted using the decision tree algorithm (J48) without performing feature selection. The results of the initial testing are presented in Table 1.

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/71817

Table 1. Results of Initial Testing.

Evaluation	Value
Accuracy	85,3%
Precision	0,850
Recall	0,853
F-Measure	0,851

Table 1 demonstrates the accuracy of the decision tree method, which was recorded at 85.3%, accompanied by a precision value of 0.850, recall value of 0.853, and F-measure of 0.851. The next step is the forward selection. The purpose of forward selection is to iteratively select the most significant features by adding them individually to the model. This process aims to identify the subset of features that contribute the most to the predictive performance, thereby enhancing the accuracy and efficiency of the model. The results of the feature selection are presented in Table 2.

Table 2. Results of Forward Selection.

No	Name of Attributes
1	no of week nights
2	required car parking space
3	lead time
4	arrival month
5	market segment type

Table 2 indicates that, out of the 17 attributes tested, forward selection identified five key attributes that significantly influenced performance. After obtaining the selected features, we applied a resampling technique to address the class imbalance. In this dataset, the attribute serving as the class was booking. The results of the testing are presented in Table 3.

Table 3. Results of Testing with Resampling.

Evaluation	Value
Accuracy	90%
Precision	0,899
Recall	0,900
F-Measure	0,899

Based on the test data presented in Table 3, the test results indicate an improvement in the accuracy, recall, precision, and F-measure. An increase in accuracy demonstrates that the model is more effective at correctly classifying instances, indicating an improvement in the model's general predictive capability. An increase in recall indicates that the model is more successful in identifying true-positive instances, reflecting enhanced sensitivity to actual positive cases within the dataset.

An increase in precision signifies that the model has improved in accurately predicting positive cases, reducing the occurrence of false positives and thereby increasing the reliability of its positive classifications. An increase in the F-measure suggests a better balance between precision and recall, indicating that the model's performance has improved in scenarios where both the identification of true positives and the reduction of false positives are critical.

The next step in the evaluation was the confusion matrix. Table 4 shows the results of Confusion Matrix for each test.

Table 4. Results of Confusion Matrix**.**

Algorithms		Predicted Class: Not Canceled	Canceled
Decision Tree	NotCanceled	1255	114
	Canceled	180	451
Decision Tree+Resampling	NotCanceled	1284	85
	Canceled	115	516

The confusion matrix in Table 4 evaluates the performance of a classification model by detailing the true positives, true negatives, false positives, and false negatives. It helps calculate various performance metrics, such as accuracy, precision, and recall, and identifies specific areas where the model makes errors. This information is crucial for comparing models and improving their predictive accuracy. Table 4 demonstrates that the implementation of the Decision Tree algorithm combined with resampling techniques leads to an increase in the number of true positive predictions.

5. Conclusion

This study demonstrates that integrating feature selection and resampling techniques with a decision tree model significantly improves the prediction of hotel reservation cancellations. The practical implications of these findings are clear: hotel managers can apply this model to optimize revenue by reducing last-minute cancellations and better managing room availability. Future research should focus on applying this model to other hotel sectors and exploring alternative algorithms, such as neural networks, to further improve accuracy and generalizability. Additionally, incorporating real-time data could enhance the model's adaptability in dynamic hotel environments.

Acknowledgment

We express our sincere gratitude to the Ministry of Education, Culture, Research, and Technology for their invaluable support and resources provided throughout this research. Their commitment to advancing educational and technological research has been instrumental in the successful completion of this study.

References

Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. In

Applied Sciences (Switzerland) (Vol. 13, Issue 12). MDPI. https://doi.org/10.3390/app13127082

Antonio, N., Almeida, A. de, & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, *13*(2), 25–39. https://doi.org/10.18089/tms.2017.13203

Antonio, N., de Almeida, A., & Nunes, L. (2019). Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. *Cornell Hospitality Quarterly*, *60*(4), 298–319. https://doi.org/10.1177/1938965519851466

Asim Shahid, M., Alam, M. M., & Mohd Su'ud, M. (2024). A fact based analysis of decision trees for improving reliability in cloud computing. *PLoS ONE*, *19*(12 December). https://doi.org/10.1371/journal.pone.0311089

Chen, S., Ngai, E. W. T., Ku, Y., Xu, Z., Gou, X., & Zhang, C. (2023). Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. *Decision Support Systems*, *170*. https://doi.org/10.1016/j.dss.2023.113959

El-Hasnony, I. M., Barakat, S. I., Elhoseny, M., & Mostafa, R. R. (2020). Improved Feature Selection Model for Big Data Analytics. *IEEE Access*, *8*, 66989–67004. https://doi.org/10.1109/ACCESS.2020.2986232

Hamdard, Asst. P. M. S., & Lodin, Asst. P. H. (2023). Effect of Feature Selection on the Accuracy of Machine Learning Model. *International Journal Of Multidisciplinary Research And Analysis*, *06*(09). https://doi.org/10.47191/ijmra/v6-i9-66

Herrera, A., Arroyo, Á., Jiménez, A., & Herrero, Á. (2024). Forecasting hotel cancellations through machine learning. *Expert Systems*. https://doi.org/10.1111/exsy.13608

Kappen, T. H., van Klei, W. A., van Wolfswinkel, L., Kalkman, C. J., Vergouwe, Y., & Moons, K. G. M. (2018). Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and Prognostic Research*, *2*(1). https://doi.org/10.1186/s41512-018-0033-6

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*, *9*, 109960–109975.https://doi.org/10.1109/ACCESS.2021.3102399

Kim, A., & Jung, I. (2023). Optimal selection of resampling methods for imbalanced data with high complexity. *PLoS ONE*, *18*(7 July). https://doi.org/10.1371/journal.pone.0288540

Kuo, C.-Y., Tung Chang, Y., & Chang, Y.-T. (2023). *Introduction of spatial analysis approaches for Emergency Medical Services data: Using the Kaggle dataset as an example Paramedicine and Emergency Response Introduction of spatial analysis approaches for Emergency Medical Ser-vices data: Using the Kaggle dataset as an example*. https://doi.org/10.30216/JPER.202310_(4).0004

Kuptametee, C., & Aunsri, N. (2022). A review of resampling techniques in particle filtering framework. In *Measurement: Journal of the International Measurement Confederation* (Vol. 193). Elsevier B.V. https://doi.org/10.1016/j.measurement.2022.110836

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*, *12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838

Sarailidis, G., Wagener, T., & Pianosi, F. (2023). Integrating scientific knowledge into machine learning using interactive decision trees. *Computers and Geosciences*, *170*. https://doi.org/10.1016/j.cageo.2022.105248

Shahrizan, M., Rahman, A., Azliana, N., Jamaludin, A., Zainol, Z., Mohd, T., & Sembok, T. (2023). The Application of Decision Tree Classification Algorithm on Decision-Making for Upstream Business. *IJACSA) International Journal of Advanced Computer Science and Applications*, *14*(8), 2023. https://dx.doi.org/10.14569/IJACSA.2023.0140873

Yoo, M., Singh, A. K., & Loewy, N. (2024). Predicting hotel booking cancelation with machine learning techniques. *Journal of Hospitality and Tourism Technology*, *15*(1), 54–69. https://doi.org/10.1108/JHTT-07-2022-0227