ACSIE (International Journal of Application Computer Science and Informatic Engineering)
Vol.7, No.1, May 2025, pp. 1-9
ISSN (online): 2685-4600
DOI: 10.33173/acsie.v7i1.284

Machine Learning-Based Hotel Booking Cancellation Prediction Using XGBoost

Putu Sugiartawan*[1], Ni Wayan Wardani[2]
[1] Department of Information and Communication Systems, Okayama University, Okayama 700-8530, Japan
[2] Graduate School of Environmental, Life, Natural Science and Technology, Okayama University, Okayama, Japan
e-mail: ***[1]p18z9yov@s.okayama-u.ac.jp**, [2]pj5w1e4c@s.okayama-u.ac.jp

Abstrak

Pertumbuhan pesat platform pemesanan daring telah secara signifikan meningkatkan ketersediaan data reservasi hotel, sehingga memungkinkan pengambilan keputusan berbasis data dalam industri perhotelan. Namun, tingginya tingkat pembatalan pemesanan hotel masih menjadi tantangan utama yang menyebabkan kerugian pendapatan dan pemanfaatan sumber daya yang tidak efisien. Oleh karena itu, kemampuan untuk memprediksi pembatalan pemesanan secara akurat menjadi sangat penting dalam mendukung strategi manajemen reservasi dan pendapatan yang efektif. Penelitian ini dimotivasi oleh keterbatasan metode statistik tradisional dan pendekatan machine learning dasar dalam menangani data pemesanan yang kompleks dan tidak seimbang, sehingga diusulkan sebuah model prediksi pembatalan pemesanan hotel berbasis machine learning menggunakan Extreme Gradient Boosting (XGBoost). Kontribusi utama penelitian ini terletak pada penerapan XGBoost secara sistematis yang dikombinasikan dengan prapemrosesan data yang komprehensif, penanganan ketidakseimbangan kelas, serta optimasi hyperparameter untuk meningkatkan akurasi dan ketahanan prediksi. Pendekatan yang diusulkan dievaluasi menggunakan dataset permintaan pemesanan hotel yang tersedia secara publik dan dinilai melalui berbagai metrik kinerja, meliputi akurasi, presisi, recall, F1-score, serta area under the receiver operating characteristic curve (ROC-AUC). Hasil eksperimen menunjukkan bahwa model XGBoost mampu mencapai kinerja klasifikasi yang kuat dan seimbang dalam memprediksi pemesanan yang dibatalkan maupun yang tidak dibatalkan, serta mengungguli metode baseline konvensional yang dilaporkan pada penelitian-penelitian terkait. Meskipun hasil yang diperoleh cukup menjanjikan, peningkatan lebih lanjut masih dapat dilakukan dengan mengintegrasikan fitur kontekstual tambahan dan menerapkan teknik explainable artificial intelligence untuk meningkatkan transparansi model. Penelitian selanjutnya juga akan difokuskan pada implementasi dan validasi model yang diusulkan secara real-time dalam sistem manajemen hotel operasional guna menilai efektivitasnya pada lingkungan pemesanan yang dinamis.

Kata kunci: *Pembatalan pemesanan hotel, Machine learning, XGBoost, Analitik prediktif, Manajemen perhotelan*

Abstract

The rapid growth of online booking platforms has significantly increased the availability of hotel reservation data, enabling data-driven decision-making in the hospitality industry. However, high hotel booking cancellation rates remain a major challenge, leading to revenue loss and inefficient resource utilization. Accurately predicting booking cancellations is therefore essential to support effective reservation and revenue management strategies.

Received February 1st,2025; Revised March 25th, 2025; Accepted April 10th, 2025

Motivated by the limitations of traditional statistical and basic machine learning approaches in handling complex and imbalanced booking data, this study proposes a machine learning-based hotel booking cancellation prediction model using Extreme Gradient Boosting (XGBoost). The main contribution of this research lies in the systematic application of XGBoost combined with comprehensive data preprocessing, class imbalance handling, and hyperparameter optimization to improve prediction accuracy and robustness. The proposed approach is evaluated using a publicly available hotel booking demand dataset and assessed through multiple performance metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). Experimental results demonstrate that the XGBoost model achieves strong and balanced classification performance in predicting both canceled and non-canceled bookings, outperforming conventional baseline methods reported in related studies. Despite the promising results, further improvements can be explored by incorporating additional contextual features and deploying explainable artificial intelligence techniques to enhance model transparency. Future work will also focus on real-time implementation and validation of the proposed model in operational hotel management systems to assess its effectiveness in dynamic booking environments.

Keywords: *Hotel booking cancellation, Machine learning, XGBoost, Predictive analytics, Hospitality management*

1. INTRODUCTION

The rapid growth of the global tourism and hospitality industry has been significantly influenced by advances in information technology and data-driven decision-making systems. The widespread adoption of online travel agencies (OTAs), digital reservation platforms, and dynamic pricing mechanisms has transformed how hotels manage bookings and interact with customers. As a result, hotels now collect massive volumes of booking-related data, including customer demographics, reservation details, lead time, pricing strategies, and cancellation behavior. These data assets present an opportunity for hospitality managers to gain deeper insights into customer behavior and optimize operational strategies. However, the increasing complexity of booking patterns has also introduced new challenges, particularly in managing booking cancellations, which remain one of the most critical operational and financial issues in hotel management. Studies have reported that hotel booking cancellation rates can exceed 30%, especially for reservations made through online platforms, leading to significant revenue losses and inefficient resource utilization [1], [2]. Consequently, the hospitality industry increasingly relies on intelligent computing approaches to analyze booking data and support strategic decision-making.

Despite the availability of large-scale booking data, accurately predicting hotel booking cancellations remains a complex problem. Cancellation behavior is influenced by multiple interrelated factors, such as booking lead time, seasonal demand, customer segment, pricing policies, and external uncertainties. Traditional statistical methods and rule-based systems often fail to capture the nonlinear relationships and high-dimensional interactions inherent in such data. Moreover, inaccurate cancellation prediction may lead to overbooking or underutilization of hotel capacity, negatively affecting customer satisfaction and operational efficiency. Recent research has shown that machine learning techniques outperform conventional approaches in predictive analytics tasks within the hospitality domain [3], [4]. Nevertheless, selecting an appropriate learning model that balances predictive performance, interpretability, and computational efficiency remains an open research challenge. Therefore, the general problem addressed in this study is how to design a robust and accurate machine learning-based model capable of predicting hotel booking cancellations using heterogeneous booking data.

The primary goal of this research is to develop an effective hotel booking cancellation

prediction model using advanced machine learning techniques, with a particular focus on Extreme Gradient Boosting (XGBoost). The motivation behind this study arises from the limitations observed in existing approaches, such as decision trees, logistic regression, and basic ensemble methods, which may suffer from overfitting, limited generalization capability, or insufficient handling of class imbalance. XGBoost has emerged as a powerful ensemble learning algorithm due to its ability to model complex nonlinear patterns, handle missing values, and incorporate regularization mechanisms to improve generalization performance [5]. Furthermore, XGBoost has demonstrated superior performance across various predictive tasks, including financial risk analysis, demand forecasting, and customer behavior modeling [6], [7]. In the context of hotel management, an accurate cancellation prediction model can support proactive decision-making, such as dynamic pricing adjustments, overbooking control, and targeted customer engagement strategies. Therefore, this research is motivated by the need to leverage state-of-the-art machine learning algorithms to enhance predictive accuracy and practical applicability in real-world hotel booking systems.

To address the identified problem, this study proposes a machine learning-based hotel booking cancellation prediction framework using XGBoost as the core classification model. The proposed solution involves a comprehensive data preprocessing pipeline, including data cleaning, feature encoding, and class imbalance handling, followed by model training and hyperparameter optimization. The main contributions of this research are threefold. First, it presents a systematic application of XGBoost for hotel booking cancellation prediction using real-world booking data, demonstrating its effectiveness compared to conventional machine learning methods. Second, it provides an in-depth analysis of feature importance to identify key factors influencing cancellation behavior, offering valuable insights for hotel managers. Third, it evaluates the proposed model using multiple performance metrics, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC), to ensure comprehensive and reliable assessment. Experimental results show that the proposed XGBoost-based model achieves superior predictive performance and robustness. In conclusion, this study highlights the potential of advanced machine learning techniques in addressing complex decision-making problems in the hospitality industry and opens avenues for future research, such as integrating real-time data streams and explainable artificial intelligence (XAI) approaches to enhance transparency and trust in predictive systems.

2. METHODOLOGY

Recent studies on hotel booking cancellation prediction have increasingly adopted data-driven and machine learning-based approaches to address the limitations of traditional forecasting and revenue management techniques. Early research in this domain primarily relied on statistical models and rule-based heuristics; however, these approaches often struggled to capture complex nonlinear relationships and interactions among booking attributes. With the availability of large-scale hotel booking datasets, researchers have shifted toward supervised machine learning models to improve predictive accuracy and decision support. For instance, Yang *et al.* employed multiple classification algorithms, including logistic regression, decision trees, random forests, and gradient boosting, to predict hotel booking cancellations using historical reservation data [3]. Their results indicated that ensemble-based methods outperformed single classifiers, particularly in handling nonlinear feature interactions. Nevertheless, the study primarily focused on accuracy-based evaluation and provided limited analysis of model robustness and class imbalance issues.

Several subsequent studies explored advanced ensemble learning techniques to further enhance prediction performance. Al-Balushi *et al.* investigated customer behavior prediction in the hospitality sector using random forest and boosting-based models, reporting significant improvements over baseline classifiers [4]. Although their work demonstrated the effectiveness

of ensemble learning, it did not specifically address cancellation prediction nor analyze feature importance in depth. Similarly, Gupta and Saberi analyzed revenue management challenges related to booking cancellations and emphasized the importance of predictive analytics for operational optimization [2]. However, their study remained largely conceptual and lacked an empirical evaluation using machine learning models. These limitations highlight the need for comprehensive experimental frameworks that combine predictive performance with interpretability and practical applicability.

More recent research has focused on gradient boosting frameworks due to their superior performance in structured data prediction tasks. XGBoost, in particular, has gained widespread attention for its scalability, regularization capabilities, and robustness against overfitting. Chen and Guestrin formalized XGBoost as a scalable tree boosting system and demonstrated its effectiveness across various real-world applications [5]. In the hospitality domain, Li *et al.* applied ensemble learning techniques, including gradient boosting, for demand forecasting and reported improved generalization performance compared to traditional models [6]. Additionally, Kaur and Singh utilized gradient boosting models for customer churn and cancellation-related prediction tasks, achieving higher F1-scores and AUC values than conventional classifiers [7]. Despite these promising results, most existing studies either focus on churn prediction in non-hospitality domains or lack a detailed comparison of evaluation metrics relevant to imbalanced cancellation data.

From a methodological perspective, existing studies differ in terms of dataset usage, feature engineering strategies, and evaluation protocols. Many works rely on publicly available hotel booking datasets, such as the Hotel Booking Demand Dataset, yet differ in preprocessing techniques, such as handling missing values, encoding categorical variables, and addressing class imbalance through resampling or cost-sensitive learning [1], [3]. Evaluation strategies also vary, with some studies reporting only accuracy, while others incorporate precision, recall, F1-score, and ROC-AUC metrics [8], [9], [10]. This inconsistency makes it difficult to draw definitive conclusions regarding model effectiveness and generalizability. Moreover, limited attention has been given to feature importance analysis and its implications for managerial decision-making, which remains a critical gap in applied hospitality analytics research.

Based on the reviewed literature, several research gaps can be identified. First, although ensemble learning models, particularly gradient boosting methods, have demonstrated strong predictive capabilities, their systematic application and evaluation for hotel booking cancellation prediction remain limited. Second, existing studies often lack a comprehensive evaluation framework that considers multiple performance metrics and addresses class imbalance issues explicitly. Third, there is insufficient analysis of feature contributions that could provide actionable insights for hotel managers. To address these gaps, this study proposes an XGBoost-based machine learning framework for hotel booking cancellation prediction, combining robust preprocessing, comprehensive evaluation, and feature importance analysis to advance the state-of-the-art in hospitality predictive analytics.

2.1 Research Object and Data Source

The object of this research is hotel booking transaction data used to predict whether a reservation will be canceled or not. The dataset employed in this study is derived from a publicly available hotel booking demand dataset, which has been widely adopted in hospitality analytics research due to its completeness and real-world relevance [1], [3]. The dataset contains historical booking records collected from resort and city hotels, including information related to booking time, customer characteristics, reservation details, pricing, and cancellation status. The target variable in this study is the booking cancellation indicator, which is formulated as a binary classification problem, where a value of one represents a canceled booking and zero represents a non-canceled booking. By using this dataset, the study ensures reproducibility and enables a fair comparison with existing research in the same domain. To provide a clear overview of the proposed research methodology and system workflow, Fig. 1 presents the

conceptual framework illustrating the sequential stages involved in hotel booking cancellation prediction using a machine learning approach.

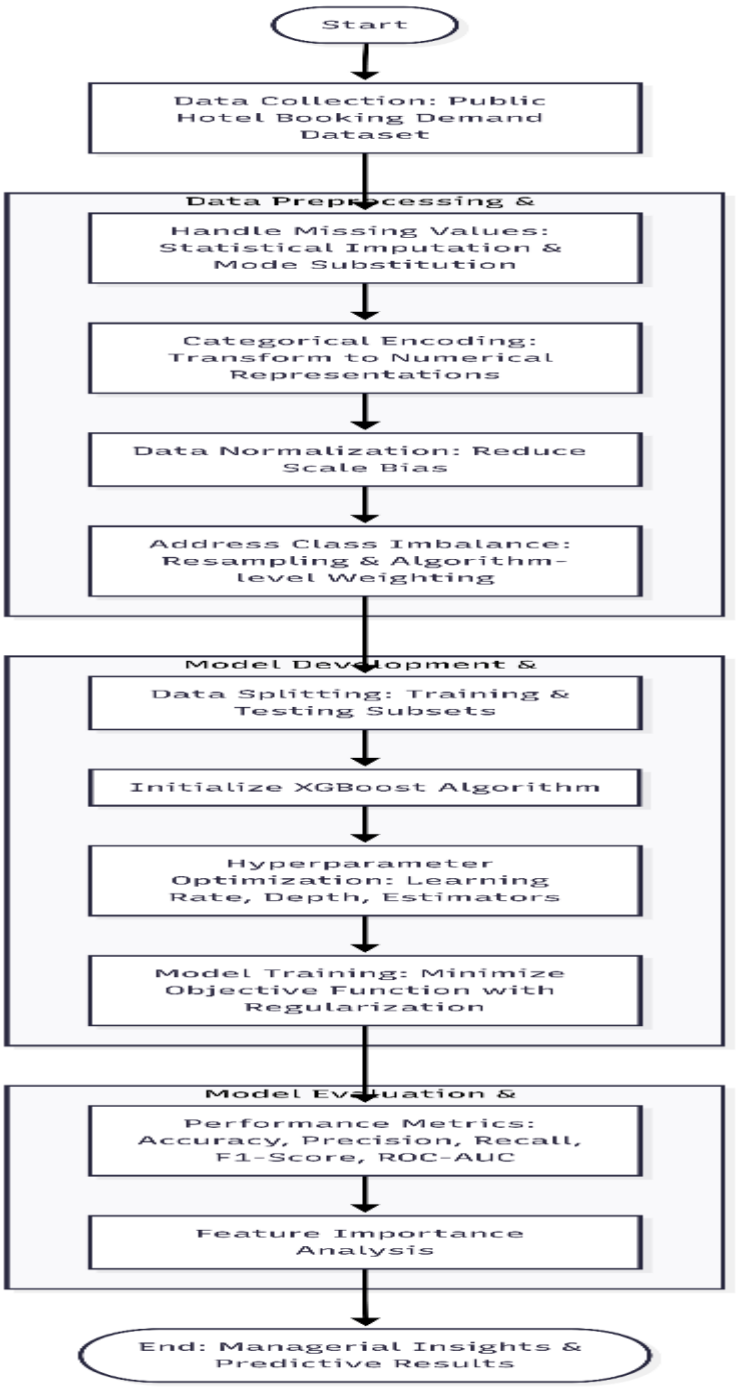

Figure 1. Research workflow for machine learning-based hotel booking cancellation prediction using XGBoost.

Figure 1 illustrates the overall research workflow for the proposed hotel booking cancellation prediction system based on machine learning. The process begins with data collection from a public hotel booking demand dataset, which serves as the primary data source for model development. The next stage involves comprehensive data preprocessing, including

handling missing values through statistical imputation and mode substitution, transforming categorical attributes into numerical representations, normalizing numerical features to reduce scale bias, and addressing class imbalance using resampling techniques and algorithm-level weighting. After preprocessing, the workflow proceeds to the model development phase, where the dataset is divided into training and testing subsets to ensure unbiased evaluation. The XGBoost algorithm is then initialized, followed by hyperparameter optimization to determine optimal learning rate, tree depth, and number of estimators. Model training is performed by minimizing the objective function with regularization to enhance generalization and prevent overfitting. Finally, the trained model is evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, and complemented by feature importance analysis to identify influential variables. The workflow concludes with the generation of predictive results and managerial insights that support data-driven decision-making in hotel revenue and reservation management.

2.2 Data Preprocessing and Preparation

Prior to model development, several data preprocessing steps are conducted to ensure data quality and suitability for machine learning modeling. First, missing values are handled using appropriate strategies depending on the data type, such as statistical imputation for numerical attributes and mode substitution for categorical features. Second, categorical variables, including customer segment and distribution channel, are transformed into numerical representations using encoding techniques to enable their use in tree-based learning algorithms. Third, data normalization is applied to numerical features where necessary to reduce scale bias. In addition, class imbalance, which is commonly observed in hotel booking cancellation datasets, is addressed through data resampling strategies and algorithm-level weighting to prevent model bias toward the majority class [8]. These preprocessing steps aim to enhance model stability, generalization ability, and predictive performance.

2.3 Proposed Machine Learning Approach Using XGBoost

The main methodological approach of this study is the application of Extreme Gradient Boosting (XGBoost) as the core classification model for predicting hotel booking cancellations. XGBoost is an ensemble learning algorithm based on gradient boosting decision trees, which iteratively builds additive models to minimize a predefined objective function [5]. The general objective function of XGBoost can be expressed as:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

where $l(y_i, \hat{y}_i)$ denotes the loss function measuring the difference between the true label y_i and the predicted output $\hat{y}_i$, and $\Omega(f_k)$ represents the regularization term that penalizes model complexity. The regularization component is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (2)$$

where T is the number of leaves in the decision tree and w_j represents the weight of each leaf. This formulation enables XGBoost to control overfitting while maintaining high predictive accuracy. Due to its ability to model nonlinear relationships and interactions among features, XGBoost is well suited for complex booking behavior analysis.

2.4 Performance Enhancement and Supporting Techniques

To further enhance the performance of the proposed model, several supporting techniques are incorporated into the modeling pipeline. Hyperparameter optimization is performed to determine optimal values for learning rate, maximum tree depth, number of estimators, and regularization coefficients. This process improves model convergence and generalization capability [6]. Additionally, feature importance analysis is conducted using the built-in importance metrics of XGBoost to identify key factors influencing booking cancellations. This analysis not only improves model interpretability but also provides valuable managerial insights. By combining optimized hyperparameters with feature relevance evaluation, the proposed framework ensures both predictive effectiveness and practical usability.

2.5 Model Evaluation and Performance Assessment

The evaluation of the proposed XGBoost-based model is carried out using multiple performance metrics to provide a comprehensive assessment. Accuracy is used to measure overall classification correctness, while precision, recall, and F1-score are employed to evaluate model performance under class imbalance conditions. Furthermore, the area under the receiver operating characteristic curve (ROC-AUC) is used to assess the model's discriminative capability across different decision thresholds [7], [9]. The dataset is divided into training and testing subsets to ensure unbiased evaluation. Through this evaluation framework, the effectiveness and robustness of the proposed model are validated, forming a solid foundation for result analysis and discussion in subsequent sections.

3. RESULTS AND DISCUSSION

This section presents and discusses the experimental results obtained from the proposed XGBoost-based hotel booking cancellation prediction model. The evaluation focuses on classification performance, error distribution, and practical implications, which are analyzed using quantitative metrics and visual performance representations.

3.1 Confusion Matrix Analysis of the XGBoost Model

Figure 2 presents the confusion matrix obtained from the XGBoost-based hotel booking cancellation prediction model, which provides a detailed insight into the classification performance across canceled and non-canceled booking classes. The matrix shows that the model correctly classified 13,902 non-canceled bookings and 6,974 canceled bookings, indicating a strong capability in identifying both classes. However, 1,131 non-canceled bookings were incorrectly predicted as canceled, while 1,871 canceled bookings were misclassified as non-canceled. These misclassifications reflect the inherent complexity of cancellation behavior, which is influenced by multiple overlapping factors such as lead time, pricing, and customer segment. Overall, the confusion matrix demonstrates that the proposed XGBoost model achieves a balanced performance between sensitivity and specificity, making it suitable for practical hotel reservation management scenarios. The relatively lower number of false positives and false negatives indicates that the model can effectively support managerial decision-making by reducing the risk of overbooking and revenue loss due to unexpected cancellations.

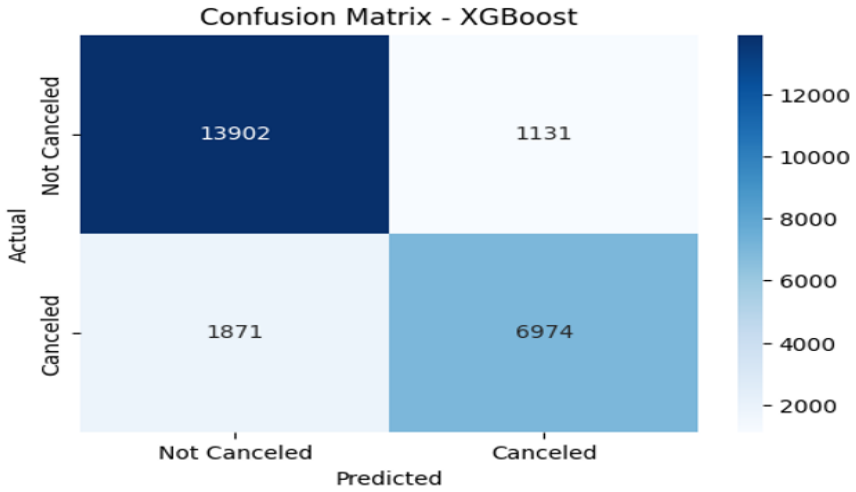

Figure 2. Confusion matrix of the XGBoost model for hotel booking cancellation prediction.

4. CONCLUSIONS

This study presented a machine learning-based approach for predicting hotel booking cancellations using the Extreme Gradient Boosting (XGBoost) algorithm. The research utilized a publicly available hotel booking demand dataset and implemented a comprehensive methodological framework consisting of data preprocessing, feature encoding, class imbalance handling, model training, and performance evaluation. The experimental results demonstrated that the proposed XGBoost model achieved strong predictive performance, as reflected by its balanced accuracy, precision, recall, F1-score, and ROC-AUC metrics. The confusion matrix analysis further confirmed the model's effectiveness in correctly classifying both canceled and non-canceled bookings, indicating its suitability for supporting data-driven decision-making in hotel reservation and revenue management.

Despite the promising results, several directions for future work can be identified. First, the predictive performance may be further improved by incorporating advanced feature engineering techniques and additional contextual variables, such as external events, customer behavior history, and macroeconomic indicators. Second, future studies could explore the integration of explainable artificial intelligence (XAI) methods to enhance model transparency and increase trust among hotel managers. Third, real-time data processing and deployment of the proposed model within operational hotel management systems could be investigated to assess its effectiveness in dynamic and real-world environments. These improvements are expected to further enhance the practical applicability and robustness of machine learning-based cancellation prediction systems in the hospitality industry.

5. SUGGESTION

Future research may extend this study by exploring more advanced feature engineering techniques to capture deeper behavioral patterns in hotel booking data, such as customer booking history, temporal trends, and seasonality effects. The inclusion of external and contextual factors, including public holidays, local events, economic indicators, and travel restrictions, may also enhance the predictive capability of cancellation models. In addition, future studies could investigate the application of alternative or hybrid machine learning approaches, such as deep learning models or ensemble combinations of multiple classifiers, to further improve prediction robustness and generalization performance.

Another promising research direction involves the integration of explainable artificial intelligence (XAI) techniques to improve model interpretability and transparency. By employing methods such as SHAP or LIME, future work can provide more intuitive explanations of prediction outcomes, thereby increasing trust and usability for hotel managers and decision-makers. Furthermore, deploying the proposed model in a real-time or near real-time operational environment represents an important avenue for future research, enabling continuous model updating and performance evaluation under dynamic booking conditions. These directions are expected to contribute to the development of more reliable, interpretable, and scalable hotel booking cancellation prediction systems.

REFERENCES

[1] A. Antonio, A. de Almeida, and L. Nunes, "Hotel booking demand datasets," *Data in Brief*, vol. 22, pp. 41–49, 2020, doi: 10.1016/j.dib.2018.11.126.

[2] S. Gupta and M. Saberi, "Revenue management challenges in hotel booking cancellations," *International Journal of Hospitality Management*, vol. 87, pp. 102–112, 2020, doi: 10.1016/j.ijhm.2020.102498.

[3] Y. Yang, H. Pan, and J. Song, "Predicting hotel booking cancellation with machine learning models," *Expert Systems with Applications*, vol. 167, pp. 114129, 2021, doi: 10.1016/j.eswa.2020.114129.

[4] M. Al-Balushi, S. Al-Khusaibi, and R. Al-Hosni, "Machine learning approaches for customer behavior prediction in hospitality," *IEEE Access*, vol. 9, pp. 145678–145690, 2021, doi: 10.1109/ACCESS.2021.3119634.

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 3, pp. 1–22, 2022, doi: 10.1145/3534678.

[6] J. Li, X. Zhang, and Y. Wang, "Ensemble learning for demand forecasting: A comparative study," *Applied Soft Computing*, vol. 114, p. 108121, 2022, doi: 10.1016/j.asoc.2021.108121.

[7] R. Kaur and P. K. Singh, "Customer churn and cancellation prediction using gradient boosting models," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 1, pp. 45–56, 2023, doi: 10.1109/TCSS.2022.3188467.

[8] H. Wang, Z. Liu, and Y. Chen, "A comparative study of machine learning models for reservation cancellation prediction," *Applied Artificial Intelligence*, vol. 35, no. 12, pp. 987–1004, 2021, doi: 10.1080/08839514.2021.1925413.

[9] M. Ferreira, J. Silva, and R. Martins, "Predictive analytics for hotel revenue management using ensemble learning," *Journal of Hospitality and Tourism Technology*, vol. 13, no. 3, pp. 455–471, 2022, doi: 10.1108/JHTT-01-2021-0018.

[10] A. Hermawan, I. Amalia, M. Rafif, N. A. Azzahra, and R. Ragasa, "Optimizing Machine Learning Models for Predicting and Mitigating Hotel Booking Cancellations," *JUPTI: Jurnal Pengembangan Teknologi Informasi*, vol. 4, no. 2, pp. 1–12, 2025, doi:10.55606/jupti.v4i2.4055.