Router: Jurnal Teknik Informatika dan Terapan
Volume 3 Number 1, March 2025
e-ISSN : 3032-3312 p-ISSN : 3026-3611, Hal 37-48
DOI: https://doi.org/10.62951/router.v3i1.400
Available Online at: https://journal.aptii.or.id/index.php/Router

Predicting Hotel Booking Cancellations Using Machine Learning for Revenue Optimization

Andy Hermawan [1]*, Aji Saputra [2], Nabila Lailinajma [3], Reska Julianti [4], Timothy Hartanto [5], Troy Kornelius Daniel [6]
[1] Universitas Indraprasta PGRI, Indonesia
[2] Universitas Khairun, Indonesia
[3,4,5,6] Purwadhika Digital School, Indonesia

TB. Simatupang, Jl. Nangka Raya No.58C Tanjung Barat,
Kec. Jagakarsa - Jakarta Selatan
Email : andy.hermawan@unindra.ac.id, aji.saputra@unkhair.ac.id

Abstract*. Hotel booking cancellations pose significant challenges to the hospitality industry, affecting revenue management, demand forecasting, and operational efficiency. This study explores the application of machine learning techniques to predict hotel booking cancellations, leveraging structured data derived from hotel management systems. Various classification algorithms, including Random Forest, XGBoost, and LightGBM were evaluated to identify the most effective predictive model. The findings reveal that XGBoost model outperforms other models, achieving F2-score of 0.7897. Key influencing factors include deposit type, total number of special requests, and marketing segment. The results underscore the potential of predictive modeling in optimizing hotel revenue strategies by enabling proactive measures such as dynamic pricing, targeted customer engagement, and improved overbooking policies. This study contributes to the ongoing advancements in data-driven decision-making within the hospitality industry, offering insights into how machine learning can mitigate financial risks associated with booking cancellations.*

Keywords*: Hotel Booking Cancellations, Machine Learning, XGBoost, Revenue Management, Classification, LightGBM.*

1. BACKGROUND

Along with the rapid development of the hospitality industry and the increasingly intense competition, the challenges faced by industry players are increasingly complex. One of the main challenges that often arise is the difficulty in accurately predicting customer demand. This is particularly evident in the case of hotel room booking cancellations which can disrupt operational planning and resource management. The hospitality industry faces significant challenges in predicting customer demand, especially in terms of canceled room bookings. Various studies have shown that booking cancellations can be influenced by many factors, both external such as weather and unexpected events, as well as internal factors relating to hotel policies and customer behavior (Zafitri & Jambak, 2023).

Unexpected cancellations can lead to revenue losses, inefficient resource allocation, and decreased room occupancy rates. Inaccuracies in predicting cancellation trends prevent hotels from adjusting pricing strategies, implementing overbooking

Received: February 05, 2025; Revised: February 20, 2025; Accepted: March 10, 2025; Published : March 12, 2025

policies, as well as ensuring optimal room utilization. With increasing competition in the hospitality sector, it is imperative for hotel managers to adopt technologies that can assist them in managing these demand changes more effectively.

In dealing with this problem, the use of Machine Learning (ML)-based predictive models offers great potential. A meta-analysis of articles related to hotel revenue management revealed that hotel revenue management teams are increasingly moving towards a more strategic approach, integrating big data analytics and ML models to predict booking cancellations (Binesh et al., 2020). By using these analytical techniques, hotels can better understand the cancellation patterns that customers have, allowing them to implement proactive policies to maximize occupancy and revenue.

Through this research, it is expected to develop a prediction model that is able to accurately identify hotel bookings that are likely to be canceled. This approach aims to reduce the risk of loss due to negligence in detecting potential cancellations, by analyzing historical datasets that include customer and booking attributes such as market segment, customer type, room type booked, deposit status, previous cancellations, as well as the number of special requests submitted. Thus, this model is expected to provide valuable insights for hotel managers in making more informed decisions regarding revenue management.

2. THEORITICAL STUDY

Hotel Revenue Management

Revenue management is an approach used to optimize revenue from hotel room bookings by considering factors such as demand, supply, price, and customer behavior (Kimes, 2011). In the hospitality industry, revenue management is a very important aspect to ensure optimal room occupancy rates and maximize revenue (Binesh et al., 2021). One of the critical aspects of revenue management is the ability to predict and manage booking cancellations. Unexpected cancellations can cause significant losses in hotel revenue due to their unpredictability (Anderson & Xie, 2016).

Booking Cancellation Behavior

Booking cancellation behavior is a frequent phenomenon in the hospitality industry and can be caused by a variety of factors, including changing customer needs, errors in room selection, or unexpected events that affect their trip (Kang, 2012). Booking cancellation behavior is a frequent phenomenon in the hospitality industry and can be

caused by a variety of factors, including changes in customer needs, errors in room selection, or unexpected events that affect their trip (Kang et al., 2012). In some cases, customers tend to cancel their bookings without sufficient advance notice, which makes it difficult for hotels to replace lost bookings. Therefore, it is important to identify patterns in cancellations to better anticipate and manage cancellations.

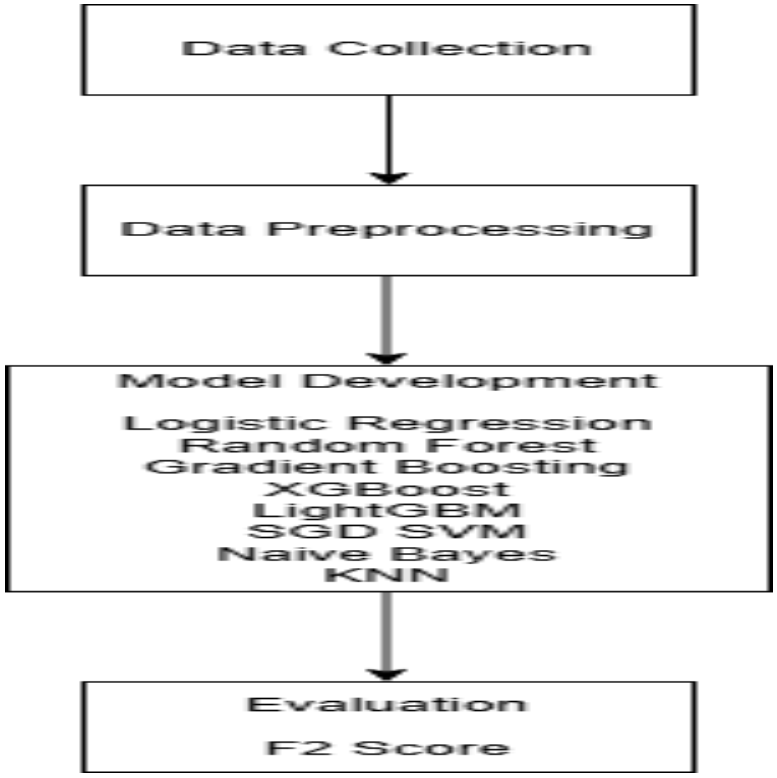
Big Data and Machine Learning in Revenue Management

Along with the development of technology, the use of big data and machine learning (ML) in revenue management has increasingly become the focus of research. Big data enables the collection and analysis of data from various sources, including transaction data, customer behavior, and market conditions. These technologies provide deeper insights into demand and cancellation trends, allowing hotels to implement more dynamic and targeted pricing strategies (Lee, 2020). Machine learning, as a branch of artificial intelligence, can identify hidden patterns in historical data that are difficult to find by conventional analysis, so it can be used to forecast booking cancellations with higher accuracy (Binesh et al., 2021).

Booking Cancellation Prediction Using Machine Learning

The use of machine learning algorithms in booking cancellation prediction has proven to be effective in improving prediction accuracy and minimizing losses caused by unexpected cancellations. Methods such as logistic regression, decision trees, and ensemble-based algorithms such as Random Forest and XGBoost are used to identify factors associated with cancellations and to classify bookings that are at high risk of being canceled (Antonio et al., 2019). With this prediction model, hotels can more effectively manage overbooking, optimize room rates, and plan more targeted marketing strategies.

3. RESEARCH METHODOLOGY



This study utilizes a truncated dataset of hotel reservations published by Antonio et al. (2019) to predict booking cancellations using machine learning techniques. The dataset includes various features related to customer behavior, booking history, and reservation details. The methodology follows the same structured approach implemented in previous studies (Antonio et al., 2017, Yaqi Lin, 2023; Rahmawati et al., 2024) encompassing data preprocessing, model selection, training, evaluation, and interpretation of classification results.

The dataset is first subjected to Exploratory Data Analysis (EDA) to identify patterns, anomalies, and missing values. Feature engineering is a process involves transforming raw data into variables that are more informative for machine learning algorithms (Hermawan et al, 2024). Feature engineering techniques, including categorical encoding and numerical transformations, are applied to enhance model performance. The dataset is then split into training and testing subsets using a stratified approach to ensure class balance. Multiple machine learning classification models, including Random Forest, XGBoost, and Bayesian Networks, are implemented due to their effectiveness in classification tasks related to booking cancellations (Antonio et al., 2017, Yaqi Lin, 2023; Rahmawati et al., 2024).

After model training, evaluation is conducted using F2-score, due to an emphasis placed on recall, as it is crucial to accurately identifying potential cancellations to optimize hotel revenue management strategies (Antonio et al., 2017). The top two models with the highest average F2-score, the lowest standard deviation of F2-score, and the lowest time to fit the model are selected for hyperparameter tuning. Then, the fine tuned model with the best scores across the three metrics discussed above (average F2-Score, lowest standard deviation of F2-score, and lowest time to fit the model) are selected for probability threshold adjustment so that the F2-score can be maximized.

4. RESULTS AND DISCUSSION

This study focuses on developing a hotel booking cancellation prediction model, which aims to identify bookings that are at risk of being canceled. The data used in this study includes a historical hotel booking dataset with 83573 booking entries and 9 key relevant features. These features include market segment, customer type, room type booked, deposit status, previous cancellation history, and number of special requests submitted. The data processing process began with data cleaning and exploration to ensure the integrity of the dataset.

In this section, the results of the data exploration, model testing, as well as the strategic implications resulting from the booking cancellation prediction analysis are described in detail. These results not only serve to improve efficiency in customer retention but also provide strategic recommendations for stakeholders in making more informed decisions.

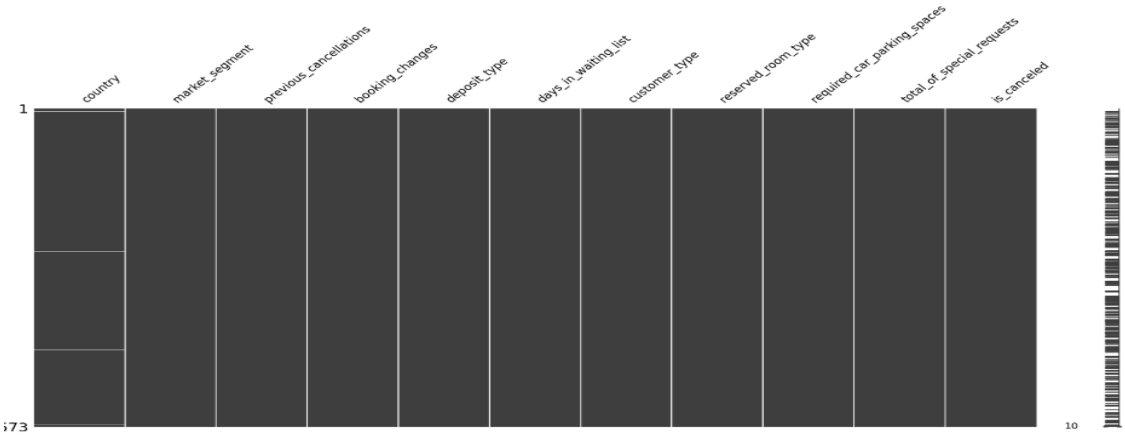
Exploratory Data Analysis

In the initial phase of the study, the hotel booking dataset is preprocessed to ensure its quality and reliability for subsequent analysis and model development. This includes handling missing values, identifying duplicate records, and detecting outliers, as recommended by prior research (Alexandropoulos, et al., 2019). The specific preprocessing steps are outlined below:

Handling Missing Values

The dataset is examined for missing values in key features such as *lead time*, *market segment,* and *deposit type*. To address these gaps, appropriate imputation strategies are applied based on the nature of the variables. Numerical features with missing values are imputed using the median, as it is less sensitive to outliers, while

categorical variables are imputed using the mode to preserve the most frequent category. These methods align with best practices in data preprocessing to ensure data integrity and minimize bias in subsequent analysis (Han, Kamber, & Pei, 2012).



Identifying and Removing Duplicate Records

Duplicate bookings, which may occur due to system errors or repeated customer reservations, are identified using a combination of customer ID, booking date, and room type features. To maintain data integrity and prevent bias in model training, these duplicate records are removed, ensuring that each booking is uniquely represented in the dataset. This step is crucial in avoiding data redundancy and improving the reliability of predictive modeling (Pratim, et al., 2008).

Outlier Detection and Treatment

Extreme values in variables such as lead time and previous cancellations are identified using the interquartile range (IQR) method and boxplot visualizations. To maintain data quality and prevent model distortion, bookings with unrealistic values—such as lead times exceeding 700 days or previous cancellations greater than 10—are either removed or capped. This approach ensures a more robust dataset and enhances the reliability of predictive modeling (Aggarwal, 2017).

Feature Analysis and Distributions

Univariate and bivariate analyses are carried out to explore the distribution and relationships of key variables in the dataset. Various visualization techniques, including histograms, boxplots, and density plots, are employed to examine features such as average daily rate (ADR), number of special requests, and booking lead time. Additionally, Pearson's correlation coefficient is used to assess the relationships between numerical variables and their influence on booking cancellations, providing insights into potential predictive factors (Zheng & Casari, 2018).

Class Imbalance Analysis

The target variable (*is_canceled*) is examined to assess class distribution and detect any significant imbalance. If an imbalance is present, strategies such as Synthetic Minority Over-sampling Technique (SMOTE) or class weighting are considered to enhance model performance and prevent bias toward the majority class. These techniques help ensure that the predictive model effectively identifies both canceled and non-canceled bookings (Chawla et al., 2002).

The results of this exploratory data analysis provide key insights into the characteristics of booking behavior, highlighting patterns that may influence cancellation likelihood. These findings serve as the foundation for feature selection and model development, ensuring that the dataset is well-structured for predictive analytics.

Modelling Results and Evaluation

As previously discussed, the models will be compared based on used

Tabel 1. Modelling with Default Parameters Results

Rank	Resampling	Classifier	F2 Score Mean	F2 Score Std	Execution Time
1	SMOTE	XGBoost	0.728732	0.004119	32.396321
2	SMOTE	LightGBM	0.728672	0.004309	32.907323
3	SMOTE	Random Forest	0.728654	0.003884	56.881753
4	RandomUnderSampler	LightGBM	0.728600	0.004471	3.424906
5	RandomUnderSampler	Random Forest	0.728535	0.003744	15.240193

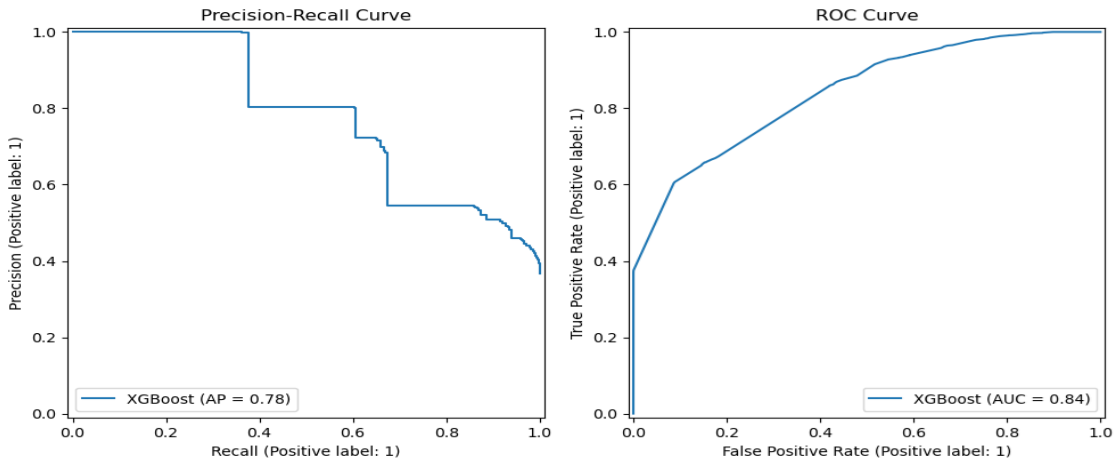
XGBoost and LightGBM models with SMOTE resampling technique appear to dominate the ranking with the highest mean of F2-score, the lowest of F2-standard deviation, and the lowest execution time.

Next, the hyperparameter tunings will be conducted on both XGBoost and LightGBM with the results posted below:

Tabel 2 Modelling Results with Hyperparameter Tuning

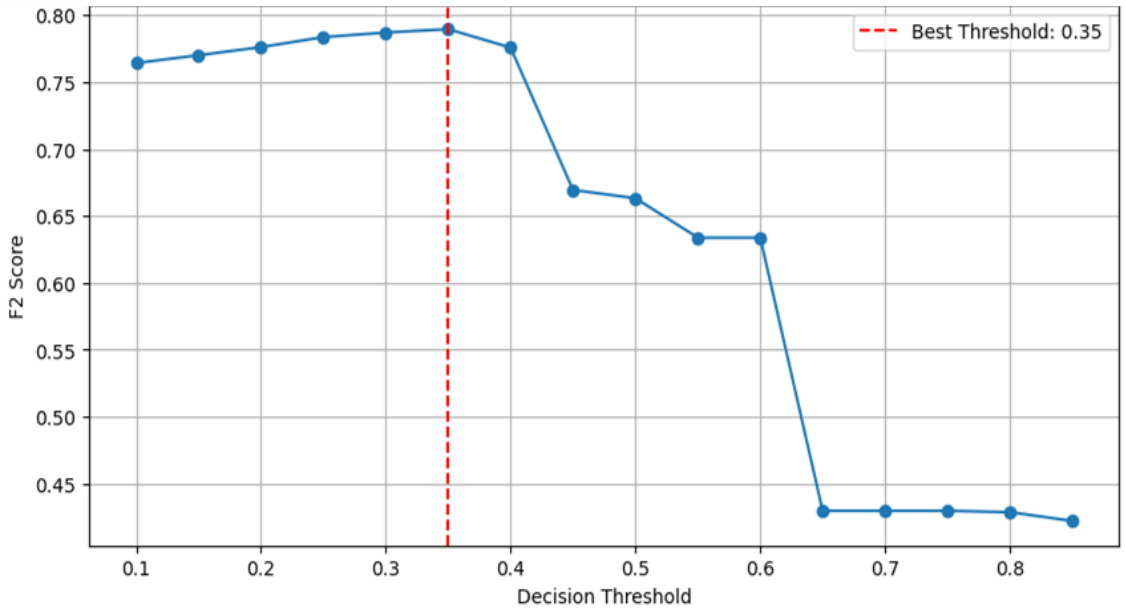
Rank	Model	Mean F2 Score	Std F2 Score	Execution Time (seconds)
1	XGBoost	0.670714	0.007394	102.668101
2	LightGBM	0.663994	0.009654	178.274542

As evidenced by the table above, XGBoost model comes out on top. The following pictures will detail visualize the Precision-Recall curve and ROC curve:



Bagan 1. Precision Recall Curve and ROC Curve of Fine Tuned XGBoost

To improve the F2-score of the fine tuned model, the model is adjusted for decision threshold with the results visualized below:



Bagan 2. Line Plot of F2 Score vs Decision Threshold

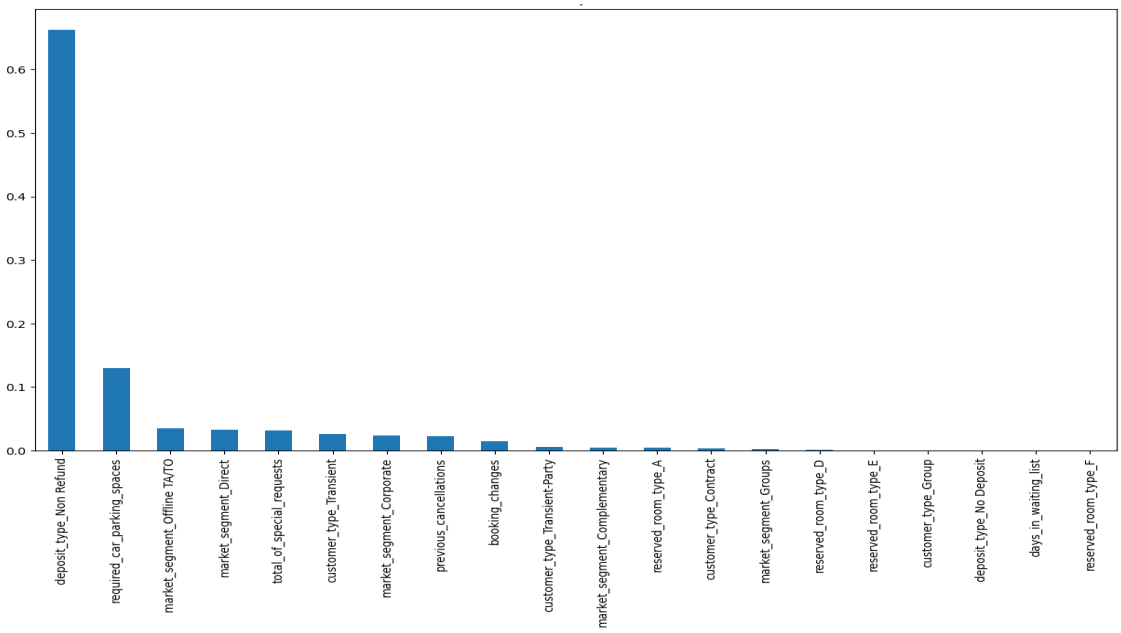
It can be seen that the optimal threshold is **0.35**, ensuring a balance between sensitivity and conservatism. The adjusted probability also increased the model's F2-Score to **0.7897**, accurately predicting **79%** of actual cancellations while minimizing false positives.

Model Evaluation

In this section, the resulting model (saved to best_model_adjusted variable) will be evaluated from its features, business outcomes, and its limitation.

Feature Evaluation

The important features from the adjusted model are visualized below:

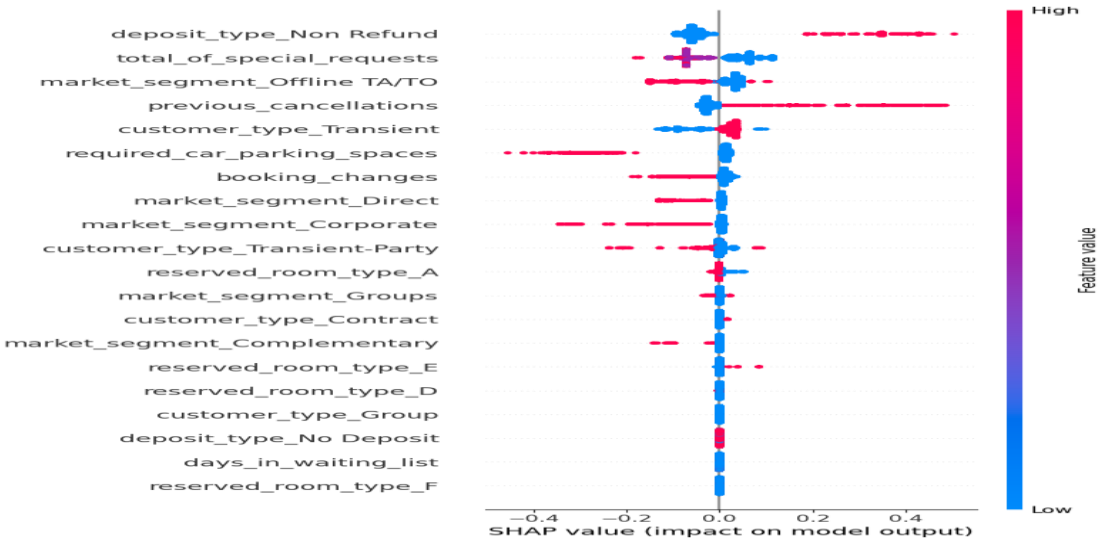


Bagan 3. Feature Importances on the Adjusted Model

There are several observations that can be made:

- Encourage Non-Refundable Deposits to reduce cancellation rates.
- Optimize Marketing & Pricing based on market segments.
- Enforce Stricter Policies for frequent cancellers while rewarding loyal customers.

Next the Shapley values will be visualized through SHAP python library:



Bagan 4. SHAP value of Adjusted Model

The model shows the following observations:

- Tailored Marketing: Flexible for transients, stricter for frequent cancellers.
- Reward Loyalty: Discounts for low-cancellation customers.

- Refine Policies: Adjust based on special request trends.

5. CONCLUSION AND RECOMMENDATION

The SHAP values presented in the plot illustrate how each feature contributes to individual predictions rather than just overall importance. The analysis highlights that "Non-Refund" deposit types have the highest impact, reinforcing that non-refundable bookings significantly reduce cancellations. Total special requests exhibit mixed effects, as they can indicate serious intent but may also correlate with higher cancellations. Market segment plays a crucial role, with offline TA/TO bookings showing a significant impact, whereas direct and corporate bookings behave differently. Additionally, previous cancellations are a strong predictor, as customers with a history of cancellations are more likely to cancel again. From a business perspective, personalized marketing strategies should be implemented based on market segments and customer types, offering flexible policies to transient and party travelers while enforcing stricter rules for frequent cancelers. Incentivizing customers with low cancellation history through discounts or loyalty perks can encourage repeat business. Moreover, analyzing and adjusting policies for special requests is essential to identify and manage those that correlate with higher cancellations. The XGBoost model, after hyperparameter tuning and probability threshold adjustment, demonstrates the best performance among nine classification methods, achieving an F2-score of approximately 0.79, effectively balancing recall and precision. Financially, by leveraging this optimized model, hotel management can save an estimated $286,800 by accurately predicting booking cancellations.

Recommendation

Business Recommendation

By analyzing both the feature importance and SHAP value plots, as well as understanding that the best model has its probability threshold adjusted to 0.35, the following business recommendations can be made:

- Leverage Non-Refundable Deposits to Minimize Cancellations: Offer discounts, early-bird deals, or room upgrades for customers choosing non-refundable bookings.
- Segment Customers and Market Accordingly: Optimize pricing, cancellation policies, and targeted promotions based on market segments and booking behaviors.
- Implement a Risk-Based Booking System: Introduce dynamic pricing or cancellation penalties for customers with a history of cancellations.

Model Improvement Recommendation

To improve the Machine Learning performance, the following recommendations are put forward:

- Incorporate Time-Based Patterns: Enrich the data by incorporating booking seasonality trends (such as yearly/monthly data) to detect seasonal spikes in cancellation, as well as incorporating rolling cancellation trends per customer to capture changing behaviors.
- Try Alternative Models for Comparison: Test other classification algorithms (e.g., Deep-Learning) to observe if other algorithms produce better classification results.

REFERENCES

Anderson, C. K., & Xie, X. (2016). Dynamic pricing in hospitality: Overview and opportunities. *International Journal of Revenue Management, 9*(2–3), 165–174.

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.

Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *The Knowledge Engineering Review, 34*, e1.

Binesh, F., Belarmino, A., & Raab, C. (2021). A meta-analysis of hotel revenue management. *Journal of Revenue and Pricing Management*, 1–13.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Ghosh, P., Gelasca, E. D., Ramakrishnan, K. R., & Manjunath, B. S. (2008). Duplicate image detection in large-scale databases. In *Advances in Intelligent Information Processing: Tools and Applications* (pp. 149–166).

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques.* Morgan Kaufmann Publishers.

Hermawan, A., Jayanti, N. R., Tabaruk, Z., Triadi, F. L. Y., Saputra, A., & Syachrudin, M. R. H. (2024). Membangun model prediksi churn pelanggan yang akurat: Studi kasus tentang TELCO company. *Merkurius: Jurnal Riset Sistem Informasi Dan Teknik Informatika, 2*(6), 67–81. https://doi.org/10.61132/merkurius.v2i6.398

Kimes, S. E. (2011). The future of hotel revenue management. *Journal of Revenue and Pricing Management, 10*, 62–72.

Kang, K. H., Stein, L., Heo, C. Y., & Lee, S. (2012). Consumers' willingness to pay for green initiatives of the hotel industry. *International Journal of Hospitality Management, 31*(2), 564–572.

Lee, M. (2022). Evolution of hospitality and tourism technology research from *Journal of Hospitality and Tourism Technology*: A computer-assisted qualitative data analysis. *Journal of Hospitality and Tourism Technology, 13*(1), 62–84.

Lin, Y. (2023). Research on the influencing factors of cancellation of hotel reservations. *Highlights in Science, Engineering and Technology, 61*, 107–117.

Rahmawati, E., Nurohim, G. S., Agustina, C., Irawan, D., & Muttaqin, Z. (2024). Development of machine learning model to predict hotel room reservation cancellations. *Jurnal Teknologi Informasi dan Terapan (J-TIT, 11*(2), 58–64. https://doi.org/10.25047/jtit.v11i2.5440

Zafitri, Z., & Jambak, M. I. (2023). Karakteristik pembatalan reservasi kamar hotel pada online travel agent menggunakan algoritma C4.5. *The Indonesian Journal of Computer Science, 12*(4).

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists.* O'Reilly Media.