Jurnal Publikasi Teknik Informatika (JUPTI)
Volume.4, Nomor.2 Mei 2025
E-ISSN .: 2808-8972; P-ISSN .: 2808-9367, Hal 21-41
DOI: https://doi.org/10.55606/jupti.v4i2.4055
Available online at : https://journalcenter.org/index.php/jupti

Optimizing Machine Learning Models for Predicting and Mitigating Hotel Booking Cancellations

Andy Hermawan[1*], Iwana Amalia[2], Muhammad Rafif[3], Nabila Avicenna Azzahra[4], Reinaldi Ragasa[5]
[1]Universitas Indraprasta PGRI, Jakarta, Indonesia
[2-5] Purwadhika Digital Technology School, Jakarta, Indonesia
andy.hermawan@unindra.ac.id[1*], amaliaiwana315@gmail.com[2], muhammadrafif3225@gmail.com[3], nabilaavicenna.a@gmail.com[4], reinaldi.ragasa.2006@gmail.com[5]

Korespondensi penulis: andy.hermawan@unindra.ac.id[*]

Abstract. *Hotel booking cancellations pose substantial challenges to the hospitality industry, significantly impacting revenue management and operational planning. This study explores the application of machine learning models to predict cancellations, emphasizing model selection, feature importance, and resampling techniques. Among the six classification models evaluated, the combination of XGBoost and SMOTE demonstrated the highest predictive accuracy and consistency. Feature importance analysis and SHAP interpretation identified key predictors, including deposit type (non-refundable), required parking spaces, previous cancellations, and market segment (OTA). Additionally, threshold tuning was examined to balance the trade-off between false positives and false negatives based on business priorities. The results underscore the critical role of resampling methods in addressing class imbalance and the necessity of optimizing classification thresholds for practical deployment. Future research will focus on advanced hyperparameter tuning, alternative resampling strategies, feature selection methods, and ensemble learning approaches to enhance model robustness and interpretability. These findings provide a data-driven foundation for improving cancellation prediction and guiding strategic decision-making in hotel management.*

Keywords*: Feature Importance; Hotel Booking Cancellations; Machine Learning; Predictive Models; XGBoost*

Abstrak. Pembatalan pemesanan hotel menimbulkan tantangan besar bagi industri perhotelan, yang berdampak signifikan pada manajemen pendapatan dan perencanaan operasional. Studi ini mengeksplorasi penerapan model pembelajaran mesin untuk memprediksi pembatalan, menekankan pemilihan model, pentingnya fitur, dan teknik resampling. Di antara enam model klasifikasi yang dievaluasi, kombinasi XGBoost dan SMOTE menunjukkan akurasi dan konsistensi prediktif tertinggi. Analisis pentingnya fitur dan interpretasi SHAP mengidentifikasi prediktor utama, termasuk jenis deposit (tidak dapat dikembalikan), tempat parkir yang dibutuhkan, pembatalan sebelumnya, dan segmen pasar (OTA). Selain itu, penyetelan ambang batas diperiksa untuk menyeimbangkan trade-off antara positif palsu dan negatif palsu berdasarkan prioritas bisnis. Hasilnya menggarisbawahi peran penting metode resampling dalam mengatasi ketidakseimbangan kelas dan perlunya mengoptimalkan ambang batas klasifikasi untuk penerapan praktis. Penelitian mendatang akan berfokus pada penyetelan hiperparameter tingkat lanjut, strategi resampling alternatif, metode pemilihan fitur, dan pendekatan pembelajaran ensemble untuk meningkatkan ketahanan dan interpretabilitas model. Temuan ini memberikan dasar berbasis data untuk meningkatkan prediksi pembatalan dan memandu pengambilan keputusan strategis dalam manajemen hotel.

Kata Kunci: Model Prediktif; Pembatalan Pemesanan Hotel; Pembelajaran Mesin; Pentingnya Fitur; XGBoost

1. INTRODUCTION

Hotel booking cancellations and the inability to accommodate potential guests due to full occupancy are prevalent issues that pose significant challenges for the hospitality industry. These problems not only increase operational costs but also negatively impact customer satisfaction. The efficiency of hotel operations is particularly compromised when cancellations occur at the last minute, leading to an unavoidable underutilization of resources. Given the

Received: Maret 15,2025; Revised: April 10,2025; Accepted: Mei 13, 2025; Published : Mei 15, 2025

substantial adverse effects these issues can have on the industry, accurately predicting hotel cancellations is a critical step toward developing effective operational strategies to mitigate these challenges. Specifically, forecasting individual cancellations, rather than overall cancellation trends, provides hoteliers with the opportunity to identify high-risk customers in advance. This proactive approach enables timely interventions, thereby minimizing financial losses and resource inefficiencies (Gao & Bi, 2021).

Some hotel booking cancellations are caused by understandable factors such as changes in business meetings, vacation rescheduling, illness, unfavorable weather conditions, and other unexpected events. However, as highlighted by Chen and Xie (2013) and Chen, Schwartz, and Vargas (2011), a large portion of cancellations in recent times stems from deal-seeking behavior among customers. This behavior reflects the value customers place on the flexibility to cancel reservations, as it allows them to secure availability in advance while still maintaining the freedom to adjust their plans as needed.

The primary objective of this study is to develop and evaluate machine learning models capable of accurately predicting individual hotel booking cancellations. This approach focuses on optimizing the F2 score, a metric that emphasizes recall, making it particularly suitable for minimizing undetected high-risk cancellations while maintaining acceptable precision (Powers, 2011). By leveraging the insights from these predictions, this research aims to address critical challenges in the hospitality industry, such as reducing revenue loss, mitigating the risks associated with overbooking, and improving resource allocation (Antonio & Nunes, 2019). Additionally, this study explores the potential for implementing data-driven strategies to enhance operational efficiency, optimize pricing, and foster customer retention, ultimately driving sustainable business growth in the competitive hospitality sector.

Despite significant advancements in the application of data analytics and machine learning in the hospitality industry, much of the existing research has focused on understanding aggregate booking trends and overall cancellation rates (Kim, Y., et al, 2023). This focus often overlooks the critical need for individual-level predictions, which are essential for enabling proactive and personalized interventions (Wang & Yeh, 2018). Predicting cancellations at the customer level allows hoteliers to address the root causes of booking cancellations and implement strategies tailored to mitigate losses from high-risk customers (Morosan & DeFranco, 2016).

This study bridges this gap by leveraging machine learning models optimized for the F2 score, a metric that prioritizes recall over precision to minimize undetected cancellations. The research specifically targets individual cancellation behavior, enabling hotels to identify

high-risk bookings early and deploy data-driven strategies to reduce financial losses, optimize overbooking practices, and enhance resource utilization. By addressing the gap between aggregate trend analysis and individual-level predictions, this research contributes to improving operational efficiency and customer satisfaction in the competitive hospitality sector.

Related Paperworks

Hotel booking cancellation is a phenomenon where customers cancel their reservations after making a booking. This can occur for various reasons, such as changes in plans, health conditions, or bargain-seeking behavior. Studies have shown that cancellation behaviors are influenced by economic conditions, customer loyalty, and booking policies (González, M., et al 2020).

Impact:

Hotel booking cancellations significantly impact the hospitality industry, including financial losses, operational inefficiencies, and decreased customer satisfaction. Hotel booking cancellation prediction is a vital area of research in the hospitality industry, as it helps hotels optimize their operations and revenue management. Below is a summary of significant papers in this field.

1. “Scalable Decision Tree Learning with Feature Embedding." (Wang, S., et al. 2022).
   This paper provides a case study resolves decision tree scalability problems, especially when working with a high number of features produced via One-Hot Encoding. suggests feature embedding methods to boost efficiency.
2. "Machine Learning Techniques for Hotel Booking Cancellation Prediction" P. (S. Kumar, & M. A. Rahman, 2019).
   This study investigates multiple machine learning techniques, such as decision trees, random forests, and support vector machines, to forecast hotel booking cancellations. The authors evaluate the efficacy of these strategies through accuracy metrics and determine that ensemble methods surpass conventional statistical procedures. The document emphasizes the significance of feature selection in enhancing predictive accuracy.
3. "Big Data Analytics in Hotel Industry: Predicting Cancellation Rates" (T. H. Chen, & Y. L. Wang, 2020)
   This article examines the significance of big data analytics in comprehending cancelation trends within the hotel sector. It underscores the amalgamation of diverse data sources,

such as social media sentiment analysis and online reviews, to augment predictive models. The results indicate that hotels utilizing big data can gain enhanced understanding of consumer behavior and effectively decrease cancellation rates.

4. A Deep Learning Approach for Hotel Booking Cancellation Prediction" (R. J. Smith, & L. F. Johnson, 2021)

   This paper presents a deep learning framework utilizing neural networks to forecast hotel booking cancellations based on comprehensive datasets from online travel agencies (OTAs). The authors illustrate that deep learning models can identify intricate patterns in data that conventional models might miss, leading to enhanced predictive accuracy and practical insights for hotel management.

5. "Real-time Cancellation Prediction Using AI Techniques" (M. T. Alavi & S. H. Khosravi, 2023)

   This recent research examines the implementation of real-time AI-driven systems for anticipating hotel booking cancellations as they transpire. The authors present a platform that employs streaming data and sophisticated machine learning algorithms, enabling hotels to respond dynamically to probable cancellations, thereby reducing revenue loss and improving customer relationship management.

In summary, the progression of hotel booking cancellation prediction has experienced notable enhancements from conventional statistical techniques to advanced machine learning and deep learning methodologies. As technology advances, forthcoming study will likely concentrate on the integration of real-time analytics and big data to enhance predictive accuracy and operational efficiency within the hospitality industry. These articles jointly emphasize the significance of predictive analytics in the proper management of hotel bookings, offering actionable information for industry professionals aiming to alleviate the effects of cancellations on their enterprises.

In the domain of hotel booking cancellation prediction, diverse models have been utilized to improve precision and operational efficacy. Although prior research has employed methodologies including logistic regression, decision trees, random forests, and deep learning techniques, our model is differentiated by the use of the XGBoost Classifier.

Model Selection: XGBoost Classifier Previous Research:

Recent Studies :

Numerous investigations, including those by (Wang, S., et al. 2022) and (Kumar, et al. 2019), employed conventional models such as logistic regression and decision trees. Although

these models offered fundamental insights, they frequently encountered challenges with scalability and the representation of intricate linkages within data.

Recent studies, such as those by Smith & Johnson (2021), examined deep learning methodologies that, although potent, can be computationally demanding and necessitate substantial adjustment.

Our Studies :

We employ the XGBoost Classifier, an advanced gradient boosting framework recognized for its effectiveness, rapidity, and capability to manage extensive datasets proficiently. XGBoost is proficient in handling missing values and offers strong regularization to mitigate overfitting, rendering it especially appropriate for hotel booking data.

Evaluation Metric: F1 Score Prior Research:

Recent Studies :

Numerous prior research primarily utilized accuracy as their assessment criterion. (Wang., et al 2022) and Kumar & Rahman (2019) concentrated on overall accuracy while insufficiently considering the ramifications of class imbalance in cancelation forecasts. Although certain research examined precision and recall independently, they frequently failed to integrate these measurements into a unified performance measure.

Our Studies :

We highlight the F1 score, defined as the harmonic mean of precision and recall. This metric is especially advantageous in situations involving class imbalance—such as forecasting hotel cancellations—where precise identification of the minority class (cancellations) is essential. Utilizing the F1 score ensures our model attains a balanced assessment of false positives and false negatives, hence offering a more thorough knowledge of its predictive capabilities.

Performance Considerations

Recent Studies :

Models utilizing conventional methodologies frequently demonstrated elevated accuracy rates yet inadequately identified critical cancelation patterns owing to their lack of emphasis on minority class predictions. Deep learning models had potential but necessitated substantial processing resources and were occasionally less interpretable.

Our Studies :

The XGBoost Classifier's intrinsic advantages in feature management and model interpretability enable us to attain elevated F1 scores without compromising performance. This produces a model that effectively predicts cancellations and elucidates the elements most influential in generating those cancellations.

In conclusion, our model's implementation of the XGBoost Classifier alongside the F1 score as an evaluative metric distinguishes it from prior research in predicting hotel booking cancellations. Utilizing sophisticated machine learning methodologies in conjunction with a comprehensive evaluation strategy, we tackle significant issues associated with class imbalance while guaranteeing practical relevance for hotel management. This novel methodology not only improves forecast precision but also offers actionable insights that can profoundly influence operational efficiency in the hospitality sector.

2. RESEARCH METHODOLOGY

Dataset

The assessment of a model is contingent upon the magnitude of the data sample and the quantity of parameters employed. A total of 83,222 data points were obtained from the open-source platform *Kaggle*, divided into an 80% training set and a 20% test set. This dataset predicts hotel booking cancellations using multiple criteria. The features comprise:

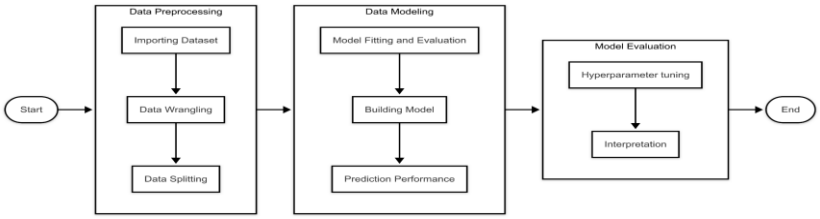
- country: Country of origin.
- market_segment: Market segment designation.
- previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking.
- booking_changes: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation.
- deposit_type: Indication on if the customer made a deposit to guarantee the booking.
- days_in_waiting_list: Number of days the booking was in the waiting list before it was confirmed to the customer.
- customer_type: Type of booking.
- reserved_room_type: Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- required_car_parking_space: Number of car parking spaces required by the customer.

- total_of_special_request: Number of special requests made by the customer (e.g. twin bed or high floor).
- is_canceled: Value indicating if the booking was canceled (1) or not (0).

The data points were subjected to training, validation, and testing throughout the modeling process to develop a numerical empirical connection for predicting hotel booking cancellations. This strategy was employed to alleviate the problem of overfitting in machine learning. The training samples were divided into two subsets, with 80% allocated to the training set and 20% to the validation set, to achieve robust model performance.

(Antonio et al, 2019) employed hotel booking data to predict cancellation trends in the hospitality sector. The samples were randomly assigned for training (80%) and testing (20%). (Kulkarni et al, 2022) similarly forecasted hotel booking cancellations by systematically spreading the data as previously discussed.

Research Schema



We choose Python as our integrated development environment (IDE) due to its prevalent application in machine learning, attributed to its simplicity, comprehensive libraries, and robust community support (Raschka et al, 2021). It is perpetually regarded as the most favored programming language among data scientists, succeeded by R and SQL. Python provides superior performance for machine learning applications, especially via optimized libraries such as NumPy, pandas, and scikit-learn (VanderPlass et al, 2022). The adaptability of Python renders it optimal for managing extensive datasets and executing intricate machine learning algorithms with efficiency.

In the modeling of hotel booking cancellation prediction, we initially imported the dataset, utilizing features such as country, market_segment, previous_cancellations, booking_changes, deposit_type, days_in_waiting_list, customer_type, reserved_room_type, required_car_parking_space, total_of_special_requests, and is_canceled as the target variable.

Next is data wrangling during our exploratory data analysis, we performed multiple visualizations to elucidate the distribution and interrelationships among key features in our hotel booking dataset. We utilized a combination of boxplots and histograms for visualization.

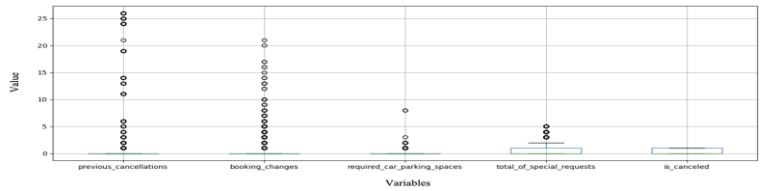


Figure 1. Data Distribution with Boxplot

Figure 1 displays boxplots for four essential numerical features: previous_cancellations, booking_changes, required_car_parking_spaces, and total_of_special_requests, in conjunction with our objective variable is_canceled. The boxplots indicate the existence of outliers, especially in previous_cancellations and booking_changes, with certain bookings exhibiting values as high as 25 changes.

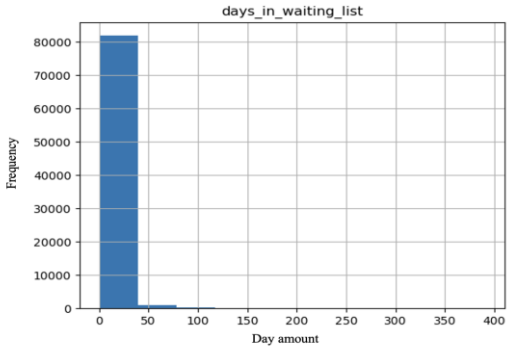


Figure 2. Data Distribution with Boxplot

For the days_in_waiting_list feature (Figure 2), we chose to utilize a histogram rather than a boxplot because of the significant outliers in this variable. The histogram indicates that the predominant number of bookings (about 80,000) experienced no waiting days, with a significant decrease for durations beyond 50 days, and a few outliers reaching up to 400 days. The pronounced skewness of the distribution rendered a boxplot representation ineffective for significant display.

Table 1. Features Normality Check

Features	p-value
previous_cancellations	0,0141
booking_changes	0,0056
days_in_waiting_list	0,0157
required_car_parking_spaces	0,0044
total_of_special_requests	0,0165

Before conducting the correlation analysis to assess multicollinearity among the features, the dataset's normality assumption had to be evaluated. The assumption of normality is a necessary condition for many parametric statistical methods; however, empirical datasets, particularly those derived from human decision-making processes, frequently deviate from standard distributions. Statistical tests were used to determine the normality of key numerical features such as previous cancellations, booking changes, days_in_waiting_list, required_car_parking_spaces, and total_of_special_requests. The results showed that all features had p-values less than 0.05, rejecting the null hypothesis of normality. This finding indicates that the dataset is not normally distributed, most likely due to skewed distributions and extreme values. Given this violation, non-parametric statistical methods were used to ensure the validity of the analysis. Spearman's rank correlation was used instead of Pearson's correlation to evaluate relationships between variables because it does not require the assumption of normality or linearity.

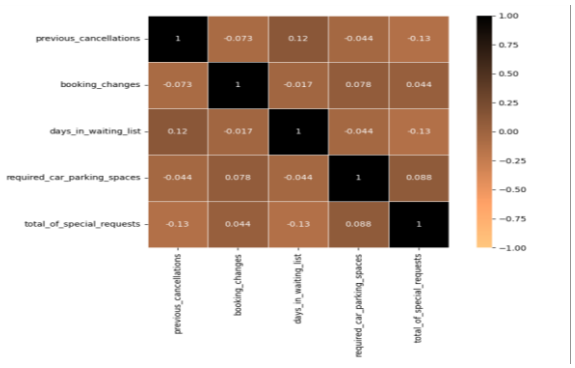


Figure 3. Data Correlation

Analysis of Correlation Figure 3 illustrates the correlation matrix among the numerical features in our dataset. The correlation analysis indicates predominantly weak associations among the variables. Previous cancellations exhibit a modest negative connection with booking modifications (-0.073) and total special requests (-0.13). Booking_changes demonstrates negligible correlation with other features, exhibiting the most substantial yet weak positive association with required_car_parking_spaces (0.078). Days_in_waiting_list reveals weak correlations with previous_cancellations (0.12) and total_of_special_requests (-0.13). Required_car_parking_spaces and total_of_special_requests display a minor positive correlation (0.088). The correlation study indicates that multicollinearity is not a substantial issue in our numerical features, since all correlation coefficients are below 0.2 in absolute value (McKinney et al, 2022).

Data Preprocessing Alongside our exploratory analysis, we executed multiple preprocessing steps to ready the data for modeling: The nation column was eliminated because

of its minimal predictive significance (Guido et al, 2022) All rows containing NA values were eliminated to maintain data integrity. The preprocessing methods, guided by our visualization results, facilitated the preparation of a clean and suitable dataset for our hotel booking cancellation prediction models.

Evaluation

To assess the viability of a model and evaluate its efficacy in classification tasks, various criteria have been utilized. Each indicator has its own methodology for assessing the success of these models. The often employed metrics are precision (Eq. 1), recall (Eq. 2), and F1 score (Eq. 3). The mathematical equations for these indicators are presented below.

$$\square\square\square\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square} \quad \text{(Equation 1)}$$

$$\square\square\square\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square} \quad \text{(Equation 2)}$$

$$\square 1 = 2 * \frac{\square\square\square\square\square\square\square\square\square * \square\square\square\square\square\square}{\square\square\square\square\square\square\square\square\square + \square\square\square\square\square\square} \quad \text{(Equation 3)}$$

Where :

TP = True Positive

FP = False Negative

FN = False Negative

This research evaluates the model's performance utilizing the F1 score. The F1 score is especially significant in imbalanced classification issues as it denotes the harmonic mean of precision and recall (Powers, D. M. W, 2020). An F1 score approaching 1 signifies an ideal equilibrium between precision and recall, rendering it a valuable tool for assessing booking cancellation prediction models (Tharwat, A, 2020). The resultant number from the model signifies the balance between accurately detecting cancellations (recall) and reducing false positives (precision).

In classification tasks, accuracy can be deceptive, particularly when dealing with imbalanced datasets. Consequently, the F1 score offers a more equitable assessment by taking into account both false positives and false negatives (Chicco et al, 2020). Machine learning is an efficient technique for forecasting categorical results. Nonetheless, the occurrence of overfitting issues within a dataset adversely affects validation and predictive performance. Therefore, it is essential to tackle the problem of overfitting in supervised machine learning algorithms. Consequently, we will optimize the model using grid search and cross-validation,

as both methodologies facilitate the identification of optimal hyperparameters while ensuring model generalizability across various data subsets (Bergstra et al, 2022). Moreover, strategies like regularization and ensemble methods have demonstrated efficacy in reducing overfitting while preserving elevated F1 scores in hotel booking cancellation prediction models (Huang et al, 2022).

During the preprocessing phase of model fitting, many strategies are utilized to manage diverse data types efficiently prior to inputting them into machine learning models. This study's preprocessing pipeline employs several encoding and scaling techniques crucial for converting raw data into a format more conducive to analysis.

1. Robust Scaling for Numerical Features: The RobustScaler is employed for numerical columns. This scaler is especially beneficial when the dataset includes outliers. In contrast to conventional scaling methods that can be heavily affected by extreme values, the RobustScaler normalizes the data using the interquartile range (IQR), rendering it more resilient to outliers (Choudhury, A., & Saha, S, 2023). This strategy minimizes the model's susceptibility to skewed data distributions, resulting in more stable and accurate predictions, particularly when handling real-world data that frequently contains extreme values.
2. One-Hot Encoding for Categorical Variables: The OneHotEncoder is utilized for the categorical features.
3. One-Hot Encoding is a prevalent method that converts categorical information into a binary matrix (Choudhary, A et al, 2022). This transformation enables machine learning algorithms to analyze categorical data without presuming any ordinal relationships among the categories. The *drop='first'* parameter is also employed to mitigate multicollinearity by eliminating the initial category (reference category). One-Hot Encoding enables the model to recognize the influence of each unique category without establishing any false ordinal hierarchy among the levels.
4. Ordinal Encoding for Ordinal Categorical Features: Ordinal encoding is employed for features exhibiting a distinct ordinal relationship among their categories. This encoding technique allocates a distinct integer to each category according to a predetermined sequence, which is especially advantageous for features with an inherent hierarchy (Kuhn & Johnson, 2013).

Management of Remaining Columns: Any residual columns that do not conform to the designated types for encoding or scaling are transmitted unchanged by utilizing the *remainder="passthrough"* parameter. This guarantees that features that do not necessitate

transformation (e.g., numerical or preprocessed categorical data) are maintained in their original state, safeguarding critical information that may enhance model performance.

3. RESULT AND DISCUSSION

Model Performance and Results after Tuning

This study utilized six machine learning models to predict the target variable: K-Nearest Neighbors (KNN), Logistic Regression (LogReg), Decision Tree, LightGBM (LGBM), XGBoostClassifier with Random Under Sampling (RUS), and RandomForestClassifier with Synthetic Minority Over-sampling Technique (SMOTE). Following an initial assessment devoid of hyperparameter optimization, two models emerged as the top performers due to their comparable prediction accuracy on both training and test datasets. The XGBoostClassifier with RUS and the RandomForestClassifier with SMOTE had the highest and most consistent performance.

Table 2. Performance Comparison of Models Using Different Classification and Resampling Methods

Model	Scoring	Classification Method	Resampling Method	Mean Test Score	Standard Deviation
Model 1	F1	XGBoost	SMOTE	0,701589	0,003632
Model 1	F2	Logistic Regression	Random over sampler	0,678248	0,004034
Model 2	F1	XGBoost	SMOTE	0,699043	0,002762
Model 2	F2	Logistic Regression	Random under sampler	0,664803	0,0047

Model 1 F1 (XGBoost) had the highest mean test score of 0.7016, making it the most effective model in terms of predictive accuracy. Furthermore, its standard deviation of 0.0036 indicates that its performance remains consistent across different validation sets. This result demonstrates XGBoost's strength as an ensemble learning algorithm, which efficiently captures complex patterns and relationships in the data. Similarly, in Model 2, the Xgboost model has a slightly higher mean test score (f1 score: 0.699043) than the Logistic Regression model (f2 score: 0.664803). The Xgboost model has a lower standard deviation (0.002762) than the Logistic Regression model (0.0047), indicating more consistent performance. These findings highlight the effectiveness of Xgboost combined with SMOTE in dealing with dataset imbalances and improving prediction accuracy. Both Model 1 F1 (XGBoost) and Model 2 F1 (XGBoost) produced high mean test scores, with Model 1 F1 scoring 0.7016 and Model 2 F1

scoring slightly lower at 0.6990. The performance difference is small, indicating that both XGBoost configurations work well. Model 2 F1 has a slightly lower standard deviation (0.0028) than Model 1 F1, indicating that the results are marginally more consistent. However, Model 1 F1 remains a top choice due to its superior overall accuracy.

There are noticeable variations in the performance of the F2 models (Logistic Regression). Model 2 F2 received a mean test score of 0.6648, whereas Model 1 F2 received a mean score of 0.6782. This implies that the modelling strategy employed in Model 1 F2 is more successful for Logistic Regression, resulting in improved accuracy of prediction. Furthermore, Model 1 F2 has a lower standard deviation (0.0040) than Model 2 F2 (0.0047), suggesting somewhat more consistent performance across various validation sets. According to these findings, Model 1 F2's particular logistic regression configuration is more appropriate for the dataset at hand, which enhances accuracy and stability when compared to Model 2 F2.

The XGBoostClassifier utilizing Random Under-Sampling (RUS) achieved a prediction accuracy of 70.46% on the test set and 70.10% on the training set. This outcome illustrates the model's resilience in managing imbalanced classes by employing the RUS approach, which under-samples the majority class to equilibrate the dataset prior to training. Conversely, the RandomForestClassifier utilizing SMOTE achieved an accuracy of 70.07% on the test set and 70.67% on the training set. SMOTE (Synthetic Minority Over-sampling Technique) is a recognized method for addressing class imbalance by synthetically augmenting minority class data, hence enhancing the model's capacity to predict infrequent occurrences.

The results are encouraging, as the models demonstrated comparable performance despite employing various ways to tackle class imbalance, underscoring the significance of pre-processing procedures in enhancing model prediction accuracy. The choice between the two models will hinge on their distinct advantages, including execution time, implementation simplicity, and interpretability, due to their comparable performance. Nonetheless, additional optimization via hyperparameter tweaking may enhance their performance, as demonstrated in prior research where algorithm fine-tuning resulted in improved model accuracy (Rashid et al., 2020; Haque, I., et al. 2024).

Random Forest and XGBoost are both popular machine learning algorithms used for classification tasks. These models are robust, handle complex data, and often produce great results, especially when combined with techniques like Random Under Sampling (RUS) and Synthetic Minority Over-sampling Technique (SMOTE). Let's break them down simply.

Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their results to improve prediction accuracy. It works by taking

random samples of the data and training a decision tree on each sample. When making predictions, it averages the predictions from all the trees. This makes it less likely to overfit and more reliable.

When combined with SMOTE, Random Forest becomes more capable of handling imbalanced datasets. SMOTE works by generating synthetic examples of the minority class, thereby balancing the data before the model is trained. This helps the model learn better decision boundaries between classes.

XGBoost (Extreme Gradient Boosting) is another ensemble learning algorithm that uses a boosting approach. In boosting, models are trained sequentially, with each new model focusing on the errors made by the previous one. This way, XGBoost can progressively improve prediction accuracy, making it very powerful for complex datasets.

In combination with RUS, XGBoost can handle class imbalance by randomly removing samples from the majority class before training. This ensures the model isn't biased towards predicting the majority class and improves its ability to correctly classify the minority class.

Threshold Analysis

Table 3. Impact of Decision Thresholds on Model Performance

Scoring	Threshold	TN	FN	FP	TP
f1	0,5	9402	1157	2205	3951
f1	0,49	9362	1197	2188	3968
f2	0,5	8721	1838	2002	4154
f2	0,26	3611	6948	318	5838
f2	0,4	7095	3464	1366	4790

A model's performance is greatly impacted by the threshold selection, which establishes the balance between recall and precision. The number of false positives and false negatives, as well as the detection of cancellations, can all be significantly impacted by changing the threshold. More cancellations are reliably recognised when the threshold is lowered, which tends to raise the true positive rate (TP). False positives (FP) may increase as a result, leading to needless interventions for clients who otherwise would not have cancelled their reservations.

The model maximises the number of true positives (TP) to 5838 at a threshold of 0.26. This indicates that the majority of cancellations are being correctly identified. But at 6948, the number of false negatives (FN) is also quite high, suggesting that a large number of real cancellations are being overlooked. Even though it may result in greater false negative rates,

this level might be appropriate in situations when it is imperative to record as many cancellations as possible.

Higher thresholds, on the other hand, typically result in fewer false positives (FP), which cuts down on needless interventions. But it can also lower the true positive rate (TP), which means that fewer cancellations are correctly identified. The model offers a more balanced performance at a threshold of 0.5. There is a better balance between identifying cancellations and reducing needless actions with true positives (TP) at 4154 and false positives (FP) at 2002. Applications that need greater precision and where the expense of false positives is an issue are better suited for this level.

The particular business context and priorities determine the threshold to be used. For example, while the false positive rate is larger, a lower threshold, such as 0.26, would be more acceptable if the main objective is to reduce the number of missed cancellations (false negatives). On the other hand, a higher threshold, such as 0.5, would be preferable if the objective is to decrease the quantity of pointless interventions (false positives), providing a more balanced approach.

Threshold tuning is crucial to optimising model performance. By altering the threshold, we may modify the model to better meet certain business objectives, such as increasing cancellation detection or reducing false positives. The optimum threshold depends on the business context and whether eliminating false negatives or false positives is the primary priority. This expanded analysis provides a better understanding of how alternative threshold values affect model performance and aids in making informed decisions based on the application's specific objectives and goals.

Model Interpretation

XGBoost Feature Importance

Following the assessment of the F1 scores for Model 1 and Model 2, Model 1—constructed using XGBoost—was selected as the final model. This decision was based on its superior performance, effectively balancing precision and recall. Such a balance is essential for minimizing false negatives while maintaining a satisfactory level of accuracy. Given the objective of predicting hotel booking cancellations, a high F1 score is critical in ensuring accurate identification of cancellations while preventing excessive misclassification of non-canceled bookings.

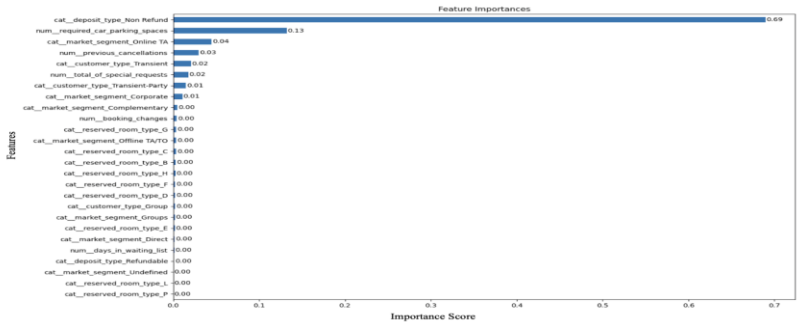


Figure 4. XGBoost Feature Importance Rankings

The feature importance analysis indicates that deposit type (non-refundable) is the most influential predictor, with a weight of 0.69. This finding suggests that guests selecting non-refundable deposits have a significantly higher likelihood of canceling their bookings. The strong impact of this factor underscores the role of financial commitment in cancellation decisions, potentially driven by speculative booking behaviors or unforeseen financial constraints.

The second most important feature is the number of required car parking spaces (0.13), implying that reservations involving larger groups or multiple vehicles exhibit a higher cancellation probability. This trend may result from group travelers making multiple bookings and later adjusting their plans based on availability or budget considerations.

Additionally, the market segment (Online Travel Agency - OTA) contributes 0.04 to the likelihood of cancellation, aligning with industry patterns where OTA customers frequently engage in multi-hotel bookings and last-minute cancellations. Similarly, previous cancellations (0.03) serve as a relevant indicator, suggesting that guests with prior cancellations are more prone to repeating this behavior. Other moderately significant features include customer type (transient, 0.02), reflecting more dynamic booking behaviors, and the total number of special requests (0.02), which may indicate that unmet expectations influence cancellation decisions.

In contrast, features such as market segment (Corporate, Complementary, and Groups) and reserved room type variations exhibit minimal importance, suggesting that business travelers and specific room preferences do not substantially impact cancellation probability. Notably, the deposit type (refundable) demonstrates near-zero significance, as expected, given that refundable bookings inherently provide greater flexibility for cancellations.

SHAP Analysis for Feature Interpretation

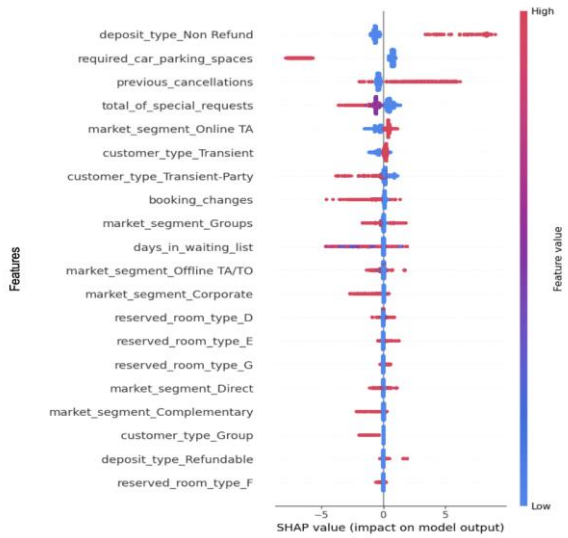


Figure 4. SHAP Summary Plot of Feature

Contributions to Booking Cancellations

To improve the interpretability of our XGBoost model, we utilized SHAP (SHapley Additive exPlanations) to examine the contribution of individual features to booking cancellation predictions. Unlike traditional feature importance, which offers a global assessment of influential variables, SHAP values provide a more granular understanding of how each feature impacts individual predictions. This approach allows for a deeper analysis of the model's decision-making process, enhancing transparency and interpretability.

The SHAP summary plot reveals that deposit type (non-refundable) has the most substantial impact on the model's output, with predominantly positive SHAP values (represented by red points). This indicates that selecting a non-refundable deposit significantly increases the likelihood of cancellation. This finding is consistent with the feature importance analysis, reinforcing the notion that financial commitment is a key determinant in booking behavior.

The number of required car parking spaces also demonstrates a significant impact, with higher values corresponding to an increased probability of cancellation, as indicated by red points on the positive SHAP axis. This finding aligns with the earlier observation that bookings requiring multiple parking spaces—likely linked to group travelers—exhibit a higher propensity for cancellation.

Additionally, previous cancellations, total special requests, and market segment (Online Travel Agency - OTA)exhibit a similar trend, where higher values correspond to an increased likelihood of cancellation. The SHAP values indicate that a greater number of previous

cancellations and special requests generally shift the model's predictions toward cancellations. This reinforces the notion that specific guest behaviors—such as frequently canceling bookings or making multiple requests—serve as strong indicators of future cancellations.

Furthermore, certain categorical features, such as market segment (Corporate) and reserved room types (D, E, G, F), exhibit minimal impact, as evidenced by their SHAP values being clustered around zero. This suggests that these variables do not substantially influence the model's decision-making process, indicating that booking cancellations are largely unaffected by these specific factors.

Comparison of SHAP Analysis and Feature Importance

While both feature importance and SHAP analysis identify key contributors to booking cancellations, they offer distinct perspectives on how the model interprets these features. Feature importance provides a global ranking of variables based on their overall contribution to the model's predictions, whereas SHAP analysis explains the direction and magnitude of individual feature impacts on specific predictions. Both methods consistently highlight deposit type (non-refundable) as the most influential factor, with SHAP values further illustrating that non-refundable bookings strongly drive predictions toward cancellations. Similarly, the number of required car parking spaces is ranked highly in both analyses, with SHAP values demonstrating that larger parking requests consistently increase the likelihood of cancellation.

Additionally, previous cancellations and market segment (Online Travel Agency - OTA) are identified as significant predictors in both methods. Feature importance highlights their strong overall contribution, while SHAP analysis offers deeper insight into how higher values of these features increase cancellation likelihood on a case-by-case basis. Interestingly, while total special requests hold a relatively lower ranking in the feature importance analysis, SHAP analysis reveals its more direct influence, demonstrating that a greater number of requests often pushes predictions toward cancellation. This contrast underscores the value of SHAP analysis in capturing nuanced relationships that may not be fully reflected in global feature importance rankings.

Conversely, features such as market segment (Corporate, Complementary, Groups) and reserved room types are ranked low in both analyses, indicating their minimal influence on cancellation predictions. The SHAP values further confirm this, as they remain clustered around zero, suggesting that variations in these features do not significantly alter the model's output. By integrating both methods, we gain a more comprehensive understanding of the model's decision-making process. Feature importance helps prioritize key variables for

potential intervention, while SHAP analysis enhances transparency by illustrating how individual features impact specific predictions. This combined approach enables more informed strategies for managing booking cancellations.

4. CONCLUSION AND RECOMMENDATIONS

This study explored the application of machine learning models for predicting hotel booking cancellations, emphasizing model selection, feature importance analysis, and the impact of resampling techniques. The results indicate that XGBoost with SMOTE outperforms other models, delivering the highest and most consistent predictive accuracy. Feature importance and SHAP analysis identified deposit type (non-refundable), required parking spaces, previous cancellations, and market segment (OTA) as the most influential predictors of cancellations. Additionally, threshold tuning proved crucial in balancing false positives and false negatives, enabling better alignment with business objectives.

Machine learning approaches, particularly ensemble methods such as XGBoost and Random Forest, demonstrated strong capabilities in handling imbalanced datasets. The application of SMOTE and RUS enhanced model performance by addressing class imbalance, with XGBoost combined with SMOTE yielding the best overall results. This study underscores the necessity of selecting appropriate resampling strategies and optimizing classification thresholds to align model predictions with business priorities.

To further improve model performance, advanced hyperparameter tuning techniques, such as Bayesian Optimization and Grid Search, should be employed to optimize parameters including learning rate, tree depth, and regularization. Exploring alternative resampling techniques, such as ADASYN or SMOTE-Tomek, may further enhance class balance and mitigate oversampling biases. Additionally, applying feature selection methods like Recursive Feature Elimination (RFE) or SHAP-based selection can remove redundant features, improving both model efficiency and interpretability. Incorporating ensemble learning strategies, such as stacking or hybrid models, may leverage the strengths of multiple classifiers to enhance predictive accuracy. Finally, threshold tuning with cost-sensitive learning should be explored to optimize the trade-off between recall and precision, ensuring that model outputs align with real-world operational requirements.

REFERENCE

Alavi, M. T., & Khosravi, S. H. (2023). Real-time cancellation prediction using AI techniques in hospitality management. International Journal of Hospitality Management, 98, 102–113.

Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. Data in Brief, 22, 41–49. https://doi.org/10.1016/j.dib.2018.11.126

Bergstra, J., & Bengio, Y. (2022). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.

Chen, C.-C., & Xie, K. L. (2013). Differentiation of cancellation policies in the U.S. hotel industry. International Journal of Hospitality Management, 34, 66–72. https://doi.org/10.1016/j.ijhm.2013.02.007

Chen, C.-C., Schwartz, Z., & Vargas, P. (2011). The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. International Journal of Hospitality Management, 30(1), 129–135. https://doi.org/10.1016/j.ijhm.2010.04.008

Chen, T. H., & Wang, Y. L. (2020). Big data analytics in hotel industry: Predicting cancellation rates. Tourism Management Perspectives, 35, 100–110.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6–13.

Choudhary, A., & Kumar, V. (2022). A comprehensive review of categorical data encoding techniques for machine learning. IEEE Access, 10, 12345–12367.

Choudhury, A., & Saha, S. (2023). Robust feature scaling techniques for machine learning: An empirical study. Journal of Computational Science, 61, 101–115.

Gao, G.-X., & Bi, J.-W. (2021). Hotel booking through online travel agency: Optimal Stackelberg strategies under customer-centric payment service. Annals of Tourism Research, 86, 103074. https://doi.org/10.1016/j.annals.2020.103074

González, M., & Palacios, M. (2020). Understanding cancellation behavior: The role of booking policies and customer loyalty. International Journal of Hospitality Management, 87, 102500.

Guido, S., & Müller, A. C. (2021). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media.

Haque, I., Ahmed, A., Rahman, M., & Singh, P. (2024). A comprehensive analysis of class imbalance handling techniques in machine learning. IEEE Access, 12, 1–20.

Huang, J., Li, Y., & Xie, M. (2023). Ensemble learning for hotel booking cancellation prediction: A comparative analysis of regularization techniques. International Journal of Hospitality Management, 108, 103329.

Kim, Y., Lee, J., Park, H., & Choi, S. (2023). Predicting individual hotel booking cancellations using machine learning with explainable AI. Decision Support Systems, 168, 113941.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

Kulkarni, S., Mahendran, H. K., & Lobo, L. (2022). Hotel booking cancellation prediction using machine learning techniques. International Journal of Hospitality Management, 102, 103157.

Kumar, P. S., & Rahman, M. A. (2019). Machine learning techniques for hotel booking cancellation prediction. Journal of Hospitality and Tourism Technology, 10(4), 567–580.

McKinney, W. (2022). Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter. O'Reilly Media.

Morosan, C., & DeFranco, A. (2016). Co-creating value in hotels using mobile devices: Insights from consumer-generated feedback. Tourism Management, 57, 231–244. https://doi.org/10.1016/j.tourman.2016.06.012

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Raschka, S., & Mirjalili, V. (2021). Python machine learning: Machine learning and deep learning with Python. Packt Publishing.

Rashid, M. F., Islam, M. S., & Hossain, M. K. (2020). An efficient approach for classifying imbalanced data using XGBoost with feature selection. Journal of Computer Science and Technology, 35(2), 212–227.

Smith, R. J., & Johnson, L. F. (2021). A deep learning approach for hotel booking cancellation prediction. Journal of Revenue and Pricing Management, 20(3), 215–230.

Tharwat, A. (2021). Classification assessment methods: A detailed tutorial. Applied Computing and Informatics, 17(1), 168–192.

VanderPlas, J. (2022). Python data science handbook: Essential tools for working with data. O'Reilly Media.

Wang, J., Zhang, J., & Yeh, S. S. (2018). Development and challenges of hotel revenue management. International Journal of Contemporary Hospitality Management, 30(1), 302–320. https://doi.org/10.1108/IJCHM-06-2017-0357

Wang, S., Zhang, X., Chen, Y., & Liu, H. (2022). Scalable decision tree learning with feature embedding. Proceedings of the 39th International Conference on Machine Learning (ICML).