Efficient Revenue Management: Classification Model for Hotel Booking Cancellation Prediction

Chin Lei[1*], Mohamed Ibrahim[2]

[1*]*Faculty of Management, Universiti Teknologi Malaysia, Malaysia.*

[2]*Faculty of Management, Universiti Teknologi Malaysia, Malaysia.*

Received: 04 January 2024; Revised: 22 January 2024; Accepted: 07 February 2024; Published: 29 March 2024

Abstract

One of the major problems in hotel business management is related to the problem of cancellation of hotel bookings, which implies causing losses of considerable revenues and disruptions of operational activities. Minimizing cancellation prediction errors can assist hotels in setting proper price models and effectively utilizing resources. The purpose of this work is to improve the current classification model for the identification of hotel booking cancellations. Failure to forecast the number of cancellations is another problem because the hotel has to lay down more inventory and loses money in the process. The data collected from kaggle involved variables like booking lead time, customer characteristics, or booking tendencies. Those include data pre-processing functions, in particular min-max normalization algorithms. In this study, the Linear Discriminant Analysis (LDA) method was used in feature extraction to classify booking cancellations. The performance of the Osprey Optimization Fine-Tuned Random Forest (O-FRF) model was assessed using different statistical measures like accuracy (92.12%), precision (88.54%), recall (91.36%), and F1-score (90.67%). The developed classification model could be used by hotel revenues as a valuable tool since it gives accurate probabilities of booking cancellation.

Keywords: Revenue Management; Hotel Booking; Cancellation Prediction; Classification Model; Risk Management.

I. INTRODUCTION

The management of revenues in the highly competitive industry is vital for sustaining the financial health and efficiency of operations since it impacts most of the Hotel's business performance and its Return on Assets (ROA) (Webb et al., 2020). Among the most important problems that hotels frequently encounter is the high cancellation rates that result in loss of revenues and logistical inefficiencies (Viglia et al., 2021). At the same time, cancellations of bookings are a real problem not only in terms of their impact on current revenues but also in terms of stock management and pricing. Customer behaviour, unfavourable economic conditions, and fluctuation in the seasons all play a part in contributing to the cancellation of bookings in hotels (Guerster et al., 2020). Such cancellations cause a lot of income losses, especially when momentary cancellations deprive the hotel of an opportunity to sell the rooms. Therefore, hotels are forced to look for policies that will help them reduce the extent of cancellations and improve their revenue management (Pereira & Cerqueira, 2022). In recent

years, data analytics and machine learning have given new leads to optimize cancellation prediction (Sekhon & Ahuja, 2024). In particular, there is a perspective of classification models to help discover a set of bookings as risky and, therefore, bring more effective actions to manage the revenue (Febrian et al., 2024). These models can classify the bookings into the various risk categories, meaning that by considering the historical booking data and customers' profile that contains the information on likely cancellation, such strategies are valuable in the sense that they offer foresight into the cancellations. The study goal is to develop an accurate classification model for predicting hotel booking cancellations to improve revenue management strategy and minimize financial loss.

II. LITERATURE REVIEW

In comparison to previous pertinent works in the field (Sánchez-Medina & Eleazar, 2020) aimed to develop a strategy for predicting cancellations of hotel reservations that utilized just 13 distinct factors. Genetic algorithms were used to optimize artificial neural networks and machine learning approaches, resulting in the highest cancellation rate. Novakovic & Turina, (2021) used machine learning techniques to anticipate passenger cancellation of hotel reservations in the tourist industry. Numerous algorithms, such as logistic regression, K-neighbour, AdaBoost, random forest classifiers, decision trees, and bagging were investigated and their performances were associated with the investigation. The findings indicated that every algorithm has a certain area of expertise and by taking into account comprehensive reservation data, they might assist hotel staff in anticipating the potential of cancellations.

Satu et al., (2020) purpose was to examine the impact of machine learning approaches on cancelled hotel reservations. There were several feature modification techniques that were employed on a dataset obtained through Kaggle. Categorical variables were explained and truncated using a number of classifiers. When focused on individual classifiers, the most frequently applied analytic approach, XGBoost, also showed the best performance. In as much as cancellation affects future income and resource management (Chen et al., 2022) focused on the efficiency and prediction of economies of cancellation in the hotel industry. It presented three potential neural network replacements: CatBoost, k-Nearest Neighbour (KNN), and logistic regression. Among all of them, CatBoost was the most effective classifier in hotel prediction. The results prove the high efficiency of CatBoost for the numerical values of reservation cancellations for the accurate prognosis of their number and the achievement of the highest scores of prognoses. Sánchez et al., (2020) employed personal name records (PNR) data to approximate the cancellation rate at the individual hotel category near service hours. It sought to minimize the difference by identifying those who were most likely to cancel their bookings in the shortest time possible. Information about the accuracy rate of cancellations made seven days prior to the date of cancellation made it possible to streamline the booking management systems as well as policies on cancellations.

III. METHODOLOGY

This section discusses the data collection for booking cancellation in the hotel, pre-processing of the data using min-max normalization and feature extraction for performing LDA. The

Osprey Optimization Fine-Tuned Random Forest (O-FRF) model is used to forecast the cancellation of bookings in hotels. Figure 1 illustrates the overview of the methodology.

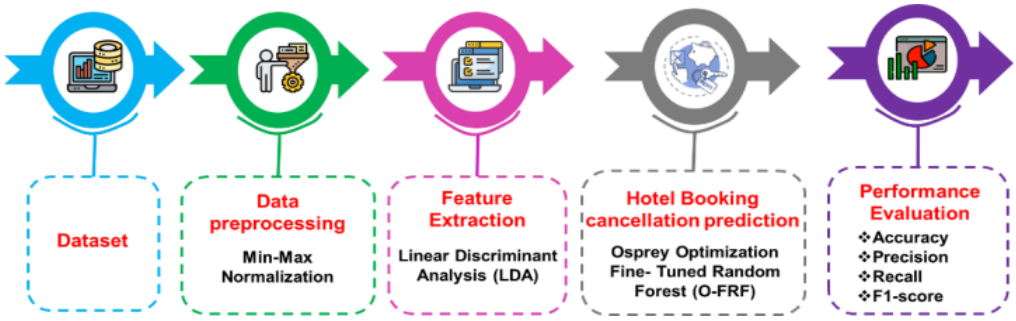

Figure 1: Workflow of Hotel Booking Cancellation

3.1. Data Collection

For this research, datasets were gathered from kaggle [https://www.kaggle.com/datasets/gauravduttakiit/reservation-cancellation-prediction]. This dataset contains detailed attributes of customer reservations to assist in predicting booking cancellations. It includes various features such as unique booking identifiers, guest demographics, details about the stay, meal plans, car parking requirements, room types, lead times, arrival dates, booking status (0/1), previous cancellations, and special guests.

3.2. Data Pre-processing

Min – Max normalization is an essential procedure for scaling features in data pre-processing, particularly for tasks such as predicting hotel booking cancellations. Features like booking lead time and amounts to acommon scale [0, 1]. Ensuring that every feature contributes equally to the categorization model improves revenue management by helping the model forecast cancellations more accurately. Min-Max normalization uses the following equation (1) to scale feature values to a certain range:

$$v' = \frac{v - min_A}{min_A - max_A}(new\_max_A - new\_min_A) + new\_min_A \qquad (1)$$

Where v'represents the normalized value and v represents the original feature value for the specified feature, max_A and min_A are the highest and lowest values, for the given feature A, and new_max_A and new_min_A define the target range [0,1]. By accelerating convergence throughout the training phase, this step improved the model's performance.

3.3. Feature Extractionusing Linear DiscriminantAnalysis (LDA)

For the dimensionality reduction problems, the best technique is called LDA that is utilized in the pre-processing step in the execution of the hotel booking cancellation prediction classification solution. LDA is used to reduce the feature space so as to investigate the class model with the highest ratio of within-class to between-class dispersion. The high-dimensional data must be mapped into the subspace L (L ≤6m-1) so as to get maximum distinction between cancelled and non-cancelled bookings. The key equations (2) to (7) used in LDA for this purpose are:

$$Tx = \sum_{w \in C_j}^{d} Tx_j = (w_j - \mu_j)(w_j - \mu_j)^S \qquad (2)$$

$$\mu = \frac{1}{M}\sum_{j=1}^{M} w_j \quad (3)$$

$$Ta = \sum_{j=1}^{d} w_j Ta_j = (m_j - m)^2 \quad (4)$$

$$(n_j - n)^2 = X^S(\mu_j - \mu)(\mu_j - \mu)^S \quad (5)$$

$$\sum_{j=1}^{m} M_i\,(n_j - n)(n_j - n)^2 \quad (6)$$

$$N_j = \frac{1}{m}\sum_{w\in C}^{m} W_l \quad (7)$$

To determine the mean value of each input value in a class (L), employ the equation below: By the eigenvectors $\lambda$of the transformation matrix$X$, the following holds: $TxX = \lambda TaX$and the mean value of the input value is calculated by dividing the total number of values by the total values of the measure. It also aids in data reduction where only the features that are relevant in distinguishing between cancelled and non-cancelled bookings are retained.

3.4. Osprey Optimization Fine-Tuned Random Forest (O-FRF)

The Osprey Optimization Fine-Tuned Random Forest (O-FRF) is a complex form of a predictive model designed for application in hotel booking cancellation. The O-FRF helps to increase the predictivity and the performance of the method since Random Forest is solid and enriched with aspects of the Osprey technique, and due to that, the forecast of cancellations in hotels is exacted and the hotel does not lose its potential revenues.

3.4.1. Fine-Tuned Random Forest (FRF)

The FRF algorithm uses the ensemble learning method, which involves building several decision trees to improve the accuracy level and model interpretability. It provides anadded splitting criterion, performs fewer calculations, and gives better algorithms. For the hotel booking cancellation prediction, this method can support big data and enhance the classification by optimized algorithms. The node splitting formula explains the amount of information gained and Gini index value when attributes 'a' to split sample set C. The equation (8, 9) is given as,

$$Gain(C, b) = Ent(C) - \sum_{u=1}^{U} \frac{|C^u|}{|C|} Ent(C^u) \quad (8)$$

$$Gini(C, b) = \sum_{u=1}^{U} \frac{|C^u|}{|C|} Gini(C^u) \quad (9)$$

C^u indicates that every sample in the C with a value of b^u on the attribute b is contained in the u branch node. The equation (10, 11) is given as,

$$Ent(C) = -\sum_{l=1}^{|z|} o_l log_2\, o_l \quad (10)$$

$$Gini(C) = \sum_{l=1}^{|z|}\sum_{l\prime\neq 1} olol' = 1 - \sum_{l=1}^{|z|} ol^2 \quad (11)$$

By integrating an adaptive parameter selection procedure with the node splitting formula, the node splitting principle seeks to improve the purity of the data set following division. The equation (12) is given as,

$$G = \min_{b,\beta \in Q} E\{c, b\} = bGin(C, b) - \beta Gain(C, b)$$
$$s.t\,\{\begin{matrix}\alpha + \beta = 1 \\ 0 \le \alpha, \beta \le 1\end{matrix} \quad (12)$$

α, β denotes the attribute splitting weight coefficient in this case. To increase the classification impact, the work focuses as node partition criteria using an adaptive parameter selection approach. A equation (13, 14) is established for the classification error rate of

sample C. Performance is determined by using accuracy and classification error rates.

$$F(e,c) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{JJ}(\mathrm{e}(w_j) \neq z_j) \quad (13)$$

$$acc(e;C)\frac{1}{n}JJ\big(e\big(w_j\big) = z_j\big) = 1 - F(e;C) \quad (14)$$

3.4.2. Osprey Optimization Algorithm (OOA)

The OOA uses natural hunting techniques of the osprey to get the best parameter estimations for model updates required to enhance accuracy and effectiveness for the case of hotel booking cancellations prediction. The OOA considers each osprey location as a possible solution to the hotel booking cancellation issue, using an $N \times D$-dimensional matrix X and Equation (15) to generate a random location for each osprey.

$$X_{i,j} = lb_j + r_{i,j}\big(ub_j - lb_j\big), i = 1,2,\cdots N; j = 1,2,\cdots, D \quad (15)$$

The i^{th}osprey's initial position in the j^{th}dimension is represented by $X_{i,j}$, where $r_{i,j}$is an integer value between 0and 1, the population number is N, the solution dimension is D, and the j^{th} dimension is referred to as j. Each osprey represents a possible solution to the hotel booking cancellation prediction issue. The fitness value is evaluated using the objective function F to determine the quality ofthe solution, using Equation (16).

$$F_i = F(X_i) i = 1,2,\cdots N \quad (16)$$

The i^{th} osprey's fitness value is given as F_i, while X_i symbolizes its location.

The global exploration phase of the osprey optimization process involves a significant shift in the osprey's location within the search space. Under the OOA paradigm, every osprey looks at where other ospreys with greater fitness levels are positioned, determining their positions using Equation (17).

$$FP_i = \{\{X_K | K \in \{1,2,\cdots N\} \cap F_k < F_i\} \cup \{X_{best}\}, i = 1,2,\cdots N \quad (17)$$

The i^{th}osprey's location set is determined by FP_i, where N represents the number of ospreys, $F_k$represents the fitness values of $K^{th}$and i^{th} ospreys, and X_{best} represents the best osprey's location. An osprey attacks a fish randomly, and Equation (18) simulates the location update procedure.

$$X_{i,j}^{p1} = X_{i,j} + r_{i,j}\big(SF_{i,j} - I_{i,j}.X_{i,j}\big), i = 1,2,\cdots N \quad (18)$$

The new i^{th}osprey's position in phase 1 is determined by the original position, with $X_{i,j}^{p1}$ representing its jth dimension. The first osprey's fish is designated as $SF_{i,j}$, with $r_{i,j}$being a random integer in the range [0, 1]. $I_{i,j}$ is randomly chosen from {1,2}. Equation (19) is used to modify the updated location if it is outside the boundary.

$$X_{i,j}^{p1} = \begin{cases} X_{i,j}^{p1}, lb_j \leq X_{i,j}^{p1} \leq ub_j \\ lb_j, X_{i,j}^{p1} < lb_j \\ ub_j, X_{i,j}^{p1} > ub_j \end{cases} \quad (19)$$

Equation (20) replaces the prior position with the updated fitness value calculated by Equation (18) and (19), resulting in the osprey's updated position. The fitness of the new location after phase 1 is represented by $F_{i,}^{p1}$.

$$X_{i,}^{1} = \begin{cases} X_{i,}^{p1}, F_{i,}^{p1} < F_i \\ X_i, F_{i,}^{p1} \geq F_{i,}^{p1} \end{cases} \quad (20)$$

The osprey moves fish to a safe spot after hunting, increasing the OOA's capability for local searches. This phase, known as the local development phase, involves boundary processing activities like Equation (21). The position update follows Equation (22), and comparable to the phase of global exploration, boundary processing activities should be carried out.

$$X_{i,j}^{p2} = X_{i,j}^{1} + \frac{lb_j + r_{i,j}(ub_j - lb_j)}{t}, i = 1,2,\cdots N;\ j = 1,2,\cdots D;\ t = 1,2,\cdots T \quad (21)$$

$$X_{i,j}^{p1} = \begin{cases} X_{i,j}^{p2}, lb_j \leq X_{i,j}^{p2} \leq ub_j \\ lb_j, X_{i,j}^{p2} < lb_j \\ ub_j, X_{i,j}^{p1} > ub_j \end{cases} \quad (22)$$

In phase 2, Equation (23) is used to calculate the new location $X_{i,j}^{p2}$of the i^{th}osprey, where $X_{i,j}^{p2}$represents the j^{th} dimension. If the revised position's fitness value is higher, the former location is replaced. The fitness value of location $X_{i,}^{p2}$is indicated by $F_{i,}^{p2}$, where $X_{i,}^{p2}$is the osprey's position after phase 2. The OOA updates Osprey locations until the optimum solution is found or the maximum iteration limit is achieved.

$$X_{i,}^{1} = \begin{cases} X_{i,}^{p2}, F_{i,}^{p2} < F_i \\ X_{i,}^{1}, F_{i,}^{p2} \geq F_{i,}^{p1} \end{cases} \quad (23)$$

The fitness value of location $X_{i,}^{p2}$ is indicated by $F_{i,}^{p2}$, where $X_{i,}^{p2}$ is the osprey's position after phase 2. Following these steps, the OOA repeatedly updates the Osprey locations until the optimum solution is discovered or the maximum iteration limit is achieved.

IV. EXPERIMENTAL RESULT

An Intel i7 CPU with 32 GB RAM and Python environment with libraries like SciKit-learn and TensaFlow are used in the experimental configuration. Utilise the following metrics: accuracy, precision, recall, and F1 score, the effectiveness of the proposed and current methodologies was evaluated. Random Forest (RF) (Andriawan et al., 2020) and Synthetic Minority Over-Sampling Technique + K-Nearest Neighbors (SMOTE +KNN) (Nababan et al., 2022) were the existing approaches compared to the proposed method. Figure 2 shows the result of training and validation.

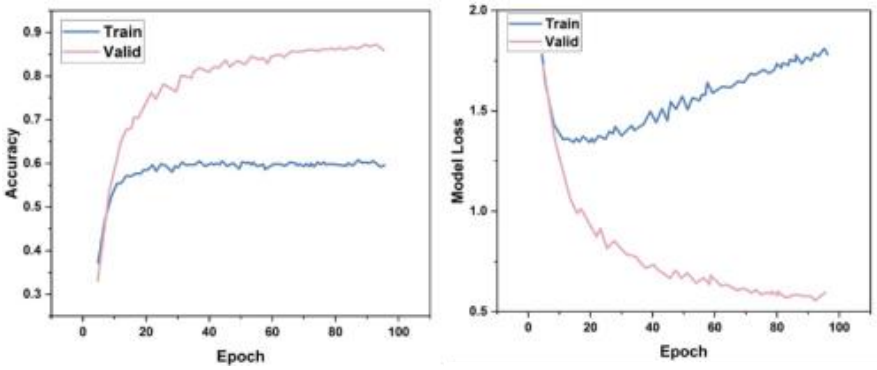

Figure 2: Outcome of Training and Validation

a) **Accuracy:** Itmeasures the proportion of correctly predicted bookings. The proposed **O-FRF (92.12%)** method outperforms existing methods like random forest (87.17%), and SMOTE +KNN (83.23%) in predicting hotel booking cancellation. Figure 3 and Table 1 show that the O-FRF achieves higher accuracy compared to these methods.

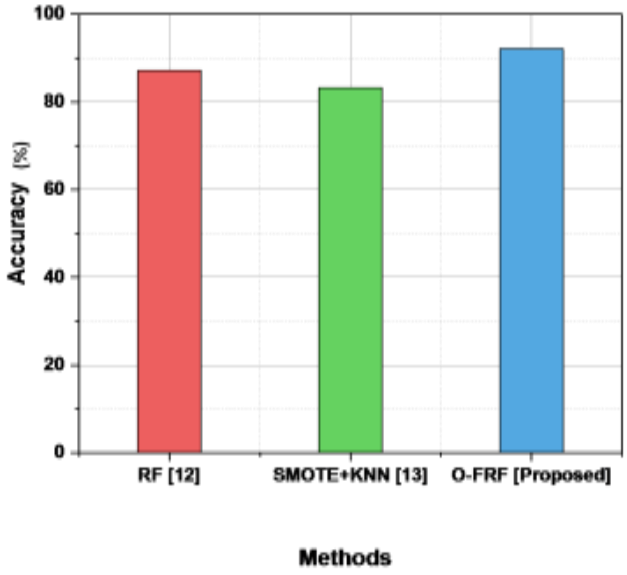

Figure 3: Outcome for Accuracy

b) **Precision:** It measures the ratio of correctly predicted bookings cancellation to total predicted cancellation. The **O-FRF (88.54%)** technique shows better precision compared to existing methods like RF (86.46%), and SMOTE+KNN (83.00%) in predicting hotel booking cancellation. Figure 4 and Table 1 show the outcomes of precision.

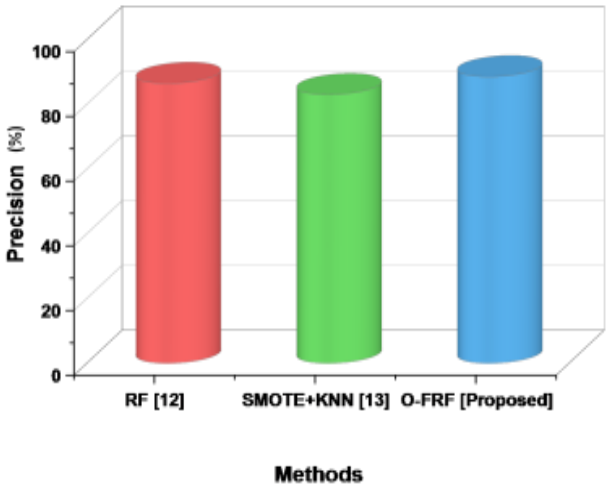

Figure 4: Outcome of Precision

c) **Recall:** It is a measure of the proportion of actualcancellations that are correctly predicted by the model. The proposed **O-FRF (91.36%)** approach surpasses RF (77.50%), and SMOTE +KNN (83.00%). Figure 5 and Table 1 show the outcomes of recall.

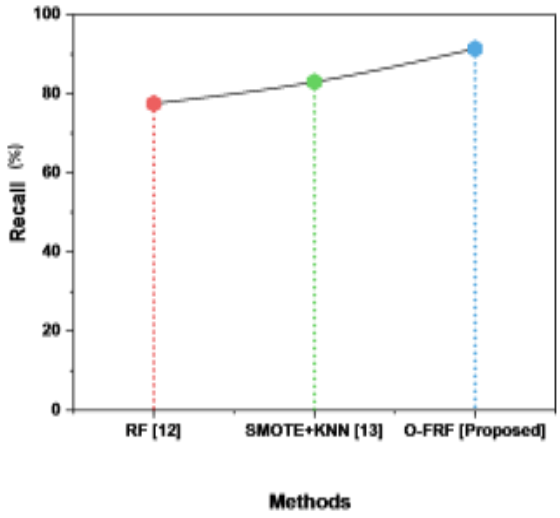

Figure 5: Outcome of Recall

d) F1score: F1score isthe average of precision and recall, a balanced measure of the model's performance of both metrics. The suggested method **O-FRF (90.67%)** outperforms SMOTE + KNN (83.00%) and random forest (81.73%). Figure 6 and Table 1 show the outcomes of F1-score.

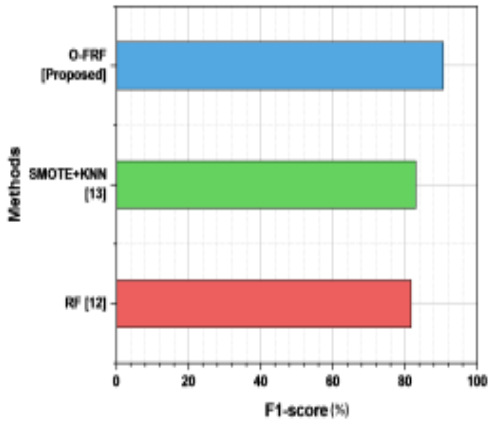

Figure 6: Outcome of F1score

Table 1: Outcome of Model Performance

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1score (%)**
RF [12]	87.17	86.46	77.50	81.73
SMOTE +KNN [13]	83.23	83.00	83.00	83.00
O-FRF [Proposed]	**92.12**	**88.54**	**91.36**	**90.67**

V. CONCLUSION

Study aimed to establish a reliable classification model for hotel booking cancellations to overcome the issues of inadequate revenue prediction and misappropriation of resources. The features of the Kaggle dataset included booking lead time, customer's age, company type,

required amenities, booking frequency, and more and to pre-process the data, the min-max normalization was used, while LDA was used for feature selection. The Osprey Optimization Fine-Tuned Random Forest (O-FRF) model was found to provide high accuracy, precision, recall, and F1-score of (92.12%, 88.54%, 91.36% and 90.67%). This model can estimate the booking cancellations accurately, which can be considered as a useful model for the hospitality industry primarily for understanding the pricing models and resources better. The performance of the model may differ based on the dataset used and other outside factors affecting cancellations not captured here. Future studies could incorporate real-time data and analyse more sophisticated models of ensembles to improve the prediction and dynamics of the model.

REFERENCES

[1] Webb, T., Schwartz, Z., Xiang, Z., & Singal, M. (2020). Revenue management forecasting: The resiliency of advanced booking methods given dynamic booking windows. *International Journal of Hospitality Management*, *89*, 102590. https://doi.org/10.1016/j.ijhm.2020.102590

[2] Viglia, G., De Canio, F., Stoppani, A., Invernizzi, A. C., & Cerutti, S. (2021). Adopting revenue management strategies and data sharing to cope with crises. *Journal of Business Research*, *137*, 336-344. https://doi.org/10.1016/j.jbusres.2021.08.049

[3] Guerster, M., Grotz, J., Belobaba, P., Crawley, E., & Cameron, B. (2020, March). Revenue management for communication satellite operators-opportunities and challenges. In *2020 IEEE Aerospace Conference* (pp. 1-15). IEEE. https://doi.org/10.1109/AERO47225.2020.9172344

[4] Pereira, L. N., & Cerqueira, V. (2022). Forecasting hotel demand for revenue management using machine learning regression methods. *Current Issues in Tourism*, *25*(17), 2733-2750. https://doi.org/10.1080/13683500.2021.1999397

[5] Sekhon, G., & Ahuja, S. (2024, March). Improving business operation in hospitality using predictive analytics and deep learning. In *AIP Conference Proceedings* (Vol. 3072, No. 1). AIP Publishing. https://doi.org/10.1063/5.0198751

[6] Febrian, Y. Y., Wijaya, D. R., & Ervina, E. (2024, February). Hotel Reservation Cancellation Prediction using Boosting Model. In *2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)* (pp. 138-143). IEEE. 10.1109/ICoSEIT60086.2024.10497479

[7] Sánchez-Medina, A. J., & Eleazar, C. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, *89*, 102546. https://doi.org/10.1016/j.ijhm.2020.102546

[8] Novakovic, J., & Turina, S. (2021). Hotel reservation cancellations: analysis and prediction using machine learning algorithms. *ACADEMIC JOURNAL*, *2(1), 4-13.*

[9] Satu, M. S., Ahammed, K., & Abedin, M. Z. (2020, December). Performance analysis of machine learning techniques to predict hotel booking cancellations in hospitality industry. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICCIT51783.2020.9392648

[10] Chen, Y., Ding, C., Ye, H., & Zhou, Y. (2022, March). Comparison and analysis of machine learning models to predict hotel booking cancellation. In *2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)* (pp. 1363-1370). Atlantis Press. https://doi.org/10.2991/aebmr.k.220307.225

[11] Sánchez, E. C., Sánchez-Medina, A. J., & Pellejero, M. (2020). Identifying critical hotel cancellations using artificial intelligence. *Tourism Management Perspectives*, *35*, 100718. https://doi.org/10.1016/j.tmp.2020.100718

[12] Andriawan, Z. A., Purnama, S. R., Darmawan, A. S., Wibowo, A., Sugiharto, A., & Wijayanto, F. (2020, November). Prediction of hotel booking cancellation using CRISP-DM. In *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICICoS51170.2020.9299011.

[13] Nababan, A. A., Jannah, M., & Nababan, A. H. (2022). Prediction Of Hotel Booking Cancellation Using K-Nearest Neighbors (K-Nn) Algorithm and Synthetic Minority Over-Sampling Technique (Smote). *INFOKUM*, *10*(03), 50-56.