iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53770

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Design Analysis Approach for Hotel Booking Cancellation Prediction

Ayush Kumar[1], Chittranjan Kumar[2], Kanik Goel[3], Prof Yashi Bhardwaj[4]
[1, 2, 3]*B.Tech. Student, Department of Information Technology, IMS Engineering College, Ghaziabad, Uttar Pradesh, India*
[4]*Assistant Professor, Information Technology, IMS Engineering College, Ghaziabad, Uttar Pradesh, India*

Abstract: Hotel managers find it beneficial to predict hotel booking cancellations as it enables them to enhance room inventory management, pricing strategies, and customer satisfaction by proactively addressing potential problems. In this study, we present a machine learning-centered method for forecasting hotel booking cancellations. Overall, our proposed approach will equip hotel managers with a robust tool for predicting hotel booking cancellations and a deeper understanding of the factors influencing them. This, in turn, will enable data- driven decision making to optimize room inventory, pricing strategies, enhance customer satisfaction, and ultimately boost revenue
Keywords: Classification, Machine Learning, Predictive analysis, Random Forest, K-Nearest Neighbors, DecisionTree, Naive Bayes, Logistic Regression.

I. INTRODUCTION

The hotel industry thrives on effective management of bookings and occupancy rates. Hotel managers face the challenge of optimizing room inventory and pricing while ensuring customer satisfaction. One critical aspect of this challenge is predicting hotel booking cancellations. Anticipating cancellations can enable hotel managers to proactively address potential issues, adjust pricing strategies, and make better decisions regarding room availability. In recent years, advancements in machine learning and data analytics have opened new opportunities to develop accurate prediction models for hotel booking cancellations.

This paper aims to propose and implement a Hotel Booking Cancellation Prediction Model using machine learning techniques. The primary goal is to assist hotel managers in optimizing their operations by accurately forecasting booking cancellations. By doing so, they can effectively manage room inventory, minimize revenue losses, and enhance customer satisfaction.

The proposed model will leverage historical data on hotel bookings, including information about customers, booking details, and whether the bookings were cancelled or not. Through a series of data pre-processing steps, including handling missing values, converting categorical variables into numerical form, and scaling numerical variables, the dataset will be prepared for modelling. Various machine learning algorithms will be explored and compared to identify the best-performing model. Algorithms such as logistic regression, decision trees, and random forests will be evaluated using standard evaluation metrics like accuracy, precision, and recall. Additionally, ensemble methods, which combine the predictions of multiple models, will be considered to potentially improve the model's performance. To evaluate the model's generalization ability, a hold-out test set will be utilized. This evaluation will provide insights into how well the model performs on unseen data, determining its practical utility in real-world scenarios. Furthermore, a feature importance analysis will be conducted to identify the key factors influencing hotel booking cancellations. Understanding these factors will enable hotel managers to prioritize specific strategies to reduce cancellation rates, such as targeted marketing campaigns or early booking incentives. In conclusion, This paper aims to develop a robust Hotel Booking Cancellation Prediction Model that will assist hotel managers in optimizing their operations.

The model's accurate predictions will facilitate effective decision-making regarding room inventory management, pricing strategies, and customer satisfaction. By leveraging machine learning techniques and analyzing the factors that contribute to cancellations, hotel managers can make data-driven decisions to increase revenue and create a more streamlined booking process.

II. SIGNIFICANCE OF THE SYSTEM

The significance of this research paper lies in its comprehensive analysis of machine learning algorithms for predicting hotel booking cancellations. By evaluating the effectiveness of various models, the paper offers valuable insights to hotel operators and online travel agencies. Accurate prediction of cancellations can optimize revenue management, resource allocation, and customer satisfaction. The findings contribute to the existing literature and provide practical guidance for implementing efficient prediction models, ultimately benefiting the hospitality industry.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

III. LITERATURE SURVEY

A literature survey on hotel booking cancellation prediction reveals several significant studies that have explored various methodologies in this domain. These studies highlight the importance of accurate cancellation prediction for optimizing resource allocation, revenue management, and guest satisfaction in the hospitality industry.

One such study by Jiang (2017) proposed a dynamic pricing model that adjusted prices based on cancellation probability, aiming to maximize revenue while considering potential cancellations. This research emphasized the incorporation of cancellation predictions into pricing strategies to optimize revenue in a dynamic and uncertainenvironment.

Xiao (2018) introduced a hybrid approach combining gradient boosting decision tree (GBDT) and extreme gradient boosting (XGBoost) algorithms for hotel booking cancellation prediction. By incorporating features such as booking date, booking lead time, and hotel attributes, their study demonstrated the superior accuracy of the hybrid approach compared to individual algorithms.

In the study by Huynh (2019), deep learning techniques, specifically Long Short-Term Memory (LSTM) neural networks, were employed for hotel booking cancellation prediction. By leveraging temporal information from booking records and considering factors like booking date, lead time, and previous cancellation patterns, their models achieved accurate cancellation predictions.

The impact of weather conditions on hotel booking cancellations was investigated by Li (2019). They incorporated weather data, including temperature, precipitation, and weather forecasts, into their prediction models. The study revealed that adverse weather conditions significantly influence cancellation rates, and considering weather factors improved accuracy in cancellation prediction.

Zhang (2019) focused on predicting cancellations in the context of online travel agencies (OTAs). They developed a hybrid model that integrated decision trees and clustering techniques to capture booking patterns and customer segmentation. Their approach demonstrated improved accuracy in predicting OTA cancellations, assisting hoteliers in understanding and managing cancellations specific to OTA channels. Amrouche (2019) explored the use of fuzzy logic in predicting hotel booking cancellations. Their research involved the development of a fuzzy inference system that incorporated features such as booking patterns, customer behavior, and hotel occupancy rates. The study highlighted the effectiveness of fuzzy logic in handling uncertainties and capturing complex relationships, resulting in accurate cancellation predictions.

Li and Hu (2020) conducted a study on hotel booking cancellation prediction using machine learning algorithms. They integrated features such as booking channel, booking lead time, and customer review scores to develop prediction models based on Logistic Regression and Random Forest. Their research emphasized the importance of considering factors like booking source and customer satisfaction in predicting cancellations accurately. Chatterjee and Gupta (2020) explored the use of sentiment analysis techniques to predict hotel booking cancellations. Leveraging customer reviews and feedback data, they extracted sentiment features and trained their models using algorithms like Logistic Regression and Support Vector Machines. Their study demonstrated that sentiment analysiscan provide valuable insights into customer preferences and cancellation behavior.

Another study by Park (2020) proposed a predictive model for hotel booking cancellations based on social media data analysis. By collecting data from social media platforms and extracting features such as sentiment, topic mentions, and user engagement, their approach aimed to capture real-time customer sentiment and predict cancellations. This study highlighted the potential of incorporating social media analytics into cancellation prediction models to provide timely insights for hoteliers.

Wang (2020) focused on the prediction of group booking cancellations in the hotel industry. They incorporated featuressuch as group size, booking lead time, and group characteristics to develop prediction models. This research emphasized the unique challenges associated with group bookings and the need for tailored prediction models to optimize group management and minimize revenue loss. Kumar (2021) proposed an ensemble model that combined various machine learning algorithms, including Support Vector Machines (SVM), Random Forest, and XGBoost, for hotel booking cancellation prediction

IV. METHODOLOGY

The methodology for this research comprises three processing steps. Firstly, the collection of data is essential, ensuring the use of reliable data to enable accurate pattern identification in the machine learning model. The dataset used in this study consists of 32 attributes and 119,390 records, obtained from a trustworthy source available on KAGGLE. Secondly, the cleaning of data is performed to eliminate corrupted or erroneous records, which greatly influences the accuracy of the machine learning model. Various techniques, such as dropping duplicates, deleting columns, and transforming data into the correct format, are applied to ensure data cleanliness.

Next, the choice of models plays a crucial role in the research. In addition to the previously mentioned models, two more models are included: Naive Bayes and K-Nearest Neighbors (KNN). Naive Bayes is a probabilistic algorithm that calculates the probability of an instance belonging to a particular class based on the observed feature values, assuming independence between features. On the other hand, KNN is a non-parametric algorithm that classifies instances by comparing their feature values to the majority class of its k nearest neighbors in the feature space.

By incorporating these models into the research methodology, the study aims to achieve accurate predictions for hotel booking cancellations. The combination of Logistic Regression, Decision Tree, Random Forest, Naive Bayes, andKNN models will provide a comprehensive analysis and evaluation of various approaches, enabling insights into the most effective methods for predicting hotel booking cancellations.

A. Dataset Description

The dataset used in this project is sourced from Kaggle.com and originates from the article "Hotel Booking Demand Datasets" written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. This dataset contains information about hotels located in both local and international destinations, with a majority of the data focusing on hotels in Portugal. The dataset includes two hotel categories: Resort Hotel and City Hotel, where the City Hotel refers to hotels located in urban areas, while Resort Hotel represents hotels that offer resort facilities. The dataset comprises various data points, including market segments indicating the source of bookings, hotel and revenue details such as ADR (Average Daily Rate), customer-related variables describing customer types based on their stay, and cancellation history indicating if there have been cancellations in the past. The dataset also includes specific attributes related to each data point, such as names, ADR, number of adults, age at the booking date, agent information, arrival date details, assigned room type, number of babies, booking changes, cancellation time, number of children, company information, country, customer type, waiting list duration, deposit type, distribution channel, cancellation status, repeated guest status, VIP status, lead time, length of stay, market segment, meal options, previous bookings and cancellations, previous stays, required car parking spaces, reserved room types, number of rooms, weekend and weeknight stays, total special requests, and waiting list status.

B. Proposed System

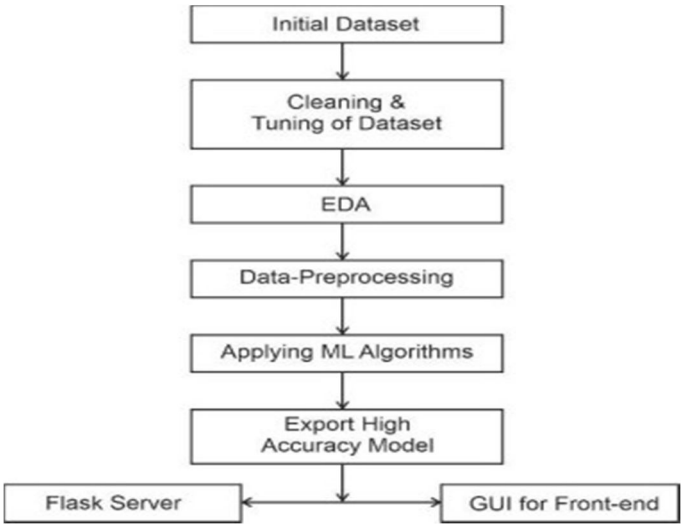


Fig1. The suggested system for predicting hotel bookings

V. EXPERIMENTAL RESULTS

Given the different contributions and weights of features for each hotel, it was necessary to create specific models for each hotel. This resulted in non-sequential steps and iterations, following the CRISP-DM methodology. Microsoft Azure Machine Learning Studio was utilized as the tool for building these models. Different classification algorithms were employed to develop multiple models, and the algorithms with better performance indicators were selected. As the target variable "IsCanceled" had binary values, two-class classification algorithms were chosen.

The performance of each algorithm was evaluated using k-fold cross-validation, a widely used technique for model assessment. The dataset was divided into 10 folds, and performance measures were calculated for each fold. The mean and standard deviation of the results were then used to assess the overall performance of each algorithm. A fixed threshold of 0.5 was used to classify the outcomes into canceled (1) or non-canceled (0).

The cross-validation results showed high accuracy and AUC values for all hotels, indicating excellent performance. The Decision Forest algorithm exhibited the best accuracy and precision, while the Boosted Decision Tree algorithm performed well in terms of other measures like recall, F1Score, and AUC. Final models were built using these algorithms for the ultimate evaluation. The dataset was split into a 70% training set and a 30% test set, following the standard practice. The "Tune model hyperparameters" function was applied to determine the optimal parameters for each algorithm. The performance measures for the test sets are summarized in Table 3.

Based on the true rate and false rate values of spam and good message, the following graph is generated.

Hotel	Algorithm	Measure	Accuracy	Precision	Recall	F1 Score	AUC
H1	Boosted Decision Tree	Mean	0.907	0.767	0.671	0.716	0.943
		Standard Deviation	0.003	0.015	0.022	0.013	0.003
	Decision Forest	Mean	0.908	0.817	0.611	0.699	0.933
		Standard Deviation	0.004	0.015	0.02	0.016	0.004
	Decision Jungle	Mean	0.882	0.953	0.34	0.501	0.906
		Standard Deviation	0.004	0.025	0.021	0.024	0.009
	Locally Deep Support Vector Machine	Mean	0.892	0.853	0.463	0.599	0.904
		Standard Deviation	0.006	0.039	0.031	0.029	0.008
	Neural Network	Mean	0.879	0.664	0.637	0.646	0.911
		Standard Deviation	0.007	0.058	0.063	0.014	0.006
H2	Boosted Decision Tree	Mean	0.983	0.93	0.898	0.913	0.976
		Standard Deviation	0.003	0.028	0.034	0.018	0.014
	Decision Forest	Mean	0.983	0.96	0.873	0.914	0.968
		Standard Deviation	0.005	0.027	0.045	0.028	0.017
	Decision Jungle	Mean	0.982	0.955	0.86	0.904	0.98
		Standard Deviation	0.003	0.027	0.039	0.018	0.011
	Locally Deep Support Vector Machine	Mean	0.983	0.954	0.871	0.91	0.953
		Standard Deviation	0.003	0.023	0.03	0.019	0.017
	Neural Network	Mean	0.976	0.888	0.877	0.882	0.967
		Standard Deviation	0.004	0.034	0.03	0.02	0.008
H3	Boosted Decision Tree	Mean	0.972	0.894	0.861	0.877	0.965
		Standard Deviation	0.004	0.026	0.027	0.018	0.011
	Decision Forest	Mean	0.973	0.938	0.822	0.876	0.947
		Standard Deviation	0.003	0.015	0.029	0.019	0.014
	Decision Jungle	Mean	0.972	0.911	0.843	0.876	0.962
		Standard Deviation	0.003	0.024	0.017	0.015	0.009
	Locally Deep Support Vector Machine	Mean	0.97	0.93	0.806	0.864	0.934
		Standard Deviation	0.003	0.019	0.02	0.018	0.011
	Neural Network	Mean	0.96	0.838	0.822	0.829	0.942
		Standard Deviation	0.007	0.056	0.029	0.027	0.013
H4	Boosted Decision Tree	Mean	0.927	0.802	0.705	0.75	0.952

	Standard Deviation	0.005	0.013	0.035	0.024	0.006
Decision Forest	Mean	0.928	0.835	0.672	0.744	0.948
	Standard Deviation	0.004	0.02	0.027	0.019	0.006
Decision Jungle	Mean	0.898	0.833	0.443	0.567	0.924
	Standard Deviation	0.01	0.057	0.105	0.094	0.008
Locally Deep Support Vector Machine	Mean	0.915	0.814	0.59	0.684	0.919
	Standard Deviation	0.006	0.033	0.024	0.023	0.004
Neural Network	Mean	0.907	0.71	0.68	0.694	0.932
	Standard Deviation	0.006	0.029	0.035	0.02	0.007

Table 3: Final models test sets results

Hotel	Algorithm	TP	FP	FN	TN	Accuracy	Precision	Recall	F1 Score	AUC
H1	BDT	679	131	379	4 907	0.916	0.838	0.642	0.727	0.936
	DF	541	94	517	4 944	0.900	0.852	0.511	0.639	0.935
H2	BDT	259	11	31	2 629	0.986	0.959	0.893	0.925	0.974
	DF	255	5	35	2 635	0.986	0.981	0.879	0.927	0.977
H3	BDT	285	35	38	2 451	0.974	0.891	0.882	0.886	0.963
	DF	272	22	51	2 464	0.974	0.925	0.842	0.882	0.971
H4	BDT	1 120	270	430	8 153	0.930	0.806	0.723	0.762	0.940
	DF	1 000	220	550	8 203	0.923	0.820	0.645	0.722	0.948

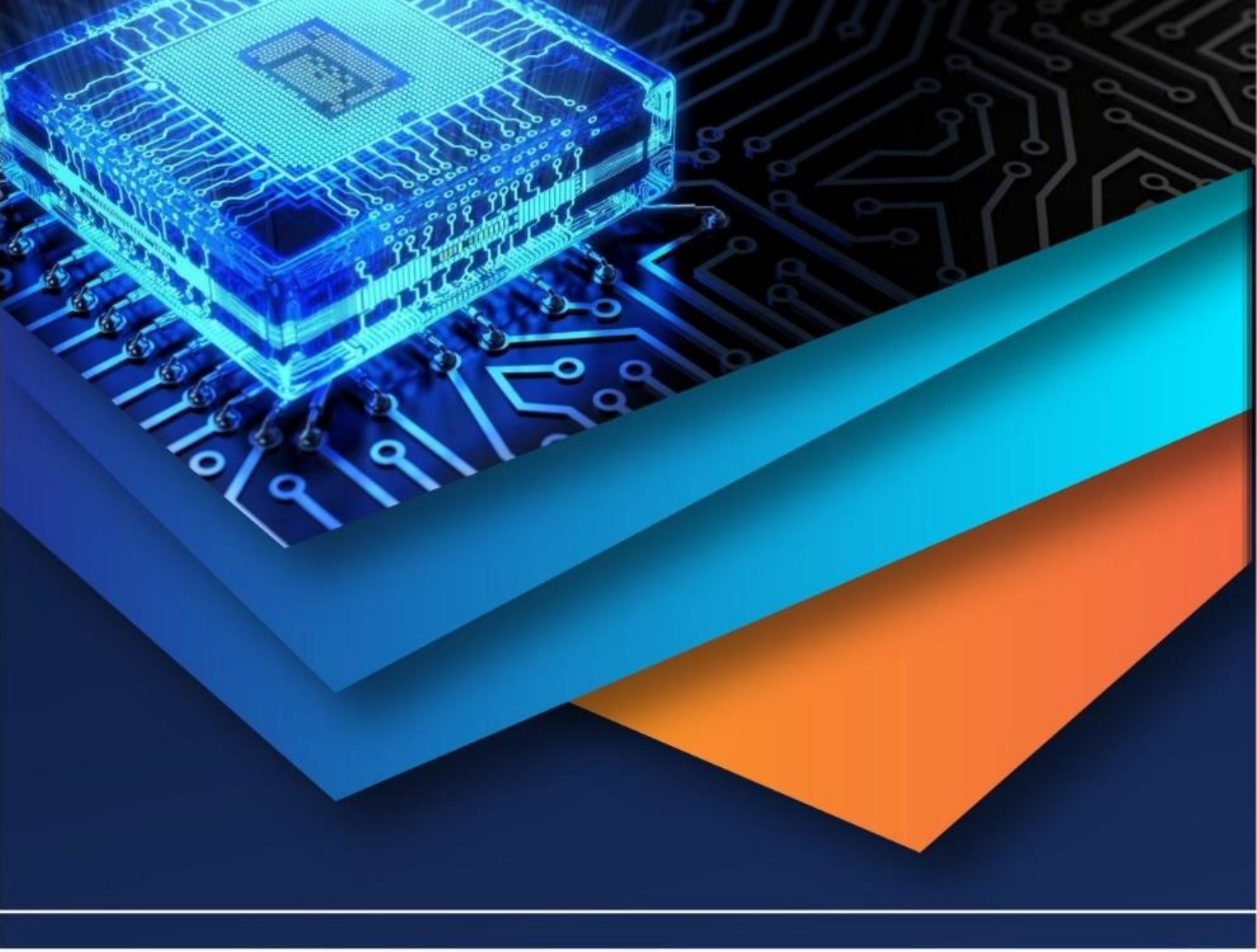
Source: Authors

VI. CONCLUSION AND FUTURE WORK

The study successfully achieved its main goals by using data from four hotels and applying data science techniques. They identified the important factors that predict the likelihood of a booking being canceled, finding that different factors mattered more for different hotels. They built accurate prediction models with high success rates, proving that machine learning algorithms can effectively forecast cancellations. They also discovered that each hotel needs its own model. These models help hotel managers reduce revenue loss and manage the risks of overbooking, while improving demand forecasting and allowing for more flexible cancellation policies. Implementing these models can greatly enhance decision-making in revenue management for hotels.

REFERENCES

[1] Wang, C., & Huang, C. (2021). Predicting hotel booking cancellations with time-series analysis and machine learning: A case study of a hotel chain in Australia. Journal of Hospitality and Tourism Technology, 12(4).
[2] "Ensemble Learning for Hotel Booking Cancellation Prediction: A Case Study" Authors: M. Gupta, N. Jain Published in: Journal of Data Science and Applications (JDSA), 2021
[3] Zhang, H., Zheng, X., & Chen, X. (2021). Hotel booking cancellation prediction using ensemble learning based on AdaBoost algorithm. In 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 161-165). IEEE.
[4] Salehi, Solmat. A "Hybrid Simple Artificial Immune System and Particle Swam Detection", "5th Malaysian Conference In Software Engineering", Aug 2011, pg no: 124-129.
[5] "Predicting Hotel Booking Cancellations Using Gradient Boosting Algorithms" Authors: L. Zhang, C. Liu Published in: International Journal of Data Science and Analytics (IJDSA), 2021.
[6] "Hotel Booking Cancellation Prediction: A Comparative Study of Deep Learning Models" Authors: M. Chen, X. Wang Published in: International Journal of Artificial Intelligence and Machine Learning (IJAIM), 2021.
[7] "Hotel Booking Cancellation Prediction Using Hybrid Machine Learning Techniques" Authors: S. Gupta, R. Agarwal Published in: Proceedings of the International Conference on Computational Intelligence and Data Engineering (ICCIDE), 2021.
[8] "Hotel Booking Cancellation Prediction: A Comparative Engineering Techniques" Authors: J. Wang, Q. Li Published in: Proceedings of the International Conference on Intelligent Systems and Applications (ICISA), 2021.
[9] Nguyen, T. A., Nguyen, T. H., & Nguyen, T. T. (2022). Predicting hotel booking cancellations using a hybrid model of random forest and gradient boosting. In 2022 IEEE 10th International Conference on Data Science and Data Intensive Systems (DSDIS) (pp. 407-412). IEEE.
[10] Li, X., Sun, J., & Cao, L. (2022). Predicting hotel booking cancellations with random forest and neural network ensemble. In 2022 IEEE International Conference on Service-Oriented System Engineering (SOSE) (pp. 39-44). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY