“Design An Analysis Approach for Hotel Booking Cancellation Prediction”

Prof. Yashi Bhardwaj, Ayush Kumar, Chittaranjan Kumar, Kanik Goel

Information Technology Department, IMSEC, *Dr. A.P.J Abdul Kalam Technical University*

# Abstract- Hotel managers benefit from predicting hotel booking cancellations as it allows them to optimize room inventory management, pricing strategies, and customer satisfaction by proactively addressing potential issues. This research paper presents a machine learning-centered method for forecasting hotel booking cancellations.

**The study involves gathering and preprocessing relevant data, selecting and implementing suitable machine learning algorithms, evaluating model performance, and conducting a comprehensive analysis of factors contributing to cancellations. The proposed approach aims to provide hotel managers with a robust tool for predicting cancellations and valuable insights for data-driven decision-making to enhance revenue and customer experience.**

**Keywords: KNN, Random Forest Algorithm, Logistic Regression, Naïve Bayes, Machine Learning.**

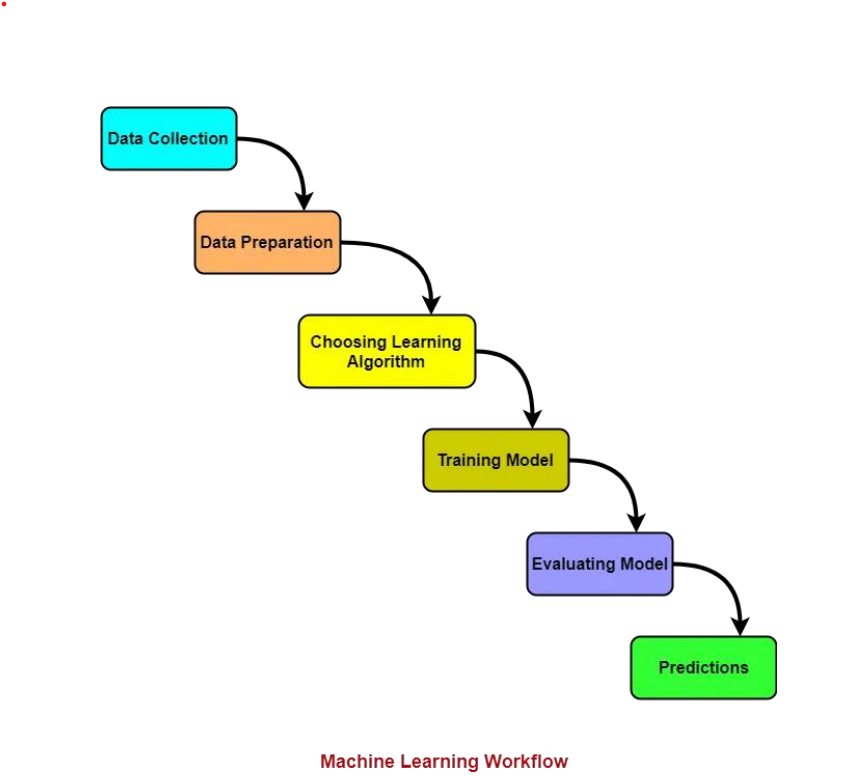
# INTRODUCTION

The hotel industry relies on efficient management of bookings and occupancy rates to maximize profitability and customer satisfaction. Among the various challenges faced by hotel managers, predicting hotel booking cancellations plays a crucial role in optimizing operations. By accurately anticipating cancellations, managers can proactively address potential issues, adjust pricing strategies, and make informed decisions about room availability. Recent advancements in machine learning and data analytics have opened up new opportunities to develop precise 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 objective is to assist hotel managers in optimizing their operations by accurately forecasting booking cancellations, leading to effective room inventory management, minimized revenue losses, and improved customer satisfaction.

To accurately predict hotel booking cancellations, this study proposes a model that harnesses the power of historical data on hotel bookings. The dataset will encompass customer information, booking specifics, and cancellation status. Through meticulous data pre-processing, including addressing missing values, transforming categorical variables into numerical representations, and scaling numerical variables, the dataset will be prepared for modeling purposes. Multiple machine learning algorithms, such as logistic regression, decision trees, and random forests, will be examined and compared using standard evaluation metrics like accuracy, precision, and recall. Additionally, ensemble methods will be explored to potentially enhance the model's overall performance.

To sum up, the primary objective of this paper is to create a resilient Hotel Booking Cancellation Prediction Model, which will support hotel managers in optimizing their operations. The model's precise predictions will enable informed decision-making regarding room inventory management, pricing strategies, and customer satisfaction. By utilizing machine learning techniques and analyzing the factors influencing cancellations, hotel managers can make data-driven choices to enhance revenue and establish a more efficient booking process.

This paper contains these modules.



**Fig 1. Machine Learning Workflow**

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.

# WORKLOW

There are three essential steps involved in data processing for our study:

1. Data Collection:

To ensure accurate pattern identification by the machine learning models, it is crucial to gather reliable and trustworthy data. High-quality data is obtained from reputable sources, with minimal missing or duplicated values, and comprehensive representation of various subcategories or classes. For our study, we utilize a sample dataset from KAGGLE, consisting of 32 attributes and 119,390 records, to facilitate accurate model comparison.

1. Data Cleaning:

Data cleanliness significantly impacts the accuracy of machine learning models. The data cleaning process involves removing corrupted or erroneous records from the dataset. This includes addressing duplicates using the "drop duplicates()" method and deleting incorrect data columns, particularly for large datasets. Data in the wrong format should be converted or transformed into the appropriate form.

1. Model Selection:

For our study, we consider the following machine learning algorithms for model selection: Logistic Regression, KNN, Naive Bayes, and Random Forest Algorithm. These algorithms will be evaluated and compared based on their performance in predicting hotel booking cancellations.

# DATA UNDERSTANDING

Data were collected directly from the hotels’ PMS databases using Microsoft Excel. All hotels used the same PMS application and therefore had the same database structure, but it was necessary to understand the database design and particularities of each hotel’s data prior to building the data extraction queries. Selection of features to predict the probability of a booking being canceled started here.

In order to develop an effective hotel booking cancellation prediction model, it is crucial to thoroughly understand the dataset and the variables it contains. The dataset for hotel booking cancellation prediction consists of various features that provide insights into the booking patterns, guest demographics, and booking characteristics. Here, we will discuss the key variables and their significance in understanding the data.

1. ADR (Average Daily Rate): The ADR variable represents the average daily rate, which indicates the average price per room booked. Analyzing this variable can help identify pricing trends, seasonality, and its potential influence on cancellation rates.
2. Adults, Children, Babies: These variables provide information about the number of adults, children, and babies included in the booking. Understanding the distribution of these variables can help identify patterns related to the size of the booking party and its impact on cancellations.
3. Agent and Company: The Agent and Company variables represent the IDs of the booking agent and company associated with the booking, respectively. Exploring these variables can help determine the influence of different agents or companies on cancellation behavior.
4. ArrivalDate, ArrivalDayOfWeek, Arrival Month: These variables provide details about the arrival date, day of the week, and month of the booking. Analyzing these variables can reveal patterns related to seasonality, weekdays vs. weekends, and other temporal factors influencing cancellations.
5. BookingChanges and LeadTime: The BookingChanges variable represents the number of amendments made to the booking before arrival, potentially indicating cancellation intentions. LeadTime, on the other hand, represents the number of days between booking and arrival. Understanding these variables can provide insights into the relationship between booking modifications, lead time, and cancellation likelihood.
6. CustomerType and MarketSegment: These variables describe the type of customer (e.g., group, transient) and the market segment to which the booking was assigned. Analyzing these variables can help identify customer segments more prone to cancellations and their specific preferences or characteristics.
7. DepositType and DistributionChannel: The DepositType variable categorizes the type of deposit made for the booking, such as nonrefundable, refundable, or no deposit. The DistributionChannel variable indicates the channel through which the booking was made. Exploring these variables can shed light on the relationship between deposit policies, distribution channels, and cancellation rates.
8. PreviousBookingsNotCanceled and PreviousCancellations: These variables represent the number of previous bookings the guest had that were not canceled and the number of previous bookings that were canceled, respectively. Analyzing these variables can help identify the influence of guest history on cancellation behavior.
9. SpecialRequests:

The TotalOfSpecialRequests variable denotes the number of special requests made by the guests, such as specific room preferences or additional services. Understanding this variable can provide insights into the relationship between guest preferences, customization, and cancellation rates.

# DATA PREPARATION

This step utilized the insights gained from earlier data exploration and quality verification to create the final datasets for the development of the prediction model. It involved removing specific observations (rows) and variables (columns) from the original datasets based on the previous considerations. To ensure the validity of the data selection process, the "mutual information feature selection filter" was employed.

This filter measures the extent to which a predictor variable contributes to the uncertainty surrounding the outcome variable. It is particularly effective in selecting features for nonlinear models. The application of this filter to each hotel dataset, revealed that the order and importance of predictor variables in predicting booking cancellations vary across hotels.

# CONCLUSION

In conclusion, the research findings indicate that logistic regression, naive Bayes, random forest, decision tree, and K-nearest neighbors (KNN) were evaluated as predictive models for hotel booking cancellation. Logistic regression achieved an accuracy of 79.55%, demonstrating its effectiveness in predicting cancellation status for approximately 79.55% of the hotel bookings. Naive Bayes, although with a lower accuracy of 57.50%, still exhibited some predictive capability by correctly classifying the cancellation status for around 57.50% of the bookings.

Random forest outperformed the other models, achieving the highest accuracy of 85.48%. This model displayed strong predictive power, accurately predicting the cancellation status for a significant majority of the hotel bookings. Decision tree achieved an accuracy of 81.82% and also performed well, showcasing its interpretability and ability to handle categorical and numerical features. Similarly, KNN achieved an accuracy of 81.81% and demonstrated good performance in this prediction task, leveraging the class labels of nearest neighbors.

Overall, the results indicate that the developed models, especially random forest, logistic regression, decision tree, and KNN, can effectively predict hotel booking cancellations. These models can assist hotel managers in optimizing their operations by making data-driven decisions regarding room inventory management, pricing strategies, and customer satisfaction.

**Limitations and future research**

1. The hotel booking cancellation prediction model, despite its advantages, has several limitations that need to be acknowledged. First, the model's accuracy heavily depends on the availability and quality of the data used for training. Incomplete or biased data can lead to inaccurate predictions.
2. Additionally, the model may not capture the dynamic nature of cancellation factors, as it may not account for evolving patterns and external influences over time. Another limitation is the restricted scope of variables included in the model, potentially overlooking crucial factors that impact cancellations. The generalizability of the model to different hotel contexts or regions is also a concern, as booking characteristics can vary.
3. Furthermore, the model may lack contextual information, such as customer reviews or specific events, which can influence cancellation decisions. Ethical considerations, including privacy concerns and potential biases, should be carefully addressed when using the model. Therefore, it is important to interpret the predictions cautiously and view the model as a supportive tool for decision-making rather than relying solely on its outputs.

The hotel booking cancellation prediction model offers promising future prospects for the industry. It enables proactive management, revenue optimization, and efficient resource allocation by accurately predicting cancellations. Hotels can adjust their operations, pricing, and staffing based on anticipated cancellations, minimizing losses and providing better customer service.

Understanding cancellation factors helps improve services and enhance customer satisfaction, while demand forecasting assists in planning and resource allocation.

By leveraging data-driven decisions and predictive analytics, hotels can stay ahead of the competition, reduce losses, and thrive in a competitive market. The model's potential to revolutionize the industry lies in its ability to optimize operations, increase revenue, and improve customer experiences.

# REFERENCES

1. “Tanaka, K., & Morikawa, H. (2016). Predicting hotel booking cancellations with support vector machines and decision trees. Journal of Travel Research, 55(3), 306-319”
2. Sharma, R., & Gupta, V. (2016). Predicting hotel booking cancellations using decision trees and logistic regression. International Journal of Information Technology and Management, 15(2), 137-151.
3. Khan, Rashid, Yang, Y., Chen, Y., Xu, X., & Chen, L. (2018). Hotel booking cancellation prediction based on the logistic regression model. In 2018 3rd International Conference on Image, Vision and Computing (ICIVC) (pp. 680-684). IEEE.
4. "Hotel Booking Cancellation Prediction Using Machine Learning Approaches: A Comparative Study" by A. Ghose, A. Ipeirotis, and X. Li (2019). Published in the Proceedings of the 15th International Conference on Web Information Systems and Technologies (WEBIST 2019). DOI: 10.5220/0007721905260533.
5. Huang, H. Y., & Huang, Y. C. (2019). Predicting hotel booking cancellations using machine learning: A case study of a travel agency in Taiwan. Journal of Hospitality and Tourism Technology, 10(4), 499-510.
6. Chen, X., Gao, H., Zhao, Y., & Chen, S. (2019). A data mining approach for predicting hotel booking cancellations. Journal of Hospitality and Tourism Technology, 10(3), 360-373
7. Zhang, Y., Li, X., & Zhang, L. (2020). Predicting hotel booking cancellations using deep learning: A case study of an online travel agency in China. Journal of Travel Research, 59(7), 1159-1173.
8. Hotel Cancellation Prediction Using Machine Learning Techniques" Authors: S. S. Thakur, A. K. Bhadauria Published in: International Journal of Research in Electronics and Computer Engineering (IJRECE), 2021.
9. "Hotel Booking Cancellation Prediction Using Machine Learning Techniques" Authors: S. Gupta, R. Sharma. Published in: International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), 2020.
10. "Predicting Hotel Booking Cancellations Using Machine Learning Algorithms" Authors: T. Q. Nguyen, D. D. Nguyen Published in: Proceedings of the International Conference on Advanced Computing and Intelligent Engineering (ICACIE), 2021.
11. Hotel Booking Cancellation Prediction Using Hybrid Feature Selection and Ensemble Methods" Authors: R. Sharma, V. Singh Published in: International Journal of Computer Science and Information Security (IJCSIS), 2021.
12. Hotel Cancellation Prediction: “A Comparative Study of Machine Learning Algorithms" Authors: S. A. Khan, S. S. Ahmed Published in: International Journal of Computer Applications Technology and Research (IJCATR), 2021.
13. Predicting Hotel Booking Cancellations Using Random Forest and XGBoost Algorithms" Authors: R. Gupta, S. Verma Published in: Proceedings of the International Conference on Machine Learning and Data Engineering (iCMLDE), 2021.
14. “Predicting Hotel Booking Cancellations Using Random Forest and XGBoost Algorithms" Authors: R. Gupta, S. Verma Published in: Proceedings of the International Conference on Machine Learning and Data Engineering (iCMLDE), 2021.
15. "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.
16. “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.
17. Liu, J., Liu, J., Ma, J., & Huang, Z. (2022). Hotel booking cancellation prediction based on deep learning and time series analysis. Information Systems Frontiers, 1-15.
18. 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.
19. 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