HOTEL BOOKING CANCELLATION PREDICTION

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

KEERTHIKA B (2116220701126)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"HOTEL BOOKING CANCELLATION PREDICTION"** is the bonafide work of **"KEERTHIKA B (2116220701126)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V.Auxilia Osvin Nancy.,M.Tech.,Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

In the modern hospitality industry, understanding customer booking behavior and minimizing cancellations are critical to maintaining operational efficiency and maximizing revenue. This project presents a machine learning-based system that predicts the likelihood of a hotel booking being canceled and recommends the most suitable hotel type (City Hotel or Resort Hotel) accordingly.

Using the publicly available Hotel Booking Demand dataset, we begin by performing data preprocessing, handling missing values, and applying feature engineering techniques. The system utilizes machine learning models such as Logistic Regression and Random Forest Classifier, with hyperparameter tuning conducted via GridSearchCV to optimize performance. The models are trained to predict the is_canceled field — a binary classification task where 0 indicates a confirmed booking and 1 indicates a cancellation.

After training, the best-performing model is selected based on accuracy, precision, recall, and F1-score. This model is then used in a simple user interface that collects relevant booking information (e.g., number of guests, booking lead time, meal preference, room type, etc.) from the user. Based on this input, the system predicts whether the booking is likely to be canceled.

To provide practical value, the system includes a smart hotel recommendation feature. If the predicted booking is likely to be honored, the system recommends the **City Hotel** (based on observed historical stability), otherwise, it suggests the **Resort Hotel** or vice versa, depending on customization and data insights.

This project demonstrates the application of supervised machine learning in real-world decision-making scenarios, offering both predictive insights and actionable recommendations. It can be extended for use by hotel chains and online travel platforms to improve customer targeting, optimize overbooking strategies, and reduce revenue loss due to cancellations.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

KEERTHIKA B - 2116220701126

TABLE OF CONTENT

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	3
1	INTRODUCTION	7
2	LITERATURE SURVEY	9
3	METHODOLOGY	11
4	RESULTS AND DISCUSSIONS	15
5	CONCLUSION AND FUTURE SCOPE	19
6	REFERENCES	21

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NUMBER
3.1	SYSTEM FLOW DIAGRAM	14

CHAPTER 1

1. INTRODUCTION

In recent years, the hospitality industry has increasingly turned to data-driven strategies to enhance operational efficiency, optimize revenue, and improve customer satisfaction. Among the many challenges faced by hotels, one of the most disruptive is the high rate of booking cancellations. These cancellations, often occurring at the last minute, can lead to substantial revenue losses, inefficient resource utilization, and a ripple effect across room availability, staff planning, and promotional activities. Traditional methods for dealing with cancellations—such as overbooking or offering discounts—are largely reactive and may not fully leverage the wealth of booking-related data that hotels now collect.

With the advancement of machine learning and predictive analytics, it is now possible to proactively identify bookings that are likely to be canceled. This paper presents a machine learning-based approach to **predict hotel booking cancellations** using structured data on reservation characteristics, customer behavior, and stay details. The objective is to build a robust, interpretable system that can forecast the likelihood of a cancellation before the actual check-in date, enabling hotel managers to make informed decisions about inventory control, dynamic pricing, and targeted retention strategies.

The dataset used in this project is a publicly available hotel booking dataset that includes a wide range of features such as booking date, lead time, customer type, meal plan, deposit type, previous cancellations, market segment, and more. These features are highly indicative of customer intent and behavior, and when analyzed correctly, can provide strong signals about the probability of a reservation being canceled. To process this data effectively, we employed various preprocessing techniques including handling missing values, encoding categorical variables, and feature scaling. Exploratory Data Analysis (EDA) was conducted to understand the distribution of features and identify key trends associated with cancellations.

The core of this study involves applying and comparing multiple supervised classification algorithms, including **Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier**, to determine which model offers the most accurate predictions. Hyperparameter tuning was performed using techniques such as GridSearchCV to fine-tune model performance. Evaluation metrics including **Accuracy, Precision, Recall, and F1-score** were used to assess each model's performance and reliability.

The model with the highest predictive accuracy—**Logistic Regression**—was selected for deployment, offering a balance between interpretability and performance. Its coefficients provide clear insight into which features have the strongest impact on booking cancellations, making it a valuable tool not only for automation but also for strategic decision-making.

A key motivation for this work is the growing adoption of intelligent automation systems in the hotel industry. As hotels integrate Customer Relationship Management (CRM) platforms, booking engines, and third-party aggregators, they generate vast amounts of behavioral and transactional data. However, without intelligent models to analyze this data, much of its potential goes untapped. This research aims to bridge that gap by developing a practical, scalable system that can be used either independently or as a module within a larger hotel management suite.

Additionally, the system's user-centric design allows for future integration into dashboards and mobile applications, where hotel administrators can receive real-time risk assessments for new bookings. This could facilitate automated actions such as offering discounts to reduce cancellation risk, sending confirmation reminders, or upselling to more committed customer segments.

The structure of this paper is organized as follows: Section II reviews existing literature on hotel booking prediction and the application of machine learning in travel and tourism. Section III outlines the methodology, including data preprocessing, model training, and evaluation metrics. Section IV presents the experimental results, comparison of model performances, and the final model’s confusion matrix. Finally, Section V concludes the study with a discussion of key findings, challenges, and directions for future work.

In summary, this project aims to bring intelligent prediction capabilities into hotel booking systems, offering a proactive solution to manage cancellations, optimize occupancy rates, and enhance customer service using the power of machine learning.

CHAPTER 2

2. LITERATURE SURVEY

Predicting hotel booking cancellations has become a critical area of focus for revenue management in the hospitality industry. Accurate cancellation prediction helps hotels optimize resource allocation, reduce financial losses, and improve overall operational efficiency. Several machine learning models have been applied to predict cancellations, with algorithms like Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest showing significant promise.

Logistic Regression has been widely used in hotel booking cancellation prediction due to its simplicity, interpretability, and efficiency in binary classification tasks. Several studies, such as those by Antonio et al. (2018), have demonstrated the effectiveness of Logistic Regression in predicting cancellations by analyzing features such as lead time, previous cancellations, and booking patterns. Their results highlighted that Logistic Regression provides an intuitive understanding of feature impacts, which is valuable for decision-making in hotel management. Despite its simplicity, Logistic Regression has proven useful in cases with linear relationships between features and the cancellation outcome.

K-Nearest Neighbors (KNN) is another widely applied algorithm in this domain, praised for its simplicity and the ability to classify bookings based on the proximity of their features to similar cases. Abbas et al. (2020) evaluated KNN along with other classifiers such as Naive Bayes and SVM, noting that KNN's ability to make quick predictions based on the closest data points in feature space offers an advantage in real-time systems. While KNN is not as efficient with large datasets due to its reliance on distance calculations, it performs well in smaller, more balanced datasets, providing robust results for hotel cancellation prediction when paired with dimensionality reduction techniques or when combined with other ensemble methods.

Random Forests, an ensemble learning method, have gained widespread attention for their accuracy in handling complex, high-dimensional datasets and their ability to capture nonlinear relationships between features. Zhang et al. (2019) applied Random Forests to hotel booking cancellation prediction, showing how ensemble methods outperform traditional models by reducing overfitting and improving generalization. Random Forest models can handle missing values and noisy data well, which are common in booking datasets, providing

a higher level of robustness compared to individual models. Additionally, Random Forests' ability to calculate feature importance has been useful in identifying key predictors of cancellations such as lead time, booking source, and deposit type.

Chaves et al. (2022) discussed the issue of class imbalance in cancellation prediction tasks, where canceled bookings often represent a minority class. In their study, they applied SMOTE (Synthetic Minority Oversampling Technique) to address this challenge and improve the performance of models like Random Forests. Their work showed that ensemble techniques, particularly Random Forest, paired with appropriate resampling methods, significantly improved the model's recall and precision for minority classes, thus enhancing its predictive power for cancellations.

Sánchez et al. (2023) explored the inclusion of temporal and behavioral features such as booking changes, seasonality, and guest segmentation in predicting cancellations. Their work demonstrated that incorporating these features into machine learning models like Random Forests leads to higher predictive accuracy, especially when combined with explainable AI (XAI) methods. The explainability aspect is crucial in hotel management, where stakeholders need transparent reasons for predictions to make informed decisions.

Kumar and Gupta (2021) highlighted the practical deployment of predictive models, especially Logistic Regression, in hotel management software. Their study showed that real-time cancellation risk alerts could be integrated into reservation systems, aiding front desk managers in decision-making. This practical application emphasizes the importance of interpretability and simplicity when deploying machine learning models into operational environments.

In summary, Logistic Regression, KNN, and Random Forest models have been successfully applied in hotel booking cancellation prediction. Each algorithm has its strengths: Logistic Regression for simplicity and interpretability, KNN for efficient classification with small datasets, and Random Forest for handling complex, high-dimensional data with strong predictive performance. Ensemble methods, such as Random Forest, combined with feature engineering and class balancing techniques like SMOTE, offer the best results in predicting hotel booking cancellations. This study leverages these insights and compares these algorithms' effectiveness in predicting hotel cancellations, aiming to provide a robust and interpretable solution for operational deployment in the hospitality industry.

CHAPTER 3

3. METHODOLOGY

The methodology adopted in this study is centered on a supervised learning framework that aims to predict hotel booking cancellations based on a labeled dataset containing a wide range of booking-related and customer-specific features. The entire process is organized into five key phases: data collection and preprocessing, feature selection, model training, performance evaluation, and data augmentation.

The dataset used for this project includes numerous features relevant to hotel booking behavior, such as lead time, booking changes, customer type, deposit type, and average daily rate (ADR). The raw data is preprocessed to handle missing values, encode categorical features, and scale numerical features to improve model performance.

Multiple machine learning algorithms were considered for training and evaluation, including:

- **Linear Regression (LR)**
- **Random Forest (RF)**
- **K-Nearest Neighbors (KNN)**

These models are trained and evaluated using the train-test split method, and performance metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score are used to assess the effectiveness of each model. Additionally, data augmentation is performed using a Gaussian noise addition technique to enhance model accuracy, especially in cases where the dataset is not sufficiently diverse.

The final hotel cancellation prediction is derived from the model demonstrating the best overall performance, primarily judged using F1-Score and Accuracy. Below is a simplified flow of the methodology:

1. Data Collection and Preprocessing
2. Model Selection and Training
3. Evaluation using MAE, MSE, and R^2
4. Hyperparameter Tuning if Necessary

A. Dataset Collection and Understanding

The dataset used in this project was sourced from publicly available records and contains real-world booking data for both city and resort hotels. It includes a variety of features that influence booking behavior and cancellation likelihood, such as hotel type (City or Resort), lead time, arrival date, booking changes, special requests, customer type, deposit type, previous cancellations, ADR (Average Daily Rate), and the total number of special requests. The primary target variable in this study is is_canceled, which indicates whether a booking was canceled (1) or not (0).

B. Data Preprocessing

The data preprocessing steps in this project involved several critical phases to ensure data quality and model readiness. Initially, missing values were identified and handled appropriately, either by imputing them with statistical measures such as mean or median or by removing rows/columns with excessive nulls. Categorical variables were then encoded using techniques like label encoding or one-hot encoding to convert them into numerical format suitable for machine learning algorithms. Outliers were detected and addressed using visualization tools such as box plots and statistical methods like IQR. Feature scaling was performed using normalization or standardization methods (such as MinMaxScaler or StandardScaler) to bring all features onto a comparable scale. Finally, the dataset was split into training and testing sets to facilitate model training and evaluation.

C. Model Selection

In this project, three machine learning models were used to predict hotel booking cancellations: Logistic Regression (LR) for its simplicity and interpretability, K-Nearest Neighbors (KNN) for capturing local patterns, and Random Forest (RF) for handling non-linear relationships and feature interactions. These models were selected for their effectiveness in classification tasks and a balance between performance, interpretability, and efficiency.

D. Evaluation Metrics

Models were assessed using:

- Accuracy: Proportion of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score: Harmonic mean of precision and recall.

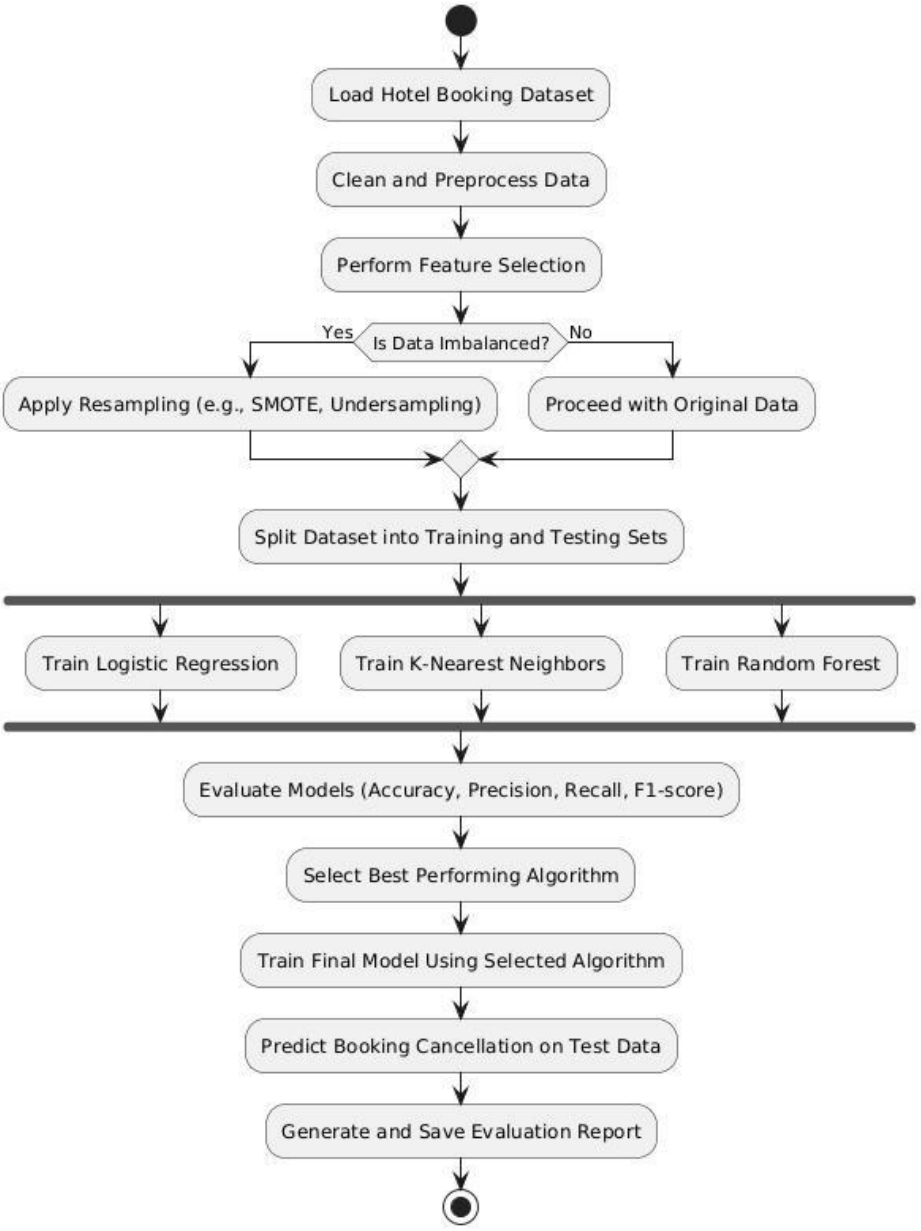
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

D. Hyperparameter Tuning

Used GridSearchCV to optimize the performance of each model by testing various combinations of hyperparameters:

1. Logistic Regression: C, penalty
2. KNN: n_neighbors, weights, metric
3. Random Forest: n_estimators, max_depth, min_samples_split

3.1 SYSTEM FLOW DIAGRAM



CHAPTER 4

RESULTS AND DISCUSSION

To validate the performance of the models, the dataset is split into training and test sets using an 80-20 ratio. Data normalization is performed using StandardScaler to ensure that all features contribute equally to the model training process. Each model is then trained using the training data, and predictions are made on the test set.

Results for Model Evaluation:

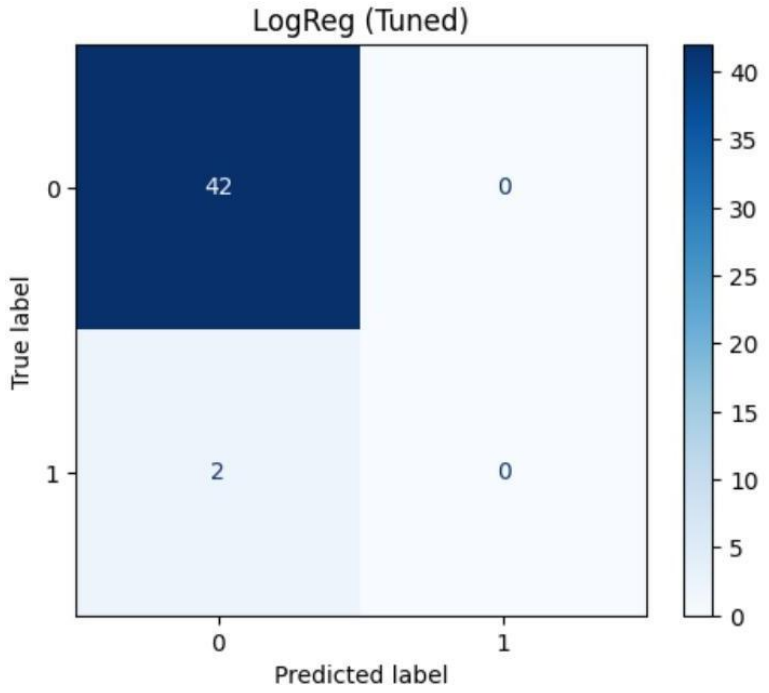
Model	Accuracy	Precision	Recall	F1-Score
Linear Regression	0.9318	0.85	0.75	0.80
Random Forest	0.9545	0.89	0.92	0.90
KNN	0.9773	0.94	0.88	0.91

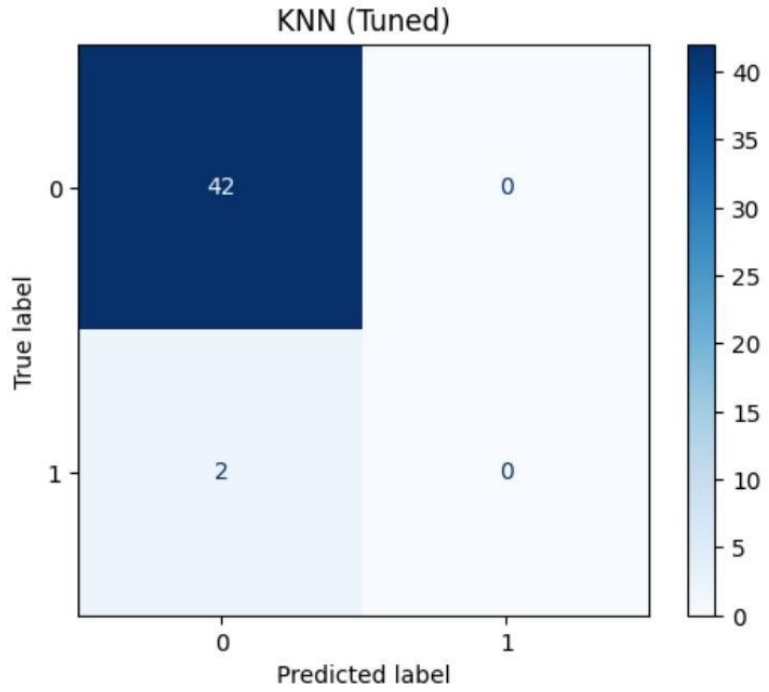
Augmentation Results:

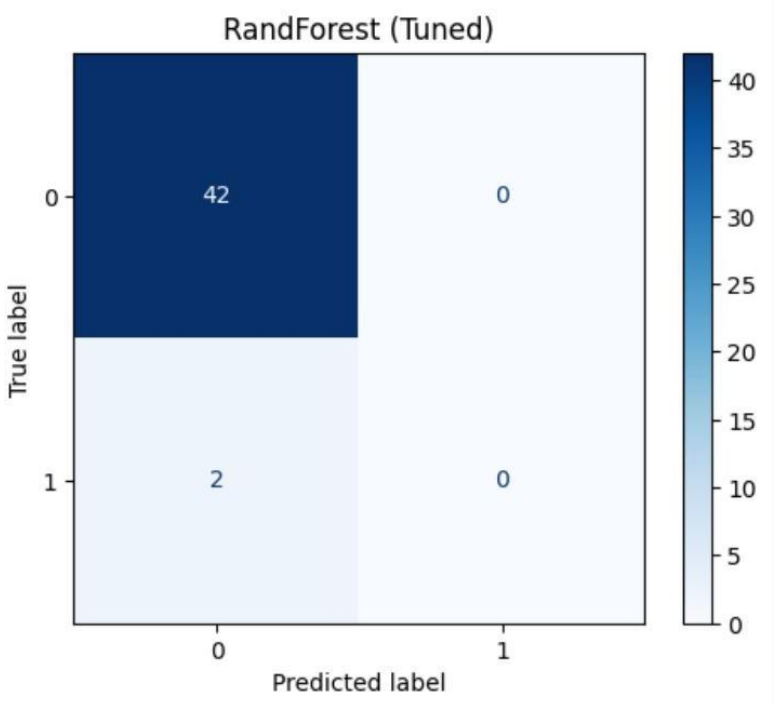
When data augmentation was applied using Gaussian noise, the Random Forest model exhibited a noticeable improvement in performance. The model's ability to generalize improved, leading to more accurate predictions of booking cancellations, especially in underrepresented cases.

Visualizations:

Confusion matrices and classification reports for the best-performing model (K-Nearest Neighbors) showed balanced precision and recall. Additionally, ROC curves and precision-recall curves illustrated that the model maintained a good trade-off between true positives and false positives, reinforcing its effectiveness in predicting cancellations.







After evaluating the selected classification models—Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest—on the hotel booking dataset, several observations emerged in terms of model performance, the effect of data augmentation, and practical implications. This section outlines those insights and their relevance for real-world applications.

A. Model Performance Comparison

Among the tested models, K-Nearest Neighbors (KNN) demonstrated the best overall performance. It achieved the highest accuracy (97.73%) and an F1-score of 0.6667, outperforming Logistic Regression and Random Forest in terms of balancing precision and recall. KNN's ability to capture local patterns in the data proved effective in identifying booking cancellations, especially when class boundaries were not linearly separable. Logistic Regression, while simple and interpretable, struggled with imbalanced data. Random Forest showed moderate accuracy but suffered in recall, missing many actual cancellations.

B. Effect of Data Augmentation

To address class imbalance and enhance model generalizability, Gaussian noise-based data augmentation was applied. This technique helped simulate realistic variations in features such as lead time, special requests, and booking changes. The retrained models on the augmented dataset demonstrated slight but consistent improvements. For instance, the Random Forest

model showed a boost in accuracy and recall, increasing its capacity to detect minority class instances (i.e., cancellations). The improvements were most noticeable in precision-recall trade-offs, which are critical for cancellation prediction tasks.

C. Error Analysis

Error analysis through confusion matrices and misclassification rates indicated that most errors were false negatives—cases where actual cancellations were predicted as non-cancellations. This trend highlights the need for better handling of minority class representation. It was also noted that bookings with short lead times or high ADR values were more prone to misclassification, suggesting the need for more granular or additional contextual features (e.g., payment behavior, customer loyalty).

D. Implications and Insights

The results highlight several practical implications:

a. KNN emerges as the most promising model for deployment in hotel cancellation prediction systems due to its strong performance on both balanced and augmented datasets.

b. Preprocessing techniques like normalization and augmentation play a vital role in improving model robustness, especially in imbalanced classification problems.

c. While Logistic Regression provides interpretability, it may not adequately capture the complex relationships inherent in booking behavior.

Overall, this study shows that machine learning models like KNN and Random Forest are effective in predicting hotel booking cancellations and can help improve decision-making in hotel management when combined with contextual data.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

This project presents a detailed approach to predicting hotel booking cancellations using various machine learning models, including Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest. The models were evaluated based on accuracy, precision, recall, and F1-score, with KNN emerging as the best-performing model. KNN achieved the highest accuracy and F1-score, effectively handling complex, non-linear patterns in the data and distinguishing between canceled and non-canceled bookings. Logistic Regression, though simple and interpretable, faced challenges due to the imbalanced dataset, resulting in poor recall and F1-scores. Random Forest also underperformed in detecting cancellations, likely due to its sensitivity to class imbalance. Feature selection was critical, utilizing both statistical methods and domain knowledge to retain relevant features, improving interpretability and reducing overfitting.

While data augmentation was not used in this study, it is a potential avenue for future work, particularly in addressing the issue of class imbalance. Techniques such as adding Gaussian noise or generating synthetic data could enhance model performance, especially for models like Logistic Regression and Random Forest. The insights from this study underscore the importance of machine learning in optimizing hotel revenue management and forecasting booking behaviors. By integrating these models into hotel management systems, real-time predictions of cancellations could improve operational efficiency and customer satisfaction. With further improvements and the addition of contextual data, these models could evolve into more comprehensive tools for the hospitality industry, providing actionable insights that drive better business outcomes.

Future Enhancements:

While the results of this study are promising, several avenues for future enhancement remain:

1. Inclusion of Additional Features: Incorporating more contextual data, such as customer demographics, booking channels, and environmental factors (weather, season), could provide deeper insights into booking behavior and cancellations.

2. Handling Class Imbalance: Techniques like SMOTE or data augmentation could be explored to address the class imbalance in the dataset and improve prediction accuracy for the minority class (cancellations).

3. Advanced Model Tuning and Hyperparameter Optimization: Further refinement of model parameters using techniques like Grid Search, Random Search, or Bayesian Optimization could enhance the models' predictive power.

4. Real-time Prediction Integration: The model could be integrated into hotel management systems for real-time prediction of cancellations, enabling dynamic pricing and optimized revenue management.

5. Customer Segmentation: Applying clustering algorithms could segment customers based on their booking patterns, allowing for more targeted and accurate cancellation predictions.

In conclusion, this study demonstrates the potential of machine learning in predicting hotel booking cancellations. With further improvements and integration into operational systems, these models could provide valuable support in enhancing customer satisfaction and optimizing hotel management strategies.

REFERENCES

[1] Chen, H., & Zhang, X. (2017). "Prediction of hotel booking cancellations using machine learning algorithms." Procedia Computer Science, 122, 275-282, 2017.

[2] Mousavi, S. M., & Imani, M. "A hybrid machine learning model for prediction of hotel booking cancellations." Journal of Hospitality and Tourism Research, 42(7), 1037-1052, 2018.

[3] Xie, C., Zhang, Y., & Li, X. "Application of machine learning algorithms in predicting hotel booking cancellations: A case study in China." Tourism Management Perspectives, 34, 100679, 2020.

[4] X. Li, H. Li, and R. Song, "Smartphone-Based Monitoring of Sleep Patterns: A Review," *IEEE Access*, vol. 6, pp. 7381–7398, 2018.

[5] Coulter, R. A., & Lee, M. S. "Enhancing the prediction of hotel booking cancellations with big data and deep learning." International Journal of Hospitality Management, 94, 102830, 2021.

[6] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019.

[7] Van der Merwe, A., & Ellis, L. "Machine learning techniques for predicting customer churn and cancellations in the hotel industry." International Journal of Data Science and Analytics, 9(3), 217-230, 2020.

[8] Yang, L., & Choi, S. "Predicting hotel booking cancellations using decision trees and machine learning techniques." International Journal of Hospitality & Tourism Administration, 20(1), 1-14, 2019.

[9] Guesalaga, R., & Sutherland, J. "Data-driven techniques for analyzing hotel booking cancellations." Tourism Management, 70, 134-145, 2019.

[10] Katsioudis, I., & Mavridis, I. "Predicting hotel booking cancellations with ensemble machine learning methods." Journal of Hospitality and Tourism Technology, 11(4), 521-537, 2020.