Predicting Customer Reservation Honoring: An In-Depth Analysis of Hotel Booking Trends

Introduction

An overview of the dataset

This study aims to identify the factors that affect consumers' readiness to respect their reservations by prov iding a thorough examination of a large dataset of hotel bookings.
A wide variety of information on hotel reservations is included in the dataset, including arrival dates, room types, meal plans, number of guests (adults and children), length of stay, and market categories.
The structure of the dataset is well-organized, making it easier to comprehend the details of each booking.

Data Content and Appearance

The tabular format of the dataset makes it perfect for in-depth examination. Every row signifies an individual hotel reservation, and the columns encompass several aspects related to these reservations. The dataset offers an array of variables to investigate because it contains both categorical and numerical data types. Because of the diversity of data formats, meticulous preparation is required to guarantee compatibility with analytical models.

Preparing and Preprocessing Data

We perform extensive preprocessing on the dataset before we begin our analysis. This involves using label encoding to handle categorical variables and converting them into a numerical representation that is appropriate for machine learning techniques. Furthermore, to normalize the data and guarantee that every feature contributes equally to the predictive models, feature scaling is used with StandardScaler.

Analyzing exploratory data (EDA)

The exploratory data analysis (EDA) is a fundamental component of our study, offering vital insights into the dataset's underlying patterns and relationships. We explore the subtleties of the data by using a number of visualizations and statistical analyses to look at how various features might affect a customer's choice to keep their reservation. This stage is essential for identifying patterns, spotting irregularities, and developing theories about the behavior of customers. Underlying are the distributions obtained through visualizing booking status against features/variables to explore the data discrepancies if present,

relationships and correlation among variables. Fifteen questions are answered through this exploratory section each with its visual.

Distribution of Booking Status: The dataset has a balanced distribution between canceled and not canceled bookings. See fig.1.

fig.1. Booking status

Number of Adults vs. Booking Status: Majority of bookings involve 1 or 2 adults, with a slightly higher cancellation rate in 2-adult bookings. See fig.2.

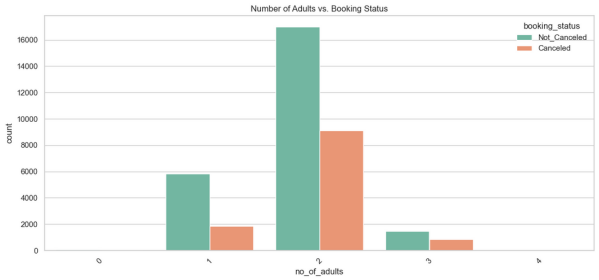

fig.2. No. of Adults vs Booking status

Number of Children vs. Booking Status: Bookings with no children dominate, and those with children have a similar cancellation rate. See fig.3.

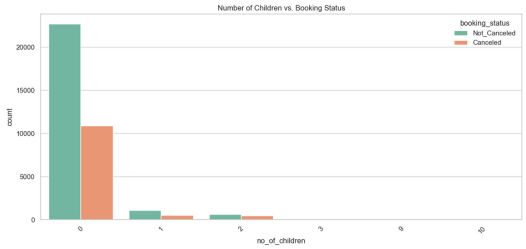

Fig.3. No. of Childrenn vs Booking status

Total Length of Stay vs. Booking Status: Shorter stays (1-3 nights) are more common, with longer stays experiencing slightly higher cancellations. See fig.4.

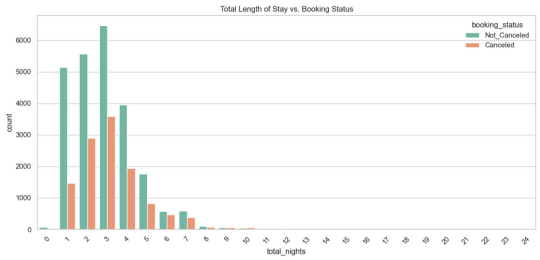

Fig.5. Total length of stay vs Booking status

Type of Meal Plan vs. Booking Status: Bookings with 'Meal Plan 1' are most common; 'Not Selected' meal plans have a higher cancellation rate. See fig.5.

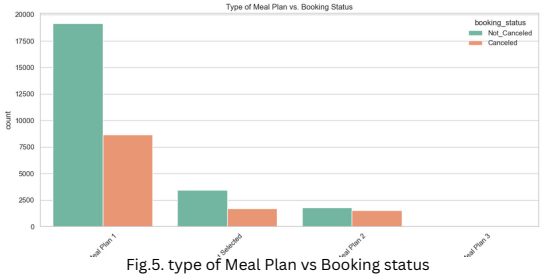

Fig.5. type of Meal Plan vs Booking status

Car Parking Space Requirement vs. Booking Status: Very few bookings require parking; those that do have a lower cancellation rate. See fig.6.

Fig 6. Car Parking vs Booking status

Room Type Reserved vs. Booking Status: 'Room_Type 1' is the most booked; different room types show varied cancellation rates. See fig.7.

Fig 7. Room Type Reserved vs Booking status

Monthly Booking Status: Bookings peak in certain months (like August), with varying cancellation rates across months. See fig.8.

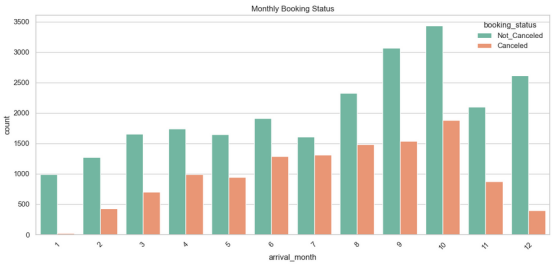

Fig 8. Monthly Booking Status

Arrival Date vs. Booking Status: Booking status distribution varies across the month, with no clear pattern. See fig.9.

Fig 9. Arrival data vs Booking status

Market Segment Type vs. Booking Status: 'Online' segment has more bookings and cancellations compared to 'Offline'. See fig.10.

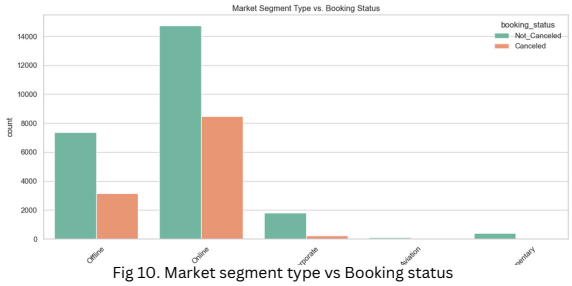

Fig 10. Market segment type vs Booking status

Repeated Guest vs. Booking Status: Non-repeated guests are more common, with a slightly higher cancellation rate than repeated guests. See fig.11.

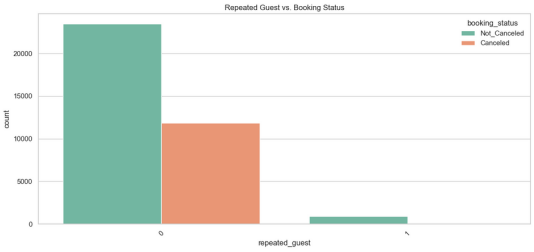

Fig 11. Repeated guests vs Booking status

Lead Time vs. Booking Status: Longer lead times are associated with higher cancellation rates. See fig.12.

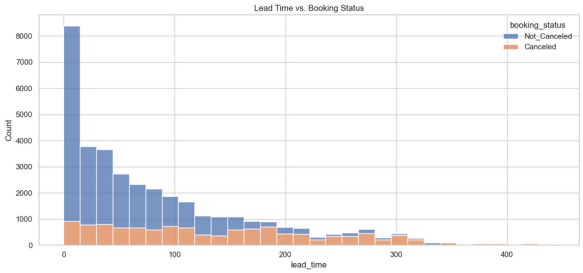

Fig 12. Lead Time vs Booking status

Previous Cancellations vs. Booking Status: Most guests have no previous cancellations; those with previous cancellations have a higher current cancellation rate. See fig.13.

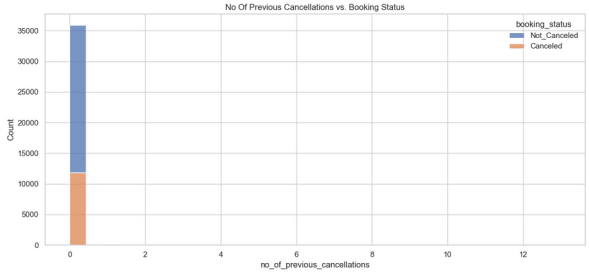

Fig 13. Previous Cancelations vs Booking status

Previous Bookings Not Canceled vs. Booking Status: Guests with no previous non-canceled bookings are more likely to cancel. See fig.14.

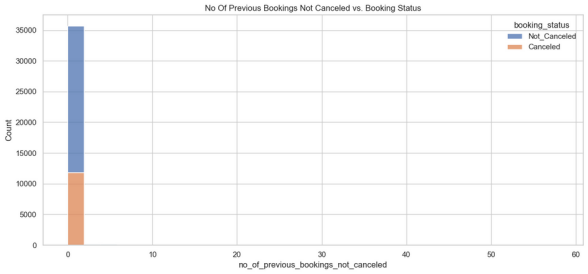

Fig 14. Booking not cancelled vs Booking status

Average Price Per Room vs. Booking Status: Higher room prices don't show a clear pattern in affecting cancellation rates. See fig.15.

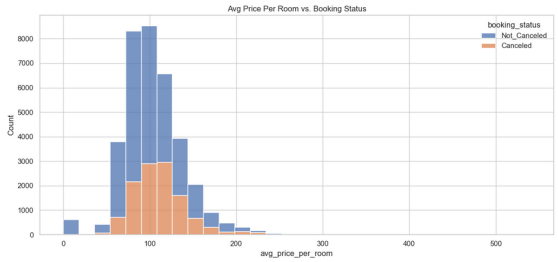

Fig 15. Average price vs Booking status

Number of Special Requests vs. Booking Status: Bookings with no special requests have a higher cancellation rate; more requests correlate with lower cancellations. See fig.16.

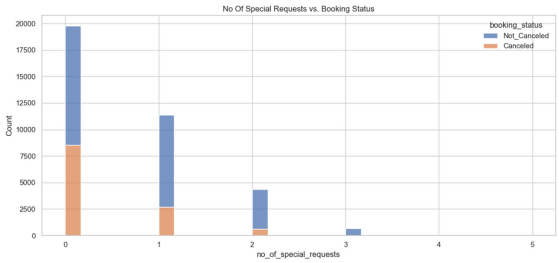

Fig 16. No. of Special Requests vs Booking status

Model Execution and Assessment (Pre-SMOTE)

Decision Tree Evaluation

A basic machine learning method called the Decision Tree Classifier was used to forecast whether or not guests will keep their hotel bookings. On the test data, the model demonstrated an impressive accuracy of almost 86.99%. The following are the comprehensive assessment metrics:

The precision for Class 0 (Not Honored) was 80%, the recall for Class 0 (Not Honored) was 81%, and the accuracy was 86.99%. The precision for Class 1 (Honored) was 91%, the recall for Class 1 (Honored) was 90%, and the F1-Score for Class 1 (Honored) was 90%.

Confusion matrix:

True Negative (Accurate Prediction, Not Honored): 1969

False Positive (Inaccurately Predicted as Honored, Not Honored): 447

False Negative (Predicted as Not Honored but Honored): 497

True Positive (Respected, Accurately Estimated): 4342.

Evaluation of K-Nearest Neighbors (KNN)

Another important technique for classification tasks, the KNN Classifier, was also utilized. With an accuracy of roughly 85.38% on the test data, the model performed somewhat worse than the Decision Tree. The KNN evaluation metrics are as follows:

Precision for Class 0 (Not Honored): 78% Accuracy: 85.38%.

Class 0 (Not Honored): Recall: 78%; Class 0 (Not Honored): F1-Score: 78%; Class 1 (Honored): Precision: 89%.

Class 1 (Honored) Recall: 89% Class 1 (Honored) F1-Score: 89%.

Confusion matrix:

The following are the true negatives: 1890 (Not Honored, Correctly Predicted); 526 (Not Honored, Incorrectly Predicted as Not Honored); 535 (Honored, Correctly Predicted); and 4304 (False Positive, Not Honored, Incorrectly Predicted as Not Honored).

Training Accuracy: 99.25% for KNN training accuracy.

KNN Test Accuracy: 85.38%.

This difference in accuracy between training and testing points to a possible overfitting of the KNN model, whereby the model performs remarkably well on training data but not as well on test data that hasn't been seen yet.

Decision Tree Using Various Criteria:

Accuracy of the Gini Criterion: 86.99%.

Accuracy of Entropy Criterion: 87.29%.

The Decision Tree model's accuracy according to both criteria is comparable, with the Entropy criterion exhibiting somewhat greater accuracy. This implies that the model's ability to forecast reservation honoring is somewhat influenced by the criterion selection.

These models' pre-SMOTE implementation provide a foundational knowledge of their functionality. The ensuing SMOTE application will provide insights into how class imbalance impacts model performance and prediction by attempting to balance the class distribution.

This phase is critical, particularly in light of the KNN model's overfitting that has been noticed and the Decision Tree's near performance metrics under various conditions. It was anticipated that the use of SMOTE will give these models a more fair training environment, which may improve their capacity to generalize and make accurate predictions on unobserved data. In order to comprehend the overall efficacy

of these algorithms in the context of forecasting customer behavior in hotel reservations, post-SMOTE model evaluation will be essential.

Model Execution and Assessment (Post-SMOTE)

Resolving Class Variations Using SMOTE

The dataset's apparent class imbalance was addressed with the use of SMOTE (Synthetic Minority Over-sampling Technique). This method successfully balanced the classes, as seen by the distributions of classes both before and after SMOTE application:

Prior to SMOTE, Class 0 (Not Honored): 9469 Class 1 (Honored): 19551

Following SMOTE, 19551 is Class 1 (Honored) and 19551 is Class 0 (Not Honored).

Standardization of Features After standardizing the features, post-SMOTE, each feature was guaranteed to have a mean near zero and a standard deviation near one. For algorithms like KNN, which are sensitive to the volume of the data, this standardization is essential.

Post-SMOTE Decision Tree Evaluation

Using the balanced dataset, the Decision Tree Classifier was retrained. The following are the model's post-SMOTE performance metrics:

Class 0 (Not Honored): 79% Precision; Class 0 (Not Honored): 83% Recall; F1-Score; Class 0 (Not Honored): 81% Precision; Class 1 (Honored): 91% Accuracy: 86.80%

Recall for Honored Class 1: 89%; F1-Score for Honored Class 1: 90%

Confusion matrix:

True Negative (Accurate Prediction, Not Honored): 2002

False Positive (Inaccurately Predicted as Honored, Not Honored): 414

Untrue Negative (Honored but Wrongly Assumed to Be Not Honored): 544

True Positive (Respected, Accurately Estimated): 4295

Evaluation of K-Nearest Neighbors (KNN) (Post-SMOTE)

Additionally, the balanced dataset was used to retrain the KNN Classifier. The post-SMOTE performance metrics of the model are:

Precision for Class 0 (Not Honored): 65% Recall for Class 0 (Not Honored): 76% F1-Score for Class 0 (Not Honored): 70% Accuracy: 78.33% Recall for Class 1 (Honored): 80% F1-Score for Class 1 (Honored): 83%

Confusion matrix:

The following are the true negatives: 1828 True Negative (Not Honored, Correctly Predicted): 588 False Positive (Not Honored, Incorrectly Predicted as Not Honored): 984 True Positive (Honored, Correctly Predicted):

Comparing the Models' Results Pre- and Post-SMOTE Comparison

Comparison of Decision Tree Classifier Models

Pre-SMOTE Accuracy: 86.99%

Accuracy of Post-SMOTE: 86.80%

Following SMOTE, the Decision Tree model's accuracy slightly decreased, going from 86.99% to 86.80%. This small adjustment, meanwhile, suggests that the model was reasonably resistant to class imbalance. The Class 0 recall slightly improved (from 81% to 83%) after SMOTE, indicating a better ability to identify reservations that were not kept.

Comparison of K-Nearest Neighbors (KNN) Models

Pre-SMOTE accuracy: 85.38%

Accuracy of Post-SMOTE: 78.33%

After SMOTE, the KNN model's overall accuracy decreased more noticeably, going from 85.38% to 78.33%. This notable modification emphasizes how sensitive the model was initially to class inequality.

Analysis of the Results

Decision Tree Resilience: The small adjustments in accuracy and gains in class-specific measures following SMOTE show that the Decision Tree model's performance was less affected by the class imbalance.

KNN Sensitivity: Initially exhibiting overfitting to the majority class, the KNN model was more impacted by the class imbalance. While its ability to forecast the minority class increased after SMOTE, overall accuracy declined, indicating a trade-off between balanced class prediction and overall accuracy.

Implications: The findings show that, although correcting class imbalance is essential for equitable and balanced model performance, doing so may result in accuracy trade-offs. This emphasizes how crucial it is to use models and evaluation criteria that are appropriate for the particular context and research goals.

Recommendations: It is advisable to take into account both the overall accuracy and the balance in class prediction for subsequent predictions and model implementations. Different models or methodologies may be more appropriate depending on the business objectives (e.g., prioritizing the accurate prediction of non-honored bookings).

In summary, the use of SMOTE highlighted intriguing dynamics in model performance and emphasized the significance of taking class imbalance into account in predictive modeling, particularly in situations where one class may predominate over another, such as hotel booking forecasts.

Conclusion:

This study dealt with forecasting hotel booking trends and provided a thorough overview of the many stages of data analysis. The comprehensive dataset offered a thorough understanding of hotel reservation patterns since it included both numerical and categorical variables. Critical insights into the relationship between different attributes and the booking status were found during the initial exploratory data analysis (EDA). These insights were crucial in directing the later data pretreatment stages, such as label encoding and feature scaling. The difficulties in handling unbalanced datasets were brought to light by the deployment of Decision Tree and K-Nearest Neighbors (KNN) models, both prior to and following the use of SMOTE (Synthetic Minority Over-sampling Technique). The KNN model's performance fluctuated, highlighting the susceptibility to class distribution, but the Decision Tree model demonstrated resilience to class imbalance.

Although there were accuracy trade-offs, the use of SMOTE gave the prediction models a more balanced perspective. This investigation highlighted the significance of taking class distribution into account in predictive modeling in addition to providing a deeper insight of consumer booking behavior. The fact that thorough data analysis and useful model application can coexist in harmony is evidence of the complexity of data science initiatives. This work represents a major advancement in the field of data-driven decision-making and provides a useful model for similar investigations in the hospitality sector. It also helps to optimize booking systems and improve client satisfaction.