The University of Texas, Arlington



**INSY 5339 – 001: PRINCIPLES OF BUSINESS DATA MINING**

### Professor Dr. Riyaz Sikora

## Project Report

## Predicting Hotel Reservation Cancellations

**Group 5** Niharika Batra

Sravani

#### Vaishnavi Sulegai Radheshyam Venkata Laxmana Srikanth Bharam

#### TABLE OF CONTENTS

INTRODUCTION 3

[DATA DESCRIPTION 3](#_bookmark0)

[OBJECTIVE 5](#_bookmark1)

[DATA PRE-PROCESSING 5](#_bookmark2)

[HYPOTHESIS TESTING 6](#_bookmark3)

[CORRELATION ANALYSIS 8](#_bookmark4)

[DESCRIPTIVE ANALYSIS 10](#_bookmark5)

[CLUSTER ANALYSIS TO IDENTIFY MARKET SEGMENTS 13](#_bookmark6)

[INITIAL RESULTS 15](#_bookmark7)

[PREDICTIVE ANALYSIS 16](#_bookmark8)

[RESULTS 18](#_bookmark9)

[IMPORTANT FEATURES OF MODEL 22](#_bookmark10)

[MODEL DEPLOYMENT 23](#_bookmark11)

[PRESCRIPTIVE MEASURES 24](#_bookmark12)

[CONCLUSION 25](#_bookmark13)

Background

The hospitality industry has undergone a remarkable evolution with the advent of online hotel reservation channels, revolutionizing the way customers interact with hotels and make bookings. This shift has undoubtedly enhanced convenience and accessibility, empowering travelers to plan their stays efficiently. However, along with these advantages, there has been a surge in reservation cancellations and no-shows, presenting a significant challenge for hotel owners and industry stakeholders.

In 2018, a staggering 40% of on-the-books revenue was canceled before the scheduled arrival, highlighting the magnitude of this issue. Furthermore, it has been observed that reservations with lead times exceeding 60 days are 65% more likely to face cancellations. The repercussions of these cancellations extend beyond the hotel itself, impacting the entire value chain of the travel industry. Not only does the customer bear the burden of cancellation fees, but travel agents also experience losses in the form of forfeited commissions. Additionally, the hotel itself suffers from the potential loss of around 30% in additional income from Food & Beverage and Amenity purchases that would have been made by the guests.

To address this pressing concern and mitigate its adverse effects, we propose a solution in the form of a machine learning model. The primary objective of this project is to develop a predictive model that can accurately estimate the likelihood of a hotel reservation being canceled, based on a comprehensive analysis of various customer and booking attributes. By leveraging data-driven insights, hotel owners can proactively anticipate cancellations, allowing them to implement strategic measures and minimize revenue losses. In this report, we will delve into the intricate workings of our machine learning approach, detailing the selection and analysis of pertinent attributes, the model development process, and the evaluation of its predictive performance. Our findings have the potential to equip the hospitality industry with a valuable tool to combat reservation cancellations, ultimately fostering improved efficiency, profitability, and customer satisfaction.

# Data Description

The dataset for this project is sourced from Kaggle ([Reservation Cancellation Prediction | Kaggle](https://www.kaggle.com/datasets/gauravduttakiit/reservation-cancellation-prediction?select=train__dataset.csv)). It provides comprehensive information about customer reservation details and includes variables related to booking characteristics, customer demographics, and previous booking history.

The dataset comprises 18 variables that capture different aspects of hotel reservations. These include the booking ID, the number of adults and children in the reservation, the duration of weekend and week nights stayed, the type of meal plan, the requirement for car parking space, the reserved room type, lead time,

arrival date information, market segment designation, customer repeatability, previous cancellations and bookings, average price per room, special requests, and the booking status (canceled or not canceled). Below is the *Table1* that provides a concise overview of data description.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Booking\_ID | Unique identifier of each booking |
| No of adults | Number of adults in the reservation |
| No of children | Number of children in the reservation |
| No of weekend nights | Number of weekend nights (Saturday or Sunday) the guest  stayed/booked |
| No of week nights | Number of week nights (Monday to Friday) the guest  stayed/booked |
| Type of meal plan | Meal plan booked by the customer |
| Required car parking space | Indicator of whether the customer requires a car parking space |
| Room type reserved | Type of room reserved by the customer |
| Lead time | Number of days between booking date and arrival date |
| Arrival year | Year of the arrival date |
| Arrival month | Month of the arrival date |
| Arrival date | Date of the month |
| Market segment type | Market segment designation |
| Repeated guest | Indicator of whether the customer is a repeated guest |
| No of previous cancellations | Number of previous bookings canceled by the customer |
| No of previous bookings not  canceled | Number of previous bookings not canceled by the customer |
| Average price per room | Average price per day of the reservation |
| No of special requests | Total number of special requests made by the customer |
| Booking status (Target variable) | Flag indicating if the booking was canceled or not |

*Table 1. Overview of the variables and their descriptions in the dataset.*

# Objective

The primary objective of this project is to develop a machine learning model that can accurately predict whether a customer will honor their hotel reservation or cancel it. By analyzing the dataset, we aim to identify the factors that influence reservation cancellations. This predictive model will enable hotel owners to proactively manage their resources, optimize revenue management strategies, and enhance customer service by better understanding customer booking behavior.

# Data Pre-processing

In this project, we performed essential data preprocessing steps to ensure the dataset's quality and preparedness for building the predictive model. One crucial aspect of data preprocessing is handling missing values. Fortunately, in our dataset, all missing values were addressed, and the dataset is entirely encoded, providing complete information for analysis.

* Unique Values Overview - Before diving into data preprocessing, we initially explored the dataset to gain insights into the unique values present in each column. This process helps in understanding the data's characteristics and identifying potential issues that might require further attention.
* Removing Rows with Zero Adults and Zero Children **-** To maintain data integrity and relevance, we excluded rows where both the number of adults and children were zero. Such entries are likely erroneous and do not represent valid hotel reservations. By performing this step, we ensured that the dataset only contains entries with non-zero adult or child guests.
* Removing Duplicate Rows- Duplicate entries in the dataset can lead to skewed analysis and affect the performance of the predictive model. To maintain the dataset's accuracy, we carefully checked for and removed any duplicate rows, resulting in a clean and deduplicated data frame of 14026 rows.
* Outlier Removal- Outliers can significantly influence statistical analysis and machine learning models. To enhance the model's robustness and accuracy, we identified and removed outliers from several columns, including "children," "adults," "weekend nights," "week nights," "room\_type\_reserved," and "type\_of\_meal\_plan."



*Figure 1. Box plots for all variables to detect outliers and distribution.*

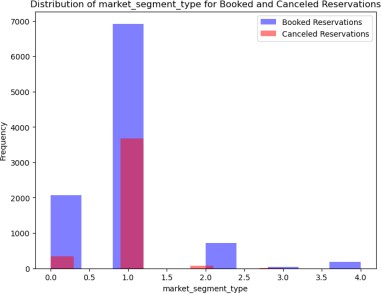
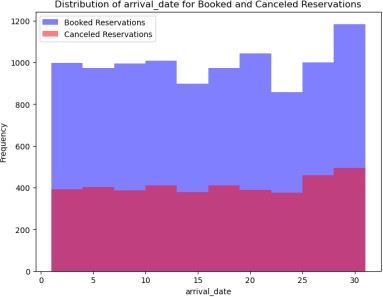
After completing the data preprocessing steps, we obtained a refined dataset with 13,983 entries and 18 columns, ready to be utilized for developing the machine learning model. The dataset is now devoid of missing values, duplicates, and significant outliers, ensuring reliable and meaningful insights in our prediction analysis.

# Hypothesis Testing

In hypothesis testing, we aimed to investigate the significance of various variables concerning the booking\_status, particularly focusing on their potential as predictors of reservation cancellations in the dataset. The null hypothesis stated that there is no significance between booking\_status and the examined variables, while the alternative hypothesis proposed that there is a significant relationship between them.

Upon conducting the hypothesis tests, we obtained the following p-values for the respective variables:

* **Market Segment Type:** The p-value for the market segment type was found to be 0.06. As this p- value is greater than the chosen significance level (alpha = 0.05), we fail to reject the null hypothesis. The lack of statistical significance suggests that the market segment type may not be a strong predictor of reservation cancellations in the dataset.
* **Arrival Date:** The p-value for the arrival date was calculated as 0.128. Again, this p-value is greater than the chosen significance level (alpha = 0.05), leading us to fail to reject the null hypothesis. Consequently, we do not find a statistically significant relationship between the arrival date and booking\_status. This indicates that the arrival date may not be a robust predictor of reservation cancellations.



*P-value : 0.128 P-value : 0.06*

*Figure 2. P-values and distribution of arrival date and market segment type*

However, it is crucial to note that certain other variables did reject the null hypothesis. These variables demonstrated a significant relationship with booking\_status, indicating their importance as potential predictors for reservation cancellations. These influential variables may encompass a wide range of factors, such as lead time, room type etc.

# Correlation Analysis

In this, we investigated the relationship between various variables and booking\_status in the dataset. Correlation coefficients indicate the strength and direction of the linear relationship between two variables. A positive correlation coefficient implies that as one variable increases, the other tends to increase as well, while a negative correlation coefficient suggests that as one variable increases, the other tends to decrease.

##### Positive Correlations with Booking Status:

`no\_of\_adults` (0.11)

`no\_of\_children` (0.07)

`no\_of\_weekend\_nights` (0.05)

`no\_of\_week\_nights` (0.12)

`type\_of\_meal\_plan` (0.06)

`room\_type\_reserved` (0.06)

`lead\_time` (0.37)

`arrival\_year` (0.16)

`avg\_price\_per\_room` (0.16)

The positive correlations indicate that an increase in these variables tends to be associated with a higher likelihood of the booking being cancelled. For instance, a larger number of adults and children in the reservation, longer lead times, and higher average price per room are positively correlated with a higher likelihood of cancellation.

##### Negative Correlations with Booking Status:

`required\_car\_parking\_space` (-0.09)

`repeated\_guest` (-0.11)

`no\_of\_previous\_cancellations` (-0.03)

`no\_of\_previous\_bookings\_not\_canceled` (-0.06)

`no\_of\_special\_requests` (-0.25)

The negative correlations suggest that an increase in these variables tends to be associated with a higher likelihood of the booking being not canceled. For example, a higher number of special requests and previous cancellations are negatively correlated with the reservation being canceled.

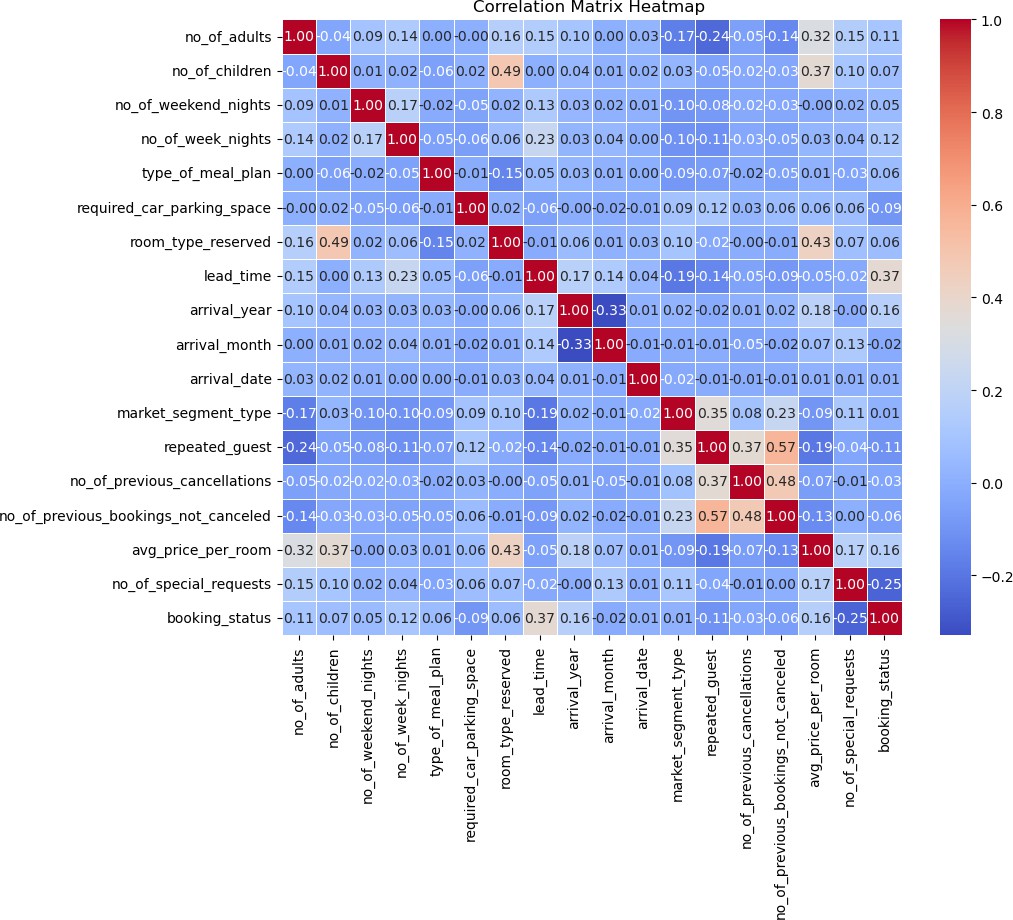
##### Weaker Correlations with Booking Status:

`arrival\_month` (-0.02)

`arrival\_date` (0.01)

`market\_segment\_type` (0.01)

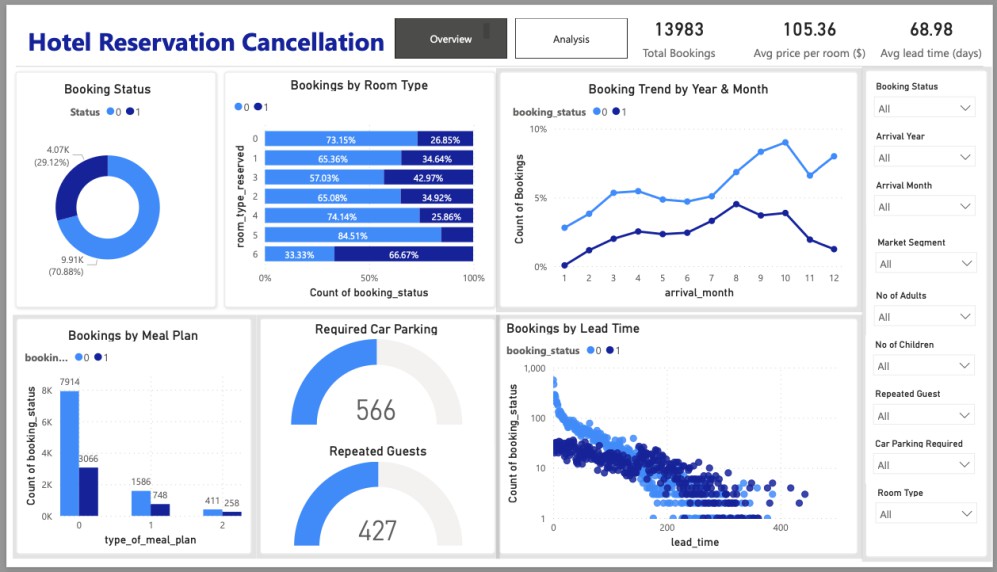
These variables exhibit weaker correlations with booking\_status, suggesting that their influence on reservation cancellations may not be as pronounced compared to the variables with stronger correlations. So, we can remove these weaker correlated variables while modelling.



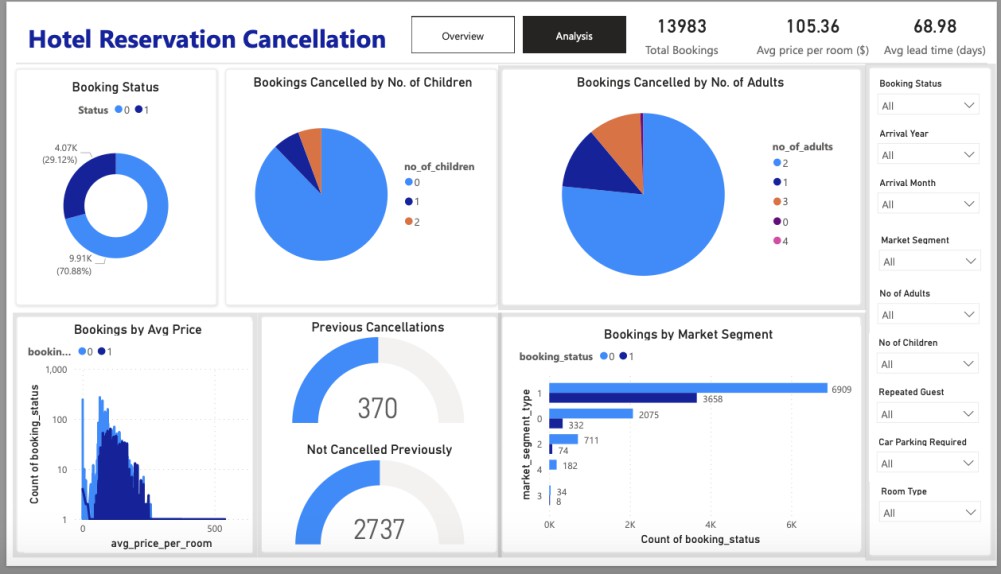
*Figure 3. Correlation Matrix*

# Descriptive Analysis

To effectively present the findings and insights from our analysis, we have developed an interactive and comprehensive dashboard using Power BI. This dashboard showcases various visualizations and essential information to provide a user-friendly interface for hotel owners and stakeholders, allowing them to gain actionable insights and make informed decisions based on the data analysis.



*Figure 4. Power BI Dashboard Overview Tab*



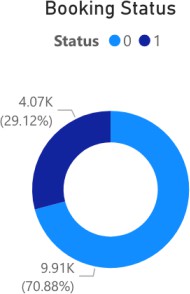
*Figure 5. Power BI Dashboard Analysis Tab*

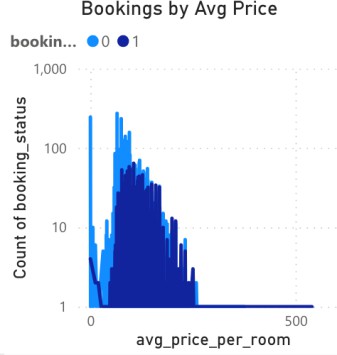
Some of the observations are:

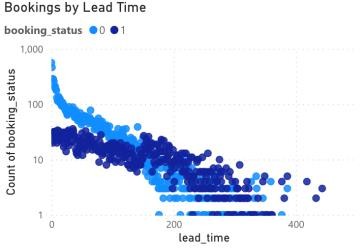
* The data contains information on 13,983 hotel bookings. The average price per room across all bookings is $105.36, and the average lead time (time between booking and check-in date) is 68.98 days.



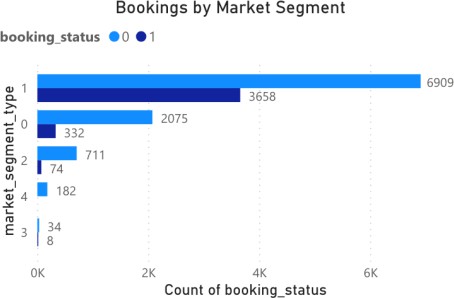
* Out of the total bookings, 9911 were confirmed (not cancelled - 0), and the rest were cancelled - 1. This indicates that the cancellation rate is significant, as more than a third of the bookings were cancelled.



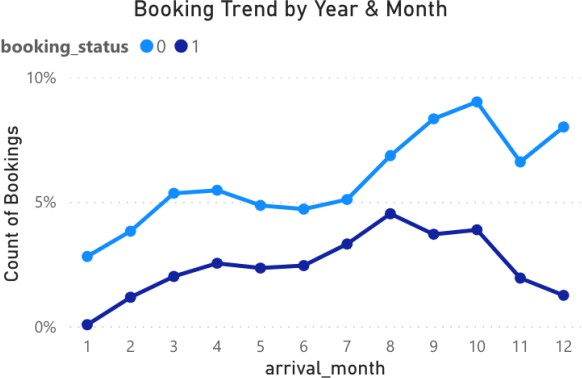
* The average room price for the cancelled bookings is higher than that of the confirmed bookings. This suggests that customers who booked higher-priced rooms are more likely to cancel their reservations.
* Bookings with less lead time (recently made bookings) are more likely to be confirmed. This implies that customers who book closer to their check-in date are more committed to their reservation and less likely to cancel.



* Market segment 1 followed by 0 has the highest number of both cancellations and bookings. It seems that these two segments are more active in making hotel reservations, but they also face a higher cancellation rate compared to other segments.



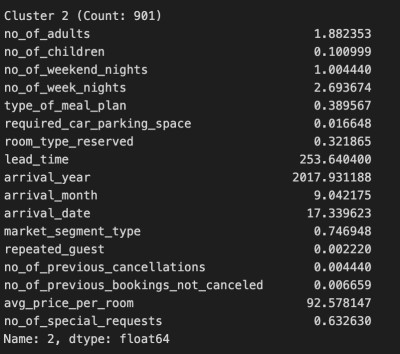
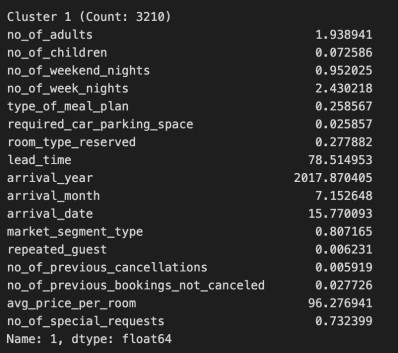
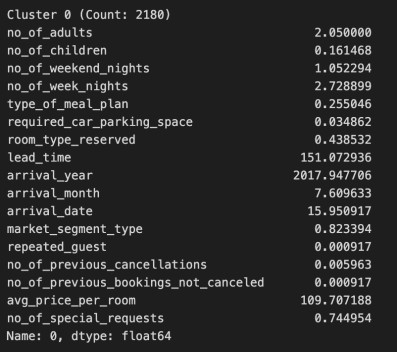
* Retained customers (presumably customers who have booked with the hotel before) have a low cancellation rate, with only 0-2 cancellations for bookings made through market segments. This indicates that loyal customers are more committed to their reservations and less likely to cancel.
* Bookings with special requests are less likely to be cancelled. This suggests that when customers make special requests, they are more invested in their stay and are less likely to cancel their bookings.
* Most of the bookings are made between the 5th and 10th month. During this period, there is a higher number of repeated guests (returning customers), and these guests are more willing to pay a higher price per room to confirm their bookings. This indicates that loyal customers are more inclined to pay a premium to secure their reservations during this time period.

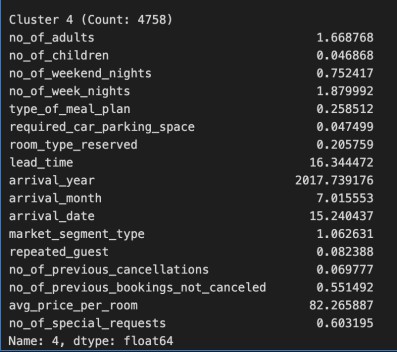


In conclusion, this descriptive analysis and interactive dashboard offer a comprehensive understanding of hotel reservation cancellations and related factors. By utilizing these insights, hotel owners and stakeholders can make data-driven decisions, optimize revenue management strategies, and enhance overall customer service, leading to improved guest satisfaction and operational efficiency.

# Cluster Analysis to identify Market Segments

To identify the market segments, we performed cluster analysis using K-Means and formed 5 different clusters.





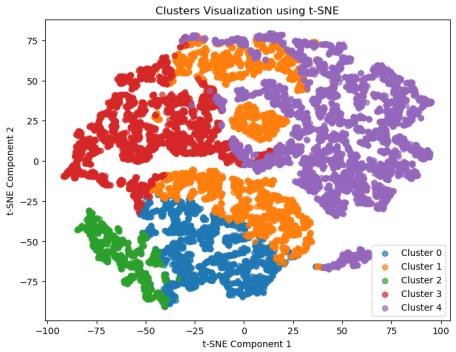
*Figure 6. Clusters*

To identify Distinct Market Segments, we analyzed the clusters description:

1. **Cluster 0 - Potential Family Travelers:** This segment consists of bookings with a relatively higher number of adults and a moderate number of children, suggesting they might be families or groups of adults

traveling together. They tend to stay for both weekend and longer weekday durations, indicating a mix of leisure and business travelers. The segment shows interest in specific meal plans and values personalized services with a considerable number of special requests.

1. **Cluster 1- Individual and Business Travelers:** This segment includes bookings with fewer adults and very few children, indicating they may consist of smaller groups or individual and business travelers. They prefer shorter weekend stays and longer weekday stays, suggesting a focus on business or extended trips. They show a moderate interest in car parking space and appreciate some personalized services with a significant number of special requests.
2. **Cluster 2 - Leisure-Oriented Travelers:** This segment comprises bookings with an average number of adults and a moderate number of children, indicating families with some children. They prefer slightly shorter weekend stays and longer weekday stays. The segment shows a distinct preference for a specific type of meal plan and appreciates **personalized** services.
3. **Cluster 3 - Loyal and Family-Oriented Travelers:** This segment includes bookings with a relatively higher number of adults and a higher number of children, suggesting families with several children. They prefer shorter weekend stays and longer weekday stays. The segment shows a distinct preference for a specific type of meal plan and values personalized services with a considerable number of special requests. They are more likely to be **loyal, repeated guests**.
4. **Cluster 4 - Individual and Budget Travelers:** This segment consists of bookings with fewer adults and very few children, suggesting they may consist mainly of individual or smaller groups of budget-conscious travelers. They prefer shorter weekend stays and longer weekday stays, suggesting a focus on business or extended trips. The segment shows a moderate interest in **car parking space** and prioritizes essential services over personalized requests.



*Figure 7. Cluster Visualization*

# Initial Results

Based on the dataset, we observed the following initial results and analysis before pre-processing as described in Table 2:

|  |  |  |
| --- | --- | --- |
| 1 | How many observations in the data set? | 18137 |
| 2 | How many binary/categorical values? | 3 |
| 3 | How many continuous variables? | 12 |
| 4 | What is the target/outcome variable? | Booking\_status |
| 5 | If binary or categorical: What percentage of the variables belong to the class? | % Distribution of booking\_status:   1. (Not canceled) - 67.24% 2. (Cancelled) - 32.76% |
| 6 | If continuous: What is the mean value of the target  variable? | N/A |
| 7 | Before doing any further processing, what would be  prediction of the target variable? | The prediction of the target variable  would be 0 (Not canceled) |

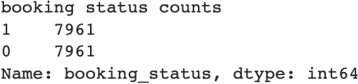
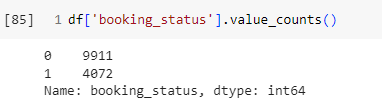
*Table 2. Initial results and analysis*

* + The dataset consists of 18 variables, including the target variable 'booking\_status' and 18137 observations.
  + Binary variables are 3 in total including ['required\_car\_parking\_space', 'repeated\_guest', 'booking\_status']
  + Continuous variables are 12 which are ['no\_of\_adults', 'no\_of\_children', 'no\_of\_weekend\_nights', 'no\_of\_week\_nights','lead\_time','arrival\_year','arrival\_month','arrival\_date','no\_of\_previous\_cancellat ions', 'no\_of\_previous\_bookings\_not\_canceled', 'avg\_price\_per\_room', 'no\_of\_special\_requests']
  + Target variable: booking\_status

# Predictive Analysis

In this project, we employed three distinct machine learning models to address the challenge of predicting hotel reservation cancellations: Logistic Regression, Decision Tree, and Random Forest and used SMOTE for balancing the dataset.

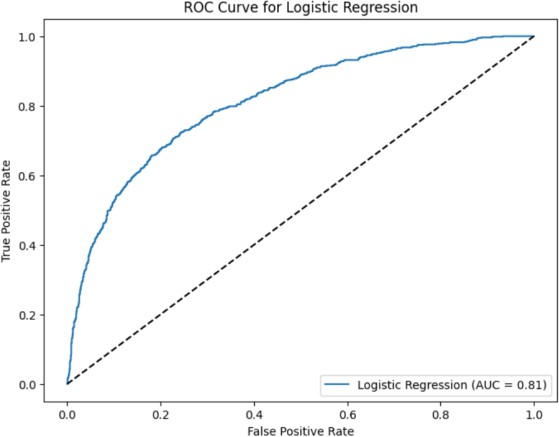
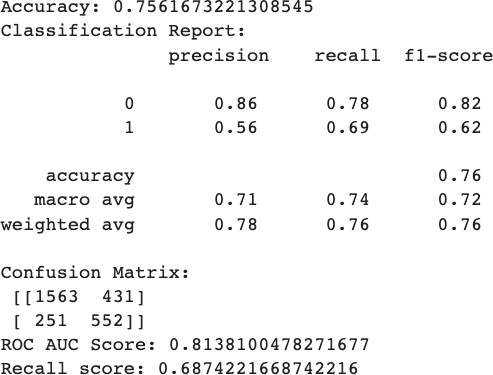
To address the class imbalance in the original dataset, which had 9,911 entries with booking\_status = 0 (not canceled) and only 4,072 entries with booking\_status = 1 (canceled), we employed the **SMOTE** (Synthetic Minority Over-sampling Technique) method. By applying SMOTE to the training dataset, which constituted 80% of the total dataset, we successfully balanced the classes for both booking\_status = 0 and booking\_status = 1. SMOTE, a powerful technique, generated synthetic samples of the minority class (booking\_status = 1) to match the quantity of the majority class (booking\_status = 0), effectively increasing the number of canceled bookings in the dataset. This process ensured a more equitable distribution between the two classes, enabling our machine learning model to receive unbiased and robust training. By mitigating the class imbalance, the model's predictive capabilities have been enhanced, facilitating accurate predictions for both canceled and not-canceled bookings, thus significantly improving the overall performance and reliability of the system.



*Figure 8. Original dataset(above) and After SMOTE on Training dataset(below)*

#### Logistic Regression

Logistic Regression is a linear model widely used for binary classification tasks. It estimates the probability of a data point belonging to a specific class, in this case, whether a hotel reservation is canceled or not. The model achieved an accuracy of approximately 75.62% on the test dataset. The classification report revealed a reasonable balance between precision and recall, with a weighted average F1-score of 0.76. The ROC AUC score of 0.82 demonstrated the model's ability to distinguish between canceled and not canceled reservations. The recall score of around 0.69 indicated the model's effectiveness in identifying canceled reservations.

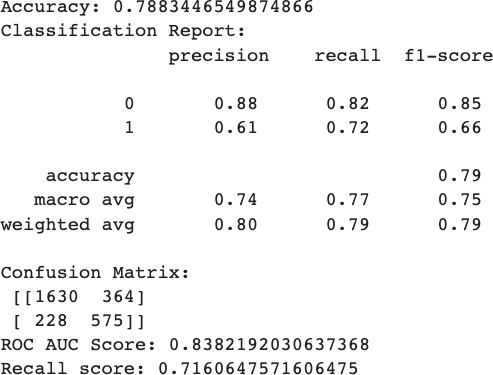
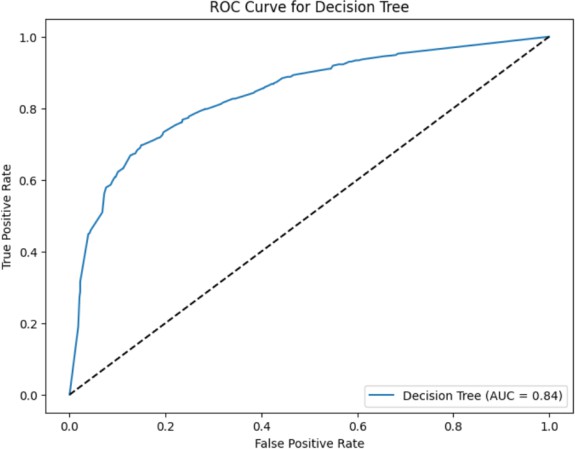
 

*Figure 9. Logistic Regression*

#### Decision Tree

The Decision Tree model is a non-linear model that learns to make decisions by splitting the data based on certain features. It is capable of handling both binary and multi-class classification tasks. We trained the Decision Tree on the dataset to predict hotel reservation cancellations. Decision Trees can capture complex relationships in the data, but they may be prone to overfitting. The model achieved an accuracy of approximately 78.83% on the test dataset. The classification report revealed a reasonable balance between precision and recall, with a weighted average F1-score of 0.79. The ROC AUC score of 0.84 demonstrated the model's ability to distinguish between canceled and not canceled reservations. The recall score of around

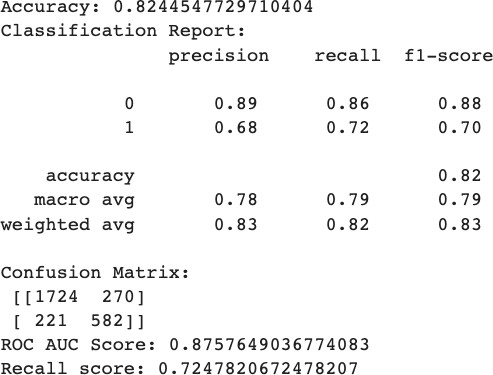
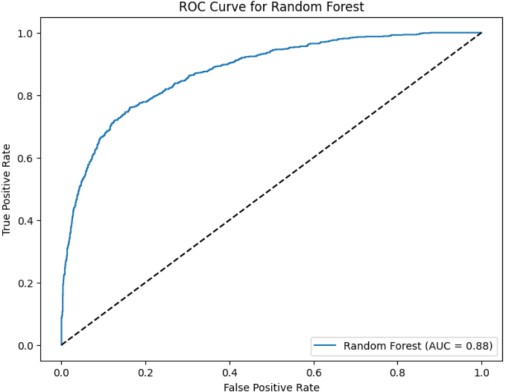
0.71 indicated the model's effectiveness in identifying canceled reservations.



*Figure 10. Decision Tree*

#### Random Forest

Random Forest is an ensemble learning method that combines multiple Decision Trees to improve predictive accuracy and reduce overfitting. It creates a multitude of decision trees and combines their predictions through voting or averaging. By using this model, we aimed to achieve higher accuracy and better generalization on unseen data. The model achieved an accuracy of approximately 82.44% on the test dataset. The classification report revealed a reasonable balance between precision and recall, with a weighted average F1-score of 0.83. The ROC AUC score of 0.88 demonstrated the model's ability to distinguish between canceled and not canceled reservations. The recall score of around 0.72 indicated the model's effectiveness in identifying canceled reservations.



*Figure 11. Random Forest*

# Results

In conclusion, after evaluating the performance of three models - Logistic Regression, Decision Tree, and Random Forest - on the hotel reservation cancellation dataset, the **Random Forest model** emerges as the most promising choice. The Random Forest model demonstrates the highest recall score of 0.72, showcasing its ability to effectively detect a significant portion of actual cancelled bookings, thereby minimizing the occurrences of false negatives. This characteristic is crucial for business, as correctly identifying cancelled reservations can enable effective resource management and customer service.

Additionally, the Receiver Operating Characteristic (ROC) curve with an AUC score of 0.88 validates the model's representation of the trade-off between true positive rate and false positive rate. The higher AUC score further confirms the Random Forest model's proficiency in making informed predictions while balancing the misclassification costs. Moreover, the Random Forest model attains the highest overall accuracy of 82%, signifying its well-rounded performance in classifying both cancelled and non-cancelled reservations.

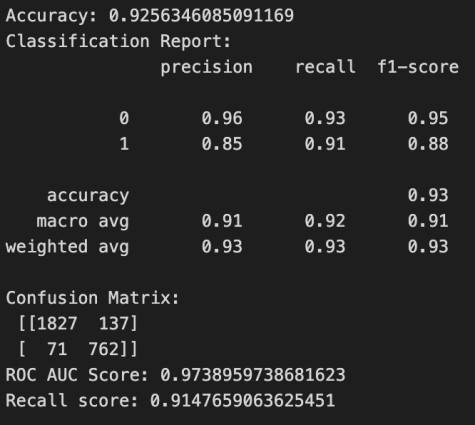
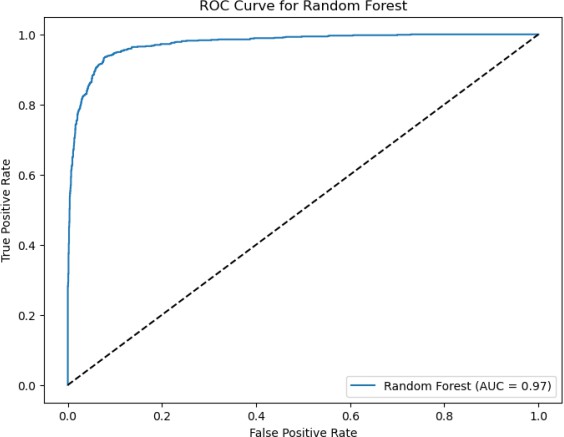
#### Impact of imbalance class on Model Predictions

The comparison of the Random Forest model's performance before and after applying SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset sheds light on the impact of class imbalance on the model's predictions. Class imbalance in the dataset can lead to skewed model performance. In this case, the high accuracy and F1-score might be misleading since the model is primarily predicting the majority class (non-canceled reservations) accurately. It might not perform as well in correctly identifying the minority class (canceled reservations), which is of particular interest in many practical scenarios.

##### Random Forest Model Before SMOTE Balancing:

Before applying SMOTE, the Random Forest model achieved a high accuracy of 92.6%. This high accuracy might suggest that the model is performing exceptionally well, but it requires further investigation, especially considering the class imbalance in the dataset.

For the positive class (canceled reservations), the model showed promising results with a recall score of 91.5%. Recall measures the model's ability to correctly identify positive instances, i.e., correctly predicting reservation cancellations. The F1-score for the positive class was 0.88, which indicates a good balance between precision and recall for this class. However, it is essential to consider that the dataset contains significantly fewer instances of canceled reservations compared to non-canceled ones.



*Figure 11. Random Forest without balancing using SMOTE*

##### Random Forest Model After SMOTE Balancing:

To address the class imbalance, we applied SMOTE, a technique that generates synthetic samples of the minority class (canceled reservations) to increase its representation in the dataset. After applying SMOTE, the dataset was balanced, and the Random Forest model's accuracy decreased to 81.6%. The decrease in

accuracy is expected as the model now has to handle a balanced dataset with equal representation of both classes.

However, the trade-off between accuracy and recall became apparent in the balanced dataset. The recall score for the positive class slightly dropped to 71.4%, indicating that the model's ability to correctly identify canceled reservations decreased. The F1-score for the positive class also decreased to 0.70, which means there is a compromise in the balance between precision and recall for the minority class.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (Class 1)** | **Recall (Class 1)** | **F1-score (Class 1)** | **ROC AUC**  **Score** |
| Before SMOTE  Balancing | 0.926 | 0.85 | 0.91 | 0.88 | 0.974 |
| After SMOTE  Balancing | 0.816 | 0.66 | 0.71 | 0.69 | 0.869 |

*Table 2. Comparison of Random Forest Model Before and After SMOTE*

Hence, the model before SMOTE balancing achieved higher performance in accuracy, precision, recall, F1- score, and ROC AUC score, indicating better overall predictive ability, especially in correctly identifying positive class instances (canceled bookings). After SMOTE balancing, the model's performance decreased slightly, with reduced recall and F1-score, but it remained effective in detecting canceled bookings.

#### Assessing False Negatives in Random Forest Model:

We can infer the characteristics of the instances where the predicted label is 0 (not cancelled) while the true label is 1 (cancelled), displaying as false negatives:

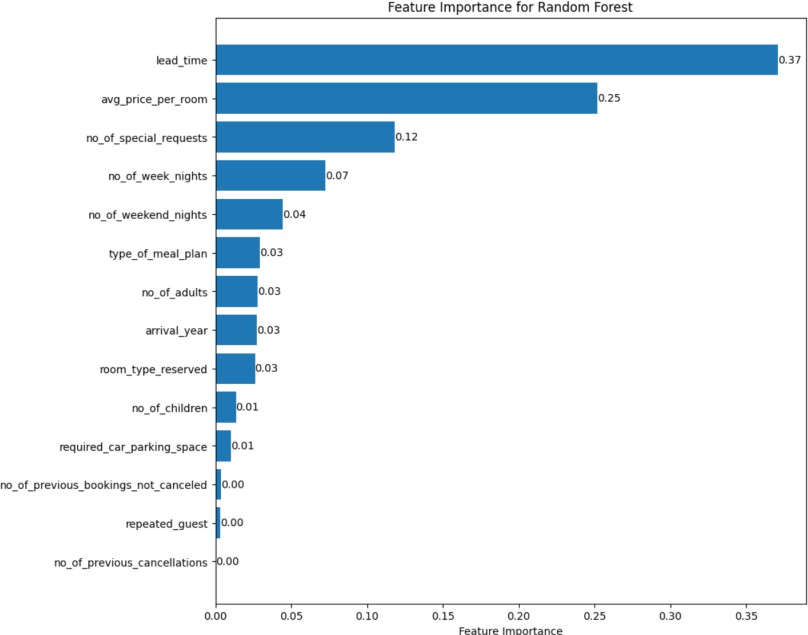
* **Booking Details:** The data shows that the bookings in question have relatively small group sizes, with an average of around 2 adults and less than 1 child. Most of these bookings were for weekend nights (around 2.5 nights) and had longer lead times (approximately 147 days on average). This suggests that these bookings were planned well in advance.
* **Type of Meal Plan and Car Parking:** The majority of these bookings had no special meal plan (0.31 on average) and required no car parking space (0.01 on average). This could imply that these customers may have preferred to dine outside the hotel or used alternative transportation methods.
* **Room Type and Repeated Guest:** The average value for room\_type\_reserved is approximately 0.48, indicating that these customers might have opted for standard room types rather than premium or upgraded options. Moreover, the average value of repeated\_guest is 0, suggesting that they were not repeat customers.
* **No Previous Cancellations:** All the bookings in this group have zero counts for no\_of\_previous\_cancellations and no\_of\_previous\_bookings\_not\_canceled, indicating that these customers had not canceled any previous bookings and did not have any prior bookings that were not canceled.

Overall, these characteristics suggest that the false negatives (predicted as not cancelled while they were actually cancelled) in the Random Forest model are associated with bookings that were well-planned, had no previous cancellations, and might not have required additional services or special room types. Further analysis and feature engineering or hyperparameter tuning may be required to improve the model's performance in predicting cancellations accurately.

# Important Features of Model

In the predictive model for booking status, feature importance analysis reveals the top influential features, which play a significant role in determining whether a booking will be canceled or not. The top 5 features are as follows:

1. **Lead Time (0.366275):** Lead time refers to the duration between booking and arrival. A longer lead time might indicate more time for potential changes in plans, leading to a higher likelihood of cancellations.
2. **Average Price per Room (0.247987):** The average price per room is indicative of the overall cost and quality of the booking. Higher prices might lead to more cautious decision-making and, thus, a lower chance of cancellations.
3. **Number of Special Requests (0.122182**): Guests with specific preferences or requirements may be more likely to finalize their bookings to ensure their needs are met. Hence, a higher number of special requests may decrease the likelihood of cancellations.
4. **Number of Week Nights (0.072669):** The duration of the stay (number of week nights) can influence cancellation probabilities. Longer stays may have more significant commitments and, therefore, a reduced likelihood of cancellations.
5. **Number of Weekends Nights (0.045264):** Similar to the number of week nights, the duration of weekend stays can also impact cancellation rates. Weekend bookings might be associated with leisure or planned events, leading to fewer cancellations.



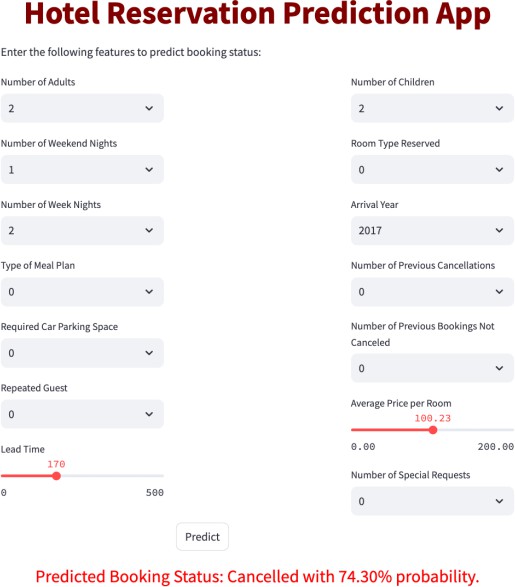
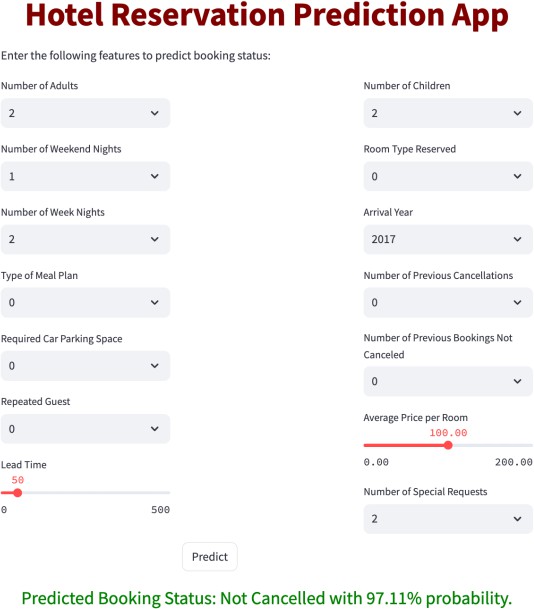
*Figure 12. Important Features for Random Forest*

# Model Deployment

In order to make our Random Forest model easily accessible and interactive, we decided to deploy it on Streamlit, a web application framework for Python. Streamlit allows us to create a user-friendly and dynamic web application that enables real-time predictions on booking cancellations.

The deployment process began by creating a Streamlit app file. We imported the required Python libraries, loaded the pre-trained Random Forest model using either joblib or pickle, and designed the user interface with input fields such as the number of adults, number of children, lead time, average price per room, etc.

To provide users with prediction results, we implemented the prediction logic within the app. When users input their booking details through the web interface, the information is fed to the Random Forest model, which promptly returns the prediction on whether the booking is likely to be canceled or not.



*Figure 13. Model Deployed on Streamlit App*

# Prescriptive Measures

By implementing these prescriptive measures, hotels can effectively manage booking cancellations, maximize revenue, and enhance overall guest satisfaction. The data-driven approach provided by our predictive model, in conjunction with these measures, empowers hotels to make strategic decisions that optimize their operations and improve customer experiences.

1. **Proactive Customer Engagement:** For customers with a high likelihood of cancellation or longer lead times, the hotel can implement proactive outreach strategies. This may include personalized communication to offer incentives like price matching, room upgrades, complimentary services, or exclusive promotions. By engaging with these customers early on, the hotel can increase the chances of retaining their bookings and ensuring customer satisfaction.
2. **Non-Refundable Advance Payments:** To mitigate the risk of cancellations, hotels can encourage customers who are likely to cancel to make non-refundable advance payments at a discounted rate. This approach not only secures revenue for the hotel but also provides cost savings for customers who are committed to their bookings. It is essential to communicate the terms and benefits clearly to customers to make this option attractive.
3. **Designing Attractive Promotional Packages:** Hotels can design attractive promotional packages that combine room rates with additional services or amenities. By creating bundled offers tailored to specific market segments, hotels can entice guests to choose comprehensive packages that align with their preferences and needs. These comprehensive packages are less likely to be canceled as they offer added value to customers.
4. **Monitor and Adjust Pricing Strategies:** Regularly monitoring market trends and demand patterns can help hotels adjust their pricing strategies accordingly. Implementing dynamic pricing based on factors like demand, seasonality, and events can optimize revenue while encouraging customers to commit to their bookings. Offering flexible pricing options can also cater to various customer preferences.
5. **Strengthen Booking Policies:** Strengthening booking policies, particularly for high-demand periods or special events, can help reduce last-minute cancellations. Clear and transparent policies regarding cancellation deadlines and associated fees can incentivize customers to honor their reservations.

# Conclusion

In conclusion, this project successfully addressed the challenges associated with hotel reservation cancellations by employing data analysis, cluster modeling, and predictive modeling techniques. The analysis provided valuable insights into the factors influencing booking cancellations, helping hotel owners and stakeholders make informed decisions.

The exploratory data analysis revealed interesting patterns in the data, highlighting the impact of lead time, room type, market segment, and other attributes on booking cancellations. Using cluster analysis, distinct market segments were identified, enabling targeted strategies for customer engagement.

The predictive models, including Logistic Regression, Decision Tree, and Random Forest, were developed to predict booking cancellations. Among these, the Random Forest model emerged as the most effective, achieving high accuracy and recall. The deployment of the Random Forest model through Streamlit demonstrated its potential for real-time application and decision-making.

Furthermore, the project showcased the significance of balancing the imbalanced dataset using SMOTE to achieve more reliable predictions, especially when the focus was on minimizing false negatives.

The project's findings and prescriptive measures offer actionable insights for hotel management to optimize revenue, reduce cancellations, and enhance customer satisfaction. By proactively engaging customers, offering attractive promotions, and adjusting pricing strategies, hotels can boost customer loyalty and overall business performance.

In conclusion, this data-driven approach equips hotels with the tools to make data-informed decisions, ensuring a seamless booking experience and fostering long-lasting relationships with their guests. As the hospitality industry evolves, leveraging data analytics will continue to be instrumental in driving success and staying ahead in a competitive market.