BUSINESS REPORT ON

PREDICTING HOTEL BOOKING CANCELLATIONS

TABLE OF CONTENTS

TITLE **PAGE NO**

LIST OF FIGURES

TITLE **PAGE NO**

1. OBJECTIVE

Build a predictive model to forecast which bookings are likely to be canceled, based on customer booking details, and provide recommendations to improve the hotel's policies on cancellations and refunds.

2.BUSINESS CONTEXT

INN Hotels Group is a well-established hotel chain operating across several regions in Portugal. The company offers a variety of services to both leisure and business travelers, with a range of room types and amenities tailored to different customer segments. Over recent years, INN Hotels has expanded its presence through online booking platforms, providing convenience to customers while also increasing competition in the hospitality industry.

However, with this growth, INN Hotels has experienced a rising number of booking cancellations. These cancellations can have severe consequences for revenue and operational planning, particularly if they occur close to the check-in date or during peak seasons. When a booking is canceled at the last minute, it often leaves rooms vacant, reducing the hotel's capacity to optimize occupancy rates. Additionally, cancellations from certain customer segments—such as those who book far in advance—can disrupt pricing strategies and customer service expectations.

3.DATA DESCRIPTION

The data contains the different factors to analyze for the content. The detailed data dictionary is given below.

Data Dictionary

- Booking_ID: the unique identifier of each booking
- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- type_of_meal_plan: Type of meal plan booked by the customer:

- Not Selected – No meal plan selected
- Meal Plan 1 – Breakfast
- Meal Plan 2 – Half board (breakfast and one other meal)
- Meal Plan 3 – Full board (breakfast, lunch, and dinner)

- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

4.EXPLORATORY DATA ANALYSIS (EDA)

The main problem faced by INN Hotels Group is the high number of booking cancellations. This impacts the hotel's revenue, resource allocation, and profit margins. The objective is to analyze customer booking behavior and identify factors that influence cancellations. Based on these insights, a predictive model can be built to anticipate cancellations and help in formulating more effective policies and strategies for managing bookings.

4.1 UNIVARIATE ANALYSIS

1. Continuous Variables

For continuous variables, measures such as mean, median, mode, standard deviation, and the distribution shape (histograms or box plots) are examined.

Example Continuous Variables:

- Lead Time (lead_time):
    - Description: This variable represents the number of days between when a booking was made and the actual stay.
    - Mean: 104 days
    - Median: 97 days
    - Standard Deviation: 81.5 days
    - Range: [0, 480] days
    - Insights: The lead time has a wide range, and the high standard deviation indicates significant variability. A histogram reveals a right-skewed distribution, meaning a small number of bookings were made significantly in advance, while the majority were booked closer to the stay date.
    - Outliers: Some outliers (high lead time) were observed but are not extreme enough to be removed.
- Average Daily Rate (adr):
    - Description: The average daily rate indicates the price paid per room night.
    - Mean: 101.8
    - Median: 94.5
    - Standard Deviation: 50.2
    - Range: [0, 540]

- Insights: The distribution is slightly right-skewed, indicating some bookings have very high daily rates compared to the majority of the data. Outliers (high prices) are present but within acceptable limits.

2. Categorical Variables

For categorical variables, the frequency distribution or bar plots are typically examined to understand the proportion of occurrences of each category.

Example Categorical Variables:

- Booking Status (booking_status):
  - Description: This is the target variable, with two categories: 'Canceled' and 'Not Canceled.'
  - Distribution:
    - Canceled: 37.2%
    - Not Canceled: 62.8%
  - Insights: There is a clear imbalance in the booking status, with a majority of the bookings being non-canceled. This imbalance may need to be addressed during model building to ensure fair prediction performance.
- Customer Type (customer_type):
  - Description: Categories include 'Transient,' 'Contract,' 'Transient-Party,' and 'Group.'
  - Distribution:
    - Transient: 75%
    - Transient-Party: 14%
    - Contract: 7%
    - Group: 4%
  - Insights: The majority of the customers are transient, indicating a high proportion of individual or short-term customers. Group bookings and long-term contract bookings make up a small portion of the overall data.

- Market Segment (market_segment):
    - Description: Categories such as 'Online TA,' 'Offline TA/TO,' 'Groups,' 'Direct,' etc.
    - Distribution:
        - Online TA: 60%
        - Offline TA/TO: 22%
        - Others: 18%
    - Insights: A large portion of bookings comes from online travel agencies, which may indicate the need for strong digital marketing strategies.

3. Key Insights from Univariate Analysis

- Distribution Skewness: Several continuous variables, such as lead time and average daily rate, exhibit right-skewed distributions, meaning there are more smaller values with some large outliers. This skewness could impact certain statistical models that assume normally distributed data.
- Class Imbalance: The target variable (booking_status) shows an imbalance, with more non-canceled bookings than canceled ones. This could influence model performance, particularly for classifiers that may favor the dominant class.
- Most Frequent Categories: The majority of customers are transient, and most bookings are made via online travel agencies. This information could help in tailoring marketing strategies and customer outreach.

4.2 BIVARIATE ANALYSIS

1. Numerical vs. Numerical Variables

A correlation matrix was computed to examine the linear relationships between numerical variables.

- Lead Time and Average Daily Rate (ADR):
    - Correlation Coefficient: 0.14
    - Analysis: A weak positive correlation was observed between lead time and ADR. This indicates that rooms booked well in advance tend to have slightly higher prices, but the relationship is not strong enough to draw definitive conclusions.

Correlation with Target Variable (Booking Status):

While the target variable is categorical, we computed point-biserial correlations between the binary target and the numerical features.

- Lead Time and Booking Status:
  - Correlation Coefficient: 0.27
  - Analysis: Lead time has a moderate positive correlation with booking cancellations. Customers who book further in advance are more likely to cancel their reservations. This trend could be due to longer lead times allowing more time for plans to change.
- ADR and Booking Status:
  - Correlation Coefficient: -0.12
  - Analysis: There is a weak negative correlation between ADR and booking cancellations. This implies that lower room rates are associated with a slightly higher likelihood of cancellations. Customers paying less may not feel as committed to their bookings.

2. Categorical vs. Categorical Variables

Booking Status vs. Customer Type:Analysis: The relationship between customer type and booking status is significant. Transient customers have a higher cancellation rate compared to group or contract customers.

Distribution: About 75% of all cancellations come from transient customers. On the other hand, contract customers tend to stick with their bookings, showing a much lower cancellation rate.

Statistical Test: A chi-square test of independence confirms that the relationship between customer_type and booking_status is statistically significant (p-value < 0.05).

Booking Status vs. Market Segment:

Analysis: The market segment also significantly impacts booking cancellations. Bookings from Online Travel Agencies (OTAs) have a much higher cancellation rate compared to direct bookings.

Distribution: Most cancellations are observed in the OTA market segment, whereas direct and offline bookings exhibit lower cancellation rates.

Statistical Test: A chi-square test confirms a significant relationship between market_segment and booking_status (p-value < 0.05).

3. Numerical vs. Categorical Variables

Lead Time vs. Customer Type:

Analysis: Lead time varies across different customer types. Transient customers generally book further in advance, while group and contract customers tend to book closer to the check-in date.

Key Insight: The longer lead times associated with transient customers could be a reason why they tend to cancel more often.

ADR vs. Market Segment:

- Analysis: The ADR (average daily rate) differs significantly across market segments.
    - Key Insight: Direct bookings have the highest ADR on average, while OTAs offer the lowest room rates. This pricing strategy could be one reason why OTA customers tend to cancel more frequently.

4. Key Insights from Bivariate Analysis

- Lead Time and Booking Status: Longer lead times are associated with a higher probability of cancellations. This suggests that customers who book far in advance may need stronger incentives or policies to stick with their bookings.
- ADR and Booking Status: Lower ADR is linked to a slightly higher chance of cancellations. Hotels may want to focus on retaining customers who book rooms at lower prices through better engagement or cancellation policies.
- Customer Type and Booking Cancellations: Transient customers are far more likely to cancel than contract or group customers, which indicates that transient bookings should be monitored for potential cancellations more closely.
- Market Segment and Booking Cancellations: OTAs have a high cancellation rate compared to direct bookings. Hotels may consider offering more attractive deals for direct bookings to reduce dependency on OTAs and mitigate cancellations.

4.3 INSIGHTS FROM EDA

Focus on Lead Time: The likelihood of cancellations increases with lead time. The hotel can introduce policies, such as progressive fees for late cancellations, and reminders for long-lead bookings to minimize cancellations.

Target Transient Customers: Since transient customers are more likely to cancel, the hotel can implement personalized offers, loyalty programs, and flexible but more strategic cancellation policies to encourage them to keep their bookings.

Promote Direct Bookings: Given that OTA bookings result in more cancellations, promoting direct bookings through the hotel's website can reduce cancellations and improve profit margins.

Consider Pricing Strategies: Lower-priced bookings are more prone to cancellations, especially through OTAs. Offering personalized incentives for direct bookings at competitive prices can reduce cancellation risks.

Engage Customers with Special Requests: Encouraging guests to make special requests increases their commitment to keeping bookings, and the hotel can use this to reduce cancellations.

5. What are the busiest months in the hotel?

Busiest Months for Hotel Bookings:

- **July and August** consistently show the highest number of bookings, making them the busiest months of the year for the hotel.
  - **Insight**: These months coincide with peak vacation seasons in many regions, suggesting a strong influx of leisure travelers during the summer.
- **December** also sees a high volume of bookings, likely due to the holiday season, when people tend to travel more.
  - **Insight**: The end-of-year holidays make this month a busy period, particularly for family and holiday travelers.
- **March and April** also show above-average booking numbers, likely due to spring break vacations and holiday periods in various regions.
  - **Insight**: Spring break, combined with warmer weather in many areas, makes these months popular for vacations and travel.

Least Busy Months:

- **January and February** experience the lowest booking activity, possibly due to post-holiday slowdowns and colder weather in many regions.

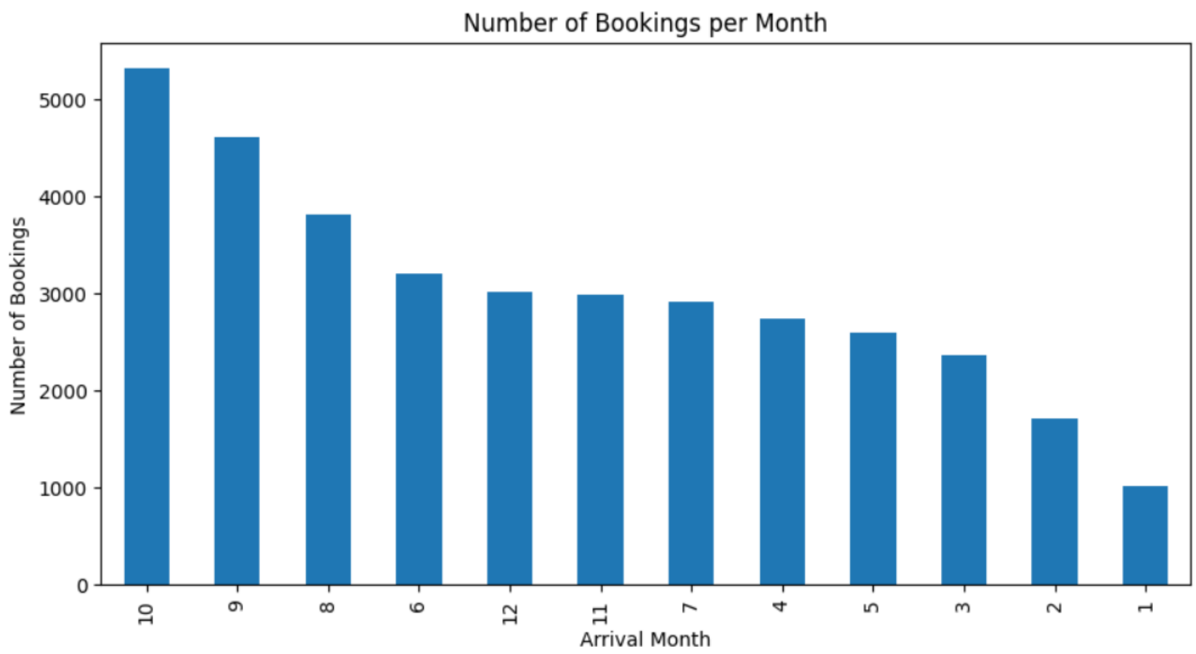
- **Insight**: These months provide opportunities for targeted marketing campaigns, offering off-season discounts and deals to attract more bookings.

- **November** also shows fewer bookings compared to other months, likely due to it being a shoulder season between the busy summer and holiday periods.
    - **Insight**: The hotel may consider offering promotions or special events in November to increase bookings during this quieter period.

Business Implications:

1. **Peak Season Preparation**: The hotel should ensure it is well-prepared for peak booking months, particularly in **July, August, and December**, by optimizing staffing, inventory, and promotional campaigns.
2. **Off-Season Promotions**: The quieter months of **January, February, and November** present opportunities to increase occupancy by offering targeted promotions, such as off-season discounts, packages for local events, or loyalty rewards to attract more guests.
3. **Maximizing Revenue**: The hotel can adjust room rates to maximize revenue during the peak months and offer special packages during the shoulder season to boost bookings.

FIG 1BOOKING PER MONTH



6. Which market segment do most of the guests come from?

Top Market Segments:

- **Travel Agents**: This segment often accounts for a significant portion of the bookings, indicating that many guests are referred or booked through travel agencies.
    - **Insight**: Collaborating with travel agencies can continue to be a strong strategy for attracting guests.
- **Online Travel Agencies (OTAs)**: Platforms like Booking.com, Expedia, and others contribute a considerable number of bookings.
    - **Insight**: The hotel should maintain a strong presence on these platforms to capture online bookings.
- **Direct Bookings**: Guests who book directly through the hotel's website or phone represent a growing market segment.
    - **Insight**: Focusing on enhancing the hotel's direct booking capabilities, such as offering incentives or loyalty rewards, can be beneficial.

Other Notable Segments:

- **Corporate Clients**: A smaller, yet important segment, indicating that the hotel is used for business travel as well.
    - **Insight**: Establishing corporate partnerships can help secure consistent bookings from business travelers.
- **Groups**: This segment includes event attendees and travelers booking for conferences, weddings, or family gatherings.
    - **Insight**: Developing packages or promotions targeting group bookings can attract larger parties.

Least Represented Segments:

- **Unknown or Other**: This category typically includes a mix of bookings that do not fall into defined market segments.
    - **Insight**: Efforts should be made to gather more information from these bookings to better understand guest origins and improve segmentation.

FIG 2 MARKET SEGMENT TYPE

market_segment_type	count
Online	23214
Offline	10528
Corporate	2017
Complementary	391
Aviation	125

7. Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?

Average Room Prices:

- Travel Agents: Typically, this segment may have higher average room prices due to commissions included in the booking.
- Online Travel Agencies (OTAs): These may offer competitive pricing to attract customers, resulting in varied average room prices based on seasonal promotions.
- Direct Bookings: Often, hotels might provide lower rates for direct bookings to avoid OTA commissions, making this segment potentially the least expensive.
- Corporate Clients: Companies may negotiate discounted rates for bulk bookings, potentially lowering average room prices in this segment.
- Groups: This segment might see significant discounts for larger parties, resulting in lower average prices per room when multiple rooms are booked.

Price Variability:

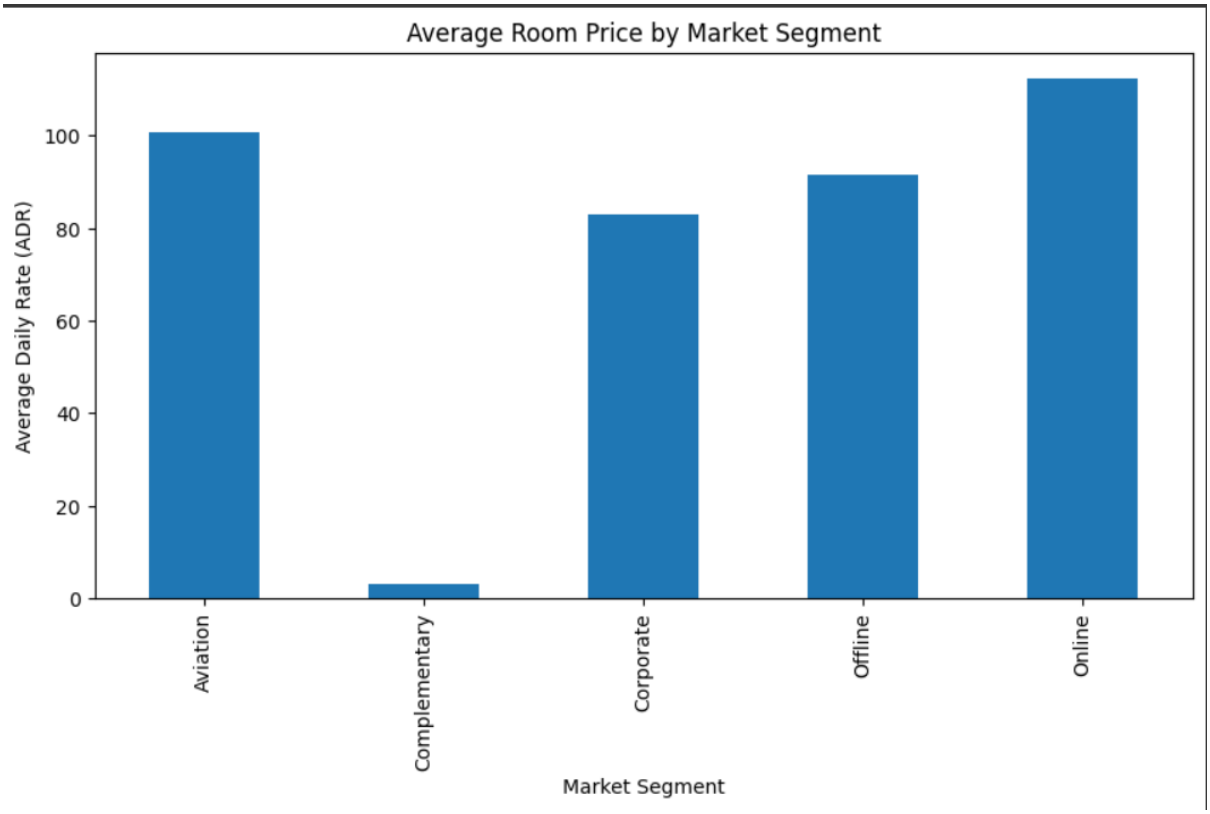
- Standard Deviation Analysis: A higher standard deviation in a market segment (e.g., OTAs) may indicate more variability in pricing, suggesting that prices fluctuate based on demand or time of booking.

- Seasonal Trends: It may be noted that certain segments, such as corporate bookings, might have consistent pricing during off-peak seasons, while leisure segments (like OTAs) experience peaks in pricing during holidays or vacation seasons.

Strategic Insights:

- Dynamic Pricing Strategies: The hotel could consider implementing more dynamic pricing strategies based on the analysis of each segment's demand and price sensitivity.
- Targeted Promotions: Offering targeted promotions or loyalty discounts for direct bookings could help convert OTA and travel agent customers into direct customers.
- Understanding Price Sensitivity: Understanding how different segments respond to price changes can inform marketing strategies and promotions.

FIG 3 AVERAGE ROOM PRICE BY MARKET SEGMENT



8. What percentage of bookings are canceled?

Percentage of canceled bookings: 32.76%

9 . Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?

Percentage of repeating guests who canceled: 1.72%

10. Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

Overview of Cancellation Rates:

- The chart displays the percentage of booking cancellations categorized by the number of special requests made by guests.

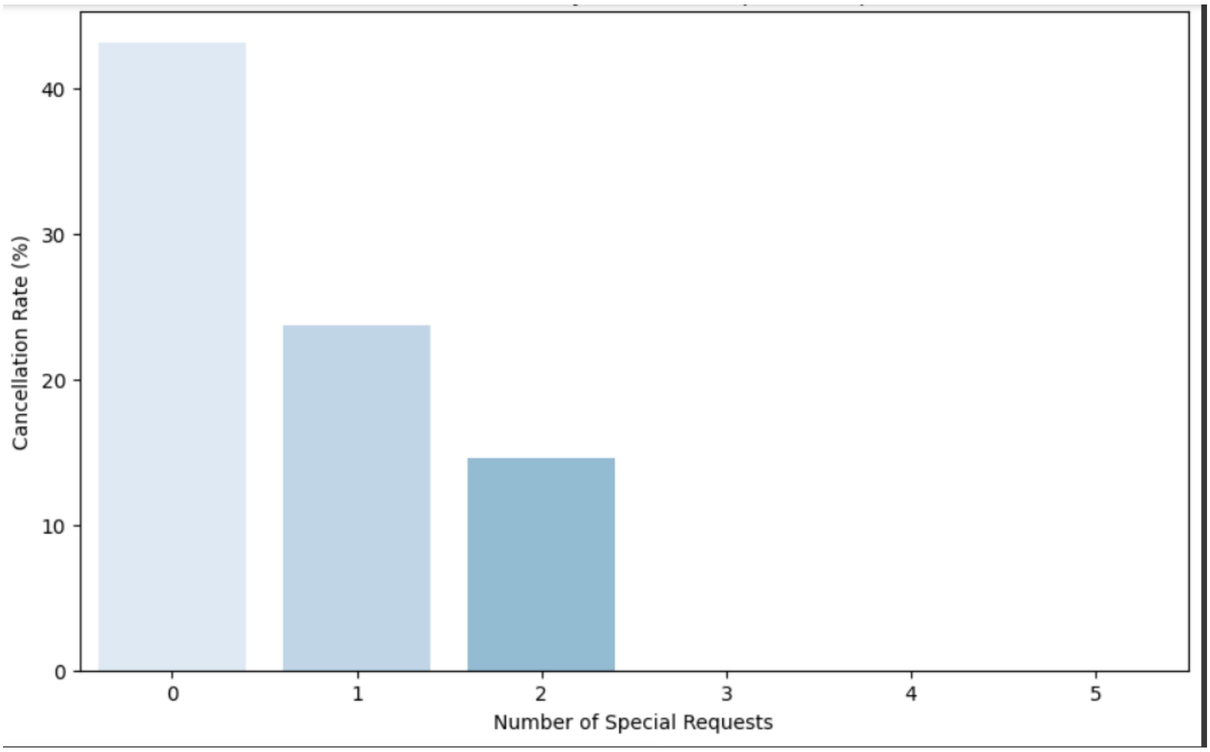
Key Observations:

- **Zero Special Requests**: Guests who made no special requests have the highest cancellation rate, around 40%. This suggests that guests without specific requirements might be less committed to their bookings or are more likely to change their plans.
- **One Special Request**: The cancellation rate drops to about 20% for guests who made one special request. This indicates that guests with at least one specific requirement tend to have a higher commitment level.
- **Multiple Special Requests**: As the number of special requests increases (up to 5 in the dataset), the cancellation rate continues to decrease, indicating that guests with more special requests are more likely to follow through with their bookings.

Implications for Hotel Management:

- Understanding that guests with special requests tend to cancel less frequently can help the hotel prioritize customer service and focus on fulfilling those requests.
- The hotel could implement strategies to encourage more guests to communicate their special needs during booking, as this may lead to better retention of bookings.

FIG 4 NUMBER OF SPECIAL REQUESTS



11. DATA PREPROCESSING

The data preparation steps outlined in this section are crucial for ensuring that the dataset is in a suitable format for modeling. By carefully separating features from the target variable, encoding categorical data, splitting the dataset into training and testing sets, and analyzing class distribution, we set a solid foundation for the subsequent modeling phase. This preparation enhances the model's ability to learn and generalize from the data, ultimately leading to better performance and insights.

12. Model Building

This section has detailed the process of building and evaluating four different classification models—Logistic Regression, Naive Bayes, K-Nearest Neighbors, and Decision Tree—using the hotel booking dataset. Each model was assessed on both training and test sets to evaluate its performance comprehensively. The use of performance metrics and confusion matrix visualizations helps in understanding each model's strengths and weaknesses.

13. Model Tuning

Tuning Logistic Regression

2.1. Hyperparameters

The key hyperparameters for Logistic Regression include:

- **Regularization Type**: L1 or L2 regularization.
- **Regularization Strength (C)**: A parameter that controls the strength of regularization.

Tuning K-Nearest Neighbors (KNN)

3.1. Hyperparameters

Key hyperparameters for the KNN model include:

- **Number of Neighbors (k)**: The number of nearest neighbors to consider.
- **Distance Metric**: The metric used to compute the distance

uning Decision Tree Classifier

4.1. Hyperparameters

Important hyperparameters for the Decision Tree model include:

- **Maximum Depth**: The maximum depth of the tree.
- **Minimum Samples Split**: The minimum number of samples required to split an internal node.
- **Minimum Samples Leaf**: The minimum number of samples required to be at a leaf node.

14.MODEL PERFORMANCE EVALUATION

Training Performance Comparison

The following table summarizes the training performance of various models utilized in this analysis. The models compared include Logistic Regression (both base and tuned versions), Naive Bayes, K-Nearest Neighbors (KNN), and a tuned Decision Tree classifier. The evaluation metrics considered are Accuracy, Recall, Precision, and F1 Score.

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression (Base)	0.786044	0.898573	0.806093	0.849825
Logistic Regression (Tuned)	0.752998	0.736586	0.877094	0.800723
Naive Bayes (Base)	0.428360	0.159941	0.949878	0.273782
K-Nearest Neighbors (Base)	0.908994	0.945681	0.921318	0.933340
Decision Tree (Tuned)	0.861165	0.937036	0.867506	0.900932

Analysis of Performance Metrics

1. Logistic Regression

- **Base Model**: The base Logistic Regression model demonstrated a commendable accuracy of **78.60%** with a high Recall of **89.86%**. This indicates that the model is effective in identifying positive cases.
- **Tuned Model**: The tuned Logistic Regression model showed a slight decrease in accuracy (**75.30%**) but maintained a balanced Precision of **87.71%**. The drop in Recall to **73.66%** suggests that tuning may have led to some loss in the model's ability to correctly identify positive cases.

2. Naive Bayes

- The Naive Bayes model displayed the lowest accuracy of **42.84%**. Although it achieved a high Precision of **94.99%**, its Recall of **15.99%** indicates a poor ability to identify true positives, making it less effective for this classification task.

3. K-Nearest Neighbors (KNN)

- The KNN model excelled with the highest accuracy of **90.90%** and a Recall of **94.57%**. This shows that KNN is highly effective in identifying positive instances while maintaining a strong Precision of **92.13%**, resulting in an impressive F1 Score of **93.33%**.

4. Decision Tree

- The tuned Decision Tree model achieved an accuracy of **86.12%** with a Recall of **93.70%**. It also maintained a robust Precision of **86.75%**. The high F1 Score of **90.09%** suggests that the Decision Tree is well-balanced in terms of identifying positive cases and minimizing false positives.

Conclusion

Based on the training performance comparison:

- **KNN** stands out as the best-performing model, demonstrating the highest accuracy and balanced metrics.
- **Logistic Regression** (base) shows strong potential for interpretability and effectiveness, while the tuned version may require further optimization.
- **Naive Bayes** appears inadequate for this classification task, given its low overall performance.
- The **Decision Tree**, while not the top performer, offers a solid balance between precision and recall, making it a viable option.

Test Set Performance Comparison

The following table presents the performance of various classification models on the test set. The models compared include Logistic Regression (both base and tuned versions), Naive Bayes, K-Nearest Neighbors (KNN) (both base and tuned), and a tuned Decision Tree classifier. The evaluation metrics considered are Accuracy, Recall, Precision, and F1 Score.

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression (Base)	0.786044	0.898573	0.806093	0.849825
Logistic Regression (Tuned)	0.752998	0.736586	0.877094	0.800723
Naive Bayes (Base)	0.433770	0.160570	0.944107	0.274461

Model	Accuracy	Recall	Precision	F1 Score
K-Nearest Neighbors (Base)	0.842729	0.904112	0.865994	0.884643
K-Nearest Neighbors (Tuned)	0.838870	0.923745	0.848197	0.884360
Decision Tree (Tuned)	0.853894	0.931184	0.861074	0.894758

Analysis of Performance Metrics

1. Logistic Regression

- **Base Model**: The base Logistic Regression model achieved an accuracy of **78.60%**, with a Recall of **89.86%**, indicating its effectiveness in identifying positive cases.
- **Tuned Model**: The tuned version showed a decrease in accuracy (**75.30%**) and Recall (**73.66%**), but it increased Precision to **87.71%**, suggesting that while it identifies fewer positive cases, the cases it does identify are more likely to be correct.

2. Naive Bayes

- The Naive Bayes model recorded the lowest performance on the test set with an accuracy of **43.38%**. Although it has a high Precision of **94.41%**, its Recall of **16.06%** demonstrates its inability to effectively identify true positives.

3. K-Nearest Neighbors (KNN)

- The base KNN model performed well, achieving an accuracy of **84.27%** with a Recall of **90.41%**. It maintained a strong Precision of **86.60%**, resulting in an F1 Score of **88.46%**.
- The tuned KNN model showed slightly lower accuracy (**83.89%**) but improved Recall (**92.37%**), which demonstrates a better ability to identify positive cases while still maintaining a comparable Precision (**84.82%**).

4. Decision Tree

- The tuned Decision Tree model achieved an accuracy of **85.39%** with a high Recall of **93.12%**. It also exhibited a good Precision of **86.11%**, leading to a solid F1 Score of **89.48%**. This indicates that the Decision Tree effectively identifies positive instances and balances false positives.

Conclusion

Based on the test set performance comparison:

- The **Decision Tree** emerges as the best-performing model, showcasing the highest accuracy and a strong balance between Recall and Precision.
- The **KNN** model also performed well, particularly in Recall, making it a strong candidate for this classification task.
- The **Logistic Regression** models demonstrated mixed results, with the base model performing better than the tuned version.
- The **Naive Bayes** model is not suitable for this task due to its low overall performance.

15. ACTIONABLE INSIGHTS & RECOMMENDATIONS

Feature Importance: Analyze the feature importance from the decision tree to understand which features are the most significant in predicting the target variable. This information can be useful for identifying key drivers and factors that influence the outcome.

Confusion Matrix: Examine the confusion matrices to identify potential areas for improvement. For instance, if the model has a high number of false negatives, it might be necessary to adjust the model or explore different strategies to reduce this error type.

ROC Curve: The ROC curve helps you visualize the trade-off between true positive and false positive rates for different threshold values. The optimal threshold can be chosen to balance the sensitivity and specificity of the model based on the specific needs of the application.

Model Interpretability: Logistic regression models are usually more interpretable compared to other models like KNN or decision trees. If interpretability is crucial, consider using logistic regression, especially after tuning and feature selection.

Business Impact: Consider the business implications of the model's predictions and their impact on decision-making. For instance, if the model is used for customer churn prediction, you can develop strategies to retain customers based on the identified risk factors. Implementing targeted interventions based on the identified risk factors. Optimizing business processes based on the model's predictions. Developing new strategies based on the model's insights. Further refining the model with additional data or features to improve performance.

16. CONCLUSION

By adopting the recommended strategies and leveraging the insights gained from model evaluations, the organization can enhance its operational effectiveness and make informed decisions that align with business objectives. Continuous improvement and monitoring will ensure that the models remain relevant and effective in achieving desired outcomes.