Machine Learning - 1 : INN Hotels

Jesmi George

Contents

List of Figures

List of Tables

1 BUSINESS PROBLEM

1.1 CONTEXT

Many hotel bookings are canceled or result in no-shows, often due to plan changes or scheduling conflicts. While free or low-cost cancellations benefit guests, they reduce hotel revenue—especially for last-minute cancellations. Online bookings have increased this challenge, leading to revenue loss, higher costs, lower profit margins, and wasted staff effort.

1.2 OBJECTIVE

INN Hotels Group in Portugal faces high booking cancellations and seeks a data-driven solution. Analyze factors influencing cancellations, build a predictive model to identify likely cancellations in advance, and recommend profitable cancellation and refund policies more shorten. Key Questions to be answered :

1. What are the busiest months in the hotel?
2. Which market segment do most of the guests come from?
3. Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?
4. What percentage of bookings are canceled?
5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?
6. Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

1.3 DATA OVERVIEW

- The dataset has 36275 rows and 19 columns.
- No duplicate records were found
- There are no null values in any columns
- There are 14 features of numeric type, rest 5 of them are objects
- *Booking_ID*, *Booking_Status* features are features of object type
- *required_car_parking_space* is numeric type in dataset but actually it denotes whether a customer requires car parking space or not, ie, Yes/No

- *repeated_guest* is numeric type in dataset but actually it denotes if the customer a repeated guest or not, ie, Yes/No

1.4 DATA DICTIONARY

The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

Field Name	Description
Booking_ID	Unique identifier for 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
required_car_parking_space	Whether parking was required (0 = No, 1 = Yes)
room_type_reserved	Type of room reserved
lead_time	Days between booking and arrival
arrival_year	Year of arrival
arrival_month	Month of arrival
arrival_date	Date of the month
market_segment_type	Market segment designation
repeated_guest	Whether the guest is a repeat customer (0 = No, 1 = Yes)
no_of_previous_cancellations	Number of previous cancellations 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 room price per day (in euros)
no_of_special_requests	Total number of special requests
booking_status	Indicates if the booking was canceled or not

TABLE 1: DATA DESCRIPTION

2 EDA

2.1 UNIVARIATE ANALYSIS

Booking Status:

- 67% of the bookings have not been cancelled.
- This shows a class imbalance of target variable in dataset.

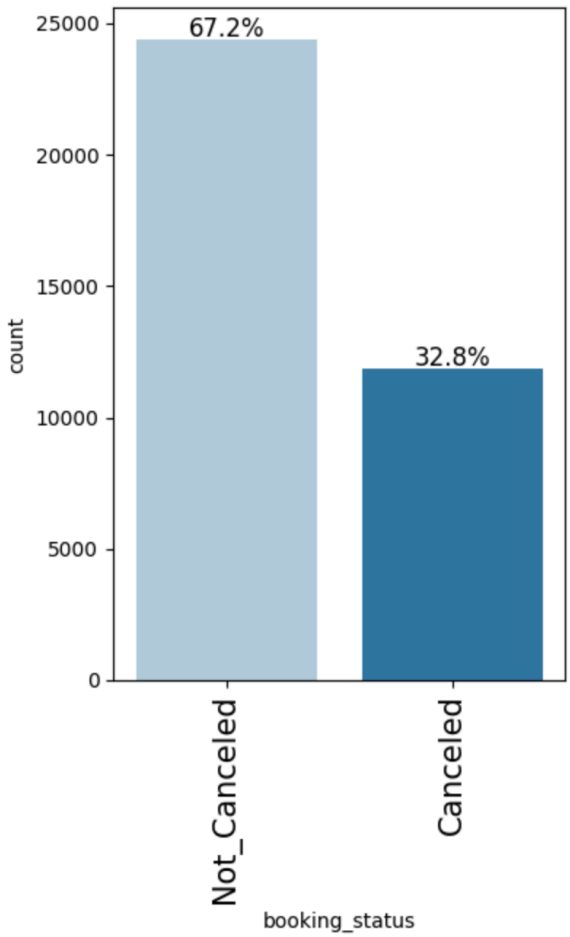

FIGURE 1: BOOKING STATUS DISTRIBUTION

Repeated Guests:

- 97% of the bookings are of new customers, only ~3% of customers book the inn again.

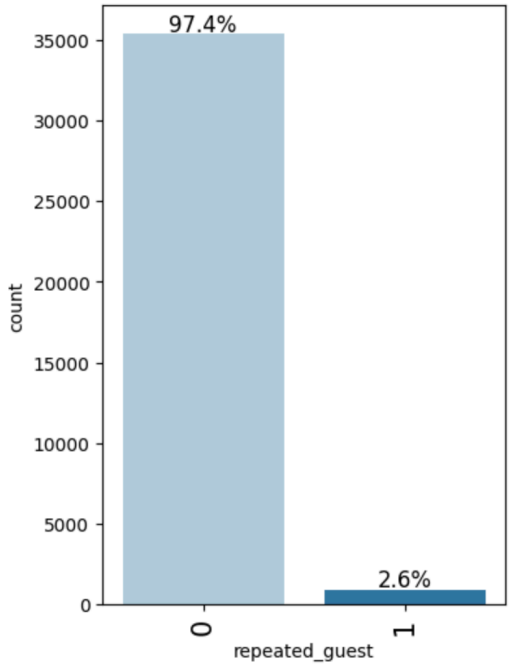

FIGURE 2: REPEATED GUEST DISTRIBUTION

Market Segment:

- 64% of the bookings are done online followed by offline bookings that account for 29%.

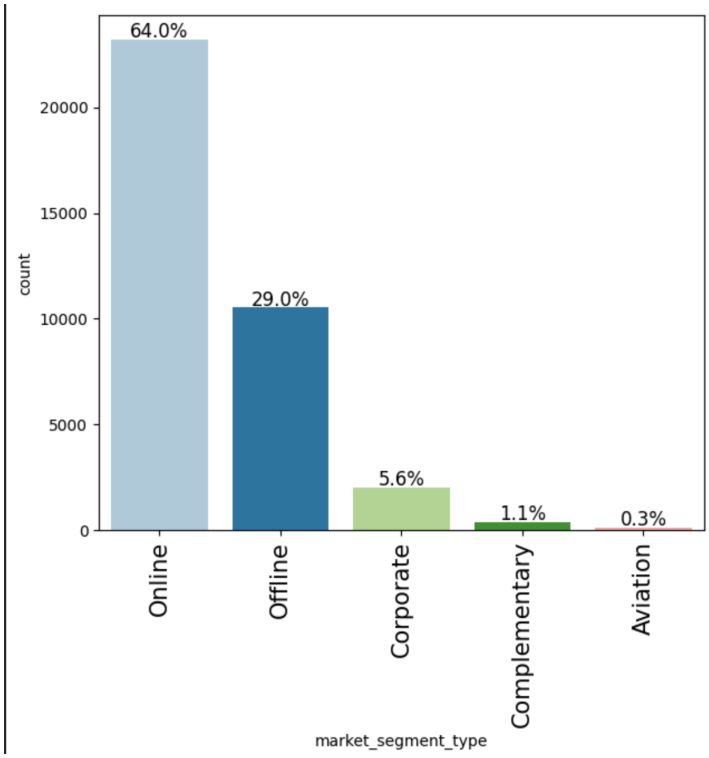

FIGURE 3: MARKET SEGMENT TYPE DISTRIBUTION

Arrival month:

- The most popular time for guests to arrive is in the second half of the year, particularly from July through December, which coincides with the fall and winter seasons.
- The most popular month for visitors is October, followed by September. This is because in September–October, Portugal's weather shifts to mild and pleasant autumn conditions.
- The festive season in November and December also seems to attract a high number of guests.

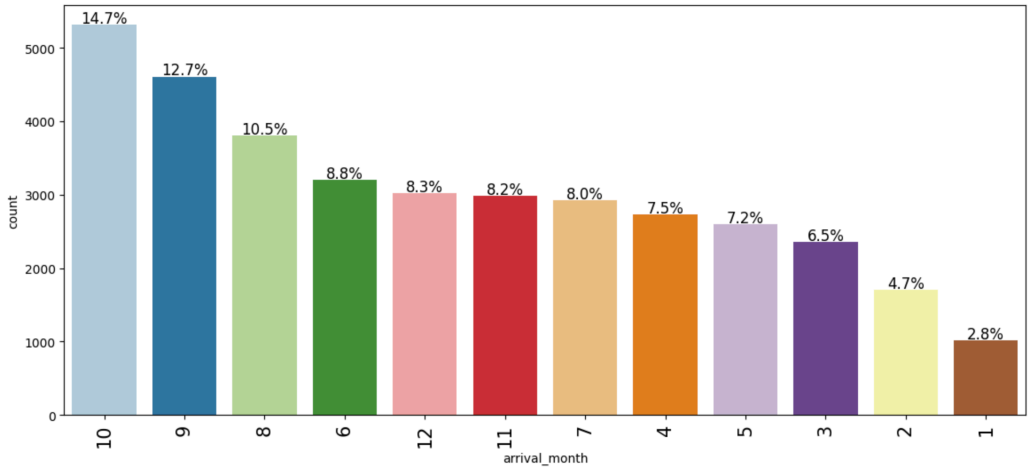

FIGURE 4: GUEST ARRIVALS BY MONTH

Type of Meal Plan:

- Majority of the guests preferred **Meal Plan 1**, followed by **Meal Plan 2**.
- Very few customers chose to book without a meal plan, indicating that most guests opt for meal-inclusive stays.

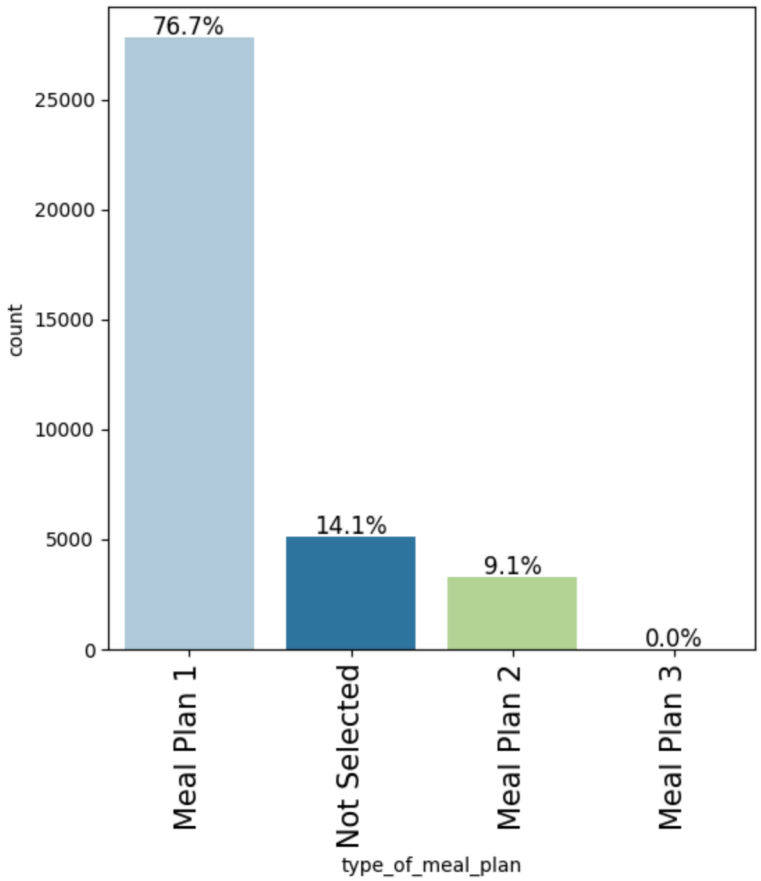

FIGURE 5: DISTRIBUTION OF TYPE OF MEAL PLAN

Required Car Parking Space:

- Most guests did not require a car parking space, suggesting that a large proportion of guests are **non-driving or local travelers**.
- Only a small fraction of bookings requested parking, possibly from guests driving in from other cities.

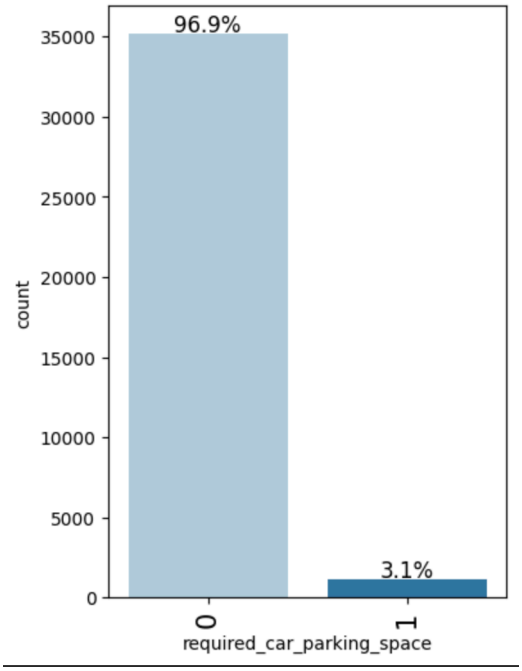

FIGURE 6: DISTRIBUTION OF REQUIRED CAR PARKING SPACE

Room Type Reserved:

- **Room Type 1** dominates the bookings, followed by a gradual decline in other room types.
- This indicates that Room Type 1 is the **most preferred or budget-friendly option** among guests.

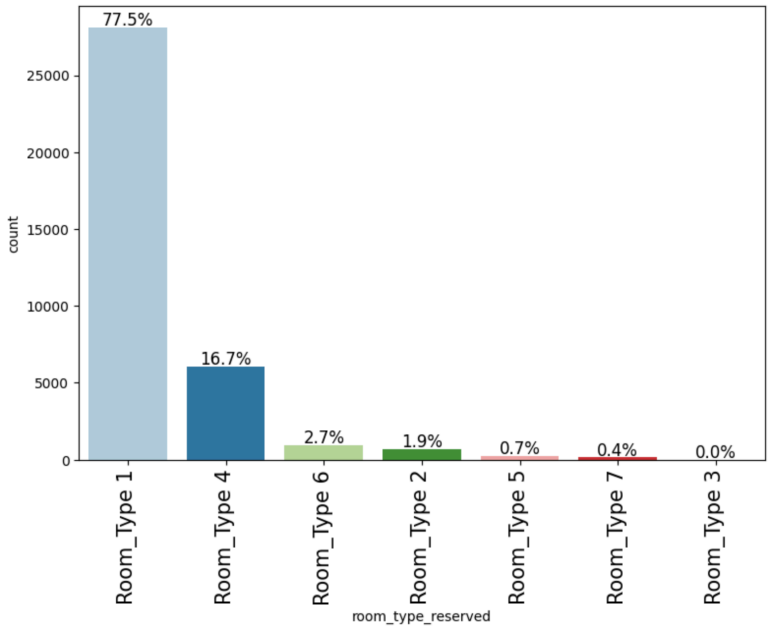

FIGURE 7: DISTRIBUTION OF ROOM TYPE RESERVED

Arrival Year:

- The data spans multiple years, mainly **2017 and 2018**.
- Booking volume remains fairly consistent across these years, showing a stable demand for hotel stays.

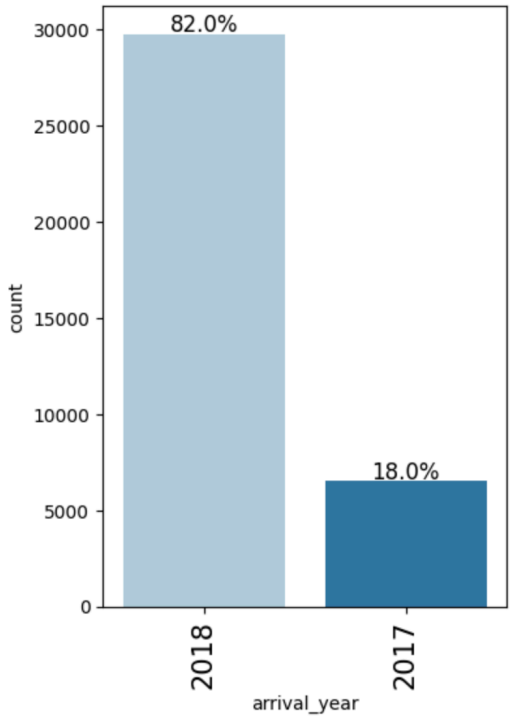

FIGURE 8: DISTRIBUTION OF ARRIVAL YEAR

Lead Time:

- Most customers made their bookings within **100 days of arrival**.
- A small number of customers booked far in advance, resulting in a **right-skewed distribution**.
- This indicates a mix of both short-term planners and early bookers.

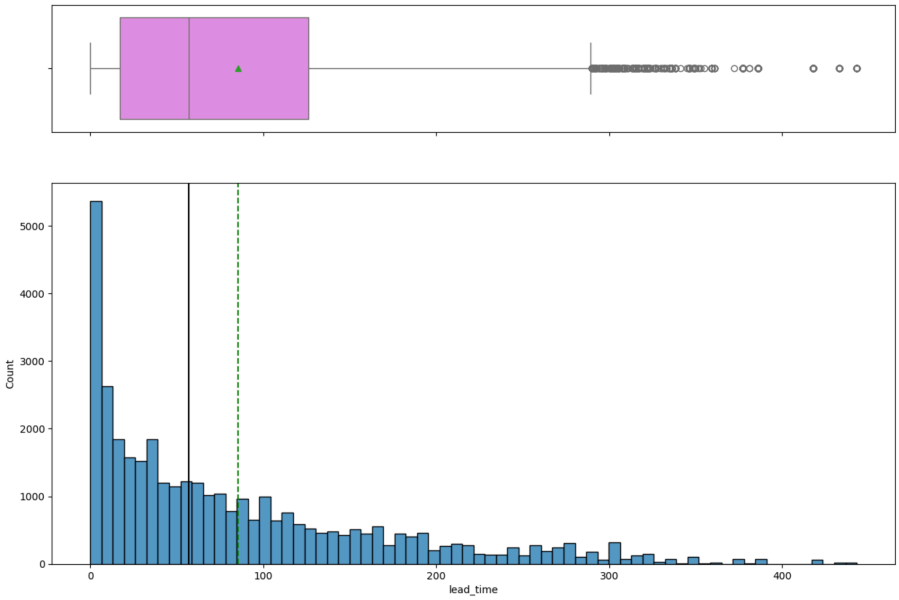

FIGURE 9: DISTRIBUTION OF LEAD TIME

Average Price per Room:

- The distribution of `avg_price_per_room` is close to normal, as the mean and median are in close range.
- Majority of rooms are priced below **200 euros**.
- Some outliers exist with room prices reaching up to **500 euros**.
- A few records show **0 euros** as the average price, which is unrealistic, but these likely belong to the **complimentary market segment**, where rooms or services are offered at no cost.

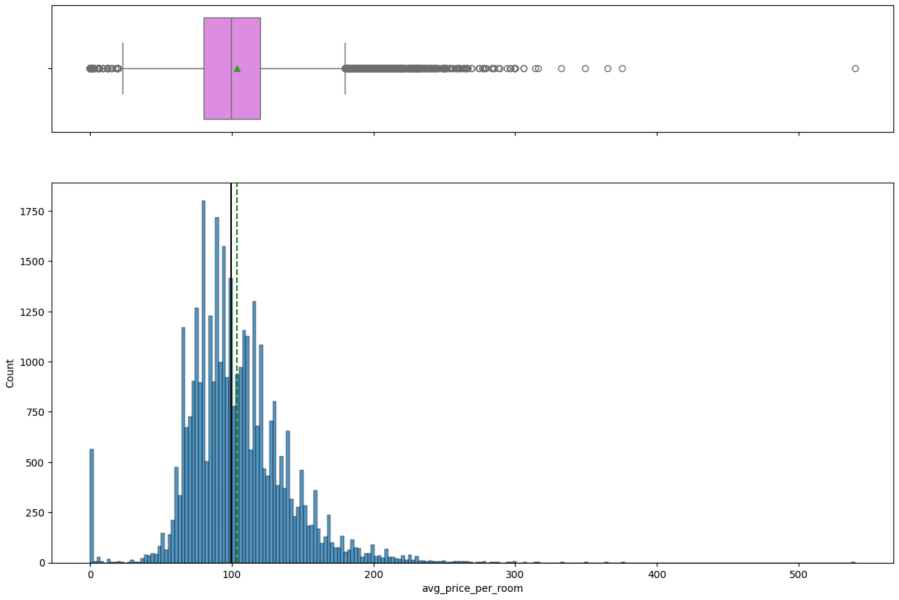

FIGURE 10: DISTRIBUTION OF AVERAGE PRICE PER ROOM

Number of Special Requests:

- Nearly half of the bookings(54.5%) don't have any special requests
- Only 43% of bookings make 1 or 2 special requests

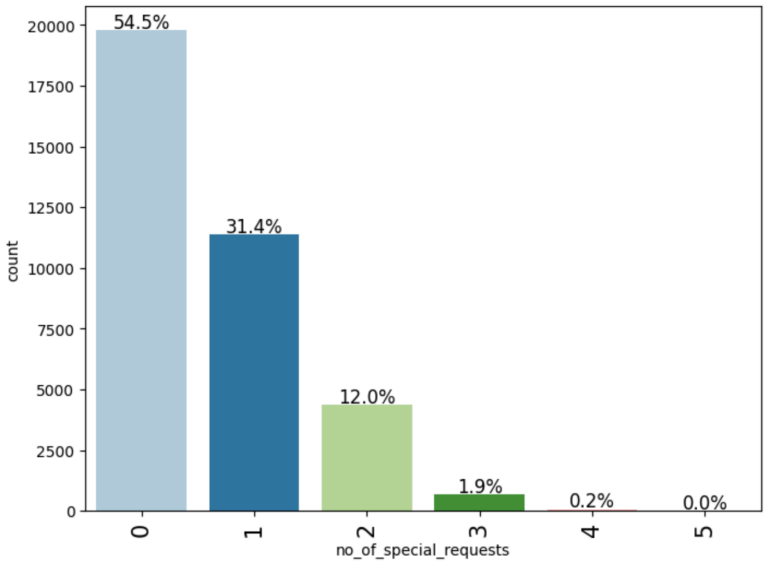

FIGURE 11: DISTRIBUTION OF NUMBER OF SPECIAL REQUESTS

Number of Previous Cancellations:

- Majority of guests have **no prior cancellations**.
- A few customers show repeated cancellations, which may indicate risky booking behavior.

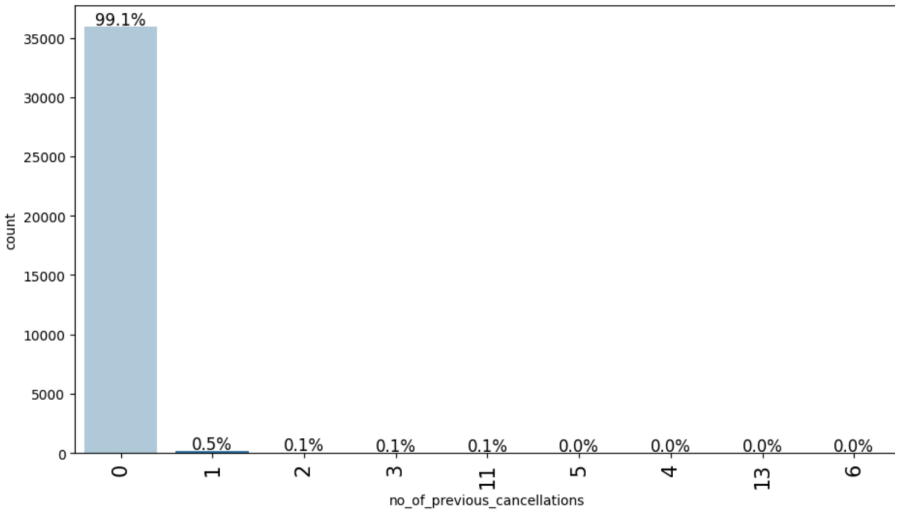

FIGURE 12: DISTRIBUTION OF NUMBER OF PREVIOUS CANCELLATIONS

Number of Adults:

- 72% of the bookings have only 2 adults.Rest of the bookings having 1 or 3 adult.Indicating that the stays are only for a short group of people
- About 0.4% of the bookings have zero adults which is an anomaly as hotel checking legally needs to done by adults

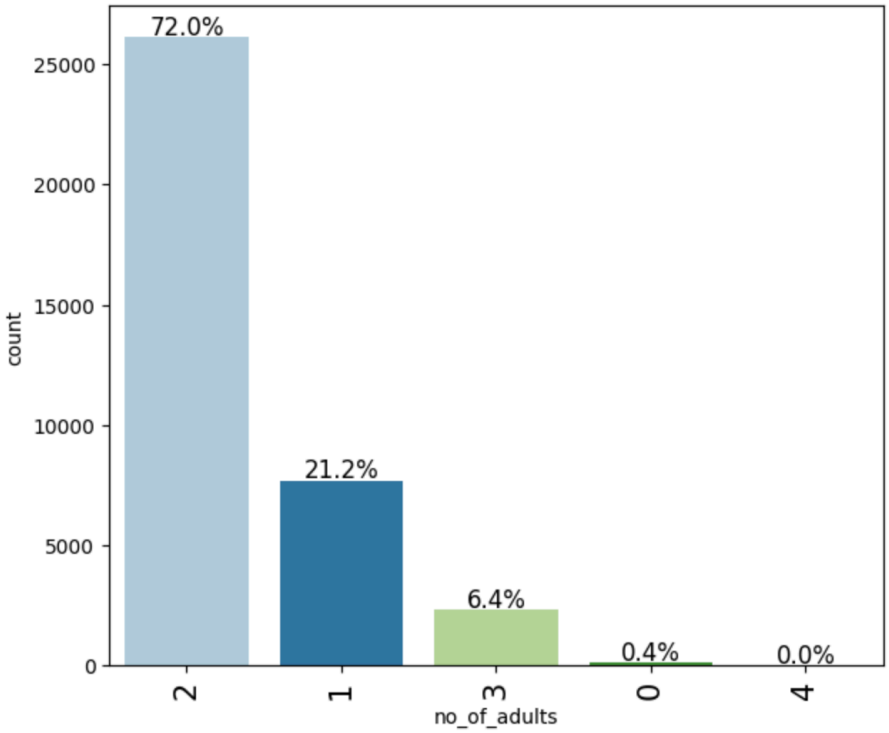

FIGURE 13: DISTRIBUTION OF NUMBER OF ADULTS

Number of Children:

- 93% of the guests don't bring in children with them

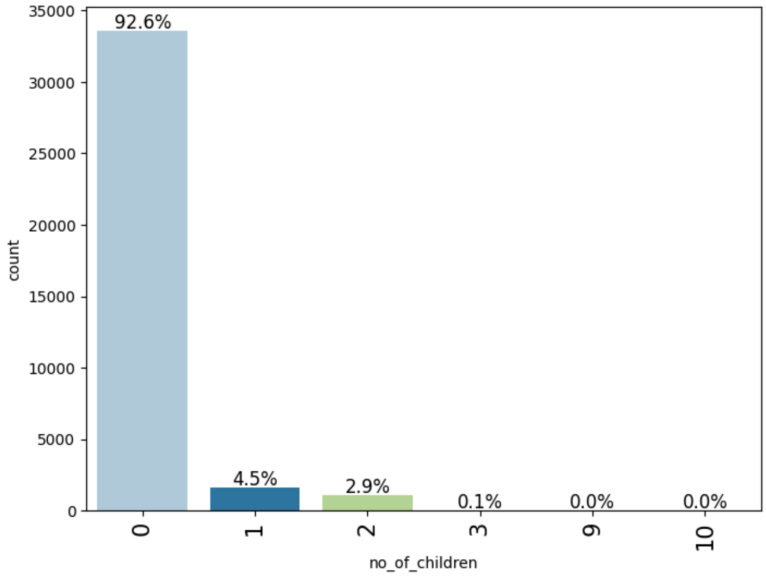

FIGURE 14: DISTRIBUTION OF NUMBER OF CHILDREN

Number of Weekend Nights:

- Nearly half of all bookings (**46.5%**) do not include a weekend stay.
- About half (~**53.5%**) include weekend stays, typically lasting up to **2 nights**, indicating guests on short trips.
- Very few bookings (**less than 1% combined**) extend beyond two weekend nights.

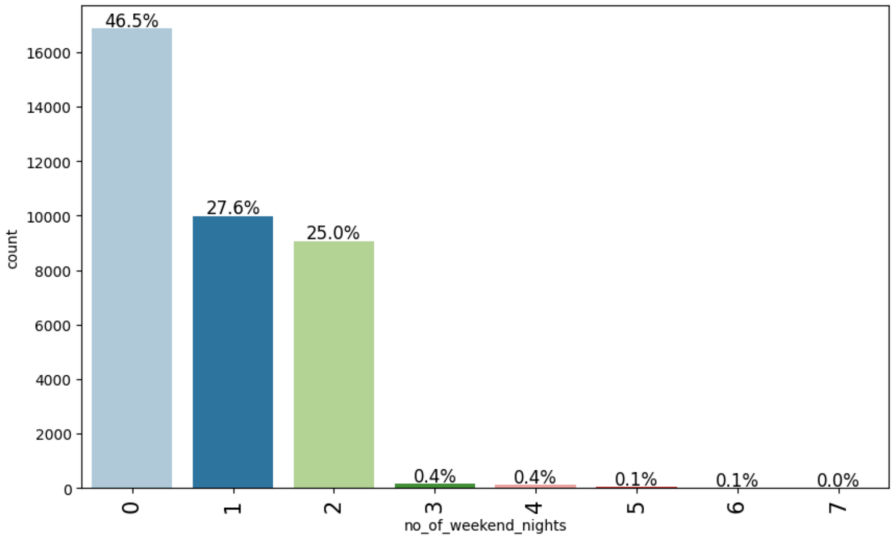

FIGURE 15: DISTRIBUTION OF NUMBER OF WEEKEND NIGHTS

Number of Week Nights:

- Most weekday bookings (**31.5%**) are for **2 nights**, followed by 1–3 nights, indicating a preference for short stays.

- Less than **2%** of bookings are for stays longer than **5 weekday nights**.
- About ~**6.6%** of bookings have **0 weekday nights**, suggesting customers who opted only for weekend stays.

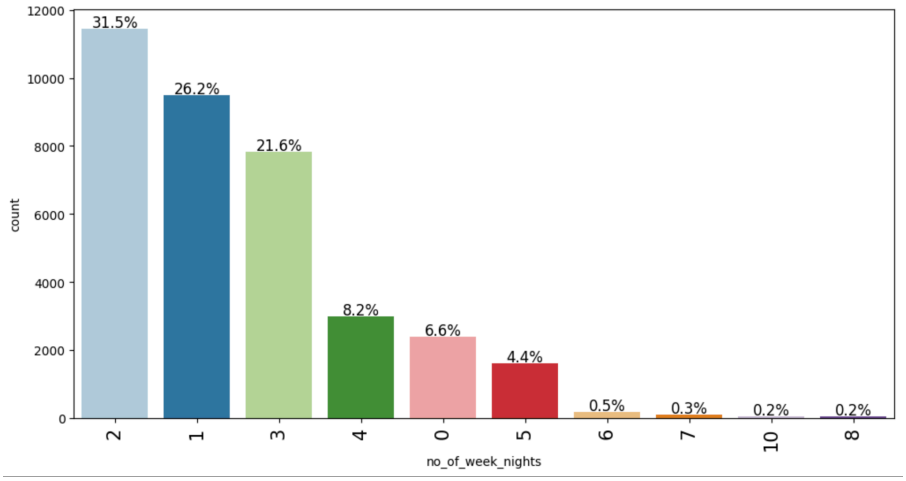

FIGURE 16: DISTRIBUTION OF NUMBER OF WEEK NIGHTS

2.2 BIVARIATE ANALYSIS

- **Lead Time vs Booking Status:** Bookings with shorter lead times (0–50 days) are far less likely to be canceled, while bookings made well in advance (over 100 days) show a much higher cancellation tendency.
- Lead time has a strong positive relationship with cancellation probability — the longer the time between booking and arrival, the greater the chance of cancellation.

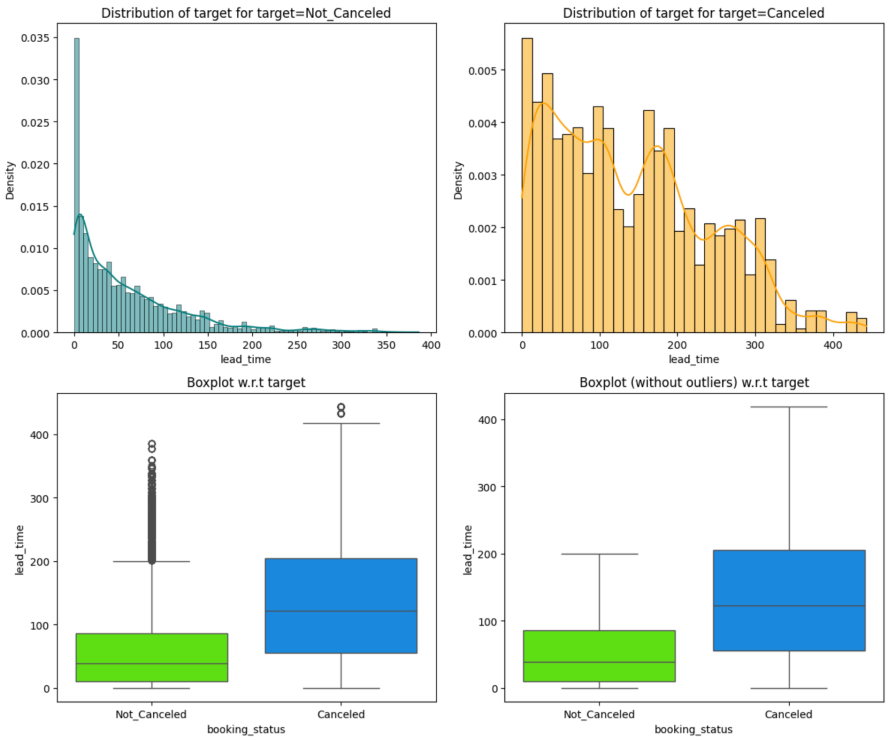

FIGURE 17: LEAD TIME VS BOOKING STATUS

- **Average Price per Room vs Booking Status:** Cancellations are more frequent for higher room prices — the median `avg_price_per_room` for canceled bookings is slightly higher than for non-canceled ones.
- Price variation is wider among canceled bookings, indicating that customers paying higher prices might be more price-sensitive or likely to find alternative deals before arrival.

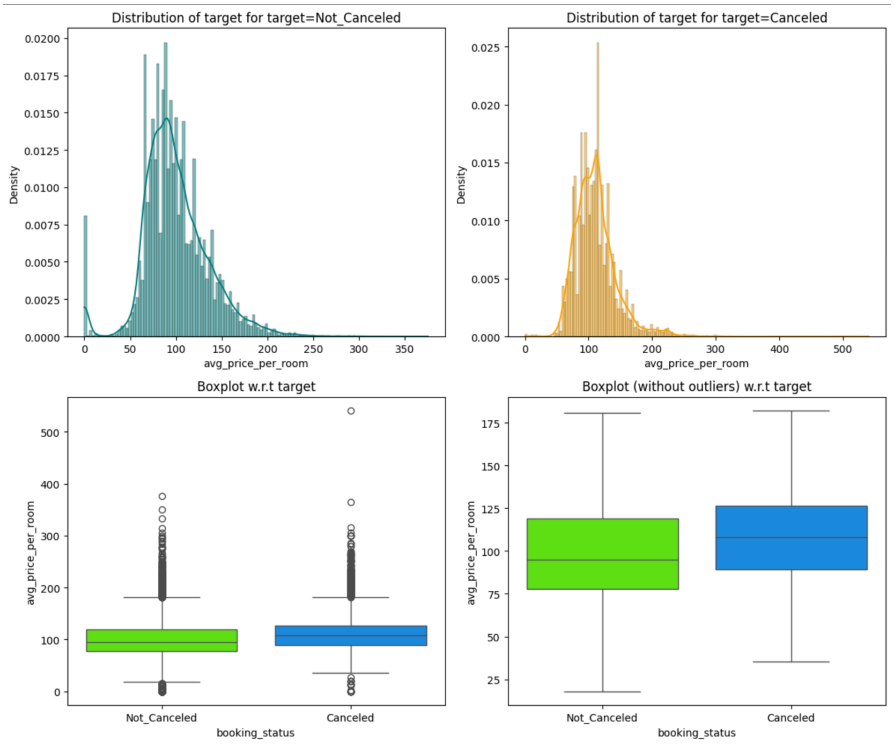

FIGURE 18: AVERAGE ROOM PRICE VS BOOKING STATUS

- **Market Segment vs Booking Status:** Online bookings show the highest chances of cancellation. Offline and Aviation bookings have lower cancellation rates, while Corporate bookings are the most stable.
- Complimentary bookings show no cancellations in the dataset.

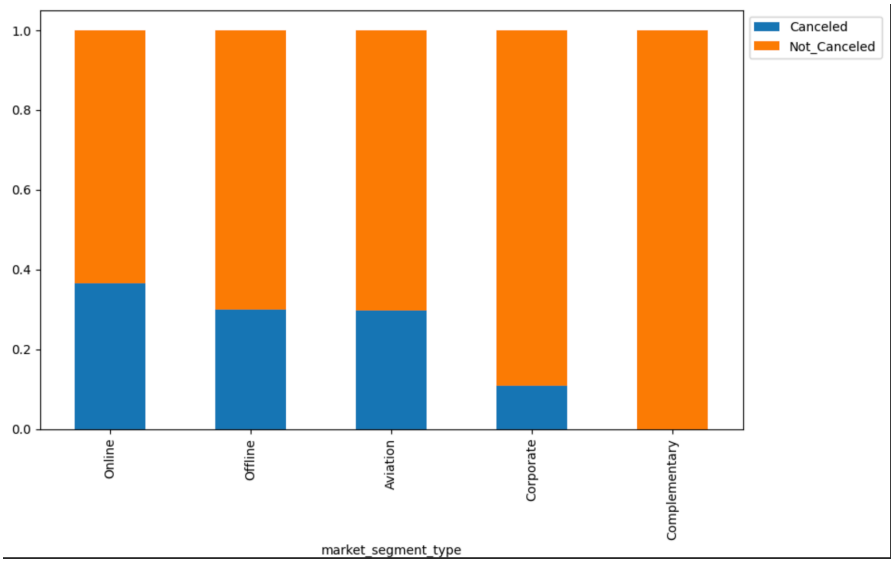

FIGURE 19: MARKET SEGMENT VS BOOKING STATUS

- **Price Variation by Segment:** Online and Offline segments have the widest variation in average room prices.

- **Lowest Price Segment:** The Complementary segment has the lowest prices, often close to zero.
- **Stable Pricing:** Aviation bookings have the most consistent rates.

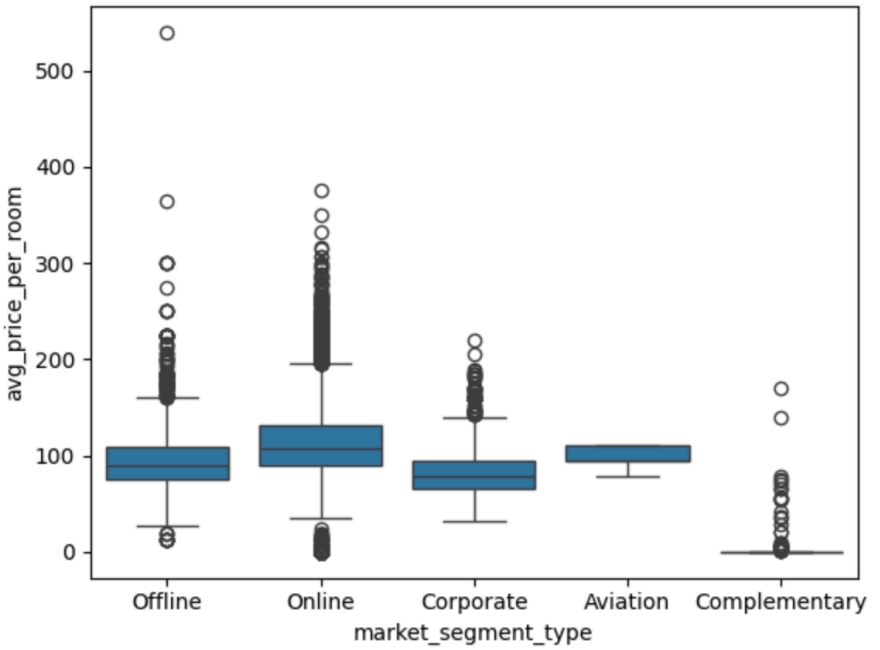

FIGURE 20: PRICE VARIATION ACROSS MARKET SEGMENTS

- **Car Parking Space vs Booking Status:** Customers who requested parking spaces have lower cancellation rates.

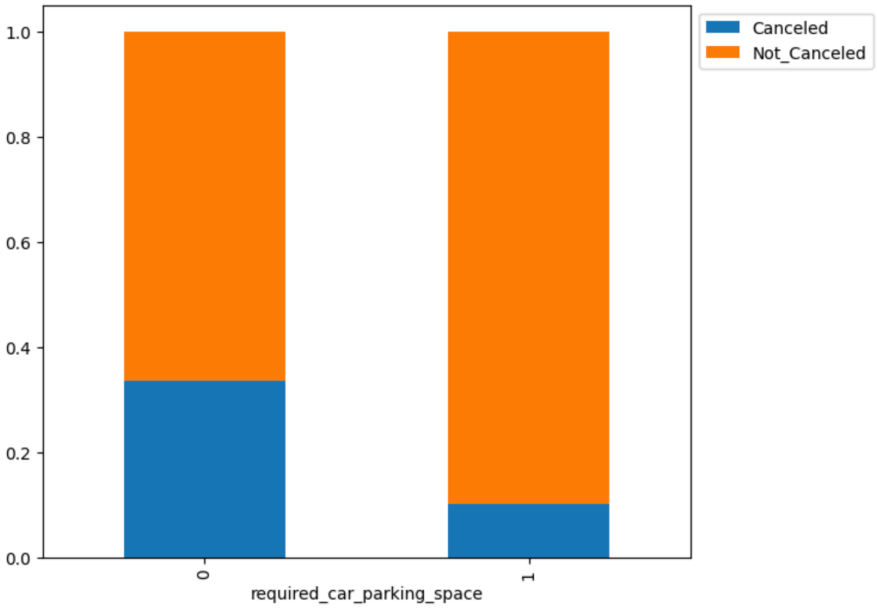

FIGURE 21: PARKING SPACE VS BOOKING STATUS

- **Meal Plan and Room Type Dependency:** Chi-square tests indicate that both meal plan and room type significantly affect booking cancellation likelihood.

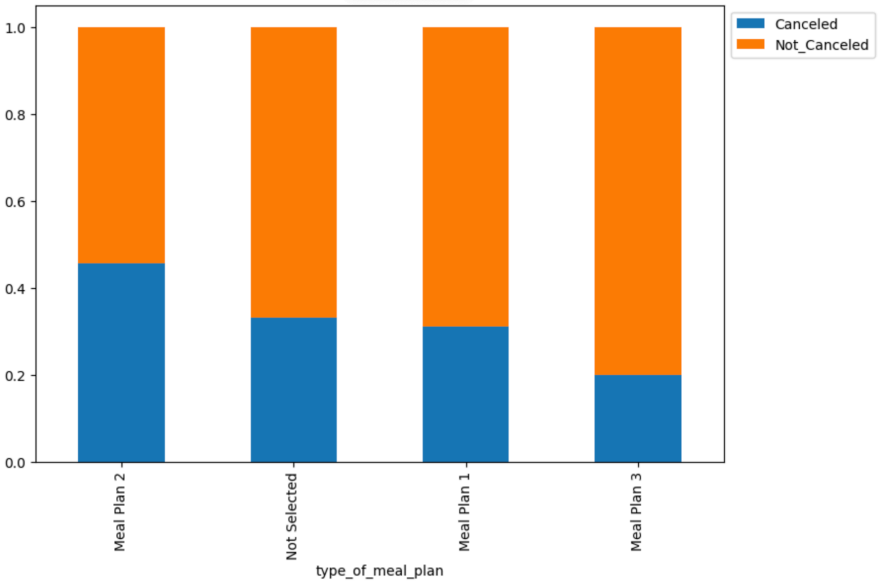

FIGURE 22: MEAL PLAN VS BOOKING STATUS

- **Arrival Year vs Booking Status:** Cancellations are higher in 2018 compared to 2017.

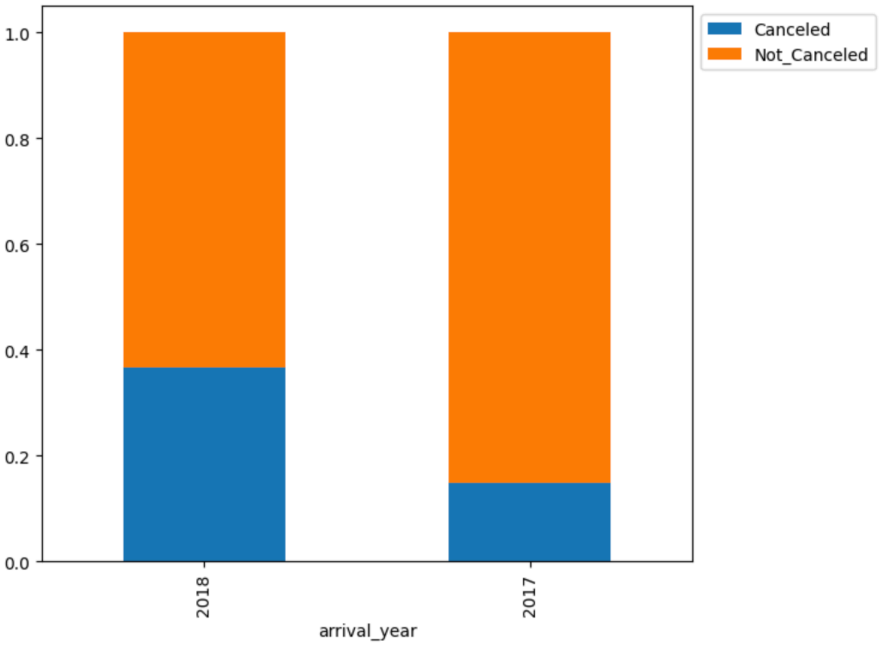

FIGURE 23: ARRIVAL YEAR VS BOOKING STATUS

- **Arrival Month vs Booking Status:** Cancellations are fewer in January–March and November–December; higher in mid-year months.

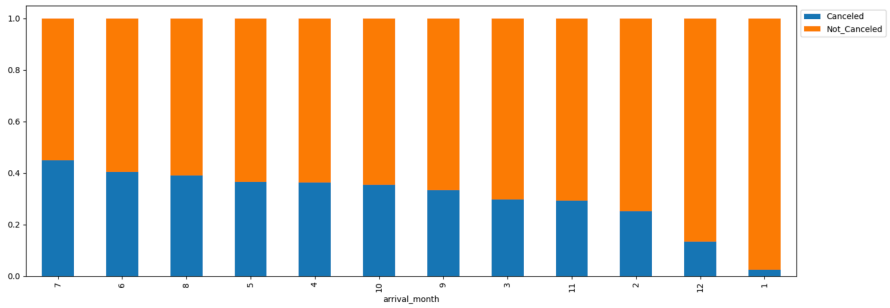

FIGURE 24: ARRIVAL MONTH VS BOOKING STATUS

- **Repeated Guest vs Booking Status:** Repeated guests rarely cancel — showing brand loyalty and satisfaction.

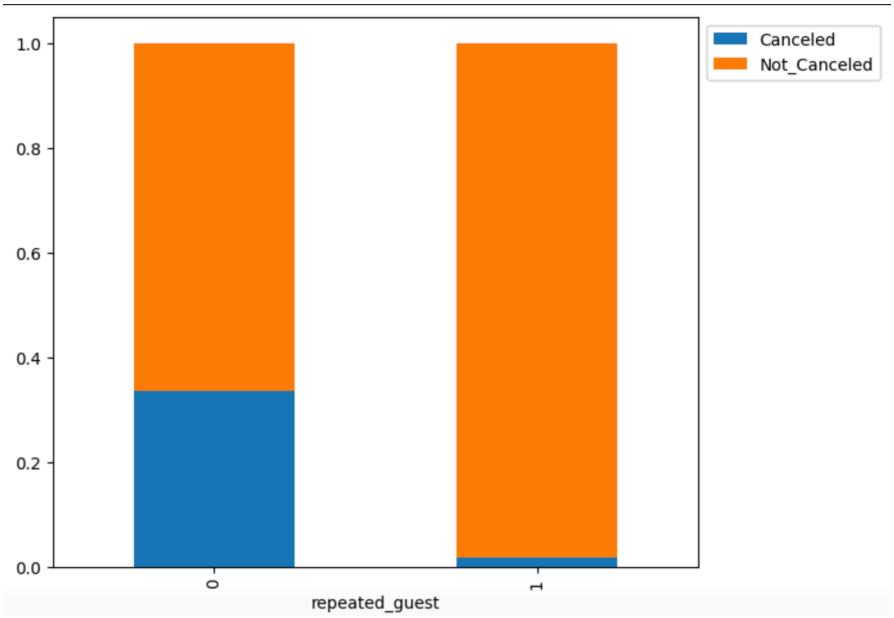

FIGURE 25: REPEATED GUEST VS BOOKING STATUS

- **Special Requests vs Booking Status:** Bookings with more special requests have much lower cancellation rates. No cancellations recorded for bookings with more than three requests.

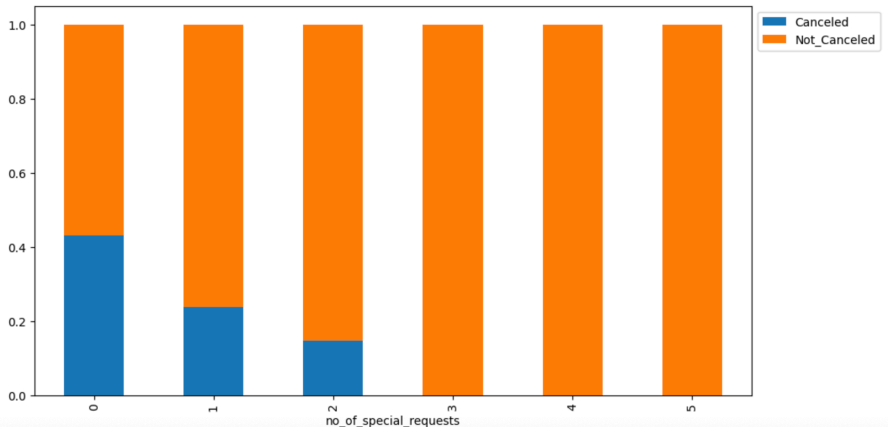

FIGURE 26: SPECIAL REQUESTS VS BOOKING STATUS

2.3 KEY EDA QUESTIONS

1. **What are the busiest months in the hotel?**

   - The most popular time for guests to arrive is in the second half of the year, particularly from July through December, which coincides with the fall and winter seasons.
   - The most popular month for visitors is October, followed by September. This is because in September - October, Portugal's weather shifts to mild autumn conditions mild and pleasant
   - The festive season in November and December also seems to attracts a high number of guests.

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

   - 64% of the bookings are done online followed by offline bookings that accounts to 29%
   - Out of remaining 7% of customers, majority are from Aviation and Corporate segments

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

   - **Price Variation by Segment** : The Online and Offline market segments show the widest variation in average room prices, with many outliers and higher median values compared to other segments.
   - **Lowest Price Segment** : The Complementary segment has the lowest average prices, with most values near zero, indicating complimentary or discounted stays.
   - **Stable Pricing**: The Aviation segment shows the most consistent pricing (narrow box, few outliers), suggesting standardized rates for this category.

4. **What percentage of bookings are canceled?**
    - 33% of the bookings are cancelled, rest 67% are not.
5. **Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?**
    - Customers who make bookings as repeated guest have very low chances of cancellations as opposed to first time customers.Out of 930 repeated guests only 16 have made cancellations.
    - Only 0.017% of the repeated guests have done cancellations
    - This represents INN Hotels Groups brand loyalty or customers who are satisfied with earlier experience of stays
6. **Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?**
    - Bookings with higer number of special requests have lower chances of cancellations
    - For bookings with > 3 special requests, no cancellations have been recorded

3 DATA-PREPROCESSING

To ensure data quality and model reliability, several preprocessing steps were performed before analysis and model building:

1. **Handling Invalid Adult Counts:**
   A total of **139 bookings** had `no_of_adults = 0`, which is not realistic for valid hotel reservations. These were corrected by setting `no_of_adults = 1` in cases where children were present (`no_of_children > 0`).
2. **Treating Zero Average Room Prices:**
   Instances with `avg_price_per_room = 0` were analyzed by market segment. Except for the `Complementary` segment (which legitimately has zero price), such values were replaced with the **mean price of that segment**, ensuring realistic pricing across bookings.
3. **Encoding Booking Status:**
   The categorical variable `booking_status` was converted into numeric form for model compatibility:
    - `Not_Canceled` $\rightarrow 0$
    - `Canceled` $\rightarrow 1$

4. **One-Hot Encoding of Categorical Variables:**
   Categorical columns — `type_of_meal_plan`, `market_segment_type`, and `room_type_reserved` — were transformed using one-hot encoding (with first category dropped to avoid multi-collinearity).

5. **Feature Engineering:**
   Two new features were derived to enhance predictive power:

   - `total_guests = no_of_adults + no_of_children`
   - `total_nights = no_of_weekend_nights + no_of_week_nights`

6. **Arrival Details:**
   The variable `arrival_date` (1–31) was retained despite its limited predictive power, while `arrival_year` contained only two values — 2017 and 2018.

7. **Outlier Analysis:**
   Variables `lead_time` and `avg_price_per_room` exhibited right-skewed distributions with extreme values. As these appeared to be legitimate business cases, no outlier removal was performed.

8. **Train-Test Split:**
   The dataset was divided into training and testing sets using a **70–30 split** with stratification on the target variable to maintain class balance:

   - Training set shape: (rows, features)
   - Test set shape: (rows, features)
   - Class distribution (Training): $\approx$ same ratio as original data
   - Class distribution (Test): $\approx$ same ratio as training

4 MODEL BUILDING

4.1 MODEL EVALUATION CRITERION

The objective of this analysis is to predict booking cancellations effectively. Two types of prediction errors can occur:

1. Predicting a booking **will be canceled** when it is actually **not canceled** (False Positive)
2. Predicting a booking **will not be canceled** when it is actually **canceled** (False Negative)

The second case is more critical for the hotel, as last-minute cancellations directly impact revenue and resource allocation. Hence, the goal is to **minimize False Negatives** — that is, to correctly identify as many actual cancellations as possible. **Recall** was therefore selected as the key performance metric to maximize.

4.2 MODEL EVALUATION APPROACH

Two supervised machine learning models were trained and evaluated:

1. Logistic Regression
2. Decision Tree Classifier

For both models, data was split into training and testing sets using a **70:30 stratified split** to maintain class balance in the train and test sets. Performance was assessed using **Accuracy**, **Precision**, **Recall**, and **F1-score**, along with a **confusion matrix** for visual interpretation.

4.3 LOGISTIC REGRESSION MODEL

A Logistic Regression model was built using class weighting to handle data imbalance. The model showed consistent performance across both training and testing sets, with no signs of overfitting.

- Training and test recall: **≈ 0.62**
- Performance was stable, indicating good generalization.

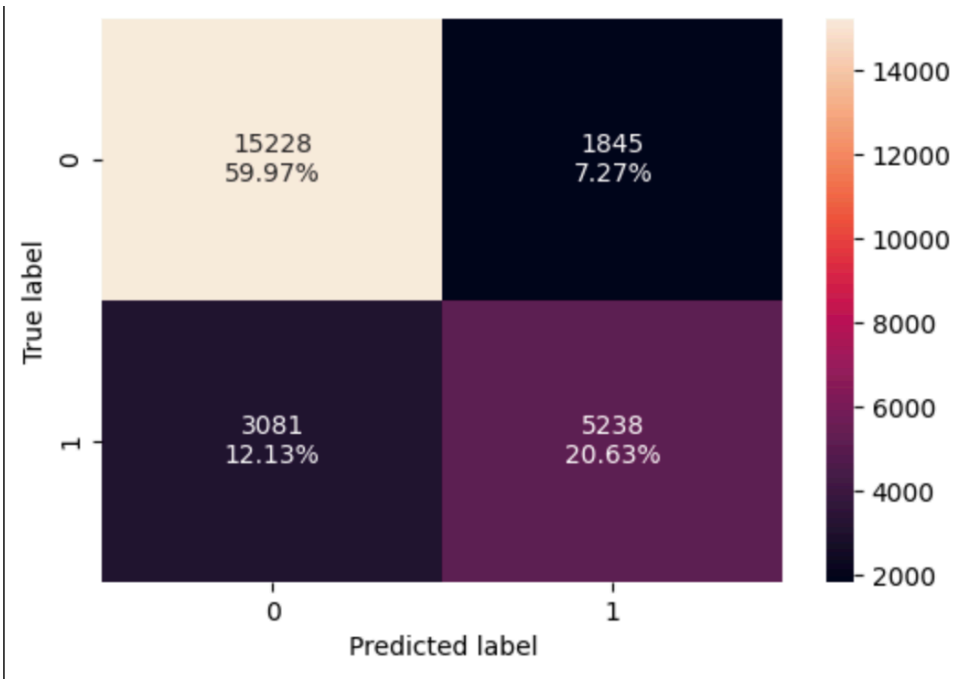

FIGURE 27: CONFUSION MATRIX FOR LOGISTIC REGRESSION TRAINING SET

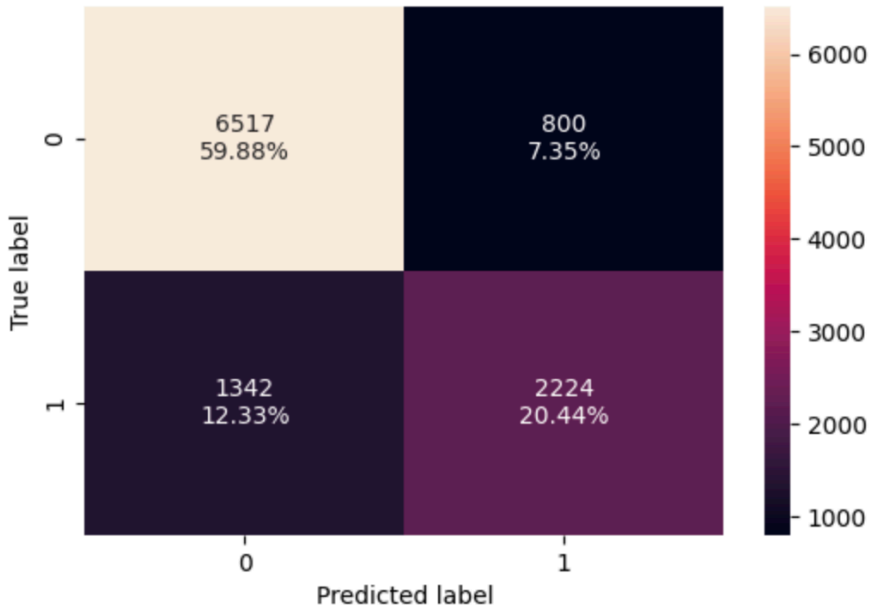

FIGURE 28: CONFUSION MATRIX FOR LOGISTIC REGRESSION TESTING SET

4.4 DECISION TREE MODEL

A Decision Tree classifier was then developed using balanced class weights. While the model performed strongly on the training data, it showed a significant performance drop on the test set, indicating **overfitting**.

- The model memorized training patterns instead of learning general trends.

- Pruning or hyperparameter tuning was identified as necessary next steps.

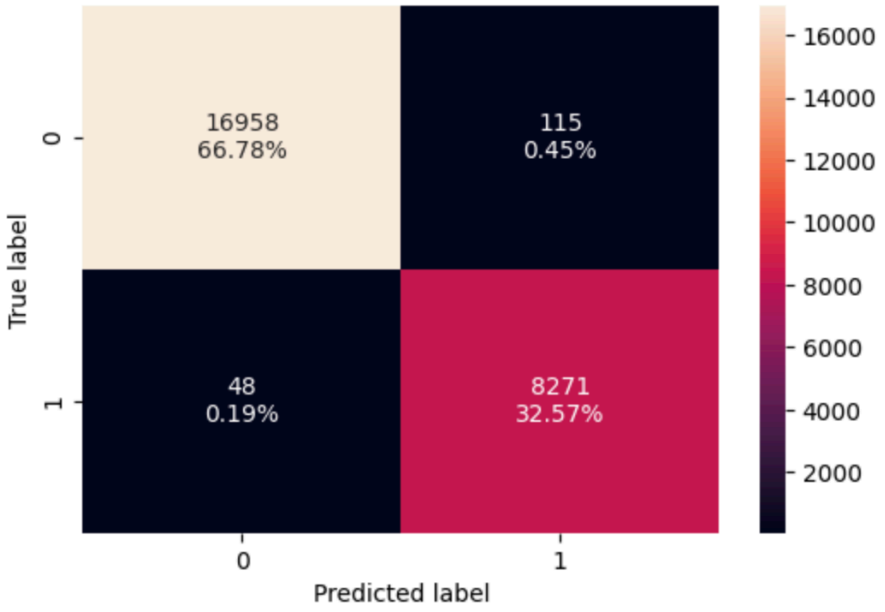

FIGURE 29: CONFUSION MATRIX FOR DECISION TREE TRAINING SET

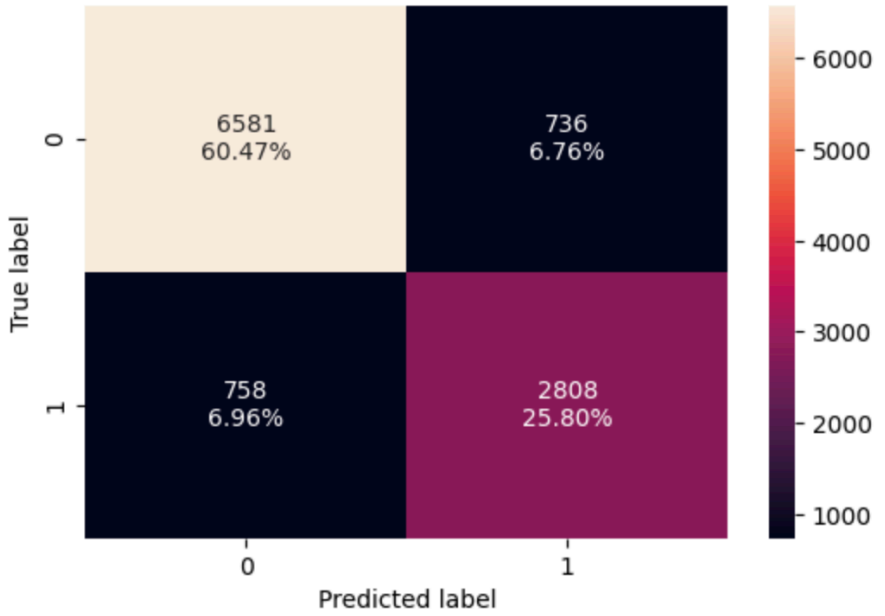

FIGURE 30: CONFUSION MATRIX FOR DECISION TREE TESTING SET

4.5 SUMMARY OF MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (Train)	0.80	0.7388	0.62	0.68
Logistic Regression (Test)	0.80	0.735	0.62	0.674
Decision Tree (Train)	0.99	0.99	0.98	0.99
Decision Tree (Test)	0.862	0.787	0.79	0.789

TABLE 2: MODEL PERFORMANCE SUMMARY

Conclusion: The Logistic Regression model demonstrates better generalization and stability, while the Decision Tree model exhibits overfitting. Further optimization such as pruning or selecting optimal threshold is required.

5 MODEL PERFORMANCE IMPROVEMENT

5.1 OVERVIEW

To enhance model performance and interpretability, both the **Logistic Regression** and **Decision Tree Classifier** models were refined through feature selection, threshold tuning, and pruning techniques. The primary goal was to **maximize Recall**, as false negatives (missed cancellations) pose a significant business risk to the hotel.

5.2 LOGISTIC REGRESSION TUNING

5.2.1 HANDLING MULTICOLLINEARITY

Variance Inflation Factor (VIF) analysis was conducted to detect multicollinearity among predictors. Sequential removal of highly correlated features — `total_guests`, `total_nights`, and `market_segment_type_Online` — resulted in all remaining features having VIF < 5, confirming that multicollinearity was no longer an issue.

```
Variance Inflation Factors:
                                Variable           VIF
0                                  const  3.936887e+07
1                          no_of_adults  1.318667e+00
2                        no_of_children  1.991451e+00
3                  no_of_weekend_nights  1.070532e+00
4                     no_of_week_nights  1.096577e+00
5            required_car_parking_space  1.034553e+00
6                             lead_time  1.404407e+00
7                          arrival_year  1.425437e+00
8                         arrival_month  1.281471e+00
9                          arrival_date  1.007704e+00
10                       repeated_guest  1.747589e+00
11         no_of_previous_cancellations  1.321954e+00
12  no_of_previous_bookings_not_canceled  1.570605e+00
13                   avg_price_per_room  2.150262e+00
14               no_of_special_requests  1.242087e+00
15        type_of_meal_plan_Meal Plan 2  1.285102e+00
16        type_of_meal_plan_Meal Plan 3  1.007975e+00
17      type_of_meal_plan_Not Selected  1.285088e+00
18    market_segment_type_Complementary  1.360579e+00
19         market_segment_type_Corporate  1.544945e+00
20           market_segment_type_Offline  1.635001e+00
21       room_type_reserved_Room_Type 2  1.093918e+00
22       room_type_reserved_Room_Type 3  1.003895e+00
23       room_type_reserved_Room_Type 4  1.362638e+00
24       room_type_reserved_Room_Type 5  1.033683e+00
25       room_type_reserved_Room_Type 6  2.001896e+00
26       room_type_reserved_Room_Type 7  1.096586e+00
```

FIGURE 31: VIF VALUES WITHOUT MULTICOLLINEARITY

5.2.2 REMOVING INSIGNIFICANT VARIABLES

Variables with high p-values $(p > 0.05)$ were iteratively dropped from the model until all remaining predictors were statistically significant. This ensured that only meaningful features contributed to the model.

5.2.3 DETERMINING OPTIMAL THRESHOLD USING ROC CURVE

The ROC curve was used to identify the optimal probability threshold that maximized the difference between True Positive Rate and False Positive Rate.

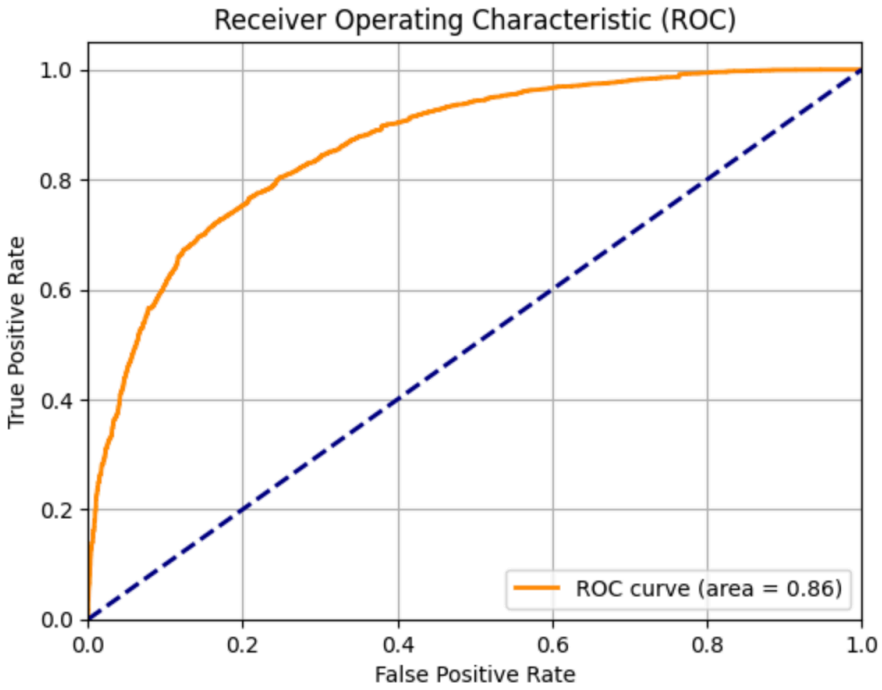

FIGURE 32: ROC CURVE FOR LOGISTIC REGRESSION MODEL

The optimal threshold was found to be approximately **0.331**.

5.2.4 MODEL EVALUATION

- Recall increased from **0.64** to approximately **0.76**, that is, improved ability to identify high-risk cancellations.
- The model generalized well across both training and test datasets.

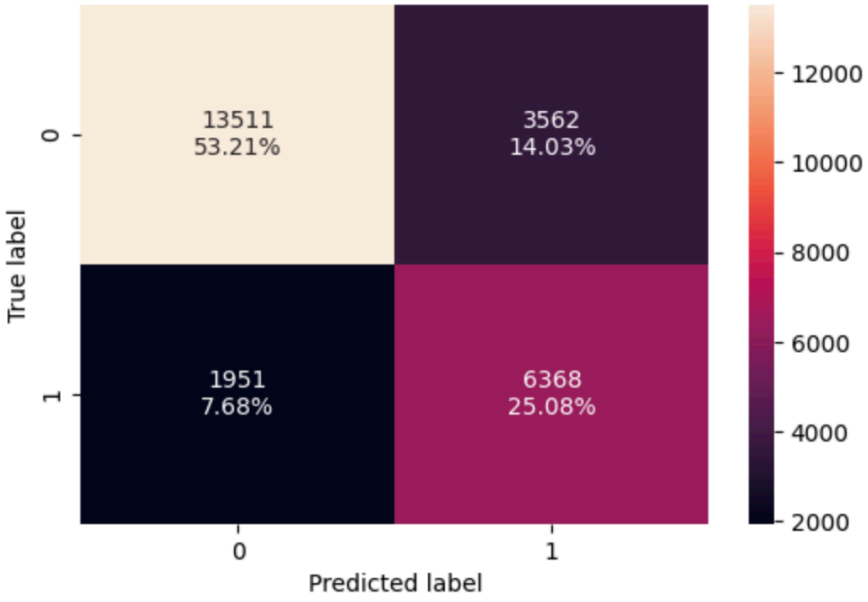

FIGURE 33: CONFUSION MATRIX FOR LOGISTIC REGRESSION(OPTIMAL THRESHOLD) TRAINING SET

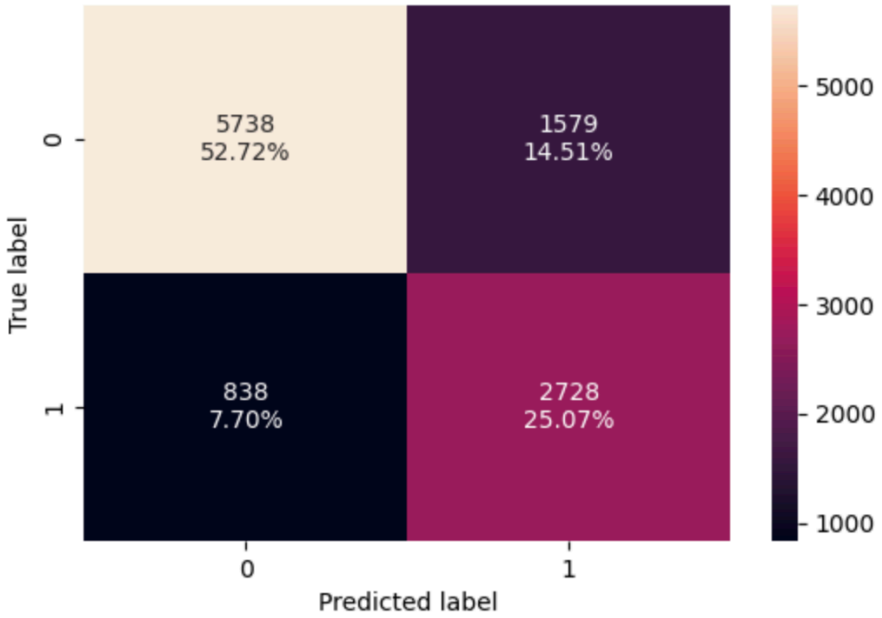

FIGURE 34: CONFUSION MATRIX FOR LOGISTIC REGRESSION(OPTIMAL THRESHOLD) TESTING SET

5.3 SUMMARY OF MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (Train)	0.782	0.765	0.64	0.697
Logistic Regression (Test)	0.77	0.765	0.63	0.69

TABLE 3: MODEL PERFORMANCE SUMMARY

5.4 DECISION TREE CLASSIFIER TUNING

5.4.1 PRE-PRUNING

Grid Search with 5-fold cross-validation was used to optimize hyperparameters such as `max_depth`, `max_leaf_nodes`, and `min_samples_split`. This helped prevent overfitting and improved generalization.

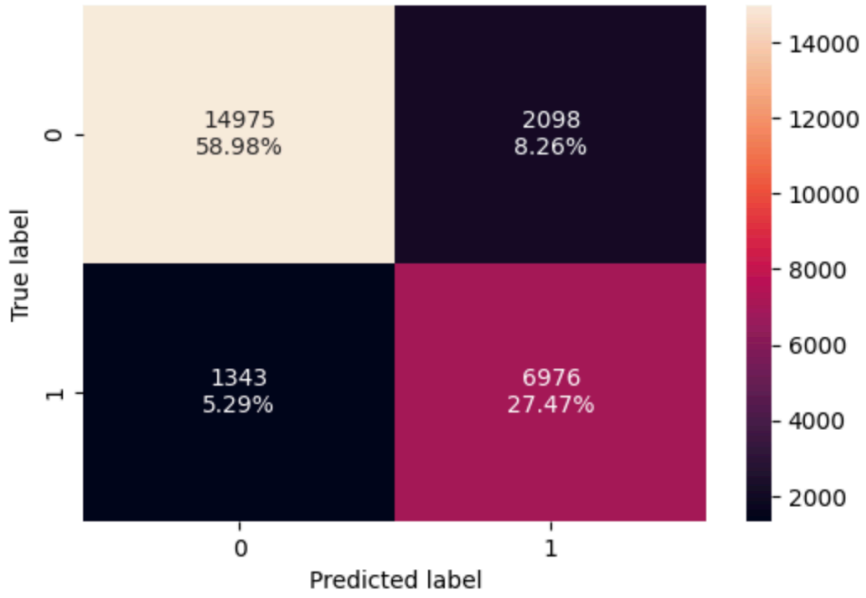

FIGURE 35: CONFUSION MATRIX FOR PRE-PRUNED DECISION TREE TRAINING SET

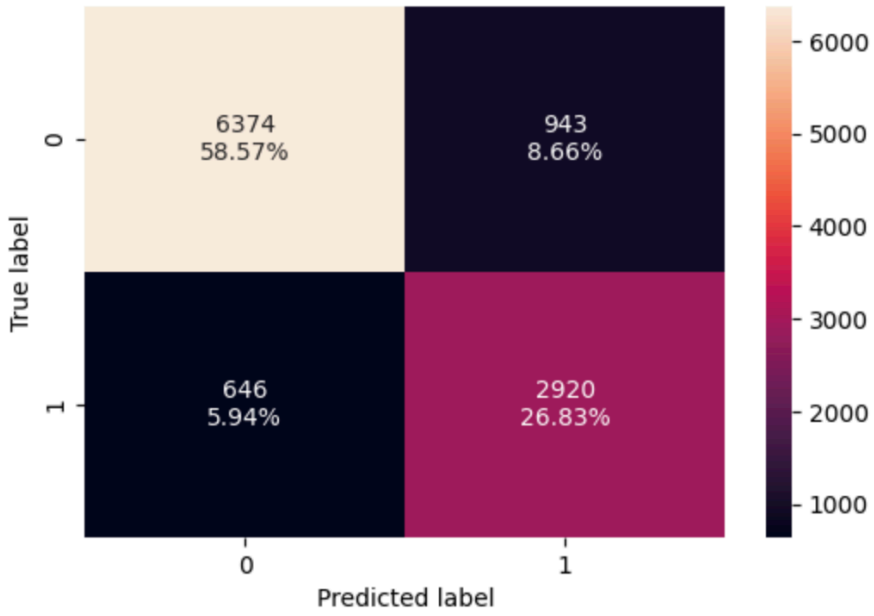

FIGURE 36: CONFUSION MATRIX FOR PRE-PRUNED DECISION TREE TESTING SET

5.4.2 POST-PRUNING

Cost complexity pruning was then applied to further simplify the tree by selecting an optimal `ccp_alpha` based on recall performance across training and testing sets.

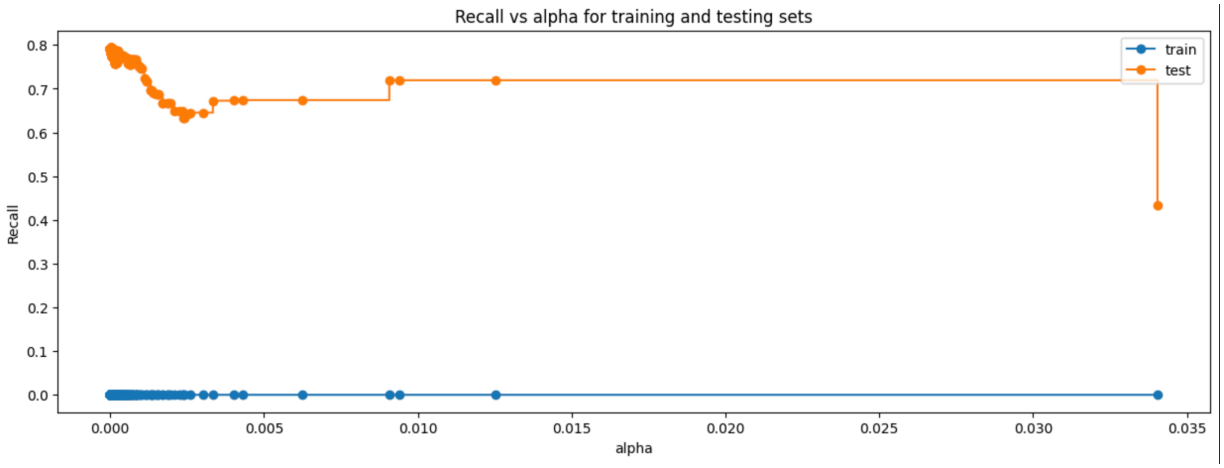

FIGURE 37: RECALL VS ALPHA CURVE FOR POST-PRUNING

5.4.3 FEATURE IMPORTANCE ANALYSIS

Feature importance visualization revealed that **lead_time**, **avg_price_per_room**, **market_segment_type_Online**, **no_of_special_requests** were the most influential predictors of booking cancellations.

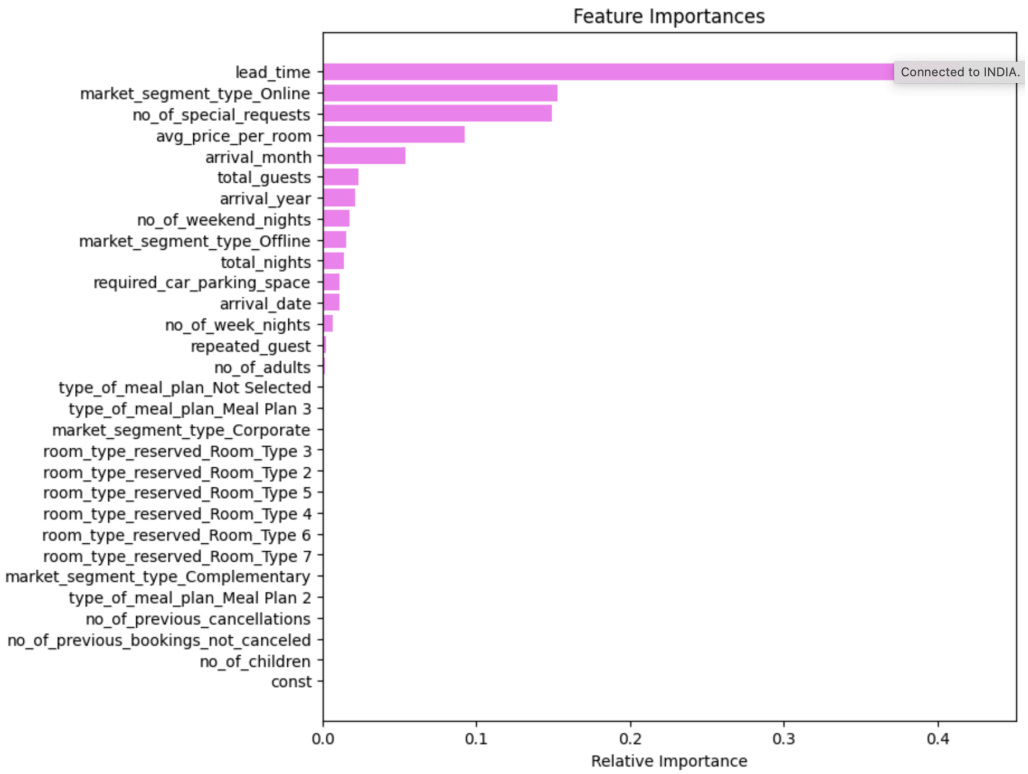

FIGURE 38: FEATURE IMPORTANCE FOR PRE-PRUNED DECISION TREE CLASSIFIER

5.4.4 MODEL EVALUATION

- The pRE-pruned Decision Tree achieved balanced recall and accuracy, eliminating prior overfitting issues.

- Recall improved notably on the test set without sacrificing too much precision.
- The model structure became more interpretable and business-friendly.

6 MODEL PERFORMANCE COMPARISON AND FINAL MODEL SELECTION

6.1 OVERVIEW

To identify the most effective predictive model for booking cancellations, performance metrics for all models were compared across both training and test datasets. The models evaluated included:

- Logistic Regression (Base)
- Logistic Regression (Tuned)
- Decision Tree (Base)
- Decision Tree (Pre-Pruned)
- Decision Tree (Post-Pruned)

6.2 TRAINING PERFORMANCE COMPARISON

The table below summarizes the performance metrics for all models on the training dataset.

Training performance comparison:

	Logistic Regression Base	Logistic Regression Improved	Decision Tree Base	Decision Tree Pre-Pruned	Decision Tree Post-Pruned
Accuracy	0.806002	0.782884	0.993581	0.864485	0.672377
Recall	0.629643	0.765477	0.994230	0.838562	0.000000
Precision	0.739517	0.641289	0.986287	0.768790	0.000000
F1	0.680171	0.697901	0.990242	0.802162	0.000000

FIGURE 39: TRAINING PERFORMANCE COMPARISON ACROSS MODELS

6.3 TEST PERFORMANCE COMPARISON

Similarly, the test set results demonstrate how well each model generalizes to unseen data.

Test set performance comparison:

	Logistic Regression Base	Logistic Regression Tuned	Decision Tree Base	Decision Tree Pre-Pruned	Decision Tree Post-Pruned
Accuracy	0.803179	0.777911	0.862722	0.853992	0.672333
Recall	0.623668	0.765003	0.787437	0.818845	0.000000
Precision	0.735450	0.633388	0.792325	0.755889	0.000000
F1	0.674962	0.693001	0.789873	0.786108	0.000000

FIGURE 40: TEST SET PERFORMANCE COMPARISON ACROSS MODELS

6.4 KEY OBSERVATIONS

- The **base Decision Tree** exhibited clear overfitting — achieving very high training accuracy but poor generalization on test data.
- The **pre-pruned Decision Tree** achieved balanced results with a Recall of **0.83** on training and **0.81** on test data, indicating effective control over overfitting.
- The **Logistic Regression (Tuned)** model performed consistently with high stability and interpretability, though with slightly lower recall than the pruned tree.

6.5 FINAL MODEL RECOMMENDATION

Based on recall performance, generalization ability, and interpretability, the following conclusions were drawn:

- The **Pre-Pruned Decision Tree** offers the best trade-off between predictive performance and explainability.
- Its structure allows hotel management to clearly understand the drivers behind cancellations, making it valuable for business decision-making.
- The model can be used to flag **high-risk bookings** early, enabling proactive retention measures such as follow-up calls, offers, or flexible rebooking options.

6.6 BUSINESS IMPACT

Deploying the pre-pruned decision tree model enables the hotel to:

- **Reduce revenue loss** from unexpected last-minute cancellations.
- **Improve resource utilization** by optimizing staffing and room allocations.
- **Enhance customer engagement** through targeted offers to high-risk segments.

Final Recommendation: The **Pre-Pruned Decision Tree Classifier** should be adopted as the operational model for predicting cancellations, supported by periodic retraining to maintain accuracy as booking patterns evolve.

7 ACTIONABLE INSIGHTS & RECOMMENDATIONS

1. **High-risk bookings:** Guests who book far in **advance**, use **online channels**, or reserve **high-priced rooms** are more likely to cancel.
   The hotel should monitor such bookings closely and consider **flexible pricing** or **partial prepayment** policies to reduce last-minute cancellations.

2. **Loyal customers: Repeated guests** and those making **special requests** show strong loyalty and are less likely to cancel.
   The hotel can **reward** these customers through loyalty points, exclusive offers, or personalized service to encourage continued engagement.

3. **Value-added services:** Guests opting for **parking spaces** or **meal plans** tend to follow through with their bookings.
   The hotel can promote **bundled offers** (e.g., stay + breakfast + parking) to increase commitment and reduce cancellation rates.

4. **Seasonal trends:** Cancellations are lower in the first and last two months of the year.
   Plan **marketing and retention campaigns** during high-cancellation periods to stabilize occupancy throughout the year. Staff can take proactive measures—such as **follow-up calls** or **special offers**—to secure high-risk bookings and minimize revenue loss.