INN Hotels

Project 4 – Supervised Learning-Classification

May 10, 2025

Submitted By:
Alex Kyeremateng Botwe

Contents / Agenda

List of Tables

List of Figures

List of Figures

List of Figures

Executive Summary

- The hospitality industry is increasingly affected by high booking cancellations, particularly those made at the last minute. While flexible cancellation policies and online booking platforms benefit guests, they present substantial challenges to hotels such as:
    - Lost Revenue: Unused rooms due to unanticipated cancellations.
    - Operational Costs: Increased costs in customer service, marketing, and rebooking.
    - Reduced Profit Margins: Necessity to lower room prices to fill vacancies last minute.
    - Increased Third-Party Dependence: Heavier reliance on costly distribution channels and promotions.
- INN Hotels Group aims at predicting cancellations in advance using data-driven techniques to enable strategic interventions.

- **Business Insights:**

  a. **Cancellations Are Predictable with High Accuracy:**

     - Both Logistic Regression and Decision Tree models predict cancellations with approximately 80–86% accuracy on test data.
     - This validates that customer and booking behavior patterns strongly correlate with cancellation likelihood

  b. **Model Performance Indicates Different Strengths:**

     i. Decision Tree Insights:

        Post-pruned tree generalizes best with the highest performance:

        - Recall (0.84044) catches the most likely cancellations
        - F1 (0.79750) provides excellent balance of precision and recall

ii. Logistic Regression Insights:

Logistic Regression (Threshold = 0.37) offers:

- Easier interpretability
- High recall (0.73964) with reasonable precision, ideal when quick business rules or auditability is needed

c. Overfitting Must Be Prevented:

- The default Decision Tree model is overfitted, with nearly perfect training scores but poorer test performance
- Proper pruning significantly improves generalization and real-world reliability

Model Comparison - Test Set		
Metric	Logistic Reg. (Thresh=0.37)	Decision Tree (Post-Pruning)
Accuracy	0.79555	0.86015
Recall	0.73964	0.84044
Precision	0.66573	0.75873
F1 Score	0.70074	0.7975

- The Decision Tree shows a stronger overall test performance than Logistic Regression

- **Business Recommendations:**

  a. Adopt Post-Pruned Decision Tree Model since it offers real-time prediction of cancellations with high accuracy

  b. The Post-Pruned Decision Tree Model helps flag risky bookings in advance, allowing proactive interventions

c. Deploy Logistic Regression with a tuned threshold (0.37) for Business Rule Design

d. Align model choice with business risk and resource constraints

e. Add cancellation risk scores to the booking management system

f. Integrate predictive output into operations to automatically trigger pre-arrival confirmations or rebooking incentives for high-risk guests, dynamic overbooking strategies based on forecasted cancellations and staff and resource planning adjustments

g. Model performance should be evaluated quarterly since seasonality, promotions, or changes in policies may shift guest behavior over time

- **Conclusions:**

  - INN Hotels can significantly reduce operational losses and improve occupancy rates through predictive modeling.

  - The Post-Pruned Decision Tree model is the optimal solution, providing actionable insights while ensuring generalizability.

  - Logistic Regression remains a robust, interpretable alternative for policy setting and stakeholder communication.

  - The predictive system empowers INN Hotels to make data-driven, proactive decisions that improve profitability and customer satisfaction.

Business Problem Overview and Solution Approach

- **Business Problem Overview:**

INN Hotels Group is facing significant financial and operational setbacks due to a high volume of booking cancellations. These affect room occupancy, staffing, marketing expenditures, and overall profitability.

To address this challenge, INN Hotels Group seeks to implement a data-driven ML solution to:

- Identify key factors influencing the likelihood of cancellation
- Predict cancellations in advance using data-driven techniques to mitigate cancellation impacts

This is purely a classification problem and the predictive framework will empower INN Hotels Group to reduce financial operational losses, improve booking management, and increase profitability using intelligent forecasting.

- ## Solution Approach

To predict the likelihood of a cancellation, **a machine learning model** was built using the following steps:

1. **Data Analysis**

   - We analyzed the provided booking dataset which included factors like booking lead time, customer type and demographics, room type, seasonal and temporary patterns and market segment type, which may influence the likelihood of cancellations.

   - Data cleaning and normalization was employed

2. **Exploratory Data Analysis (EDA)**

- Identify trends and cancellation prone profiles using visual and statistical tools.
- Identified potential predictors and multicollinearity risks

3. **Data Preprocessing**

- Converted and encoded of categorical variables to derive new features
- Outlier detection, Feature engineering and Data preparation for modeling were performed

4. Model Development & Refinement (Logistic Regression)

- Applied Logistic Regression (Logit) Model
- Iteratively removed high VIF variables to address multicollinearity
- Refined model using p-value-based backward elimination
- Checking model performance on the training set and test set
- Train predictive model to estimate cancellation probabilities

5. Model Evaluation & Validation

- Evaluated performance on train and test split using: Accuracy, Recall, Precision and F1
- Plotted confusion matrix and ROC curve for further evaluation

6. **Model Development & Refinement (Decision Tree)**

- Applied Decision Tree Model (DecisionTreeClassifier() from sklearn)
- Tree splits data recursively using Gini Impurity or Entropy (Information Gain)

7. **Model Evaluation & Validation**

- Evaluated performance on train and test split using: Accuracy, Recall, Precision and F1
- Visualize the tree for business explainability
- Check overfitting: a deep tree might perform perfectly on training data but poorly on test data

EDA – Univariate Analysis

- **Lead Time**

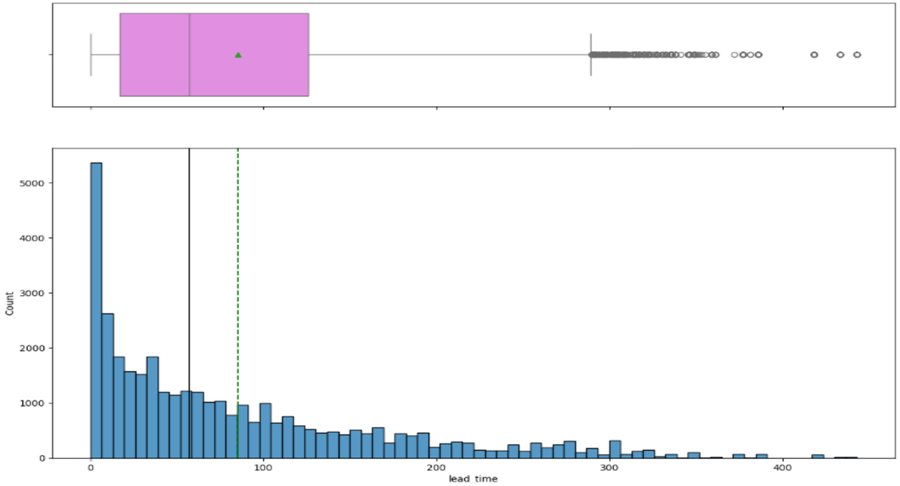


Fig 1: Univariate Analysis of Lead Time

Observations:

- The distribution of lead times is positively skewed indicating a lead time distribution that is concentrated at lower values with a long tail of higher values and several outliers. This suggest a tendency for longer-than-usual lead times to occur

Link to Appendix slide on data background check

- ## Average Price per Room

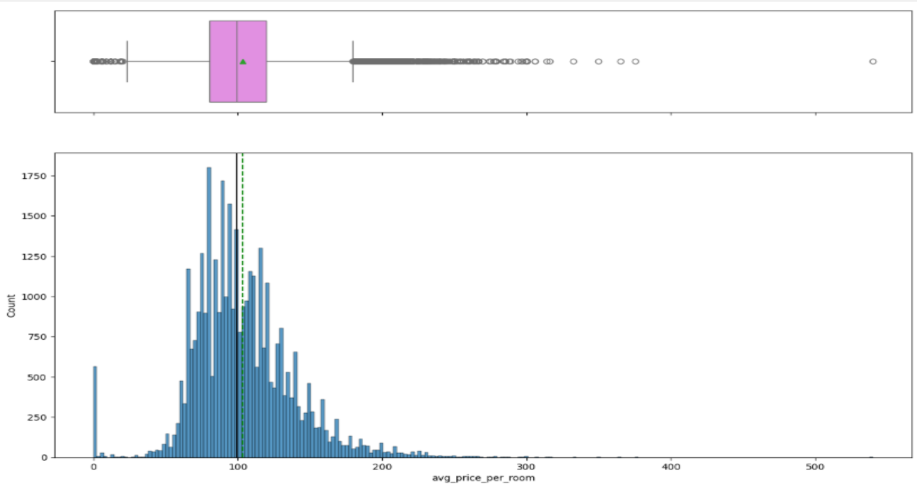


Fig 2: Univariate Analysis of Average Price per Room

Observations:

- The histogram also suggests a mild positive skew. The distribution has a longer tail extending to the right (higher prices) compared to the left.

- This suggest that the average price per room tends to be in a specific lower range, but there's a noticeable presence of more expensive rooms contributing to a slightly higher average overall

Link to Appendix slide on data background check

- ## Previous Booking Cancellations

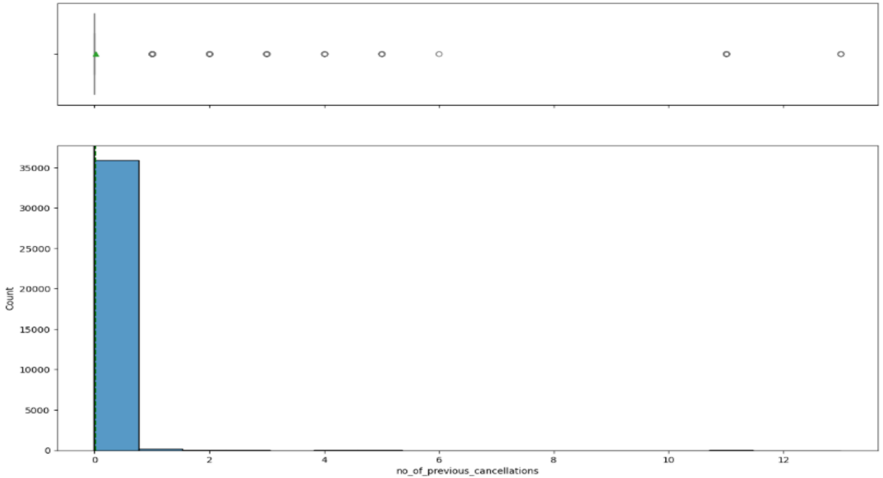


Fig 3: Univariate Analysis of Previous Booking Cancellations

Observations:

- The distribution is highly positively skewed. The vast majority of the data is concentrated at the lowest value of zero(0).
- This means that most customers are new or have not cancelled previously.
- A small fraction of customers account for the instances of one or more previous cancellations, with a few individuals having a notably high number of past cancellations.

Link to Appendix slide on data background check

• Number of Previous Booking not Canceled

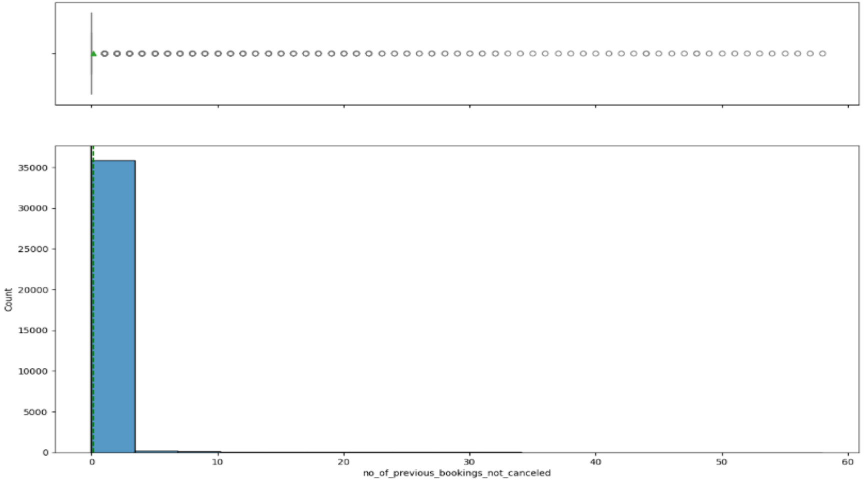


Fig 4: Univariate Analysis of Number of Previous Booking not Canceled

Observations:

- The distribution is highly positively skewed towards zero(0)
- This indicates that most customers have no previous bookings that were not canceled which is quite surprising and unusual.
- This could imply that the dataset may primarily focus on new customers or customers whose previous bookings were all canceled. This therefore warrants further investigation

[Link to Appendix slide on data background check]

- **Number of Adults**

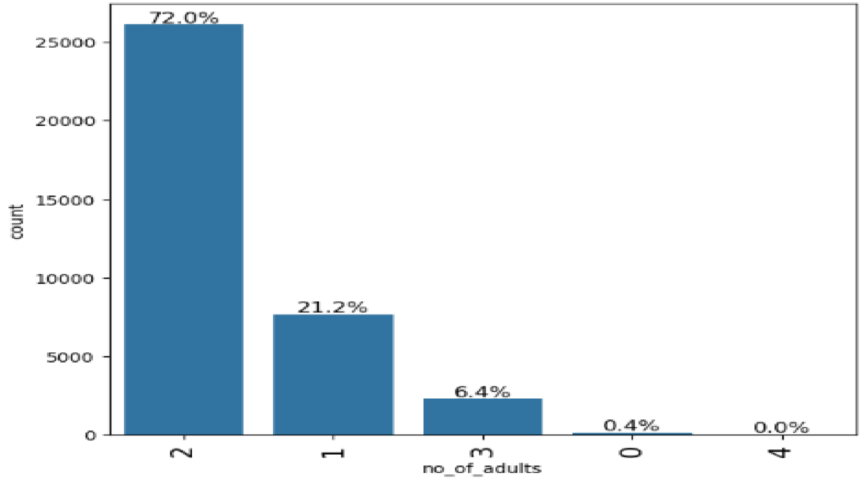


Fig 5: Univariate Analysis of Number of Adults

Observations:

- The vast majority of bookings are for two adults, accounting for approximately 72.0% of all reservations
- 21.2% of the bookings were made by one adult
- Bookings for three adults constitute just a small percentage.
- This suggest that most people book for themselves and one other person

Link to Appendix slide on data background check

- **Number of Children**

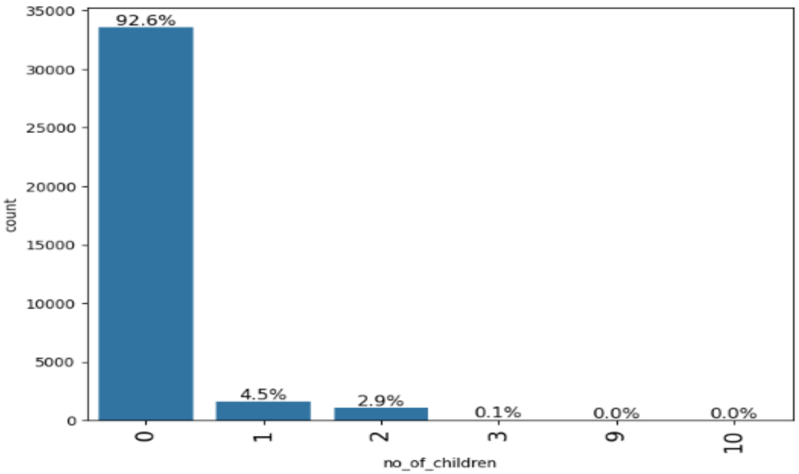


Fig 6: Univariate Analysis of Number of Children

Observations:

- The vast majority of bookings, 92.6%, were made with zero children. This indicates that most reservations are made by adults-only parties.
- While the majority of bookings are adults-only, the presence of bookings with one or two children suggests some demand for family-friendly amenities, but perhaps on a smaller scale.

[Link to Appendix slide on data background check]

Number of Week Nights

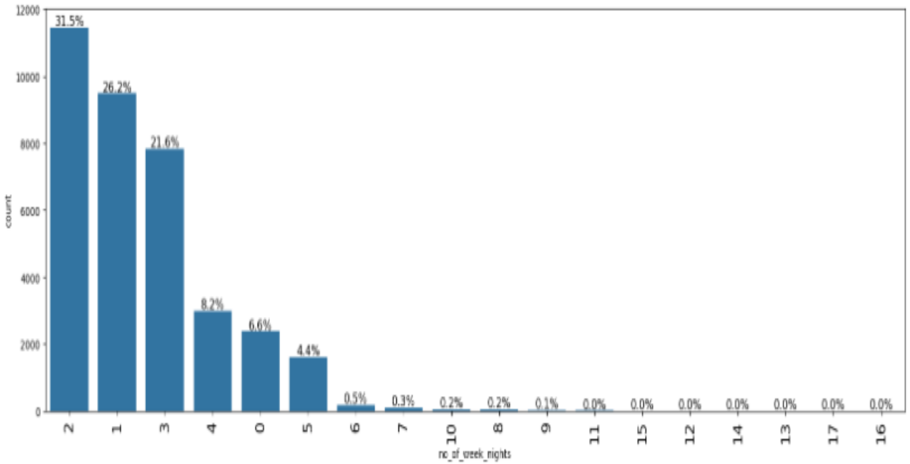


Fig 7: Univariate Analysis of Number of Week Nights

Observations:

- The distribution indicates that most guests booking week nights tend to stay for short durations (1-3 nights), with a smaller but still significant portion staying for around 4 nights or booking stays that don't include full week nights.

- Longer weekday stays are quite uncommon

[Link to Appendix slide on data background check](#)

- ## Number of Weekend Nights

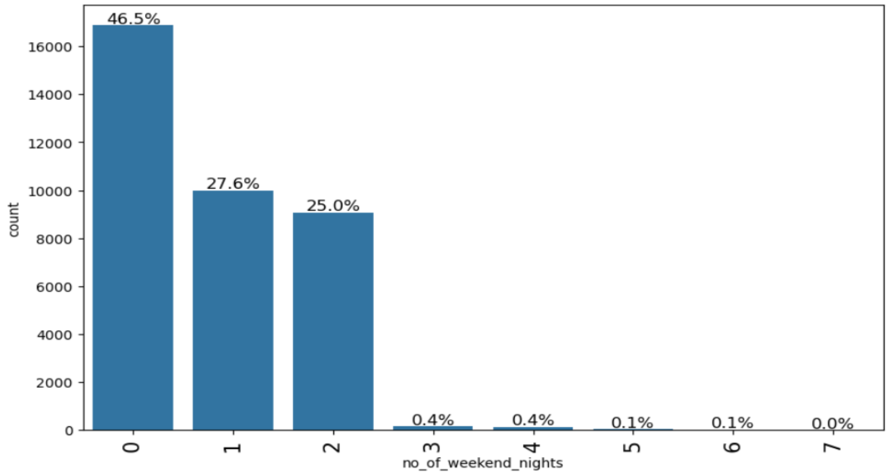


Fig 8: Univariate Analysis of Number of Weekend Nights

Observations:

- The distribution indicates that a large portion (46.5%) of stays occur entirely during the weekdays likely driven by business travelers or short weekday getaways.
- Extended weekend stays beyond two nights are very infrequent.

Link to Appendix slide on data background check

- **Required Car Parking Space**

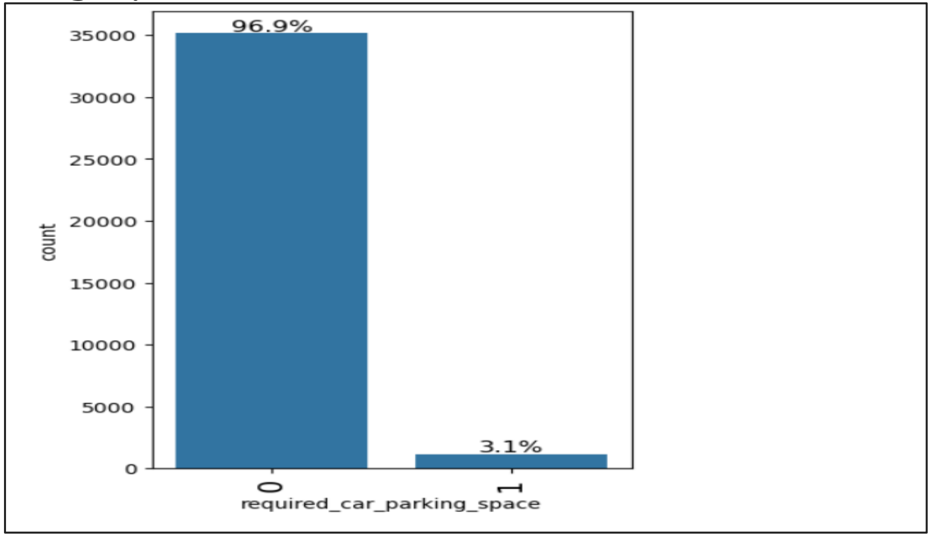


Fig 9: Univariate Analysis of Required Car Parking Space

Observations:

- 96.9% of the bookings did not require a car parking space indicating that car parking is not a common requirement for the vast majority of hotel bookings, with only a small percetage (3.1%) of guests needing it.

Link to Appendix slide on data background check

- ## Type of Meal Plan

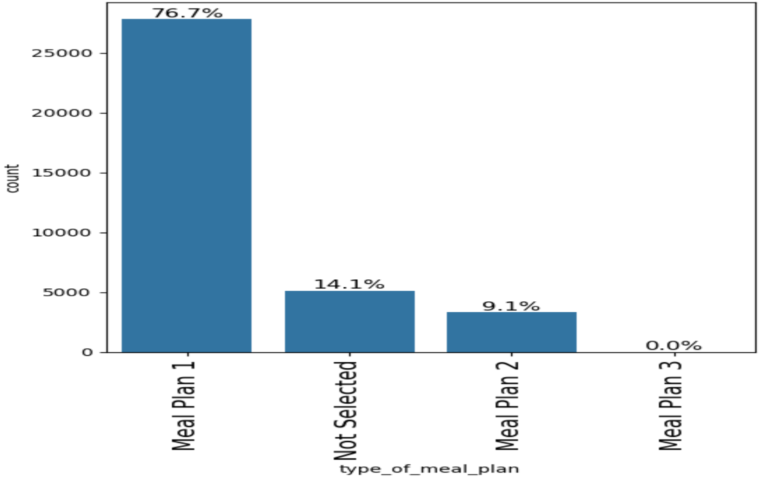


Fig 10: Univariate Analysis of Type of Meal Plan

Observations:

- 76.7%, selected "Meal Plan 1". This indicates a strong preference for this particular meal option
- Meal Plan 1 is clearly the most popular option among guests.
- Meal Plan 2 has limited popularity
- Meal Plan 3 is virtually unused
- A significant portion (14.1%) of guests either didn't choose a meal plan or their choice wasn't recorded.

Link to Appendix slide on data background check

Room Type Reserved

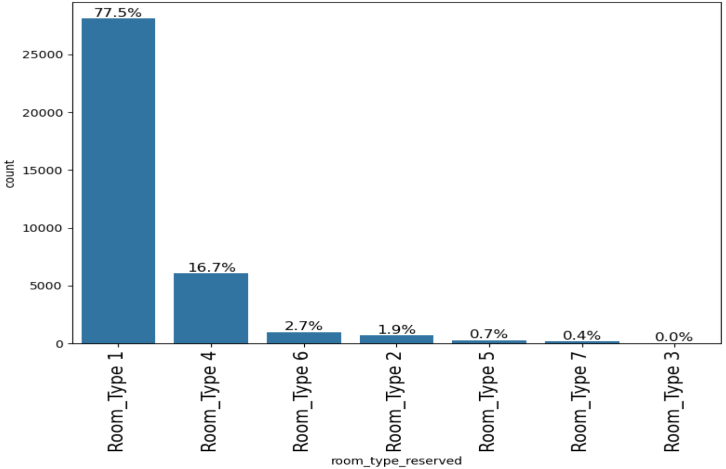


Fig 11: Univariate Analysis of Room Type Reserved

Observations:

- The vast majority of bookings, 77.5%, reserved "Room Type 1". This the most popular choice among guests. This could probably be due to factors like price, size, amenities, or availability
- Room Type 4 being the second most popular
- The other room types have significantly lower reservation rates, with Room Type 3 being almost non-existent in the bookings.

[Link to Appendix slide on data background check]

- **Arrival Month**

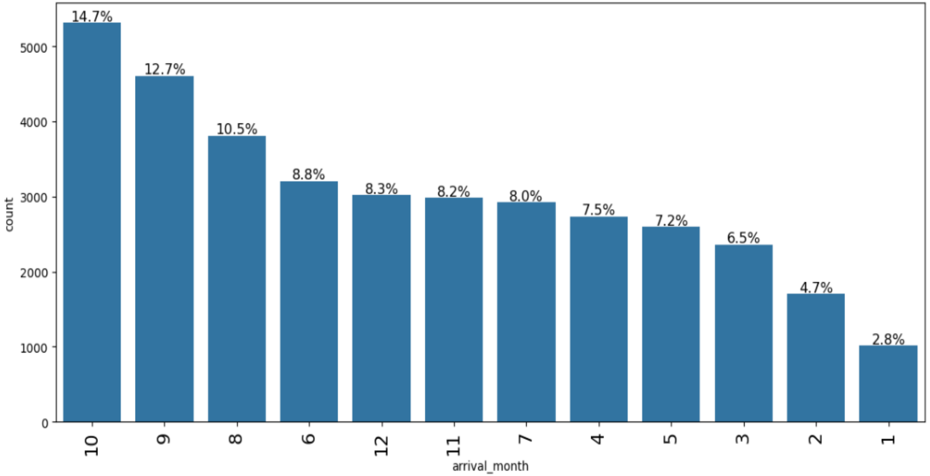


Fig 12: Univariate Analysis of Arrival Month

Observations:

- October has the highest number of arrivals, accounting for 14.7% of all bookings. This suggests October is the busiest month for arrivals.
- September follows as the second most popular arrival month, with 12.7% of bookings.
- This suggest that the hotel experiences a significant peak in arrivals during October and September, with a general trend of higher arrivals in late summer and autumn, moderate arrivals in mid-seasons, and the lowest arrivals during the winter months.

Link to Appendix slide on data background check

Market Segment Type

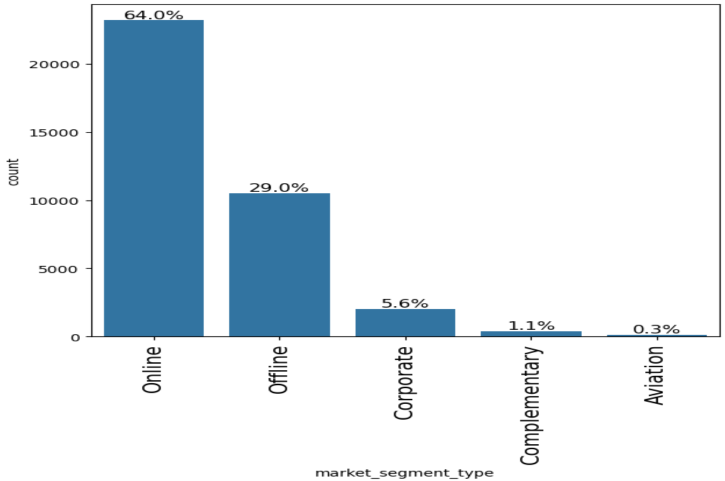


Fig 13: Univariate Analysis of Market Segment Type

Observations:

- The distribution indicates that the hotel heavily relies on online platforms for the majority of its bookings.
- Traditional booking methods still contribute a significant portion of business.
- Bookings from corporate clients and the complementary/aviation sectors are a relatively small part of the overall booking volume.

Link to Appendix slide on data background check

- ## Number of Special Requests

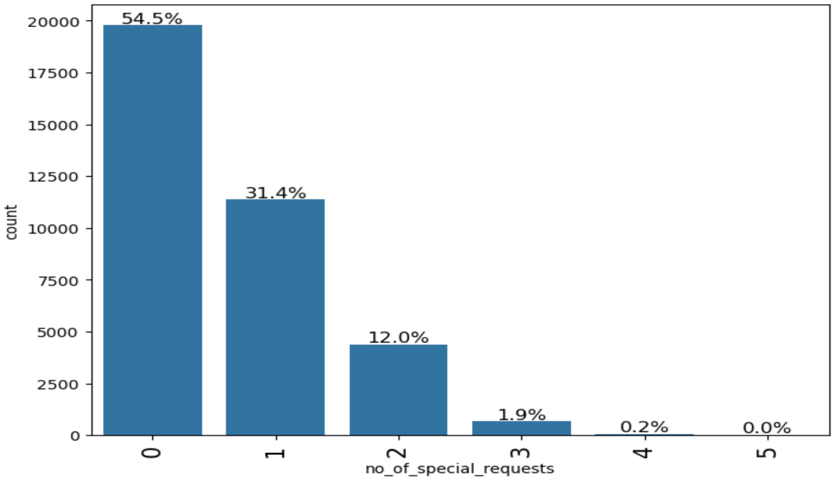


Fig 14: Univariate Analysis on Number of special Requests

Observations:

- The distribution indicates that over half of the guests(54.5%) do not have any specific needs or preferences beyond the standard booking.

- While a significant portion includes special request, multiple special requests are quite rare.

[Link to Appendix slide on data background check](#)

- ## Booking Status

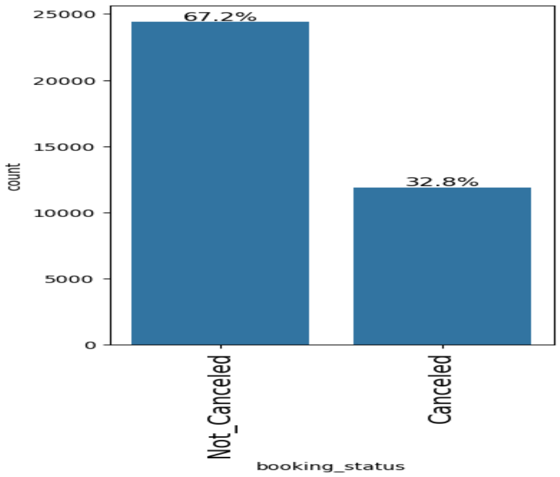


Fig 15: Univariate Analysis of Booking Status

Observations:

- The distribution indicates that the majority of reservations were completed with 67.2% of bookings not cancelled.
- However, 32.8% of bookings cancelled is a very significant figure which may lead to financial and operational challenges for the hotel, as previously noted (loss of revenue, staffing issues, wasted marketing efforts, etc.)

Link to Appendix slide on data background check

EDA – Bivariate Analysis

- **Correlation Check**

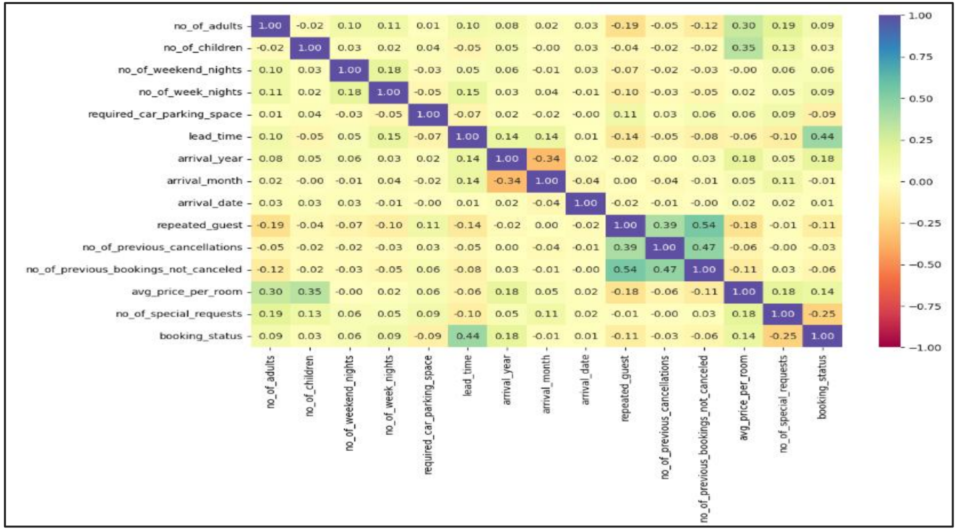


Fig 16: Correlation Plot

Observations:

- Repeated guest has a strong positive correlation with no of previous cancellations. This could imply that guests who book frequently are also more likely to have canceled at some point.
- Repeated Guest has a strong positive correlation with number of previous bookings that were not cancelled. This makes logical sense as repeated guests are those who have completed previous bookings.
- Booking Status has a moderate negative correlation with number of previous booking not cancelled. This indicates that guests with more previous completed bookings are less likely to cancel their current booking.

[Link to Appendix slide on data background check](#)

- **Average Price per Room vs Market Segment**

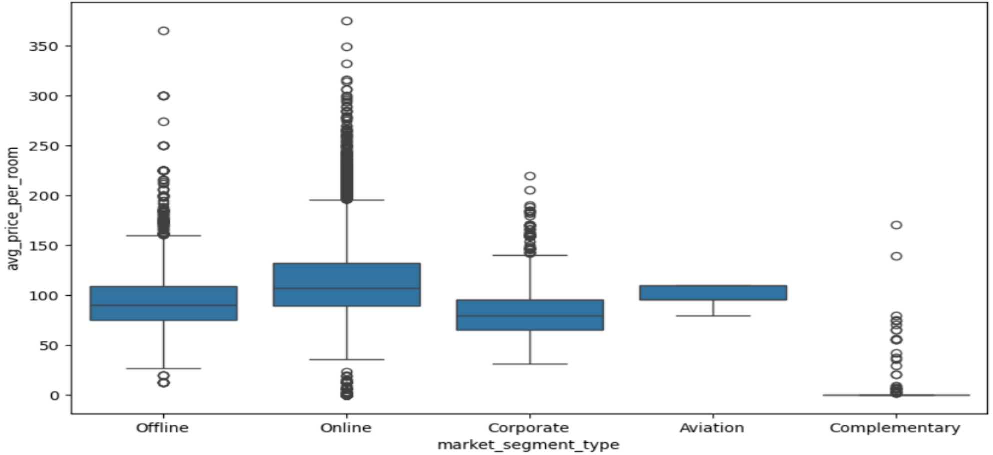


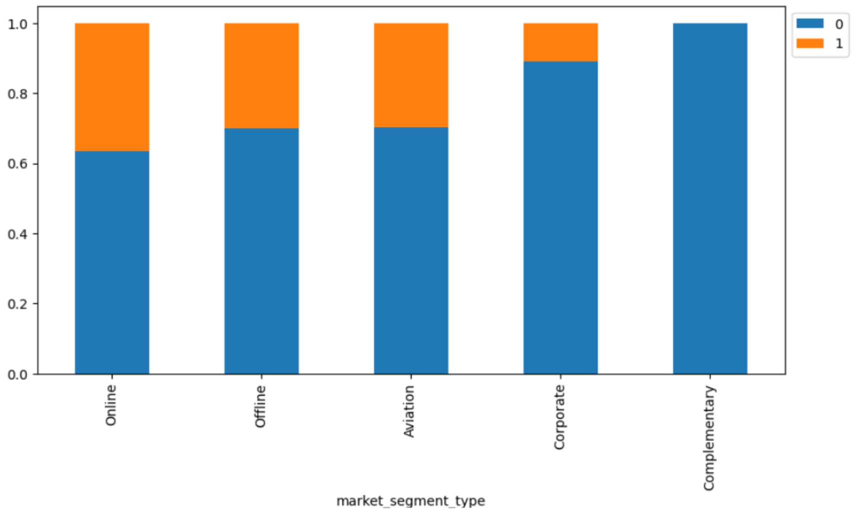
Fig 17: Bivariate Analysis of Average Price per Room vs Market Segment

Observations:

- Both online and offline market segments tend to have higher average room prices compared to corporate bookings
- Corporate bookings generally have lower average room prices and less variability.
- The average price per room for aviation related bookings is quite consistent.
- Complementary bookings are primarily free
- The presence of outliers suggests that there are bookings with significantly higher or lower average room prices within each market segment

Link to Appendix slide on data background check

- **Booking Status vs Market Segments**



booking_status	0	1	All
market_segment_type			
All	24390	11885	36275
Online	14739	8475	23214
Offline	7375	3153	10528
Corporate	1797	220	2017
Aviation	88	37	125
Complementary	391	0	391

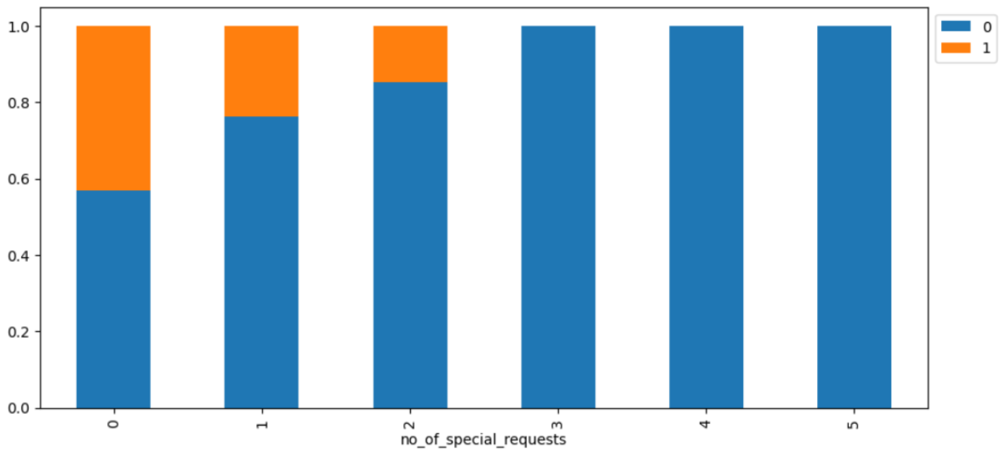
Fig 18: Booking Status vs Market Segments

Observations:

- Cancellation rates vary significantly across various market segment
- A larger percentage of bookings from online channels and the aviation segment tend to be canceled
- Corporate bookings are the most reliable, with the lowest proportion of cancellations.
- The cancellation rate for offline bookings is lower than online and aviation but higher than corporate
- Complementary bookings have a zero-cancellation rate

Link to Appendix slide on data background check

- **Booking Status vs Special Request**



booking_status	0	1	All
no_of_special_requests			
All	24390	11885	36275
0	11232	8545	19777
1	8670	2703	11373
2	3727	637	4364
3	675	0	675
4	78	0	78
5	8	0	8

Fig 19: Booking Status vs Special Request

Observations:

- As the number of special requests increases, the proportion and count of canceled bookings decrease significantly
- Bookings made without any special requests have the highest cancellation rate
- Bookings with three or more special requests have a very low to zero cancellation rate
- This suggest that the number of special requests could be a valuable feature in predicting booking cancellations

[Link to Appendix slide on data background check]

- **No. of Special Request vs Average Price per Room**

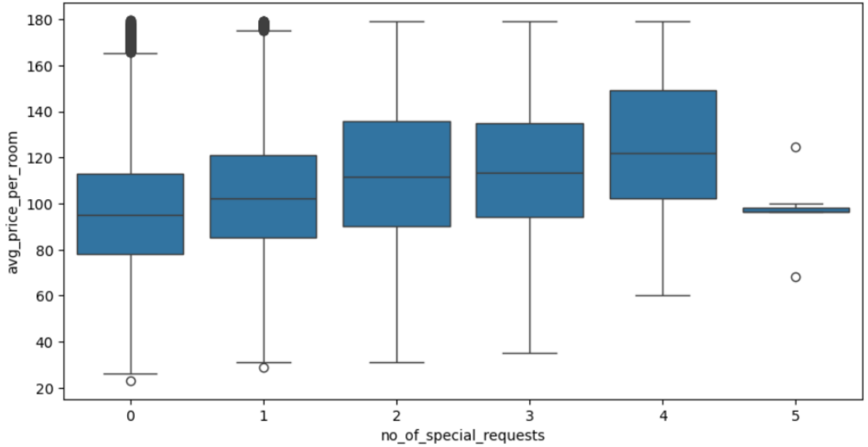


Fig 20: No. of Special Request vs Average Price per Room

Observations:

- The overall distribution of the average price per room increasing as the number of special requests goes up from 0 to 4 suggesting that guests booking more expensive rooms might be more likely to make special requests
- There is more price variability for bookings with more requests
- The unusual pattern for bookings with special request of 5 requires further scrutiny

[Link to Appendix slide on data background check]

Distribution Plot of Booking Status vs Average Price per Room

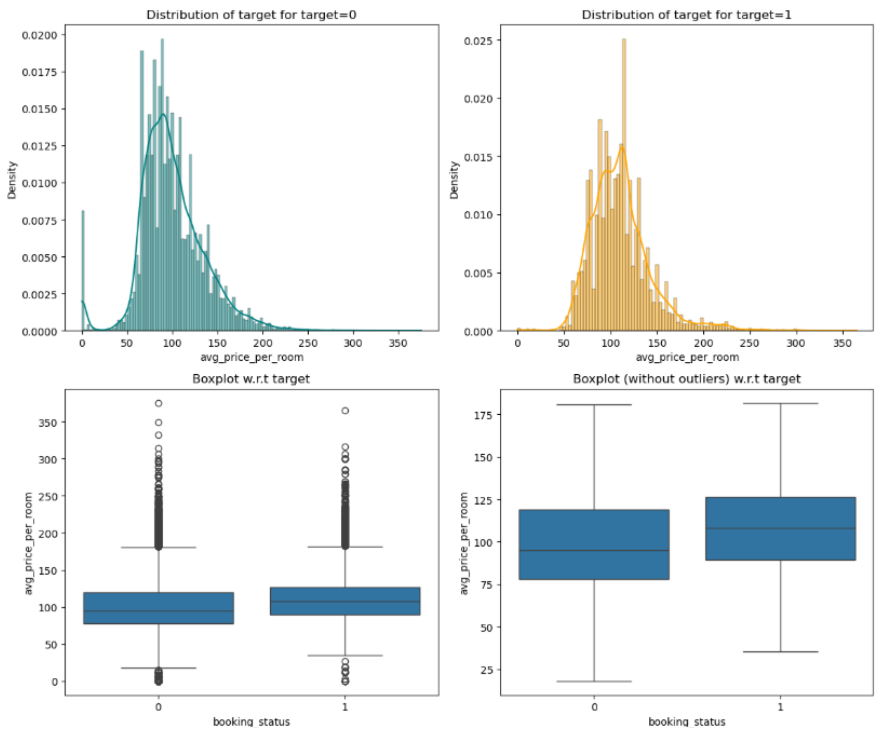


Fig 21: Booking Status vs Average Price per Room

Observations:

- Canceled bookings might have a slightly higher average price per room on average
- The distributions of average prices for both canceled and non-canceled bookings exhibit significant overlap. Cancellations occur across a broad range of prices.
- This aligns with the weak positive correlation coefficient, suggesting that average price per room alone is not a strong predictor of booking cancellation

Link to Appendix slide on data background check

Distribution Plot of Booking Status vs Lead Time

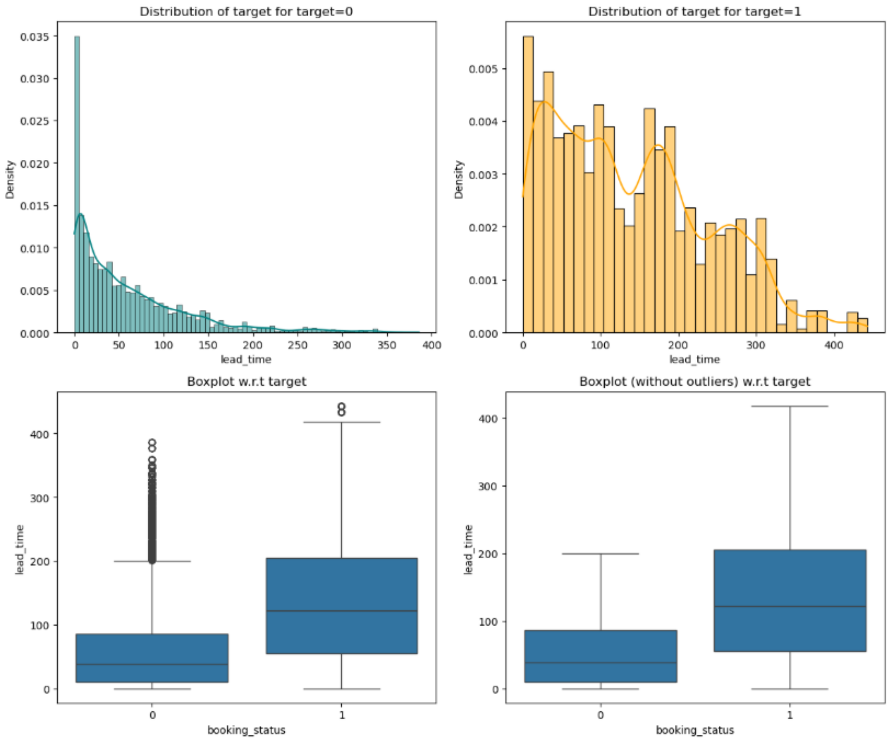


Fig 22: Booking Status vs Lead Time

Observations:

- There is a clear positive relationship between lead time and the likelihood of a booking being canceled
- Bookings made further in advance (longer lead times) have a higher tendency to be canceled compared to bookings made closer to the arrival date (shorter lead times)
- Bookings with short lead times are much more likely to be kept

Link to Appendix slide on data background check

- ## Booking Status vs Number of Family Members

booking_status no_of_family_members	0	1	All
All	18456	9985	28441
2	15506	8213	23719
3	2425	1368	3793
4	514	398	912
5	11	6	17

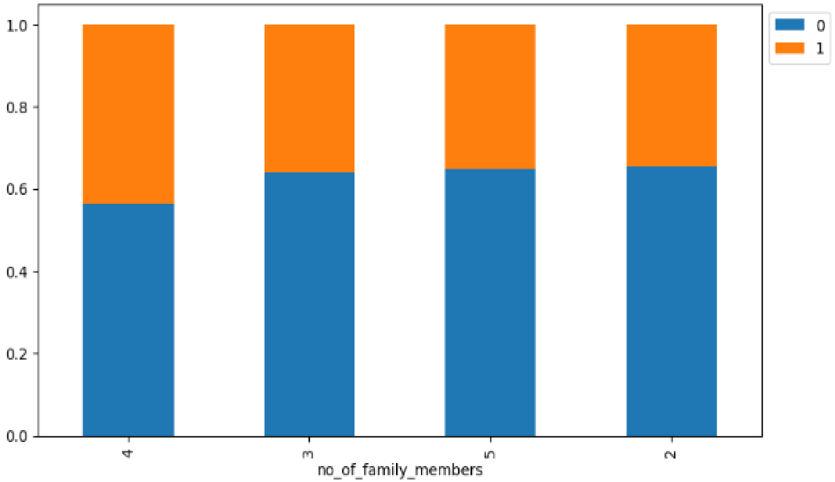


Fig 23: Booking Status vs Number of Family Members

Observations:

- The number of family members appears to have some influence on the booking status
- Bookings with 4 family members show a noticeably higher cancellation rate compared to bookings with 2, 3, or 5 family members
- The cancellation rates for 2, 3, and 5 family members are relatively similar to the overall cancellation rate

Link to Appendix slide on data background check

- **Booking Status vs Total Days**

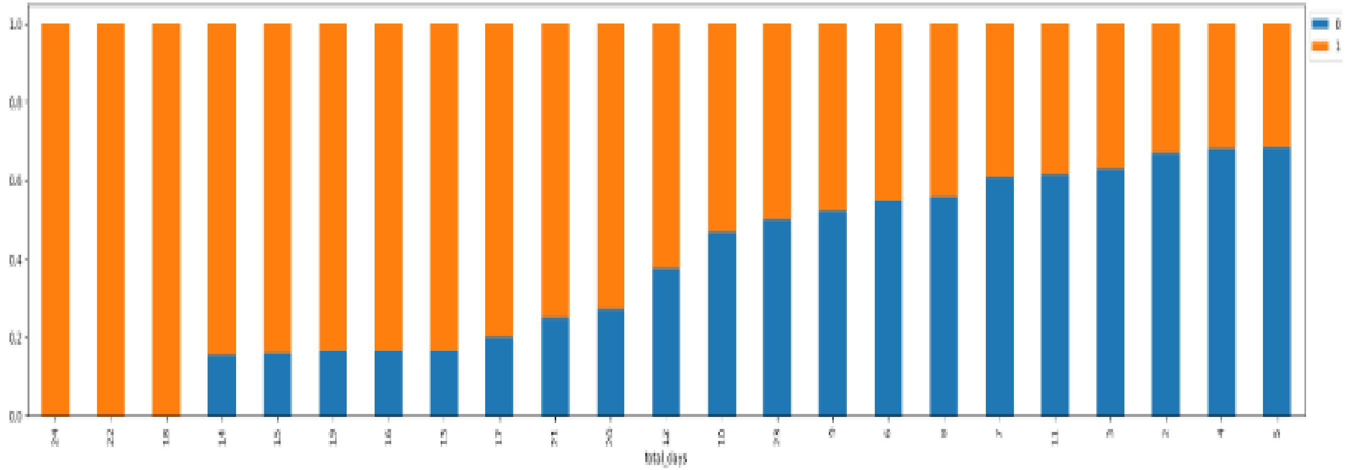


Fig 24: Booking Status vs Total Days

booking_status	0	1	All
total_days			
All	10979	6115	17094
3	3689	2183	5872
4	2977	1387	4364
5	1593	738	2331
2	1301	639	1940
6	566	465	1031
7	590	383	973
8	100	79	179
10	51	58	109
9	58	53	111
14	5	27	32
15	5	26	31
13	3	15	18
12	9	15	24
11	24	15	39
20	3	8	11
19	1	5	6
16	1	5	6
17	1	4	5
18	0	3	3
21	1	3	4
22	0	2	2
23	1	1	2
24	0	1	1

Observations:

- Short Stays between 1-3 days have a significant cancellation rate
- Cancellation rates tend to decrease as the stay duration increases from 1 to around 4 days.
- There's a notable increase in cancellation rates for stays around 5-10 days
- Longer stays beyond 10 days generally have a much lower proportion of cancellations

Link to Appendix slide on data background check

- **Booking Status vs Repeated Guest**

booking_status	0	1	All
repeated_guest			
All	24390	11885	36275
0	23476	11869	35345
1	914	16	930

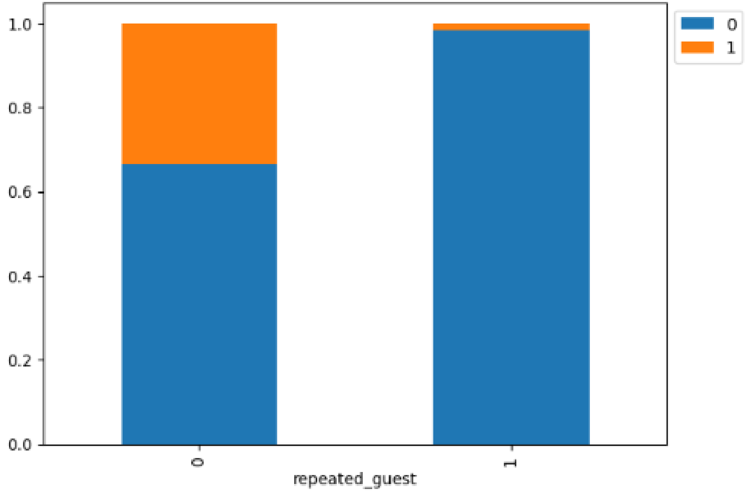


Fig 25: Booking Status vs Repeated Guest

Observations:

- Percentage of repeated guests who canceled their booking: 1.72%
- Whether a guest is a repeat customer has a very strong influence on the booking status.
- Repeated guests are less likely to cancel their bookings compared to new guests
- Non-repeated guests have a cancellation rate that is significantly higher than that of repeated guests and close to the overall cancellation rate
- This highlights the value of customer loyalty
- The "repeated_guest" feature would likely be a very important predictor in a booking cancellation model.

Link to Appendix slide on data background check

- **Busiest Months in the Hotel**

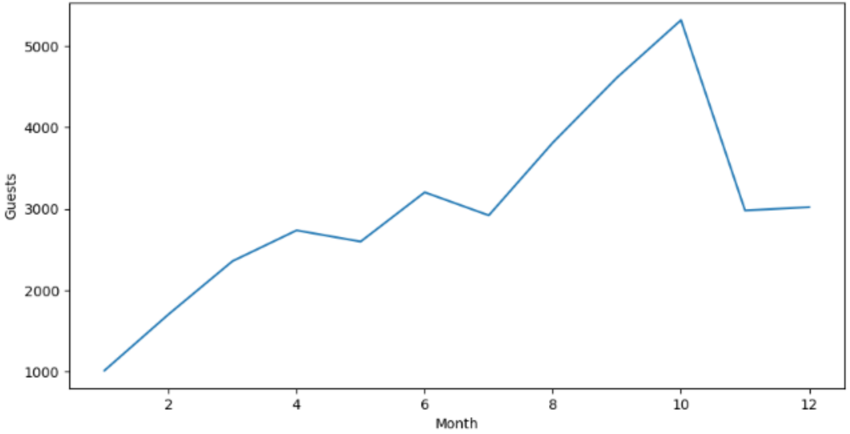


Fig 26: Month vs Number of Guest

Observations:

- There is a clear seasonal pattern in the number of guests
- Month of October stands out as the peak season with the highest number of guests.
- The period leading up to the peak (roughly July to September) also experiences high guest numbers.
- January usually shows the lowest number of guests, likely representing the off-season
- The sharp decline after the peak in Month of October suggests a strong end to the high season.

[*Link to Appendix slide on data background check*]

● Percentage of Bookings Canceled in each Month

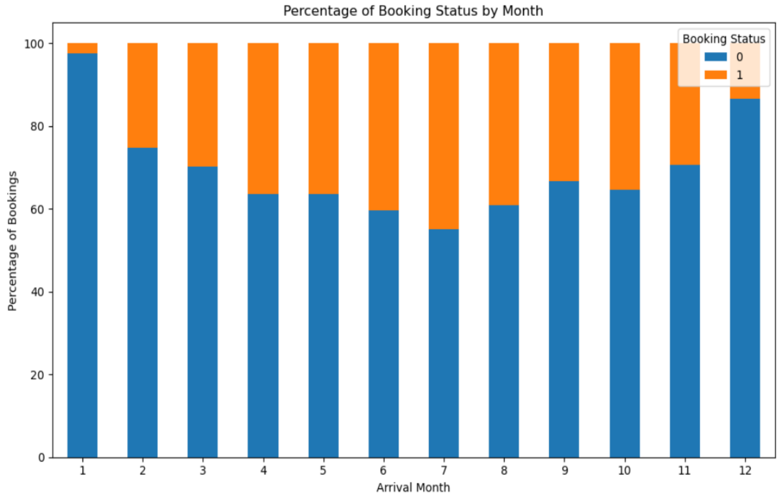


Fig 27: Arrival Month vs Booking Status (%)

arrival_month	Not_Canceled	Canceled	% Not_Canceled	% Canceled
1	990	24	97.63314	2.36686
2	1274	430	74.76526	25.23474
3	1658	700	70.31383	29.68617
4	1741	995	63.63304	36.36696
5	1650	948	63.51039	36.48961
6	1912	1291	59.69404	40.30596
7	1606	1314	55.00000	45.00000
8	2325	1488	60.97561	39.02439
9	3073	1538	66.64498	33.35502
10	3437	1880	64.64172	35.35828
11	2105	875	70.63758	29.36242
12	2619	402	86.69315	13.30685

Observations:

- The cancellation rates vary significantly across different arrival months
- The summer months (June, July, August) tend to have the highest cancellation rates, while the beginning and end of the year (January and December) have the lowest
- This seasonal pattern in cancellations could be related to various factors such as travel trends, weather conditions, or specific events occurring in those months

Link to Appendix slide on data background check

Prices Across different Months

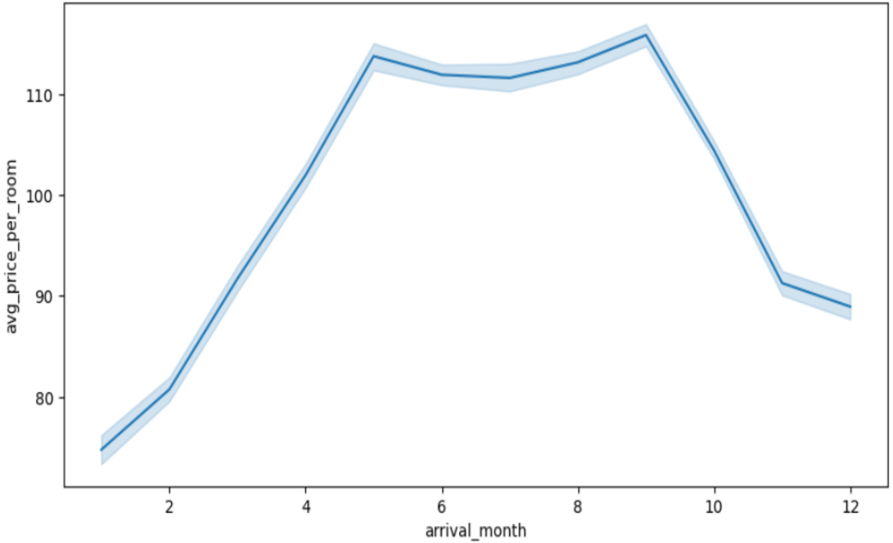


Fig 28: Average Price per Room vs Arrival Month

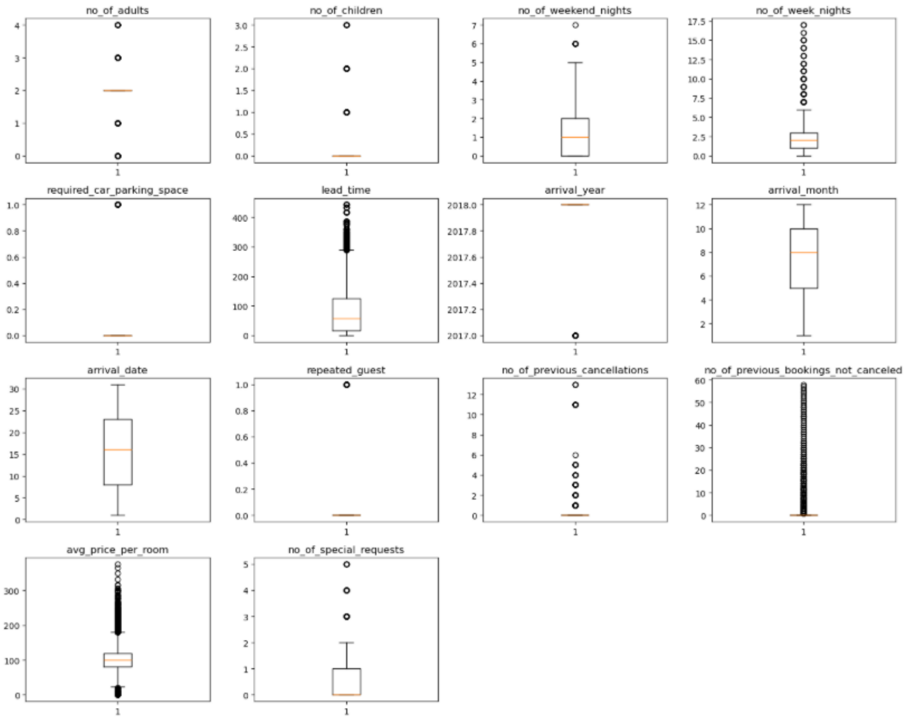
Observations:

- There is a clear seasonal pattern in the average price per room, indicating that demand and pricing strategies vary throughout the year.
- May/June and around September are the periods with the highest average room prices.
- January represents the time with the lowest average room prices
- The higher prices likely correlate with periods of higher demand (as seen in the previous graph of guest numbers), while lower prices correspond to periods of lower demand

Link to Appendix slide on data background check

Data Preprocessing

- **Outlier Check**



- Number of adults appear as having potential outliers.
- Number of children also seems to have potential outliers.
- Numerous bookings with very long lead times (far beyond the typical range) are identified as outliers.
- A significant number of bookings with extended weekday stays appear as outliers
- Bookings with more than 2 weekend nights are flagged as potential outliers
- Guests with a history of multiple cancellations are identified as outliers
- Bookings with a very high number of previous completed stays appear as outliers
- Both unusually high and unusually low average room prices are flagged as outliers

Model Evaluation Criterion

- If we predict that a booking will not be canceled and the booking gets canceled then the hotel will lose resources and will have to bear additional costs of distribution channels.

- If we predict that a booking will get canceled and the booking doesn't get canceled the hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be canceled. This might damage the brand equity.

- In order to reduces losses, the hotel would want F1 Score to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives

Data Preparation for Modeling (Logistic Regression)

Defining Dependent and Independent Variables

```
    const  no_of_adults  no_of_children  no_of_weekend_nights  \
0 1.00000       2.00000         0.00000               1.00000
1 1.00000       2.00000         0.00000               2.00000
2 1.00000       1.00000         0.00000               2.00000
3 1.00000       2.00000         0.00000               0.00000
4 1.00000       2.00000         0.00000               1.00000

   no_of_week_nights  required_car_parking_space  lead_time  arrival_year  \
0            2.00000                     0.00000  224.00000    2017.00000
1            3.00000                     0.00000    5.00000    2018.00000
2            1.00000                     0.00000    1.00000    2018.00000
3            2.00000                     0.00000  211.00000    2018.00000
4            1.00000                     0.00000   48.00000    2018.00000

   arrival_month  arrival_date  repeated_guest  no_of_previous_cancellations  \
0       10.00000       2.00000         0.00000                       0.00000
1       11.00000       6.00000         0.00000                       0.00000
2        2.00000      28.00000         0.00000                       0.00000
3        5.00000      20.00000         0.00000                       0.00000
4        4.00000      11.00000         0.00000                       0.00000

   no_of_previous_bookings_not_canceled  avg_price_per_room  \
0                               0.00000            65.00000
1                               0.00000           106.68000
2                               0.00000            60.00000
3                               0.00000           100.00000
4                               0.00000            94.50000
```

```
   no_of_special_requests  type_of_meal_plan_Meal Plan 2  \
0                 0.00000                        0.00000
1                 1.00000                        0.00000
2                 0.00000                        0.00000
3                 0.00000                        0.00000
4                 0.00000                        0.00000

   type_of_meal_plan_Meal Plan 3  type_of_meal_plan_Not Selected  \
0                        0.00000                         0.00000
1                        0.00000                         1.00000
2                        0.00000                         0.00000
3                        0.00000                         0.00000
4                        0.00000                         1.00000

   room_type_reserved_Room_Type 2  room_type_reserved_Room_Type 3  \
0                         0.00000                         0.00000
1                         0.00000                         0.00000
2                         0.00000                         0.00000
3                         0.00000                         0.00000
4                         0.00000                         0.00000

   room_type_reserved_Room_Type 4  room_type_reserved_Room_Type 5  \
0                         0.00000                         0.00000
1                         0.00000                         0.00000
2                         0.00000                         0.00000
3                         0.00000                         0.00000
4                         0.00000                         0.00000
```

```
   room_type_reserved_Room_Type 6  room_type_reserved_Room_Type 7  \
0                         0.00000                         0.00000
1                         0.00000                         0.00000
2                         0.00000                         0.00000
3                         0.00000                         0.00000
4                         0.00000                         0.00000

   market_segment_type_Complementary  market_segment_type_Corporate  \
0                            0.00000                        0.00000
1                            0.00000                        0.00000
2                            0.00000                        0.00000
3                            0.00000                        0.00000
4                            0.00000                        0.00000

   market_segment_type_Offline  market_segment_type_Online
0                      1.00000                     0.00000
1                      0.00000                     1.00000
2                      0.00000                     1.00000
3                      0.00000                     1.00000
4                      0.00000                     1.00000

0    0
1    0
2    1
3    1
4    1
Name: booking_status, dtype: int64
```

const	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date
1.00000	2.00000	0.00000	1.00000	2.00000	0.00000	224.00000	2017.00000	10.00000	2.00000
1.00000	2.00000	0.00000	2.00000	3.00000	0.00000	5.00000	2018.00000	11.00000	6.00000
1.00000	1.00000	0.00000	2.00000	1.00000	0.00000	1.00000	2018.00000	2.00000	28.00000
1.00000	2.00000	0.00000	0.00000	2.00000	0.00000	211.00000	2018.00000	5.00000	20.00000
1.00000	2.00000	0.00000	1.00000	1.00000	0.00000	48.00000	2018.00000	4.00000	11.00000

Intercept added to data and creating dummies for independent features

required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled
0.00000	224.00000	2017.00000	10.00000	2.00000	0.00000	0.00000	0.00000
0.00000	5.00000	2018.00000	11.00000	6.00000	0.00000	0.00000	0.00000
0.00000	1.00000	2018.00000	2.00000	28.00000	0.00000	0.00000	0.00000
0.00000	211.00000	2018.00000	5.00000	20.00000	0.00000	0.00000	0.00000
0.00000	48.00000	2018.00000	4.00000	11.00000	0.00000	0.00000	0.00000

Input attributes converted into float type for modeling

Splitting the data in 70:30 ratio for train to test data

Number of rows in train data = 25392
Number of rows in test data = 10883

Percentages of classes in Train and Test set: Booking Status

```
Percentage of classes in training set:
booking_status
0    0.67064
1    0.32936
Name: proportion, dtype: float64
Percentage of classes in test set:
booking_status
0    0.67638
1    0.32362
Name: proportion, dtype: float64
```

Model Performance Summary

- **Overview of ML model and its parameters**

Objective: To develop a predictive machine learning model that can accurately identify hotel bookings with a high likelihood of cancellation prior to the check-in date in order to proactively manage high-risk reservations, reduce lost revenue, optimize resource allocation and support data-driven policy decisions regarding cancellation terms and guest segmentation

Model Type: Logistic Regression (Logit)

- The model used for this business problem is a Supervised Learning algorithm, specifically a Binary Classification model

- The goal is to estimates the probability of cancellation based on historical patterns and input features. This makes it not only predictive but also interpretable, helping INN Hotels Group understand which factors contribute most to cancellations.

Key Model Parameters:

- **Dependent Variable**: booking_status

- **Independent Variables**: no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, type_of_meal_plan, required_car_parking_space, room_type_reserved, lead_time, arrival_year, arrival_month, arrival_date, market_segment_type, repeated_guest, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests

- **fit_intercept, penalty, C, solver, class_weight, random_state and Max_iter**

- **Summary of most important features used by the ML model for prediction**

    - **Market segment type** is the strongest predictor in the model. “Complementary” and “Offline” guests are far less likely to cancel.

    - **Lead time** is a major positive predictor. Bookings made far in advance are more likely to be canceled

    - **Room type and special requests and previous behavior** reflects guest commitment—certain room types and more requests imply reduced cancellation risk

    - **Price and Amenities** like meal plans and parking have moderate effects:

    - **Demographic features** (adults/children) have minimal influence

Summary of key performance metrics for training and test data (Logistic Regression)

Training performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Test performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

- The performance metrics (Accuracy, Recall, Precision, F1-score) are relatively consistent between the training and test sets for each threshold. This is a good sign, suggesting that the model is generalizing reasonably well to unseen data and not overfitting dramatically

- Changing the probability threshold for classification significantly affects Recall and Precision, and consequently the F1-score

- Accuracy is less sensitive to these threshold changes in this case.

• Summary of key performance metrics for training and test data (Decision Tree)

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99437	0.83554	0.90981
Recall	0.98570	0.78339	0.91862
Precision	0.99708	0.73299	0.82572
F1	0.99136	0.75735	0.86969

Test performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.86529	0.83212	0.86015
Recall	0.79445	0.76921	0.84044
Precision	0.79445	0.73205	0.75873
F1	0.79445	0.75017	0.79750

- The Decision Tree shows perfect or near-perfect performance on the training data, however, its performance drops significantly on the test data. This is a clear indication of severe overfitting
- The tree has learned the training data too well, including its noise, and doesn't generalize well to new, unseen data.
- Both pre-pruning and post-pruning techniques appear to have a positive impact on the model's ability to generalize
- The Decision Tree (Post-Pruning) consistently shows better performance than the Decision Tree (Pre-Pruning) on both the training and test sets across all metrics

APPENDIX

Data Background and Contents

- **Data Overview**

The dataset consists of 36275 rows and 19 columns, representing data information about hotel booking records for the INN Hotels Group in Portugal offering a detailed look at customer booking behavior, preferences, and outcomes

Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time
INN36271	3	0	2	6	Meal Plan 1	0	Room_Type 4	85
INN36272	2	0	1	3	Meal Plan 1	0	Room_Type 1	228
INN36273	2	0	2	6	Meal Plan 1	0	Room_Type 1	148
INN36274	2	0	0	3	Not Selected	0	Room_Type 1	63
INN36275	2	0	1	2	Meal Plan 1	0	Room_Type 1	207

Table 1: Top 5 rows of the Dataset

- **Data Background**

The dataset INNHotelsGroup was used in the preparation of machine learning-based model for predicting booking cancellations. The goal is to analyze booking patterns, identify factors that lead to cancellations and build a model to predict the likelihood of cancellation in advance

- **Data Contents**

```
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Booking_ID                            36275 non-null  object 
 1   no_of_adults                          36275 non-null  int64  
 2   no_of_children                        36275 non-null  int64  
 3   no_of_weekend_nights                  36275 non-null  int64  
 4   no_of_week_nights                     36275 non-null  int64  
 5   type_of_meal_plan                     36275 non-null  object 
 6   required_car_parking_space            36275 non-null  int64  
 7   room_type_reserved                    36275 non-null  object 
 8   lead_time                             36275 non-null  int64  
 9   arrival_year                          36275 non-null  int64  
 10  arrival_month                         36275 non-null  int64  
 11  arrival_date                          36275 non-null  int64  
 12  market_segment_type                   36275 non-null  object 
 13  repeated_guest                        36275 non-null  int64  
 14  no_of_previous_cancellations          36275 non-null  int64  
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64  
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64  
 18  booking_status                        36275 non-null  object 
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

Table 2: Information on the Data Set

There are three datatypes namely: float64(1), int64(13) and object(5) with 14 numerical and 5 categorical (strings). The target variable is the booking_status, which is of object type.

- Booking_ID: 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 week nights (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)
- no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- 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.
- lead_time: Number of days between the date of booking and the arrival date

- type_of_meal_plan: Type of meal plan booked by the customer:

- no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

- 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.

- 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.

Statistical Summary

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000	10.00000
no_of_weekend_nights	36275.00000	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
required_car_parking_space	36275.00000	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
lead_time	36275.00000	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
repeated_guest	36275.00000	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000	540.00000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000

Table 3: Statistical Summary of the Dataset

- **Observations**

  - **Adults:** Average number is ~1.84 (mostly solo or couple travelers), Max is 4, indicating small group bookings.

  - **Children:** Mean is very low (~0.11), and 75% of bookings have 0 children. Outliers exist, with some bookings having up to 10 children

  - **Length of Stay**: Mean is ~0.81; stays tend to include 0–2 weekend nights and max is 7 nights, possibly full week bookings starting on the weekend whereas mean is ~2.2; median is 2, indicating short weekday stays. Some stays extend up to 17 nights

  - **Parking & Requests:** Only ~3% of bookings request parking. Special request has a mean ~0.62 meaning most bookings have 0–1 request. Max of 5 indicates highly customized guest expectations for some.

  - **Booking Behavior:** Highly variable (std ~86 days), ranging from same-day to 443 days in advance. Median is 57 days: most bookings are made 2 months in advance. Repeated guest show extremely low mean indicating that most booking are from new guest. Previous Cancellations has very low mean (~0.02); however, some guests have up to 13 previous cancellations.

```
no_of_adults                            0
no_of_children                          0
no_of_weekend_nights                    0
no_of_week_nights                       0
type_of_meal_plan                       0
required_car_parking_space              0
room_type_reserved                      0
lead_time                               0
arrival_year                            0
arrival_month                           0
arrival_date                            0
market_segment_type                     0
repeated_guest                          0
no_of_previous_cancellations            0
no_of_previous_bookings_not_canceled    0
avg_price_per_room                      0
no_of_special_requests                  0
booking_status                          0
dtype: int64
```

- There are no missing values in the dataset

Model Building - Logistic Regression

- **First Model (lg)**

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25364
Method:                           MLE   Df Model:                           27
Date:                Fri, 09 May 2025   Pseudo R-squ.:                  0.3292
Time:                        04:06:42   Log-Likelihood:                -10794.
converged:                      False   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                -922.8266    120.832     -7.637      0.000   -1159.653    -686.000
no_of_adults                            0.1137      0.038      3.019      0.003       0.040       0.188
no_of_children                          0.1580      0.062      2.544      0.011       0.036       0.280
no_of_weekend_nights                    0.1067      0.020      5.395      0.000       0.068       0.145
no_of_week_nights                       0.0397      0.012      3.235      0.001       0.016       0.064
required_car_parking_space             -1.5943      0.138    -11.565      0.000      -1.865      -1.324
lead_time                               0.0157      0.000     58.863      0.000       0.015       0.016
arrival_year                            0.4561      0.060      7.617      0.000       0.339       0.573
arrival_month                          -0.0417      0.006     -6.441      0.000      -0.054      -0.029
arrival_date                            0.0005      0.002      0.259      0.796      -0.003       0.004
repeated_guest                         -2.3472      0.617     -3.806      0.000      -3.556      -1.139
no_of_previous_cancellations            0.2664      0.086      3.108      0.002       0.098       0.434
no_of_previous_bookings_not_canceled   -0.1727      0.153     -1.131      0.258      -0.472       0.127
avg_price_per_room                      0.0188      0.001     25.396      0.000       0.017       0.020
no_of_special_requests                 -1.4689      0.030    -48.782      0.000      -1.528      -1.410
type_of_meal_plan_Meal Plan 2           0.1756      0.067      2.636      0.008       0.045       0.306
type_of_meal_plan_Meal Plan 3          17.3584   3987.836      0.004      0.997   -7798.656    7833.373
type_of_meal_plan_Not Selected          0.2784      0.053      5.247      0.000       0.174       0.382
room_type_reserved_Room_Type 2         -0.3605      0.131     -2.748      0.006      -0.618      -0.103
room_type_reserved_Room_Type 3         -0.0012      1.310     -0.001      0.999      -2.568       2.566
room_type_reserved_Room_Type 4         -0.2823      0.053     -5.304      0.000      -0.387      -0.178
room_type_reserved_Room_Type 5         -0.7189      0.209     -3.438      0.001      -1.129      -0.309
room_type_reserved_Room_Type 6         -0.9501      0.151     -6.274      0.000      -1.247      -0.653
room_type_reserved_Room_Type 7         -1.4003      0.294     -4.770      0.000      -1.976      -0.825
market_segment_type_Complementary     -40.5975   5.65e+05  -7.19e-05      1.000   -1.11e+06    1.11e+06
market_segment_type_Corporate          -1.1924      0.266     -4.483      0.000      -1.714      -0.671
market_segment_type_Offline            -2.1946      0.255     -8.621      0.000      -2.694      -1.696
market_segment_type_Online             -0.3995      0.251     -1.590      0.112      -0.892       0.093
```

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80600	0.63410	0.73971	0.68285

- **Converged False:** The model did not converge successfully, meaning there might be high correlation between independent variables
- **Model Fit:** Pseudo R-squared (0.3292) indicates that the model explains approximately 32.92% of the variation in the log-odds of booking cancellation.
- **Statistical Significance:** The LLR p-value (0.000) suggests the model is statistically significant
- **Accuracy:** The model correctly predicts the outcome (cancellation or not) for 80.6% of the training data
- **Recall:** Of all the actual cancellations, the model correctly identified 63.41%. This indicates the model's ability to capture the positive class (cancellations)
- **Precision:** Out of all the bookings that the model predicted as cancellations, 73.971% were actually cancellations. This shows the model's reliability when it predicts a cancellation
- **F1 Score:** The F1 score is 0.68285. It provides a balanced measure of precision and recall.

In summary therefore the model shows a good overall accuracy, however, there's a noticeable difference between precision and recall. The model is more precise in predicting cancellations but has a moderate recall, indicating it misses a significant portion of actual cancellations

Checking Logistic Regression Assumptions

Test for Multicollinearity

	feature	VIF
0	const	39497686.20788
1	no_of_adults	1.35113
2	no_of_children	2.09358
3	no_of_weekend_nights	1.06948
4	no_of_week_nights	1.09571
5	required_car_parking_space	1.03997
6	lead_time	1.39517
7	arrival_year	1.43190
8	arrival_month	1.27633
9	arrival_date	1.00679
10	repeated_guest	1.78358
11	no_of_previous_cancellations	1.39569
12	no_of_previous_bookings_not_canceled	1.65200
13	avg_price_per_room	2.06860
14	no_of_special_requests	1.24798
15	type_of_meal_plan_Meal Plan 2	1.27328
16	type_of_meal_plan_Meal Plan 3	1.02526
17	type_of_meal_plan_Not Selected	1.27306
18	room_type_reserved_Room_Type 2	1.10595
19	room_type_reserved_Room_Type 3	1.00330
20	room_type_reserved_Room_Type 4	1.36361
21	room_type_reserved_Room_Type 5	1.02800
22	room_type_reserved_Room_Type 6	2.05614
23	room_type_reserved_Room_Type 7	1.11816
24	market_segment_type_Complementary	4.50276
25	market_segment_type_Corporate	16.92829
26	market_segment_type_Offline	64.11564
27	market_segment_type_Online	71.18026

There are multiple columns with very high VIF values, indicating presence of strong multicollinearity.

Removing Multicollinearity

Removing p values >0.05

Below are the selected_features after removing p>0.05:

['const', 'no_of_adults', 'no_of_children', 'no_of_weekend_nights', 'no_of_week_nights', 'required_car_parking_space', 'lead_time', 'arrival_year', 'arrival_month', 'repeated_guest', 'no_of_previous_cancellations', 'avg_price_per_room', 'no_of_special_requests', 'type_of_meal_plan_Meal Plan 2', 'type_of_meal_plan_Not Selected', 'room_type_reserved_Room_Type 2', 'room_type_reserved_Room_Type 4', 'room_type_reserved_Room_Type 5', 'room_type_reserved_Room_Type 6', 'room_type_reserved_Room_Type 7', 'market_segment_type_Corporate', 'market_segment_type_Offline’]

• Second Model (lg1)

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25370
Method:                           MLE   Df Model:                           21
Date:                Fri, 09 May 2025   Pseudo R-squ.:                  0.3282
Time:                        04:07:14   Log-Likelihood:                -10810.
converged:                       True   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=========================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                           -915.6391    120.471     -7.600      0.000   -1151.758    -679.520
no_of_adults                       0.1088      0.037      2.914      0.004       0.036       0.182
no_of_children                     0.1531      0.062      2.470      0.014       0.032       0.275
no_of_weekend_nights               0.1086      0.020      5.498      0.000       0.070       0.147
no_of_week_nights                  0.0417      0.012      3.399      0.001       0.018       0.066
required_car_parking_space        -1.5947      0.138    -11.564      0.000      -1.865      -1.324
lead_time                          0.0157      0.000     59.213      0.000       0.015       0.016
arrival_year                       0.4523      0.060      7.576      0.000       0.335       0.569
arrival_month                     -0.0425      0.006     -6.591      0.000      -0.055      -0.030
repeated_guest                    -2.7367      0.557     -4.916      0.000      -3.828      -1.646
no_of_previous_cancellations       0.2288      0.077      2.983      0.003       0.078       0.379
avg_price_per_room                 0.0192      0.001     26.336      0.000       0.018       0.021
no_of_special_requests            -1.4698      0.030    -48.884      0.000      -1.529      -1.411
type_of_meal_plan_Meal Plan 2      0.1642      0.067      2.469      0.014       0.034       0.295
type_of_meal_plan_Not Selected     0.2860      0.053      5.406      0.000       0.182       0.390
room_type_reserved_Room_Type 2    -0.3552      0.131     -2.709      0.007      -0.612      -0.098
room_type_reserved_Room_Type 4    -0.2828      0.053     -5.330      0.000      -0.387      -0.179
room_type_reserved_Room_Type 5    -0.7364      0.208     -3.535      0.000      -1.145      -0.328
room_type_reserved_Room_Type 6    -0.9682      0.151     -6.403      0.000      -1.265      -0.672
room_type_reserved_Room_Type 7    -1.4343      0.293     -4.892      0.000      -2.009      -0.860
market_segment_type_Corporate     -0.7913      0.103     -7.692      0.000      -0.993      -0.590
market_segment_type_Offline       -1.7854      0.052    -34.363      0.000      -1.887      -1.684
=========================================================================================
```

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

- **Successful Convergence:** The model converged successfully, which is good
- **Model Fit:** The Pseudo R-squared (0.3282) indicates the model explains a moderate amount of the variance in booking cancellations.
- **Statistical Significance:** The LLR p-value (0.000) suggests the model is statistically significant.
- **Accuracy:** The model correctly predicts booking status about 80.55% of the time on the training data.
- **Recall:** The model correctly identifies 63.27% of actual cancellations
- **Precision:** When the model predicts a cancellation, it's correct about 73.91% of the time
- **F1:** The F1 score balances precision and recall

No categorical feature has p-value greater than 0.05, so we'll consider the features in X_train1 as the final ones and lg1 as final model

Converting coefficients to odds

Below is the output after converting coefficients to odds:

	const	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month
Odds	0.00000	1.11491	1.16546	1.11470	1.04258	0.20296	1.01583	1.57195	0.95839
Change_odd%	-100.00000	11.49096	16.54593	11.46966	4.25841	-79.70395	1.58331	57.19508	-4.16120

Coefficient Interpretations

- **lead_time:** For each additional day of lead time (how far in advance the booking was made), the odds of cancellation increase by approximately 1.58%
- **required_car_parking_space:** Bookings that require a car parking space are significantly less likely to be canceled. The odds of cancellation are reduced by approximately 79.70%
- **no_of_adults:** For each additional adult in the booking, the odds of cancellation increase by approximately 11.49%
- **no_of_children:** For each additional child in the booking, the odds of cancellation increase by approximately 16.55%
- **no_of_weekend_nights:** For each additional weekend night in the stay, the odds of cancellation increase by approximately 11.47%
- **no_of_week_nights:** For each additional week night in the stay, the odds of cancellation increase by approximately 4.26%
- **arrival_year:** Bookings made in a later year have significantly higher odds of cancellation
- **arrival_month:** The odds ratio is 0.95839. The odds of cancellation slightly decrease

repeated_guest	no_of_previous_cancellations	avg_price_per_room	no_of_special_requests	type_of_meal_plan_Meal Plan 2	type_of_meal_plan_Not Selected
0.06478	1.25712	1.01937	0.22996	1.17846	1.33109
-93.52180	25.71181	1.93684	-77.00374	17.84641	33.10947

Coefficient Interpretations

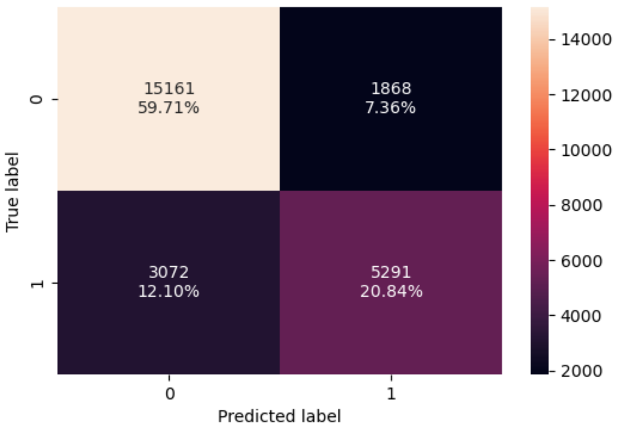
- **repeated_guest:** The odds of cancellation for a repeated guest are about 93.52% lower than for a non-repeated guest
- **no_of_previous_cancellations:** For each additional previous cancellation by the guest, the odds of cancellation for the current booking increase by approximately 25.71%
- **avg_price_per_room:** For each unit increase in the average price per room, the odds of cancellation increase by approximately 1.94%
- **no_of_special_requests:** For each additional special request made by the guest, the odds of cancellation decrease significantly by approximately 77%
- **type_of_meal_plan_Meal Plan 2:** Bookings with "Meal Plan 2" have about 17.85% higher odds of cancellation compared t the reference meal plan
- **type_of_meal_plan_Not Selected:** Bookings where a meal plan was not selected have about 33.11% higher odds of cancellation compared to the reference meal plan.Sources and related content

room_type_reserved_Room_Type 2	room_type_reserved_Room_Type 4	room_type_reserved_Room_Type 5	room_type_reserved_Room_Type 6	room_type_reserved_Room_Type 7
0.70104	0.75364	0.47885	0.37977	0.23827
-29.89588	-24.63551	-52.11548	-62.02290	-76.17294

Coefficient Interpretations

- **room_type_reserved_Room_Type_2: B**ookings for "Room Type 2" are less likely to be canceled compared to the reference room type. The odds of cancellation are reduced by approximately 29.90%
- **room_type_reserved_Room_Type_4:** Bookings for "Room Type 4" are also less likely to be canceled. The odds of cancellation are reduced by approximately 24.64%
- **room_type_reserved_Room_Type_5:** Bookings for "Room Type 5" are significantly less likely to be canceled. The odds of cancellation are reduced by approximately 52.12%
- **room_type_reserved_Room_Type_6:** Bookings for "Room Type 6" are even less likely to be canceled. The odds of cancellation are reduced by approximately 62.02%
- **room_type_reserved_Room_Type_7:** Bookings for "Room Type 7" are the least likely to be canceled among these room types. The odds of cancellation are reduced by approximately 76.17%
- **market_segment_type_Corporate:** Bookings from the "Corporate" market segment are less likely to be canceled compared to the reference market segment. The odds of cancellation for corporate bookings are reduced by approximately 54.67%
- **market_segment_type_Offline**: Bookings from the "Offline" market segment are significantly less likely to be canceled. The odds of cancellation for offline bookings are reduced by approximately 83.23%.

- **Model Performance Check**
  - Checking Model Performance on the Training Set (Confusion Matrix)



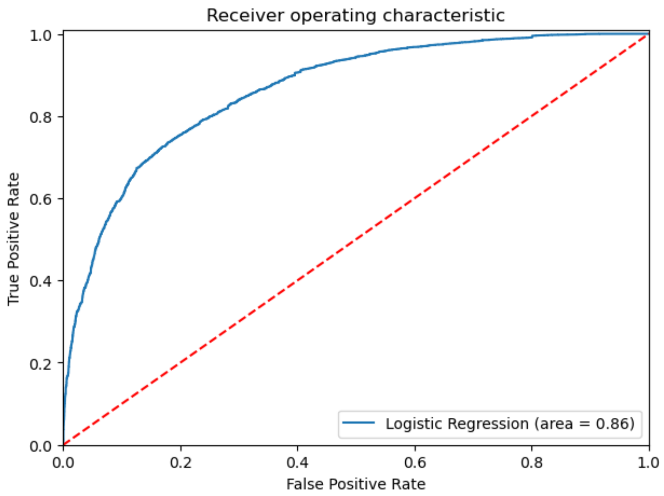
Training performance:

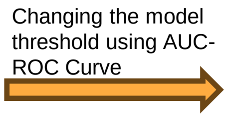
	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

- **Accuracy:** The model correctly predicts the booking status (canceled or not) for approximately 80.55% of the training samples.
- **Recall:** The model identifies 63.27% of all the actual canceled bookings
- **Precision:** When the model predicts a booking will be canceled, it is correct about 73.91% of the time
- **F1 Score:** The F1 score, which balances precision and recall, is 0.68174

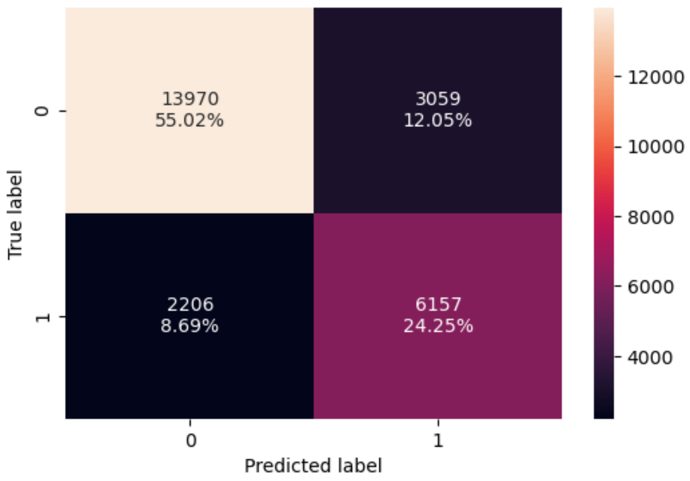
We will now try to improve the performance of the model

Model Performance Improvement (ROC Curve and ROC-AUC) – Training Set





Confusion Matrix using optimal_threshold_auc_roc=0.37

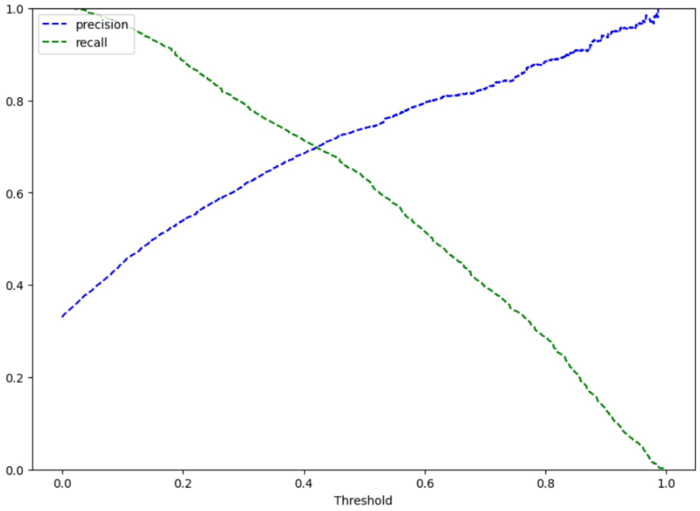


Logistic Regression model is giving a good performance on training set

Training performance:

	Accuracy	Recall	Precision	F1
0	0.79265	0.73622	0.66808	0.70049

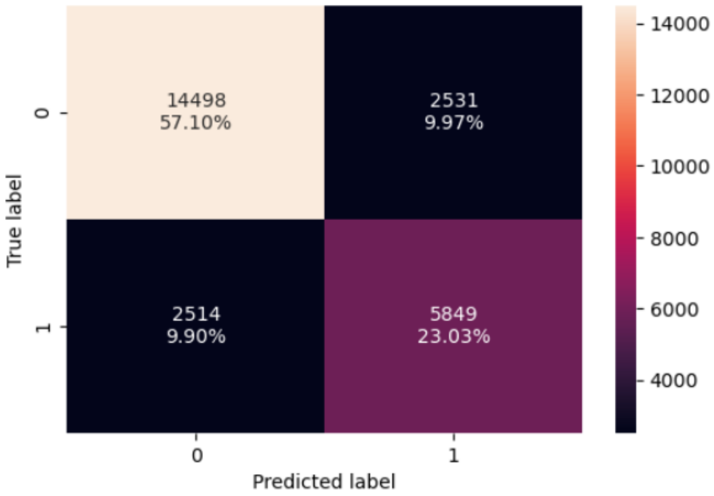
- Model Performance Improvement (Precision-Recall Curve) – Training Set



Precision-Recall curve given a threshold =0.42

Changing the model threshold using Precision-Recall curve

Confusion Matrix using optimal_threshold_curve=0.42

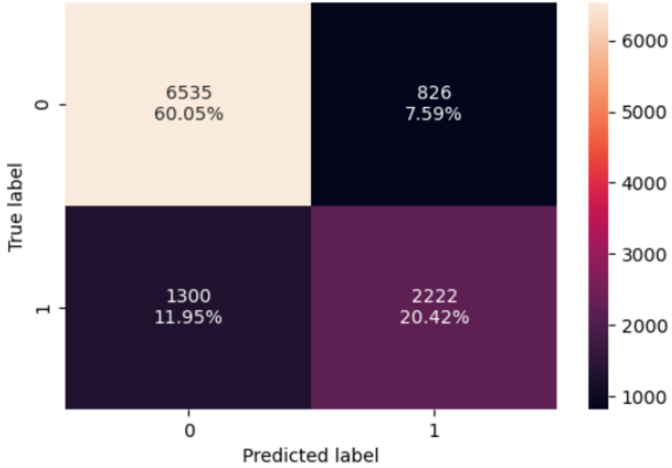


Training performance:

	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

Checking Model Performance on the Test Set (Confusion Matrix)

Confusion Matrix using default threshold =0.5



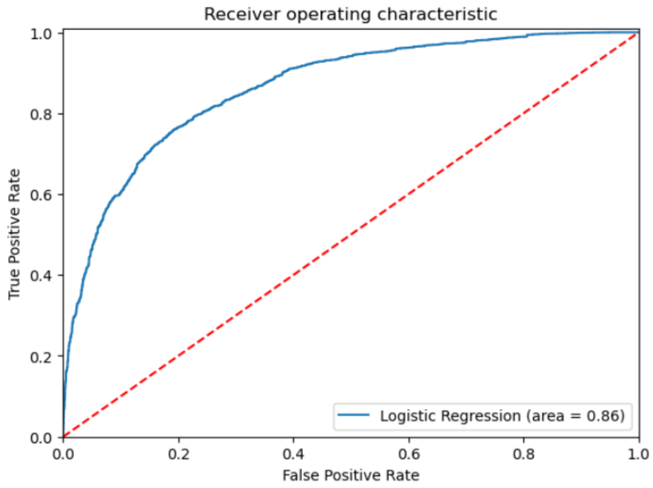
Test performance:

	Accuracy	Recall	Precision	F1
0	0.80465	0.63089	0.72900	0.67641

- **Accuracy:** The model correctly classified 80.47% of all bookings in the test set
- **Recall:** The model correctly identified 63.09% of all actual canceled bookings. This indicates the model's ability to capture the positive class (cancellations)
- **Precision:** Out of all the bookings that the model predicted as cancellations, 72.90% were actually cancellations. This shows the model's reliability when it predicts a cancellation
- **F1 Score:** The F1 score of 0.67641 provides a balanced measure of precision and recall

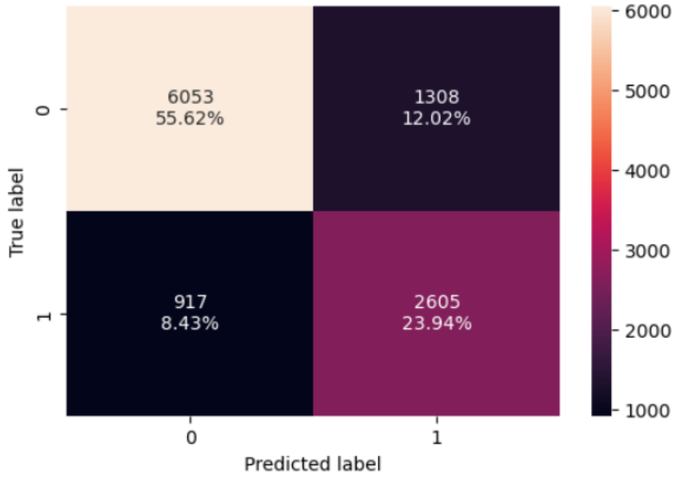
We will now try to improve the performance of the model

- Model Performance Improvement (ROC Curve and ROC-AUC) – Test Set



Changing the model threshold using AUC-ROC Curve

Confusion Matrix using optimal_threshold_auc_roc=0.37

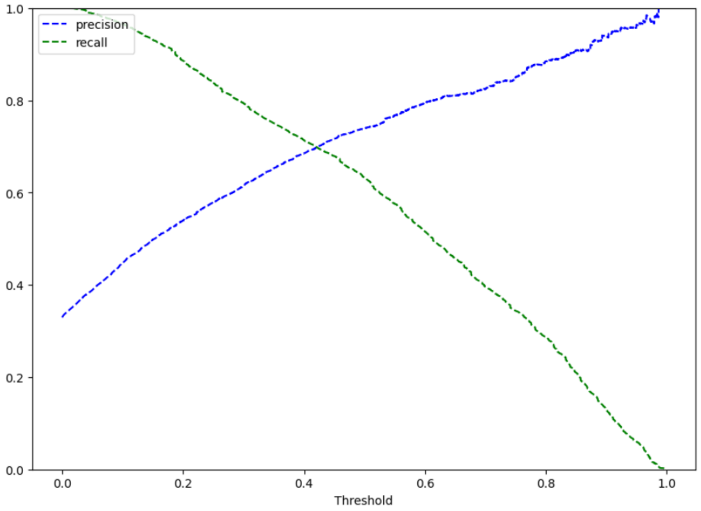


Test performance:

	Accuracy	Recall	Precision	F1
0	0.79555	0.73964	0.66573	0.70074

Logistic Regression model is giving a good performance on training set

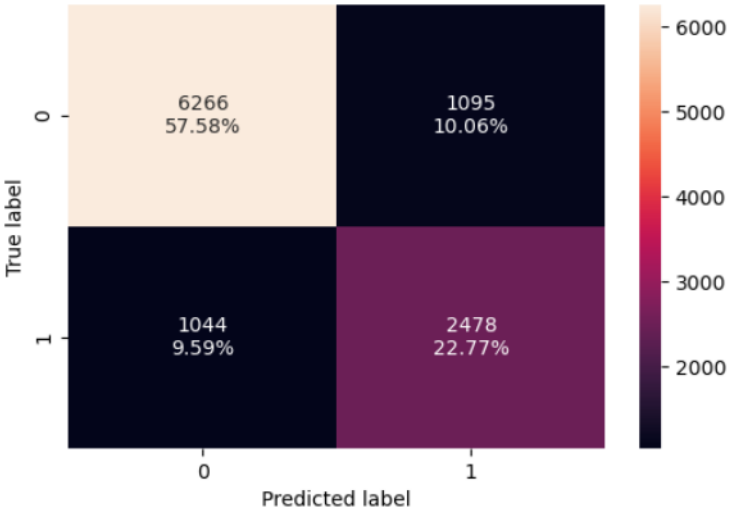
Model Performance Improvement (Precision-Recall Curve) - Test Set



Precision-Recall curve given a threshold =0.42

Changing the model threshold using Precision-Recall curve

Confusion Matrix using optimal_threshold_curve=0.42



Test performance:

	Accuracy	Recall	Precision	F1
0	0.80345	0.70358	0.69353	0.69852

Model Performance Summary – Logistic Regression

Training performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Test performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

- **Threshold Impact:** As the threshold decreases, recall increases, and precision decreases. This is a typical trade-off. Lowering the threshold makes the model more sensitive to identifying booking cancellations but also increases the number of false positives.
- **F1 Score:** The F1 score, which balances precision and recall, varies across the thresholds. On the test set, the 0.37 threshold achieves the highest F1 score (0.70074), suggesting it provides the best balance between precision and recall.
- **Overall Performance:** The model achieves a reasonable accuracy of approximately 80% on both the training and test sets. However, there's room for improvement, particularly in recall, even at the lower thresholds.

Model Building - Decision Tree

- ## Model Building Steps of Decision Tree

   i. **Data Preparation:** Identified the features (independent variables) that will be included in your model

   ii. **Data Preprocessing:** Converted categorical features into a numerical format using one-hot encoding

   iii. **Split the data:** Split the data into train and test (70:30) to be able to evaluate the model that was build on the train data with random_state = 1

Below is the shape of train and test data after split:

```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set:
booking_status
0    0.67238
1    0.32762
Name: proportion, dtype: float64
Percentage of classes in test set:
booking_status
0    0.67233
1    0.32767
Name: proportion, dtype: float64
```

iv. **Tree Growing:** The algorithm recursively repeats the following steps:

- Start at the root node with the entire training dataset.
- For each feature, evaluate the splitting criterion to find the best split.
- Split the node into child nodes based on the best split.
- Continue splitting the child nodes until a stopping criterion is met

v. **Stopping Criteria:** Determine when to stop splitting a node by

- Stop splitting if a node has fewer than a specified number of samples
- Limit the maximum depth of the tree
- Stop splitting if a node is pure
- Stop splitting if the best split does not significantly improve the purity of the node.

vi. **Tree Pruning:** Used to prevent overfitting, which occurs when the tree is too complex and learns the noise in the training data. This is done by Pre-Pruning and Post-Pruning

vii. **Model Evaluation:** Evaluate the performance of the decision tree model on the test set using appropriate metrics, such as:

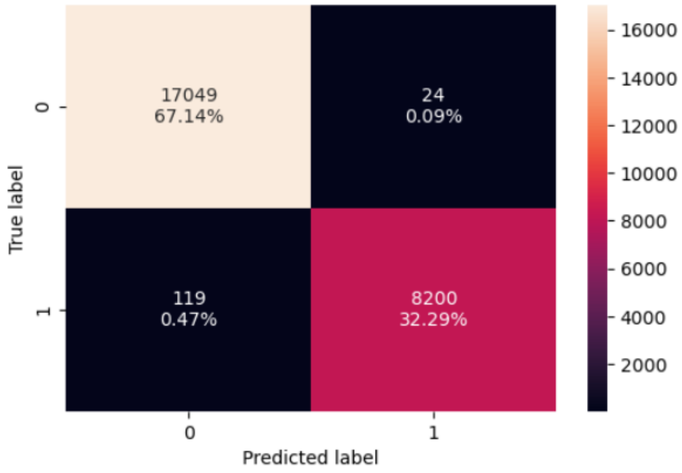
- Accuracy
- Precision
- Recall
- F1-score

Model Performance Evaluation and Improvement - Decision Tree

- ## Model Performance Check – Decision Tree

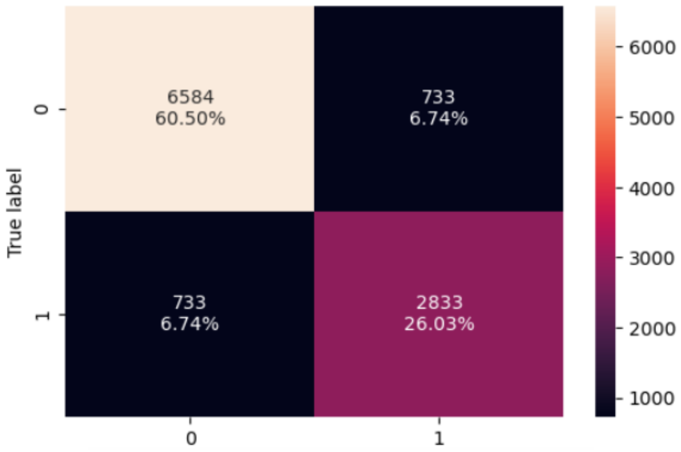
```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=1)
```

Model Performance on the Training Set



	Accuracy	Recall	Precision	F1
0	0.99437	0.98570	0.99708	0.99136

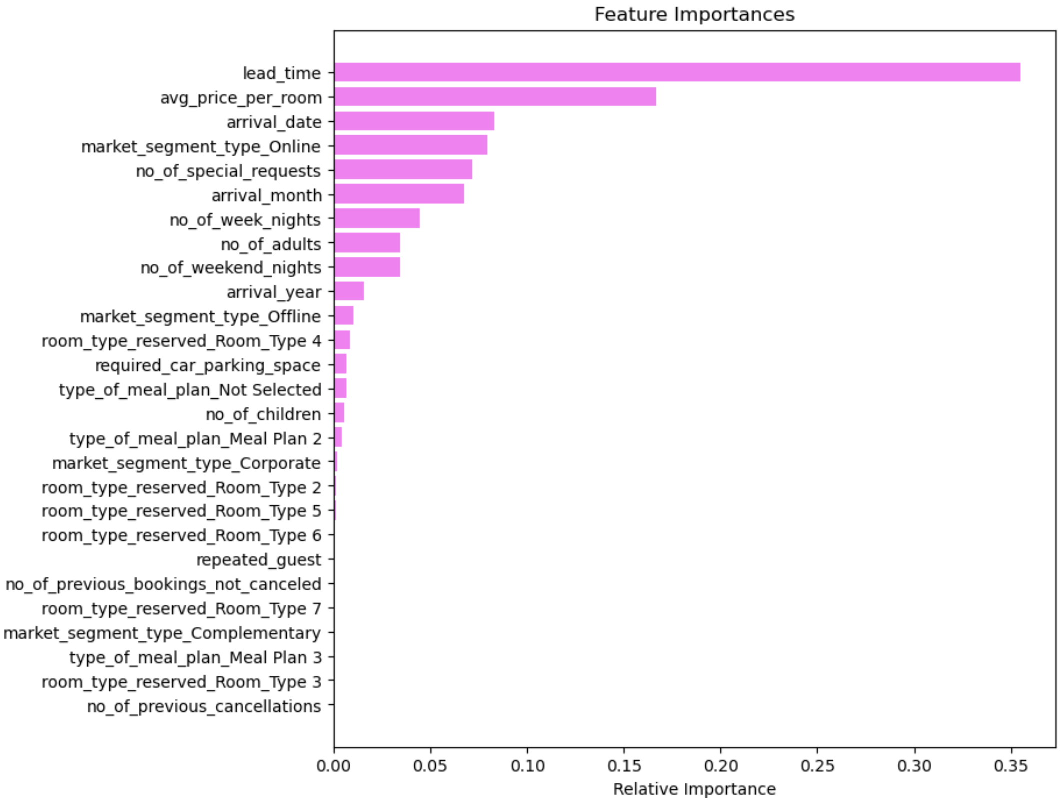
Model Performance on the Test Set



	Accuracy	Recall	Precision	F1
0	0.86529	0.79445	0.79445	0.79445

The results indicate a classic case of overfitting. The model has learned the training data extremely well, including its noise, but it fails to generalize to new, unseen data

- **Check the important features for the Decision Tree before Pruning**

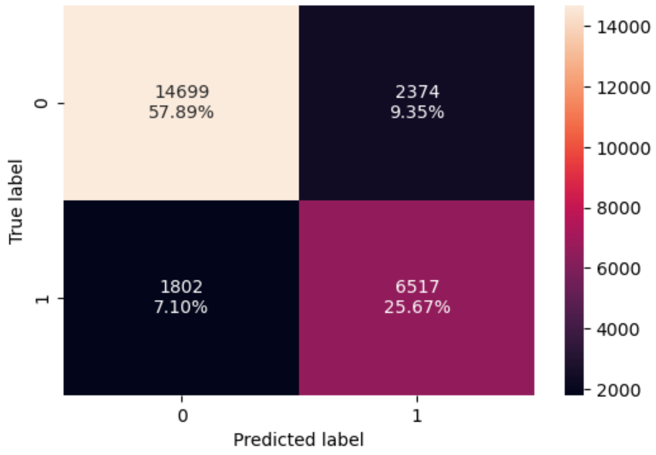


In the pre tuned decision tree, lead time and average price per room are the most important features followed by arrival date.

Pre-Pruning

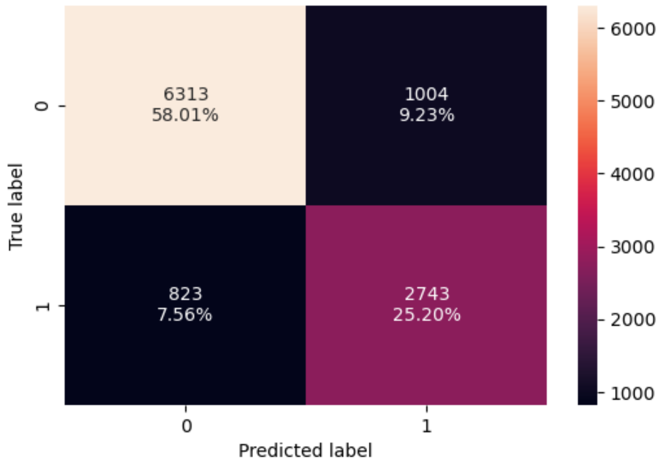
```
DecisionTreeClassifier
DecisionTreeClassifier(class_weight='balanced', max_depth=6, max_leaf_nodes=50,
                       min_samples_split=70, random_state=1)
```

Model Performance on the Training Set



	Accuracy	Recall	Precision	F1
0	0.83554	0.78339	0.73299	0.75735

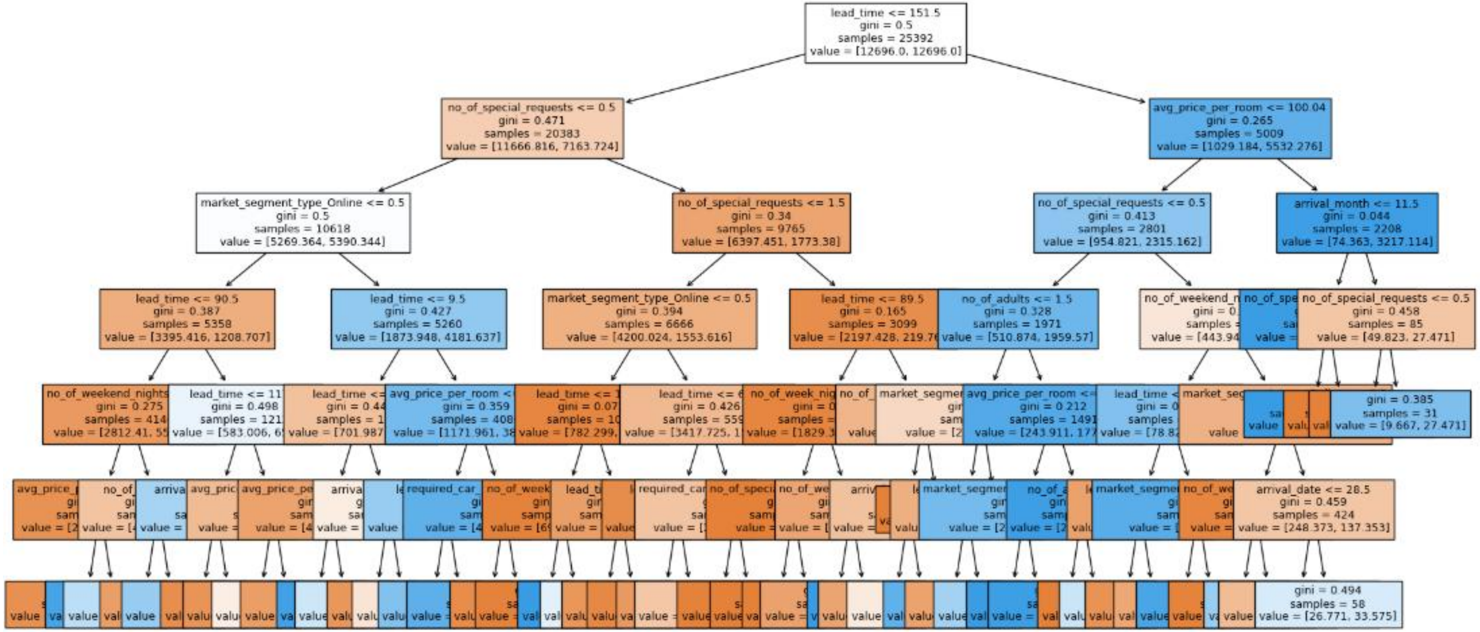
Model Performance on the Test Set



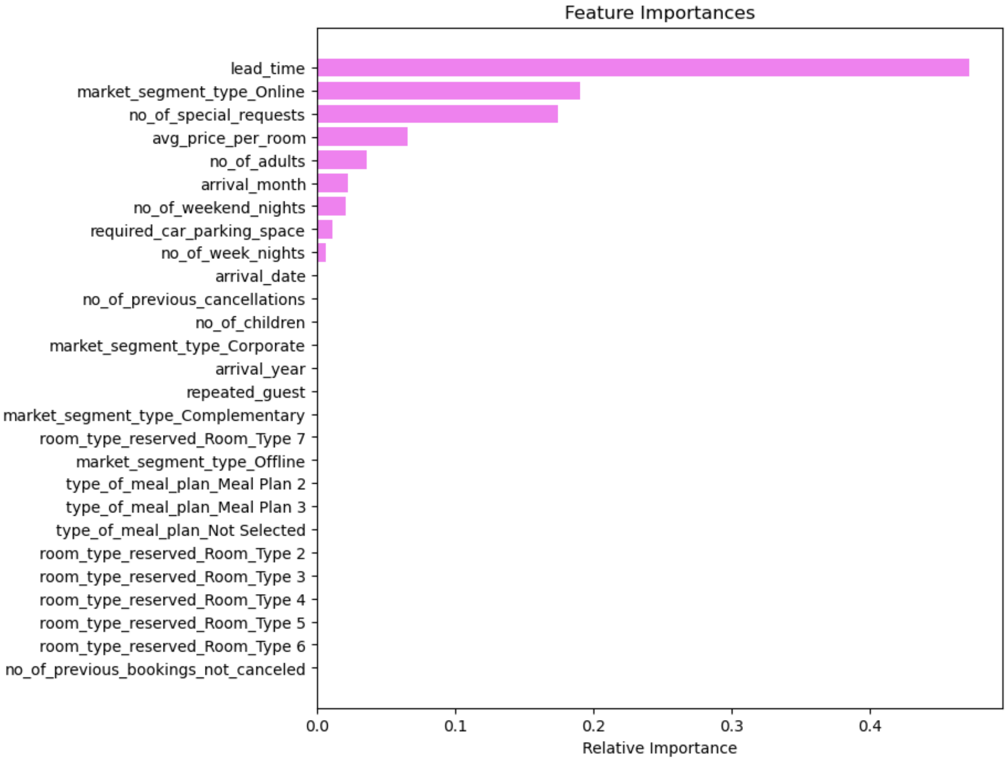
	Accuracy	Recall	Precision	F1
0	0.83212	0.76921	0.73205	0.75017

The model demonstrates good generalization. It is performing almost as well on unseen data as it did on the data it was trained on, which is a positive sign

Visualizing the Decision Tree – Pre Pruning



- **Check the important features for the Decision Tree before Post Pruning**



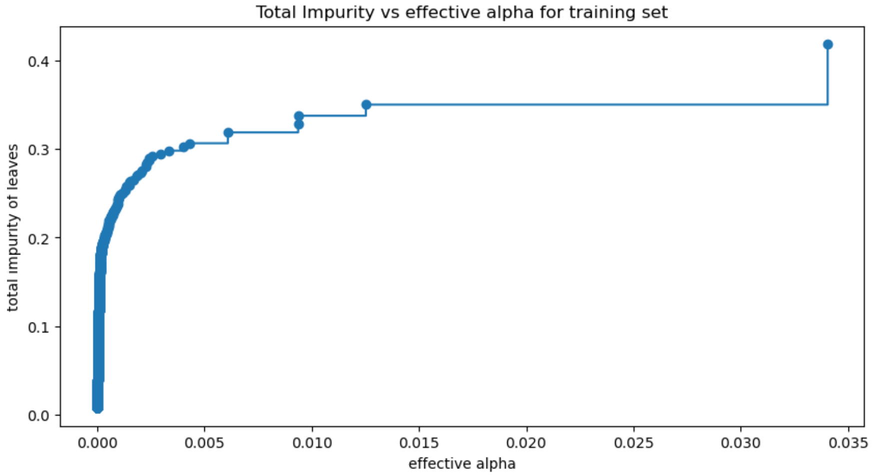
In the post tuned decision tree, lead time and market segment type online are the most important features.

Post Pruning - Cost Complexity Pruning

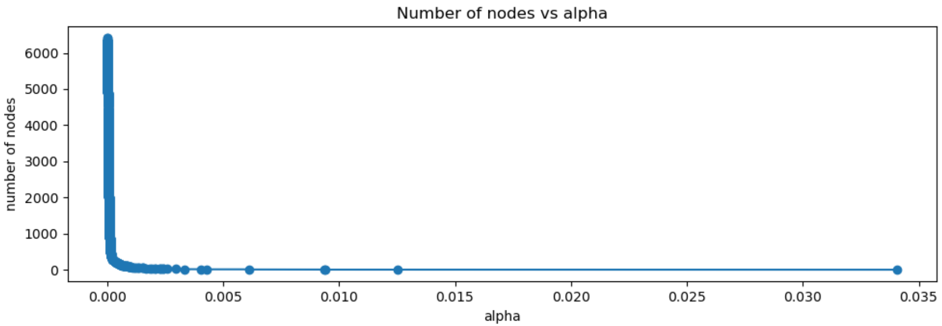
Training trees with different ccp_alpha values

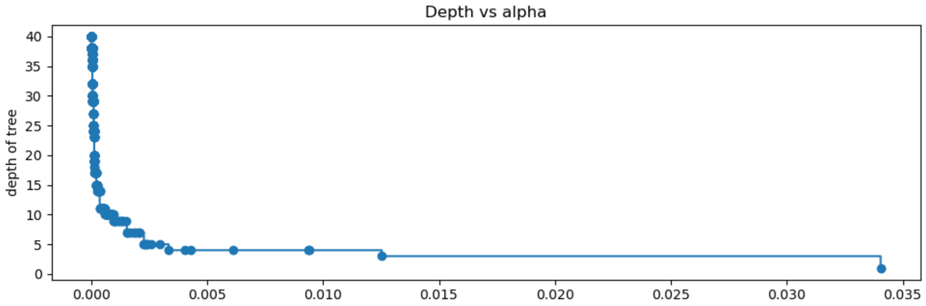
	ccp_alphas	impurities
0	0.00000	0.00833
1	-0.00000	0.00833
2	0.00000	0.00833
3	0.00000	0.00833
4	0.00000	0.00833
...	...	...
1648	0.00938	0.32791
1649	0.00941	0.33732
1650	0.01253	0.34985
1651	0.03405	0.41794
1652	0.08206	0.50000

1653 rows × 2 columns



- As effective alpha increases, the total impurity of the leaves generally increases as well
- The steepness of the graph indicates that during pruning, branches of the tree are collapsed, and the total impurity remains the same until the next branch is pruned
- The initial part of the graph shows a rapid increase in total impurity of leaves as effective alpha increases, then the increase slows down. This suggests that initially, pruning has a significant impact on impurity, but later on, the impact is less severe

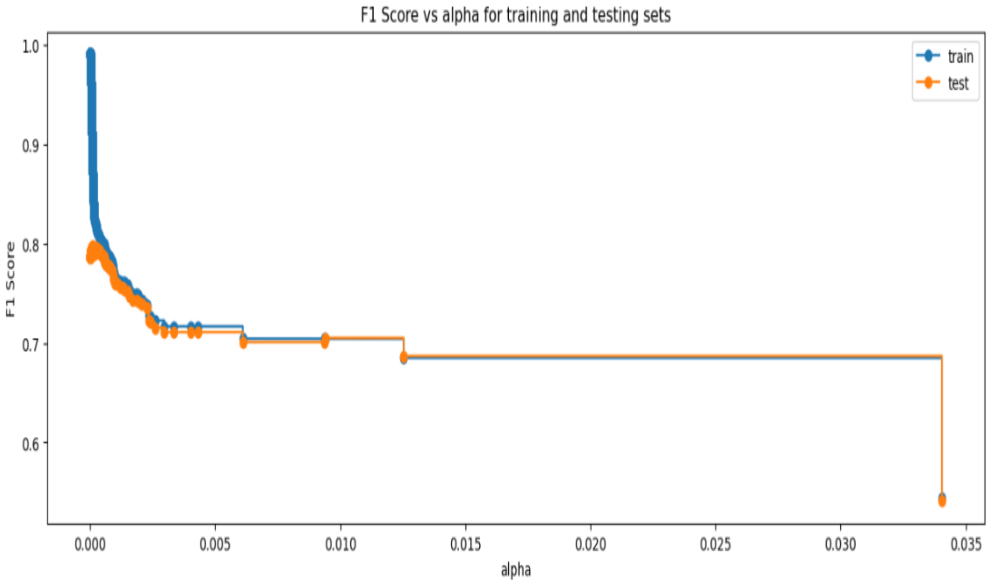




The graphs show that as you prune more (increase alpha):

- The bush gets smaller with fewer branches and leaves (number of nodes decreases)
- The bush gets shorter (depth decreases).

F1 Score vs Alpha for Training and Testing Sets

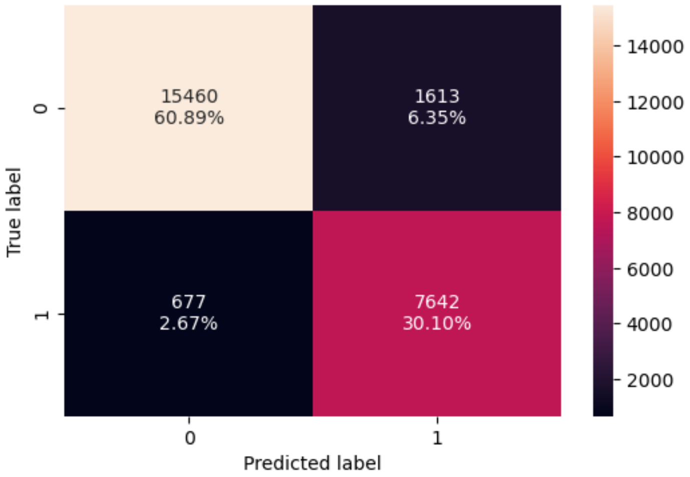


The graph shows how the F1 score changes as the value of alpha increases for both the training and the testing datasets.

- **Training Performance:** The F1 score for the training set is very high when alpha is close to zero indicating that the model fits the training data very well when the tree is complex
- **Test Performance:** The F1 score for the testing set initially increases as alpha increases, reaching a peak at a certain alpha value, and then decreases as alpha continues to increase

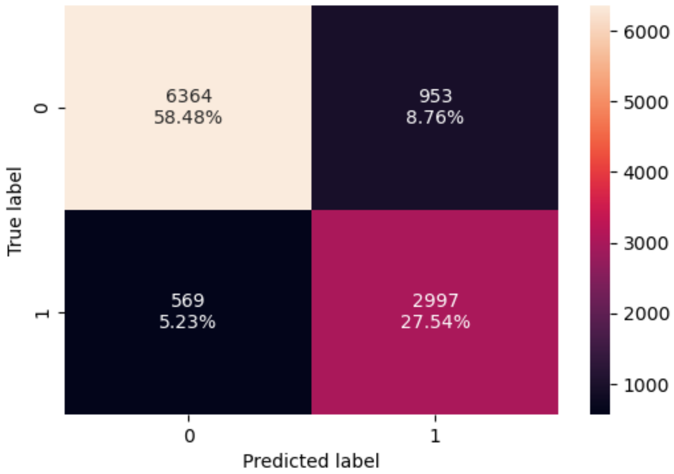
Post-Pruning - Cost Complexity Pruning

Model Performance on the Training Set



	Accuracy	Recall	Precision	F1
0	0.90981	0.91862	0.82572	0.86969

Model Performance on the Test Set

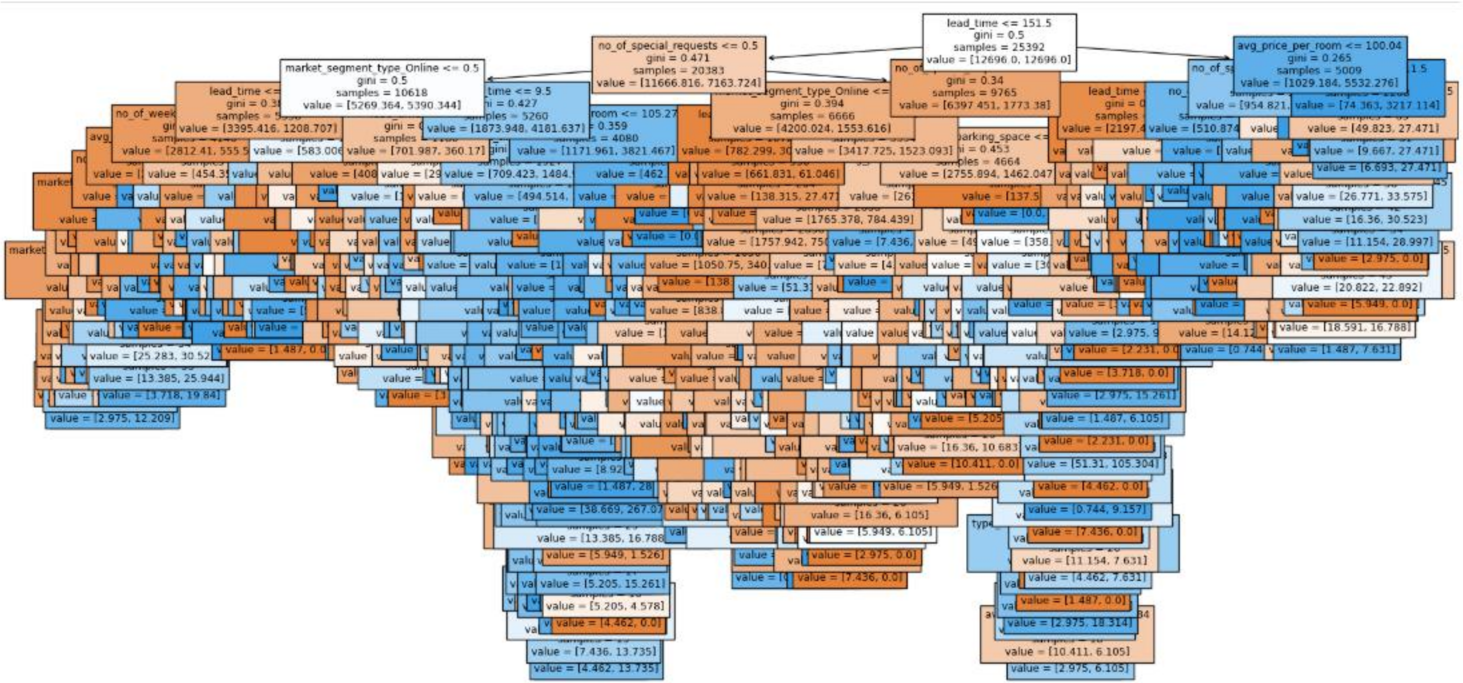


	Accuracy	Recall	Precision	F1
0	0.86015	0.84044	0.75873	0.79750

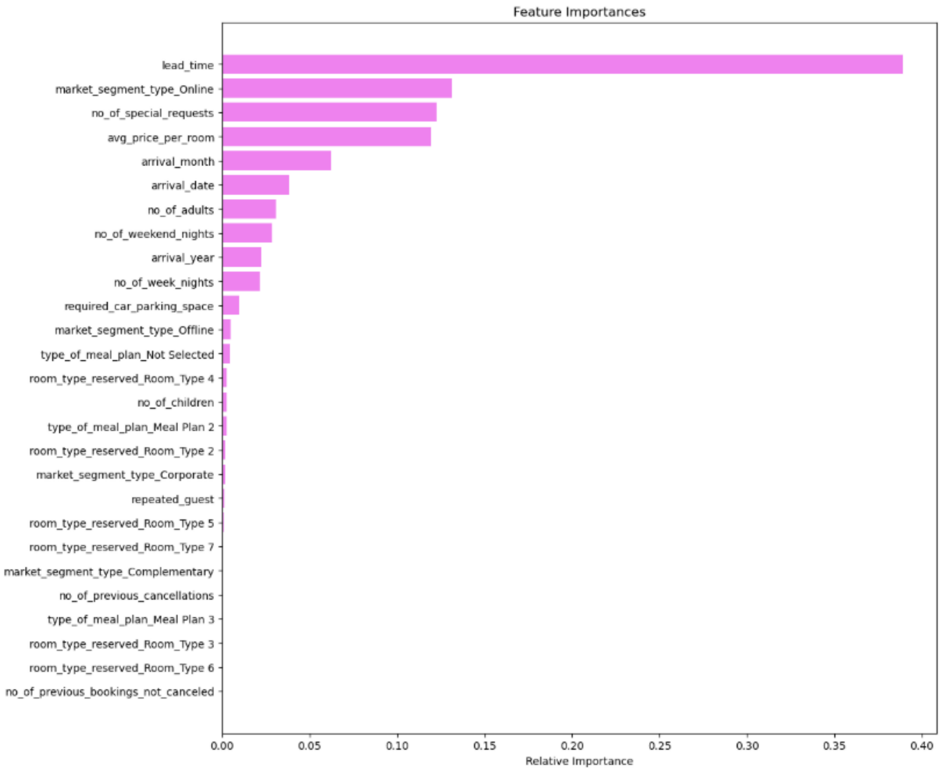
If we assume the first scenario (which is more likely), then:

- The model performs well on the training data, with accuracy above 90% and good balance between precision and recall.
- The performance drops slightly on the test data, which is expected. However, the drop isn't very large, suggesting that the model generalizes reasonably well to unseen data.

Visualizing the Decision Tree – Post Pruning



- **Check the important features for the Decision Tree after Post Pruning**



In the post tuned decision tree, lead time and market segment type online are the most important features.

Model Performance Summary – Decision Tree

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99437	0.83554	0.90981
Recall	0.98570	0.78339	0.91862
Precision	0.99708	0.73299	0.82572
F1	0.99136	0.75735	0.86969

Test performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.86529	0.83212	0.86015
Recall	0.79445	0.76921	0.84044
Precision	0.79445	0.73205	0.75873
F1	0.79445	0.75017	0.79750

- The standard Decision Tree (Decision Tree sklearn) overfits the training data, leading to poor generalization.
- Pre-pruning and post-pruning are effective techniques to prevent overfitting
- Post-pruning appears to be the most effective in this scenario, as it provides the best trade-off between fitting the training data and generalizing to unseen data

Final Model Summary

Model	Accuracy	Recall	Precision	F1
Logistic Regression (Default Threshold)	0.80465	0.63089	0.72900	0.67641
Logistic Regression (0.37 Threshold)	0.79555	0.73964	0.66573	0.70074
Logistic Regression (0.42 Threshold)	0.80345	0.70358	0.69353	0.69852
Decision Tree (sklearn)	0.86529	0.79445	0.79445	0.79445
Decision Tree (Pre-Pruning)	0.83212	0.76921	0.73205	0.75017
Decision Tree (Post-Pruning)	0.86015	0.84044	0.75873	0.79750

Logistic Regression:

- Demonstrates a trade-off between precision and recall depending on the probability threshold
- The 0.37 threshold provides a better balance than the default, achieving a higher F1 score

Decision Trees:

- The unpruned Decision Tree (sklearn) overfits the training data, resulting in the highest training performance but lower test performance
- Pruning techniques (pre-pruning and post-pruning) mitigate overfitting and improve generalization
- Post-pruning slightly outperforms pre-pruning in this case

Best Model

Considering the F1 score, which balances precision and recall, the Decision Tree with Post-Pruning and the Decision Tree (sklearn) achieve the highest F1 scores (0.79750 and 0.79445, respectively) on the test set. However, the post-pruned Decision Tree is preferred as it is less prone to overfitting.

References

Great Learning. (n.d.) *Supervised Learning - Classification*. **Great Learning.** https://olympus.mygreatlearning.com/courses/124966/modules/items/6397578

Happy Learning !