# INN Hotels Project

## Supervised Learning Classification

Julie Kistler

Date: December 15, 2023

# Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

# Executive Summary

The objective was to leverage existing booking data to develop a predictive model capable of forecasting cancellations and enabling the formulation of profit-driven policies. Two robust models, a logistic regression model and a decision tree model, were successfully constructed.

The models identified critical features influencing higher cancellation rates, lead time, market segment type, average price per room, and the number of special requests. These insights form the foundation for targeted strategies to minimize cancellations and enhance operational efficiency.

By implementing these models, it is anticipated to reduce the cancellation rates, resulting in decreased commissions to channel partners and a corresponding reduction in refund payouts. This, in turn, is expected to contribute to increased revenue and heightened operational efficiency.

**Insights and Recommendations:**

- **Utilizing Dynamic Pricing Strategies:** During peak demand periods implement a pricing strategy that keeps room rates competitive. Lowering prices during high-demand months attracts more potential guests. Mitigating cancellations and optimizing occupancy rates.
- **Implement a Customer Loyalty Program**: Using a customer reward system will foster brand loyalty and increase repeat bookings. This will also help with customer retention and potential referrals among their friends with great word-of-mouth marketing.

# Executive Summary cont...

**Insights and Recommendations cont:**

- **Customized Room Packages:** Build a system in the booking system that allows guests to request and designate specific room requests as well as build add-on essentials to create a more personalized booking and add additional revenue streams with add-on packages. This should significantly reduce cancellations.
- **Create a Marketing Campaign that Encourages Stay and Share**. Guests that stay in the properties are encouraged to share their experience with others online using a specific hashtag campaign. Each share or hashtag they use will enter them into a contest for a free stay at a future date or a complimentary add-on package to their stay. This should build a great organic marketing campaign while building goodwill within the customer base.
- **Reduce Calendar Options:** Consider reducing the calendar availability to only reflect a maximum lead time of six months. This will reduce those bookings that may cancel due to plans changings, market conditions, and competitive offerings that may lead to an increase in cancellations due to long lead times.

# Business Problem Overview and Solution Approach

INN Hotels Group in Portugal has been impacted by a higher number of cancellations. They have reached out to our team to help them understand and tackle this trend with data-driven solutions. The increase in cancellations has had an impact on their revenue, resource allocation, pricing model, and distributions channel commission payouts. Our data team has been charged to identify and understand the factors that are causing an increase in cancellations. There are various factors that may contribute to the increase in cancellations. They may include customer behavior, customer loyalty, lead time, market trends, pricing, and other variables. Identifying the root causes are crucial for developing an accurate predictive model.

The solution approach:

- Identify key factors that may influence booking cancellations.
- Develop a predictive model by implementing the supervised learning classification methodology to include logistic regression and decision trees classification.
- Interpret the outputs of the classification models developed to forecast which booking are likely to be cancelled.
- Recommend data-driven policies to reduce booking cancellations/refunds to increase profitability.

# DATA OVERVIEW

| Data Dictionary | |
|---|---|
| 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* |
| 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 |


| * Type of Meal Plan | |
|---|---|
| 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) |

# DATA OVERVIEW CONT...

| Column | Dtype |
|---|---|
| Booking_ID | object |
| no_of_adults | int64 |
| no_of_children | int64 |
| no_of_weekend_nights | int64 |
| no_of_week_nights | int64 |
| type_of_meal_plan | object |
| required_car_parking_space | int64 |
| room_type_reserved | object |
| lead_time | int64 |
| arrival_year | int64 |
| arrival_month | int64 |
| arrival_date | int64 |
| market_segment_type | object |
| repeated_guest | int64 |
| no_of_previous_cancellations | int64 |
| no_of_previous_bookings_not_canceled | int64 |
| avg_price_per_room | float64 |
| no_of_special_requests | int64 |
| booking_status | object |

| Columns | Rows |
|---|---|
| 36275 | 19 |

- There are no missing values
- There are no duplicate values


- **5 object data types:** (Booking_ID, type_of_meal_plan, room_type_reserved, market_segment_type, booking_status)
- **13 integer data types:** (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, repeating_guest, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, no_of_special_requests)
- **1 float data type:** (avg_price_per_room)

# EDA Results_Univariate Analysis

>>>> Booking_ID column was dropped from the dataframe <<<<

- Lead Time is skewed right with several outliers – average lead time is ~ 85 Days
- Average Price Per Room Distribution appears normal with many outliers
  - Average Price Per Room is ~ $103
  - There are 545 with a average price per room to be equal to $0
  - Average price per room equal to or above $500 were treated (Max was $540)
- The average number of previous cancellations is ~ 023%
- 72% of booking had 2 Adults
- Approximately 93% of booking did not have Children (booking with 9 to 10 Children were treated)
- A majority of week night bookings include stays between 1 – 3 Days
- Approximately 28% of all bookings in one weekend night
- 25% of bookings in 2 weekend nights
- Approximately 46% of all bookings do not include a weekend night
- Approximately 97% of all bookings do not request a required parking spot
- Meal plan 1 appears to be the top choice for guests (freq of 27835 ~ 77%)
- Room type 1 appears to be the top choice for guests (freq of 28130 ~ 78%)
- The years analyzed are 2017 to 2018

*Link to Appendix statistical summary and supporting univariate analysis slides*

# EDA Results_Univariate Analysis

- Arrivals are strongest in October, September and August. (~38% of all bookings).  January and February have the lowest bookings.
- The online market segment appears to be the dominant booking choice (freq of 23214 ~  64%)
- Over half (~55%) of the bookings do not have any special requests.
- Approximately 31% of all bookings have one special request.
- A little over 2/3 of all booking are not cancelled.
- Approximately 1/3 of all booking are cancelled

*Link to Appendix statistical summary and supporting univariate analysis slides*

# EDA Results_Bivarate Analysis

- There are no strongly correlated variables with the exception of repeat guest and number of previous booking not cancelled.
- Online bookings appears to have the highest Average Price per Room.  Most bookings are made online and pricing is based upon demand.  As demand increased prices rise.
- Corporate booking appear to have the lowest Average Price per Room – most falling under the $100 mark.
- The largest portion of cancellations appear to be from online bookings with little to none at the complimentary segment.
- Booking with no special requests appears to have the largest amount of cancellations. The more special requests the less chance for cancellations.
- There appears to be a small spike on the Average Price per Room as special request rise.
    - The most significant spike in Average Price Per Room is 2 Special Requests
- Average Price per Room may have an impact on cancellation ratios.  It appears a higher Average Price per Room may lead to a higher cancellation rate.
- It appears the longer the lead time the increased chance that the booking will be cancelled.  It appears bookings with lead times over 100 days have a higher chance of cancellations.
- Number of family members does not seem to affect the number of cancellations.
- It appears the booking with more days may result in more cancellations
- The summer months appear to have the highest level of cancellation where the winter months tend to have less.
- The highest Average Price per Room appears to be in the summer months.
- The lowest Price per Room is in the winter months.

*Link to supporting bibvarate analysis slides*

# Data Preprocessing

- There were no duplicates found. - no treatment necessary
- There were no missing values founds – no treatment necessary
- Outlier check completed – all data appears to be in the dataset no treatment was necessary
- Featured engineering: We  previously treated some data while performing the EDA:
  - Booking_ID was dropped from the dataframe
  - The the upper level outliers in the Average Price Per Room that had values be equal to or greater than $500 were treated and replaced with the upper whisker price of $179.55
  - There were some outliers in the number of children – bookings with 9 and 10 children were replaced with 3
- Data preparation for modeling – Encoding booking status: Non Cancelled = 0 and Cancelled = 1
  - We want to predict which booking will be cancelled before we build the model. We split the data using a 70/30 split into train and test to evaluate the model and build the train data.

| Booking Status | Train | Test |
|---|---|---|
| 0 | 67% | 67.6% |
| 1 | 32.9% | 32.3% |

*Link to supporting data processiong slides*

# Model Performance Summary

- The Logistic Regression showed good performance and no signs of overfitting in the final models.
- For the Logistic Regression model it was determined that the model with the .37 threshold is the best model to use with the best F1 Score.

Training performance comparison:

| | Logistic Regression-default Threshold | 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 | 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 model was treated for multicollinearity – removing high P-values
- The model converted coefficients to odds

*Link to supporting logistic regression slides*

# Model Performance Summary

- The Decision Tree models both showed good performance and no signs of overfitting in the final models. For the Decision Tree model it was determined that the post-pruning model is the best model to use with the best F1 Scores.

Training performance comparison:

| | Decision Tree sklearn | Decision Tree (Pre-Pruning) | Decision Tree (Post-Pruning) |
|---|---|---|---|
| **Accuracy** | 0.99421 | 0.83097 | 0.89954 |
| **Recall** | 0.98661 | 0.78608 | 0.90303 |
| **Precision** | 0.99578 | 0.72425 | 0.81274 |
| **F1** | 0.99117 | 0.75390 | 0.85551 |

Test performance comparison:

| | Decision Tree sklearn | Decision Tree (Pre-Pruning) | Decision Tree (Post-Pruning) |
|---|---|---|---|
| **Accuracy** | 0.87118 | 0.83497 | 0.89954 |
| **Recall** | 0.81175 | 0.78336 | 0.90303 |
| **Precision** | 0.79461 | 0.72758 | 0.81274 |
| **F1** | 0.80309 | 0.75444 | 0.85551 |

- Some of the most important features the model used for predicting is lead time, market segment type, average price per room, and number of special requests.

*Link to supporting decision tree slides*

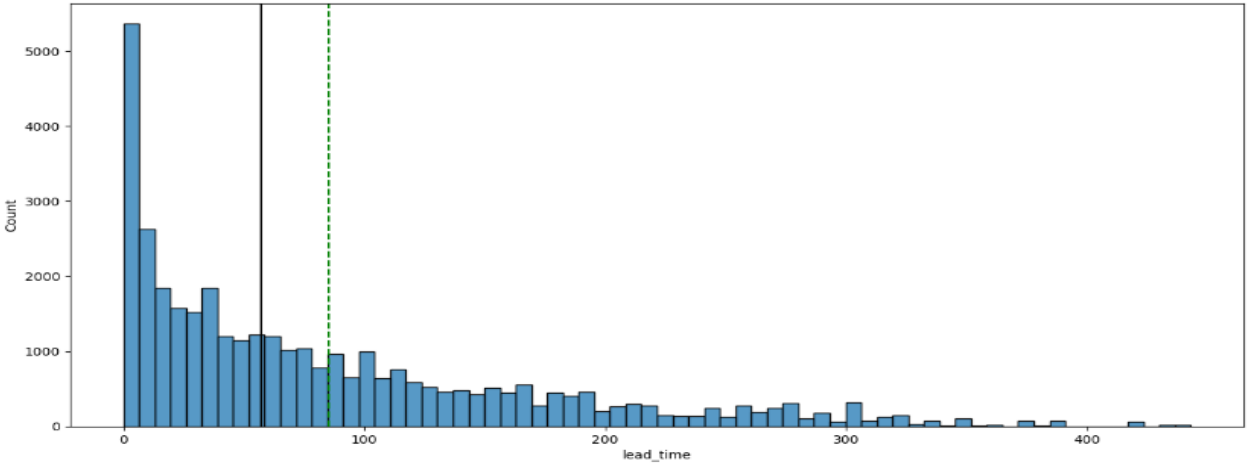

# APPENDIX

# Statistical Summary - EDA

| | count | unique | top | freq | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| no_of_adults | 36275.00000 | NaN | NaN | NaN | 1.84496 | 0.51871 | 0.00000 | 2.00000 | 2.00000 | 2.00000 | 4.00000 |
| no_of_children | 36275.00000 | NaN | NaN | NaN | 0.10528 | 0.40265 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 10.00000 |
| no_of_weekend_nights | 36275.00000 | NaN | NaN | NaN | 0.81072 | 0.87064 | 0.00000 | 0.00000 | 1.00000 | 2.00000 | 7.00000 |
| no_of_week_nights | 36275.00000 | NaN | NaN | NaN | 2.20430 | 1.41090 | 0.00000 | 1.00000 | 2.00000 | 3.00000 | 17.00000 |
| type_of_meal_plan | 36275 | 4 | Meal Plan 1 | 27835 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| required_car_parking_space | 36275.00000 | NaN | NaN | NaN | 0.03099 | 0.17328 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 1.00000 |
| room_type_reserved | 36275 | 7 | Room_Type 1 | 28130 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| lead_time | 36275.00000 | NaN | NaN | NaN | 85.23256 | 85.93082 | 0.00000 | 17.00000 | 57.00000 | 126.00000 | 443.00000 |
| arrival_year | 36275.00000 | NaN | NaN | NaN | 2017.82043 | 0.38384 | 2017.00000 | 2018.00000 | 2018.00000 | 2018.00000 | 2018.00000 |
| arrival_month | 36275.00000 | NaN | NaN | NaN | 7.42365 | 3.06989 | 1.00000 | 5.00000 | 8.00000 | 10.00000 | 12.00000 |
| arrival_date | 36275.00000 | NaN | NaN | NaN | 15.59700 | 8.74045 | 1.00000 | 8.00000 | 16.00000 | 23.00000 | 31.00000 |
| market_segment_type | 36275 | 5 | Online | 23214 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| repeated_guest | 36275.00000 | NaN | NaN | NaN | 0.02564 | 0.15805 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 1.00000 |
| no_of_previous_cancellations | 36275.00000 | NaN | NaN | NaN | 0.02335 | 0.36833 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 13.00000 |
| no_of_previous_bookings_not_canceled | 36275.00000 | NaN | NaN | NaN | 0.15341 | 1.75417 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 58.00000 |
| avg_price_per_room | 36275.00000 | NaN | NaN | NaN | 103.42354 | 35.08942 | 0.00000 | 80.30000 | 99.45000 | 120.00000 | 540.00000 |
| no_of_special_requests | 36275.00000 | NaN | NaN | NaN | 0.61966 | 0.78624 | 0.00000 | 0.00000 | 0.00000 | 1.00000 | 5.00000 |
| booking_status | 36275 | 2 | Not_Canceled | 24390 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

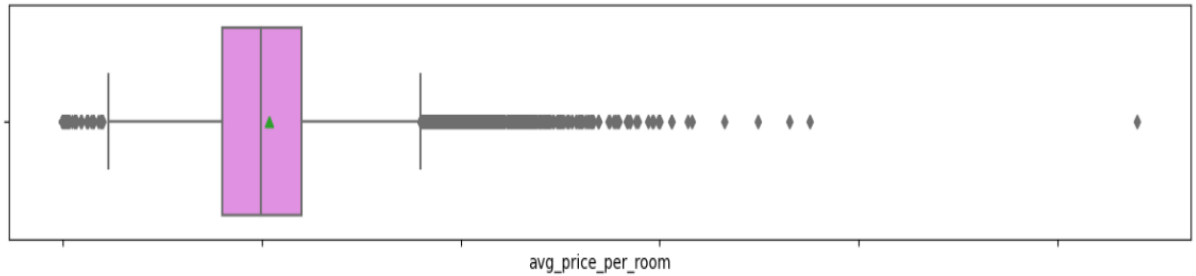
# Univariate Analysis

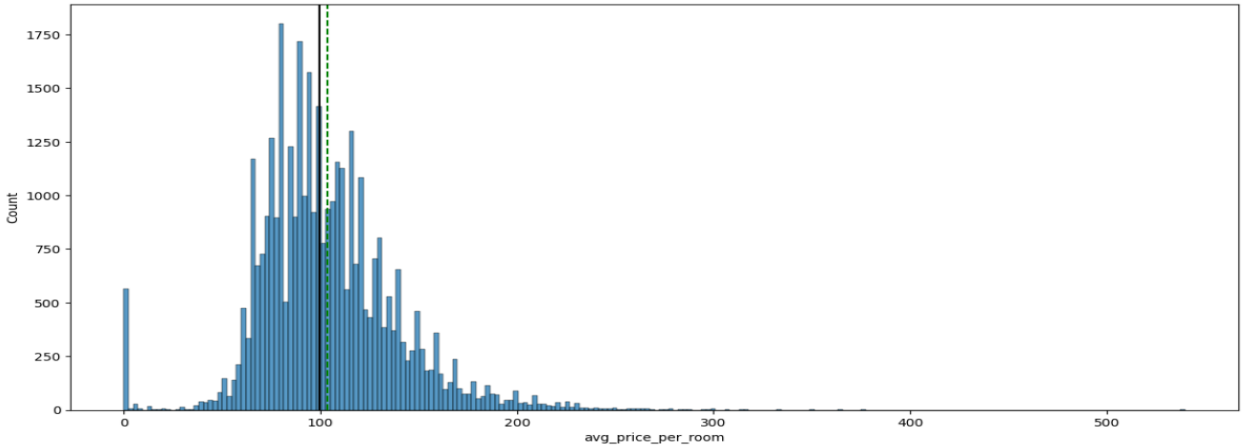
# Lead Time


Skewed right with several outliers.

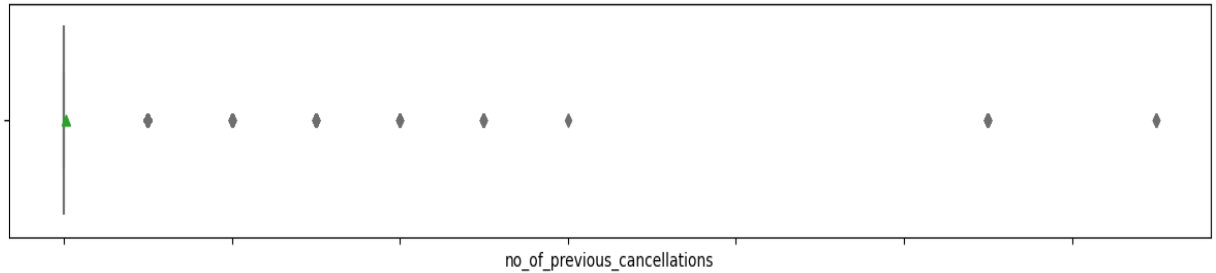
# Average Price Per Room

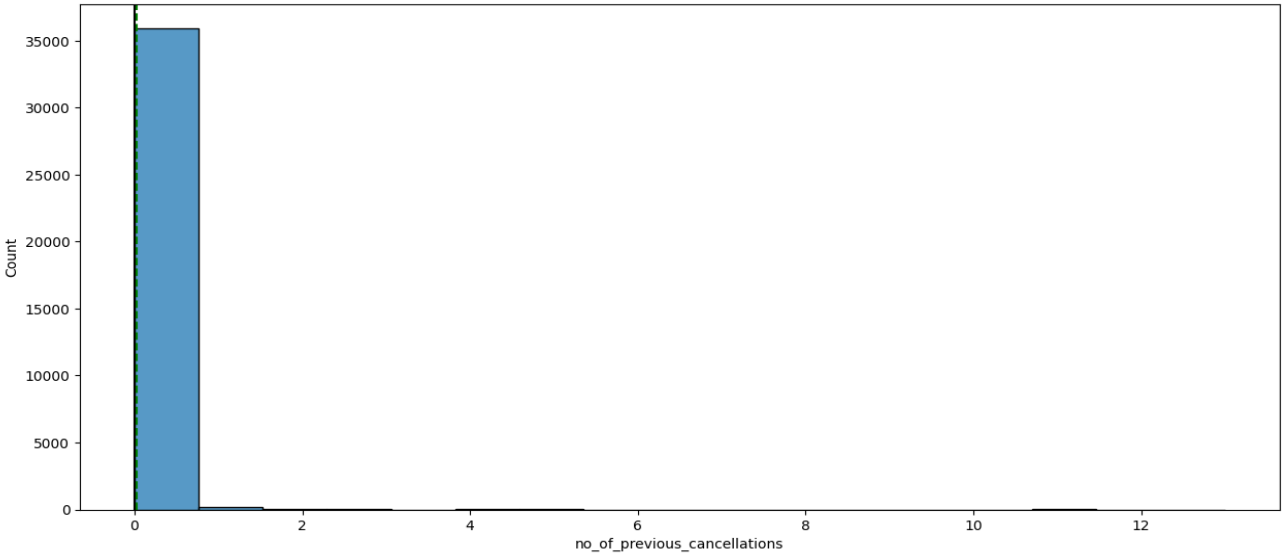



- Distribution appears relatively normal with many outliers.
- There are 545 bookings with an Average Price Per Room to be equal to 0:
  - *Complementary: 354
  - *Online: 191
- Treated the outliers with an Average Price Per Room equal to our greater than $500.00 with the upper whisker: $179.55.

# Number of Previous Booking Cancellations



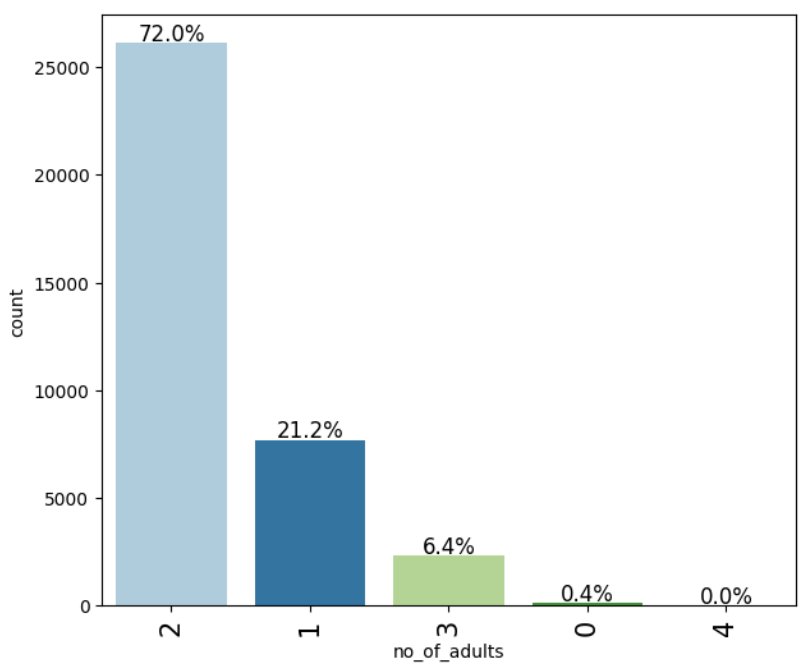



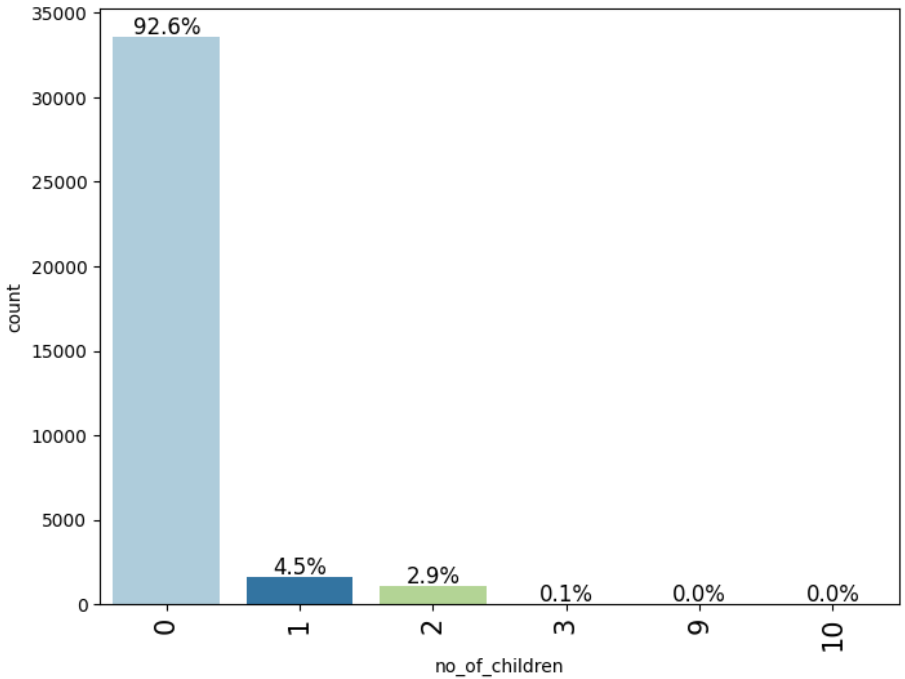

A majority of bookings do not show previous booking cancellations.

# Number Adults and Number of Children

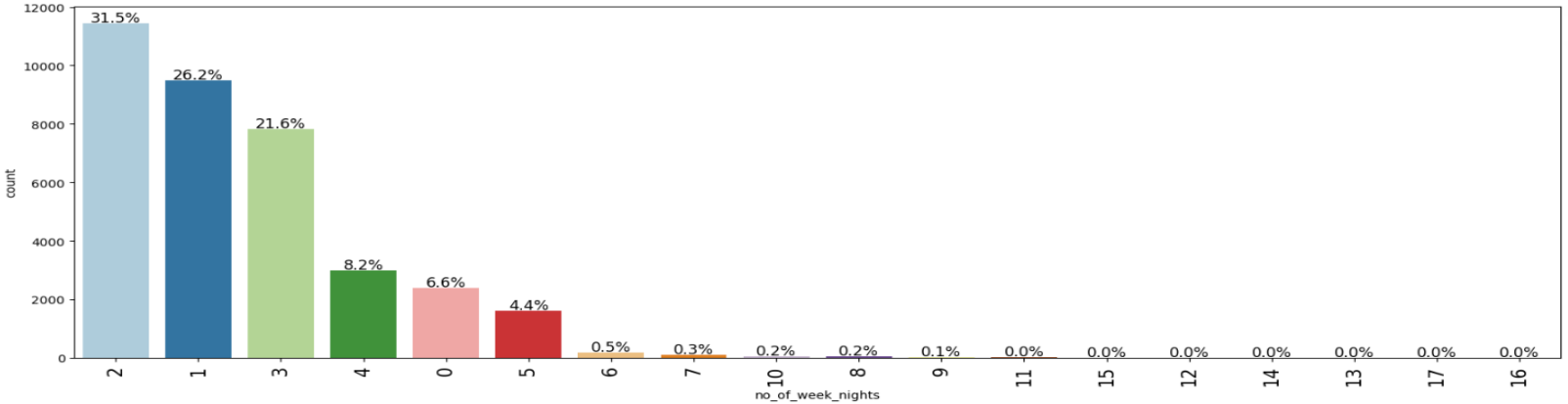
- 72% of the bookings had 2 adults in booking.
- 92.6% bookings had no children listed in the booking:

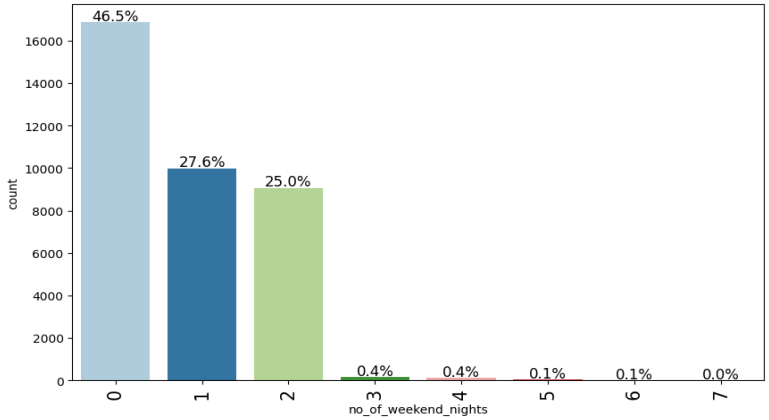
  *There were some outliers in the number of children – bookings with 9 and 10 listed were replaced with 3



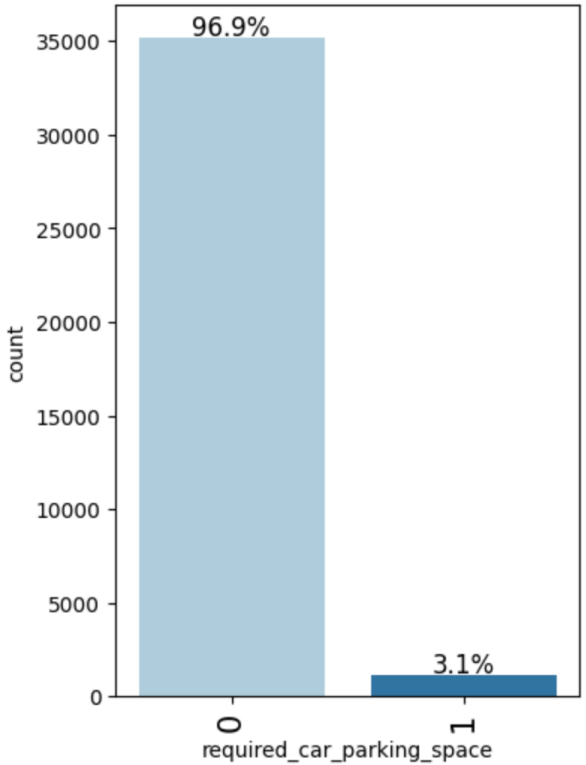
# Number of Week Nights vs. Number of Weekend Nights



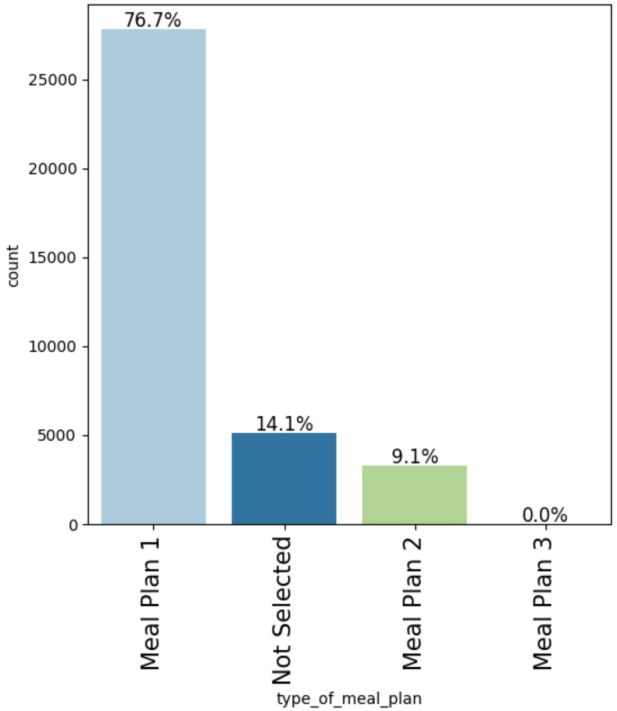

- It appears a majority of Week Night bookings include stays between 1 – 3 days.
- 27.6% of bookings include one weekend night.
- 25% of bookings include 2 weekend nights.
- 46.5% of bookings do not include a weekend night.
- There are some outliers.

# Required Parking Spot



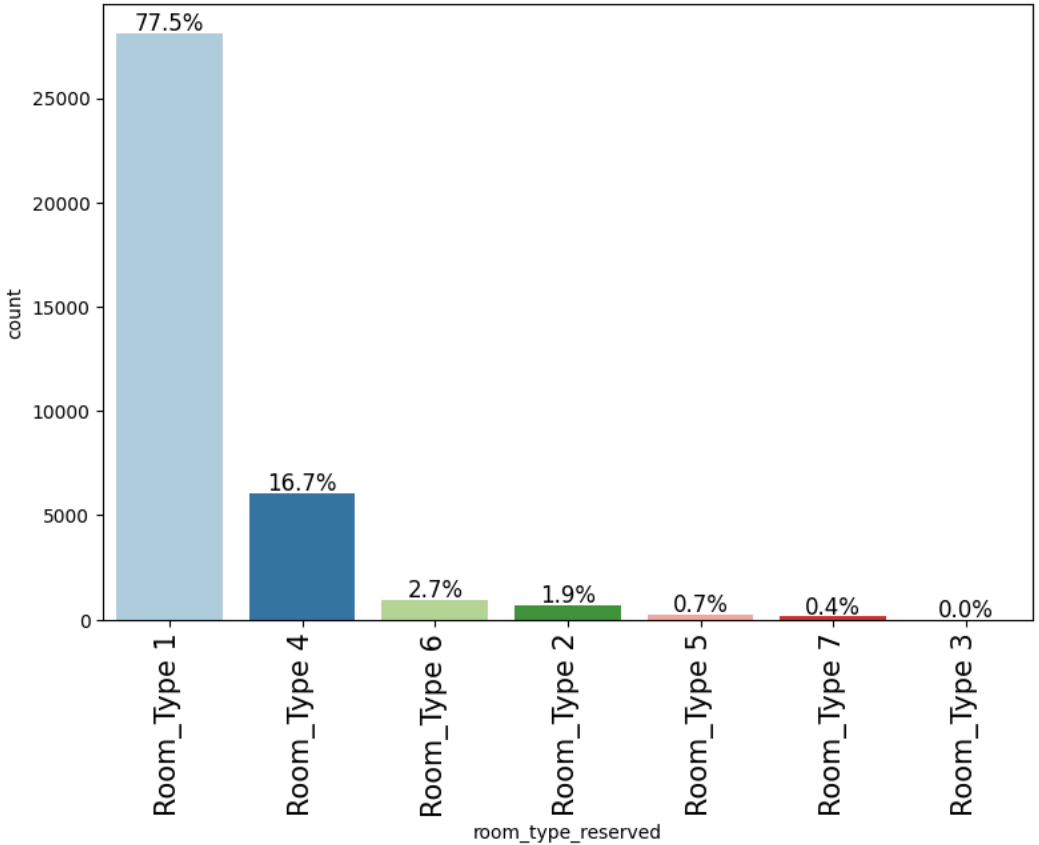

Most guests are not requesting a required parking spot.

# Type of Meal Plan



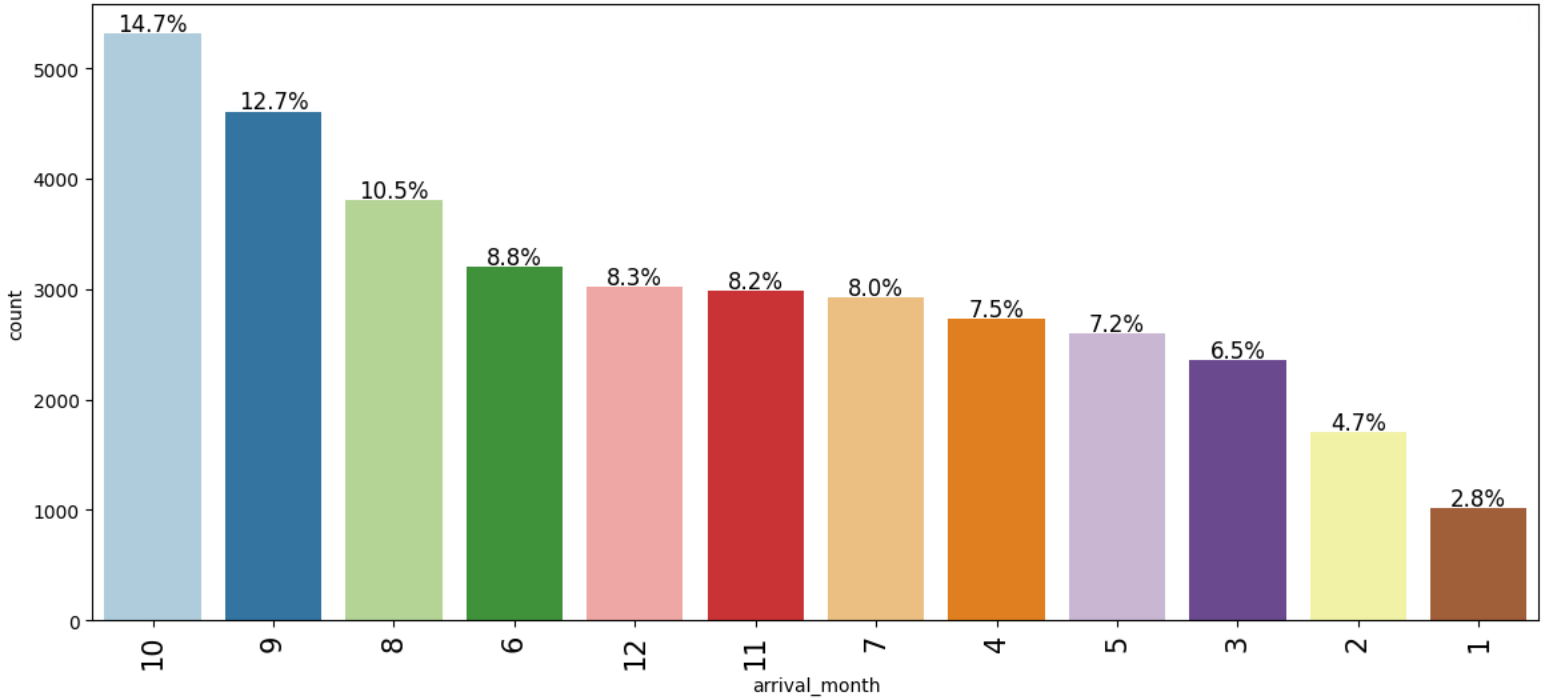

- Most guests are opting for meal plan 1.
- Meal plan 3 does not appear to be in demand.

# Room Type Reserved



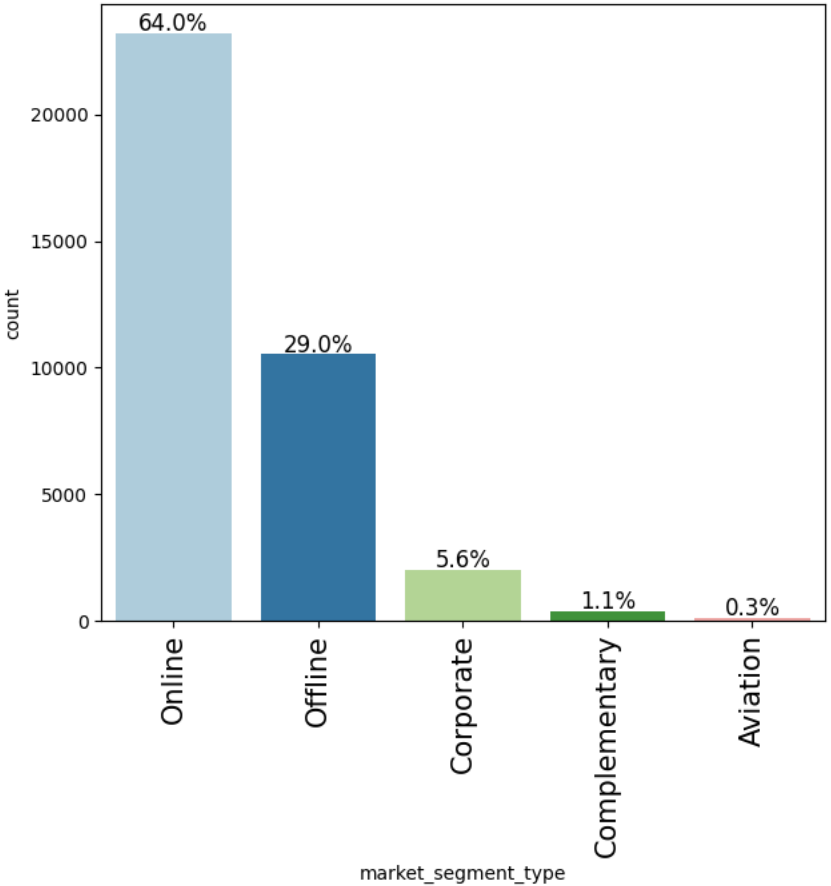

- Room Type 1 is the clear leader in rooms requested.
- Room Types 5, 7, 3 demonstrate little to no demand.

# Arrival Month



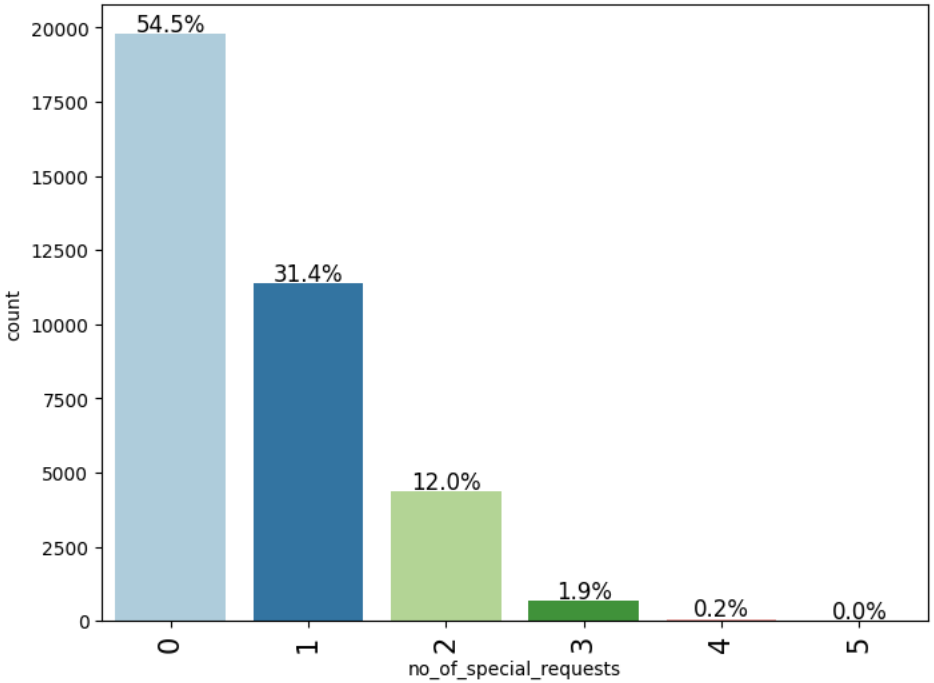

Arrivals are strongest in months October, September and August, respectively, with 37.9% of all bookings to occur during these months

# Market Segment Type



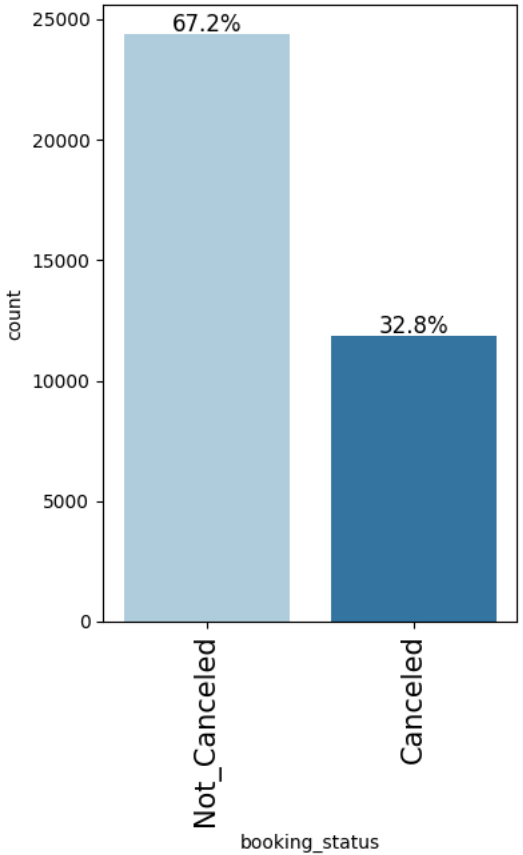

Most bookings are made online with 64% of total bookings.

# Number of Special Requests




- Over half (54.5%) of the bookings do not have any special requests.
- 31.4% of bookings have one special request.

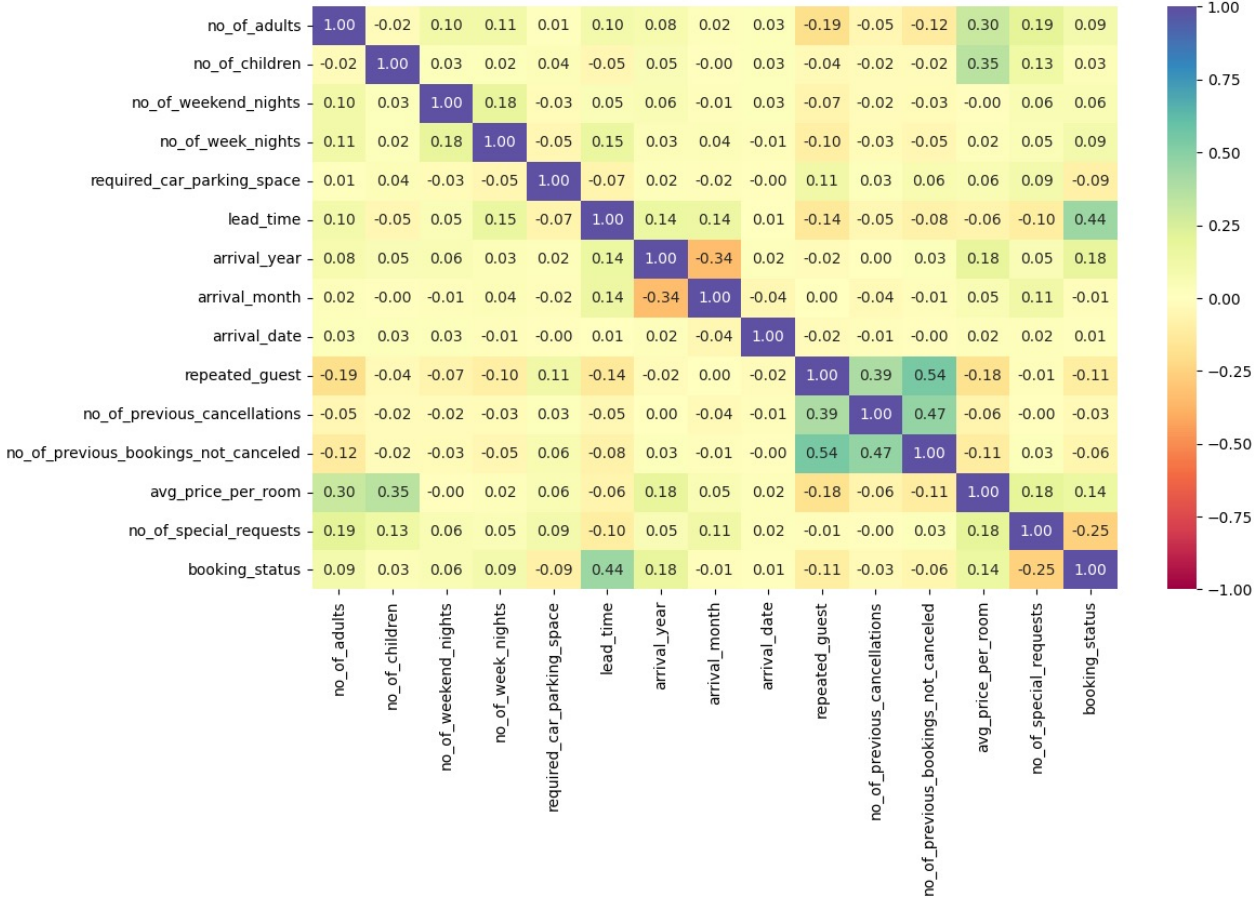
# Booking Status




- A little over 2/3 of the booking are not cancelled.
- Almost a 1/3 of bookings are cancelled.

# Bivariate Analysis

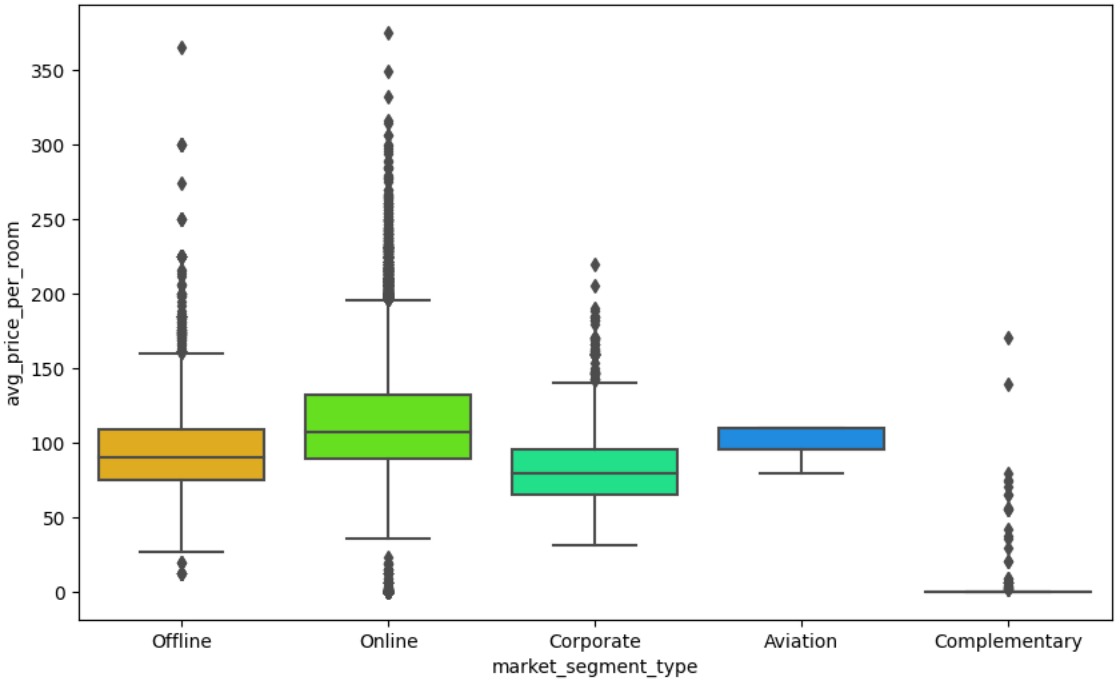
# Correlation




There are no strongly correlated variables with the exception of repeat guest and number of previous booking not cancelled.

Further analysis into lead time and booking status may be needed.

# Average Price per Room and Market Segment

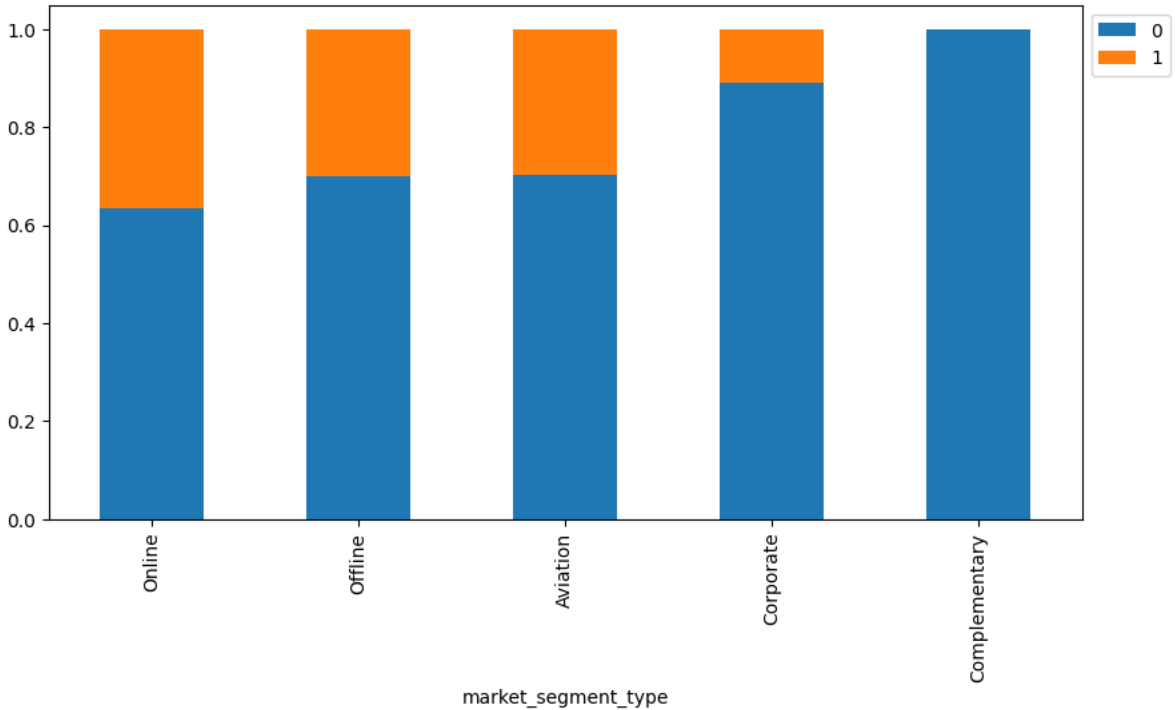



Online bookings appears to have the highest Average Price per Room. Most bookings are made online and pricing is based upon demand. As demand increased prices rise.

Corporate booking appear to have the lowest Average Price per Room – most falling under the $100 mark.

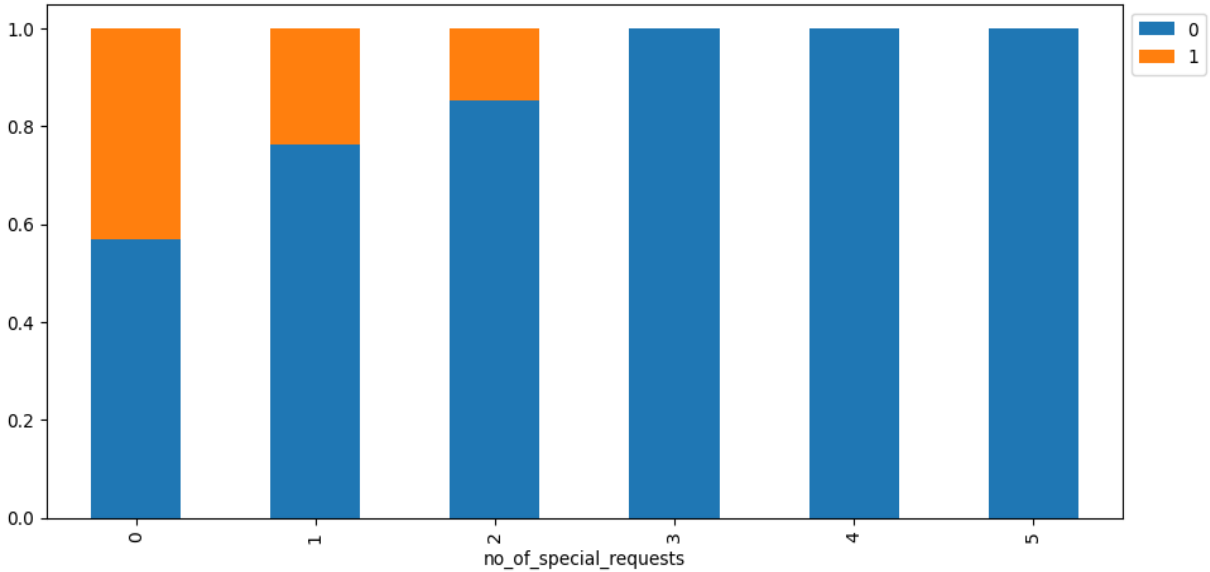
There are many outliers.

# Booking Status and Market Segment



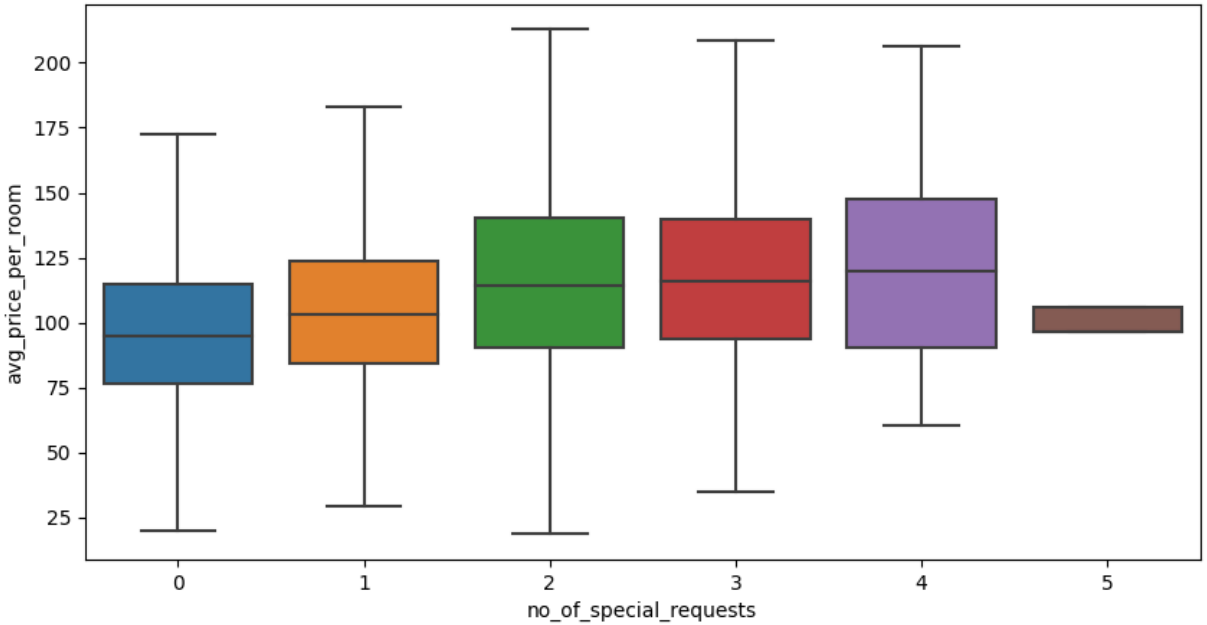

The largest portion of cancellations appear to be from online bookings with little to none at the complimentary segment.

# Number of Special Requests and Booking Status



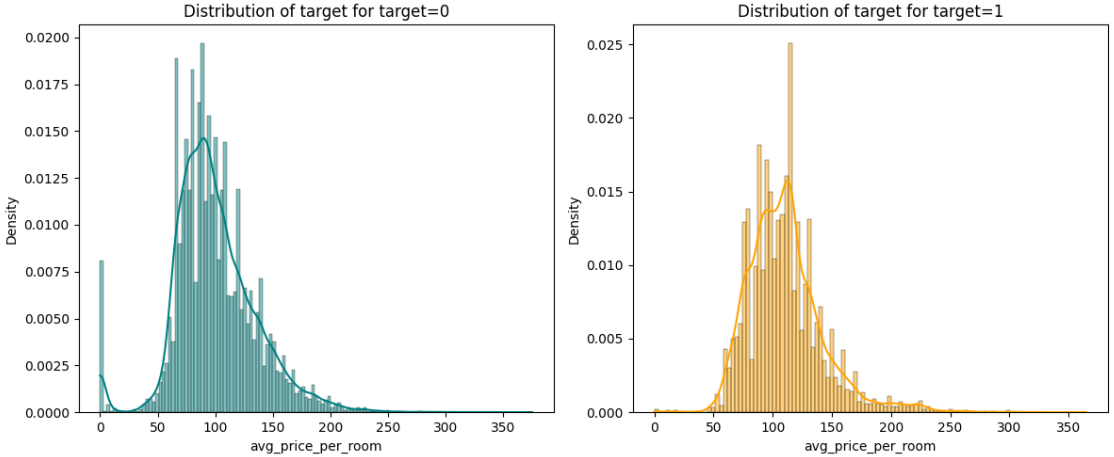

- Booking with no special requests appears to have the largest amount of cancellations.
- The more special requests the less chance for cancellations.

# Number of Special Requests and Average Price per Room



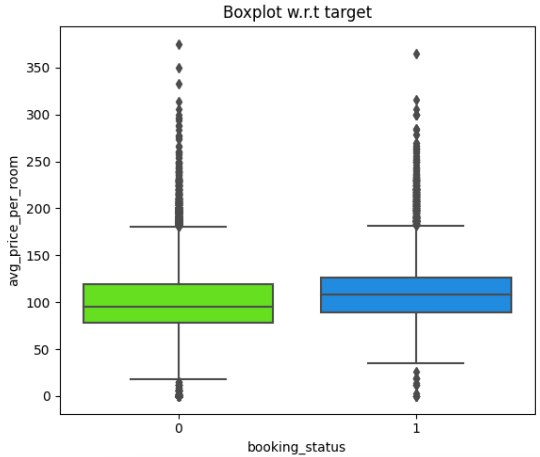

- There appears to be a small spike on the Average Price per Room as special request rise.
- The most significant spike in Average Price Per Room is 2 Special Requests
- 5 or more special requests appear to have little/neg impact and may loose variance.

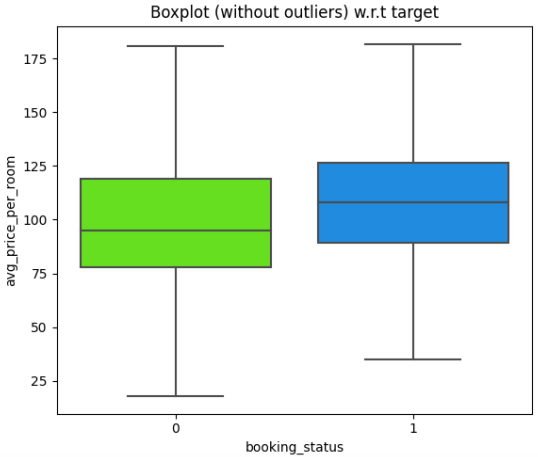
# Average Price per Room and Booking Status _ Correlation





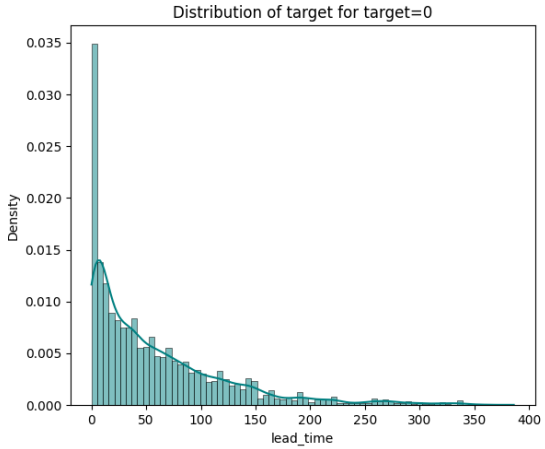


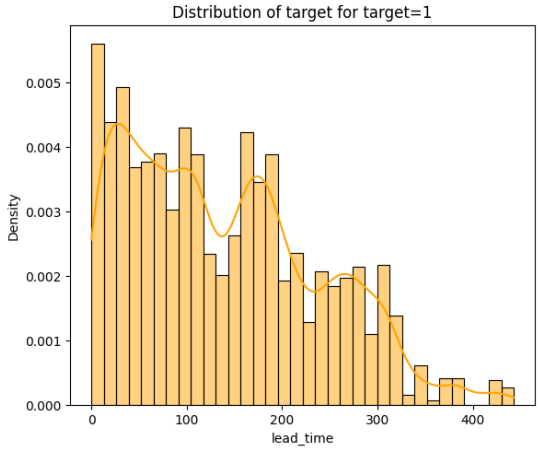


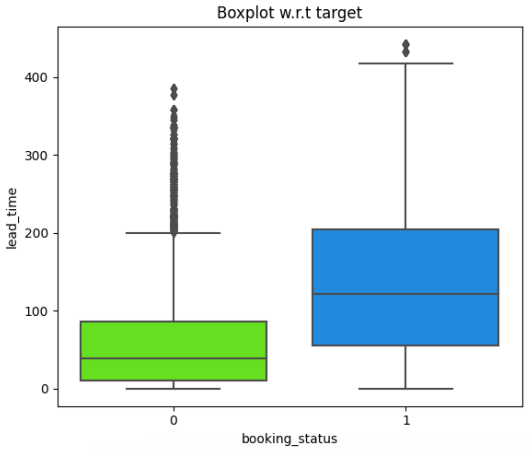



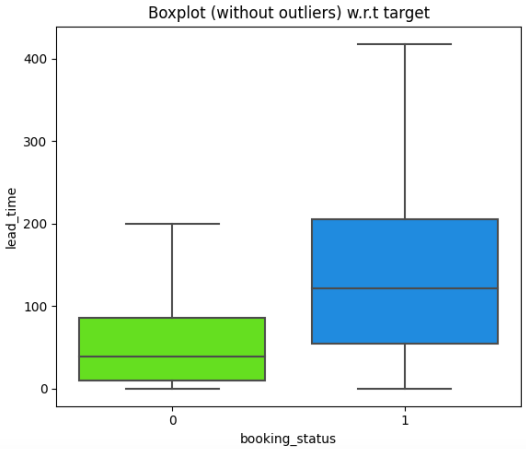

Average Price per Room may have an impact on cancellation ratios. It looks like a higher Average Price per Room may lead to a higher cancellation rate.

# Lead Time and Booking Status _ Correlation



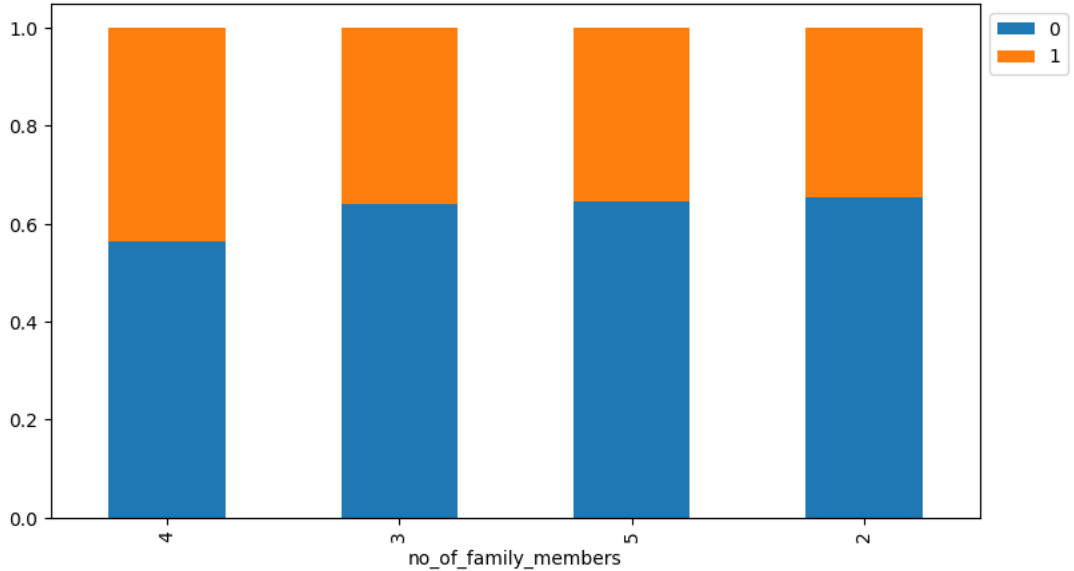



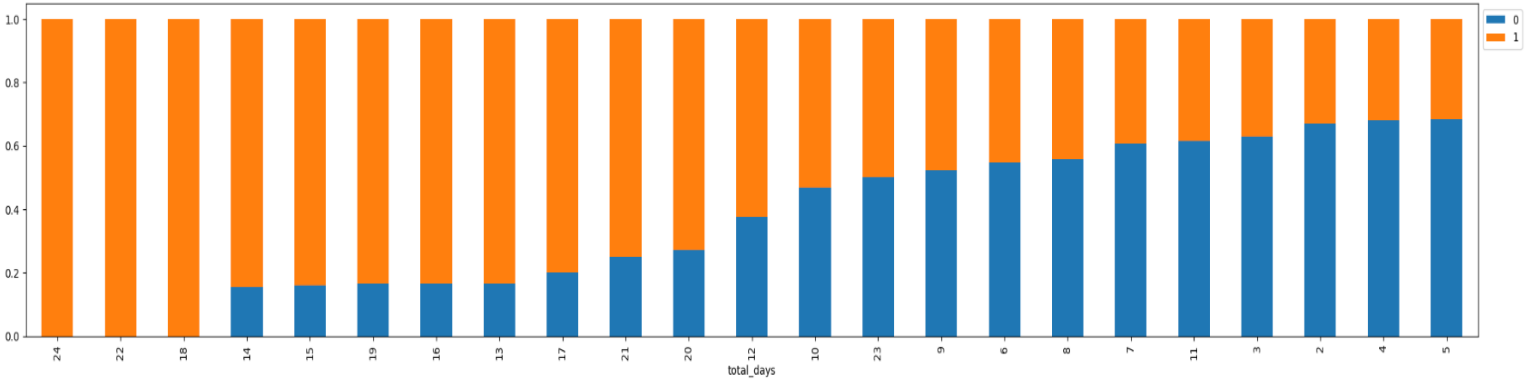

Lead time and booking status appear to have a correlation. It appears the longer the lead time the increased chance that the booking will be cancelled. It appears bookings with lead times over 100 days have a higher chance of cancellations.

# Number of Family Members and Booking Status_ Correlation



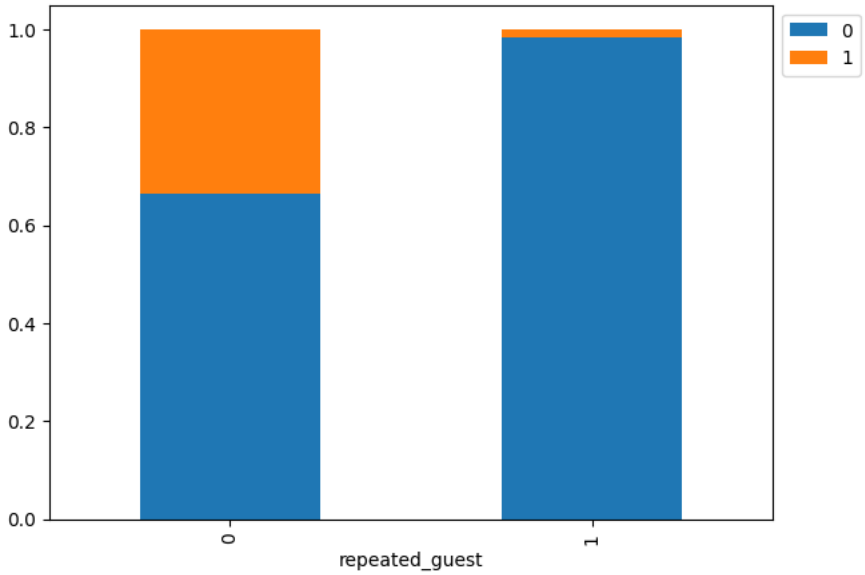

Number of family members does not seem to affect the number of cancellations.

# Number of Days and Booking Status_ Correlation



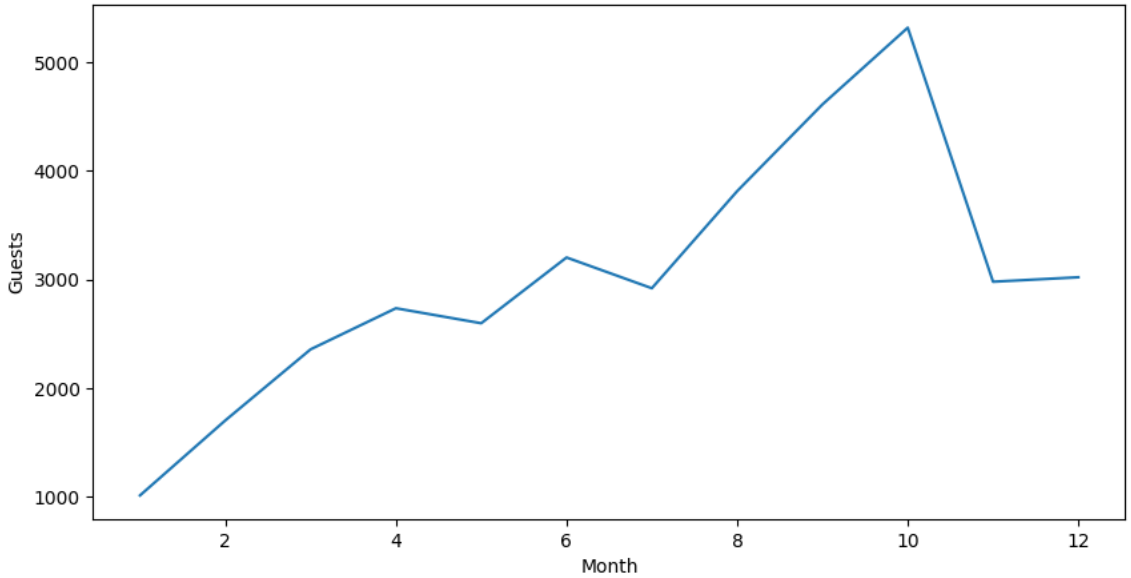

- It appears the booking with more days may result in more cancellations.
- Bookings single digits number of nights appear to have less cancellations.

# Repeat Guest vs. Booking Status _ Correlation



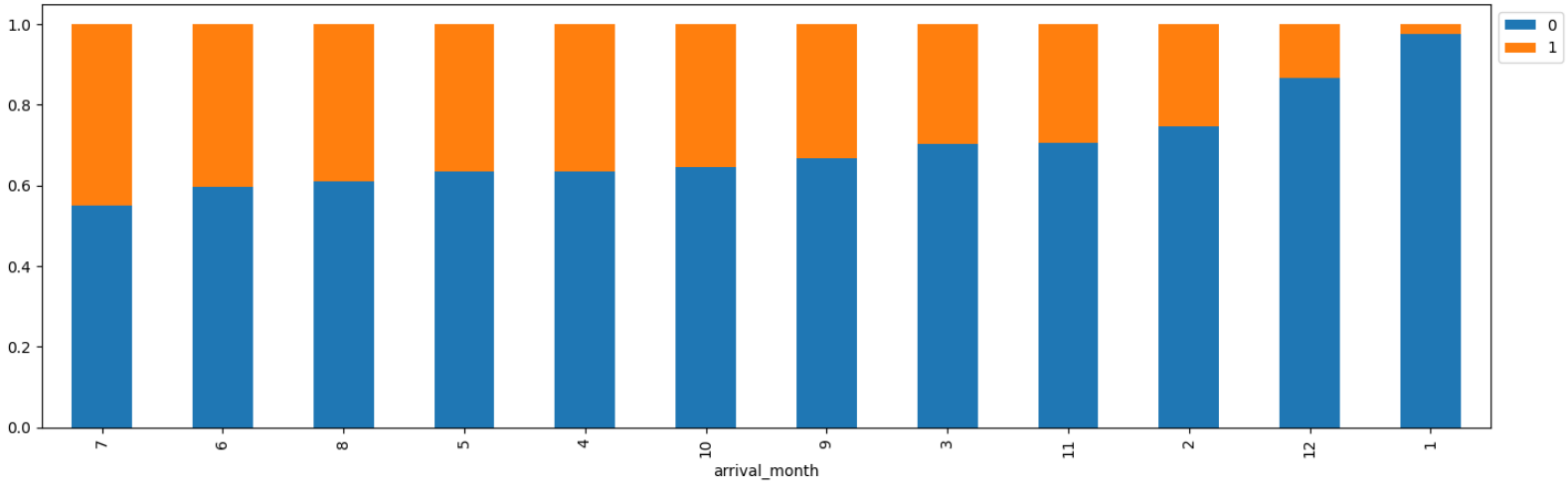

It appears repeat guests are less likely to cancel their bookings.

# Busiest month in hotel – number of guests



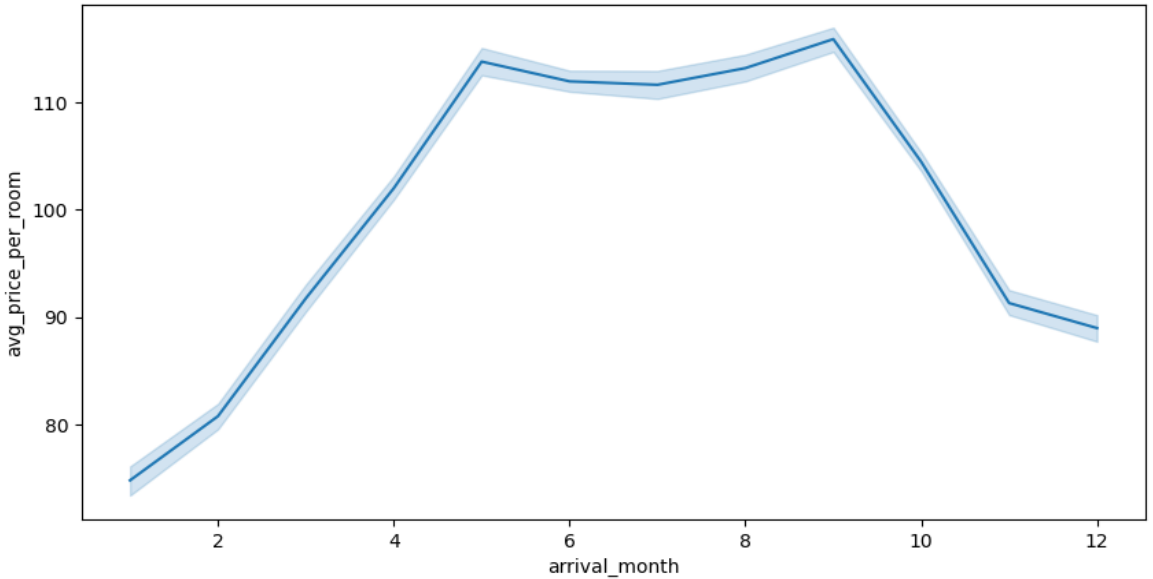

October appears to be the busiest month with a sharp drop in January and February.

# Percentage of bookings canceled each month




The summer months appear to have the highest level of cancellation where the winter months tend to have less.

# Average Price per Room and Arrival Month

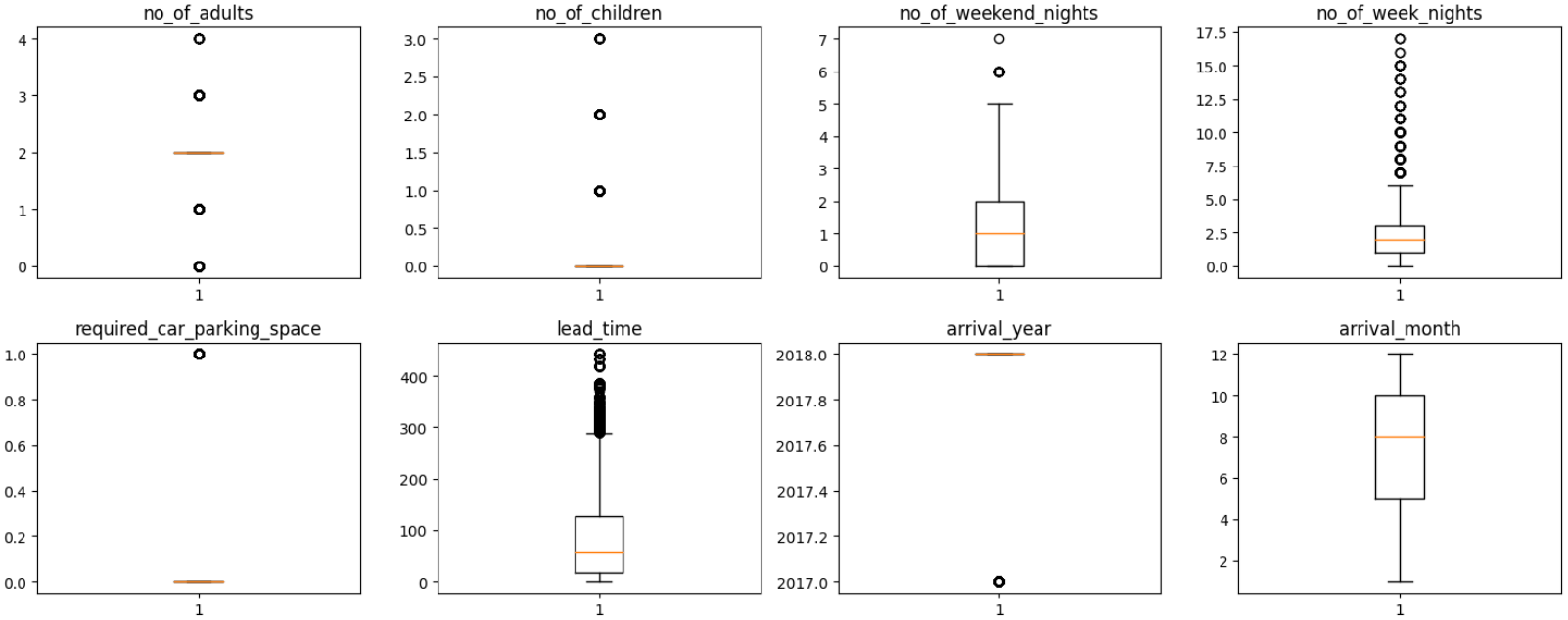



- The highest Average Price per Room appears to be in the summer months.
- The lowest Price per Room is in the winter months.
- Correlating with cancellations higher price and cancellations

# Data Processing

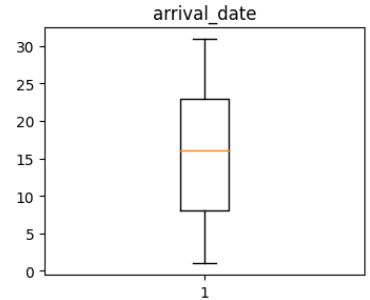
# Data Processing _ Outlier Check

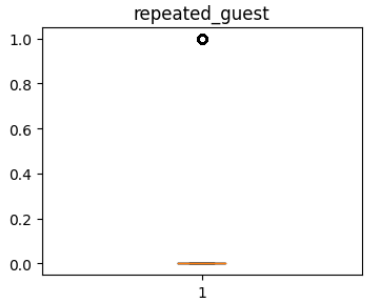
- There are quite a few outliers in the data- these all seem to appear in the dataset.
- However, we will not treat them as they are proper values

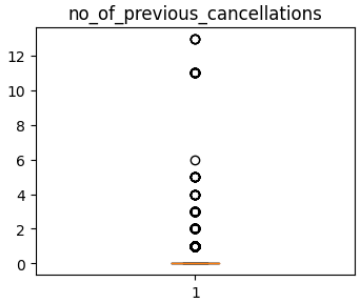


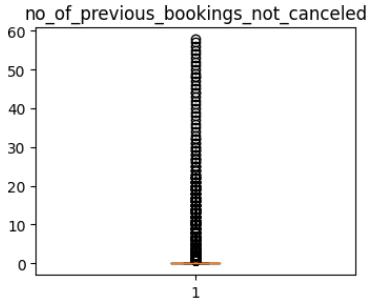

*Link to Appendix statistical summary and supporting univariate analysis slides*

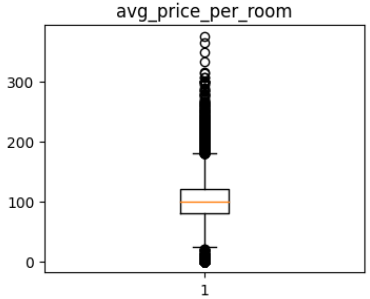
# Data Processing _ Outlier Check

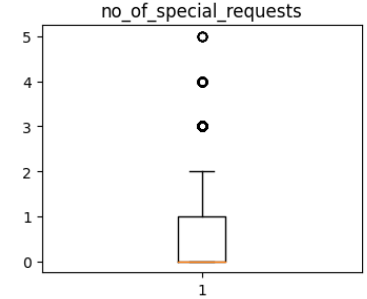



*Link to Appendix statistical summary and supporting univariate analysis slides*

# Model Building - Logistic Regression

# Data Preparation for modeling…Build Logistic Regression

```
Shape of Training set :  (25392, 28)
Shape of test set :  (10883, 28)
Percentage of classes in training set:
0   0.67064
1   0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0   0.67638
1   0.32362
Name: booking_status, dtype: float64
```

| Booking Status | Train | Test |
|---|---|---|
| 0 | 67% | 67.6% |
| 1 | 32.9% | 32.3% |

*Link to Appendix statistical summary and supporting univariate analysis slides*

# Model Building – Original Logistic Regression Model

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | booking_status | No. Observations: | 25392 |
| Model: | Logit | Df Residuals: | 25364 |
| Method: | MLE | Df Model: | 27 |
| Date: | Tue, 12 Dec 2023 | Pseudo R-squ.: | 0.3292 |
| Time: | 23:51:12 | 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 |

- Use statsmodels to check the validity of our data
- Negative coefficients reflect the a decreased probability of the booking being canceled.
- Positive coefficients reflect and increased probability of the booking being cancelled.
- There are some high p-values that should be addressed.
- F1 score is low – we will make adjustments. Starting with multicollinearity.

# Model Building – New Logistic Regression Model

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | booking_status | No. Observations: | 25392 |
| Model: | Logit | Df Residuals: | 25370 |
| Method: | MLE | Df Model: | 21 |
| Date: | Tue, 12 Dec 2023 | Pseudo R-squ.: | 0.3282 |
| Time: | 23:51:16 | 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 |

- After treating for multicollinearity - there is not a significant change in the logistic regression.
- The F1 score reduced slightly.

# Converting Coefficients to Odds – Increasing Odds

Table 1

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Odds** | 0.00000 | 1.11491 | 1.16546 | 1.11470 | 1.04258 | 0.20296 | 1.01583 | 1.57195 | 0.95839 | 0.06478 | 1.25712 | 1.01937 | 0.22996 |
| **Change_odd %** | -100.00000 | 11.49096 | 16.54593 | 11.46966 | 4.25841 | -79.70395 | 1.58331 | 57.19508 | -4.16120 | -93.52180 | 25.71181 | 1.93684 | -77.00374 |
| | | 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 | | | |
| | | 1.17846 | 1.33109 | 0.70104 | 0.75364 | 0.47885 | 0.37977 | 0.23827 | 0.45326 | 0.16773 | | | |
| | | 17.84641 | 33.10947 | -29.89588 | -24.63551 | -52.11548 | -62.02290 | -76.17294 | -54.67373 | -83.22724 | | | |

- Number of Adults increases the odds of cancelling the booking by 11.5%
- Number of Children increases the odds of cancelling the booking by 16.5%
- Number of Weekend Nights increased the odds of cancelling the booking by 11.5%
- Number of Week Nights increases the odds of cancelling the booking by 4.3%
- Lead Time increases the odds of cancelling the booking by 1.6%
- Arrival Year increases the odds of cancelling the booking by 57.2%
- Number of Previous Cancellations increases the odds of cancelling the booking by 25.7%
- Average Price Per Room increases the odds of cancelling the booking by 1.9%
- Type of Meal Plan-Meal Plan 1 increases the odds of cancelling the booking by 17.8%
- Type of Meal Plan- Not Selected  increases the odds of cancelling the booking by 33.1%

# Converting Coefficients to Odds - Decreasing Odds

Table 1

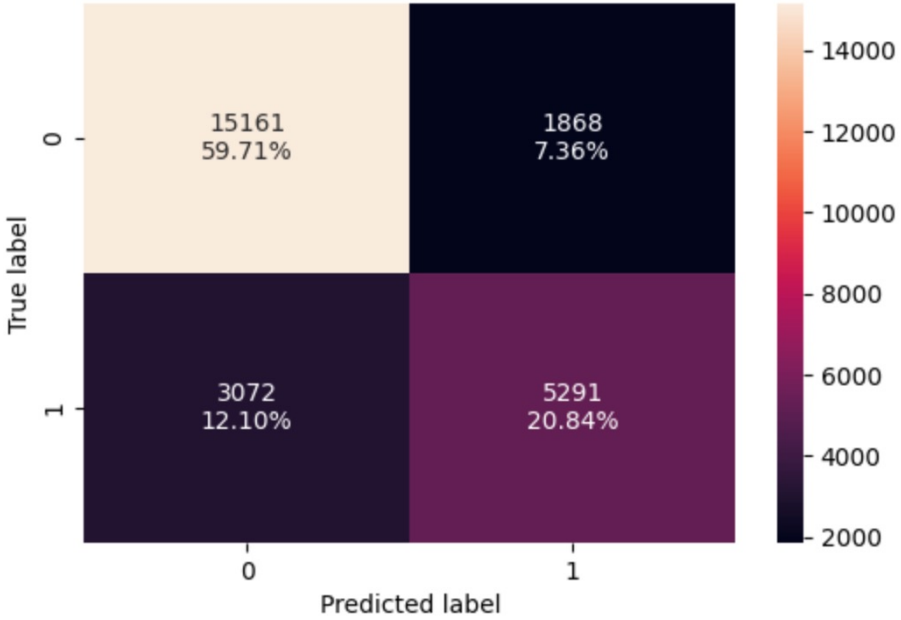
| | const | no_of_adults | no_of_children | no_of_weekend_nights | no_of_week_nights | required_car_parking_space | lead_time | arrival_year | arrival_month | repeated_guest | no_of_previous_cancellations | avg_price_per_room | no_of_special_requests |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Odds** | 0.00000 | 1.11491 | 1.16546 | 1.11470 | 1.04258 | 0.20296 | 1.01583 | 1.57195 | 0.95839 | 0.06478 | 1.25712 | 1.01937 | 0.22996 |
| **Change_odd %** | -100.00000 | 11.49096 | 16.54593 | 11.46966 | 4.25841 | -79.70395 | 1.58331 | 57.19508 | -4.16120 | -93.52180 | 25.71181 | 1.93684 | -77.00374 |
| | | 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 | | | |
| | | 1.17846 | 1.33109 | 0.70104 | 0.75364 | 0.47885 | 0.37977 | 0.23827 | 0.45326 | 0.16773 | | | |
| | | 17.84641 | 33.10947 | -29.89588 | -24.63551 | -52.11548 | -62.02290 | -76.17294 | -54.67373 | -83.22724 | | | |

- Required Car Parking Space decreases the odds of cancelling the booking by 79.7%
- Arrival Month decreases the odds of cancelling the booking by 4.2%
- Repeated Guest decreases the odds of cancelling the booking by 93.5%
- Number of Special Requests decreases the odds of cancelling the booking by 77%
- Room Type Reserved – Room Type 2 decreases the odds of cancelling the booking by 29.9%
- Room Type Reserved – Room Type 4 decreases the odds of cancelling the booking by 24.6%
- Room Type Reserved – Room Type 5 decreases the odds of cancelling the booking by 52.1%
- Room Type Reserved – Room Type 6 decreases the odds of cancelling the booking by 62%
- Room Type Reserved – Room Type 7 decreases the odds of cancelling the booking by 76.1%
- Market Segment Type_Corporate decreases the odds of cancelling the booking by 56.7%
- Market Segment Type_Offline decreases the odds of cancelling the booking by 83.2%

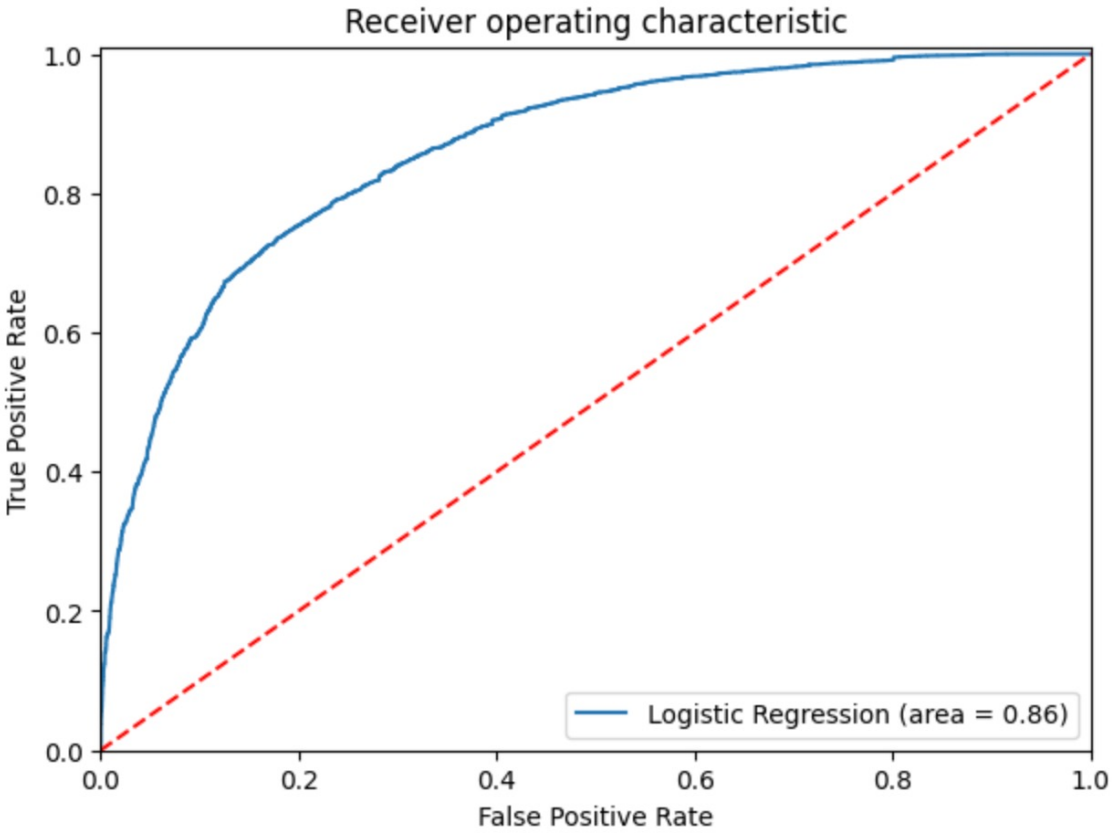
# Model Performance on Training Set

Training performance:

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.80545 | 0.63267 | 0.73907 | 0.68174 |



# ROC-AUC on Training Set



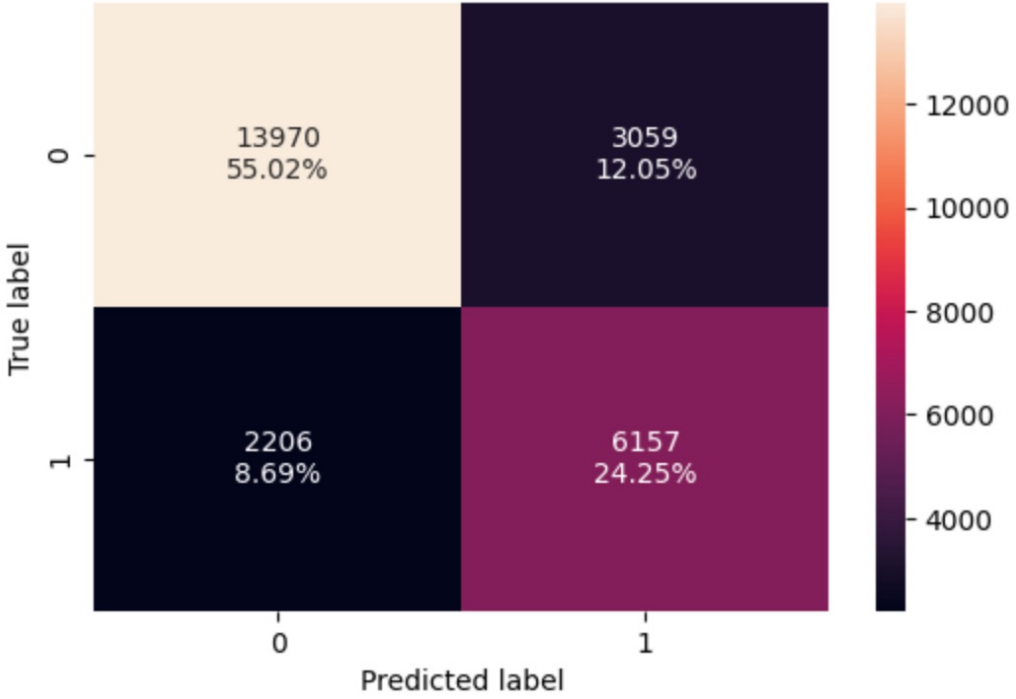

Logistic Regression model is giving a good performance on training set.

# Can Recall score be improved using the AUC-ROC Curve?

Optimal threshold using AUC-ROC curve

- **0.3700522558708252**

Confusion Matrix using the Threshold

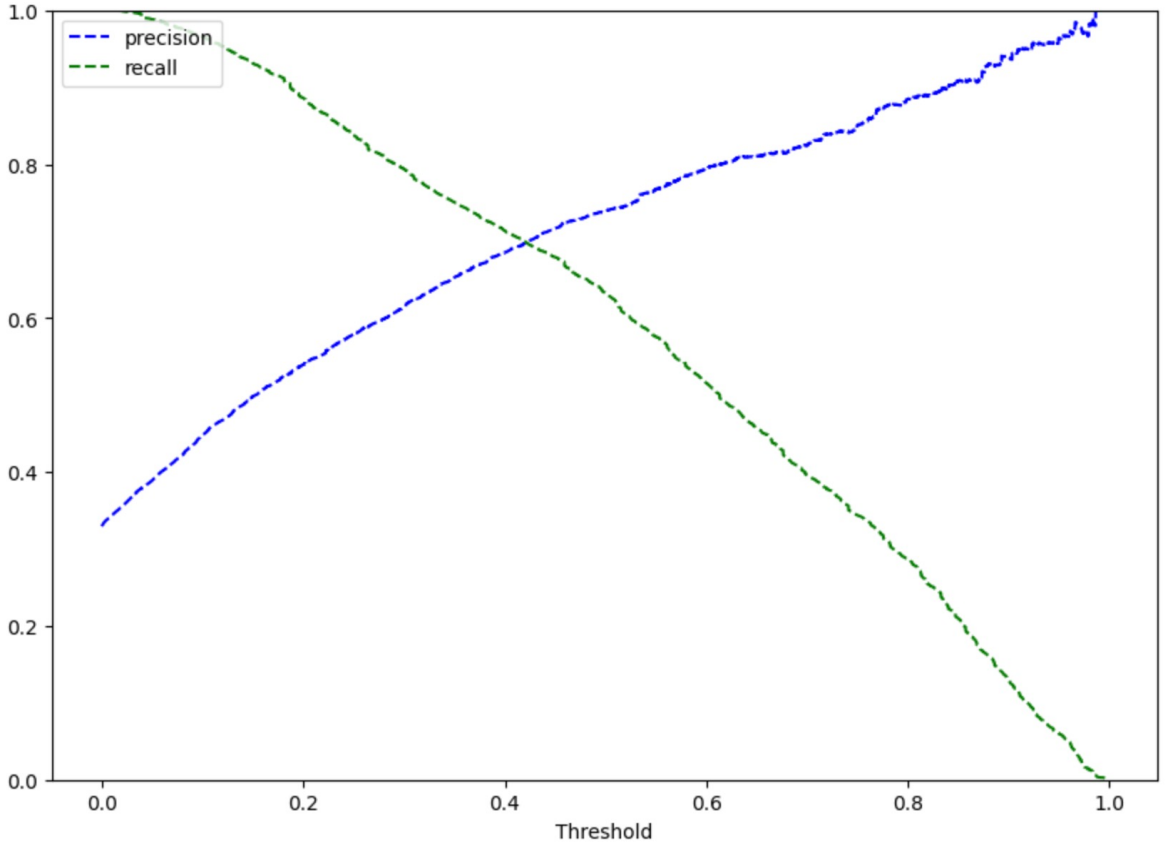



Training performance:

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| **0** | 0.79265 | 0.73622 | 0.66808 | 0.70049 |

- Recall and F1 has increased but the Accuracy and Precision have reduced.
- The model is still giving a good performance.

# Used Precision-Recall curve & see if we could find a better threshold




Optimal Threshold Curve = 0.42

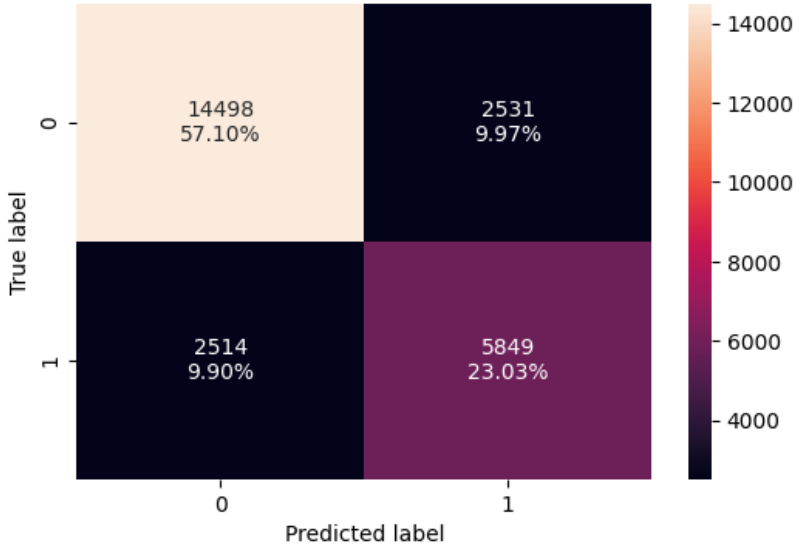
INN Hotels

# Checking performance of the training set

Optimal threshold

- **0.42**

Confusion Matrix using the Threshold




Training performance:

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.80132 | 0.69939 | 0.69797 | 0.69868 |

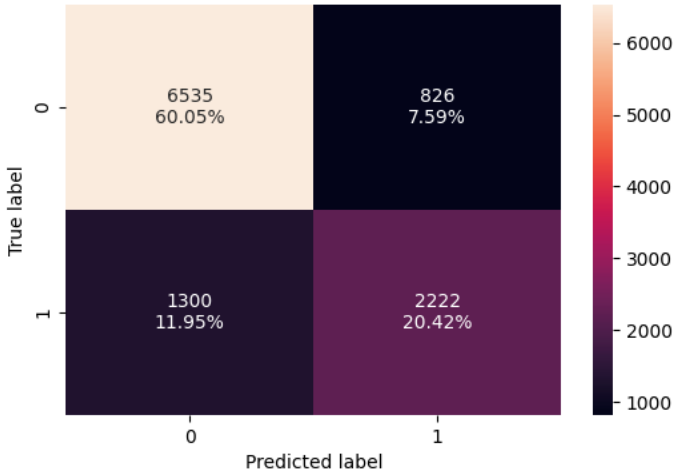
- Accuracy and Precision has increased but the Recall and F1 have reduced.
- The model is still giving a good performance.

# Checking performance of the test set

Optimal threshold

- **0.42**

Confusion Matrix using the Threshold

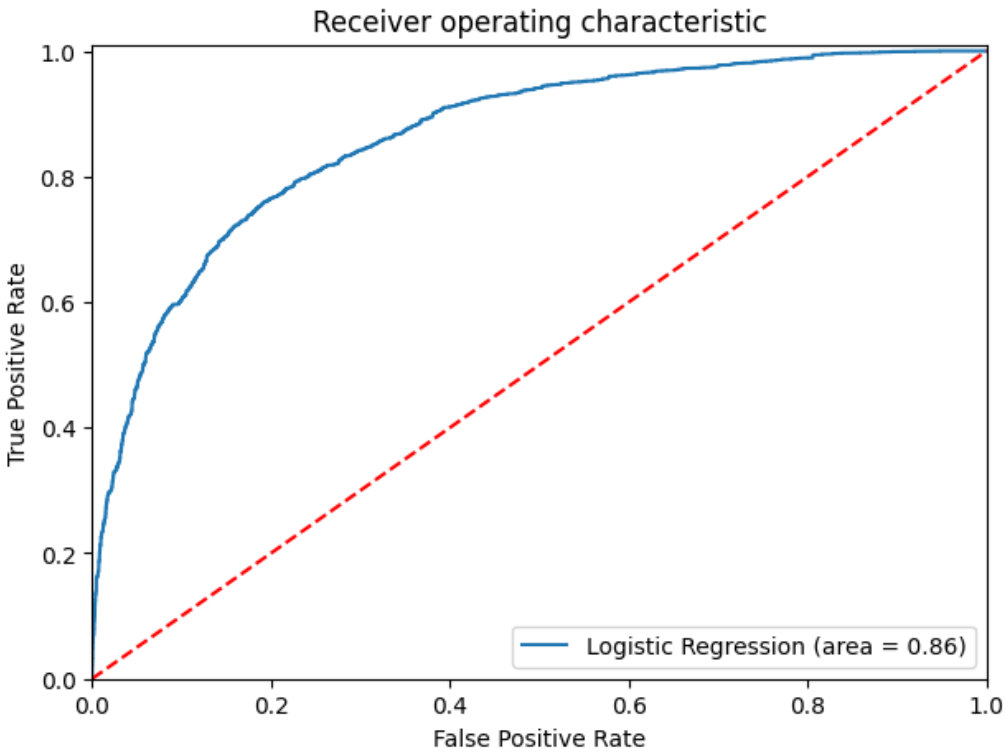



Test performance:

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| **0** | 0.80465 | 0.63089 | 0.72900 | 0.67641 |

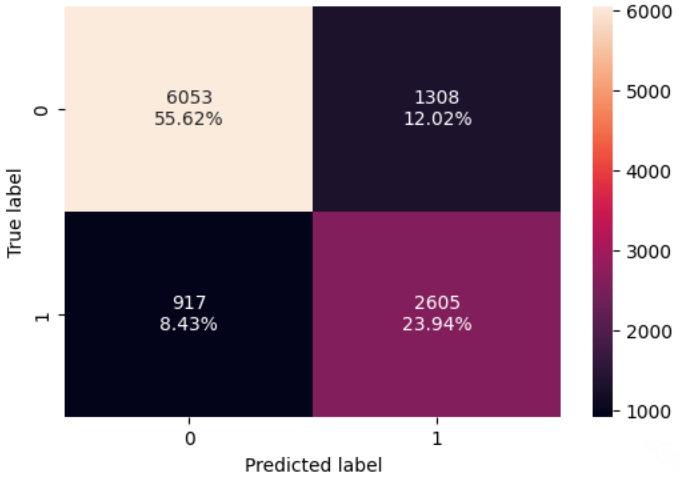
- Accuracy and Precision has increased but the Recall and F1 have reduced.
- The model is still giving a good performance.

# ROC-AUC on Test Set




Logistic Regression model is giving a good performance on training set.

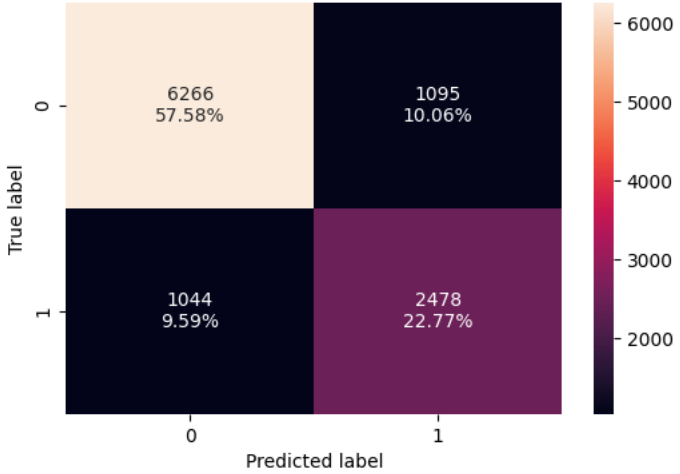
## Checking performance of the test set
## Optimal threshold .37




Test performance:

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.79555 | 0.73964 | 0.66573 | 0.70074 |

## Checking performance of the test set
## Optimal threshold .42




Test performance:

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.80345 | 0.70358 | 0.69353 | 0.69852 |

# Model Performance Evaluation & Improvement- Logistic Regression

- The training and testing set are both performing well without overfitting
- The model with the .37 threshold is giving the best F1 score.  Recommend selecting this model

Training performance comparison:

| | Logistic Regression-default Threshold | 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 | 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 |

# Model Building - Decision Tree

# Data Preparation for modeling…Decision Tree

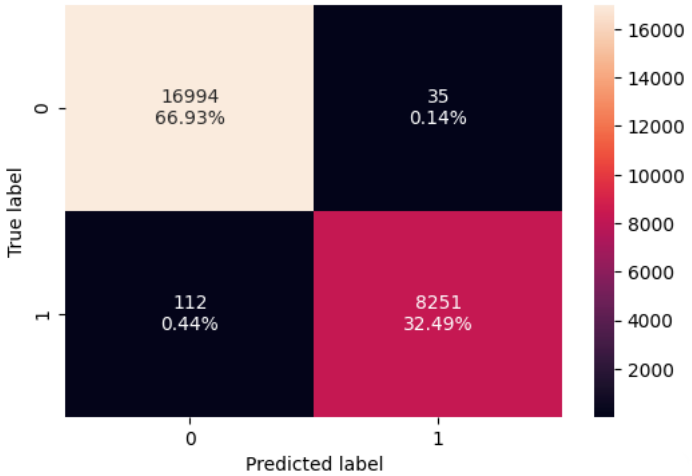
```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set:
0   0.67064
1   0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0   0.67638
1   0.32362
Name: booking_status, dtype: float64
```

| Booking Status | Train | Test |
|---|---|---|
| 0 | 67% | 67.6% |
| 1 | 32.9% | 32.3% |

*Link to Appendix statistical summary and supporting univariate analysis slides*

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

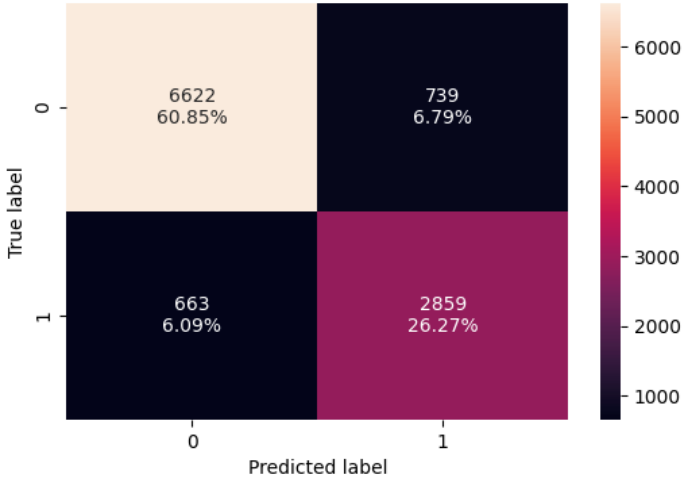
## Checking performance of the Training set




| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.99421 | 0.98661 | 0.99578 | 0.99117 |

F1 score is almost 100%

## Checking performance of the test set

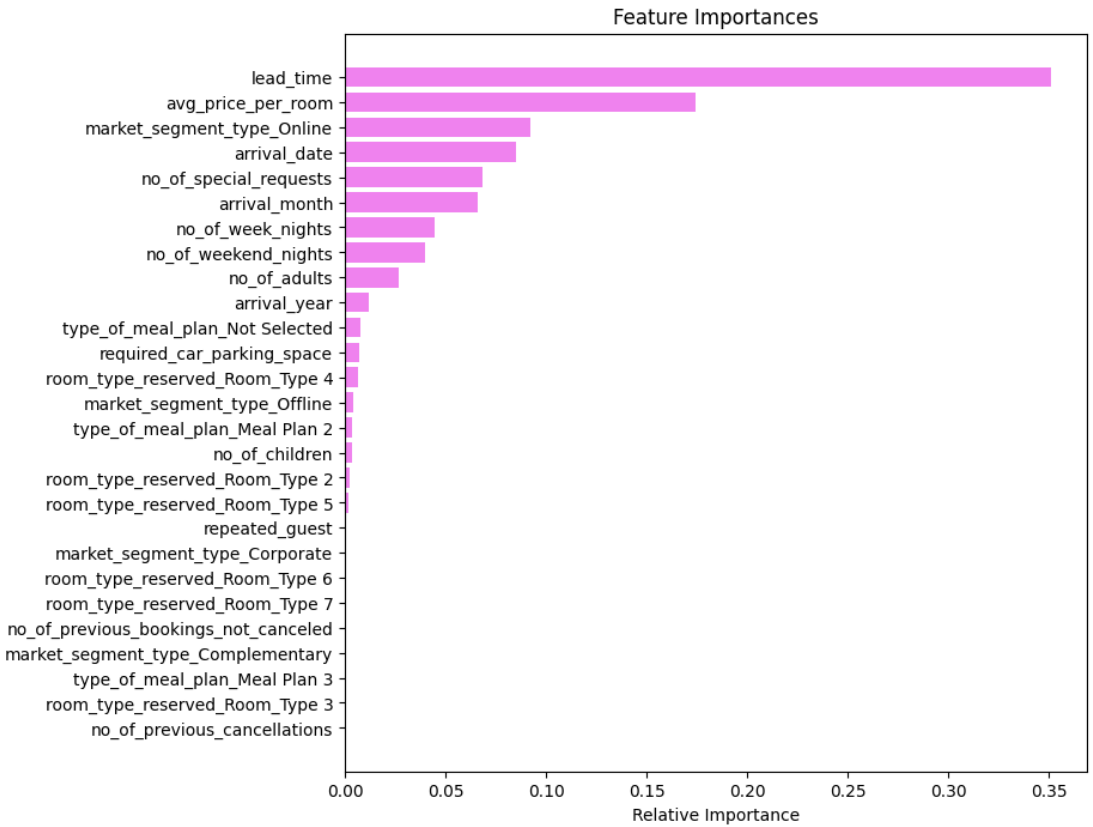



| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.87118 | 0.81175 | 0.79461 | 0.80309 |

F1 score is much less than 100%

Overfitting is occurring

# Feature Importance Pre-Pruning

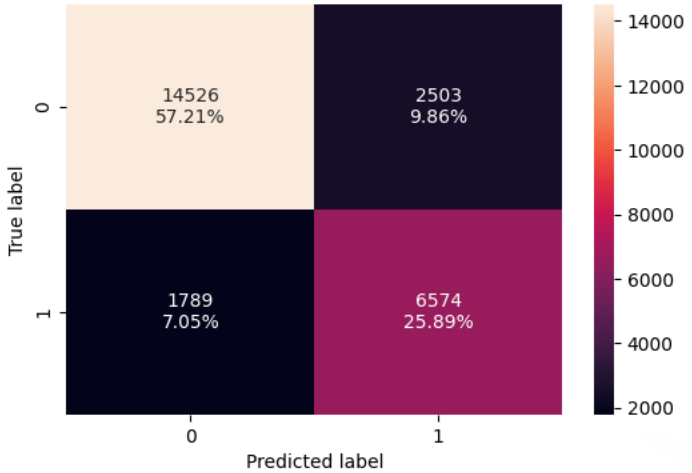



It appears Lead Time is the most important feature for predicting cancellations.

# Pre-Pruning

```
DecisionTreeClassifier
DecisionTreeClassifier(class_weight='balanced', max_depth=6, max_leaf_nodes=50,
                       min_samples_split=10, random_state=1)
```

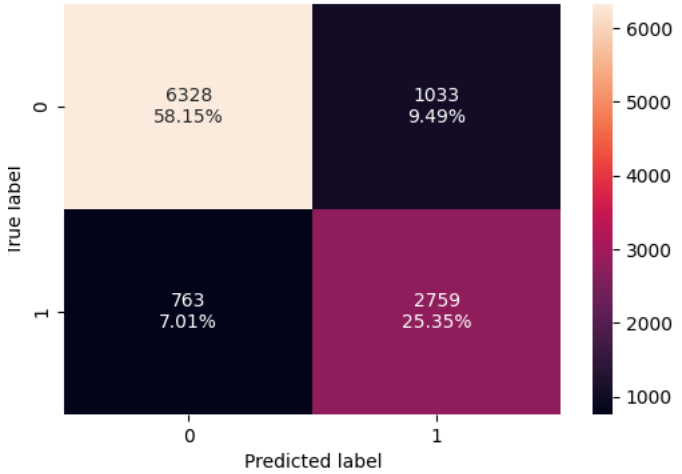
## Checking performance of the Training set




| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.83097 | 0.78608 | 0.72425 | 0.75390 |

F1 score is ~ 75.4%

## Checking performance of the test set

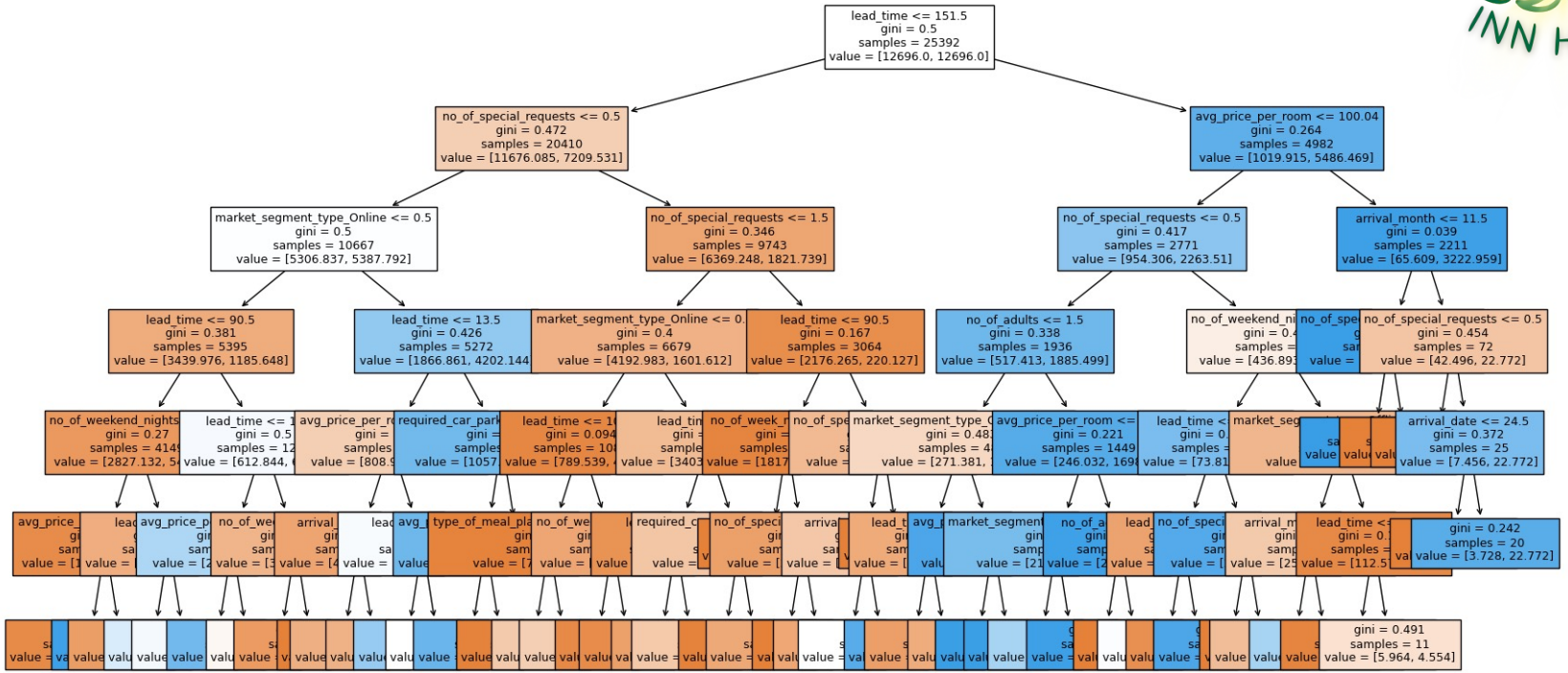



| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.83497 | 0.78336 | 0.72758 | 0.75444 |

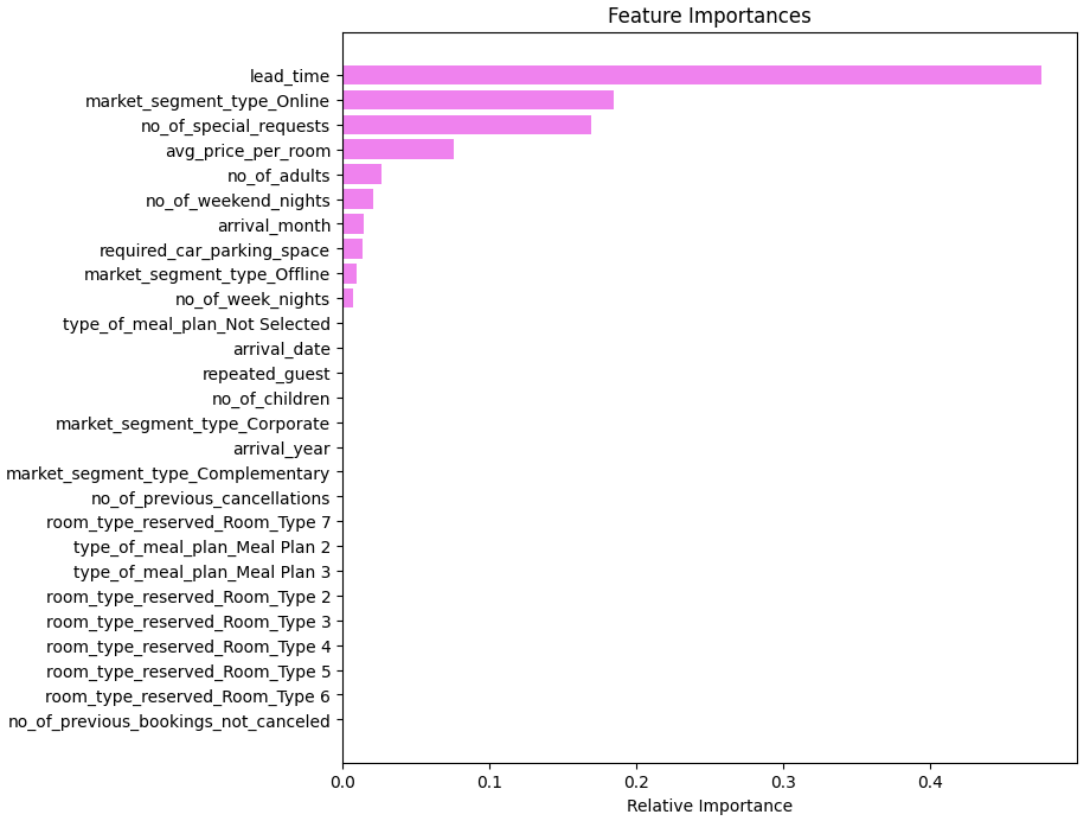
F1 score is ~ 75.4%

It appears Overfitting has been corrected

# Visualizing the Decision Tree – Pre Pruning



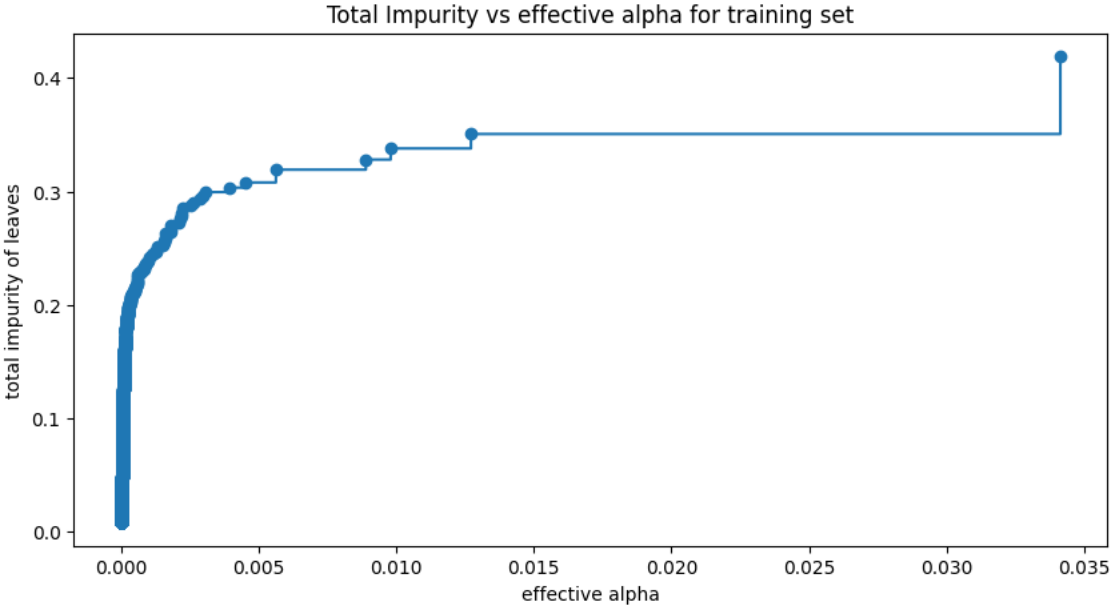
# Feature Importance After Pre-Pruning




It appears Lead Time is still the most important feature for predicting cancellations.

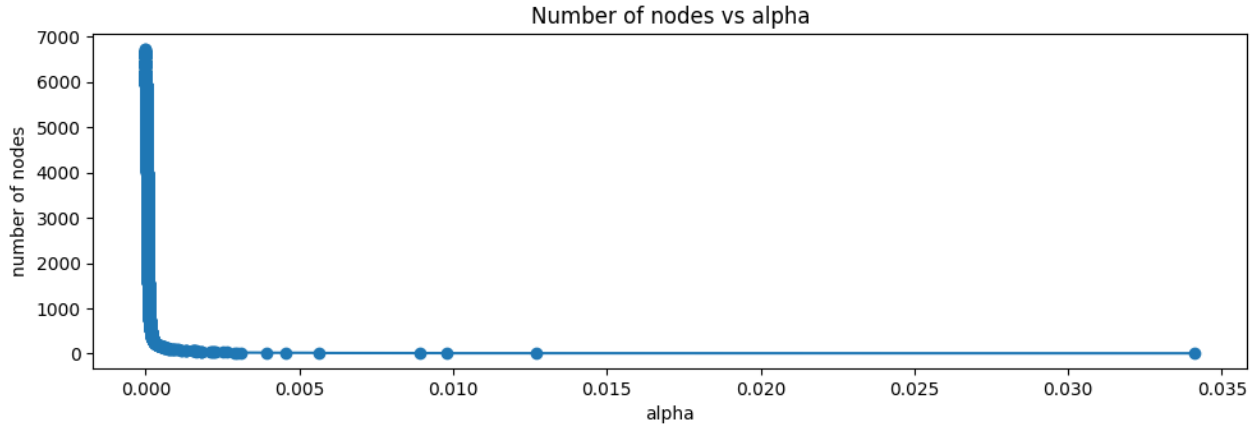
However, Market Segment Type (Online) and Number of Special Requests have moved up in level of importance.

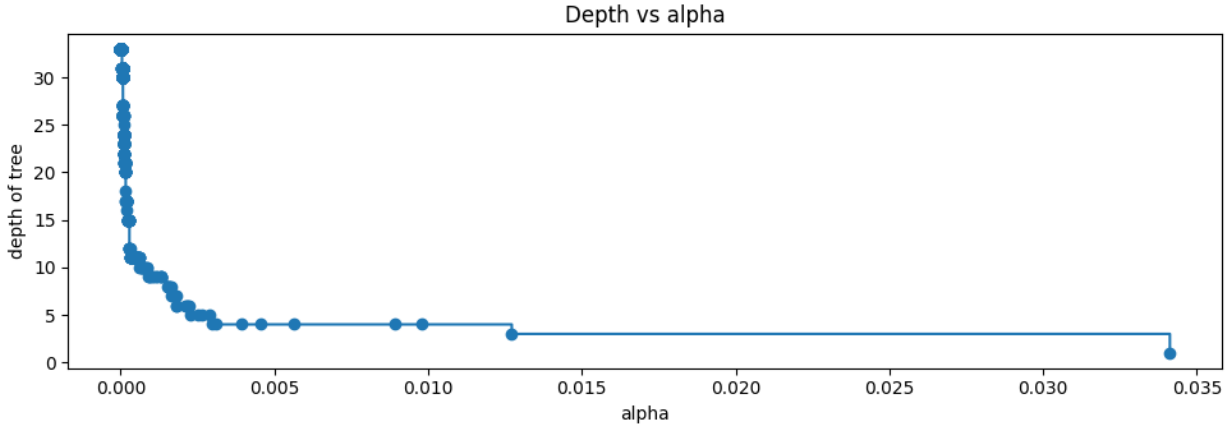
# Cost Complexity Pruning



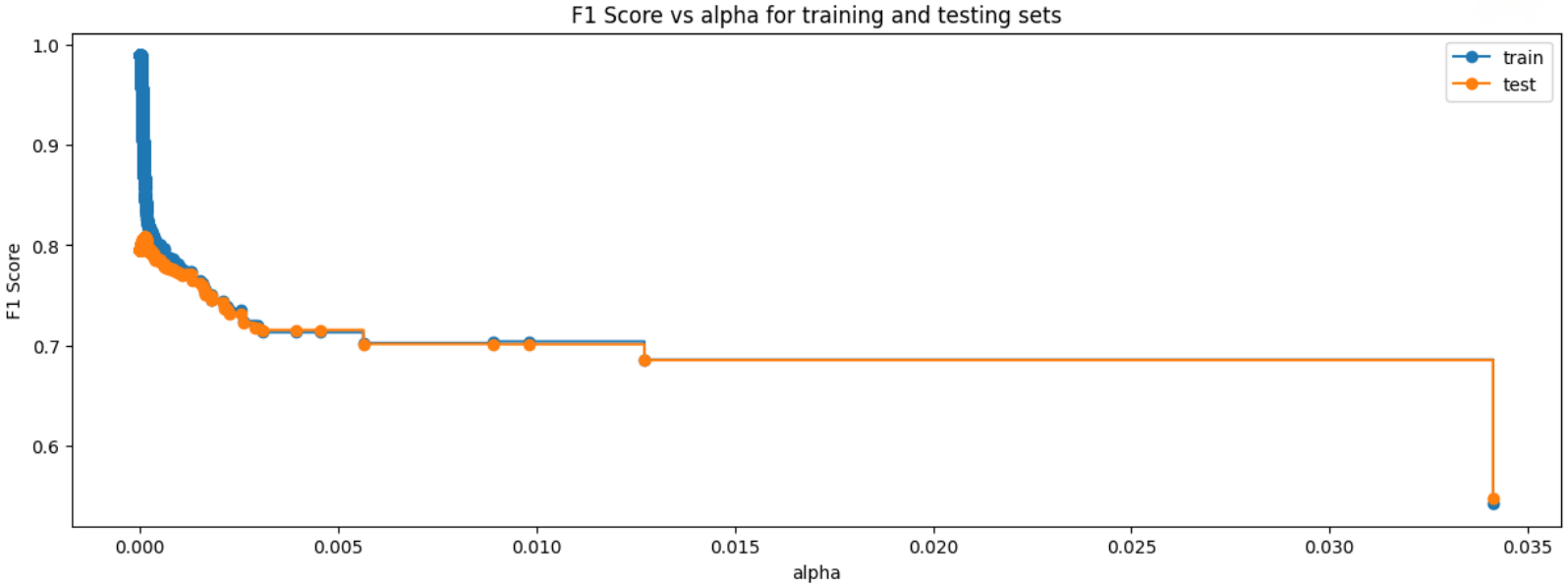

Number of nodes in the last tree is: 1 with ccp_alpha: 0.0811791438913696

# Cost Complexity Pruning Cont…



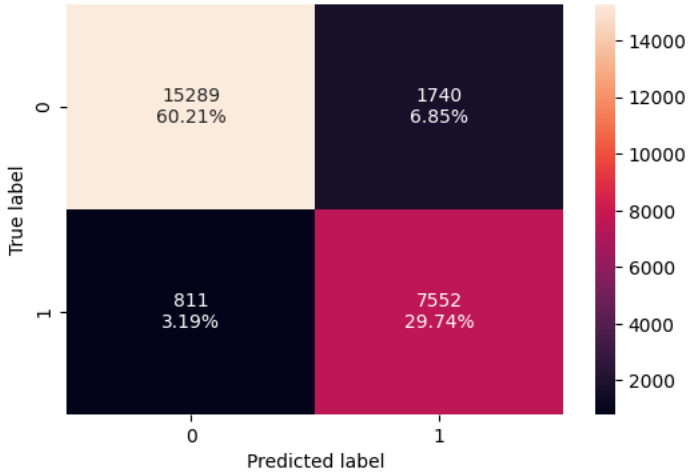



# Cost Complexity Pruning Cont…




DecisionTreeClassifier(ccp_alpha=0.00012267633155167043, class_weight='balanced', random_state=1)

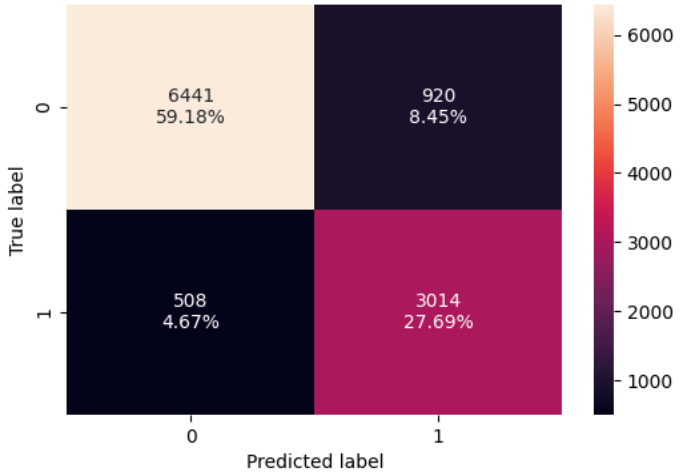
## Checking performance of the Training set Post-Pruning




| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.89954 | 0.90303 | 0.81274 | 0.85551 |

F1 score is ~ 85.6%

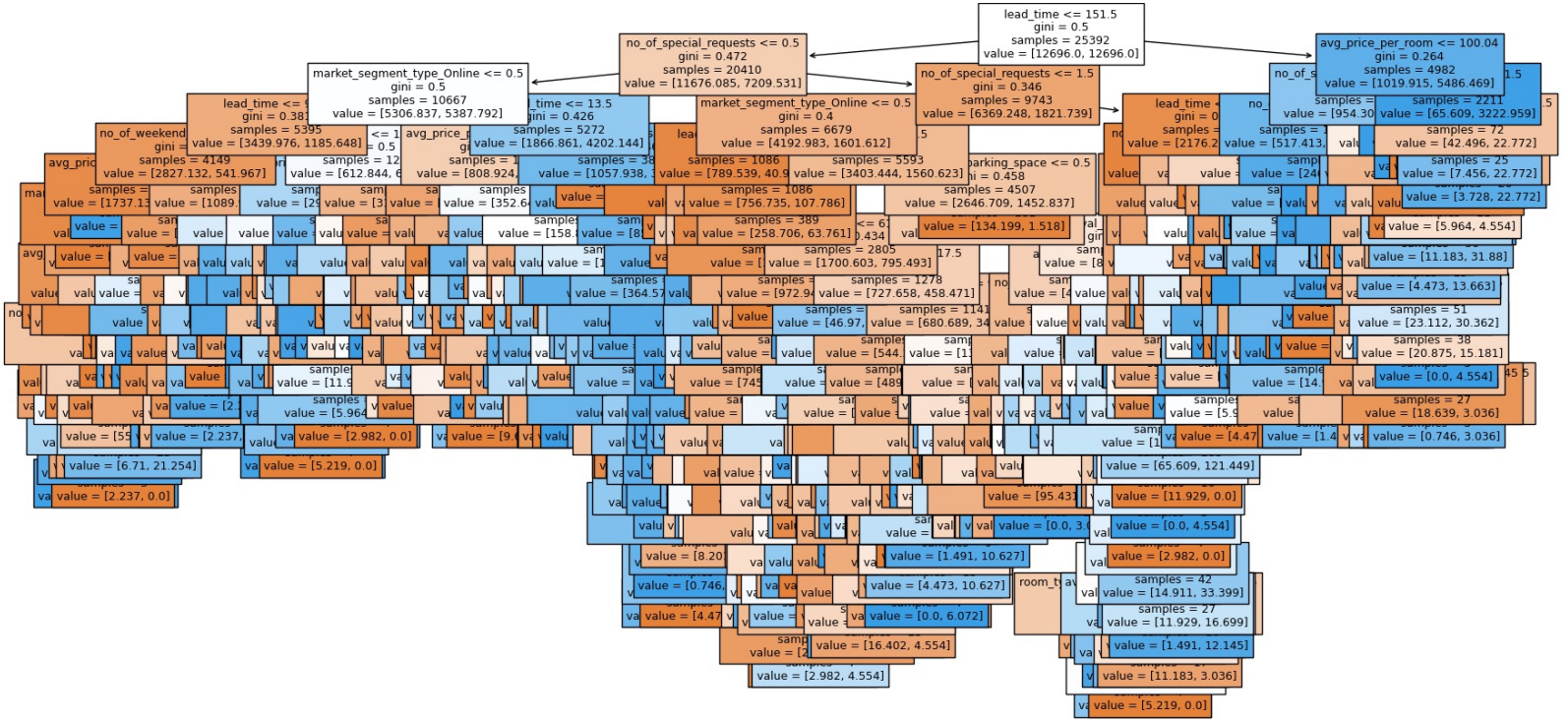
## Checking performance of the Test set Post- Pruning



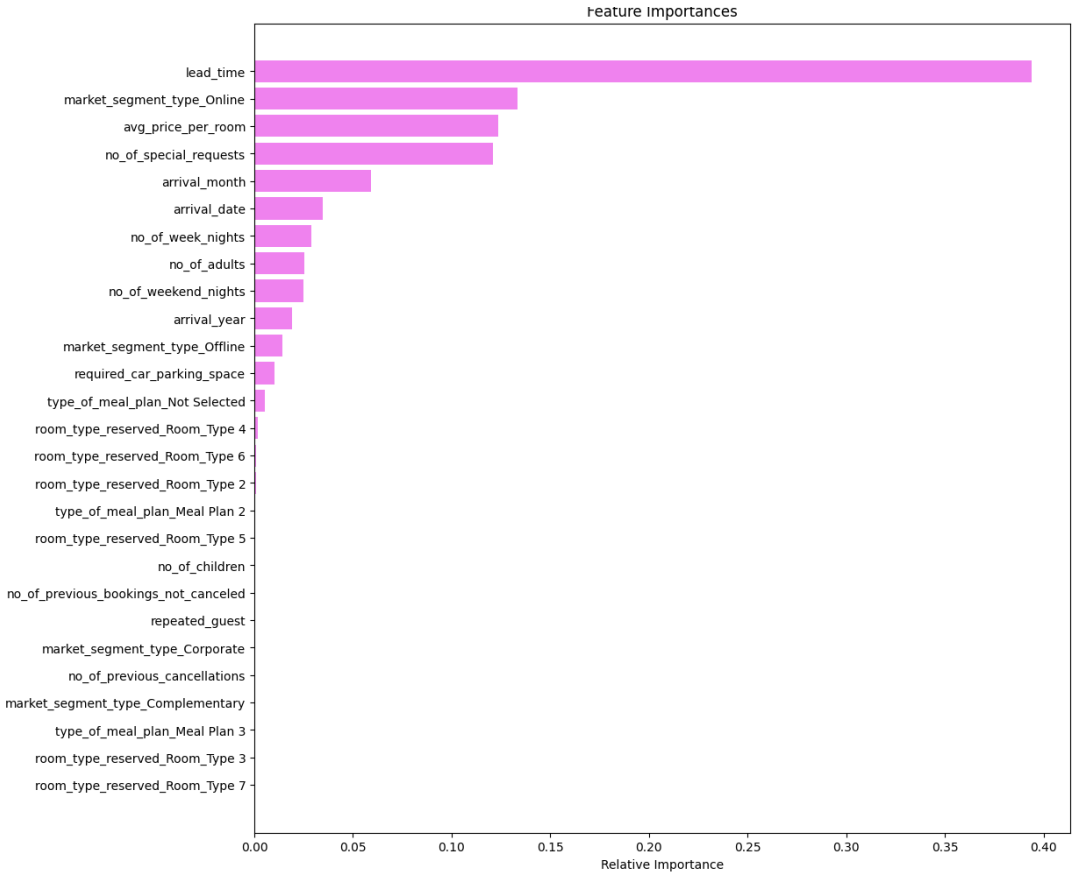

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| 0 | 0.89954 | 0.90303 | 0.81274 | 0.85551 |

F1 score is ~ 85.6%

# Visualizing the Decision Tree – Post-Pruning



# Feature Importance Post-Pruning




It appears Lead Time is still the most important feature for predicting cancellations.

# Model Performance Evaluation & Improvement – Decision Tree

- The training and testing set are both performing well without overfitting
- The post-pruning decision tree appears to have the best F1 score - Recommend selecting this model

Training performance comparison:

| | Decision Tree sklearn | Decision Tree (Pre-Pruning) | Decision Tree (Post-Pruning) |
|---|---|---|---|
| **Accuracy** | 0.99421 | 0.83097 | 0.89954 |
| **Recall** | 0.98661 | 0.78608 | 0.90303 |
| **Precision** | 0.99578 | 0.72425 | 0.81274 |
| **F1** | 0.99117 | 0.75390 | 0.85551 |

Test performance comparison:

| | Decision Tree sklearn | Decision Tree (Pre-Pruning) | Decision Tree (Post-Pruning) |
|---|---|---|---|
| **Accuracy** | 0.87118 | 0.83497 | 0.89954 |
| **Recall** | 0.81175 | 0.78336 | 0.90303 |
| **Precision** | 0.79461 | 0.72758 | 0.81274 |
| **F1** | 0.80309 | 0.75444 | 0.85551 |