Influencing Factors on Booking Cancellations

Classifications - Pam Lozano

December 16, 2023

Contents / Agenda

- Data Information
- Business Problem Overview and Solution Approach
- Data Background and Contents
- EDA Results
- Data Preprocessing
- Logistics Regression
- Decision Tree
- Pre Pruning
- Post Pruning
- Executive Summary

Objective

Analyze hotel booking data to identify factors influencing cancellations and build a predictive model for advanced cancellation prediction, aiding in the formulation of effective cancellation management policies.

Key Focus Areas:

1. Identify influential factors affecting booking cancellations.
2. Build a predictive model for advanced cancellation prediction.
3. Formulate policies based on data-driven insights to optimize cancellation management for INN Hotels Group.

Data Information

The data contains the different attributes of customers' booking details.

Booking_ID	the unique identifier of each booking
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
no_of_week_nights	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
type_of_meal_plan	Type of meal plan booked by the customer:
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)
room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
arrival_month	Month of arrival date
arrival_date	Date of the month
market_segment_type	Market segment designation.
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
no_of_special_requests	Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
booking_status	Flag indicating if the booking was canceled or not.

Business Problem Overview

- INN Hotels Group, a prominent hotel chain in Portugal, is seeking to optimize its approach to booking cancellations for enhanced revenue and operational efficiency.
- The company must address A comprehensive analysis of factors influencing booking cancellations.
- The company should look at Strategic positioning of products in the market based on market dynamics and customer preferences.
- Implementation of data-driven strategies to predict cancellations in advance is necessary.
- Formulation of efficient cancellation policies to improve customer satisfaction and loyalty.
- Exploration of eco-friendly practices to minimize waste and enhance sustainability is part of the overarching objective.

The ultimate aim is to dominate the hospitality market by maximizing profit and implementing sustainable practices.

Solution Approach

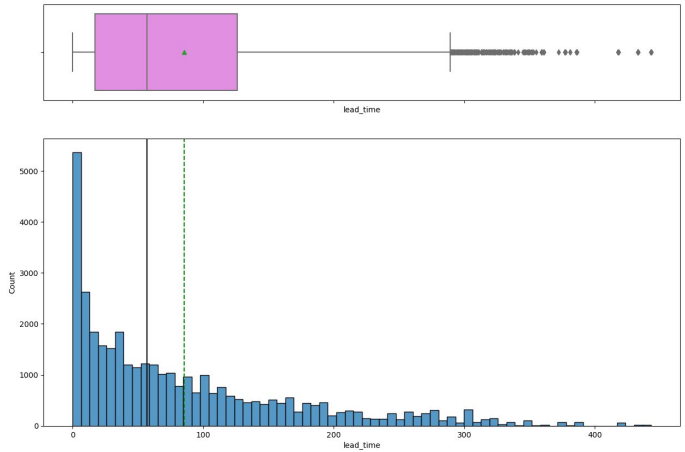
- **Data Analysis and Feature Identification:** Analyze data for booking cancellation influencers like lead time, room types, and customer history.
- **Exploratory Data Analysis (EDA):** Conduct EDA to position hotel offerings by studying the market and understanding guest preferences.
- **Predictive Models:** Develop models using past data to predict future cancellations, considering past cancellations, guest preferences, and market trends.
- **Customer Engagement and Loyalty Programs:** Implement programs for guest engagement and loyalty, including eco-friendly initiatives.
- **Machine Learning Model :** Choose various machine learning models for price prediction.
- **Efficient Cancellation Policies:** Establish clear and transparent cancellation policies, offering flexible options to meet revenue goals.
- **Technology Integration:** Utilize data tools and AI for real-time insights and explore technologies like chatbots for enhanced guest interactions.
- **Collaborate with Partners:** Work with industry partners to stay updated on trends and form partnerships for wider visibility.
- **Recommendations:** Based on insights, provide actionable recommendations for improving the landing page and increasing conversions.

Data Background and Contents

- Our dataset comprises 36,275 rows and 19 columns, providing a comprehensive foundation for our analysis.
- There are a mix of data types in our dataset: integers, objects, and floats, contributing to the diversity of information available for our analysis.
- There are no missing values in our dataset.
- The average no of previous cancellations is .02 serving as a central reference point in understanding the pricing dynamics of our dataset.
- The lowest and median no of previous cancellations is 0, providing a measure of central tendency in our dataset.
- The highest no of previous cancellations in our dataset is 13, representing the upper limit within our analyzed data.

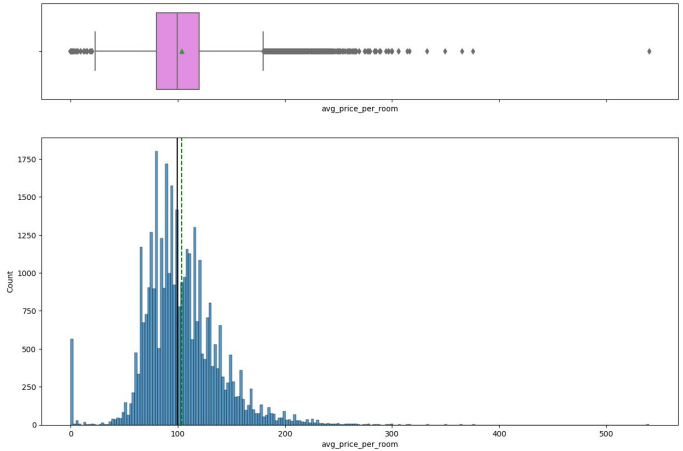
Exploratory Data Analysis

Univariate Analysis - Lead Time



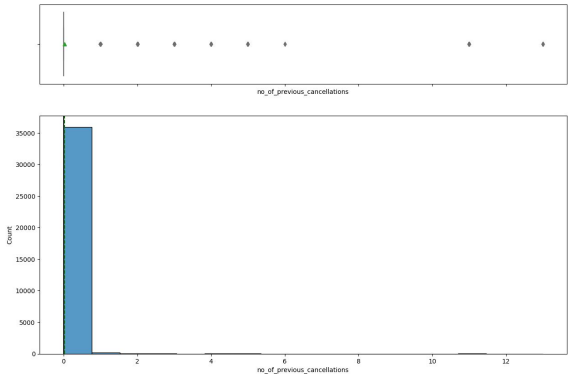
- Outliers are evident on the right tail of the distribution, suggesting potential extreme values.
- The distribution of lead time is right-skewed, with a median of around 60 days.

Univariate Analysis - Average Price per Room



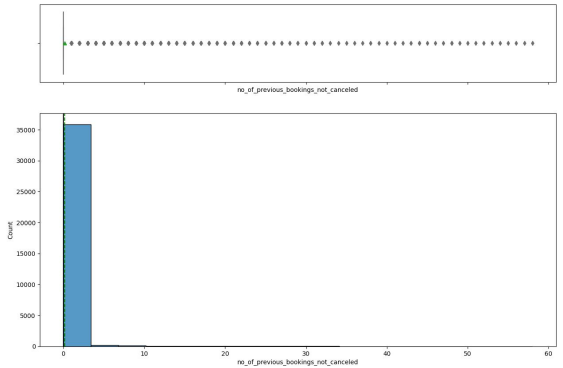
- The distribution exhibits several outliers, indicating potential unusual data points.
- The median average price per room is 100 euros.

Univariate Analysis - No of Previous Booking Cancellations



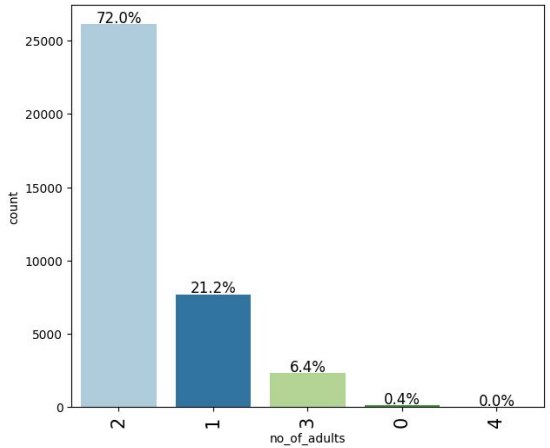
- We've identified some outliers in this distribution, suggesting potential deviations from the overall pattern
- The average number of previous booking cancellations is 0, warranting further in-depth analysis.

Univariate Analysis - Number of Previous Booking Not Cancelled



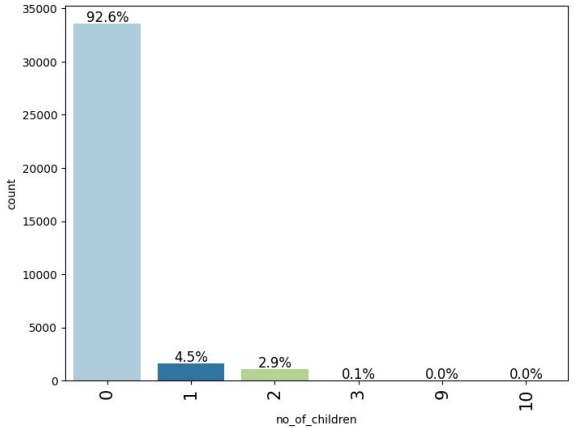
- There are outliers present in this distribution.
- The count of previous bookings not canceled is 0.

Univariate Analysis - Number of Adults



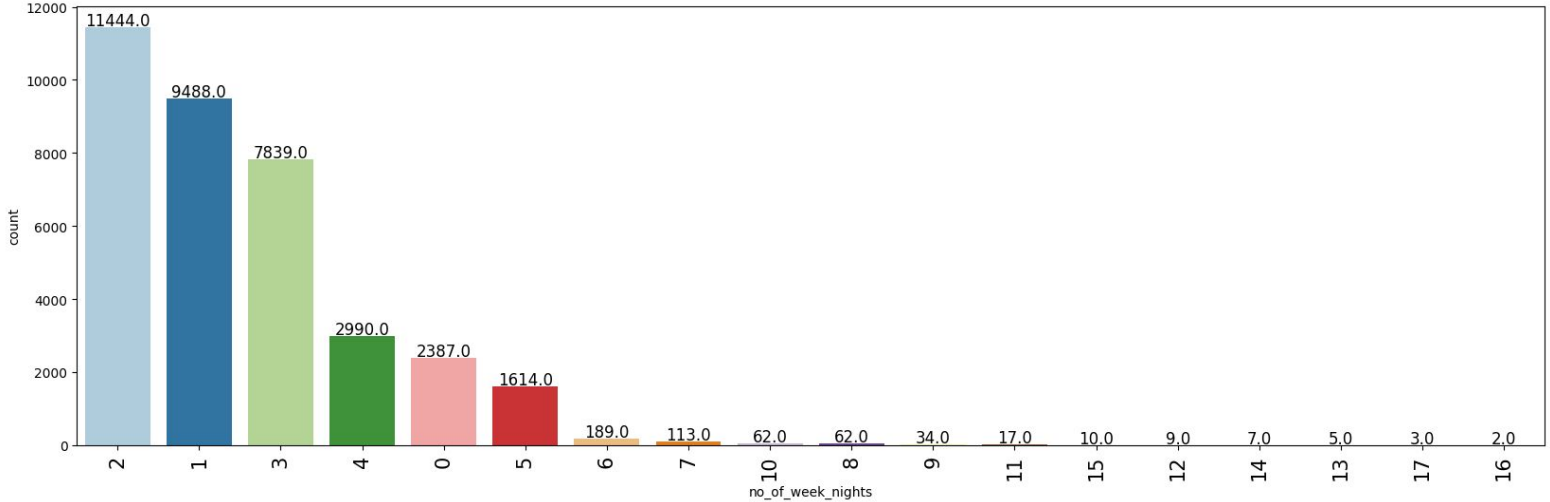
- 72% of reservations involve two adults.

Univariate Analysis - Number of Children



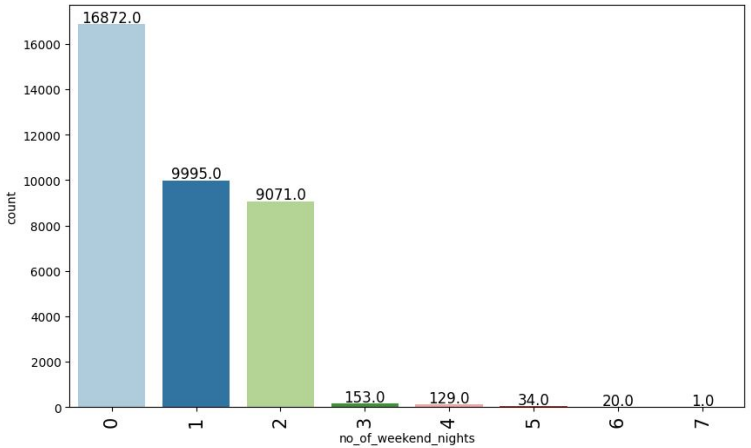
- The majority of reservations (92.6%) involve zero children.

Univariate Analysis - Number of Week Nights



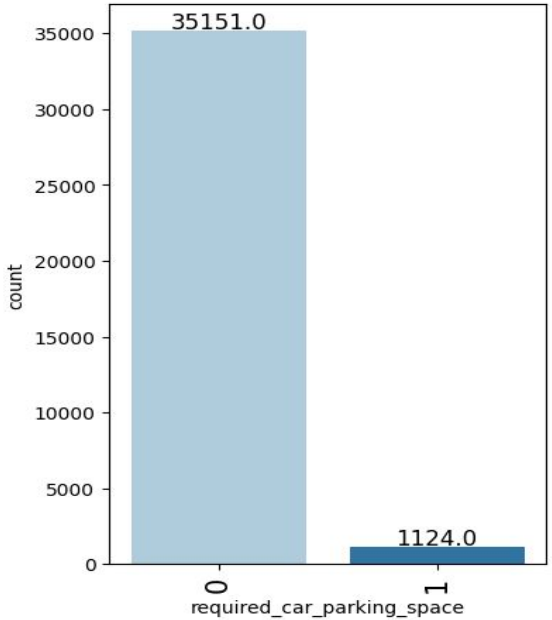
- The most common number of weeknights is 2, with 11,444 occurrences.

Univariate Analysis - Number of Weekend Nights



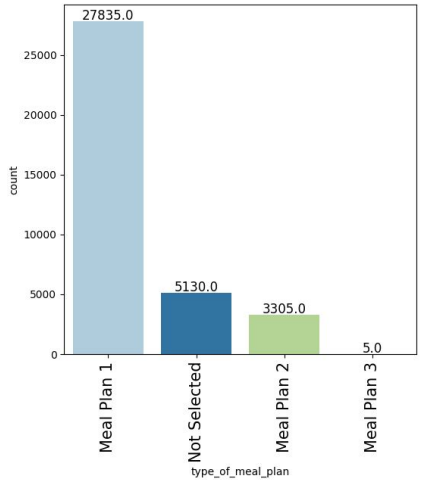
- The highest count for weekend nights is 0, with 16,872 occurrences.

Univariate Analysis - Required Car Parking Space



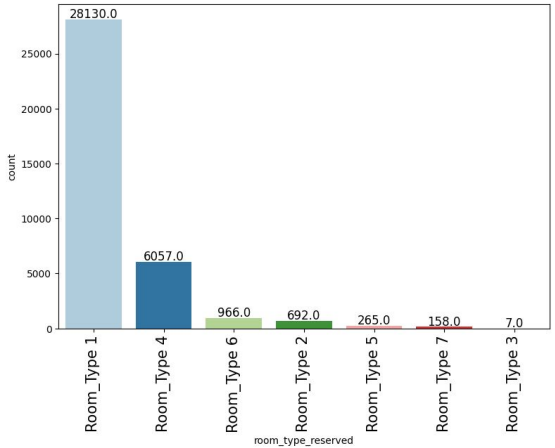
- Most customers do not require a car parking space.

Univariate Analysis - Type of Meal Plan



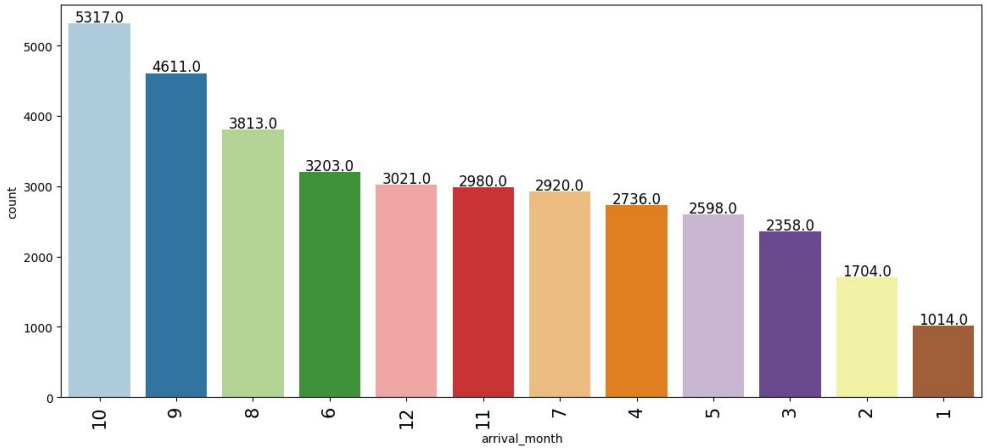
- Most customers selected breakfast as their preferred meal plan.

Univariate Analysis - Room Type Reserved



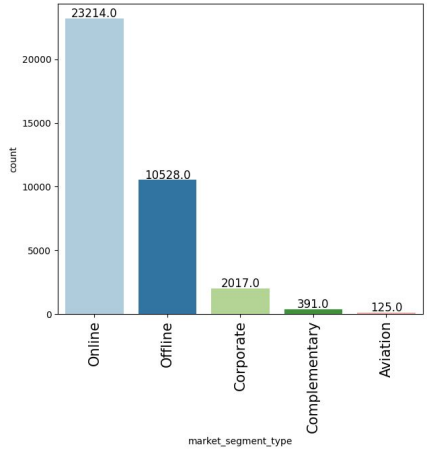
Most customers reserved room type 1.

Univariate Analysis - Arrival Month



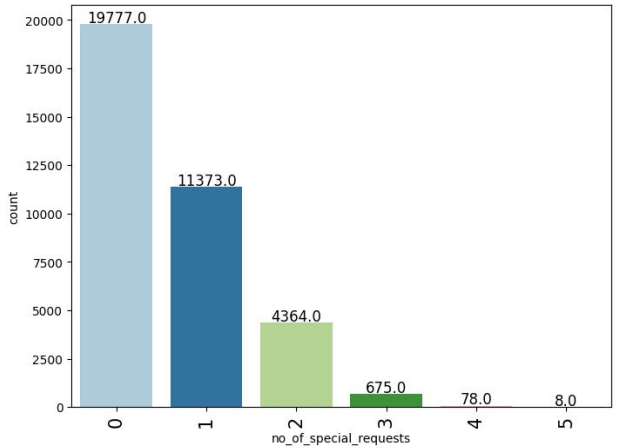
- The peak month for customer arrivals is October, with January having the fewest arrivals.

Univariate Analysis - Market Segment Type



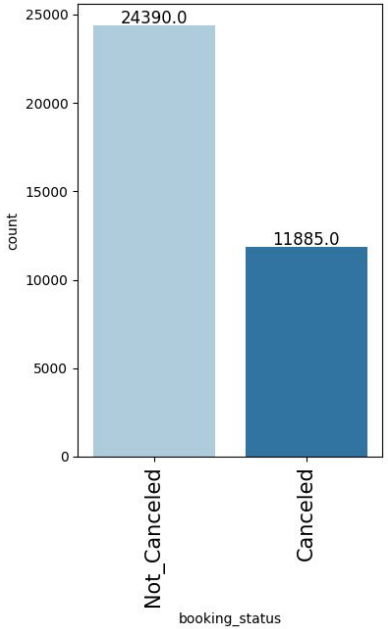
- Online channels are the most popular choice for reservations, whereas aviation channels are the least preferred.

Univariate Analysis - Number of Special Requests



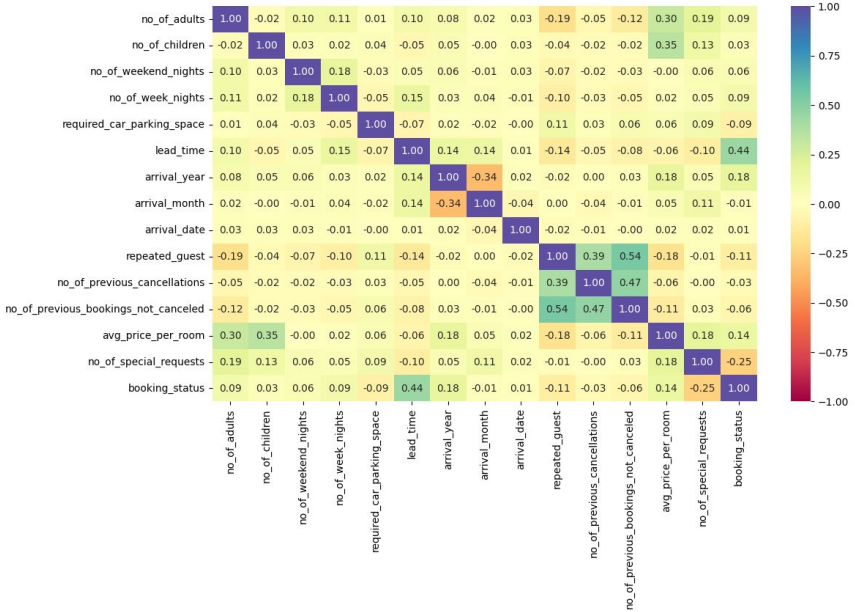
- The majority of customers do not make any special requests for their reservations.

Univariate Analysis - Booking Status



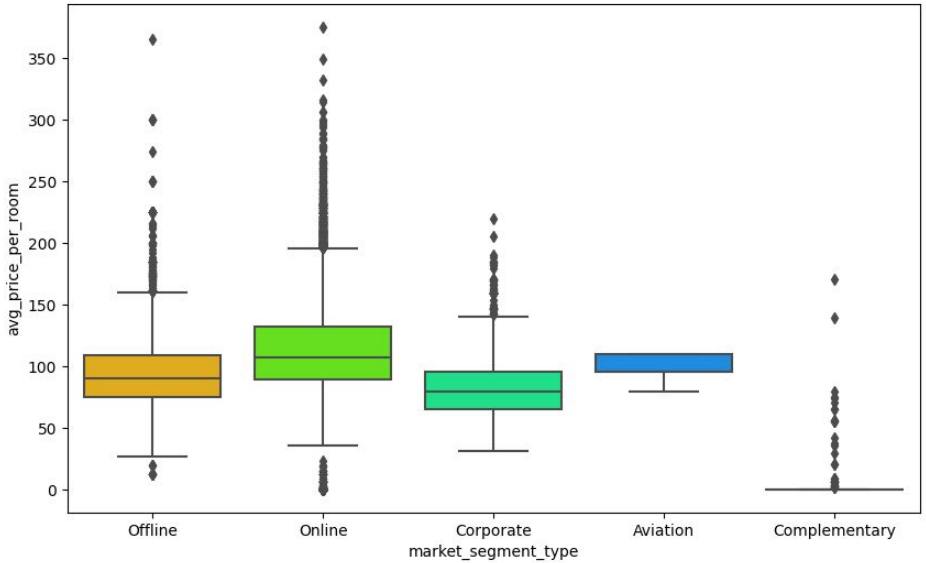
- The majority of customers do not cancel their reservations.

Bivariate Analysis



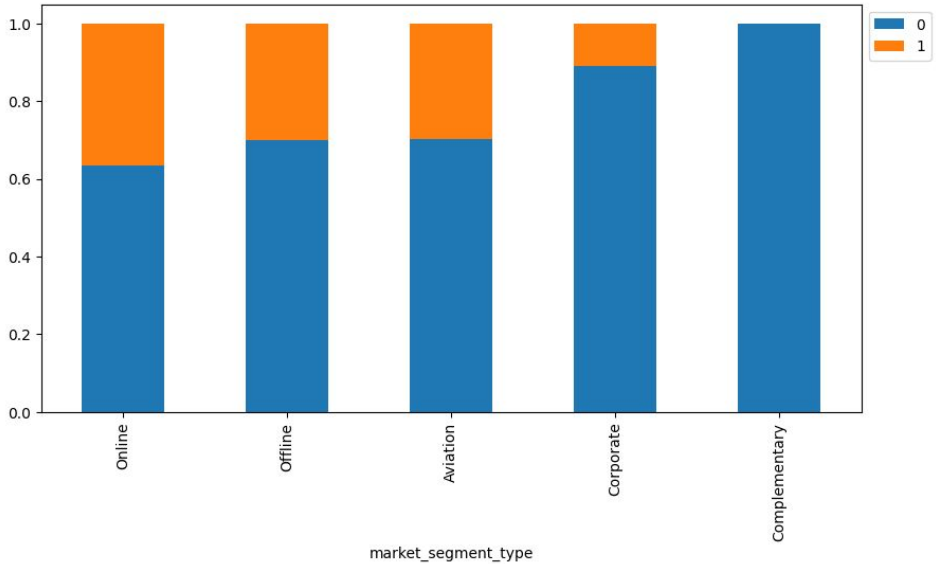
- There's a strong correlation between repeated guest, number of previous cancellations, and number of previous booking not cancelled
- The booking status shows strong correlation with lead time.
- There is a negative correlation between arrival year and arrival month.

Bivariate Analysis - Market Segment Type vs Avg Price Per Room



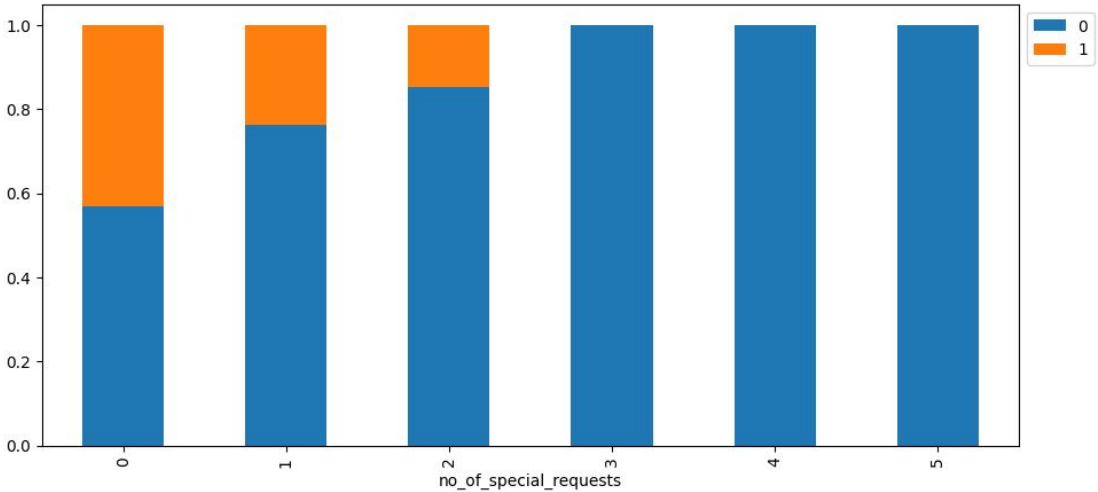
- Online is a more popular choice than complimentary to make a reservation.

Bivariate Analysis - Market Segment Type vs Booking Status



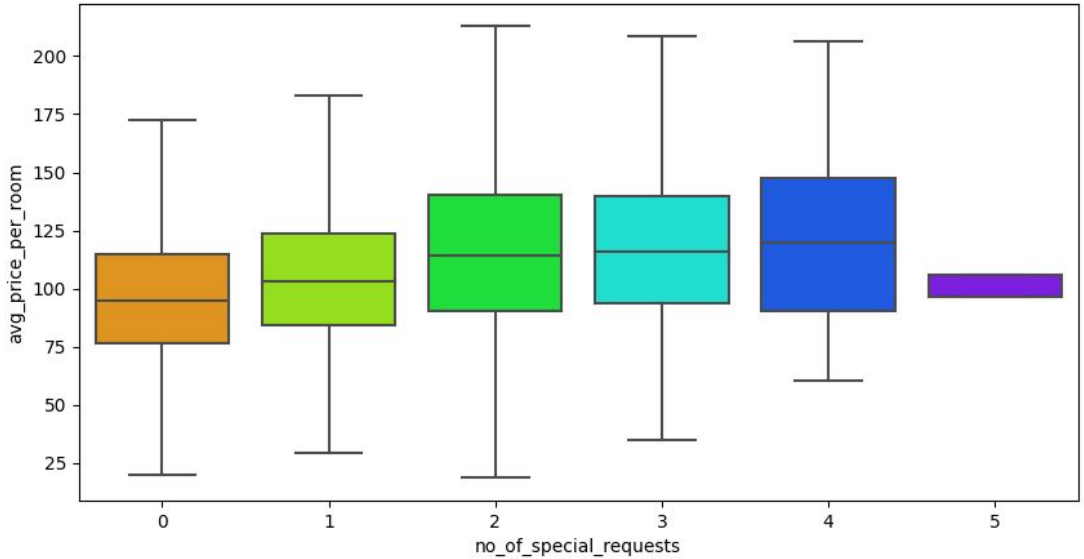
- Most customers prefer to book their reservations online rather than choosing a complimentary room.

Bivariate Analysis - No of Special Request vs Booking Status



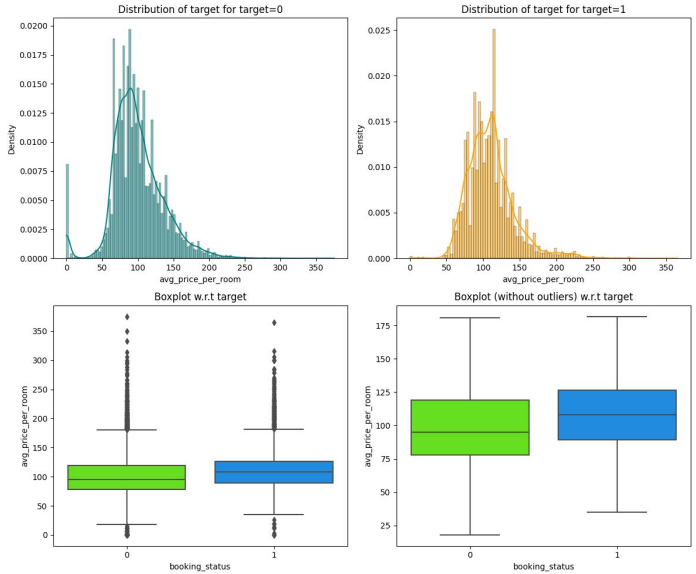
- The majority of customers do not make special requests.

Bivariate Analysis - No of Special Requests vs Avg Price per Room



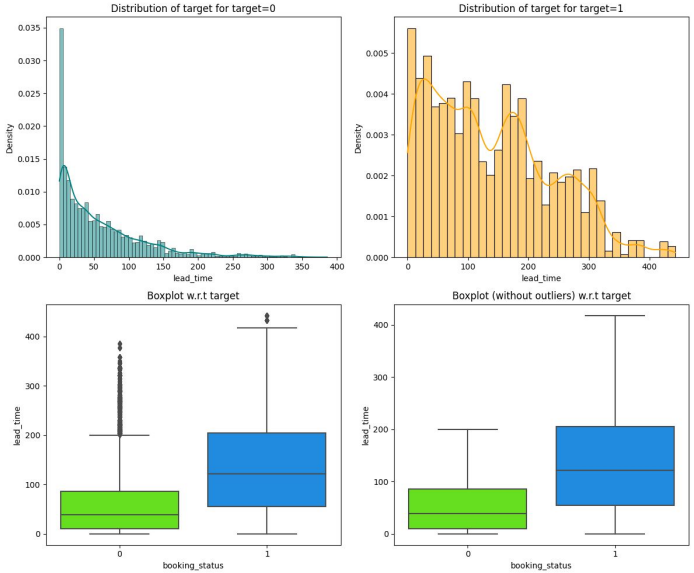
- The average price per room tends to increase with a higher number of special requests.

Bivariate Analysis - Avg Price Per Room vs Booking Status



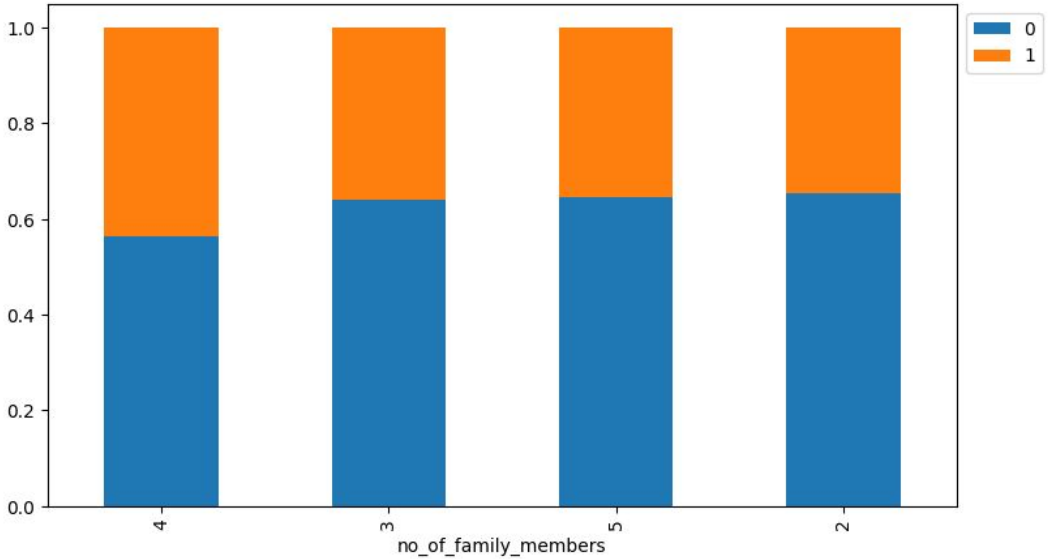
- Cancellations do not appear to have a significant impact on the average price per room.

Bivariate Analysis - Lead Time vs Booking Status



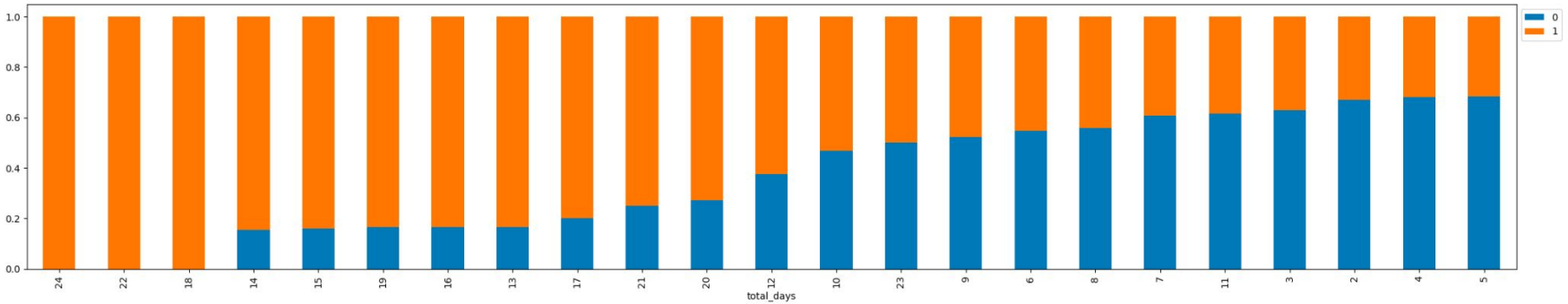
- There is a positive correlation between lead time and the likelihood of cancellations.

Bivariate Analysis - No of Family Members vs Booking Status



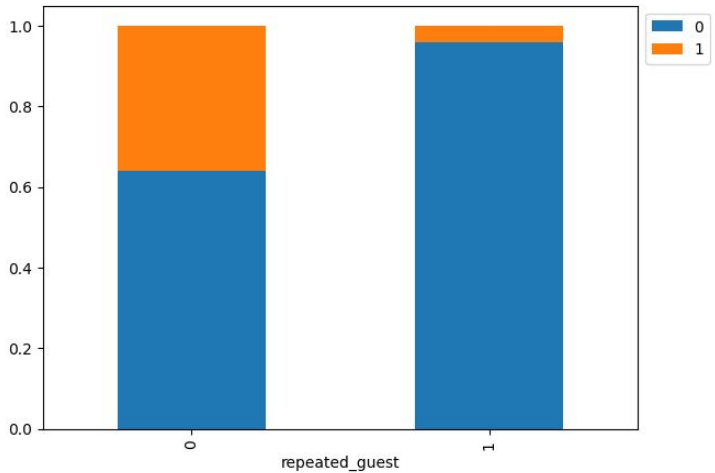
- Family size does not significantly influence the likelihood of cancellations.

Bivariate Analysis - Total Days vs Booking Status



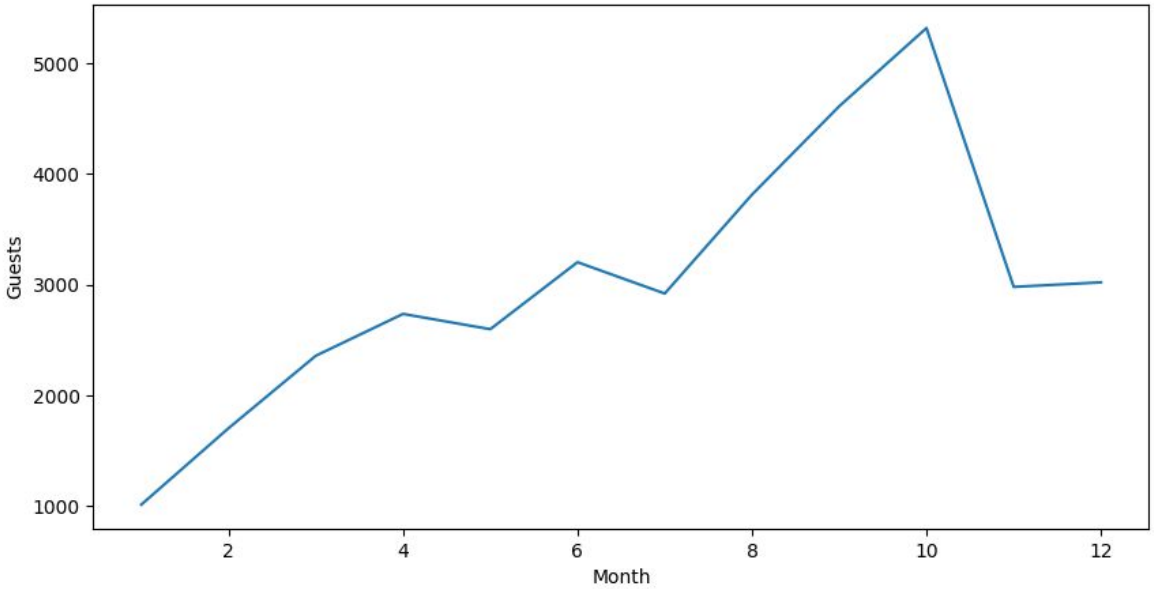
- An increase in the number of days leads to a higher rate of cancellations.

Bivariate Analysis - Repeated Guest vs Booking Status



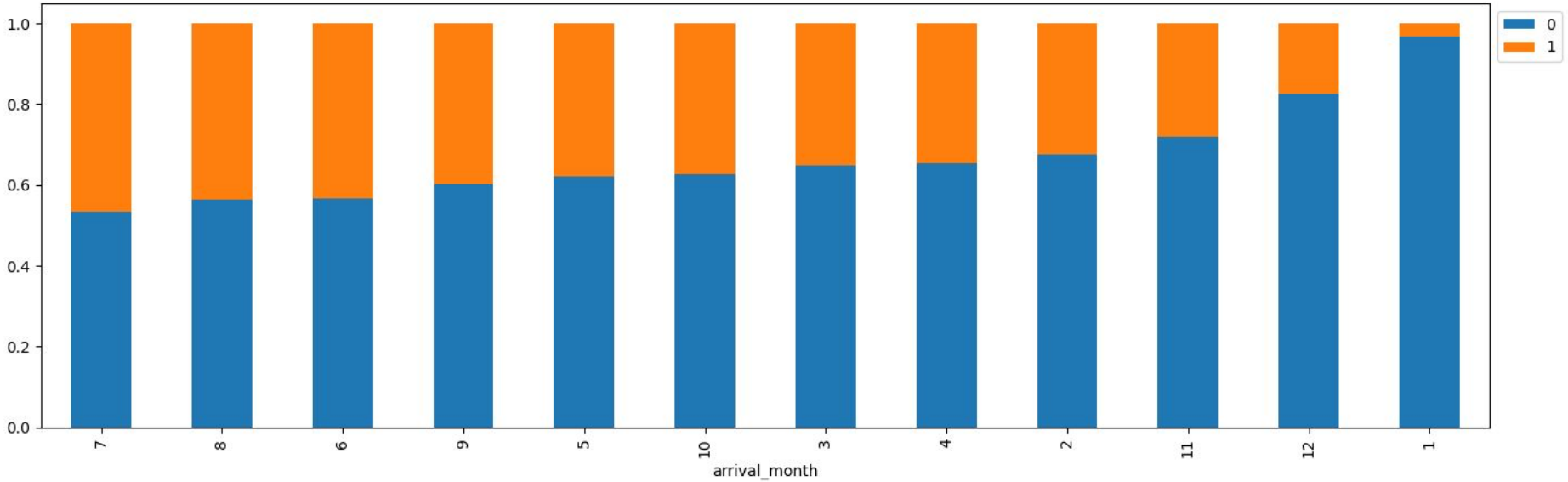
- Repeated guest are more likely to make a reservation.

Bivariate Analysis - Arrival Month vs Booking Status



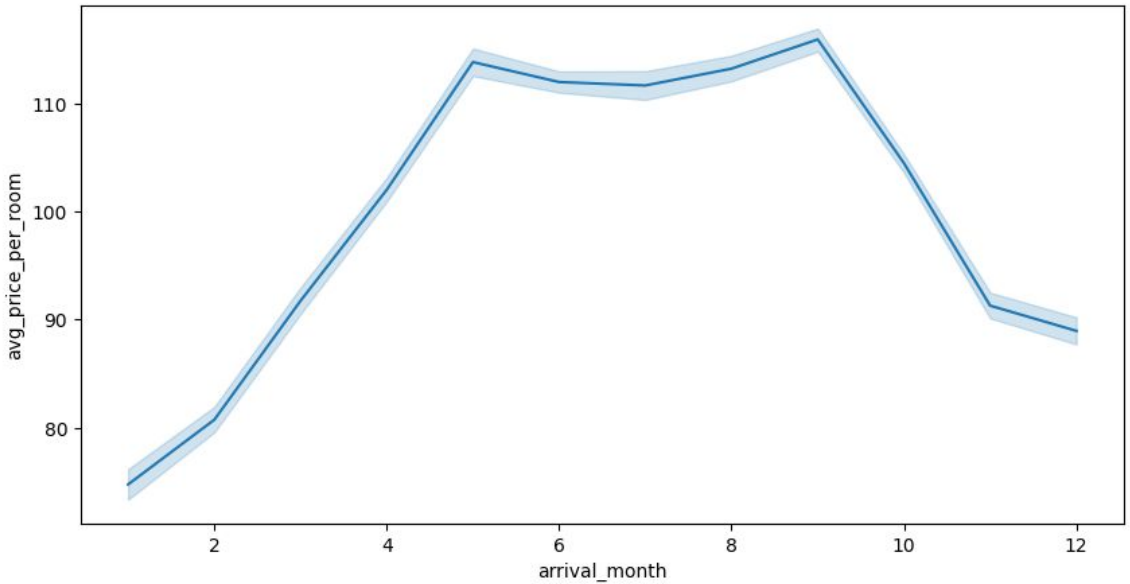
- The busiest month for arrivals is October.

Bivariate Analysis - Arrival Month vs Booking Status



- Cancellations tend to increase as the year progresses.

Bivariate Analysis - Arrival Month vs Avg Price Per Room



- Between the months of June and October have higher prices per room.

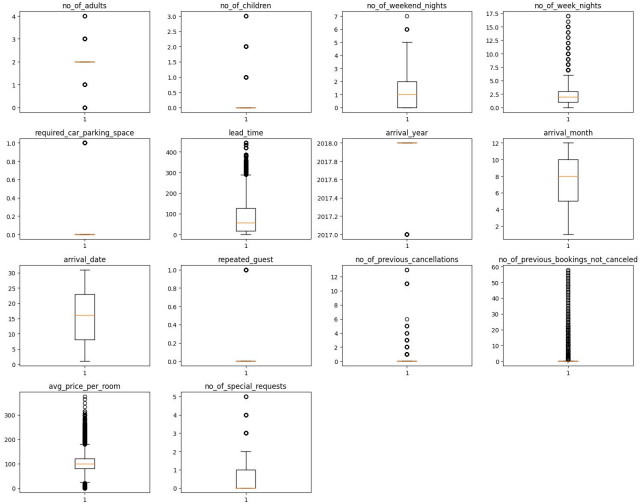
Data Preprocessing - Logistic Regression

Data Preprocessing

We want to predict which bookings will be cancelled. To achieve this:

- We'll encode categorical features.
- The data will be split into train and test sets for model evaluation.
- A Logistic Regression model will be built using the train data, and we'll assess its performance.

Data Preprocessing



- We've identified several outliers in the data, but no treatment will be applied.
- We will drop the booking status column.

Logistic Regression

Logistic Regression - Multicollinearity

We will use the VIF to test for multicollinearity.

- There were a few variables, market_segment_type_corporate, market_segment_type_offline, and market_segment_type_online that exhibit high multicollinearity with a VIF score > 5.

Logistic Regression - Training

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

We have built different models at thresholds of .37 and .42 to compare with the original model. There isn’t much of a change between the different thresholds.

Logistic Regression - Training

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:	01:36:39	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

- The negative coefficients, which include, required car parking space, arrival month, repeated guest, average price per room, and room type have little impact on cancellations.
- Positive coefficients number of adults and children, weekend and weeknights, average price per room have an impact.

Model Performance Summary

- Number of adults : Holding all other features constant a unit change in number of adults will increase the odds of a customer cancelling by 1.11 times or a 11.49% increase in odds.
- Number of weekend nights: Holding all other features constant a unit change in weekend nights will increase the odds of a customer cancelling by 1.11 times or a 16.461% increase in the odds.
- Arrival Month: The time of year affects a customer by 0.95 times or 4.16 decrease in the odds.
- Repeated Guest: The odds of a customer cancelling as a repeat guest is .06 or a decrease or 93.53 decrease in odds.

Decision Tree

Data Preprocessing

We want to predict which bookings will be cancelled. To achieve this:

- We'll encode categorical features.
- The data will be split into train and test sets for model evaluation.
- A Decision Tree model will be built using the train data, and we'll assess its performance. The steps are:
  - Calculate different metrics and confusion matrix
  - Check important features
  - Pre-pruning
  - Visualize the tree
  - Check performance on training and test set

Data Preparation for Modeling

We want to predict which bookings will be cancelled. To achieve this:

a. We'll encode categorical features.
b. The data will be split into train and test sets for model evaluation.
c. A Decision Tree Regression model will be built using the train data, and we'll assess its performance.

Pre Pruning

Checking Model Performance

Training

	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Testing

	Accuracy	Recall	Precision	F1
0	0.87182	0.80522	0.80000	0.80260

There is a little bit of a discrepancy in the performance of the model on the training and testing set, suggesting the model is overfitting,

Checking Performance

Training

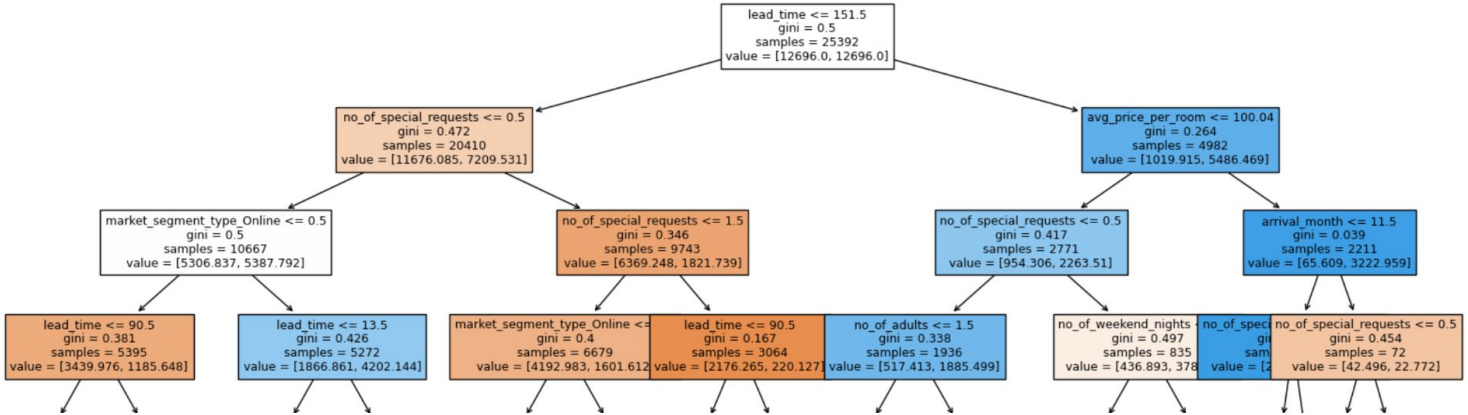
	Accuracy	Recall	Precision	F1
0	0.83109	0.78608	0.72449	0.75403

Testing

	Accuracy	Recall	Precision	F1
0	0.83488	0.78308	0.72751	0.75427

The model shows good results on the training and test set once the pruning process was started.

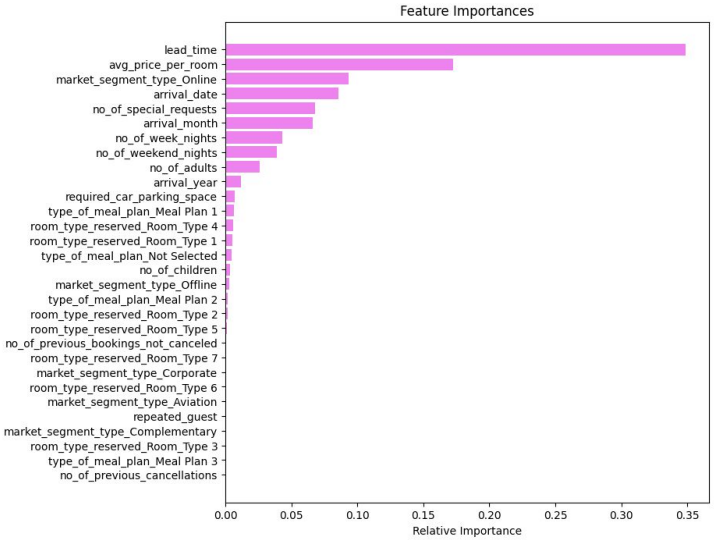
Model Building - Decision Tree



- If the lead time is less than or equal to 151.50, the number of special request is less than or equal to 0.5, the market segment type is less than or equal to 0.5 and the lead time is greater than 90.5, then the customer will cancel.

Post Pruning

Important Features



In the pre tuned decision tree, the top three important features are lead time, average price per room, and market segment type.

Checking Performance - Best Model

Training

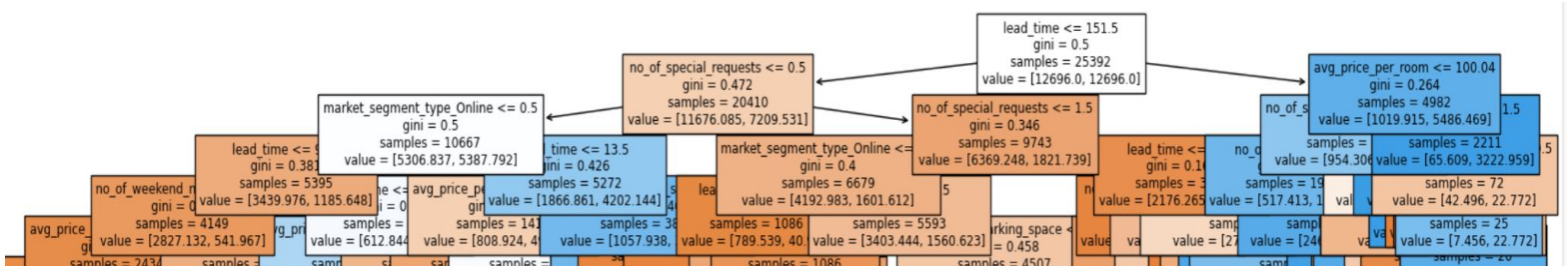
	Accuracy	Recall	Precision	F1
0	0.89414	0.89669	0.80435	0.84802

Testing

	Accuracy	Recall	Precision	F1
0	0.86750	0.85406	0.76423	0.80665

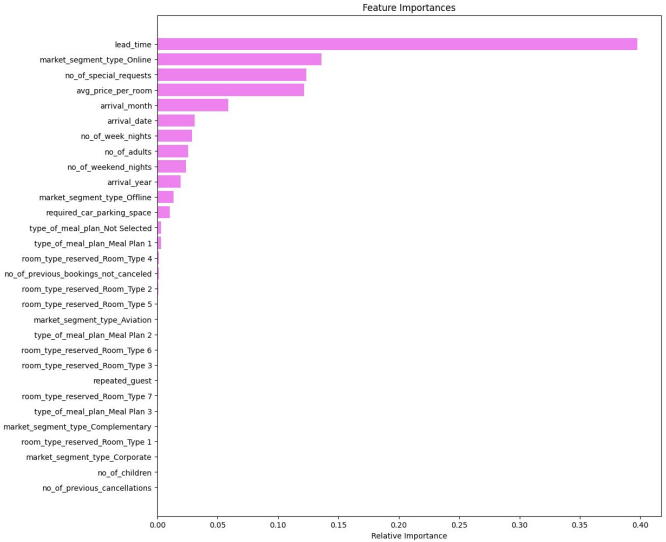
- The Decision Tree best model achieved high accuracy and recall on the training set (89.41% and 89.67%, respectively).
- On the test set, the model maintained good performance with an accuracy of 86.75% and a recall of 85.41%.
- Precision on the test set was 76.42%, and the F1 score was 80.67%, indicating a balanced performance between precision and recall.

Model Building - Post Pruning



- The observations we got from the pre-pruned tree is close with the decision tree rules of the post pruned tree.
- If the lead time is less than or equal to 151.50, the number of special request is less than or equal to 0.5, the market segment type is less than or equal to 0.5 and the lead time is greater than 90.5, then the customer will cancel.

Important Features - Post Pruning



- Lead Time, Market Segment Type, and Number of special requests are the most important features for the post pruned tree.

Comparing Decision Tree Models

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83109	0.89414
Recall	0.98661	0.78608	0.89669
Precision	0.99578	0.72449	0.80435
F1	0.99117	0.75403	0.84802

Testing performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87182	0.83109	0.86750
Recall	0.80522	0.78608	0.85406
Precision	0.80000	0.72449	0.76423
F1	0.80260	0.75403	0.80665

- Decision tree models with pre-pruning and post-pruning both are giving equally high recall scores on both training and test sets.

Executive Summary

Executive Summary - Conclusions

- The logistic regression model demonstrates varying performance with different thresholds, influencing accuracy, recall, precision, and F1 score.
- Key predictors for cancellations include the number of adults, weekend nights, and arrival month, as well as whether the guest is a repeated guest.
- Decision tree models exhibit strong accuracy and recall, emphasizing the importance of pre- and post-pruning for generalization.
- Recommendations focus on dynamic pricing, targeted marketing, loyalty programs, and operational improvements for enhanced customer satisfaction and reduced cancellations.
- Continuous monitoring of market dynamics and customer behaviors is crucial for adapting strategies and maintaining competitiveness.

Executive Summary - Recommendations

- Implement dynamic pricing strategies based on observed predictors, optimizing rates for specific customer segments and periods.
- Introduce targeted marketing campaigns to incentivize repeat bookings, focusing on customer loyalty programs and personalized offers.
- Enhance operational efficiency to reduce lead time and increase customer satisfaction, potentially lowering cancellation rates.
- Periodically review and adjust cancellation policies to align with market trends and customer preferences.
- Leverage customer data for proactive communication, offering relevant incentives or alternatives to potential cancellations.
- Invest in continuous monitoring and analysis of market dynamics, enabling timely adjustments to pricing and marketing strategies.
- Explore collaborations with online platforms and agencies to expand reach and diversify customer acquisition channels.

These recommendations are designed to guide INN Hotels Group toward enhanced customer satisfaction, revenue growth, and operational effectiveness.