Star Hotels

Travis Bruns

Table of Contents

- Core business idea
- Problem to tackle
- Financial implications
- How to use ML model to solve the problem

Business Problem Overview

A significant number of hotel bookings are called-off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impact a hotel on various fronts:

- Loss of resources (revenue) when the hotel cannot resell the room.
- Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
- Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
- Human capital to make arrangements for the guests.

Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. Star Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. The objective is to analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

Data Overview

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

- 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)
- 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 Star 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.
- Import initial Libraries

Raw Data Summary

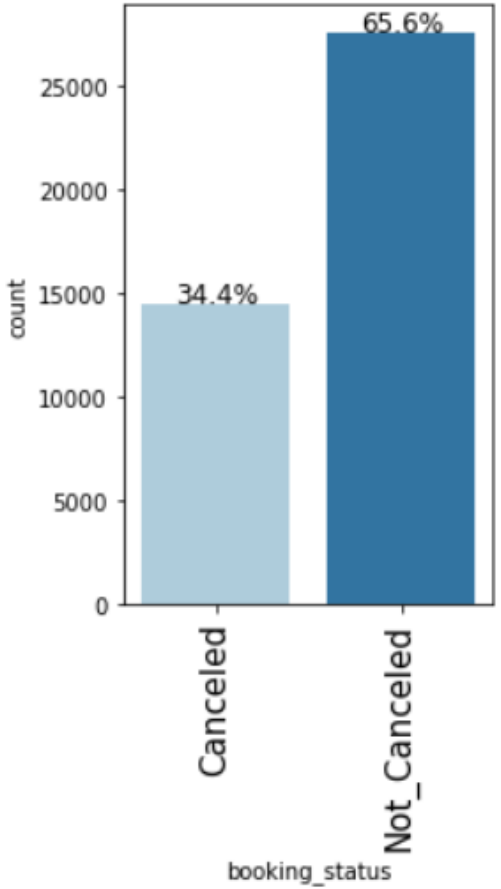
- The Source Data includes 56,926 rows and 18 columns.
- The Source Data included 14,350 duplicates. These duplicates were removed to reduce the rows to 42,576.
- The Data types are 1 float, 13 integers, and 4 objects.
    - Float: Average Price per Room
    - Integer: no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, required_car_parking_space, lead_time, arrival_year, arrival_month, arrival_date, repeated_gues, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, and no_of_special_requests.
    - Object: type_of_meal_plan, room_type_reserved, market_segment_type, and booking_status
- The Data does not include any missing values.
- The Dependent Variable is Booking Status.

Data Summary

- Number of Adults: Average adults is 1.917 ranging from 0 to 4.
- Number of Children: Average children is .142 per visit ranging from 0 to 10.
- Number of weekend nights: Average of .895 weekend nights per visit ranging from 0 to 8.
- Number of Week Nights: Average of 2.321 week nights per visit ranging from 0 to 17.
- Required Car Parking Space: Average of .034 vistors require a parking space.
- Lead Time: Average lead time is 77.316 days ranging from 0 to 521.
- Arrival Year: Average year of arrival for the data is 2018 ranging from 2017 to 2019.
- Arrival Month: Average arrival month is June ranging between January and December.
- Arrival Date: Average arrival date is the 16th ranging from the 1st to the 31st.
- Repeat Guest: A guest is a repeat visitor approximately 3.1% of the time.
- Number of previous Cancellations: Number of previous cancellations averages .025 per booking ranging from 0 to 13 instances.
- Number of previous bookings not canceled: Number of previous bookings not canceled averages .223 per booking ranging from 0 to 72 instances.
- Average Price per Room: Average price per room is 112.38rangingfrom112.38rangingfrom85 to $540.
- No of Special Requests: Special requests average .768 of visits ranging from 0 to 5 special requests.

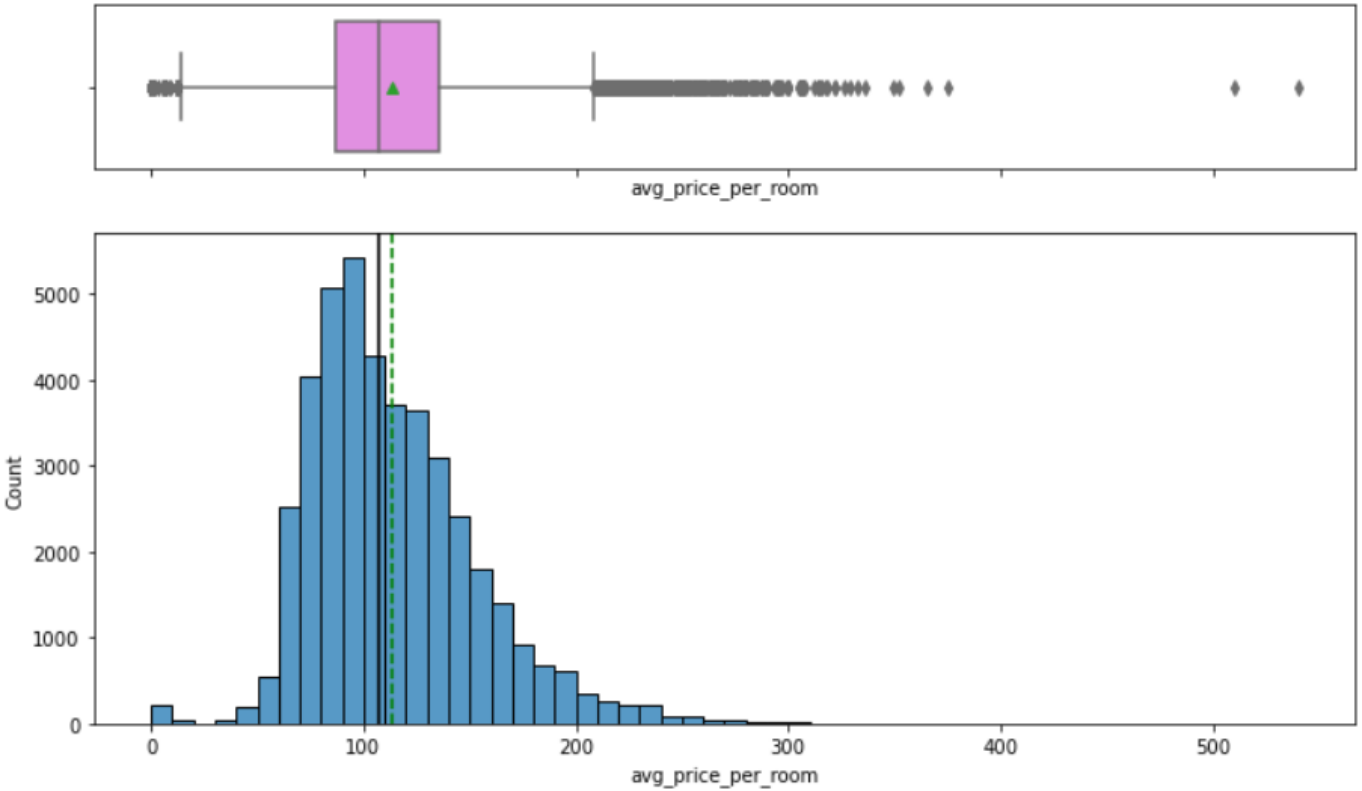
Categorical Data Summary

- Meal Plan 1 is the most common meal plan.
- Room Type 1 is the most requested room.
- Online booking is the most common market segment.
- Cancellations occur 21,548/56,926 or 37.85% of the time.
- Complementary bookings do not lead to cancellations and can be removed from the analysis. This reduces the data population to 42,080.
- Once the data is preprocessed, the cancellation rate for the data being analyzed is 14,487/42,080 or 34.4%



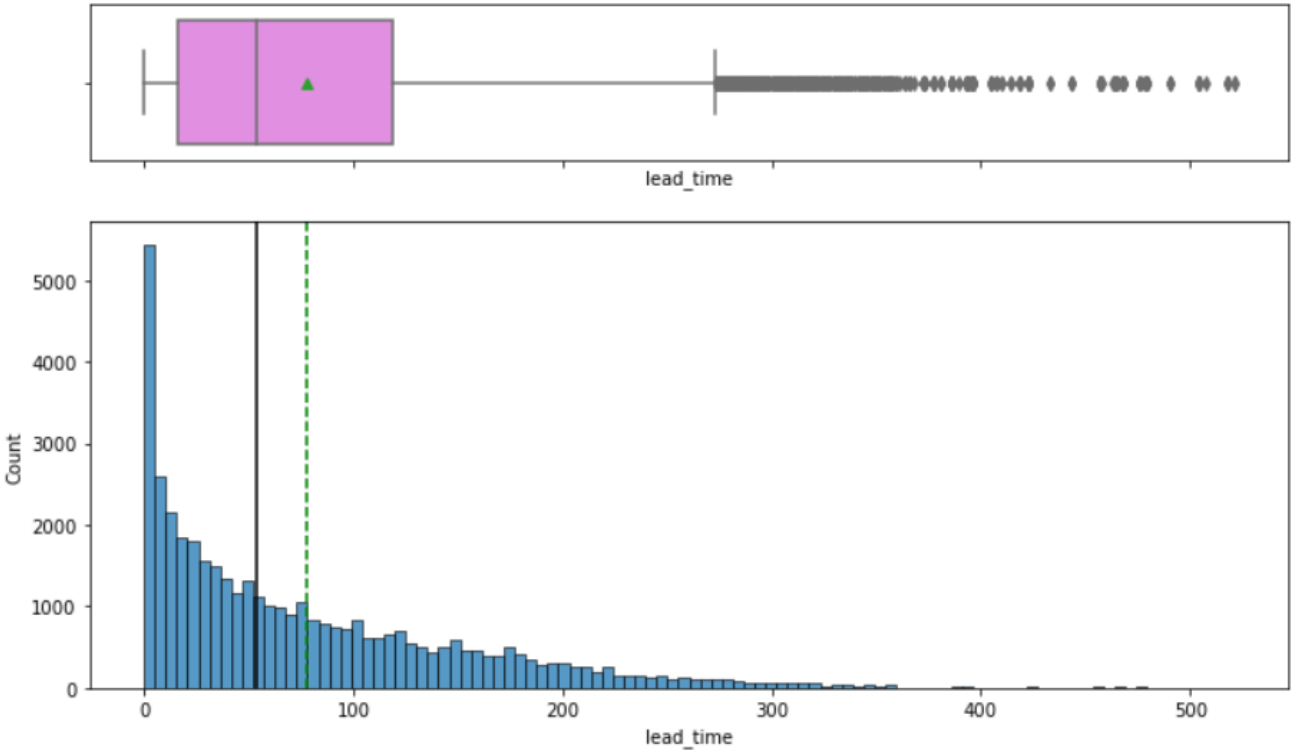
Univariate Data Analysis – Booking Lead Times

- The data is right skewed with 75% of bookings occurring with 300 days.



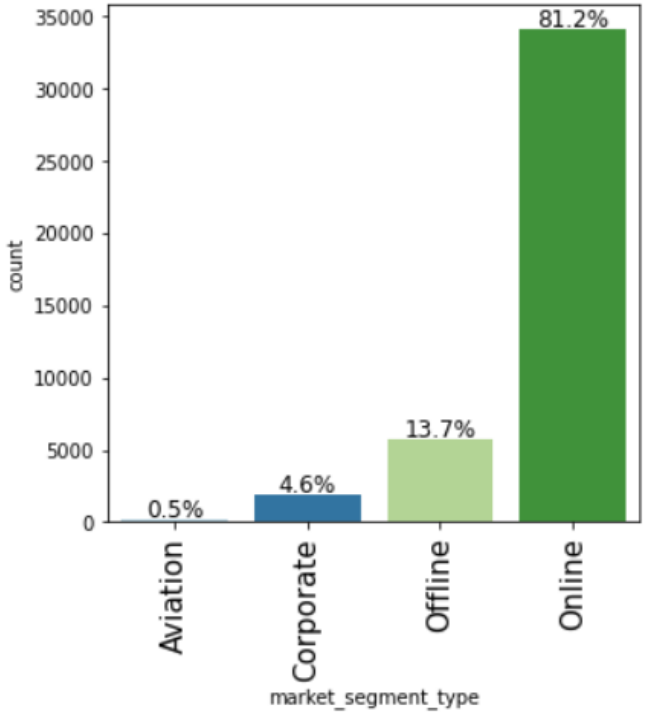
Univariate Data Analysis – Booking Lead Times

- The data is right skewed with 75% of rooms costing less than $210.



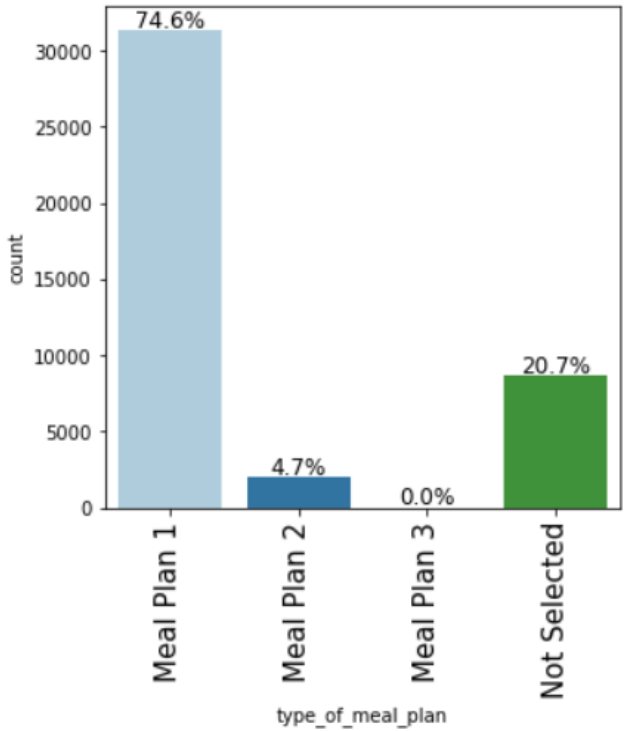
Univariate Data Analysis – Market Segments

- Online and Offline bookings account for most guest acquisition.



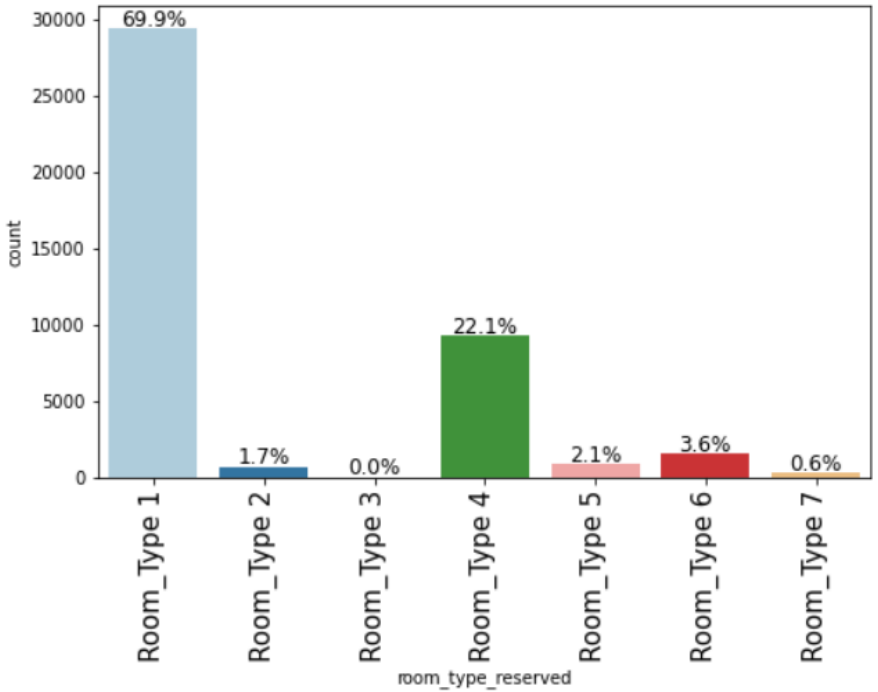
Univariate Data Analysis – Meal Plans

- Meal Plan 1 Accounts for the Majority of Meal Plan selections.



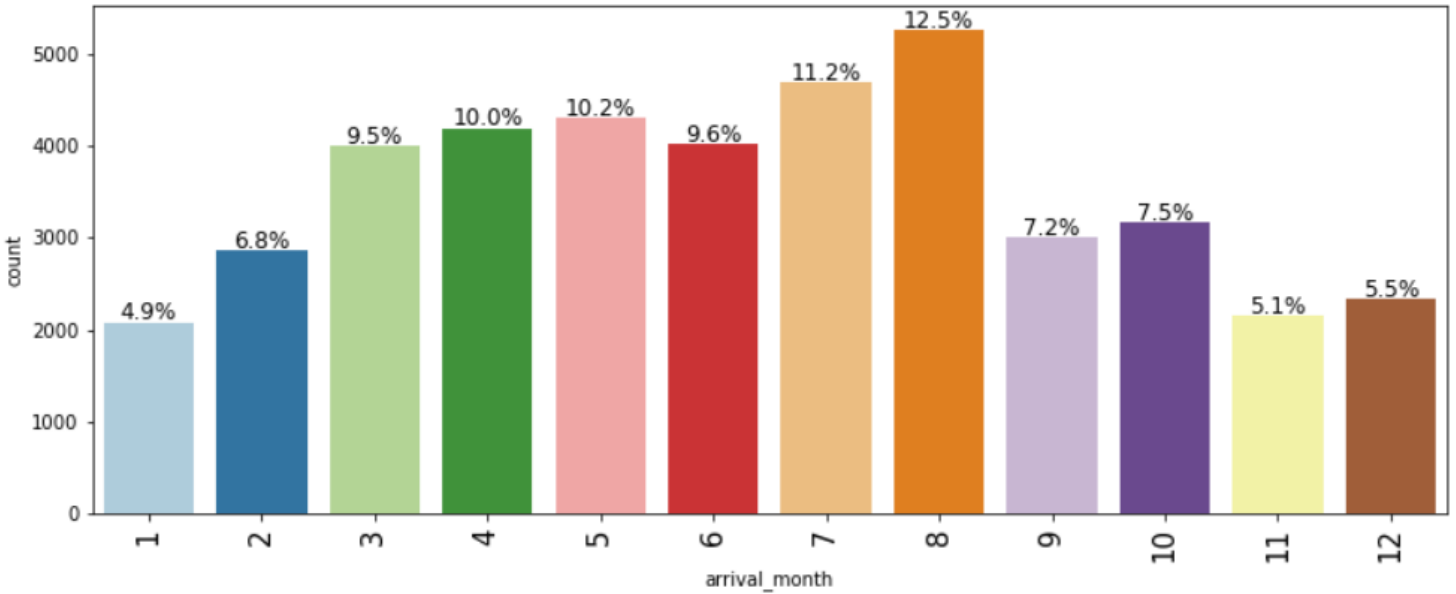
Univariate Data Analysis – Room Types

- Room type 1 and Room type 2 are the most favored room types.



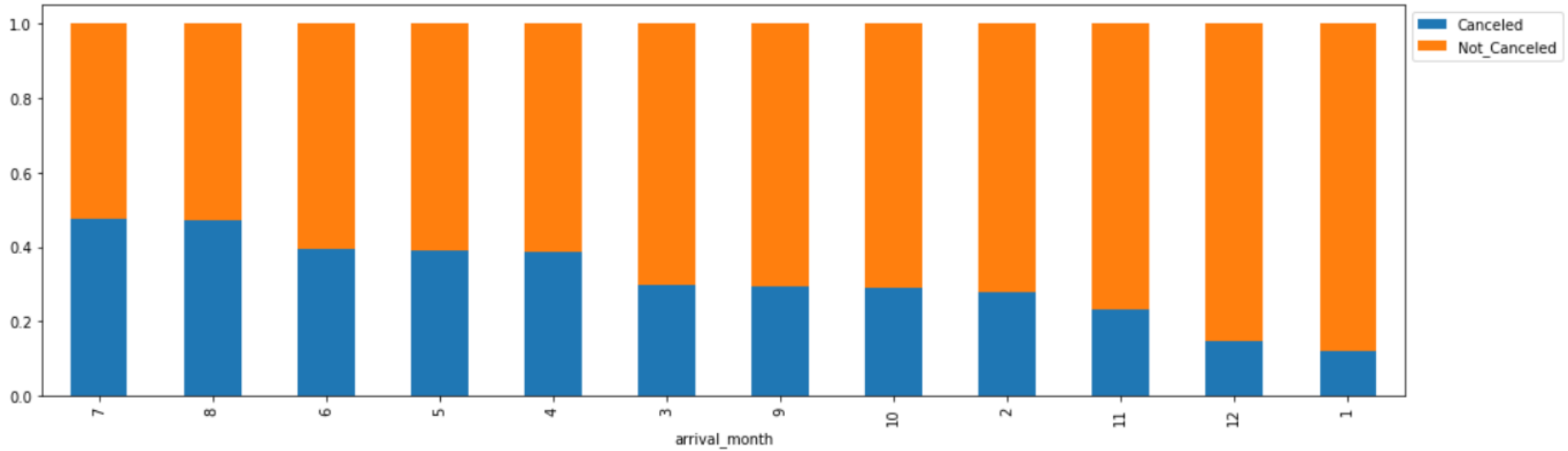
Univariate Data Analysis – Monthly Bookings

- The summer months are popular to be booked with August being the most popular booking month.



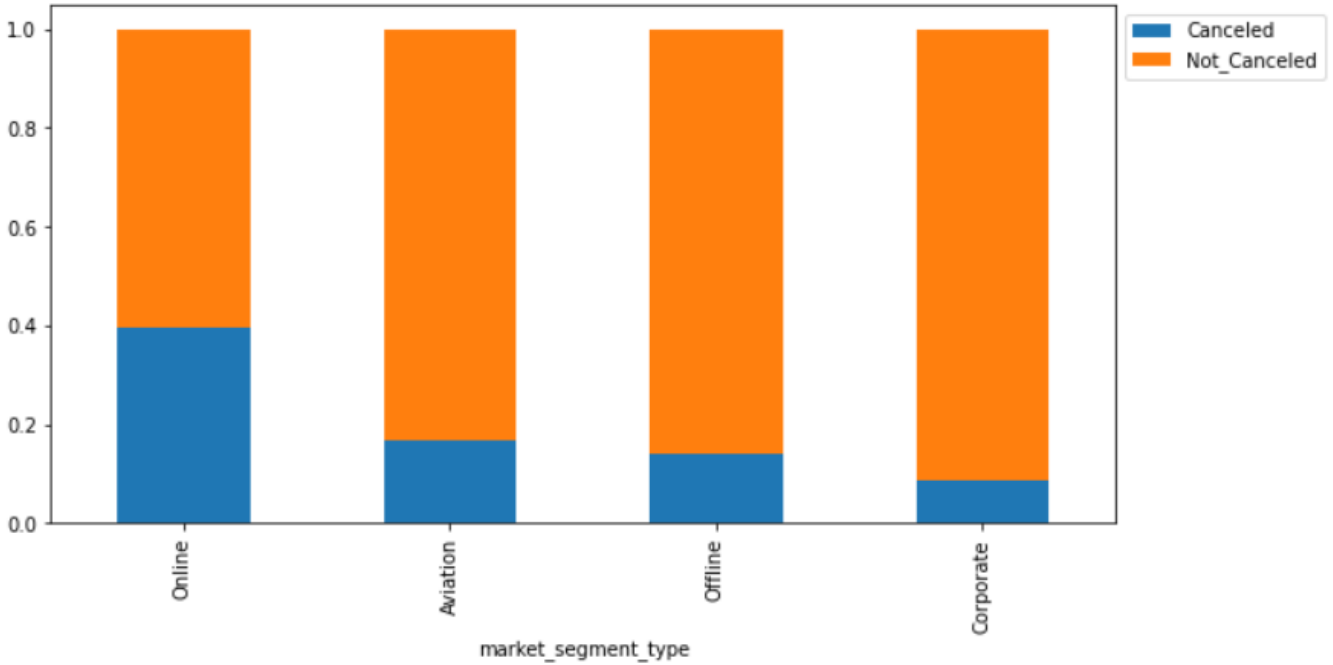
Bivariate Data Analysis – Cancellations per Month

- July bookings are canceled roughly half the time.
- Summer Months lead to more cancellations than winter months.



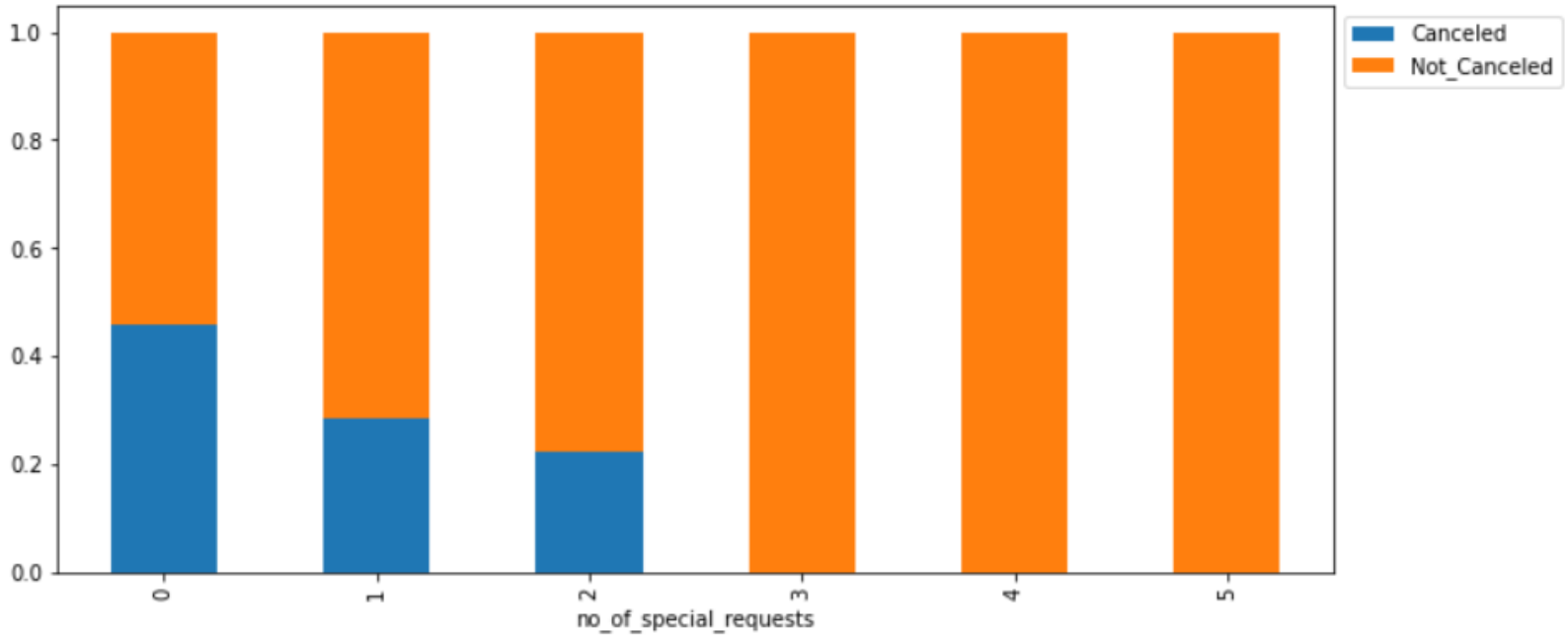
Bivariate Data Analysis – Cancellations per Market Segment

- Corporate employees cancel the least amount of time.
- Online bookings are canceled ~40% of the time.



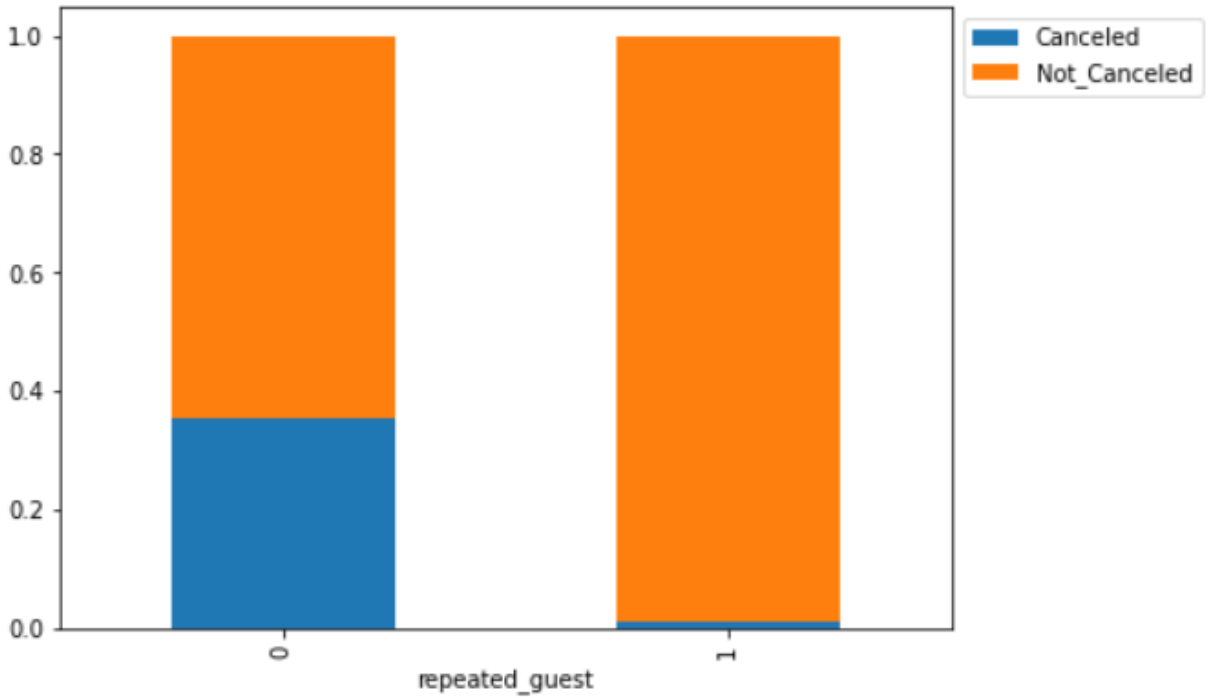
Bivariate Data Analysis – Cancellations per Special Requests

- Individuals with 3 or more special requests do not cancel bookings.
- Roughly 50% of cancellations are from guests without special requests.



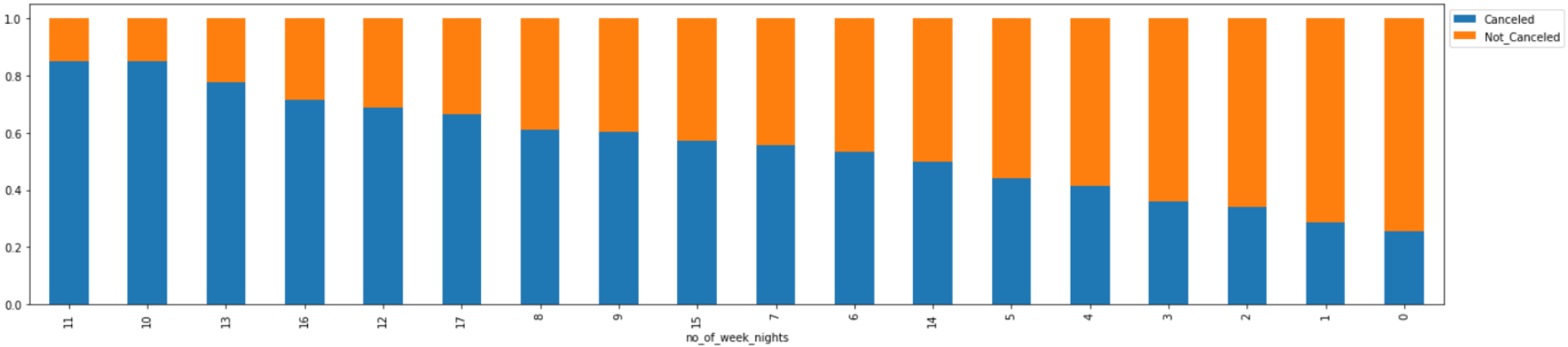
Bivariate Data Analysis – Cancellations per Repeat Guests

- Repeat Guests rarely cancel.



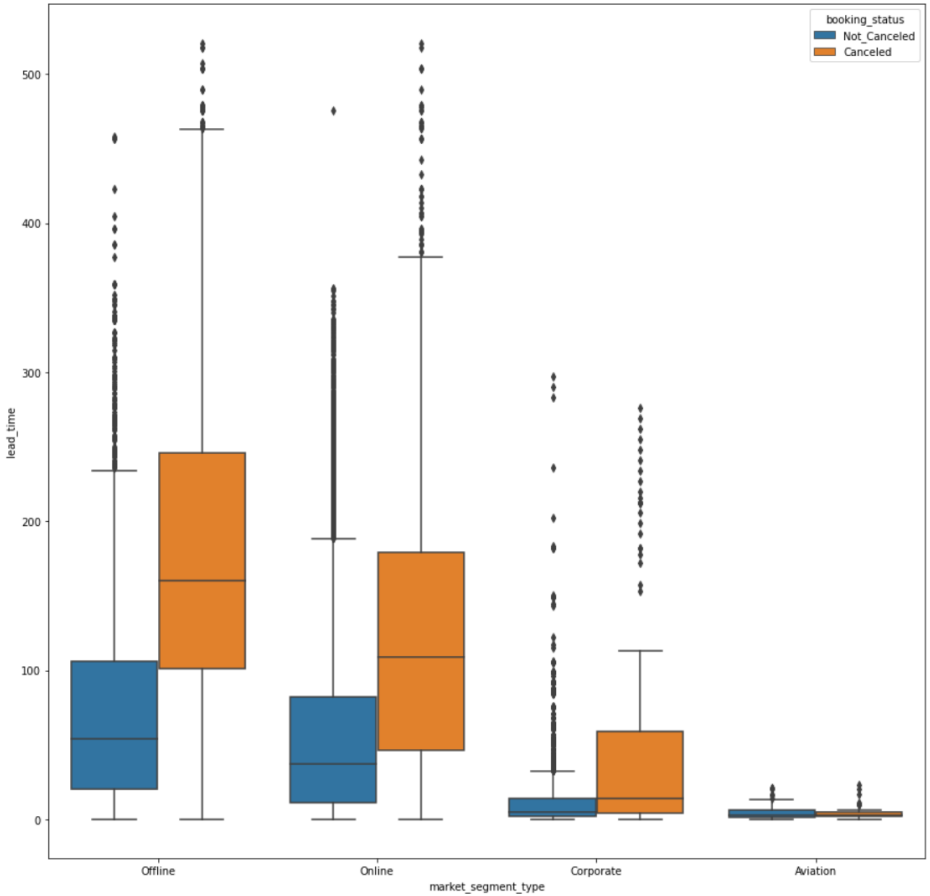
Bivariate Data Analysis – Cancellations per Month

- Visits for weekend nights and week nights greater than 5 days appear to lead to a greater than 50% cancellation rate.
- Longer scheduled bookings increase the risk of cancellations.



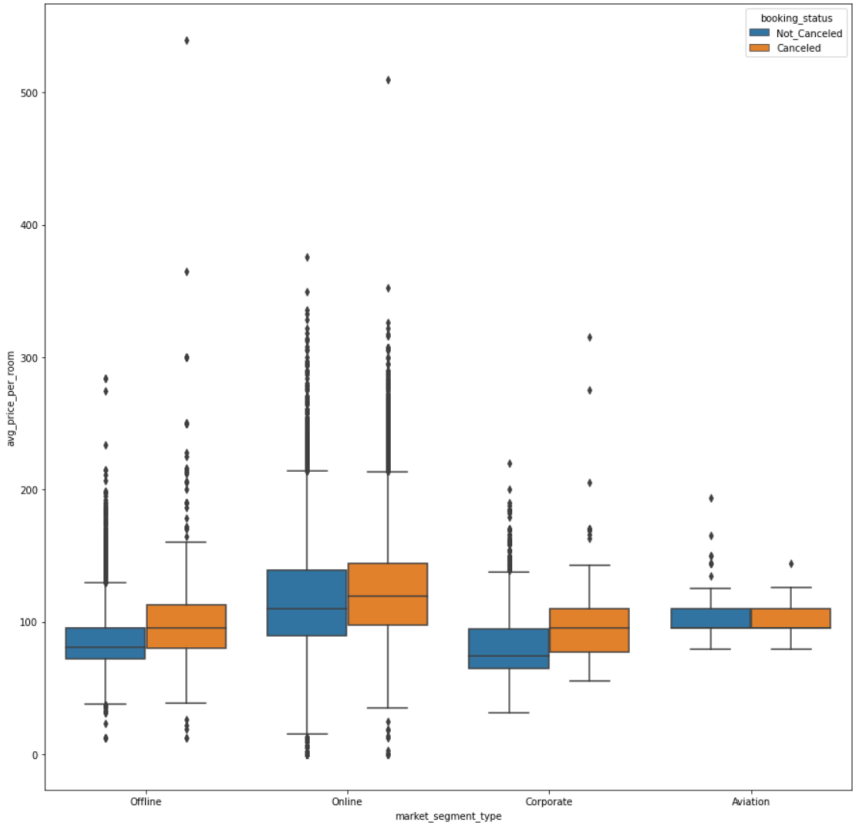
Multi-Variate Data Analysis – Market Segment Lead Times

- Longer lead times appear to increase cancellation rates.
- Offline bookings often have the longest lead times.
- Aviation bookings have the shortest lead times.



Multi-Variate Data Analysis – Market Segment Avg Price

- Online bookings pay the most.
- Higher price per booking appears to lead to more cancellations.
- Online bookings have the greatest range in pricing.
- Aviation pricing appears to be consistent for cancels and not cancels.



Logistic Regression Model Performance Summary

Building the Model

Model evaluation criterion

Model can make wrong predictions as:

- Predicting a person will cancel but in reality they do not cancel.
- Predicting a person will not cancel but in reality they do cancel.

Which case is more important?

- Both the cases are important as:
- If we predict a person will cancel but actually will not cancel then that person will not be able to book a room constituting an opportunity loss.
- If we predict a person will not cancel but actually will cancel then the hotel might not be able to fill the vacancy resulting in a loss due to vacancy.

Logistic Regression Model Performance Summary

Observations

- The training and testing f1_scores are 0.66 and 0.67, respectably.
- f1_score on the train and test sets are comparable.
- This shows that the model is showing generalized results.
- We have build a logistic regression model which shows good performance on the train and test sets but to identify significant variables we will have to build a logistic regression model using the statsmodels library.
- We will now perform logistic regression using statsmodels, a Python module that provides functions for the estimation of many statistical models, as well as for conducting statistical tests, and statistical data exploration.
- Using statsmodels, we will be able to check the statistical validity of our model - identify the significant predictors from p-values that we get for each predictor variable and remove variables with p-value > .05 from the model.

Logistic Regression Model Performance Summary

Observations

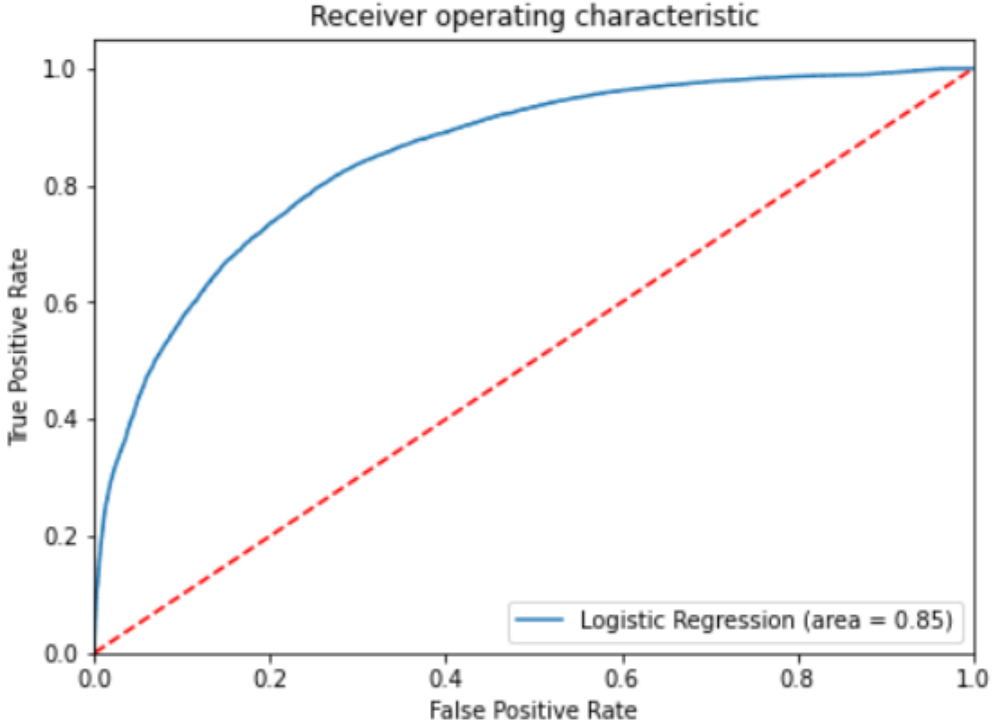
- The training and testing f1_scores are 0.66 and 0.67, respectably.
- f1_score on the train and test sets are comparable.
- This shows that the model is showing generalized results.
- No variables were identified as high multi-collinearity.
- No changes to the data are necessary due to multicollinearity.
- High P-Values were identified and removed from the model. These variables are number of children, number of weekend nights, and arrival date.

Coefficient Interpretations

- Coefficient of number of weeknights, types of meal plans, lead time, and average price per room are positive. Due to the inversion the assigned values for cancellations, increases to these coefficients will increase the risk of cancellation.
- Coefficients of number of adults, required car parking space, arrival year, arrival month, market segment type, repeat guest, and number of special requests are negative so increase in these will lead to decrease in chances of a person canceling their booking.

Logistic Regression Model Performance Summary

- Logistic Regression model is giving a good performance on training set based on the Receiver Operating Characteristic.



Logistic Model Summary

- We have been able to build a predictive model that can be used by the hotel to find the bookings expecting to be cancelled with an f1_score of 0.66 on the training set and formulate policies accordingly.
- All the logistic regression models have given a generalized performance on the training and test set.
- Coefficient of number of weeknights, types of meal plans, lead time, and average price per room are positive an increase in these will lead to increase in chances of a person canceling.
- Coefficient of number of adults, required car parking space, arrival year, arrival month, market segment type, repeat guest, and number of special requests are negative. Increases in these will lead to decreases in cancellations.

Decision Tree Model Performance Summary

Model evaluation criterion

Model can make wrong predictions as:

- Predicting a customer will not cancel but in reality the customer cancels.
- Predicting a customer will cancel but in reality the customer would not cancel.

Which case is more important?

Both are important. If we predict a booking will cancel, but the customer would not cancel in reality an opportunity loss occurs due to ignoring a customer profile. If we predict a booking will not cancel, but the customer does cancel, the hotel may not be able to find another booking to replace the expected revenue.

Decision Tree Model Performance Summary

Model evaluation criterion

Model can make wrong predictions as:

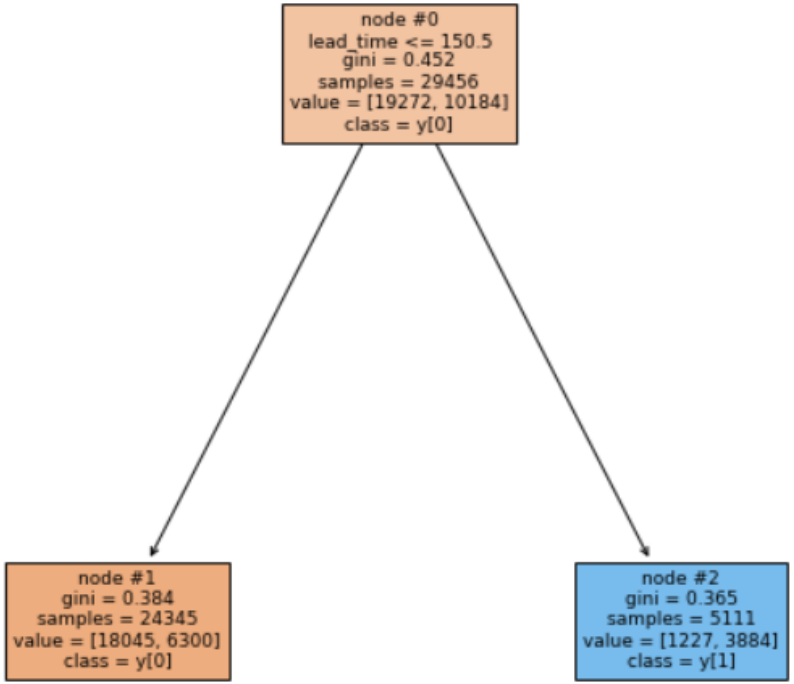
- Predicting a customer will not cancel but in reality the customer cancels.
- Predicting a customer will cancel but in reality the customer would not cancel.

Which case is more important?

Both are important. If we predict a booking will cancel, but the customer would not cancel in reality an opportunity loss occurs due to ignoring a customer profile. If we predict a booking will not cancel, but the customer does cancel, the hotel may not be able to find another booking to replace the expected revenue.

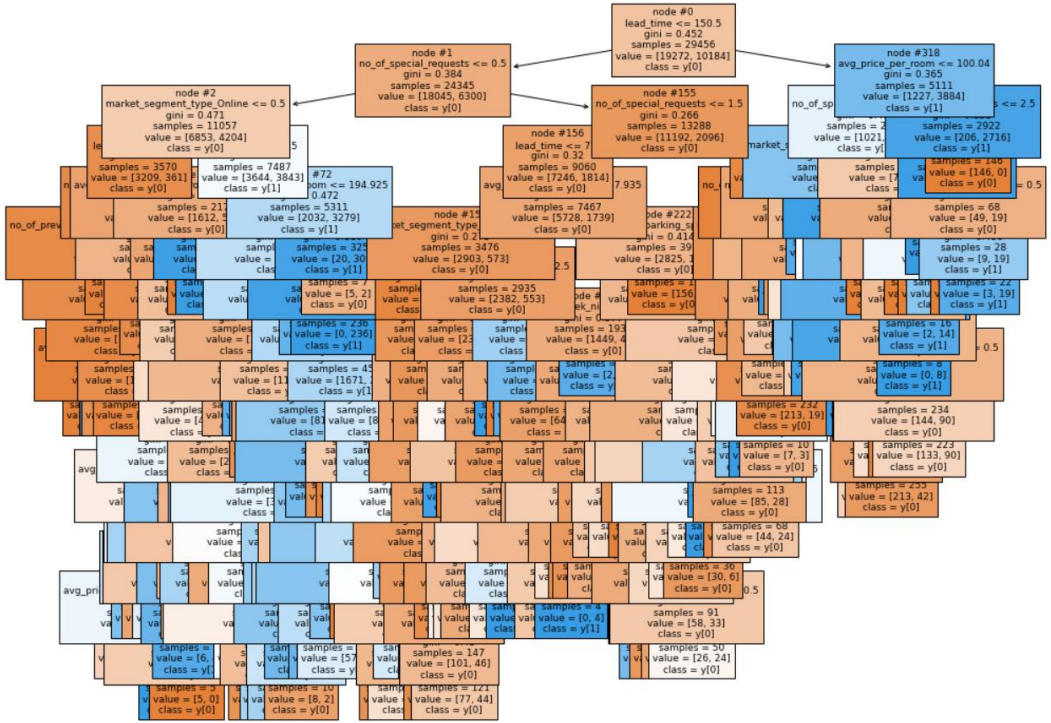
Decision Tree Visual – Best Model using Post Pruning

- Lead time is the most important feature to anticipate cancellations.



Decision Tree Visual – Pre - Pruning

- Pre-Pruning leads to a more complex but more precise model



Decision Tree Comparison

- To Avoid False Negatives, the Post-Pruning Model should be used.

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.996062	0.849912	0.744466
Recall	1.000000	0.920039	0.936333
Precision	0.994017	0.860269	0.741220
F1	0.996999	0.889151	0.827430

Business Insights

- The lead time of bookings is the greatest influence on cancellations.
- Average price per room, arrival dates, number of special requests, online market segment, arrival month, and number of nights stayed are importance for predicting cancellations.
- Previous cancellations do not imply causation of future cancellations.
- Online bookings are the most commonly canceled market segment.
- A customer requiring a parking space is less likely to cancel a booking.
- Repeat guests are less likely to cancel their bookings.
- More special requests will decrease the chances of cancellations.
- Longer stays are predictors for cancellations.

Business Recommendations

- Bookings should not be allowed for more than a year in advance of a visit.
- More expensive room bookings should lead to the concierge contacting the guests to ask for special requests.
- Summer months may need to reduce the allowable lead time before visits.
- Online bookings should not be allowed for stays greater than seven nights.
- Incentives should be created for repeat guests.
- Meal plan 3 should be discontinued.

greatlearning
Power Ahead

Happy Learning !