Roman Dzadzic

01/15/2022

Presentation to The Data Science Lead

INN Hotels

Contents

We will cover the following topics

- Business Problem Overview
- Data Overview
- Exploratory Data Analysis (EDA)
- About the Model
- Business Insights and Recommendations:
- Appendix: various data points

Business Problem Overview

- The INN Hotels Group has a chain of hotels in Portugal and is experiencing a significant number of cancellations or no-shows, and these impact the hotel in various ways. For example, there is a loss of revenue or resources when the hotel room cannot be resold. Canceled rooms present additional cost of distribution and advertisement plus increased commissions when the rooms cannot be sold. Profit margins typically drop when the cancelled room needs to be resold. And last but not least, it is difficult to arrange for various human resources the last minute necessary to support the guests of these last-minute changes.

- For context, the typical cancellation reasons include change of plans, scheduling conflicts, and more. Cancelations are often made easier by the option to do it free of charge or at a low cost, which is great for hotel guests, but per the above paragraph it is not desirable for hotels. Such above-described losses are particularly high on the last-minute cancellations.

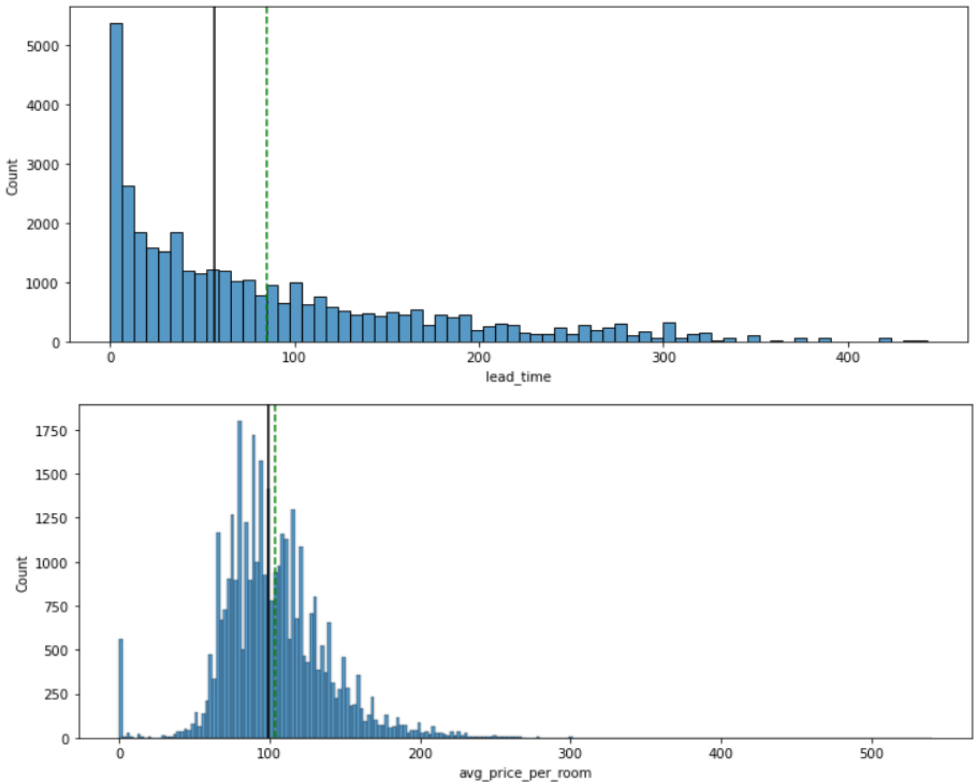
- Additionally, the new online booking channels/technologies have changed customers' booking options and behaviors. This adds another layer of difficulties in how the hotel handles cancellations, which are no longer limited to just traditional booking channels and typical guest (user) behaviors.

- The increasing number of cancellations has forced the INN Hotels Group to seek out a solution that can help them predict which types of bookings are likely to be canceled. The task is to analyze the provided data and find the factors that are likely to have the strongest influence on booking cancellations. A Machine Learning model will be built to help predict which booking is going to be canceled in advance, This will further assist in formulating profitable policies for the hotel group's cancellations and refunds.

Data Overview

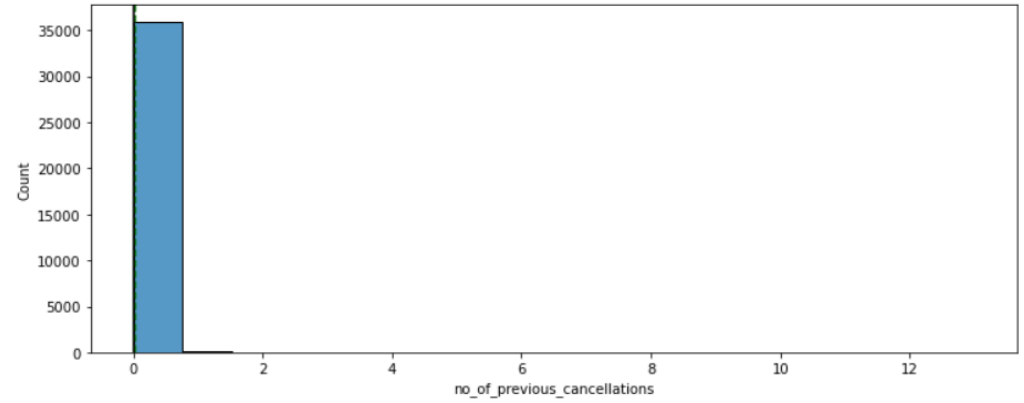
- The dataset has 36,275 rows and 19 columns of data with bookings and their various characteristics/details.
- Booking characteristics include data such as number of adults, number of children, number of weekend vs. weeknights, type of meal plans, required parking space, room type, lead reservation time, arrival dates, customer market segments, repeated guests, number of previous cancelations and separately the number of bookings not previously canceled, average price per room, number of special stay requests, and booking status.
- Data types are 14 numeric columns and 5 object data points.
- No discrepancies in the available records for each of the columns, so no missing values, and there are no duplicate values.
- Average room prices go from $0 - $540 dollars with a right skew and average price of just about $100.
- The top market segment is 'Online', top booking status is 'Not_Canceled', and top room type is Room_Type 1.
- There are 4 types of meals, 7 room type reservations, 5 market segments, and 2 booking statuses.
- Some guests have as many as 13 canceled reservations (max), and previous bookings not canceled 58 (max).
- In the following segment, we will observe bookings and booking cancelations against other booking data points.

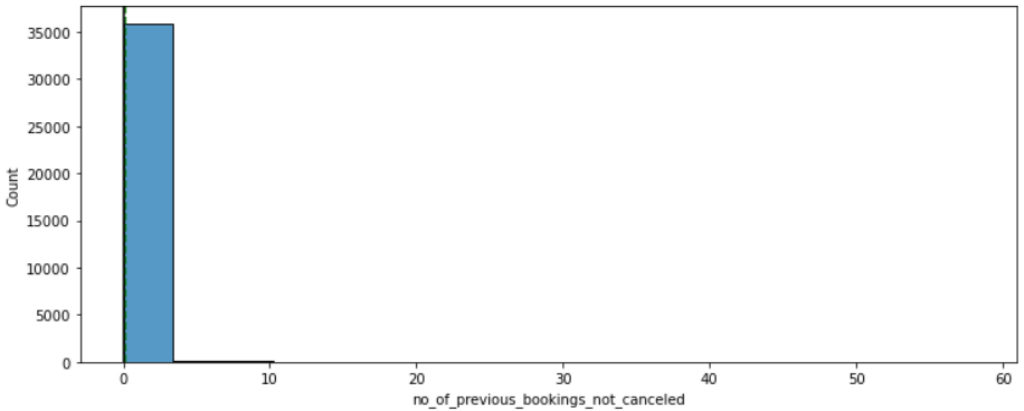
Exploratory Data Analysis



- Lead time is heavily right skewed with mean of around 80 days, and max lead times exceeding 480 days – entirely too long.
- Average price is almost bell shaped but with many outliers. Mean room price is just around $100.
- Looking at $0 priced rooms and market segments, ‘Complementary’ and ‘Online’ are the biggest segments with Complimentary being almost 2/3rds off the 0 priced rooms.
- The IQR for the room price is between $50 - $179.55 for the upper whisker.

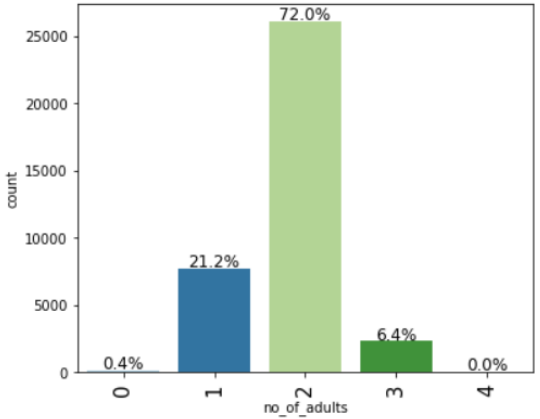
Exploratory Data Analysis

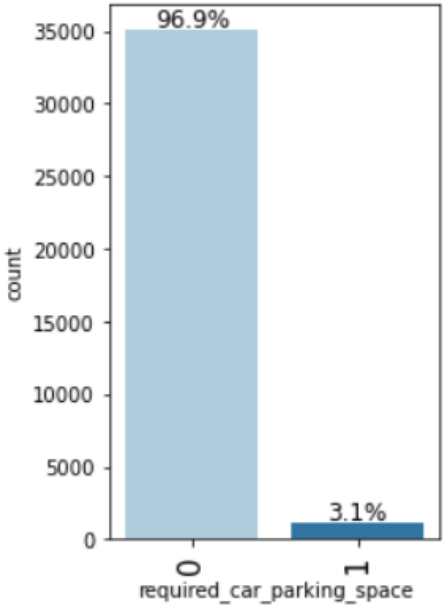


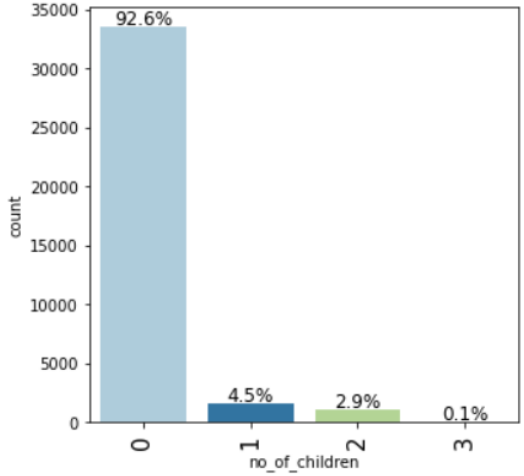


- The mean for the number of previous bookings that were canceled by the customer prior to the current booking is 0.02335 - so fairly small.
- The count of customers who have canceled 0 times is high compared to the total number of records -- 35k for 0 cancelations vs. the total records of 36.3k.
- Some of the customers have canceled up to 13 bookings prior to the current booking, but that's a very small count.

Exploratory Data Analysis

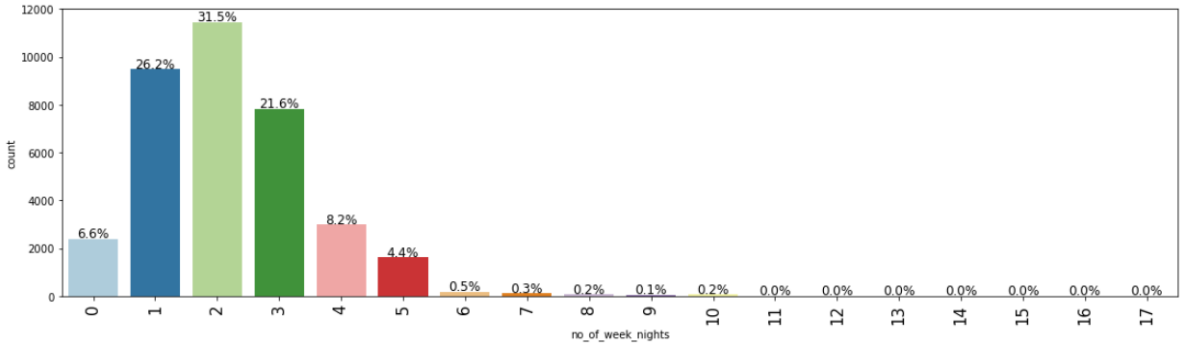


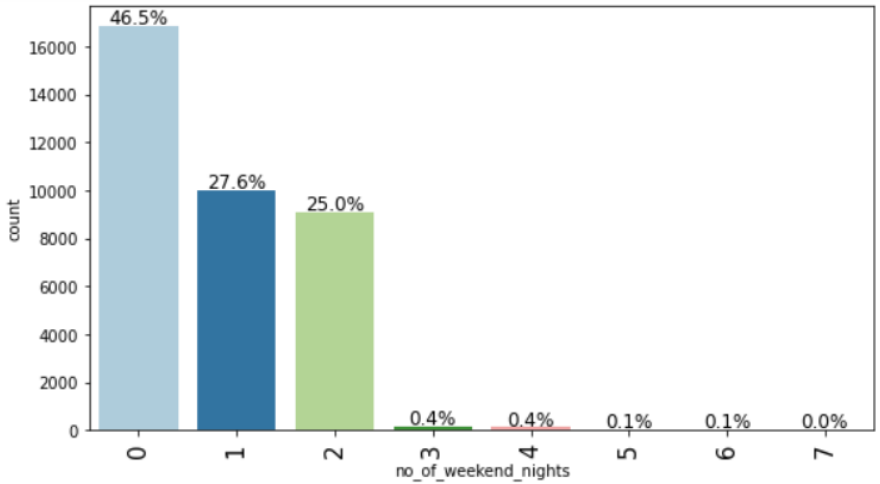




- 72% of all bookings are 2 adults, while 2 or less adults take up jointly over 93% of all bookings.
- Childless adults also take up 92.6% of all bookings, while 1-child visitors take up 4.5%, and 2 or more –child visitors take up just over 3% of all bookings.
- It is also apparent that nearly 97% of all bookings do not require a parking space.

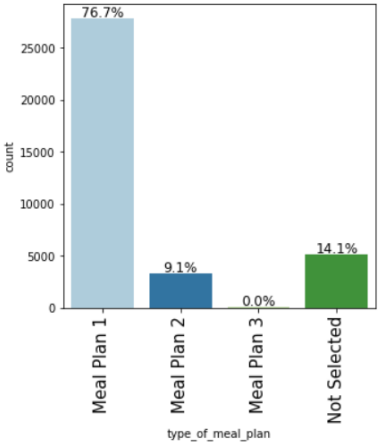
Exploratory Data Analysis

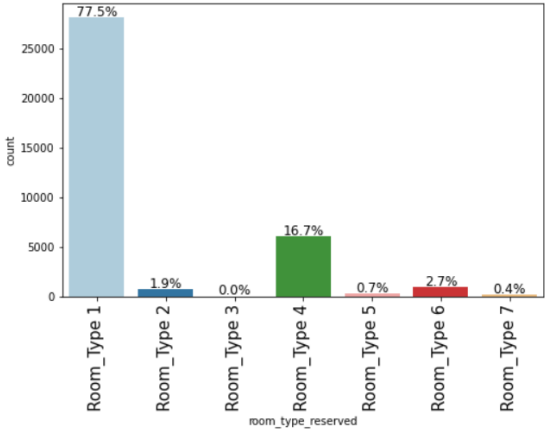


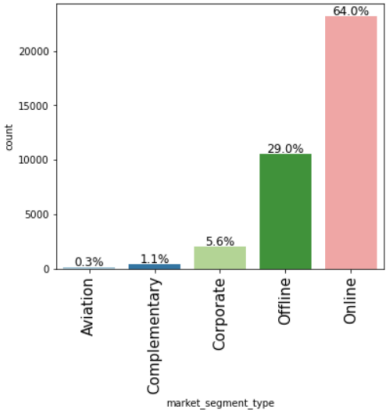


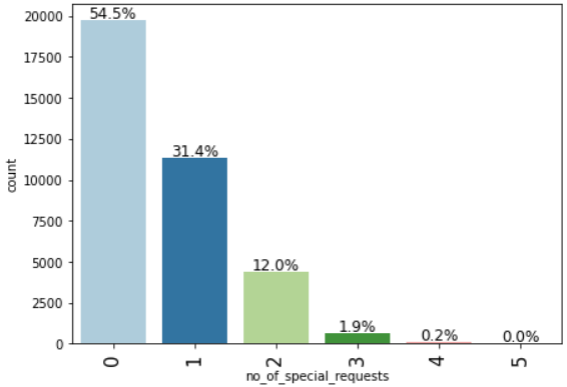
- Around 80% of all weeknight stays accounts for staying 1, 2, or 3 weeknights.
- Less than 15% of all weeknight stays stay longer than 3 weeknights.
- 46.5% of all bookings do not stay on weekends, while nearly 53.5% of all bookings stay on weekends.
- Just about 1% of all bookings stays longer than 2 weekend nights (e.g., 3 or more weekend nights).

Exploratory Data Analysis



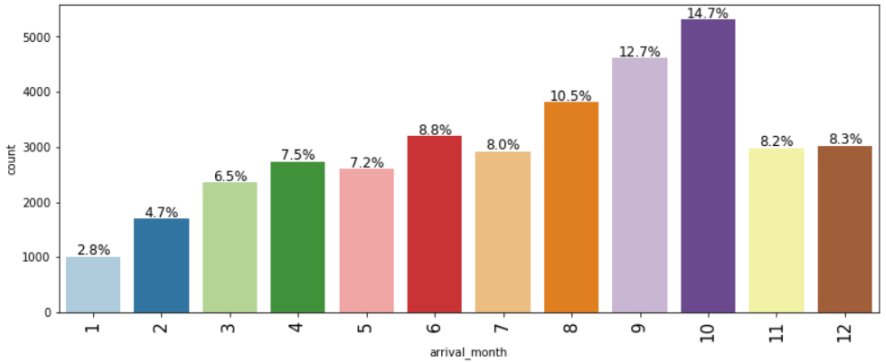


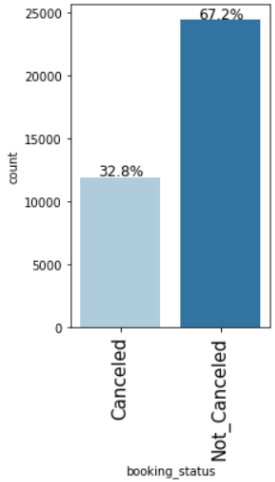




- Meal Plan 1 is nearly 77% of all plans followed by Not-Selected option being around 14%.
- The most commonly selected room type is 1 at almost 78% of all rooms followed by room type 4 at almost 17% of all bookings. All other room types together are close to 5%.
- Online market segment is the largest at 64% followed by offline at 29% and corporate at 5.6%. Complementary rooms are 1.1%.
- More than half of all bookings are without special requests. Bookings with 1 requests are at 31.4%, 2-requests at 12%, and 3-requests are at 1.9%.

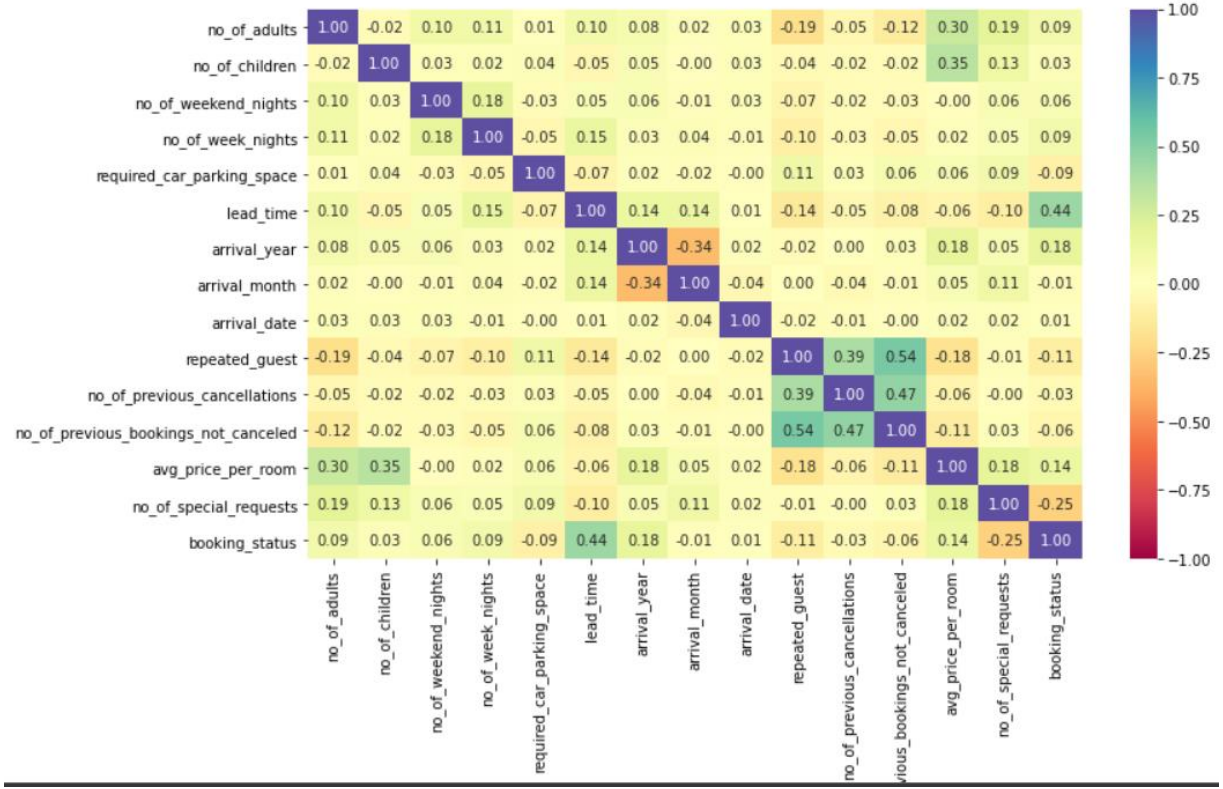
Exploratory Data Analysis





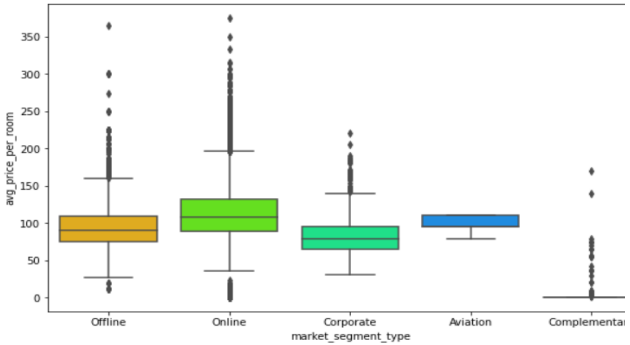
- Strongest bookings months are August, September, and October.
- April through July, and November and December are all between 7-8% of all bookings.
- January, February and March are the least visited/booked months.
- As many as 32.8% of all bookings are canceled in a given time period, while 67.2% of all bookings remain (not-canceled).

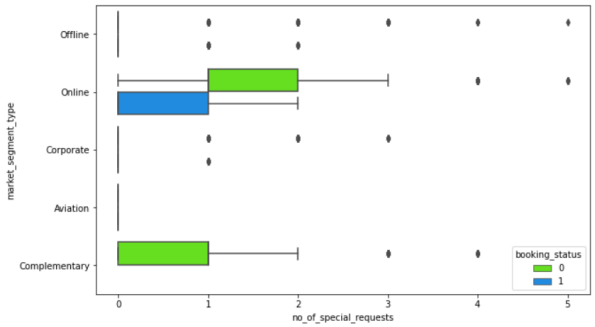
Exploratory Data Analysis

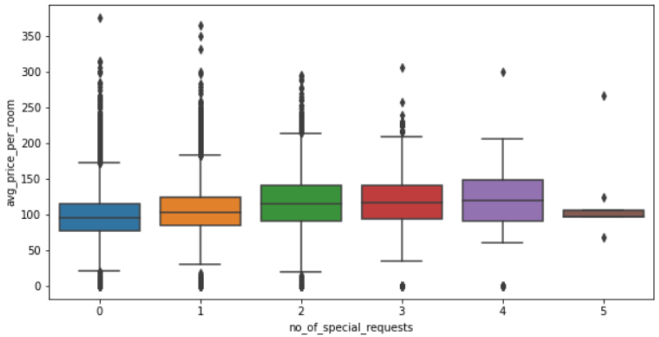


- The strongest of the correlations with the booking_status is the lead_time at 0.44.
- Repeated guests have an increased positive correlation with both, the no_of_previous_bookings_not_canceled and with the no_of_previous_cancellations, which would make sense since they're repeat customers.
- Another intuitively logical positive correlation is the number of adults and number of children correlation with the average price per room.

Exploratory Data Analysis

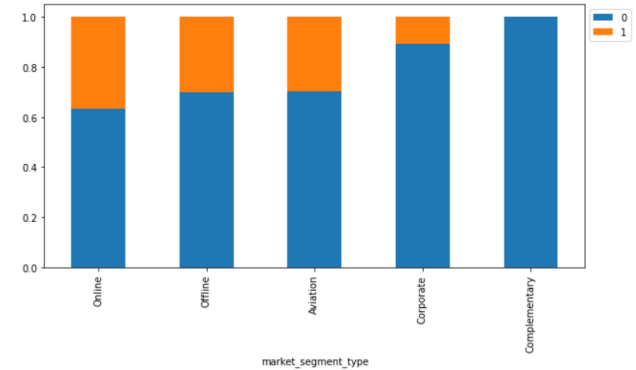


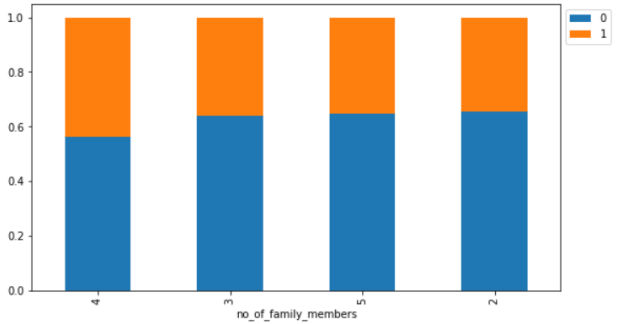


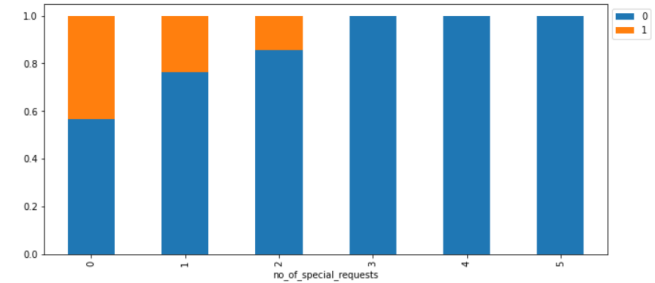


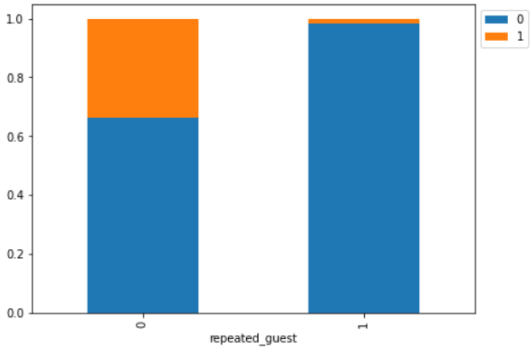
- Interestingly, it seems that the Online market segment pays the most per room. This is counterintuitive. One would hope that with all online-available travel research tools would help shoppers find the best deals.
- Offline and aviation segments are in the middle, while the Corporate segment has the lowest rates (excluding complementary).
- Average price per room seems to go up with the increase in the number of special requests, which is the correlation visible even in the heatmap. The only interesting things is the 5 requests, which seem to drop down in price.
- Also very interesting to see that most of canceled bookings come from Online customers with 0 special requests.

Exploratory Data Analysis



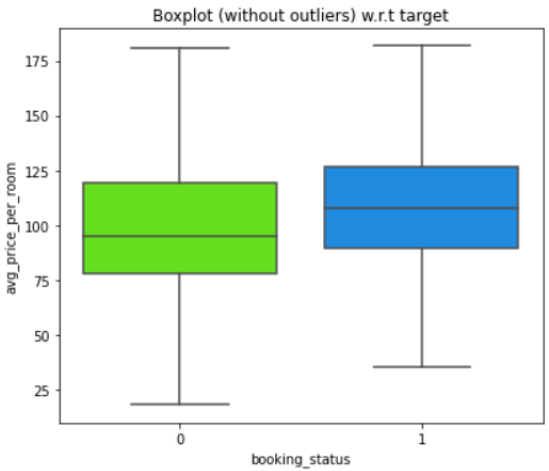


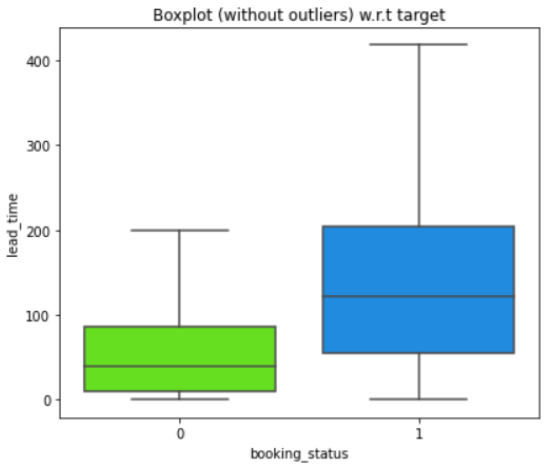




- Online followed by Offline, followed by Aviation have the highest number of cancelations.
- 4 family member packages have the highest number of cancelations. The 3, 5, and 2-member packages have the least number of cancelations.
- Seems that 0-special requests bookings have the largest number of cancelations, followed by 1 and 2, while 3, 4, and 5-special requests bookings appear to have no cancelations.
- Important: Repeated guests have hardly any cancelations compared to new customers.

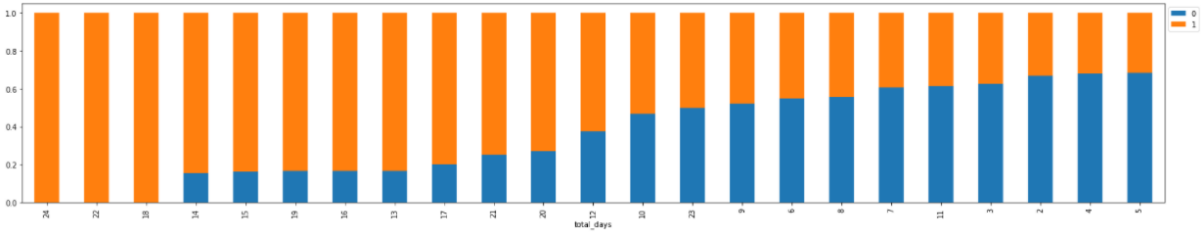
Exploratory Data Analysis

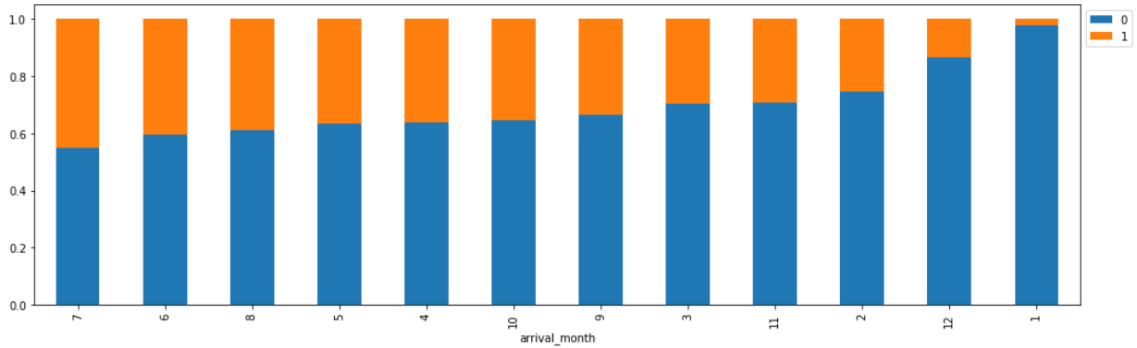


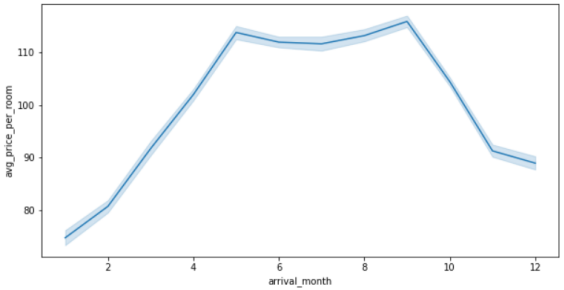


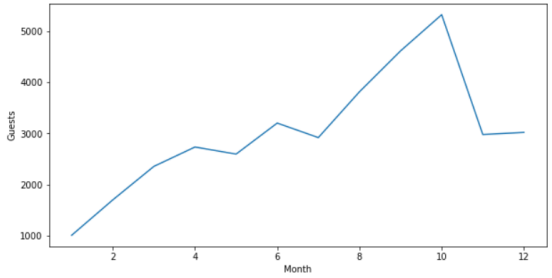
- Observing booking status against the lead time, it is apparent that longer lead times (IQR of 70 to 200 days) is associated with canceled bookings, while lower lead times (IQR of 10 to 85 days or so) is associated with non-canceled bookings.
- Observing booking status against the average room price, it is also somewhat intuitively visible that canceled bookings IQR of about $90 to $125 dollars, while non-canceled booking status is associated with average room price of about (IQR) $80 to $120 dollars. We can assume that higher room prices cause customers to look for and find cheaper room prices and therefore cancel their bookings.

Exploratory Data Analysis









- When booking status is observed against the total number of stay days, it is apparent that the larger number of stay days is corelated or associated with booking cancelations, while 5, 4, 2, and 3-day stays are associated with the smallest number of booking cancelations.
- In terms of the arrival month, we see that the smallest number of cancelations is from November through March, with December and January leading in the smallest number of cancelations. On the other end of that scale, it's June, July and August that lead as the months with the top number of booking cancelations.
- Additionally, we see that while busiest months are in September and October, the average room prices remain high from May through October.

About the Model

Data Pre-Processing and Preparation for Modeling

- There were no missing values or duplicate values.
- We checked for outliers (see screenshot in Appendix). Most outliers seem to linger around 5 data points: no_of_week_nights, lead_time, no_of_previous_bookings_not_canceled, and avg_price_per_room. I didn't treat the outliers due to me not understanding (yet) the hotel business model, so wanted to check for how outliers impact the overall data and models.
- I wanted to predict which bookings will be canceled.
- Before building the model, I encoded categorical features.
- Data was split into train and test at 70:30 to be able to evaluate the model built on the train data. The booking_status data was dropped from X.
- Percentage of classes were confirmed in the training and testing set.

About the Model – Logistic Regression

Model Building, Performance, Evaluation and Improvement – Logistic Regression

- We want to predict which bookings will be canceled based on various data categories/characteristics provided to us.
- The model can make wrong predictions if:
  - A) we predict that a customer WILL NOT cancel their booking, but the customer cancels it, or
  - B) we predicting that a customer WILL cancel their booking, but the customer does not cancel it.

- Both above cases are important, because the hotel will experience some losses in both cases. For example, in the case A) above, the hotel will lose resources as they repost the room and even lower the price for the last moment purchases. And in the case B) above, the hotel might not have the necessary resources to service the customers, hence this might create unsatisfied customers and cause other damages including brand equity damages.

- To reduce the losses, the INN Hotel Group would want their model's F1 Score to be maximized, because higher F1 scores will minimize both cases, False Negatives and False Positives.
- To calculate this, we first created functions to calculate different metrics and confusion matrix so as not to repeat the same code for each model. We created the model_performance_classification_statsmodels function to check the model performance, and the confusion_matrix_statsmodels function to plot the confusion matrix.

About the Model – Logistic Regression

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

Testing performance comparison:

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

Model Building, Performance, Evaluation and Improvement – Logistic Regression

- We built the logistic regression with statsmodels library (see Appendix), and after using logit to fit logistic regression, the first run showed a few high “P” values.
- Through Multicollinearity and checking the VIF we found a dropped a few columns with high p-values that were greater than 0.05.
- We then converted coefficients to odds log(odd) by taking the exponential of the coefficients -- the percentage change in odds is given as odds = (exp(b) - 1) * 100
- We checked the model performance on the training set through confusion matrix, and received false positive at 7.36%, and false negative at 12.10%.
- We utilized the ROC-AUC on the same training set with optimal_threshold_auc_roc = 0.37, and used precision-recall curve to find a better threshold (optimal_threshold_curve=0.42).
- The final training and testing performance comparison shows the best F1 score for 0.37 threshold & well balanced with accuracy.

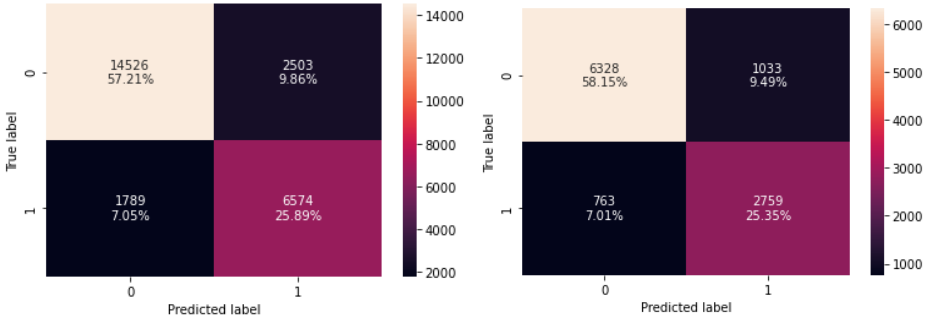
About the Model – Decision Tree

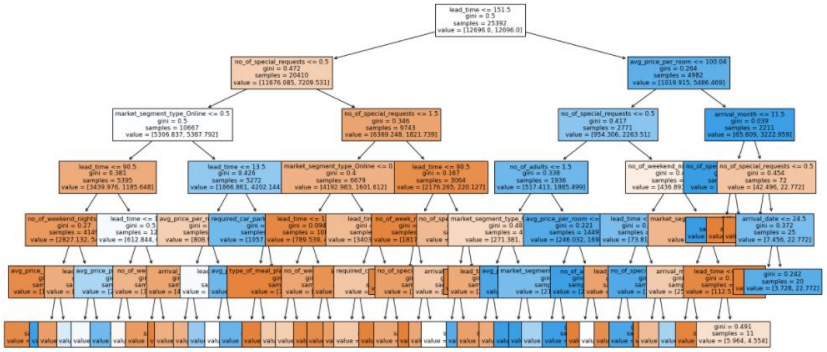
Model Building, Performance, Evaluation and Improvement – Decision Tree

- To build the tree, we first created two functions to calculate different metrics and confusion matrix to prevent using the same code for each of the models. We used A) the model_performance_classification_sklearn function to check the model performance, and B) the confusion_matrix_sklearnfunction to plot the confusion matrix.
- When we first checked the model performance on the train and test data, Accuracy, Recall, Precision and F1 were unrealistically good on the train set, but less favorable on the test data.
- We then pre-pruned the tree using *f1_score* and the following *parameters*:

"max_depth": np.arange(2, 7, 2),
"max_leaf_nodes": [50, 75, 150, 250],
"min_samples_split": [10, 30, 50, 70],

Results on training (left) & testing (right) were similar

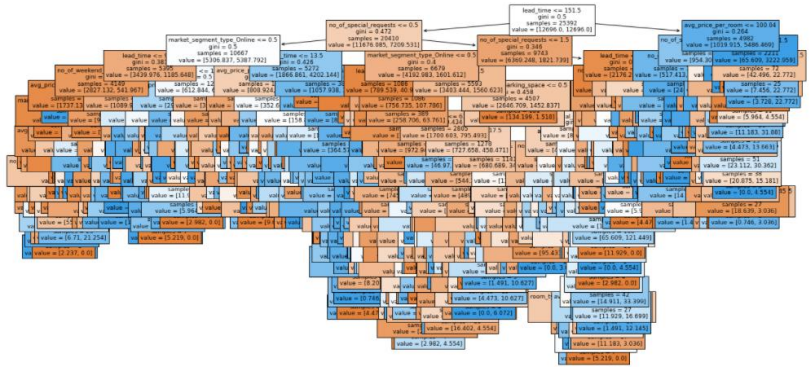




About the Model – Decision Tree

Model Building, Performance, Evaluation and Improvement – Decision Tree

- We used the Cost Complexity pruning where the cost complexity parameter *ccp_alpha* is observed. Greater values of ccp_alpha increase the number of nodes pruned. The nodes with the smallest effective alpha are pruned first. Then we train a decision tree using the *effective alphas*. The result here shows that the number of nodes in the last tree is: 1 with *ccp_alpha: 0.08117914389136954*
- We then compare F1 Score to alpha (see left chart), and found the most balanced ccp_alpha=0.0001226763315516706
- The resulting tree is more complex, visually speaking, compared to the prior model (see right) →



- However, training (left) and testing (right) F1 values look better than the prior pruning model

	Accuracy	Recall	Precision	F1
0	0.89989	0.90303	0.81353	0.85594

	Accuracy	Recall	Precision	F1
0	0.86888	0.85576	0.76634	0.80858

About the Model – Decision Tree

Final Decision Tree Comparison

- In the final comparison of decision tree models, we see that for the most part both sets, training and testing are similar with exception of the Sklearn, where the training set is higher vs. testing set. Otherwise, the pre- and post-pruning models for both, training and testing sets are quite similar.
- F1 score is the highest in the Sklearn model of the training set (0.9912), but within the testing set of data F1 score is highest within the post-pruning model (0.8086).
- The most important features that we can use as predictors (see Appendix):
  - Lead time
  - Market segment type Online
  - Average price per room
  - Number of special requests

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89989
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81353
F1	0.99117	0.75390	0.85594

Testing performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86888
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76634
F1	0.80309	0.75444	0.80858

Business Insights and Recommendations

Actionable Insights

- We analyzed the INN Hotels booking data using different techniques and we used Decision Tree to build a predictive model, which would provide insights into which bookings are likely to cancel (or not).
- We visualized different trees and their confusion matrix to get a better understanding of the model.
- We found the “Lead time”, “Market segment type Online”, “Average price per room”, and the “Number of special requests” to be the most important variables (in that order) in predicting the booking status (bookings that are likely to cancel or not).
- We were able to build a predictive model that can be used by the INN Hotels Group to find the bookings that are likely to cancel with an f1_score of 0.808 on the training set and formulate policies accordingly.
- No traveling customer should be allowed to book a room beyond a lead time of around 151.5 days. If such lead time is used, it should then be considered a very likely booking to cancel over time.
- If a lead time <= 151.5 days and avg_price_per_room <= $100.04, the booking is likely to NOT cancel. Intuitive logic, which is also visible in the Decision Tree model, is that more recent bookings that are not too expensive do not give customers many reasons to look for better deals over time (e.g., less expensive rooms), while the opposite is likely true too.

Business Insights and Recommendations

Benefits of implementing the solution

- The INN Hotel Group could use the prediction model in various efforts to control costs, influence revenue, improve its brand equity, and even improve its people operations (HR) model in busy season with reduced cancelations. For example:
- The hotel should cater to those customers who are more certain of their travel times in shorter time spans (shorter than lead time of 151 days on average), since lead time is such an important predictor of booking cancelations.
- With 64% of its clientele being in the Online market segment, which is simultaneously an important predictor, the hotel should focus on how they cater to that audience. For example, and based on decision tree and EDA research (slide 12), they should make every attempt to push online customers into more than 1 special requests. With nearly half of all stays without a special request, it's apparent that online customers with 0 special requests show much less commitment to their stay, hence a likely booking cancelation.
    - One thing, though, to keep in mind with the above suggestion based on the Decision Tree model is that, per EDA analysis, larger number of special requests are typically associated/correlated with the higher average room rates. And since higher room prices can lead to more price sensitivity and possible booking cancelations, the hotel should carefully find the right balance between the number of special requests and room prices.
- The hotel should explore how they regulate their price in busy and slow seasons to adjust the price for the price sensitive online segment, which has much better tools to find a better deal if the booked room appears overpriced. Overpriced rooms are likely to lead to cancelations as online customers find better deals elsewhere, which results in larger costs to hotel when the booking is canceled.
- Conversely, the hotel could use the prediction model to focus on other market segments in different seasons that are less price sensitive or lead time sensitive. For example, offline segment seems to be more tolerant towards longer lead times and higher prices.

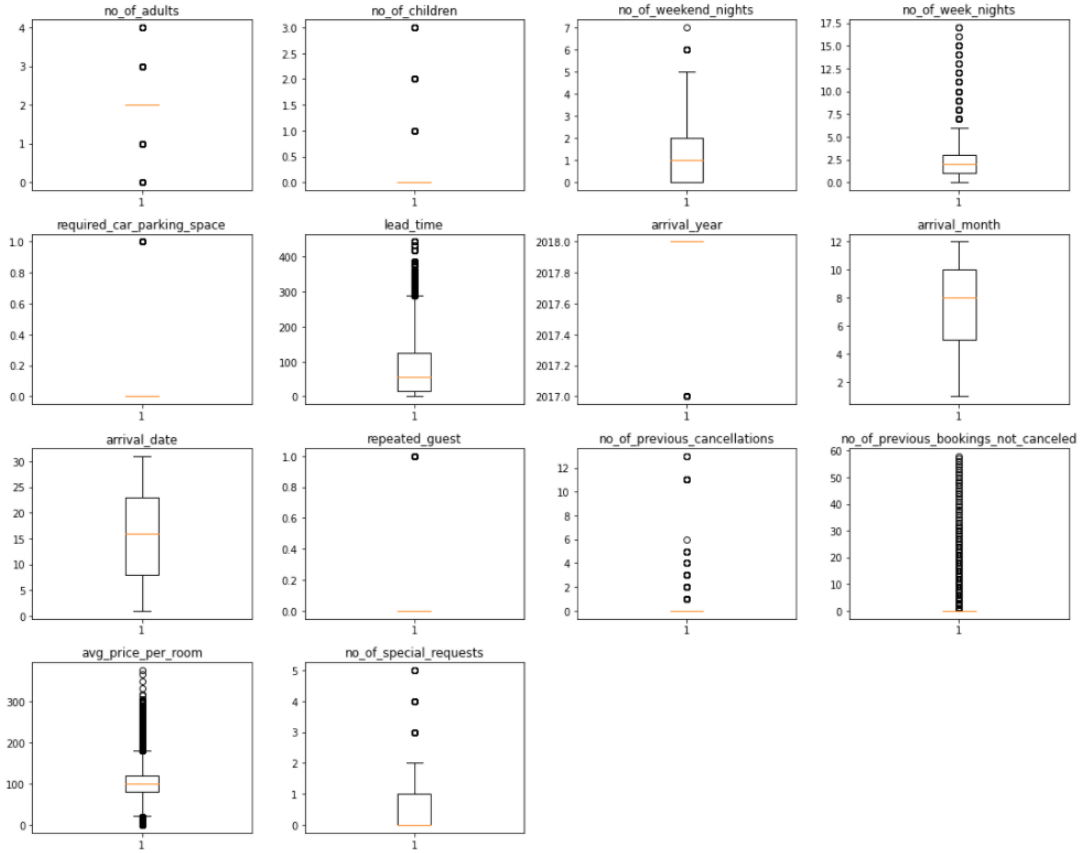
greatlearning
Power Ahead

Happy Learning !

Appendix – Data Dictionary

- 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:
- 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 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.

Outliers



Logistic Regression

In [76]:
```
logit1 = sm.Logit(
    y_train, X_train1.astype(float)
)  ## Complete the code to train logistic regression on X_train1 and y_train
lg1 = logit1.fit()  ## Complete the code to fit logistic regression
print(lg1.summary())  ## Complete the code to print summary of the model
```

```
Optimization terminated successfully.
         Current function value: 0.425731
         Iterations 11
                          Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25370
Method:                           MLE   Df Model:                           21
Date:                Tue, 11 Jan 2022   Pseudo R-squ.:                  0.3282
Time:                        21:58:38   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
=================================================================================================
```

Feature Importance