INN Hotels - Supervised Machine Learning

Michail Mersinias

04/13/2023

Contents / Agenda

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

Executive Summary

- The executive summary (insights and recommendations) of the INN Hotels Supervised Machine Learning presentation is as follows:
    - **Model of Choice:** A post-pruning Decision Tree Classifier has the ability to help the business predict the bookings with high risk of cancellation, and the factors that lead to it.
    - **Lead Time Matters:** This was the most important factor for cancellations. The company should send a reconfirmation request, as well as reminders, close to the date of the booking.
    - **Online Bookings:** Online bookings have a higher risk of being cancelled. Thus, cancellation policies should be stricter, with less refund offers, when it comes to online bookings.
    - **Length of Stay:** When the length of stay is over 11 days, the risk of cancellation is high. A limit of 10 days should be applied, with the ability to extend the stay with a new booking.
    - **Loyalty Program:** Repeated guests have a much lower chance to cancel their booking. Thus, a carefully designed loyalty program should be launched in order to retain customers and turn them into repeated guests.

Business Problem Overview and Solution Approach

- Problem Definition:
  - INN Hotels Group has a chain of hotels in Portugal, and they are facing problems with the high number of booking cancellations.
  - The goal 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.
- Solution Approach and Methodology:
  - For EDA, both univariate and multivariate analysis will be performed to find insights.
  - For Data Preprocessing, missing value imputation and feature engineering will be performed.
  - For Model Building, we will use a both a Logistic Regression model and a Decision Tree model.

EDA Results: Data Overview

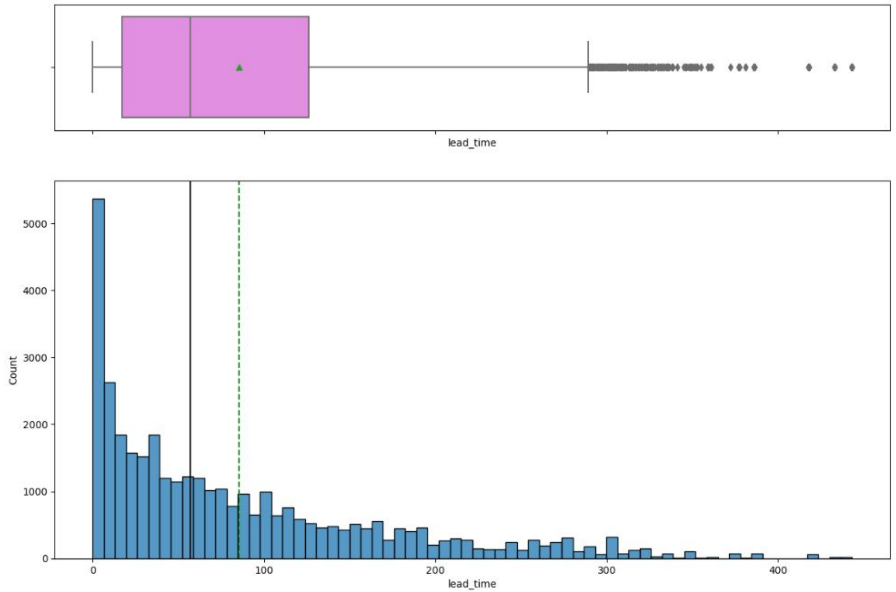
- The dataset is comprised of 36275 rows and 19 columns.
- The columns are as follows: Booking_ID, no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, type_of_meal_plan, required_car_parking_space, room_type_reserved, lead_time, arrival_year, arrival_month, arrival_date, market_segment_type, repeated_guest, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests, booking_status.
- There are no duplicate values in the dataset.
- There are no missing values in the dataset.

EDA Results: Univariate Analysis

- In this section, we perform univariate analysis on the data.
  - For each attribute, we perform a descriptive statistical analysis where only that attribute is involved as a variable.
  - We also analyze the corresponding data and present the distribution of the attribute, with confidence intervals to signify variance and statistical significance.
  - Finally, we write the conclusion based on both quantitative and qualitative observations.

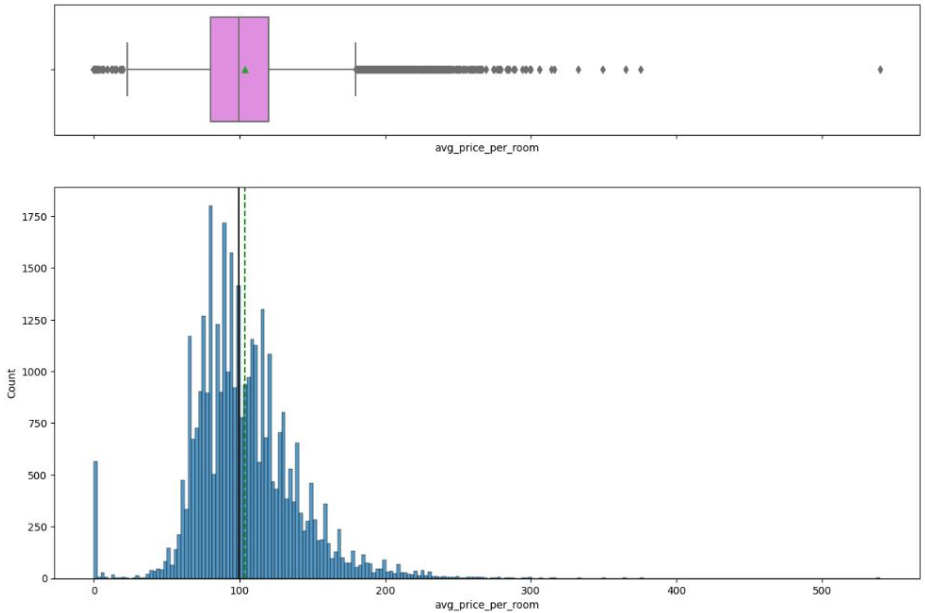
EDA Results: Univariate Analysis

- Analyzing the lead_time attribute, we report a mean value of 85.23, with a standard deviation of 85.93. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



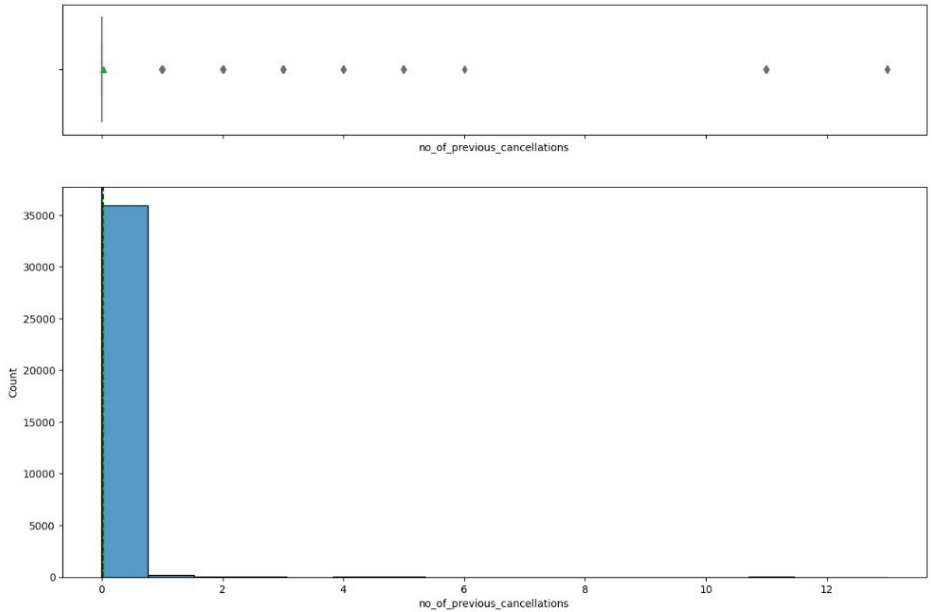
EDA Results: Univariate Analysis

- Analyzing the avg_price_per_room attribute, we report a mean value of 103.42, with a standard deviation of 35.08. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



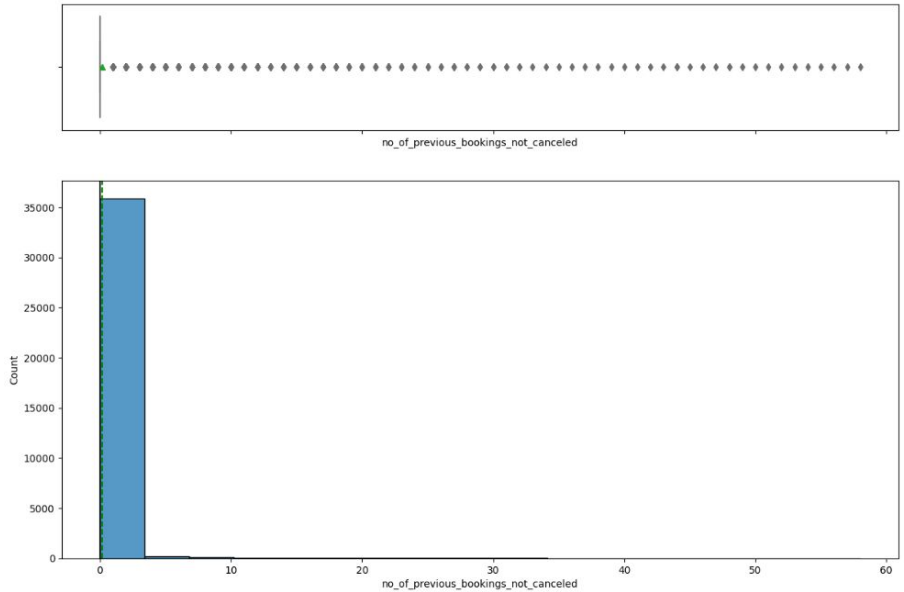
EDA Results: Univariate Analysis

- Analyzing the no_of_previous_cancellations attribute, we report a mean value of 0.023, with a standard deviation of 0.368. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



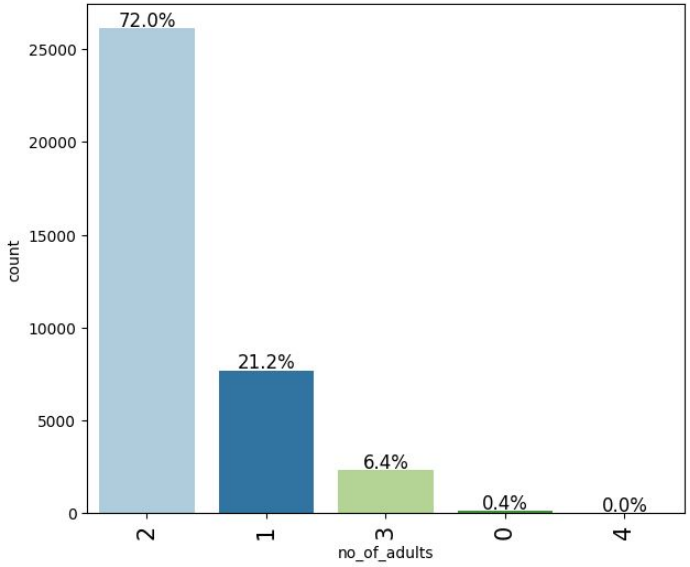
EDA Results: Univariate Analysis

- Analyzing the no_of_previous_bookings_not_canceled attribute, we report a mean value of 0.153, with a standard deviation of 1.754. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



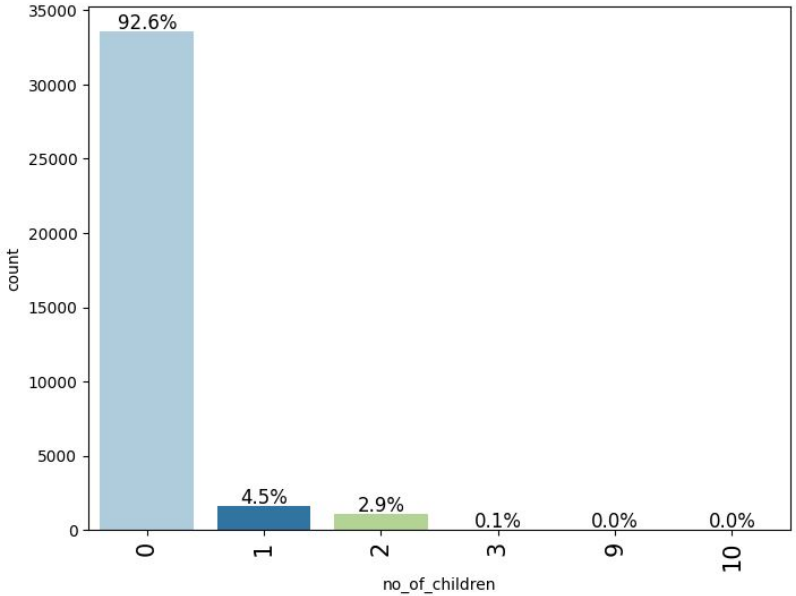
EDA Results: Univariate Analysis

- Analyzing the no_of_adults attribute, we report a mean value of 1.84, with a standard deviation of 0.51. A countplot is presented to depict this more clearly.



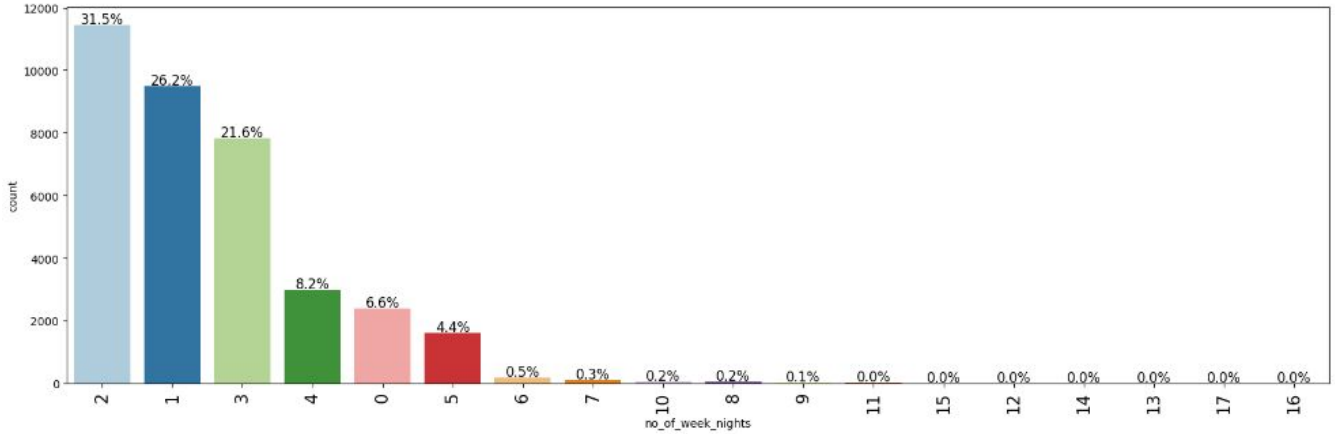
EDA Results: Univariate Analysis

- Analyzing the no_of_children attribute, we report a mean value of 0.105, with a standard deviation of 0.402. A countplot is presented to depict this more clearly.



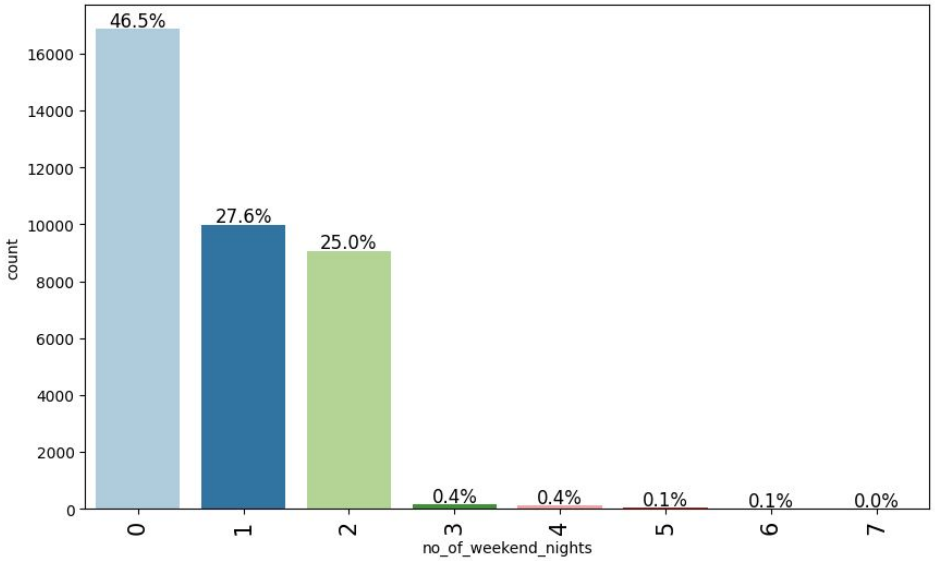
EDA Results: Univariate Analysis

- Analyzing the no_of_week_nights attribute, we report a mean value of 2.20, with a standard deviation of 1.41. A countplot is presented to depict this more clearly.



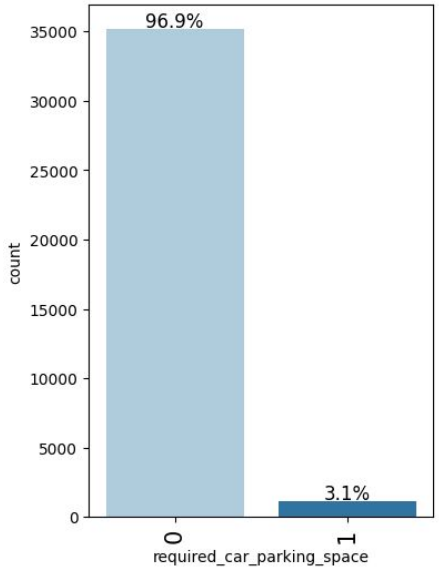
EDA Results: Univariate Analysis

- Analyzing the no_of_weekend_nights attribute, we report a mean value of 0.810, with a standard deviation of 0.870. A countplot is presented to depict this more clearly.



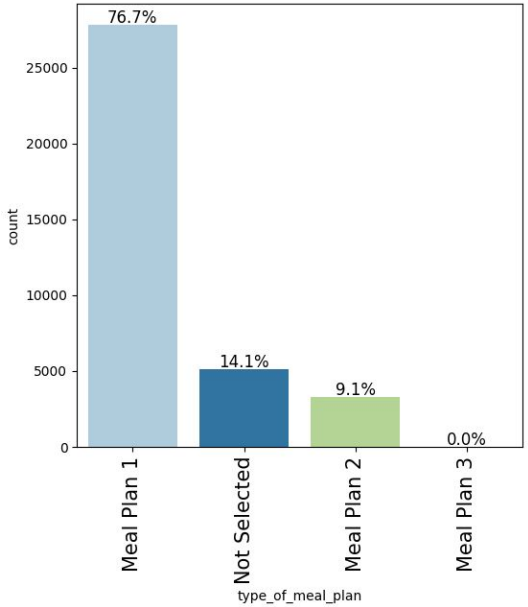
EDA Results: Univariate Analysis

- Analyzing the required_car_parking_space attribute, we that most guests (96.9%) require no car parking space. A countplot is presented to depict this more clearly.



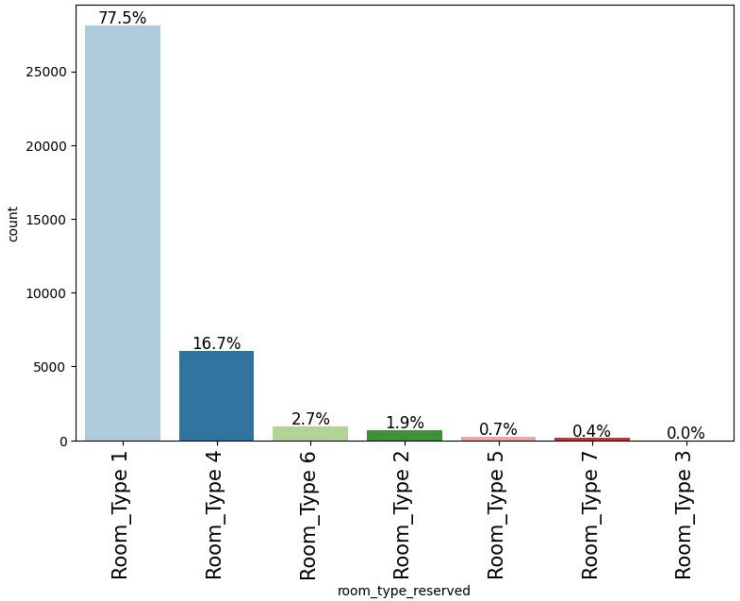
EDA Results: Univariate Analysis

- Analyzing the type_of_meal_plan attribute, we report Meal Plan 1 as the most common (76.7%) meal plan. Furthermore, 14.1% did not select a meal plan. A countplot is presented to depict this more clearly.



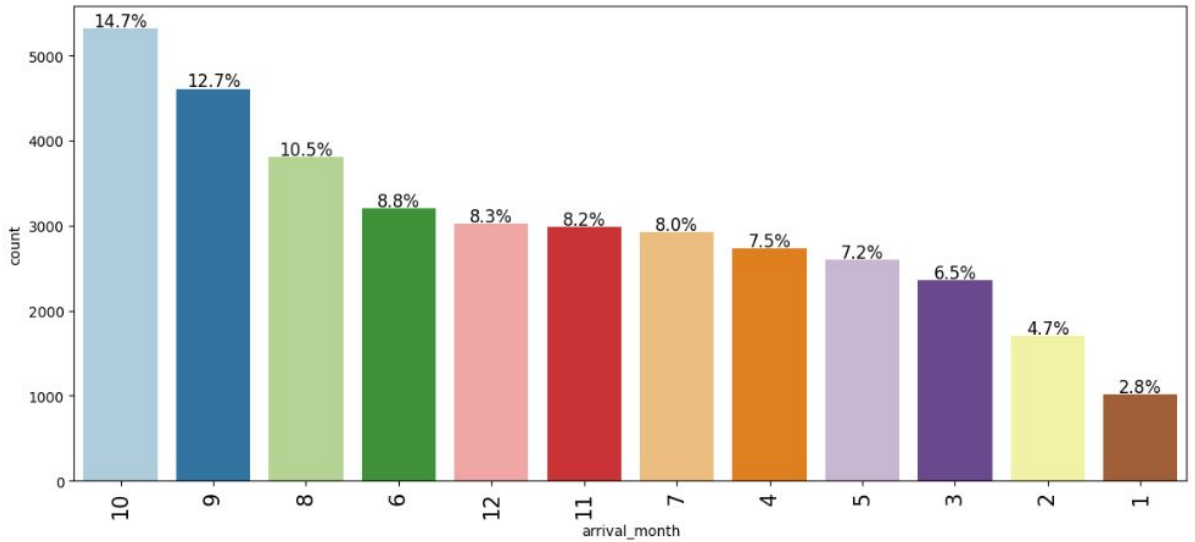
EDA Results: Univariate Analysis

- Analyzing the room_type_reserved attribute, we report Room Type 1 as the most common (77.5%) room plan, followed by Room Type 4 (16.7%). A countplot is presented to depict this more clearly.



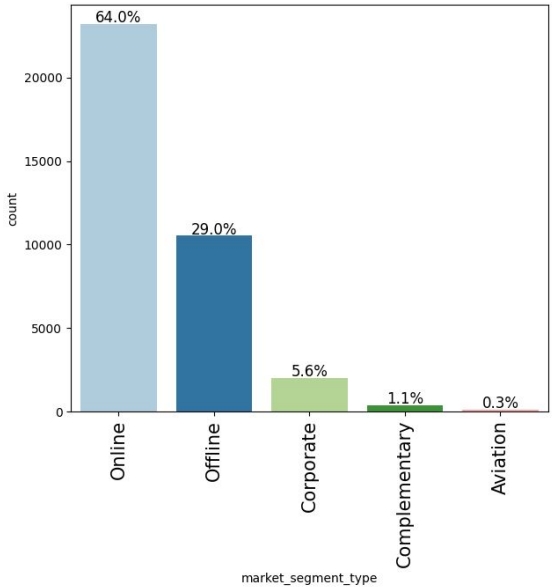
EDA Results: Univariate Analysis

- Analyzing the arrival_month attribute, we report the months of October, September and August as the most popular ones. The first three months of the year - January, February, March - are the least popular. A countplot is presented to depict this more clearly.



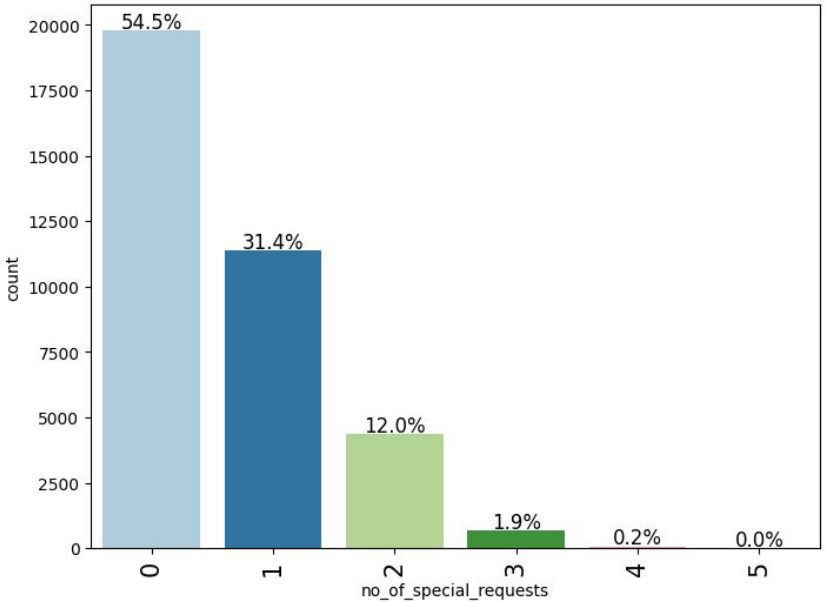
EDA Results: Univariate Analysis

- Analyzing the market_segment_type attribute, we report that the most common market segment is Online (64%). A countplot is presented to depict this more clearly.



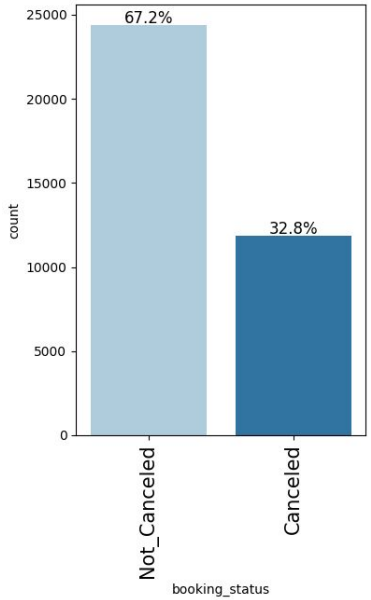
EDA Results: Univariate Analysis

- Analyzing the no_of_special_requests attribute, we report that most (54.5%) people had no special requests. A countplot is presented to depict this more clearly.



EDA Results: Univariate Analysis

- Analyzing the booking_status attribute, we report that 67.2% did not cancel their booking, but 32.8% cancelled their booking. A countplot is presented to depict this more clearly.

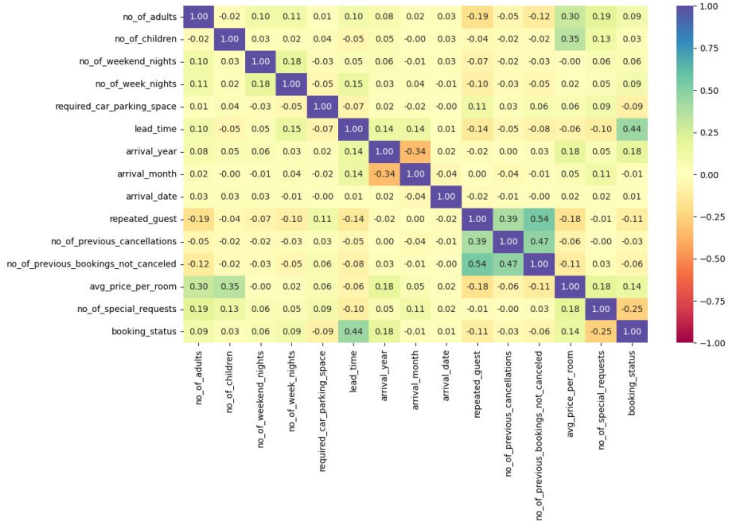


EDA Results: Multivariate Analysis

- In this section, we perform multivariate analysis on the data.
    - For selected attributes, we perform a descriptive statistical analysis where each attribute is jointly examined along with other attributes in order to evaluate their relationship and correlation.
    - We also analyze the corresponding data and present the joint distribution of the attributes, with confidence intervals to signify variance and statistical significance.
    - Finally, we write the conclusion based on both quantitative and qualitative observations.

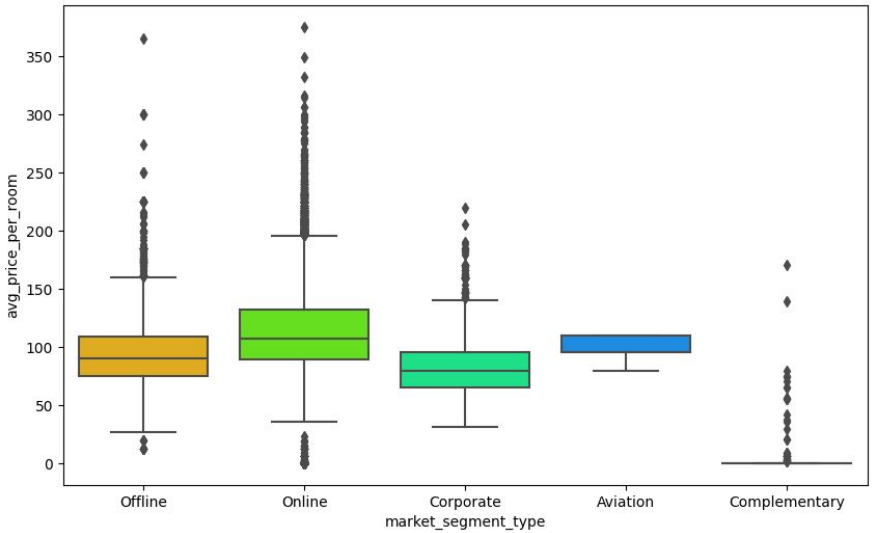
EDA Results: Multivariate Analysis

- Analyzing the correlation matrix between variables, we highlight the strong positive correlation between booking_status and lead_time, the medium positive correlations between booking_status and avg_price_per_room, and booking_status and arrival_year. Also, the medium negative correlations between booking_status and no_of_special_requests, and booking_status and repeated_guest.



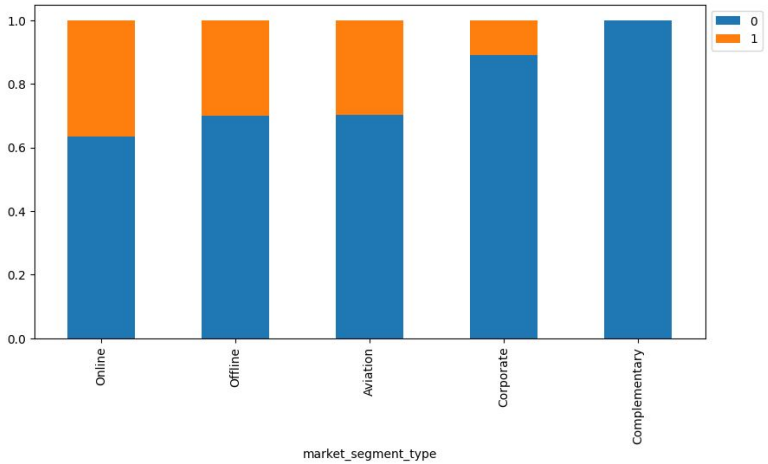
EDA Results: Multivariate Analysis

- Analyzing the relationship between the market_segment_type and avg_price_per_room attributes, we observe that online bookings lead to a particularly higher average price per room. The aviation category also leads to a high average price per room.



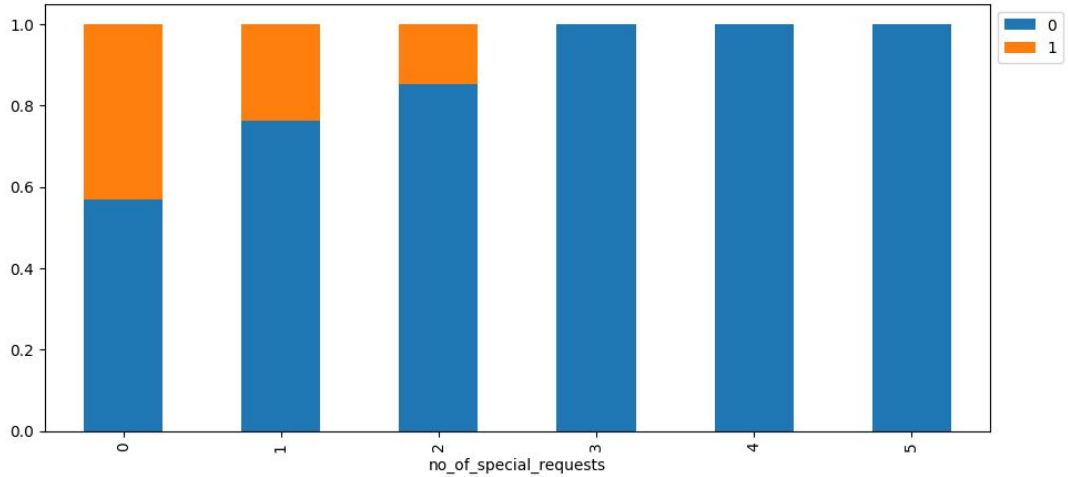
EDA Results: Multivariate Analysis

- Analyzing the relationship between the booking_status and market_segment_type attributes, we observe that successful bookings come from Corporate and Complementary, while Online bookings have the highest chance to be cancelled.



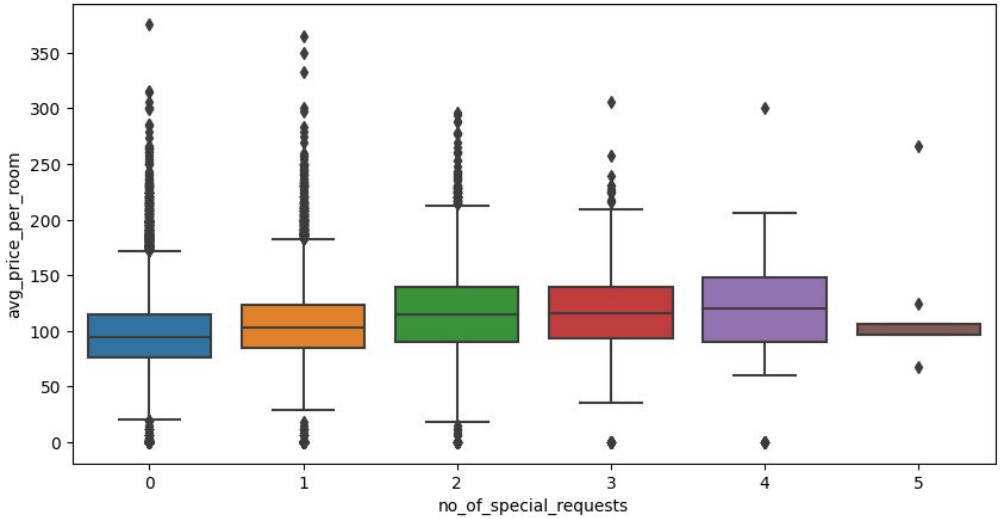
EDA Results: Multivariate Analysis

- Analyzing the relationship between the booking_status and no_of_special_requests attributes, we observe that successful bookings include 2 or more special requests, while those with 0 or 1 special requests have a very high chance to be cancelled.



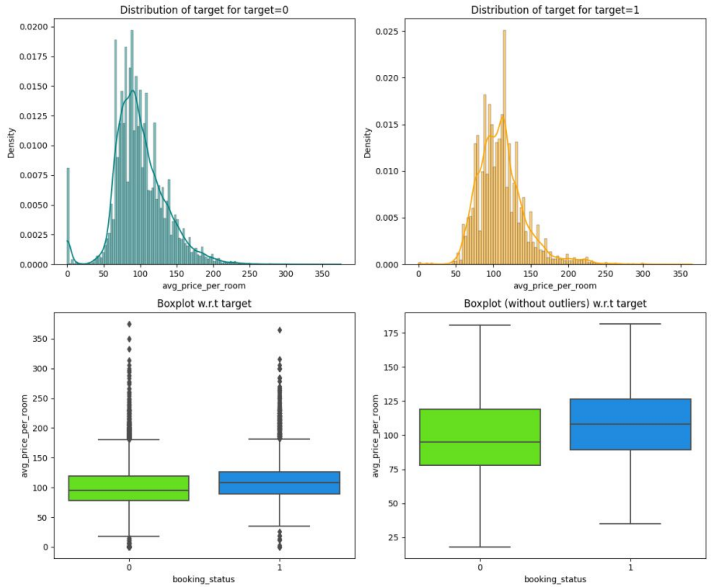
EDA Results: Multivariate Analysis

- Analyzing the relationship between the avg_price_per_room and no_of_special_requests attributes, we observe that as the number of special requests increases, there is also a slight increase in the average price per room, which is however not significant.



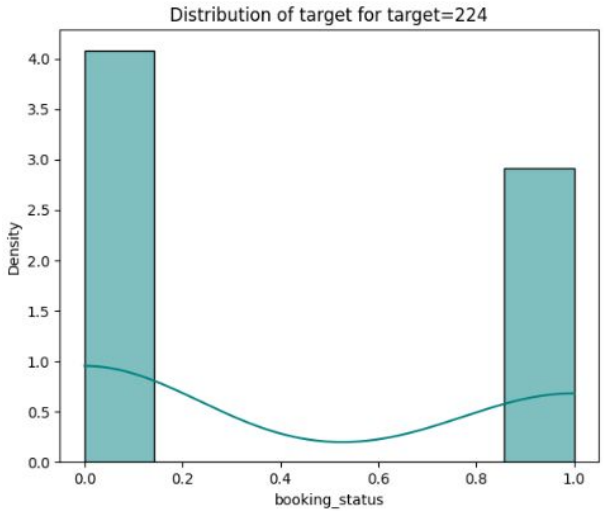
EDA Results: Multivariate Analysis

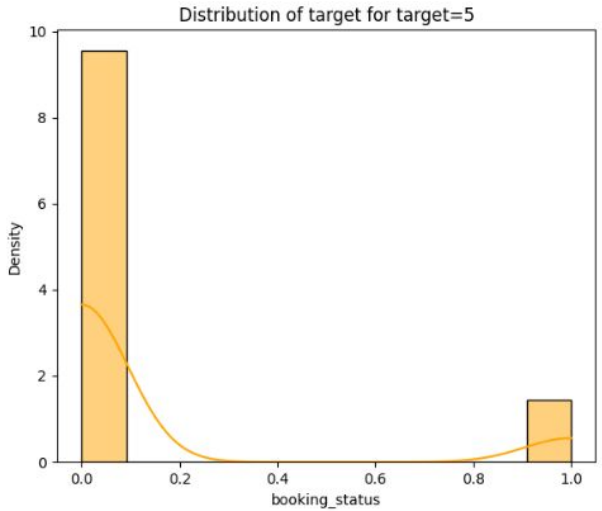
- Analyzing the relationship between the booking_status and avg_price_per_room attributes, we observe that rooms that get cancelled have a slightly higher average price, however this finding is not statistically significant.



EDA Results: Multivariate Analysis

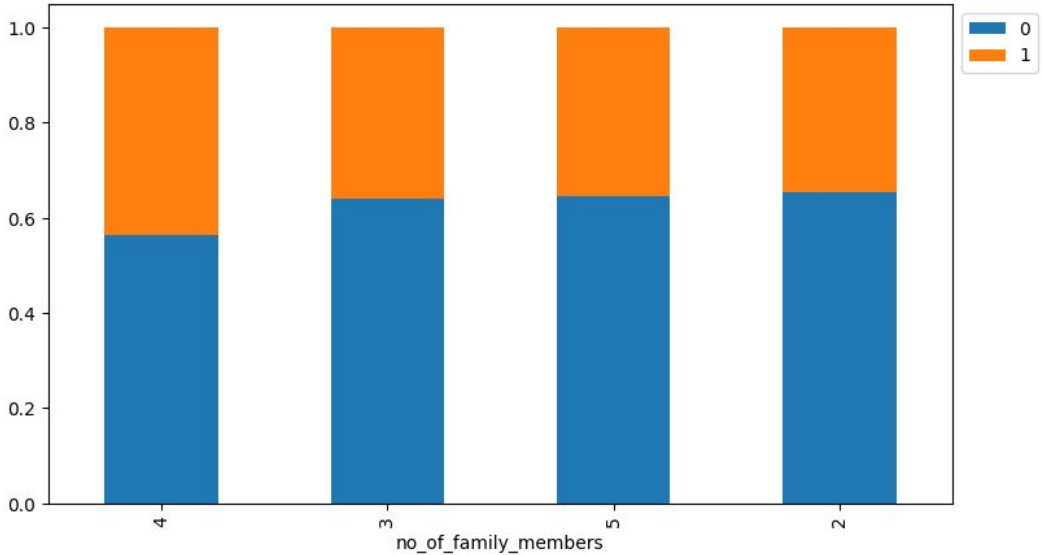
- Analyzing the relationship between the booking_status and lead_time attributes, we observe that a higher lead time leads to a greater chance of a booking cancellation.





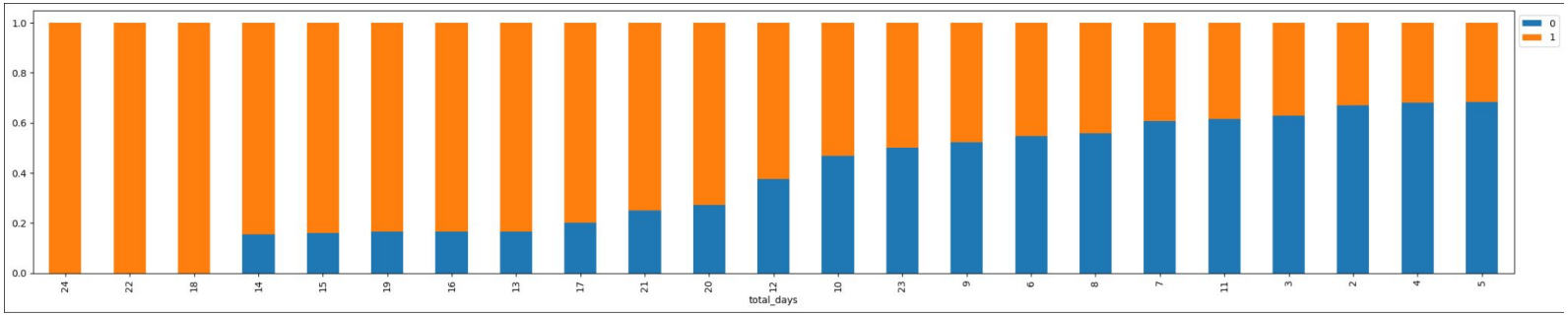
EDA Results: Multivariate Analysis

- Analyzing the relationship between the booking_status and no_of_family_members attributes, we observe no significant correlation between number of family members and cancellation risk. Large families (4 family members) have a slightly higher risk, however.



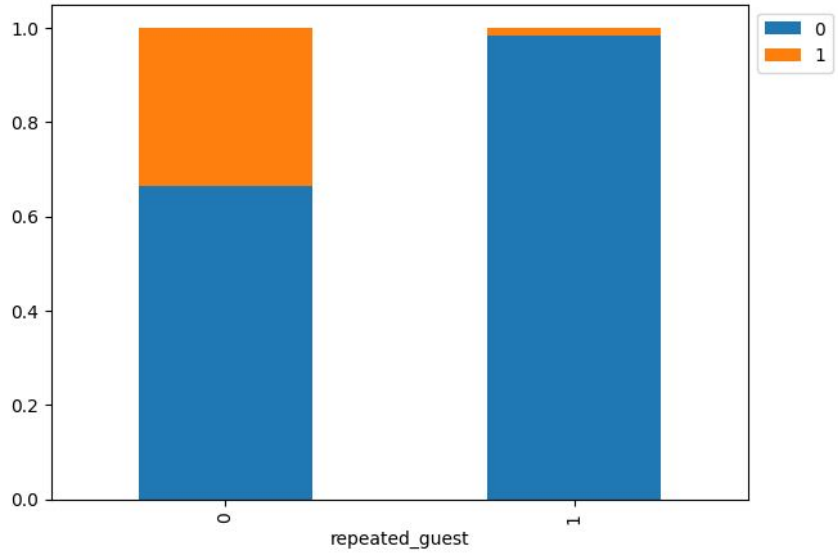
EDA Results: Multivariate Analysis

- Analyzing the relationship between the booking_status and total_days attributes, we observe that successful bookings are those that correspond to 11 days or less. Generally, as the number of total days increases beyond that point, so does the risk of cancellation.



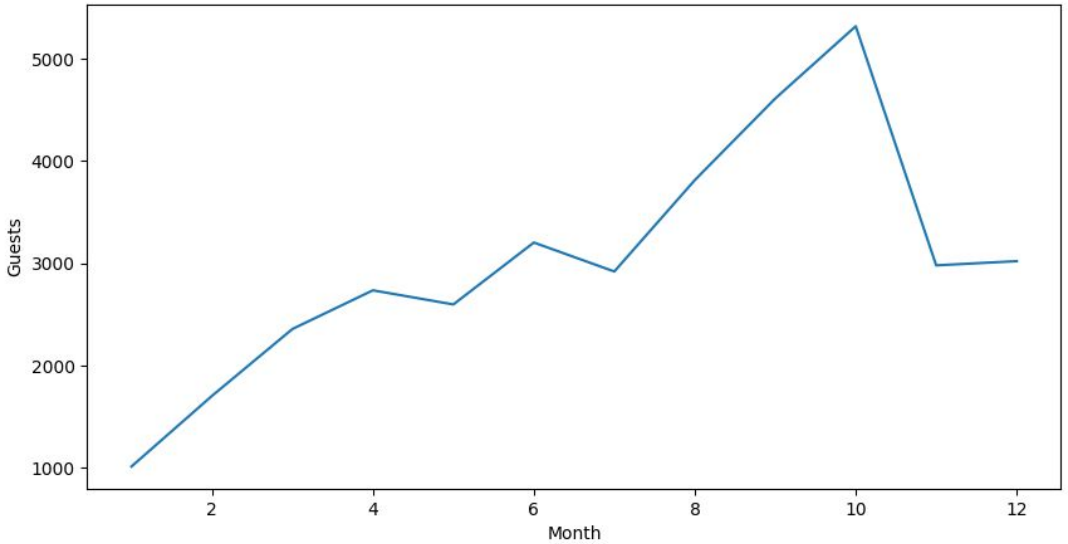
EDA Results: Multivariate Analysis

- Analyzing the relationship between the booking_status and repeated_guest attributes, we observe that a repeated guest has a significantly lower chance to cancel their booking.



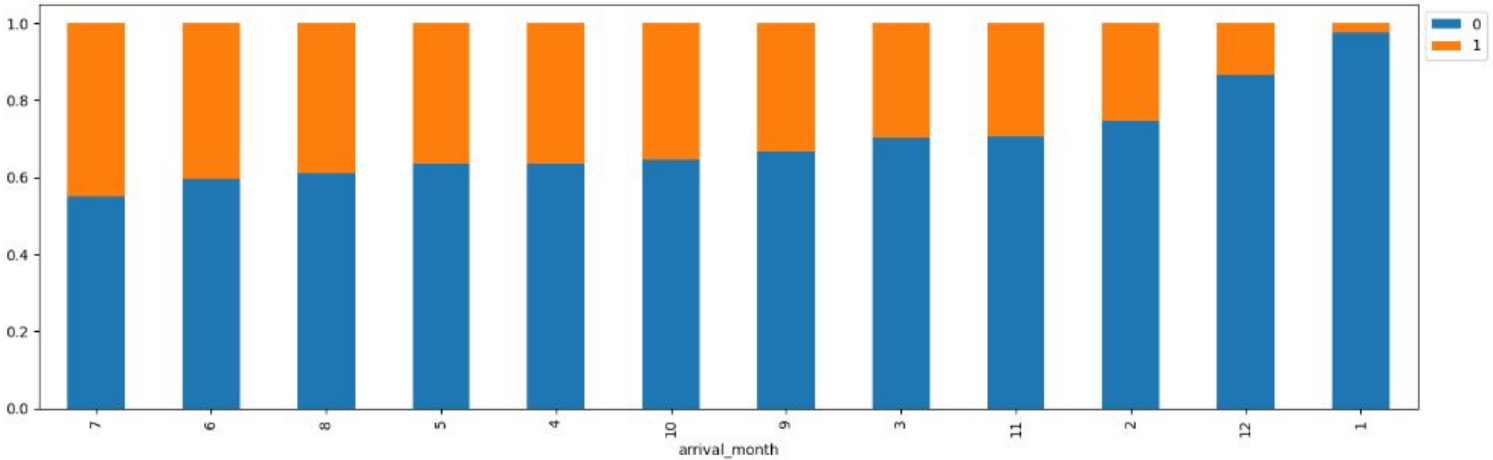
EDA Results: Multivariate Analysis

- Analyzing the relationship between the booking_status and arrival_month attributes (grouped by as total guests per month), we observe that the winter and spring months are less busy, while the summer and fall months are significantly busier. However, December is also busy, probably due to the Christmas holidays.



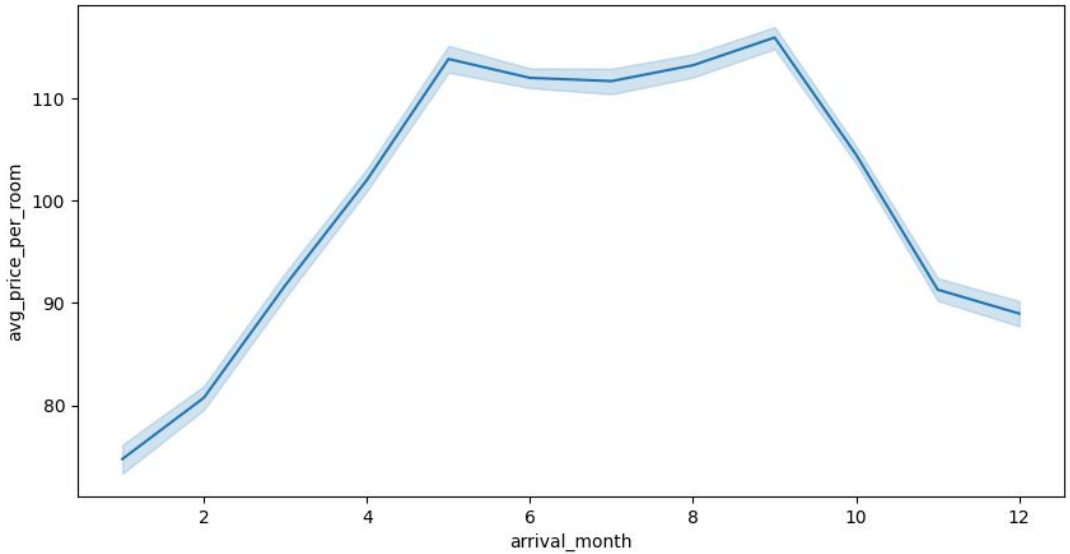
EDA Results: Multivariate Analysis

- Analyzing the relationship between the booking_status and arrival_month attributes, we observe that the highest risk of cancellation exists for the summer months of Juny, July and August. The risk of cancellation for the winter months of December, January and February is very low.



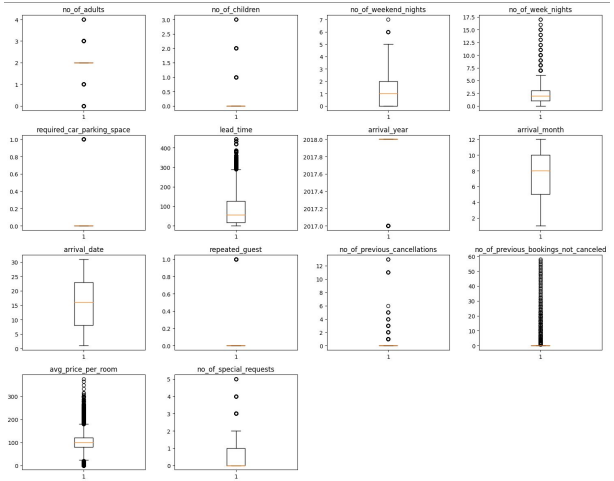
EDA Results: Multivariate Analysis

- Analyzing the relationship between the avg_price_per_room and arrival_month attributes, we observe that the time period between May and September, which corresponds to the tourist season, displays the highest average prices per room.



Data Preprocessing

- Outlier Check: There are numerous outliers in this most columns of this dataset, however no action will be taken as they are accurate values that contain important information about each feature.



- Preparing Data for Modeling: We use booking_status as our target variable which we seek to predict accurately, and all the other columns as features to train our model.
- Train/Test Dataset Split: We split the data in 70:30 ratio for train to test data.

Model Performance Summary

- Modeling Setup:
  - First, we choose our model, which is Logistic Regression.
  - Then, we drop any high p-value variables and thus, only keep a subset of the features.
  - The selected features are the following:

```
['const', 'no_of_adults', 'no_of_children', 'no_of_weekend_nights', 'no_of_week_nights', 'required_car_parking_space', 'lead_ti
me', 'arrival_year', 'arrival_month', 'repeated_guest', 'no_of_previous_cancellations', 'avg_price_per_room', 'no_of_special_re
quests', 'type_of_meal_plan_Meal Plan 2', 'type_of_meal_plan_Not Selected', 'room_type_reserved_Room_Type 2', 'room_type_reserv
ed_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']
```

  - Finally, we define functions for the following metrics of performance: Accuracy, Precision, Recall and F1 Score.

Model Performance Summary

- The results on the training data are the following:
  - Vanilla Logistic Regression:

	Accuracy	Recall	Precision	F1
0	0.80600	0.63410	0.73971	0.68285

  - Logistic Regression after dropping high p-values to solve multicollinearity:

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

  - Logistic Regression after converting coefficients to odds:

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

  - Logistic Regression with optimal AUC ROC threshold (0.37):

	Accuracy	Recall	Precision	F1
0	0.79265	0.73622	0.66808	0.70049

  - Logistic Regression with optimal Recall-Precision Curve threshold (0.42):

	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

Model Performance Summary

- The results on the test data are the following:

	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

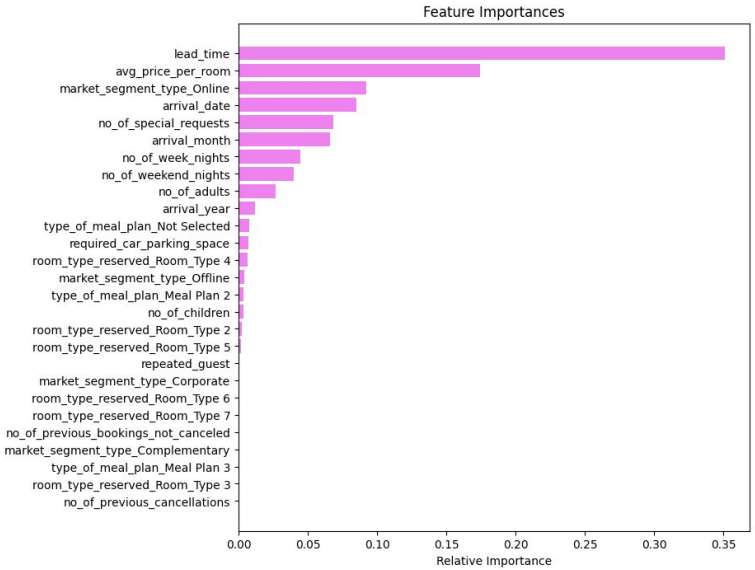
- Therefore, a Logistic Regression with either the 0.37 threshold (which leads to the highest Recall) or the 0.42 threshold (which leads to the highest Precision) perform the best and similarly well, as they lead to an F1 Score of approximately 70%.

Model Performance Summary

- Modeling Setup:
  - First, we choose our model, which is a Decision Tree Classifier.
  - Then, we tune the parameters of the model using Grid Search, and find the best combination.
  - Afterwards, we prune the tree to only keep the important features and thus, only keep a subset of the features.
  - Finally, we define functions for the following metrics of performance: Accuracy, Precision, Recall and F1 Score.

Model Performance Summary

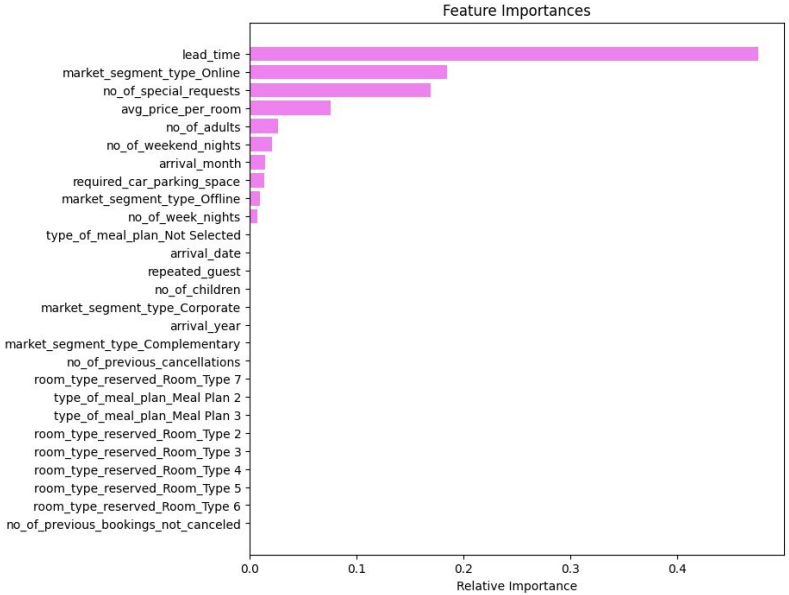
- The feature importance of the Vanilla Decision Tree Classifier is the following:



- Important features: Lead time, average price per room, online market segment type, number of special requests, arrival month, number of week night, number of weekend nights, number of adults.

Model Performance Summary

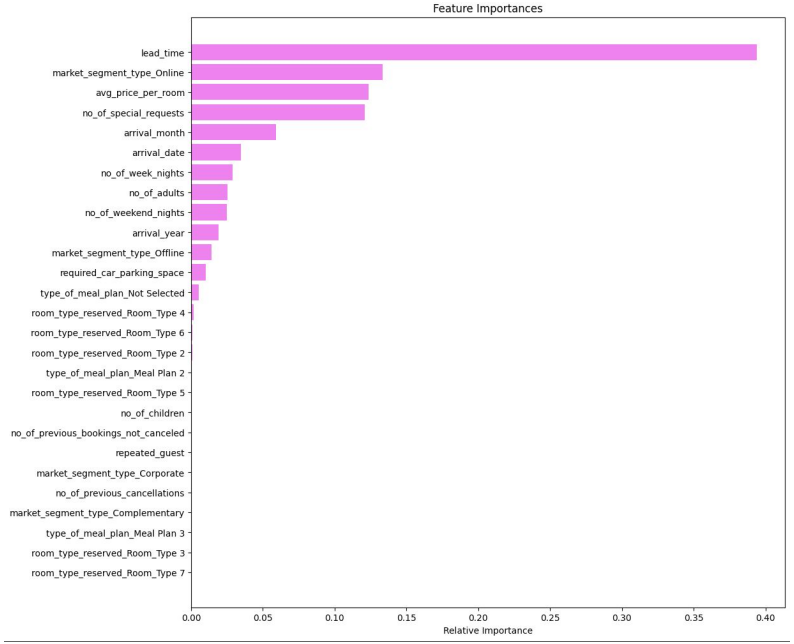
- The feature importance of the fine-tuned Decision Tree Classifier is the following:



- Important features: Lead time, online market segment type, number of special requests, average price per room, number of adults, number of weekend nights.

Model Performance Summary

- The feature importance of the post-pruning Decision Tree Classifier is the following:



- Important features: Lead time, online market segment type, average price per room, number of special requests, arrival month, arrival date, number of week nights, number of adults, number of weekend nights.

Model Performance Summary

- The results on the training data are the following:

	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

- The results on the test data are the following:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86879
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

- Therefore, the post-pruning Decision Tree Classifier manages to tackle the overfitting issue and achieves the highest overall performance on the test data with an F1-Score of approximately 81%. The most important features are: Lead time, online market segment type, average price per room, number of special requests, arrival month.

Executive Summary

- The executive summary (insights and recommendations) of the INN Hotels Supervised Machine Learning presentation is as follows:
    - **Model of Choice:** A post-pruning Decision Tree Classifier has the ability to help the business predict the bookings with high risk of cancellation, and the factors that lead to it.
    - **Lead Time Matters:** This was the most important factor for cancellations. The company should send a reconfirmation request, as well as reminders, close to the date of the booking.
    - **Online Bookings:** Online bookings have a higher risk of being cancelled. Thus, cancellation policies should be stricter, with less refund offers, when it comes to online bookings.
    - **Length of Stay:** When the length of stay is over 11 days, the risk of cancellation is high. A limit of 10 days should be applied, with the ability to extend the stay with a new booking.
    - **Loyalty Program:** Repeated guests have a much lower chance to cancel their booking. Thus, a carefully designed loyalty program should be launched in order to retain customers and turn them into repeated guests.

Thank you for your time!

Michail Mersinias

04/13/2023