INN HOTELS PROJECT

COURSE TITLE: SUPERVISED LEARNING – CLASSIFICATION

DATE: JUNE 3RD, 2022

CONTENTS

BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

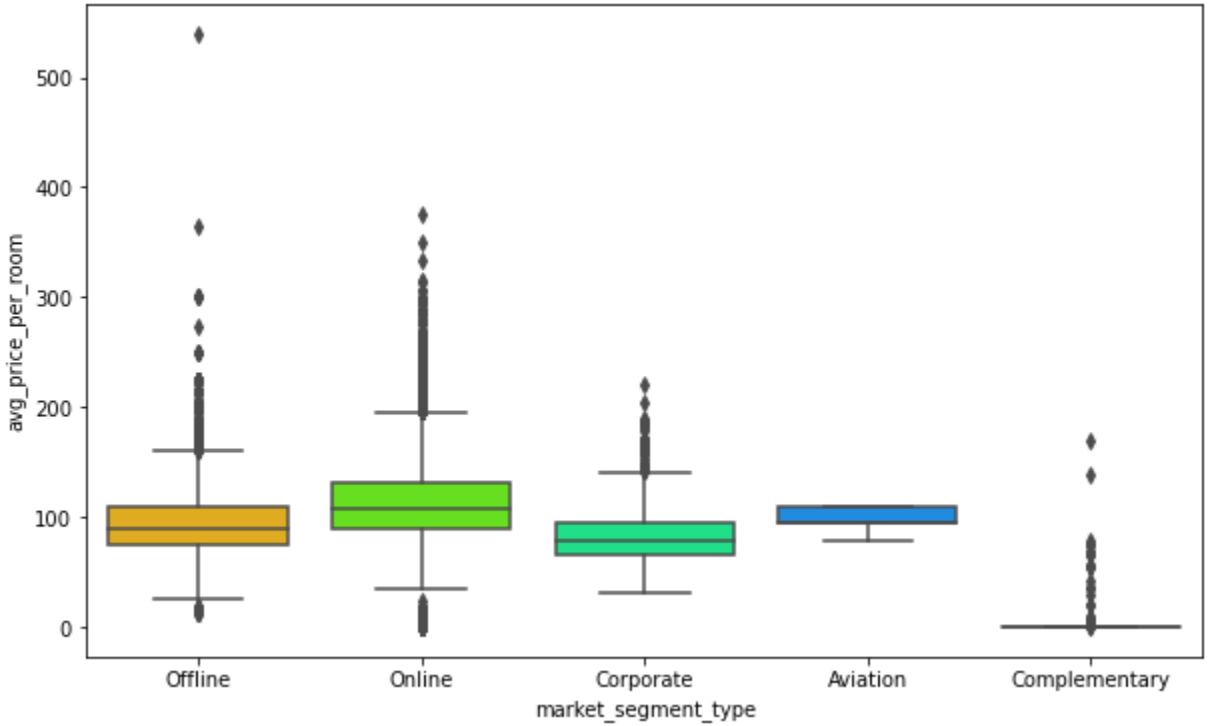
- A significant number of hotel bookings are called-off due to cancellations or no-shows due to change of plans, scheduling conflicts, etc
- These cancellations are done free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and a revenue-diminishing factor for INN hotel and such losses are particularly high on last-minute cancellations.
- The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior
- To analyze the data, 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

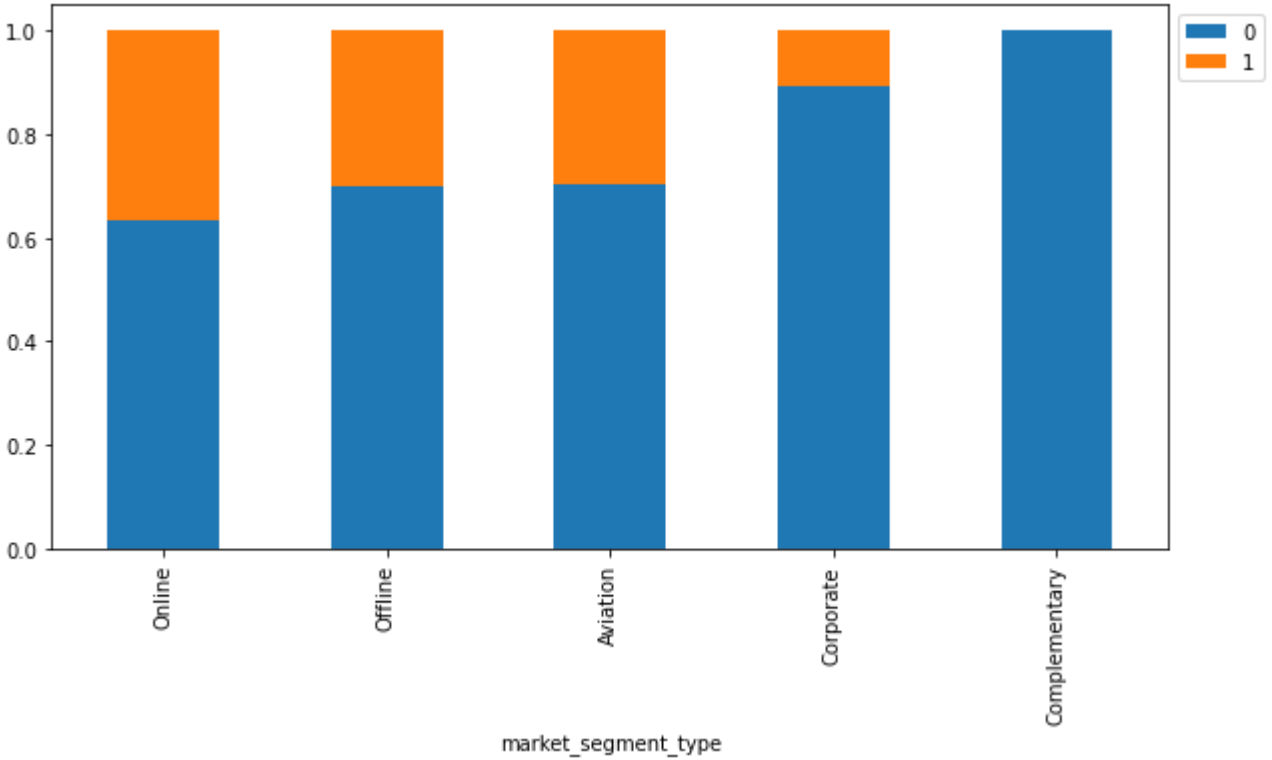
- The data set consist of 36,275 rows and 19 columns
- Most of the columns in the data are numeric in nature (integer or float).
- we have four (5) object type, which means they have text values and 1 float
- there are no missing or duplicated value in the dataset
- the maximum number of previous cancellations is 13
- the maximum number of special request is 5
- the max number of previous bookings not cancelled is 58
- The average lead time is 85 days

EXPLORATORY DATA ANALYSIS(EDA)

- the offline & online market segment has the highest average price per room
- Corporate, aviation and complementary have the lowest

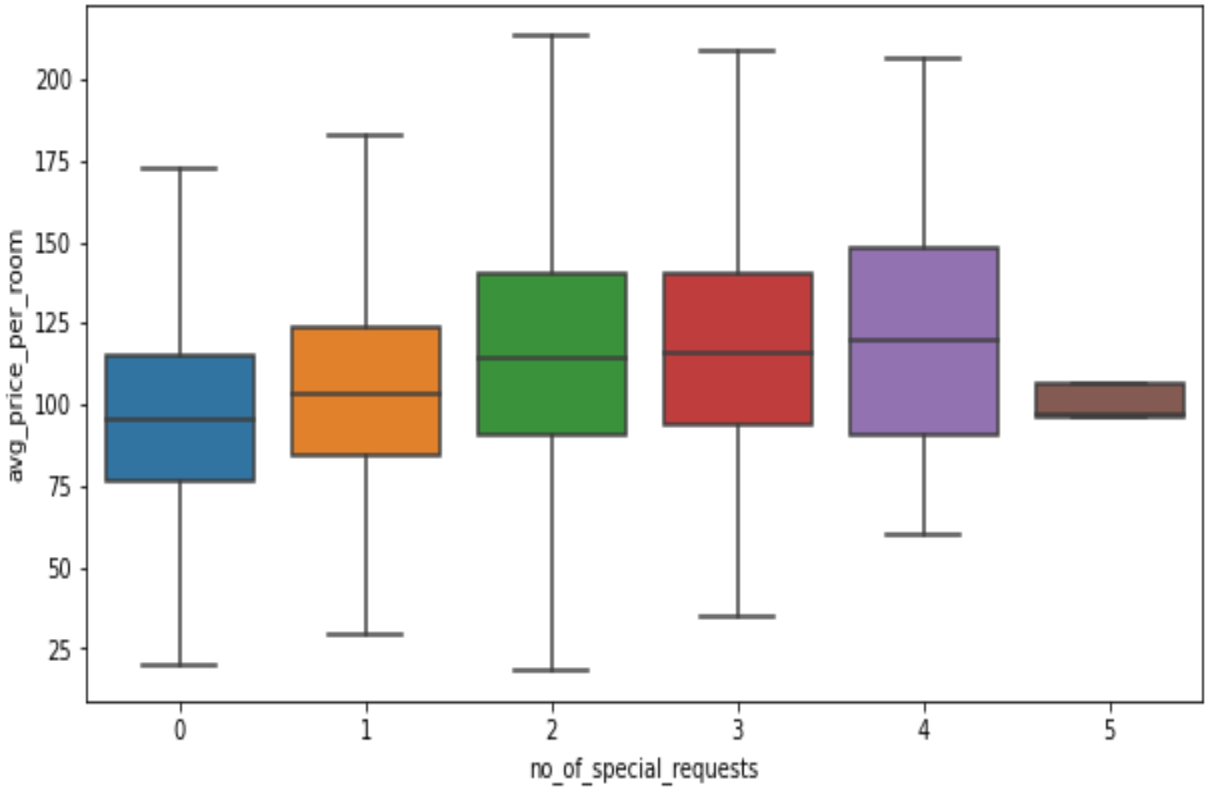


- the offline & online market segment has the highest number of book status
- Corporate, aviation and complementary have a very low booking status

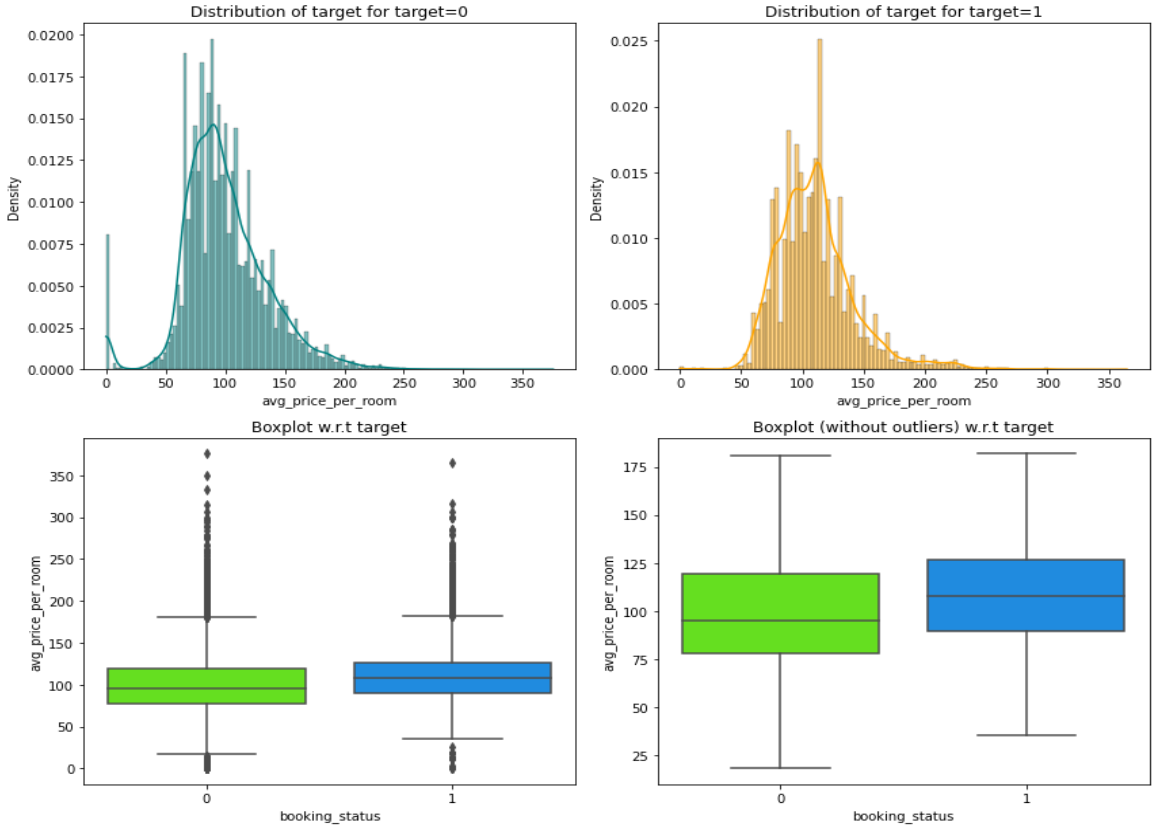


EXPLORATORY DATA ANALYSIS(EDA)

- There is very little difference in the prices of the room that make special request compared to the rooms with no request

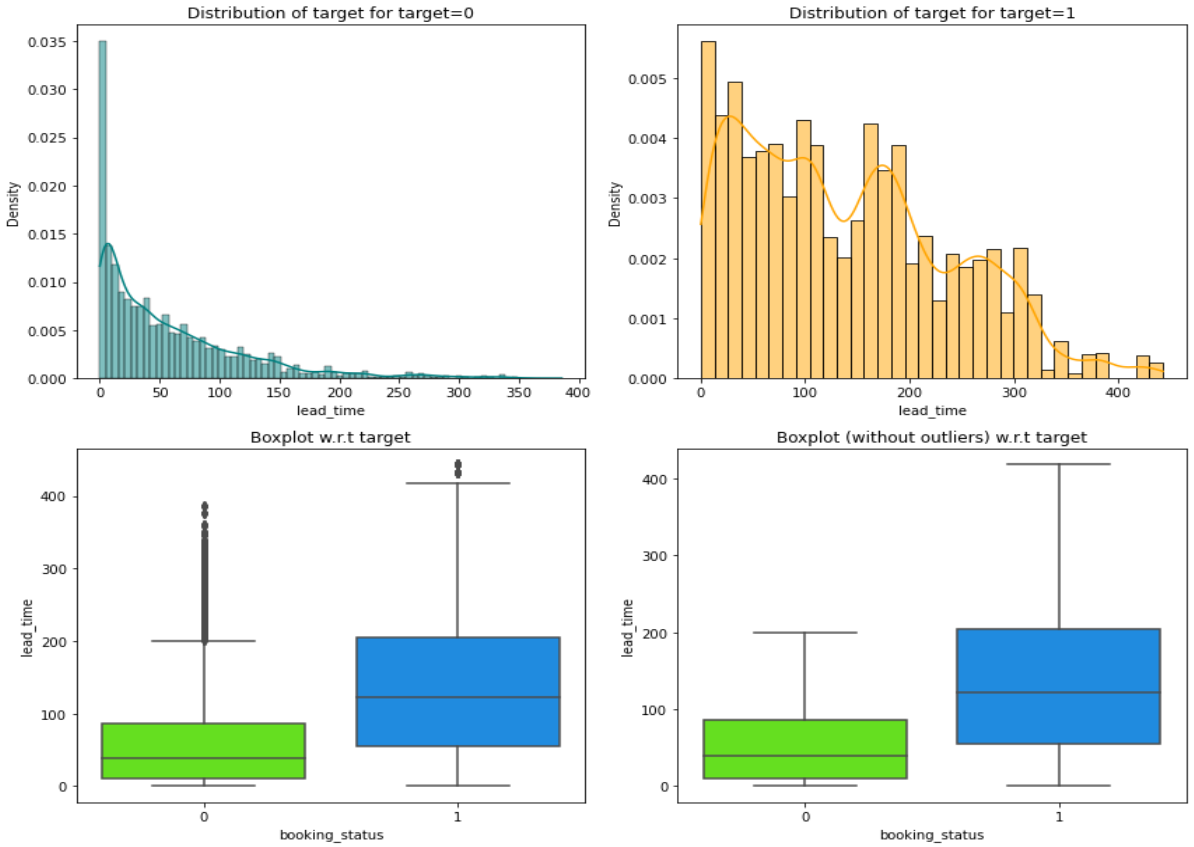


- The average price per room for the bookings canceled and not cancelled are similar
- There are outliers in boxplots of both class distributions

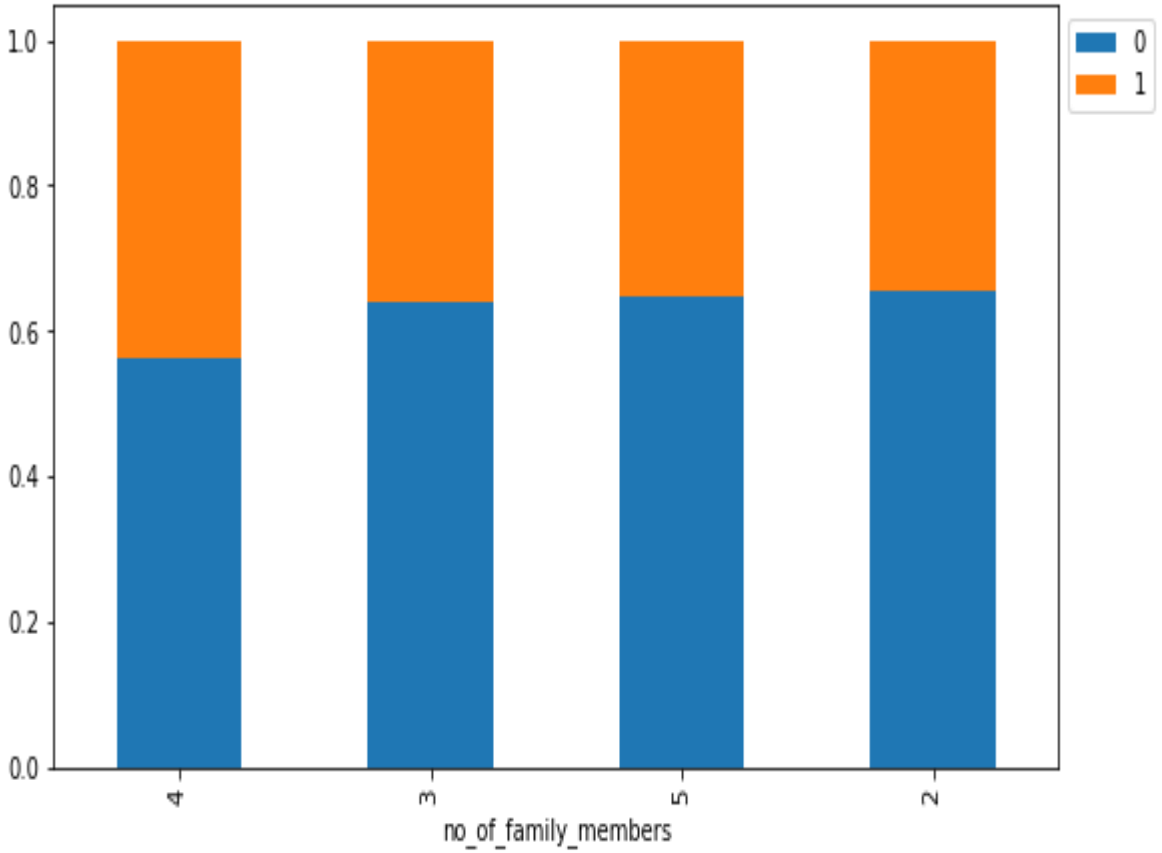


EXPLORATORY DATA ANALYSIS(EDA)

- The difference between the lead time if booking cancelled and not cancelled is higher
- The higher the lead time the higher the probability of a the booking getting cancelled

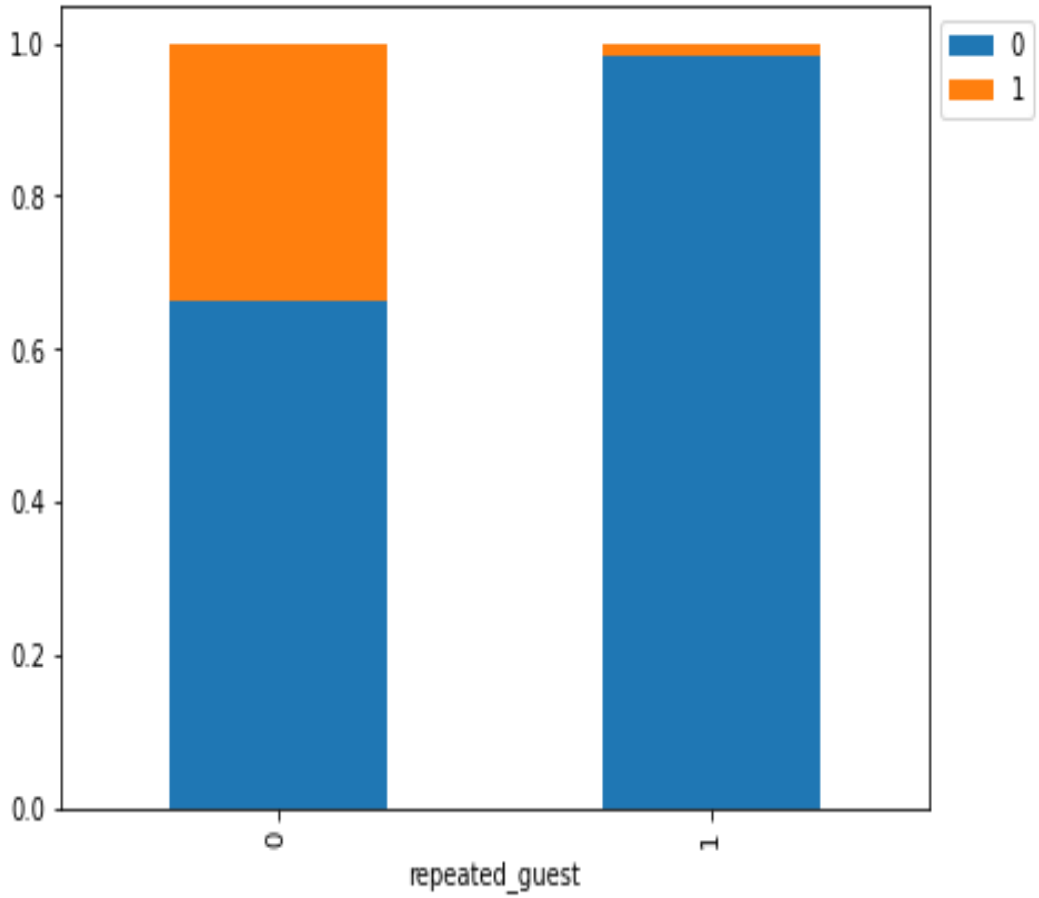


- The family of two have a higher booking status
- The higher the number of family the lower the booking getting cancellled

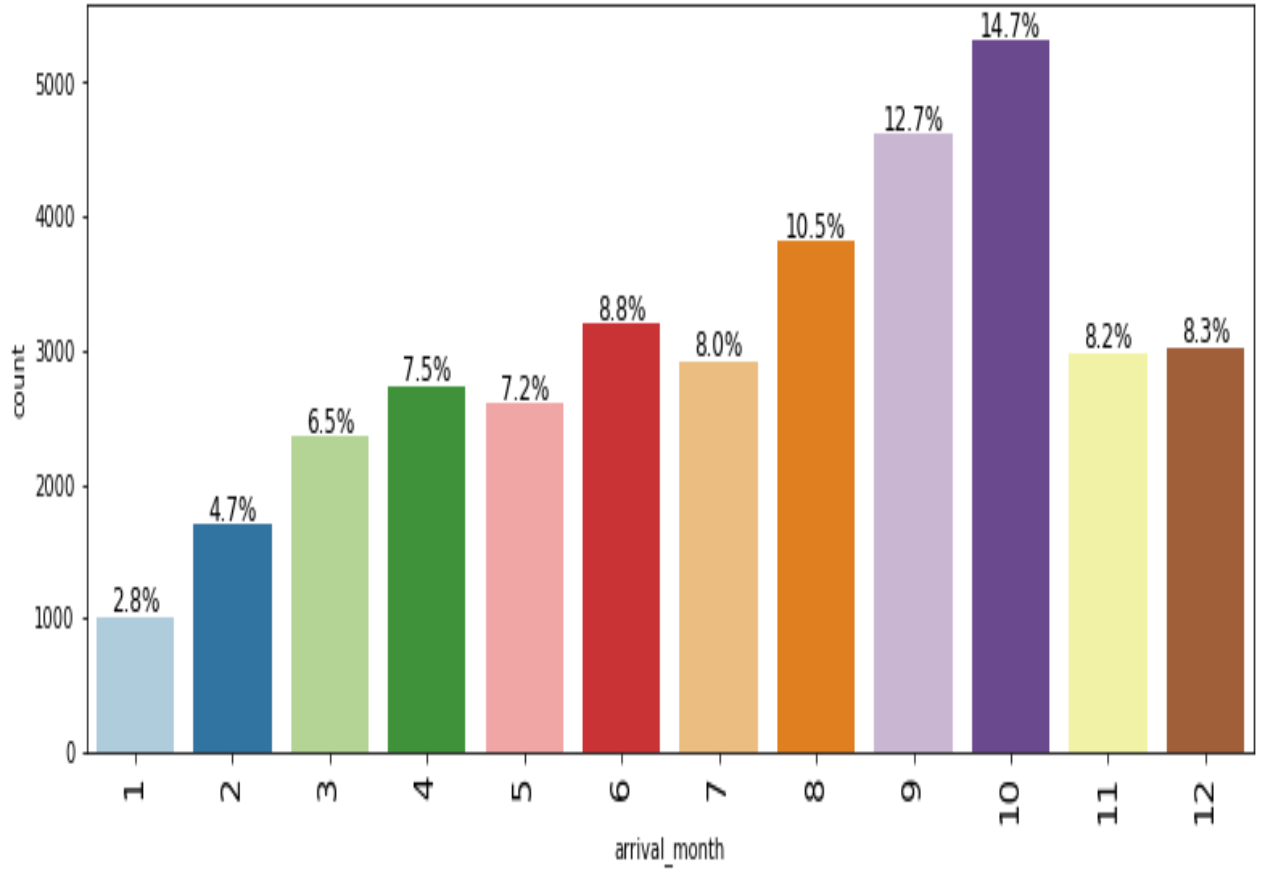


EXPLORATORY DATA ANALYSIS(EDA)

- A very low amount of repeated guest cancelled

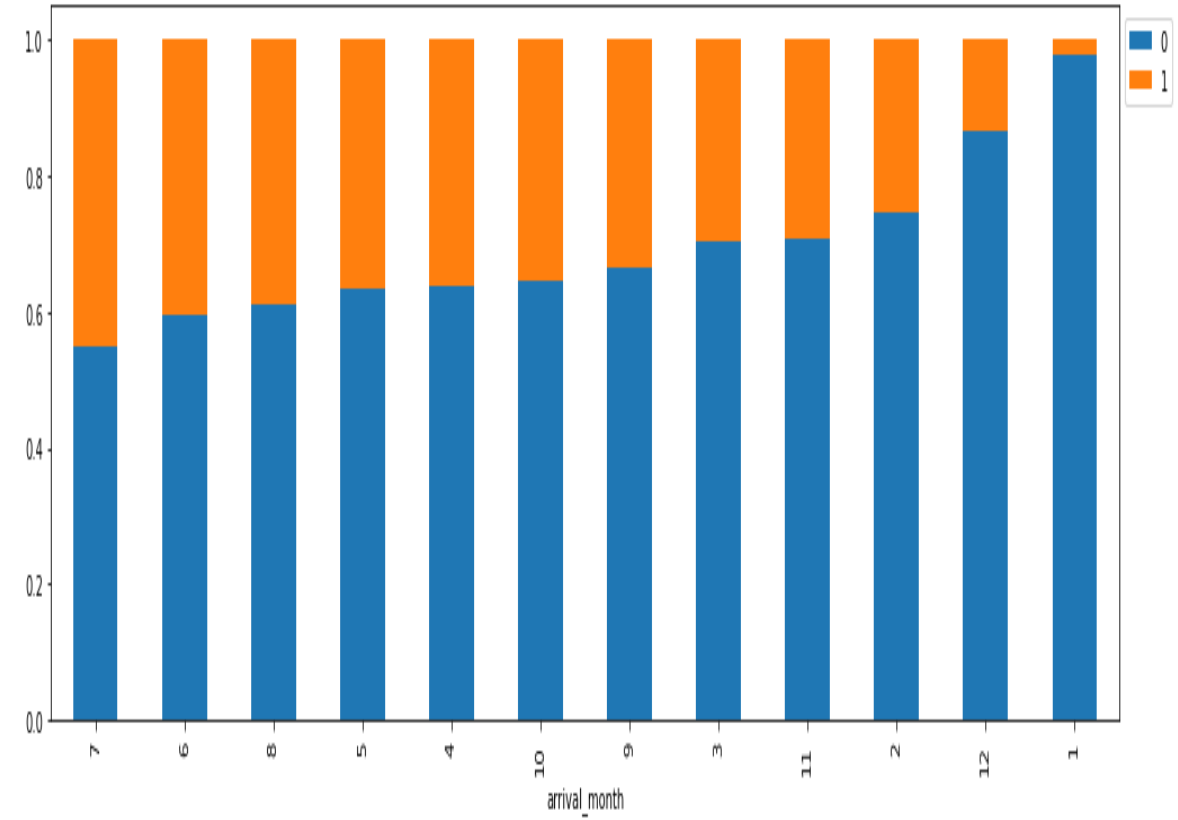


- The busiest month in the hotel October with 14.7% followed by September & august

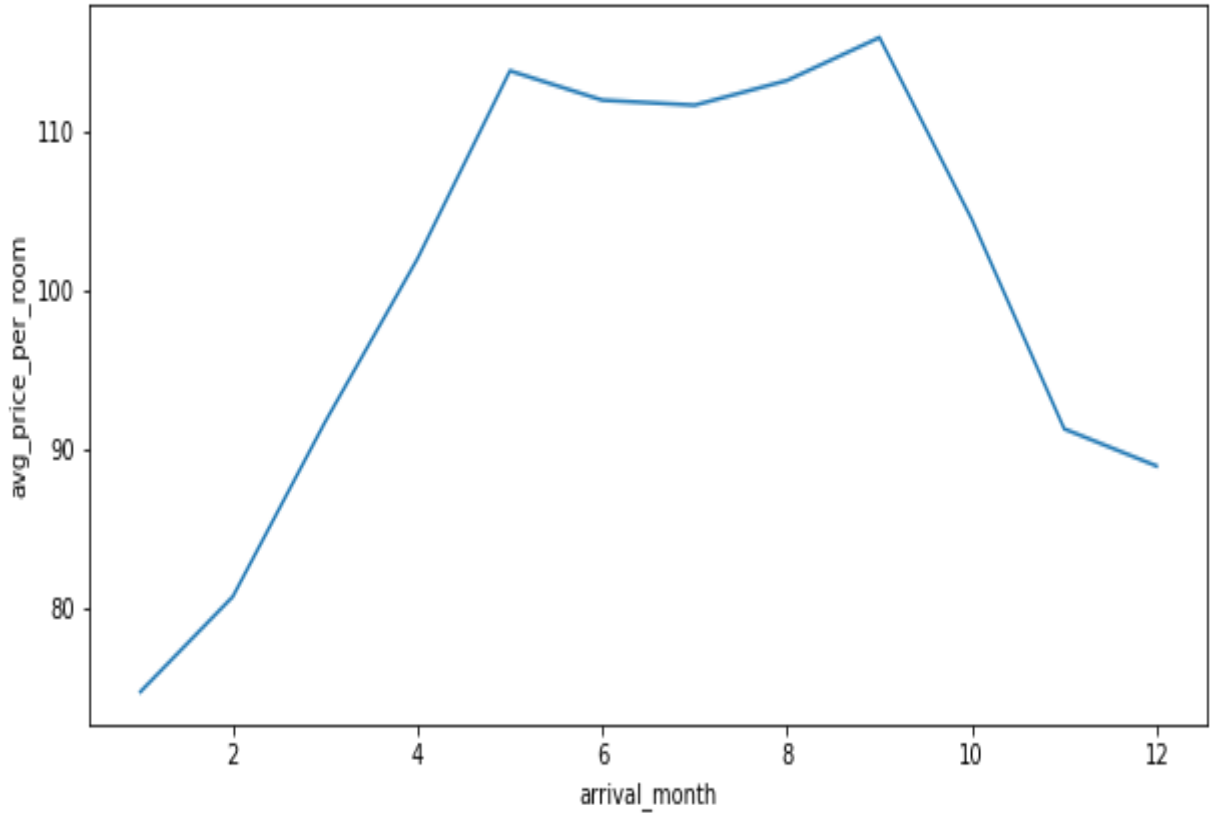


EXPLORATORY DATA ANALYSIS(EDA)

- There is a high percentage of cancellation from june to October
- October has the highest with approximately 16%



- The display shows from may to September, hotel room prices were higher



MODEL PERFORMANCE SUMMARY

- Build a predictive model that can predict which booking to be cancelled in advance
- The decision tree is giving the highest recall on the training set - 0.98
- Using the AUC-ROC curve, there is a significant increase in recall and decrease in precision in the training set compared to the initial model which means we can predict cancellation but if it doesn't get cancelled we lose customers
- Using the optimal threshold curve, there is a decrease in recall and increase In precision in the training set compared to the initial model which means we cannot predict cancellation but if booking get cancelled we loose profit and resources
- The decision tree pre-pruning shows improvements on all ranges compared to the initial model which gives the best model to predict if a booking is going to be cancelled in advance
- The decision tree post pruning is giving an increase in both recall and precision compared to pre-pruning

MODEL PERFORMANCE SUMMARY

Model	Train accuracy	Test Accuracy	Train recall	Test recall	Train precision	Test precision	Train F1	Test F1
Logistic regression	0.80	0.80	0.63	0.63	0.73	0.73	0.68	0.68
Logistic regression(threshold=0.37)	0.79	0.79	0.73	0.73	0.66	0.66	0.70	0.70
Logistic regression(threshold=0.42)	0.80	0.80	0.69	0.70	0.69	0.69	0.69	0.69
Decision tree	0.99	0.87	0.98	0.81	0.99	0.79	0.99	0.80
Decision tree – pre pruning	0.83	0.83	0.78	0.78	0.72	0.72	0.75	0.75
Decision tree – post pruning	0.89	0.86	0.90	0.85	0.81	0.76	0.85	0.80

BUSINESS INSIGHTS AND RECOMMENDATIONS

The following are the insights the data displayed

- it shows 32.8% bookings were cancelled
- Lead time, market segment(online) and no of special requests had a high influence in booking cancellations

The following will be recommended for INN hotels to help in formulating profitable policies for cancellation;

- Offer non-refundable rates/cancellation fees to guests, this will help reduce cancellations and also improve customers commitment especially on the online market segment
- Send guest email reminders about the bookings, from the data it shows the average lead time is 85 days, a monthly reminder, 7 days daily reminder to the arrival date
- The data also shows guests who have 1 or 2 special requests are more likely to cancel, a higher charge should be implemented when a guest ask for this requests

BUSINESS INSIGHTS AND RECOMMENDATIONS

- August , September, & October experienced a higher number of booking cancellation, INN hotels should check the cancellation behavior in this region comparing with competitive markets. This can help understand the problem and if there is not issue with the region, INN hotels should look inward to reduce impact on these months