INN Hotels Project Business Presentation

Contents

Business overview Problem and Solution Approach

- 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:

1.Loss of resources (revenue) when the hotel cannot resell the room.

2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.

3.Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.

4.Human resources to decide for the guests.

Objective

- Statistical and Exploratory Data Analysis of the data provided by the INN hotel management .
- To build a machine learning model to discover the patterns in the user data and the make predictions based on these and intricate patterns for answering business questions.
- We majorly focus on building 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

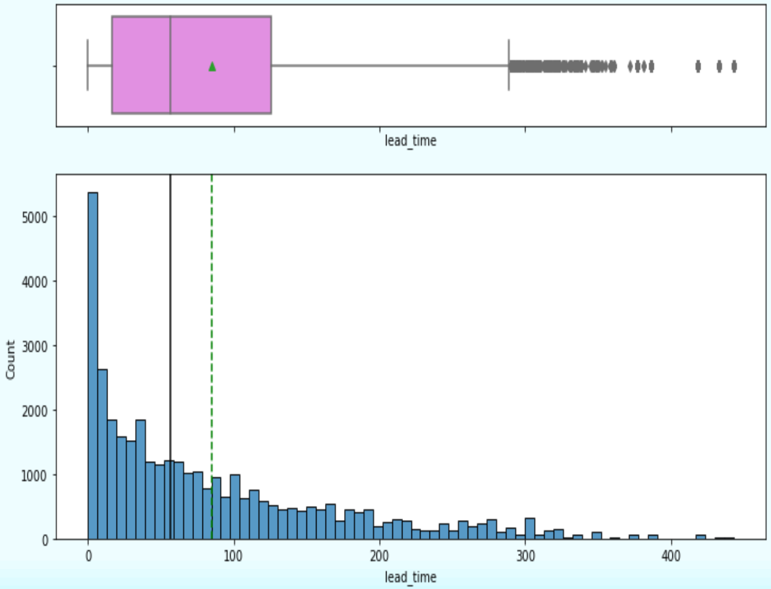
Observations	Variables
36275	19

- Booking_ID: 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 weeknights: 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.
- 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.

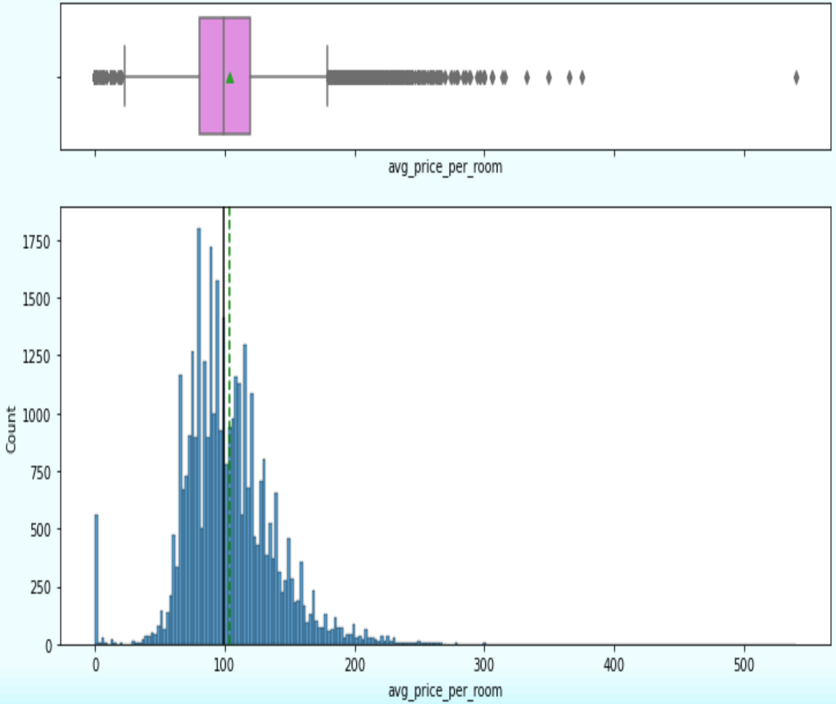
Exploratory Data Analysis (EDA)

Univariate Analysis of lead time



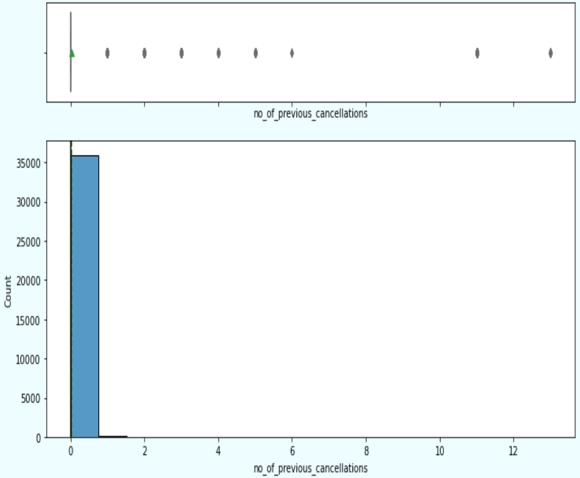
- The distribution looks right skewed.
- The mean is greater than the median.
- There are many outliers as we can clearly identify from the boxplot-histogram.
- The lead time is the number of days between the booked and the arrival date, we can clearly see that maximum bookings are done very close to the arrival date .
- This might play an important role in our statistical analysis.

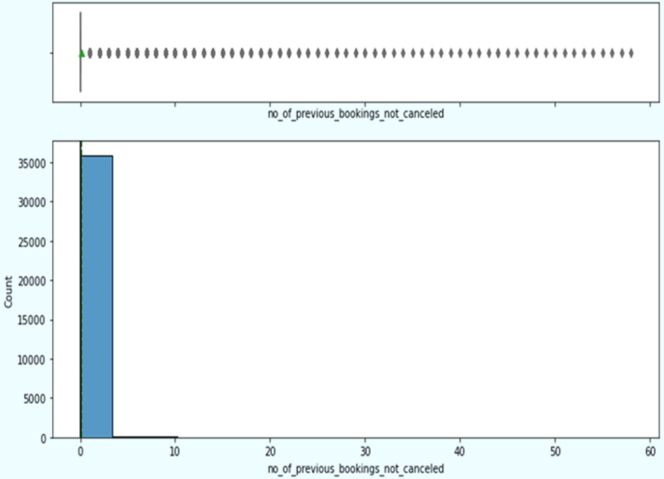
Univariate Analysis – Average price per room



- The mean is very close to the median, the distribution is almost normal .
- There are many outliers.
- There are a few outliers with average price per room as 0, whose Market segment type shows them as :
  - Complimentary
  - Online
- We can calculate the IQR and replace the outliers with the upper whisker for further evaluation.

Univariate analysis-no of previous cancellations no of previous bookings not cancelled

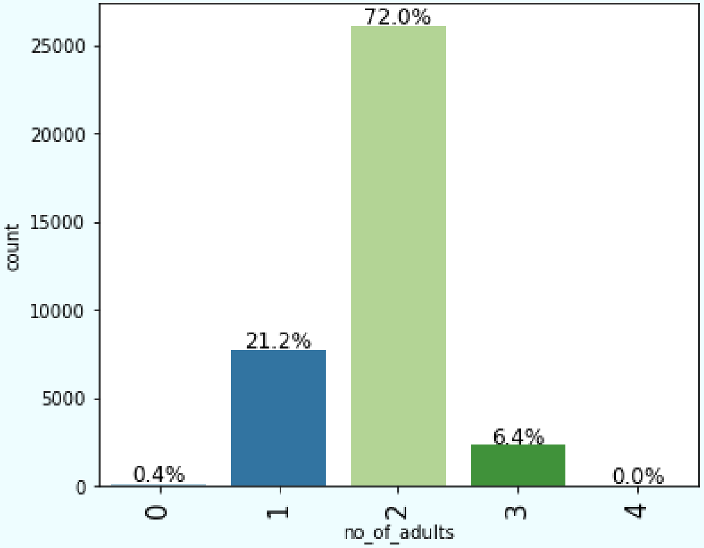


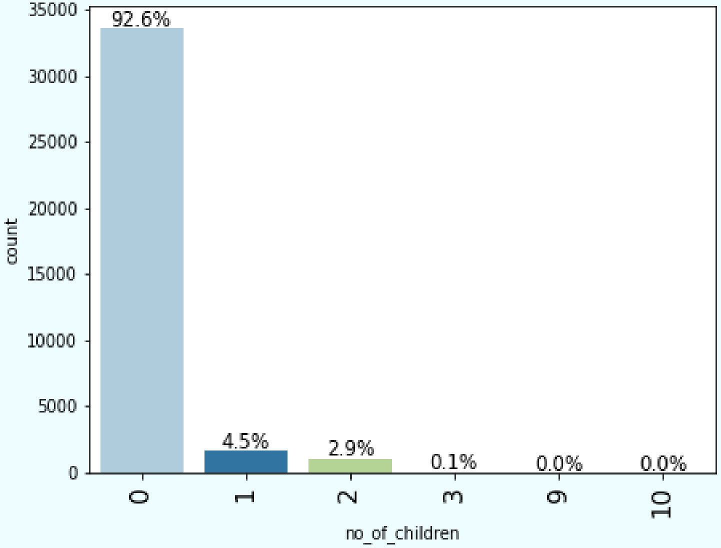


- The mean and the median are 0.
- The analysis clearly shows that the no of previous bookings cancelled by the customer prior to the latest bookings are mostly 0 except for some outliers.

- The mean and the median are 0.
- There are many outliers.
- The number of previous bookings not cancelled by the customer prior to the current booking are mostly 0.

Univariate analysis –no of adults, no of children

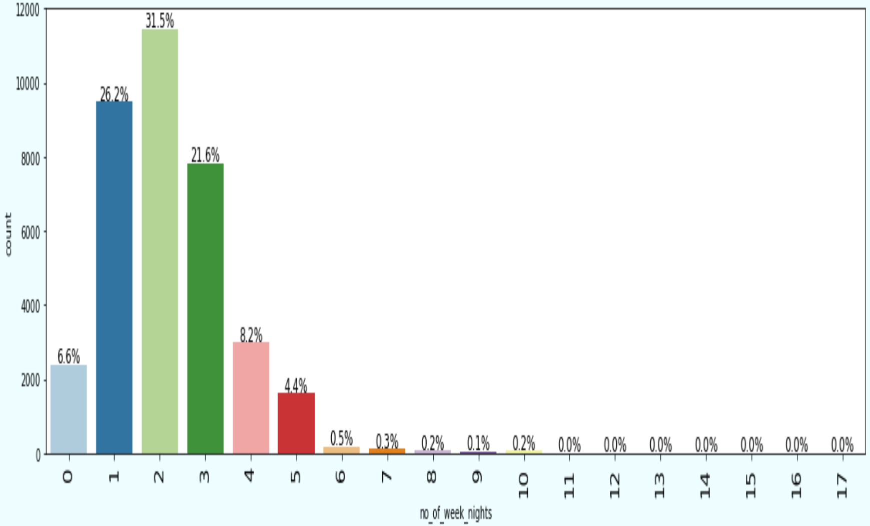


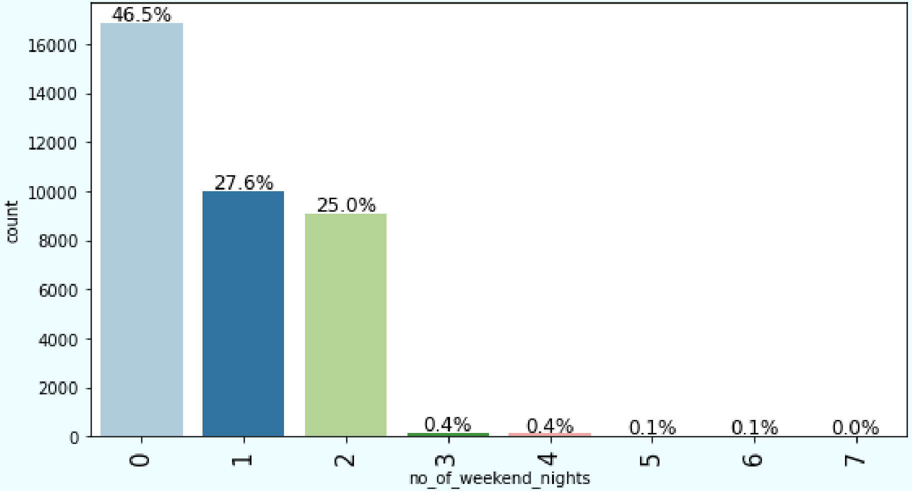


- From the total population, the number of adults (2) has the maximum percentage 72%.
- The percentage of people traveling alone is 21.2% ,but greater than 3 people 6.4%.

- The number of children is 0, nearly 92.6%.
- The people who book the hotel are not accompanied by children most of the time.

Univariate analysis-no of weeknights, no of weekend nights

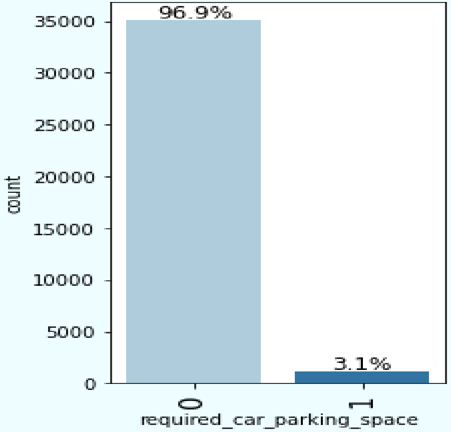


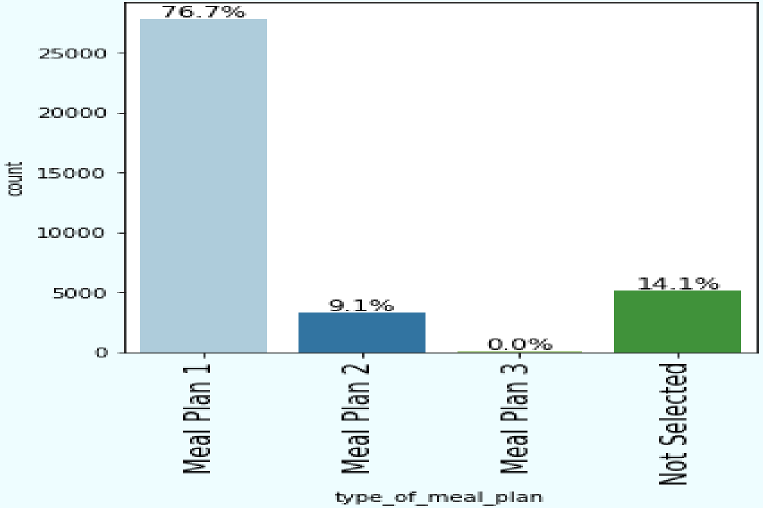


- 31.5% of the customers booked 2 nights during weekdays, 26.2% of the customers booked 1 night during weekday and 21.6% booked for 3 nights.
- There is more business for the hotel during the weekdays than in the weekend .

- 46.5% of the weekend nights are not booked , 27.6% booked 1 night on a weekend and 25% for 2 nights.
- Definitely, the hotel business must improve during the weekends.

Univariate analysis-required car parking space, type of meal plan

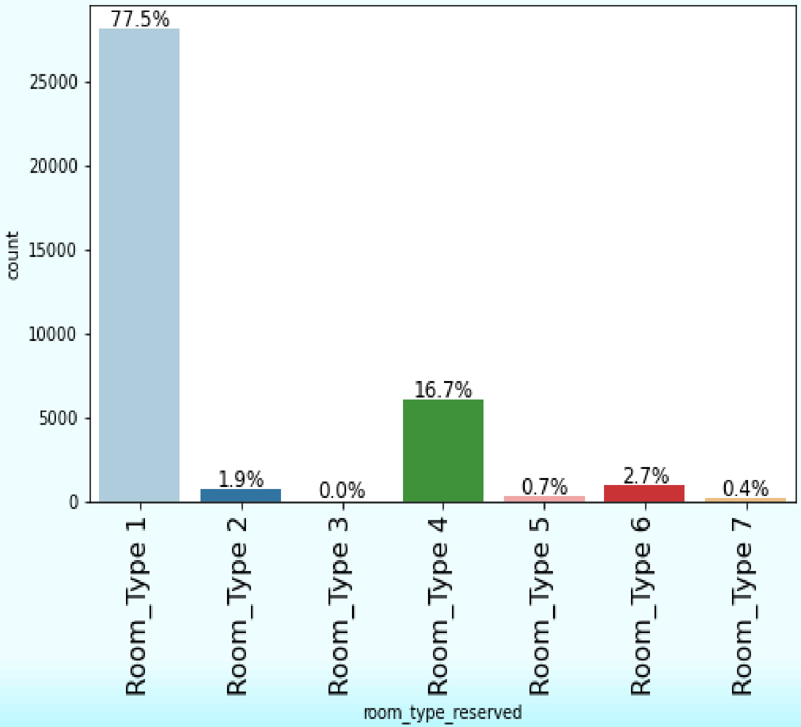




96.9% of the customers does not require car parking , maybe they flied from different places.

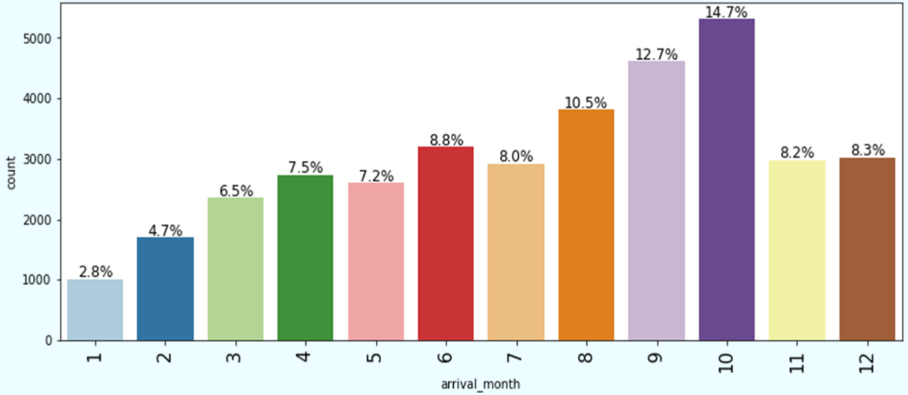
- 76.7% of the customers chose Meal plan 1, which includes only breakfast.
- 9.1% chose half board and 14.1% did not select any option.
- Breakfast seems to be the most preferred meal .

Univariate analysis –room type reserved

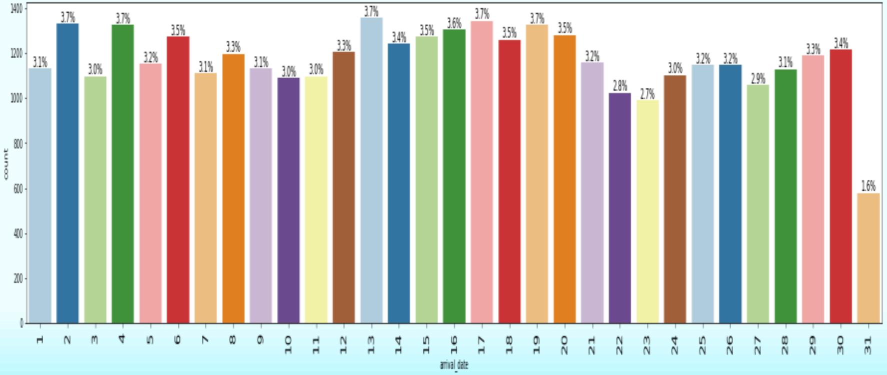


- There are 7 types of rooms available in the INN hotels for the customers to select.
- Room type 1 ,is the most preferred room by the customers.
- 77.5% of the customers selected room type 1, maybe it is most affordable and convenient.
- Room type 4 is the next preferred one, maybe it is a family room or has bunk beds for children.

Univariate analysis- arrival month and arrival date

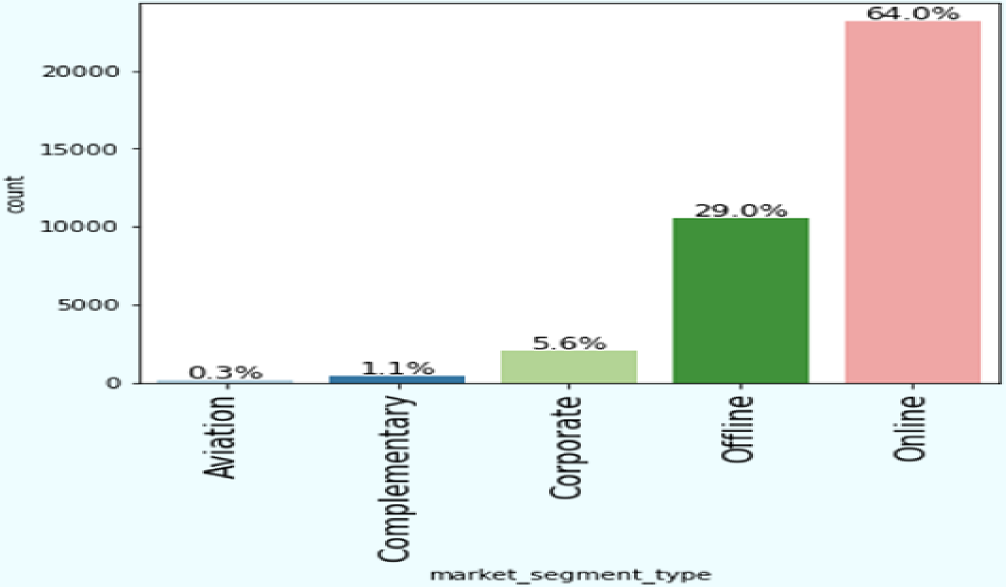


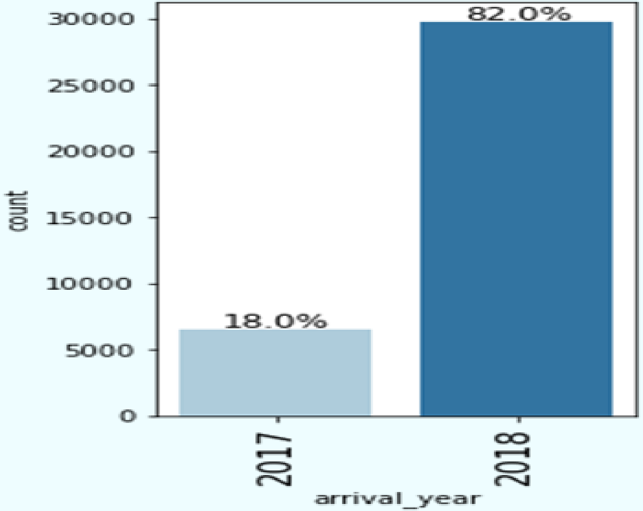
- October has the highest number of bookings, 14.7%
- September stands next with 12.7% followed by August 10.5%.
- All other months don’t differ much.
- Special offers may be one of the reason for attracting more customers in October.



- The arrival date is almost evenly distributed.
- It does not contribute much for our analysis.

Univariate analysis-market segment type and arrival year

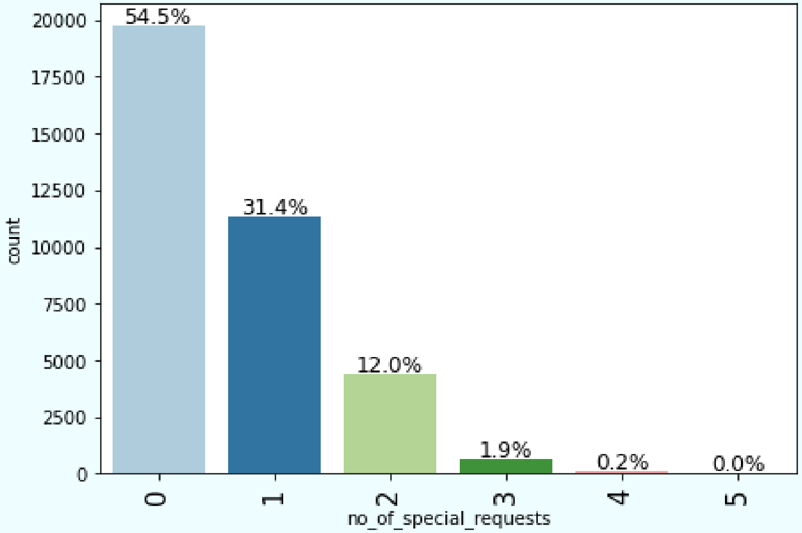




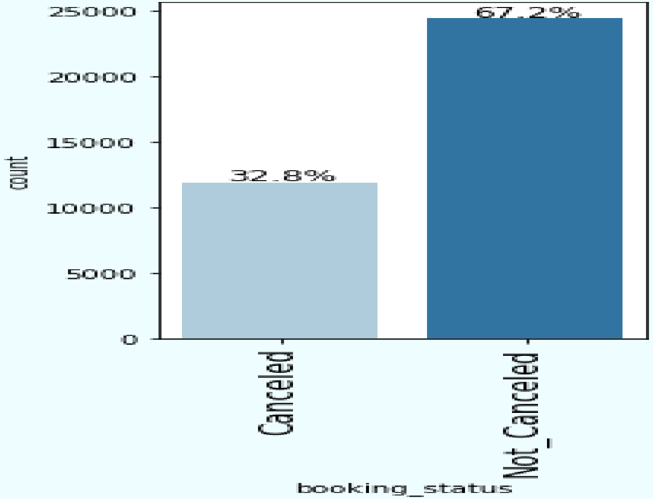
- 64% of the bookings are made online, 29% are made offline.
- Corporate bookings contribute only 5.6% followed by complimentary which is only 1.1%.

82% of the bookings are made in 2018 and 18% in 2017.it does not contribute much for our analysis.

Univariate analysis- no of special requests, booking status



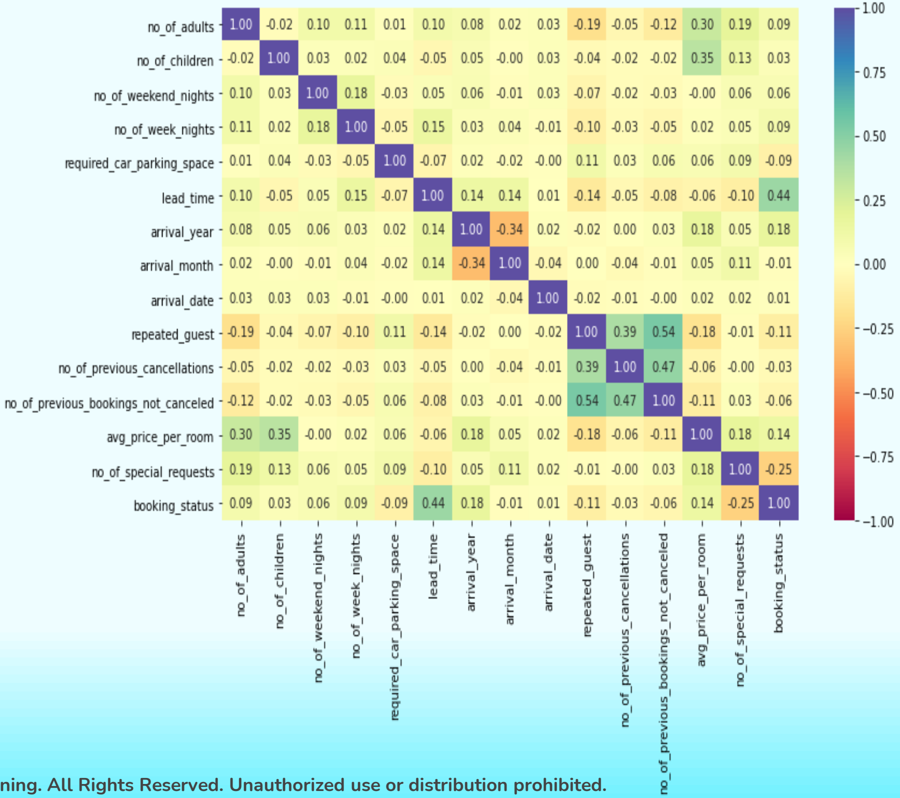
- 54.5% of the customers made no special requests.
- 31.4% of customers made only 1 special request and 12% made 2 special requests.
- The special requests can be for additional bed, view from the room etc.



- Booking status is the target variable, which mainly matters for our business problem.
- As we can see, the number of cancellations are less(32.8%) compared to not cancelled.

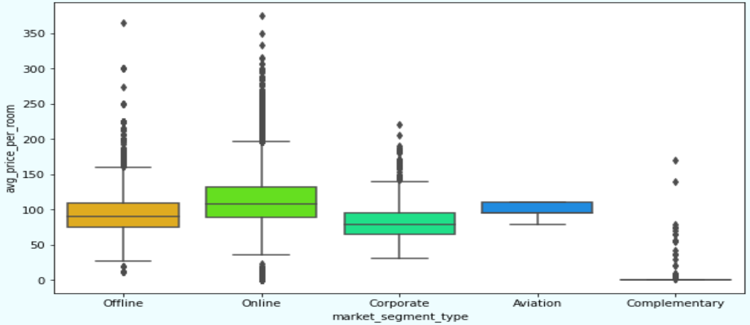
Bivariate analysis

- Bivariate analysis using heat map for the numerical variables of the given data set clearly shows , there is no high correlation between any of the variables.
- There is positive correlation between lead time and booking status.
- There is a positive correlation between booking status and average price per room.



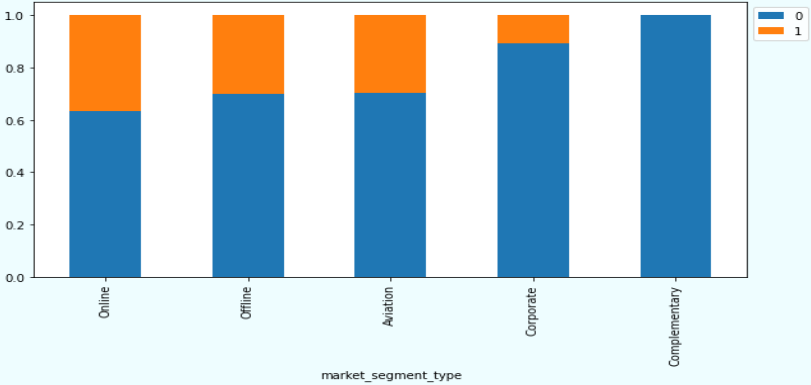
	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
no_of_adults	1.00	-0.02	0.10	0.11	0.01	0.10	0.08	0.02	0.03	-0.19	-0.05	-0.12	0.30	0.19	0.09
no_of_children	-0.02	1.00	0.03	0.02	0.04	-0.05	0.05	-0.00	0.03	-0.04	-0.02	-0.02	0.35	0.13	0.03
no_of_weekend_nights	0.10	0.03	1.00	0.18	-0.03	0.05	0.06	-0.01	0.03	-0.07	-0.02	-0.03	-0.00	0.06	0.06
no_of_week_nights	0.11	0.02	0.18	1.00	-0.05	0.15	0.03	0.04	-0.01	-0.10	-0.03	-0.05	0.02	0.05	0.09
required_car_parking_space	0.01	0.04	-0.03	-0.05	1.00	-0.07	0.02	-0.02	-0.00	0.11	0.03	0.06	0.06	0.09	-0.09
lead_time	0.10	-0.05	0.05	0.15	-0.07	1.00	0.14	0.14	0.01	-0.14	-0.05	-0.08	-0.06	-0.10	0.44
arrival_year	0.08	0.05	0.06	0.03	0.02	0.14	1.00	-0.34	0.02	-0.02	0.00	0.03	0.18	0.05	0.18
arrival_month	0.02	-0.00	-0.01	0.04	-0.02	0.14	-0.34	1.00	-0.04	0.00	-0.04	-0.01	0.05	0.11	-0.01
arrival_date	0.03	0.03	0.03	-0.01	-0.00	0.01	0.02	-0.04	1.00	-0.02	-0.01	-0.00	0.02	0.02	0.01
repeated_guest	-0.19	-0.04	-0.07	-0.10	0.11	-0.14	-0.02	0.00	-0.02	1.00	0.39	0.54	-0.18	-0.01	-0.11
no_of_previous_cancellations	-0.05	-0.02	-0.02	-0.03	0.03	-0.05	0.00	-0.04	-0.01	0.39	1.00	0.47	-0.06	-0.00	-0.03
no_of_previous_bookings_not_canceled	-0.12	-0.02	-0.03	-0.05	0.06	-0.08	0.03	-0.01	-0.00	0.54	0.47	1.00	-0.11	0.03	-0.06
avg_price_per_room	0.30	0.35	-0.00	0.02	0.06	-0.06	0.18	0.05	0.02	-0.18	-0.06	-0.11	1.00	0.18	0.14
no_of_special_requests	0.19	0.13	0.06	0.05	0.09	-0.10	0.05	0.11	0.02	-0.01	-0.00	0.03	0.18	1.00	-0.25
booking_status	0.09	0.03	0.06	0.09	-0.09	0.44	0.18	-0.01	0.01	-0.11	-0.03	-0.06	0.14	-0.25	1.00

Hotel rates are dynamic and change according to demand and customer demographics. Let's see how prices vary across different market segments



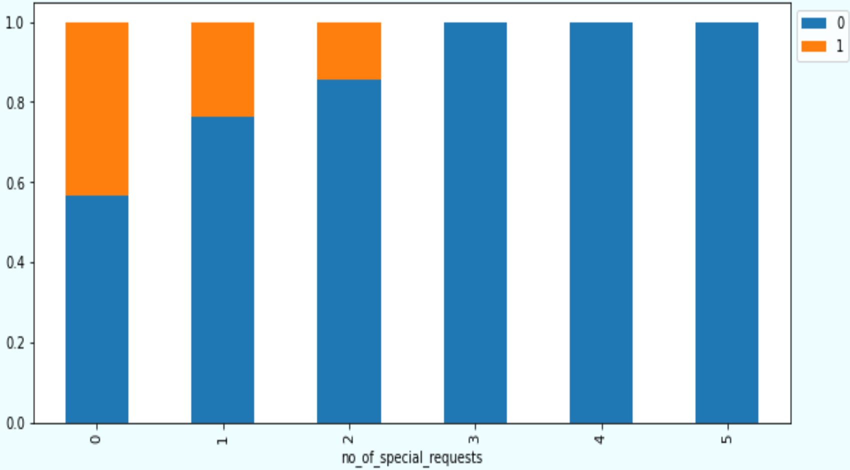
- Average price per room looks higher for online and aviation reservations
- There are less complimentary reservations
- The price of offline and corporate reservations looks almost similar.
- The aviation ,market segment type has no outliers, looks like their prices are fixed.
- All the other segments have many outliers.

Let's see how booking status varies across different market segments. Also, how average price per room impacts booking status



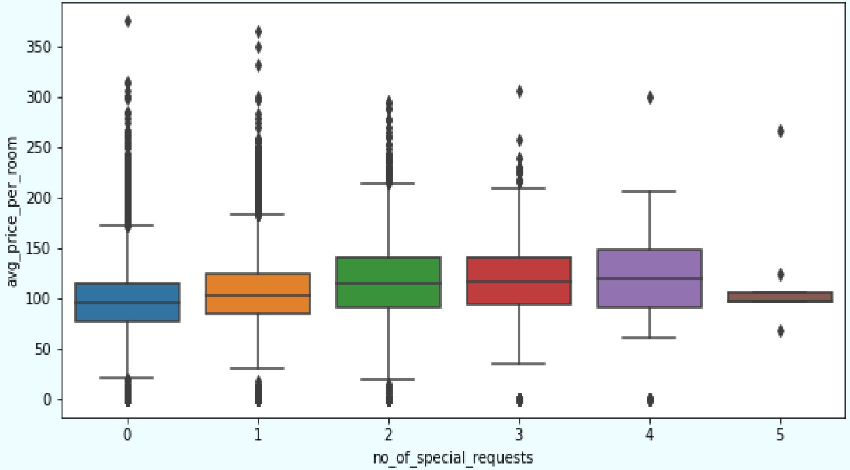
- We have encoded Canceled bookings to 1 and Not Canceled as 0 for our analysis.
- Online segment has the highest number of cancellations followed by offline and aviation.
- Corporate has less cancellations , may be due to business commitments .
- Complimentary segment has 0 cancellations.

Many guests have special requirements when booking a hotel room. Let's see how it impacts cancellations



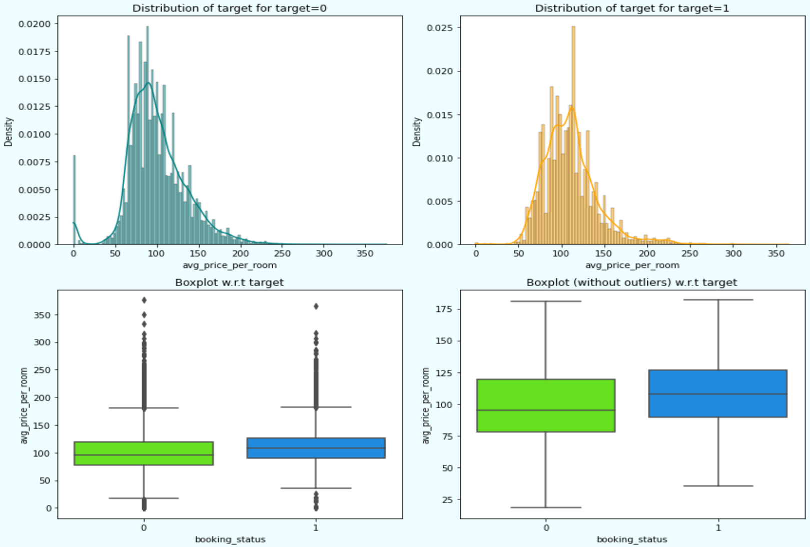
- There are maximum number of cancellations with no special requests.
- Number of cancellations are less as the number of special requests increase; this indicates that customers are quite sure of their travel.

Let's see if the special requests made by the customers impacts the prices of a room



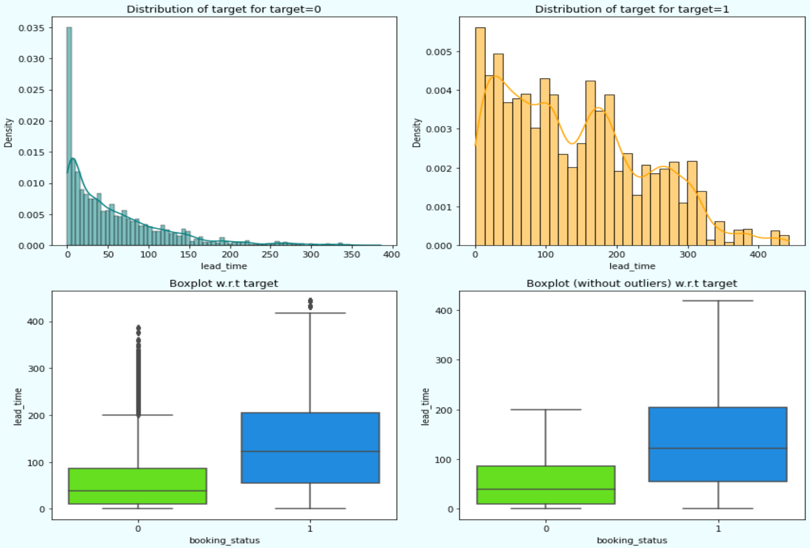
- The average price of the room increase with the number of special requests.
- There are outliers.

Bivariate analysis of booking status and average price per room.



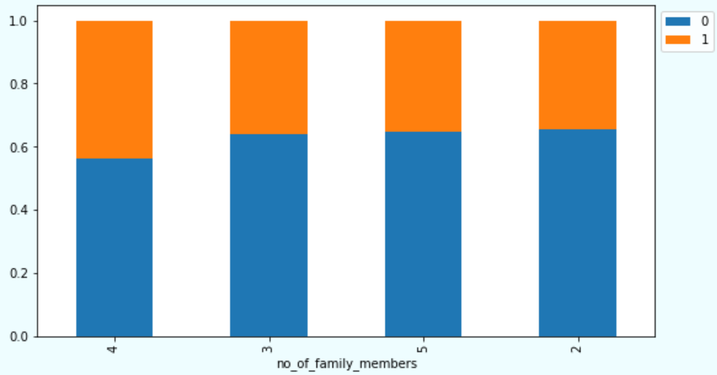
- Analysis with and without outliers, clearly shows that there is a positive correlation between the 2 variables.
- Booking status is cancelled when the average price per room tends to increase.

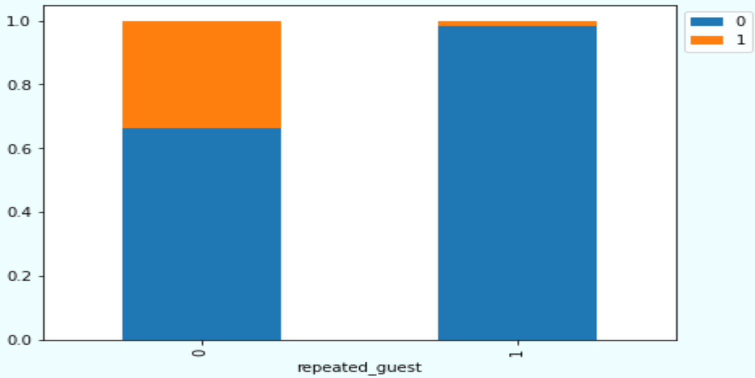
Bivariate analysis of booking status and lead time .



- There is positive correlation between lead time and booking status.
- Number of cancellations are more when the lead time increases.

Bivariate Analysis of Booking Status with the number of people, repeating guests

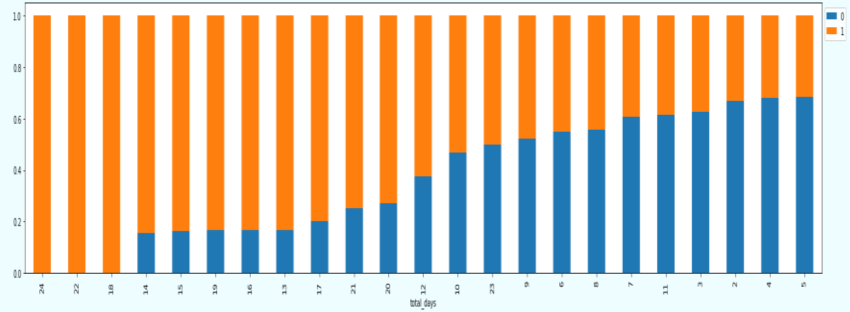




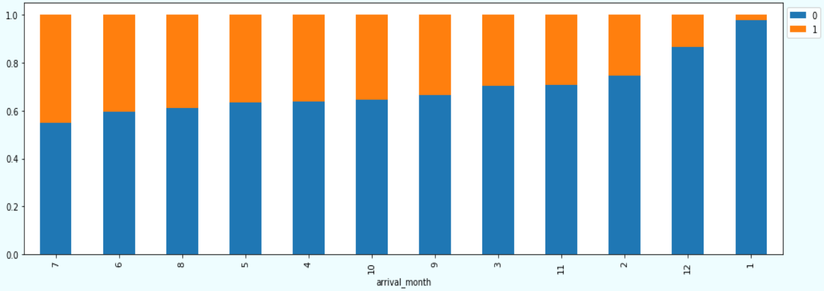
- Generally, people travel with their spouse and children for vacations or other activities. Let's see how customers who traveled with their families impact the booking status.
- The number of cancellations is almost the same with the no of family members.

- Repeating guests are the guests who stay in the hotel often and are important to brand equity
- Repeating guests make less cancellations compared to the nonrepeating guests .

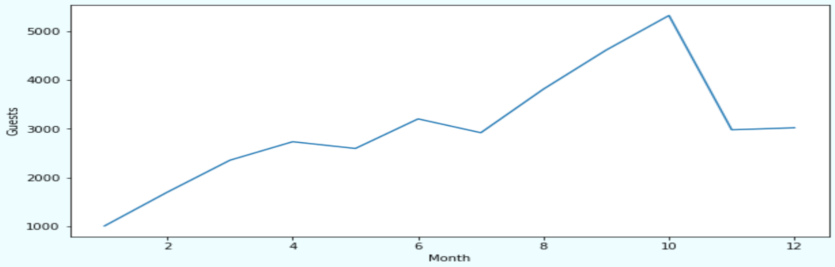
Bivariate analysis continued.



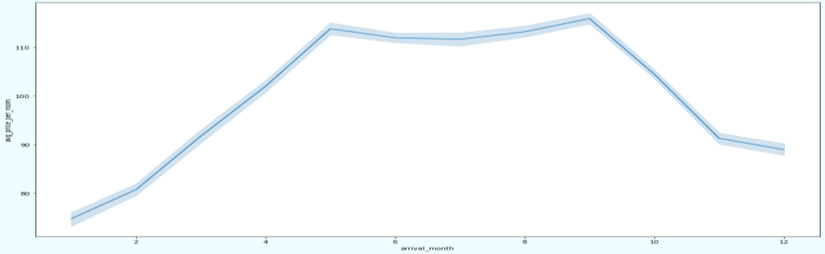
Analysis for the customer who stays for at least a day at the hotel shows that the number of cancellations decreases as the number of days decrease.



Analysis of number of cancellations for each month shows that, January had the least number of cancellations and July had the most .



October seems to be the busiest month at the hotel, and the sales tend to decrease during the year end.



Average price per room looks higher between the month of May and October.

Logistic Regression and Decision tree model Evaluation

We must predict the booking status ,which is a Binary classification and hence we are using Logistic Regression.

General steps:

- split the data into “training” and “test” sets.
- Use regression/classification results from the training set to predict test set.
- Compare predicted values to actual values.

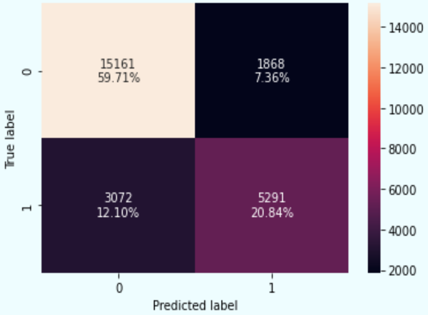
Model can make wrong predictions which leads to
Loss of resources
Damage of brand equity

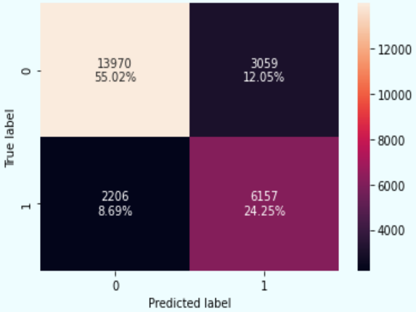
In order to reduce the loses ,F1 score should be maximized, greater the F1 score higher are the chances of minimizing false negatives and false positives.

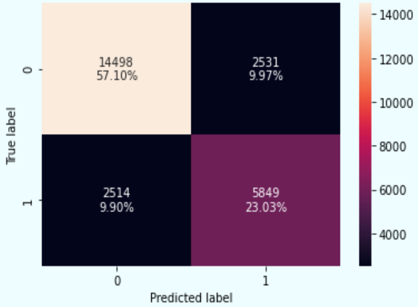
Task	• Model the probability that y(target) belongs to a particular category.
Functions used	•Model_performance_classification_statsmodel •Confusion_matrix_statsmodel •Model_performance_classification_sklearn •Confusion_matrix_sklearn
Performance measures	• Accuracy,Precision,Recall,F1 score, ROC curve ,Confusion Matrix

Logistic Regression Model Performance (Training and Test Data)

Training data



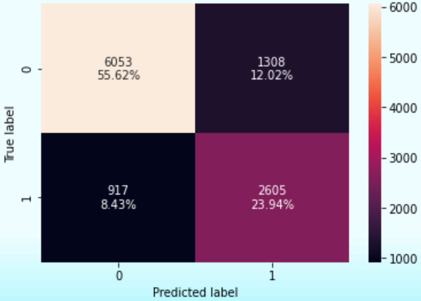


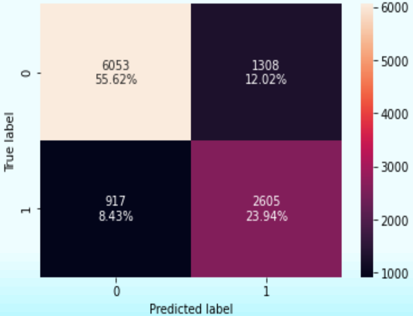


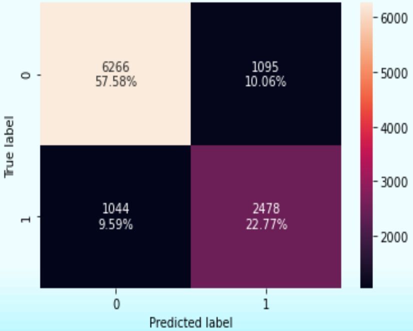
Confusion matrix with optimal threshold(default).

Confusion matrix with optimal threshold of 0.37

Confusion matrix with optimal threshold of 0.42







Test data

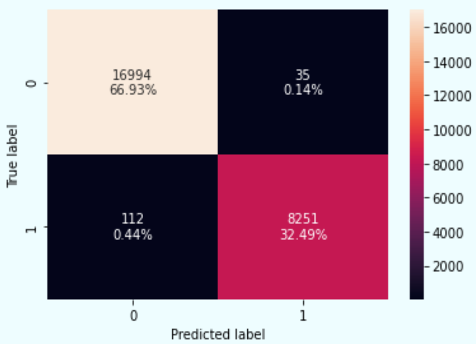
Logistic Regression Model Performance Summary

Training data	Logistic regression (default threshold)	Logistic regression (0.37 threshold)	Logistic regression (0.42 threshold)
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73662	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Test data	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

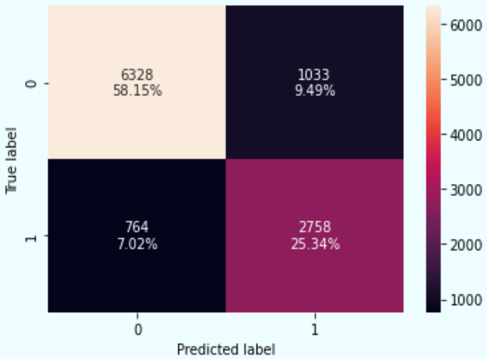
- We have been able to build a predictive model that can be used by INN hotels to predict a customer will cancel their booking with an F1 score of 0.70049 on the training data.
- We have considered F1 score as our performance metric, because both false positives and false positives are important in this case. Logistic regression at threshold of 0.37 tends to perform better .
- All the logistic regression models have given a generalized performance on the training and test set.

Decision Tree Model Performance(before Pruning)



Confusion matrix of the train data

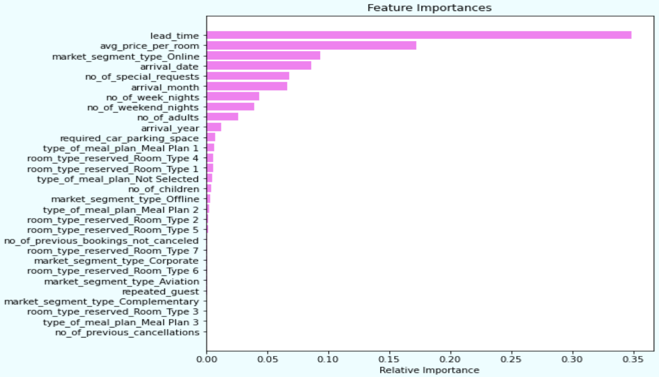
accuracy	Recall	Precision	F1
0.99421	0.98661	0.99578	0.99117



Confusion matrix of the test data

accuracy	Recall	Precision	F1
0.87182	0.80522	0.800	0.80260

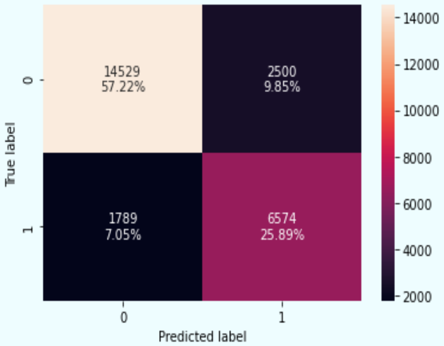
- The decision tree is overfitting and hence it has the perfect accuracy and F1 score.
- The model performance for the test data has reduced , F1 is 0.80 for the test data. The tree needs to be pruned to improve the test data performance.



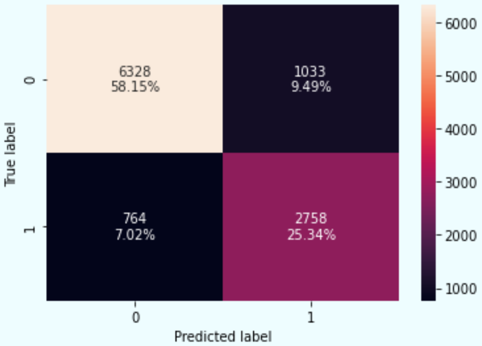
Important Features

- Lead time is the most important feature for building the model.
- Average price per room, market segment type online, arrival date, number of special requests are the important features for building the tree.
- The significance of all other features is negligible.

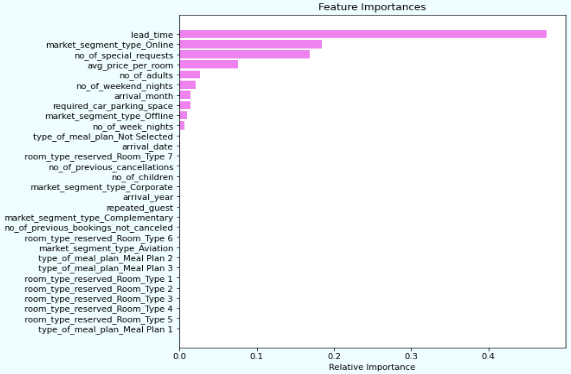
Decision tree Model Performance –Pre-Pruning



Confusion matrix of the train data after pre-pruning



Confusion matrix of the test data after pre-pruning



Important Features after pre-pruning

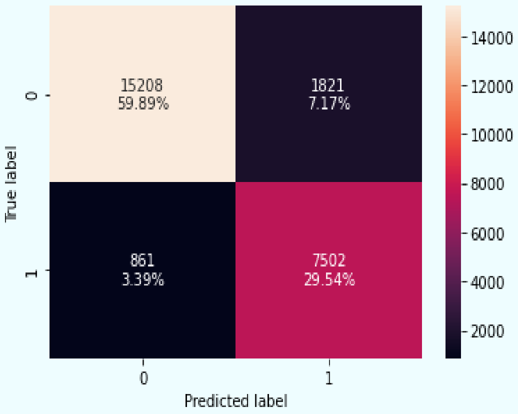
accuracy	Recall	Precision	F1
0.83109	0.78608	0.72449	0.75403

accuracy	Recall	Precision	F1
0.83488	0.78308	0.72751	0.75427

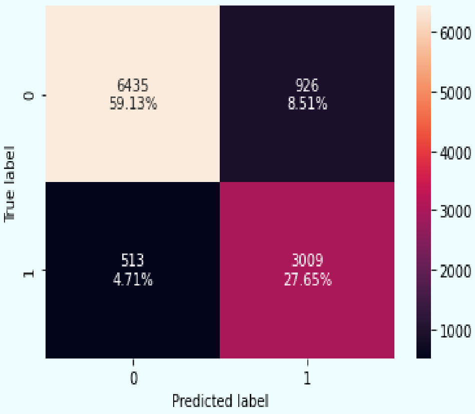
- Set the Parameters, maximum depth, maximum leaf nodes, minimum leaf split for pre-pruning the tree. set the class weight as balanced.
- Decision tree classifier chose the best tree with max depth of 6, max leaf nodes as 75 , min sample split as 10 with random state 1.
- The performance of the test sample and train sample are almost same with F1 of 0.75427.

- Lead time, market segment type online, number of special requests and average price per room are the most important feature for building the model.
- Model after pre-pruning shows fewer important features.

Decision Tree Model Performance-cost complexity pruning



Confusion Matrix for training data with alpha=0.00013,class weight “balanced” and random state 1.

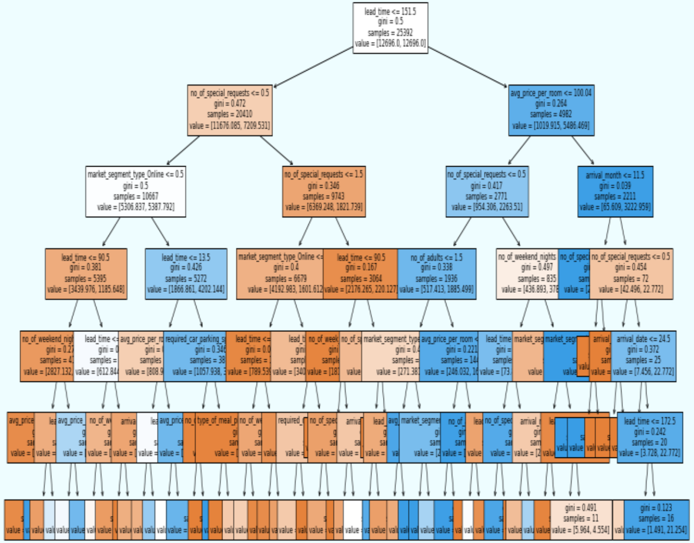


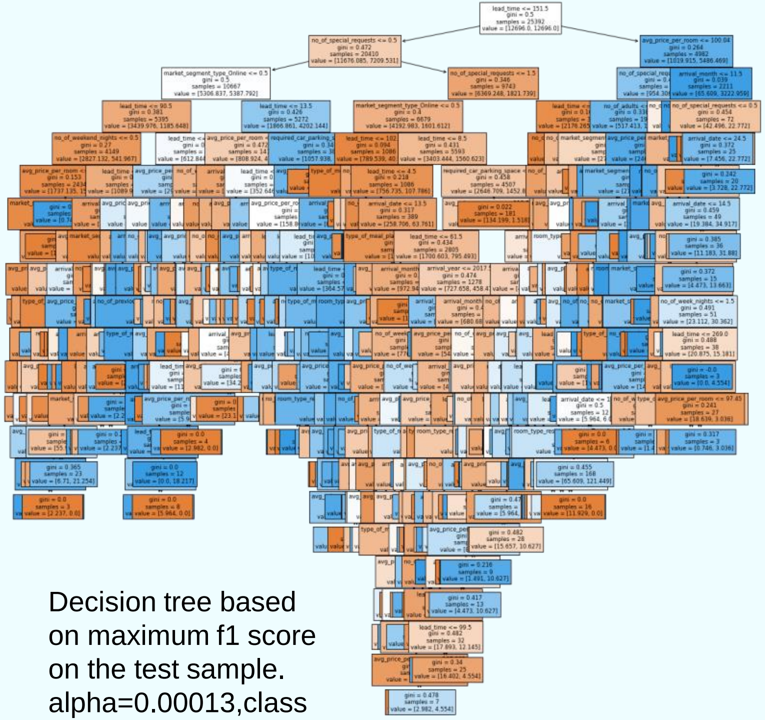
Confusion Matrix for test data with alpha=0.00013,class weight “balanced “and random state 1

Cost complexity pruning provides another option to control the size of the decision tree, this technique is parameterized by the cost complexity parameter-ccp alpha.
Greater values of ccp-alpha increase the number of nodes pruned.
F1 score train=0.84835
F1 score test=0.80703
As we can see the performance of the test sample does not differ largely from the train sample , but the ccp-alpha value we chose is the minimum as the F1 test score is high when alpha is minimum(refer Appendix), which leads to a complex tree, but is more accurate on all parameters as compared to the pre-pruned tree.

Decision tree after pre-pruning

Decision tree after post-pruning




lead_time <= 151.5
gini = 0.5
samples = 25392
value = [12696.0, 12696.0]
no_of_special_requests <= 0.5
gini = 0.472
samples = 20410
value = [11676.085, 7209.531]
avg_price_per_room <= 100.04
gini = 0.264
samples = 4982
value = [1019.915, 5486.469]
market_segment_type_Online <= 0.5
gini = 0.5
samples = 10667
value = [5306.837, 5387.792]
no_of_special_requests <= 1.5
gini = 0.346
samples = 9743
value = [6369.248, 1821.739]
arrival_month <= 11.5
gini = 0.039
samples = 2211
value = [65.609, 3222.959]

Decision tree based on the Grid of parameters:

- Class weight-balanced
- Max depth-6
- Max leaf nodes 75
- Min samples split-10
- Random state-1

Decision tree based on maximum f1 score on the test sample. alpha=0.00013,class weight “balanced” and random state 1.

Decision tree Model performance –Comparing the models

Training data	Decision tree sklearn	Decision tree(pre-pruning)	Decision tree(post pruning)
Accuracy	0.99421	0.83109	0.89438
Recall	0.98661	0.78608	0.89705
Precision	0.99578	0.72449	0.80468
F1	0.99117	0.75403	0.84835

Test data	Decision tree sklearn	Decision tree(pre-pruning)	Decision tree(post pruning)
Accuracy	0.87182	0.83488	0.86778
Recall	0.80522	0.78308	0.85434
Precision	0.80000	0.72751	0.76468
F1	0.80260	0.75427	0.80703

- We tried the fit the model using decision tree classifier, which resulted in an overfitting model with F1 score of 0.99 for train data and only 0.80 for the test data.
- We pre-pruned the model using hyperparameter tuning ,which resulted in consisted F1 score 0f 0.75 for both the train and test samples, and the tree is also simple with a depth of 6.
- We post pruned the model using cost complexity parameter ccp-alpha , the F1 score is much better(0.80) ,but the size of the tree is complex, but more accurate.
- Post -pruned decision tree looks like a better option for model building, as the performance metrics for both the train and test are giving almost similar information.

Business Insights and Recommendations

- We have been able to build a predictive model using logistic regression that can be used by INN hotel management to find the customers who might cancel their bookings with an f1_score of 0.70 on the training set and formulate actions accordingly.
- All the logistic regression models have given a generalized performance on the training and test set.
- Coefficient of some levels of education, number of adults , number of children, number of weekend nights, number of weekday nights, lead time, arrival year, number of previous cancellations, average price per room are positive, an increase in these will lead to increase in chances of a customer cancelling the bookings.
- Coefficient of required car parking space, arrival month, repeated guest, number of special requests , type of room and market segment type are negative increase in these will lead to decrease in chances of a customer cancelling the bookings.

- We analyzed the “INN hotel" using different techniques and used Decision Tree Classifier to build a predictive model for the same.
- The model built can be used to predict if a customer is going to cancel the booking or not.
- We visualized different trees and their confusion matrix to get a better understanding of the model.
- Lead time, average price per room, market segment type online, arrival date number of special requests are the most important variables in predicting the customers that cancel the booking.
- Analyzing the decision tree, if the lead time<=151.5, number of special requests<=0.5,market segment type online <=0.5 and average price per room <=196.50 , then the booing might not be cancelled by the customer.
- If the lead time >151.5, average price per room <=100,number of special requests <=0.50,then there is a chance that customer might cancel the booking.

Business Insights and Recommendations

- The INN hotel Group should review the booking window for cancellations. if a lot of these bookings are made far ahead, consider offering early bookers a nonrefundable rate to increase their occupancy and secure revenue.
- Verify guests credit card details when they initially receive bookings and request a prepayment in advance.
- If they get a last-minute cancellation, review and relax some restrictions to make sure they are visible to as many potential guests as possible.
- Try to offer a mix of rates and policies that appeal to all kind of guests.
- Offer guests multiple payment options to increase their chances of booking and securing revenue.
- By setting a grace period and waving cancellation fees for guests who cancel within a day or two after booking, they can reduce no-shows significantly. This will reduce the manual work involved in managing cancellations and reselling rooms and help secure revenue as quickly as possible.

greatlearning
Power Ahead

Happy Learning !

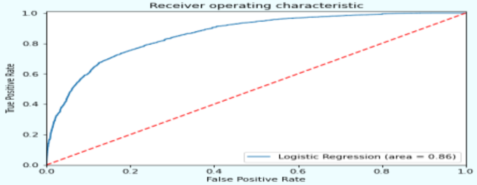
Appendix- Model assumptions

- Logistic Regression

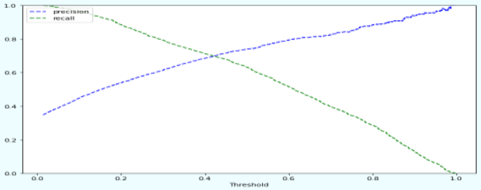
We have checked for Multicollinearity using VIF

1. All numerical variables with VIF >5 were dropped.
2. VIF for dummy variables were ignored.

- Optimal Threshold value of 0.37 was taken from the AUC-ROC curve.

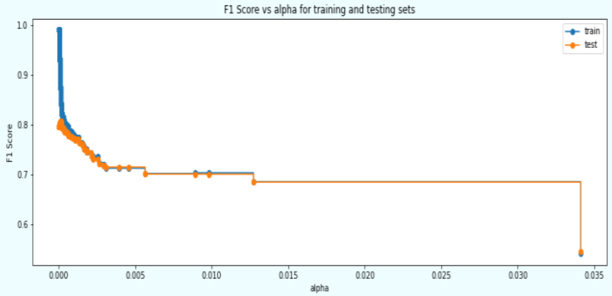


- Optimal Threshold value of 0.42 was taken from precision-recall curve.



Appendix

- Decision tree, cost complexity parameter alpha(ccf) was selected, based on the highest F1 score of the test data


F1 Score vs alpha for training and testing sets
train
test
F1 Score
alpha