greatlearning
Learning for Life

Business report

INN Hotels Group

DATA SCIENCE AND BUSINESS ANALYTICS

greatlearning
Learning for Life

Content:

4.2. Logistic regression (deal with multicollinearity, remove high p-value variables, determine optimal threshold using ROC curve)

4.3. KNN classifier

4.4. Decision tree classifier (pre-post pruning)

4.5. Check and comment on tuned model performance across different metrics

5. Model performance comparison and Final model selection

5.1. Compare all the models and choose the best model
5.2. Comment on all the model performance and provide rationale for selecting the best model

6. Actionable Insights & Recommendations

6.1. Actionable Insights & Recommendations

greatlearning
Learning for Life

1. Exploratory Data Analysis

1.1. Problem definition:

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist have 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.

1.2. Data background and content.

Context

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 make arrangements for the guests.

Data Description

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

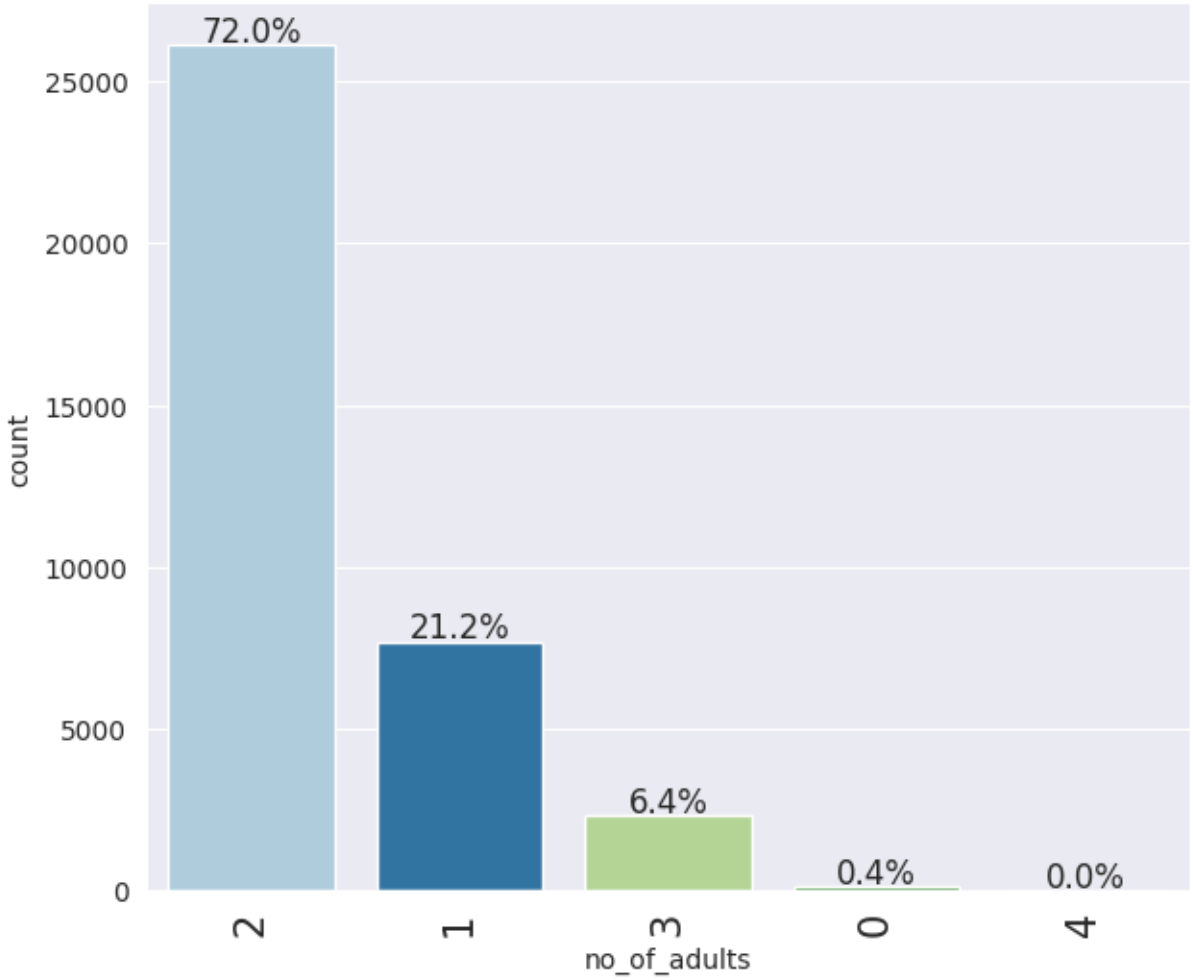
Data Dictionary:

- 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.

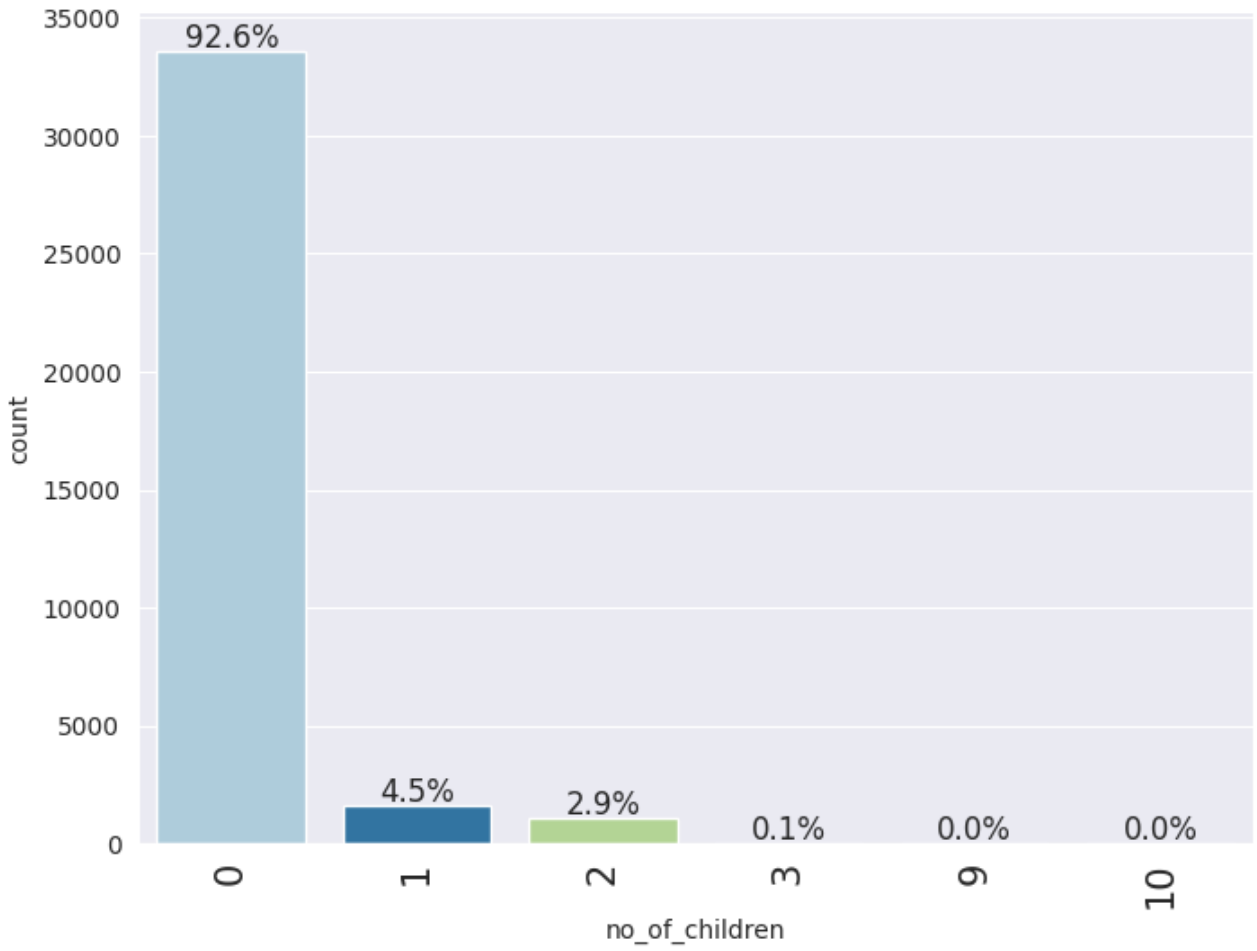
1.3. Univariate analysis.

- **no_of_adults :**



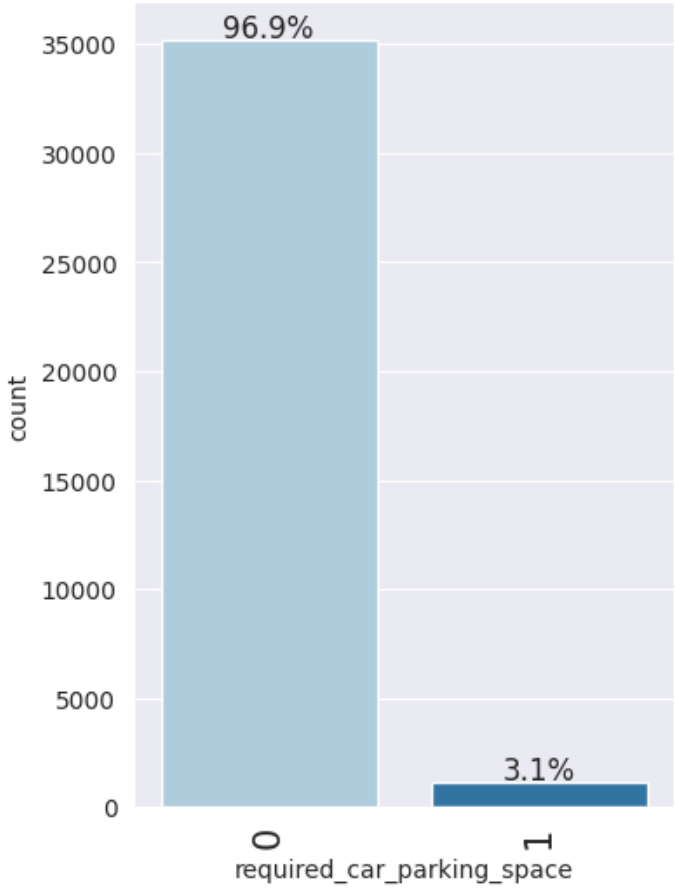
- As per above plot only 2 adults frequently visit hotels,the percentage of visiting only 2 adults is 72%.
- secondly a single adult often visit to the hotel,one adult vising percentage is 21.2%.

- **No of children :**



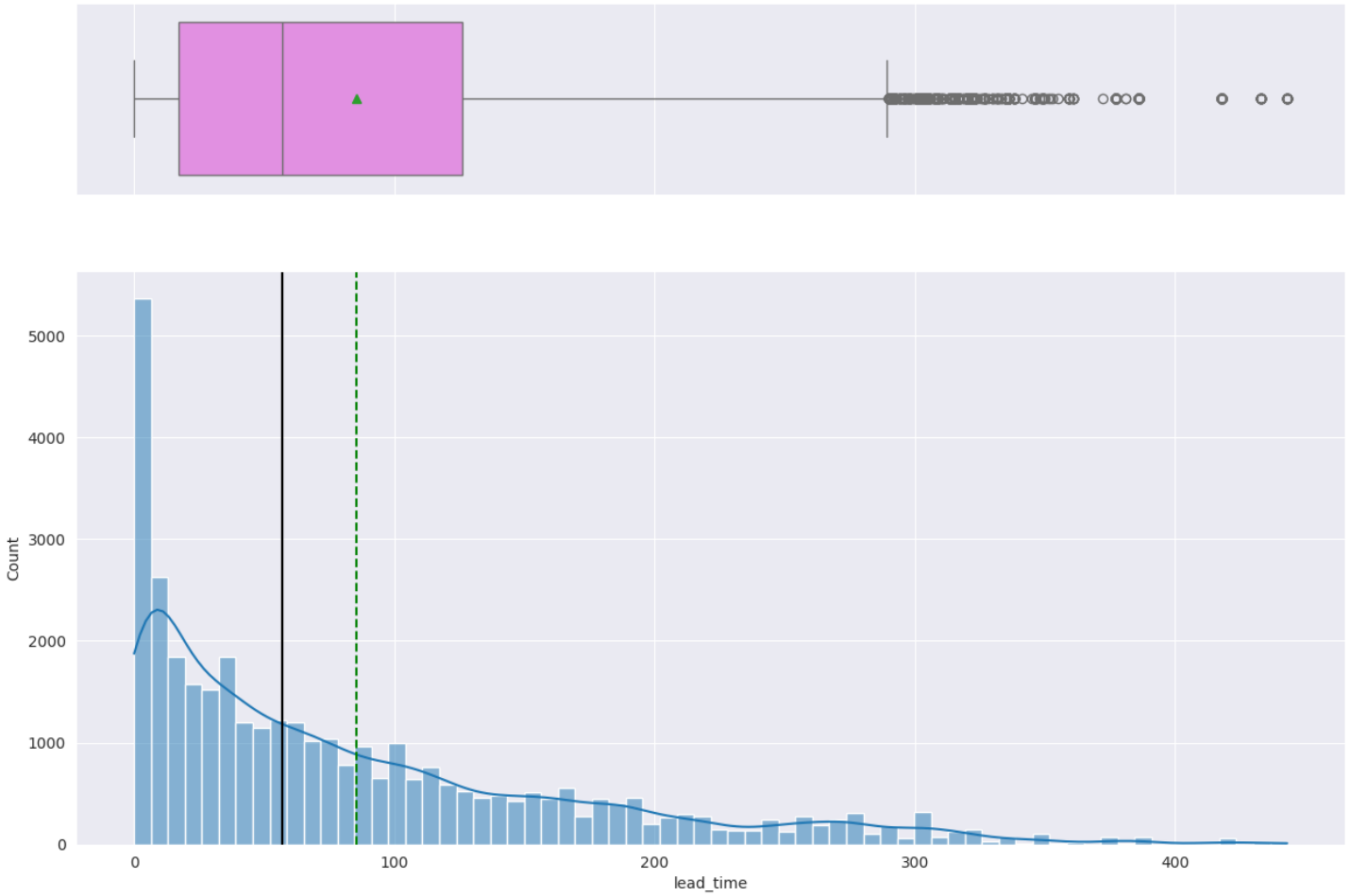
- According to the above plot children are not used to visit hotel often, 92.6% of people of visit hotel are not children.
- Only a single child visiting to hotel is 4.5% which is higher than 2, 3, 9 or 10

- **Required car parking space :**



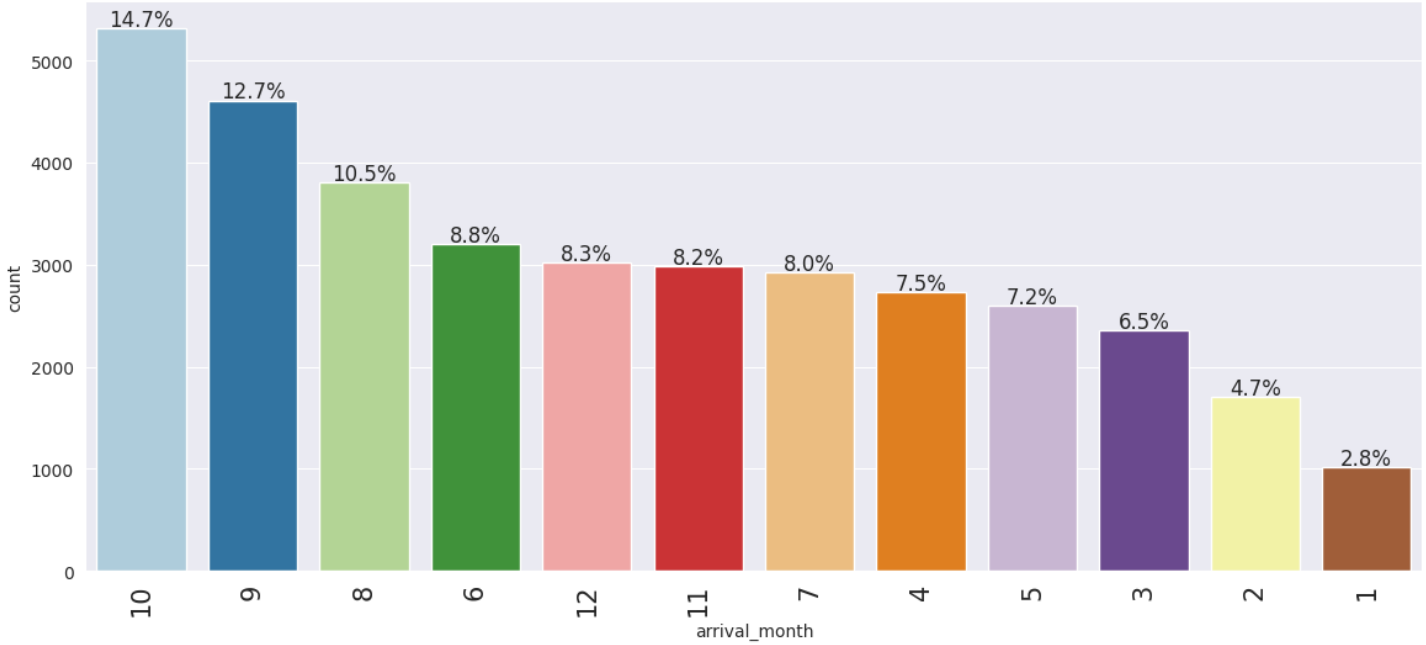
- 96.9% of the visitors has no demand for the car parking space, which shows 96.9% of the visitors do not bring cars to the hotel.
- 3.1% of the visitors required car parking space.

- **Lead time :**



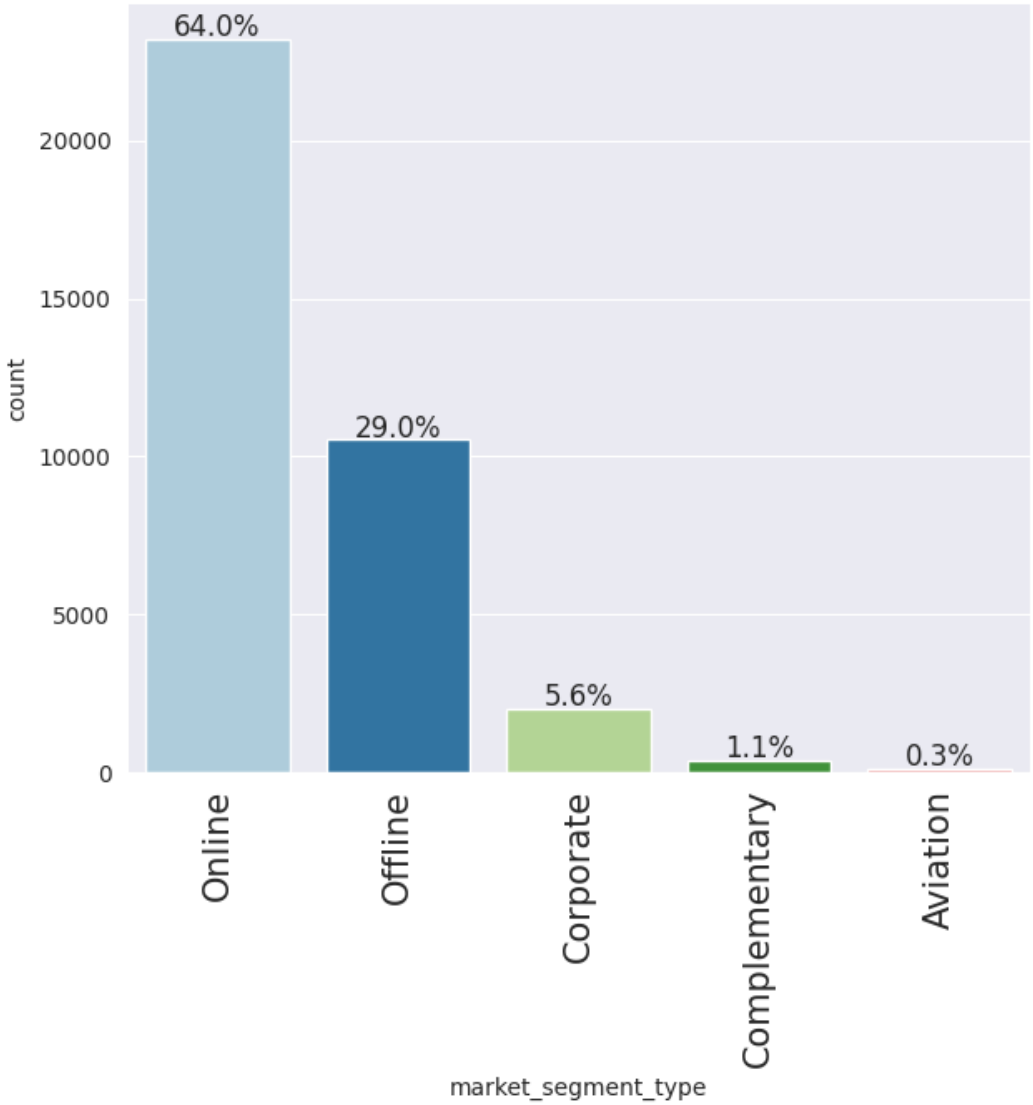
- Most the people dont have lead time, that means most of the visitors do not book their room early.
- Boxplot have outliers
- sometimes people used to book their room so early , that means visitors make their plan so early that they visit to hotel.

- **Arrival month :**



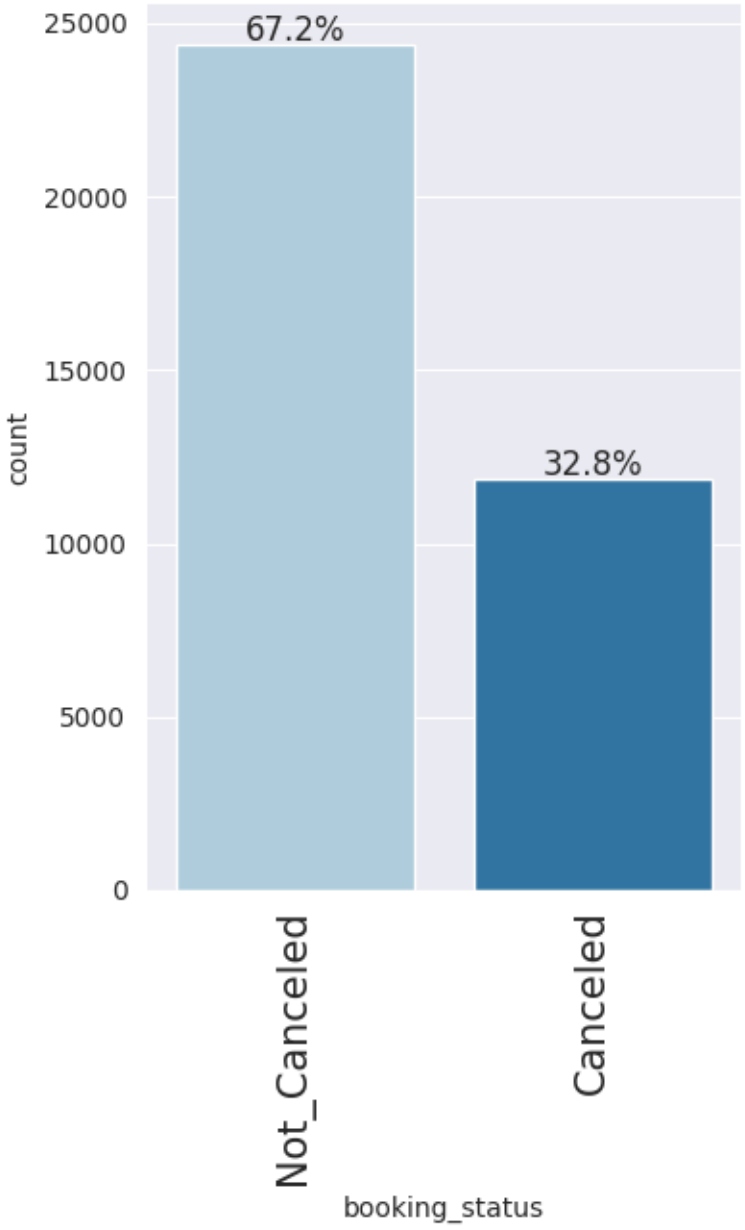
- During the month of october the visitors used to visit hotel more often, 14.7% of the visitors used to visit on october.
- 2nd most preferable month is september ,12.7% of the people visits hotel in this month.
- 3rd preferable month is august , where 10.5% of the people visits.
- **OCTOBER, SEPTEMBER, AUGUST ARE THE BUSIEST MONTHS IN THE HOTEL.**

- **Market segment type :**



- Most used market segment is online , 64% of marketing is done through the online.
- Offline is second more preferable market segment type, 29% of marketing is done.
- 3rd most important one is corporate which is of 5.6%.
- **ONLINE IS THE MARKET SEGMENT DO MOST OF THE GUESTS COME FROM.**

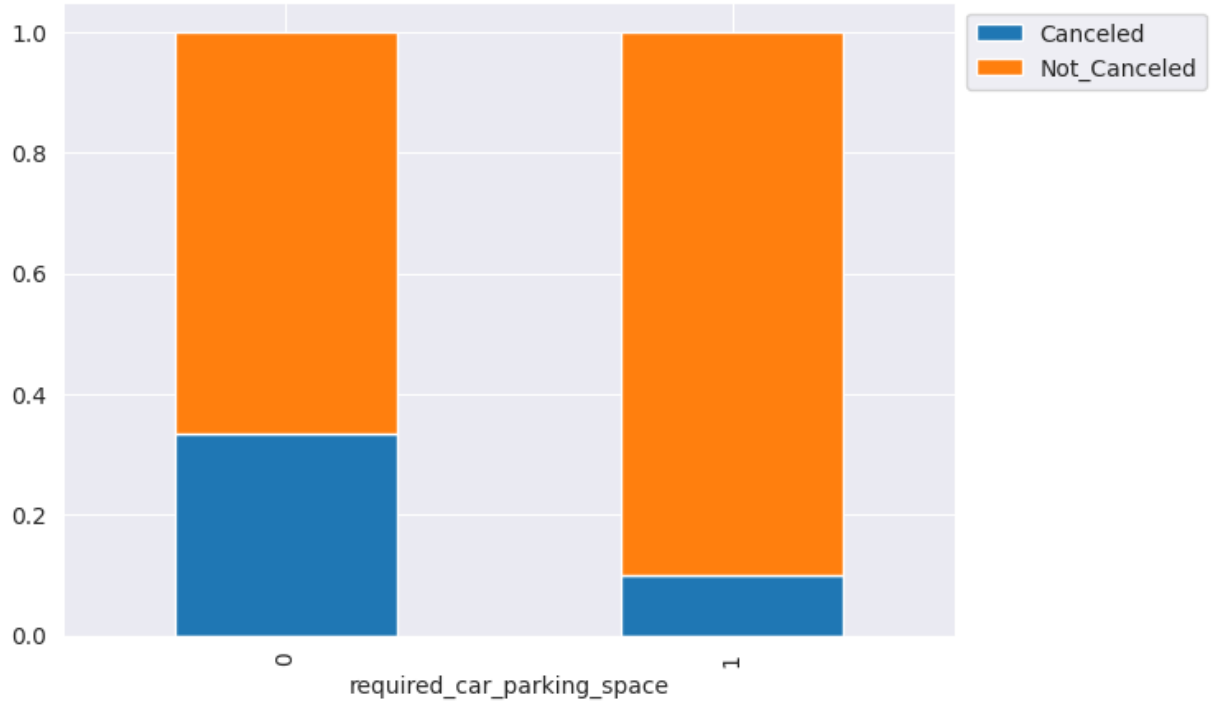
- **Booking status**



- 67.2% of booking is not cancelled
- 32.8% of booking is cancelled which has to predict using given variables.

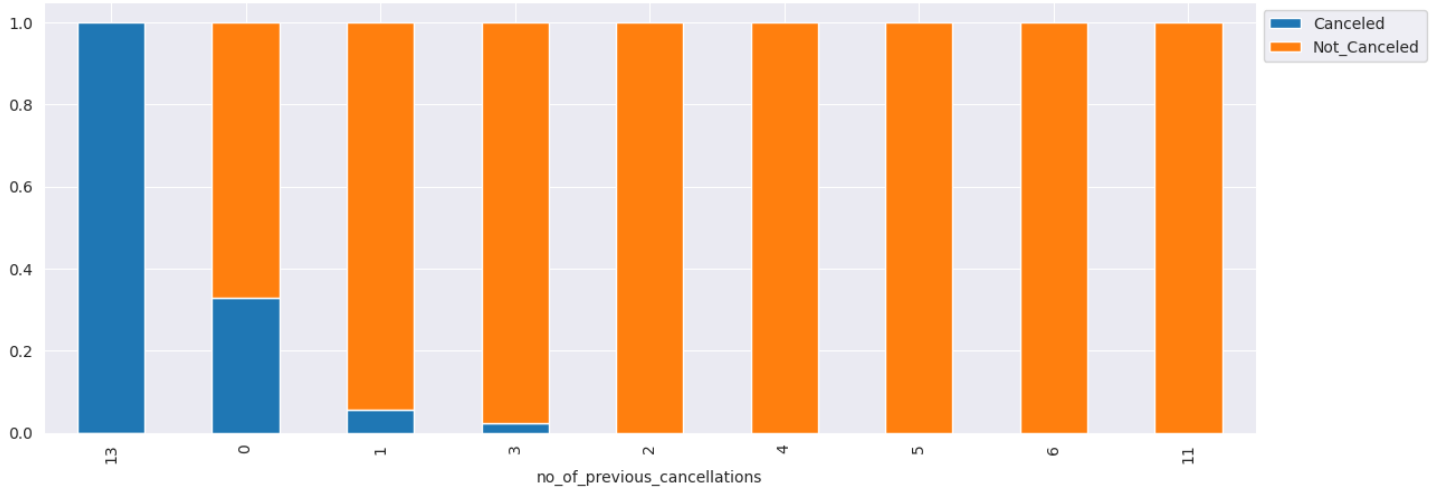
1.4. BIVARIATE ANALYSIS

- **Required parking space and the booking status:**



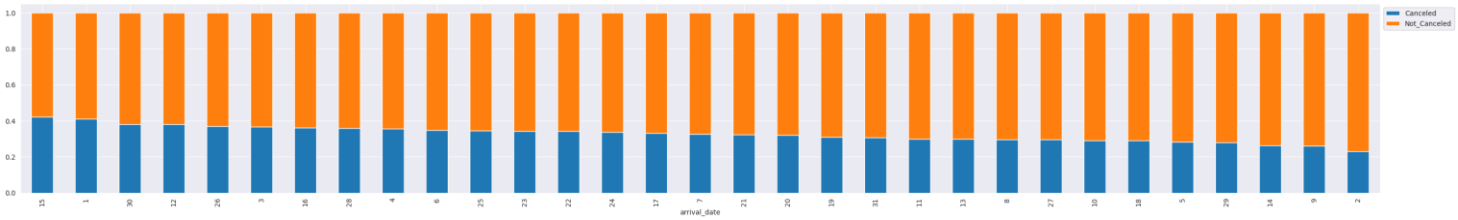
- Those visitors who do not required parking space are tend to cancel the booking.
- The cancelation is less when visitor demand to have a parking space than who dont demand.

- **No of previous cancelation vs booking status :**



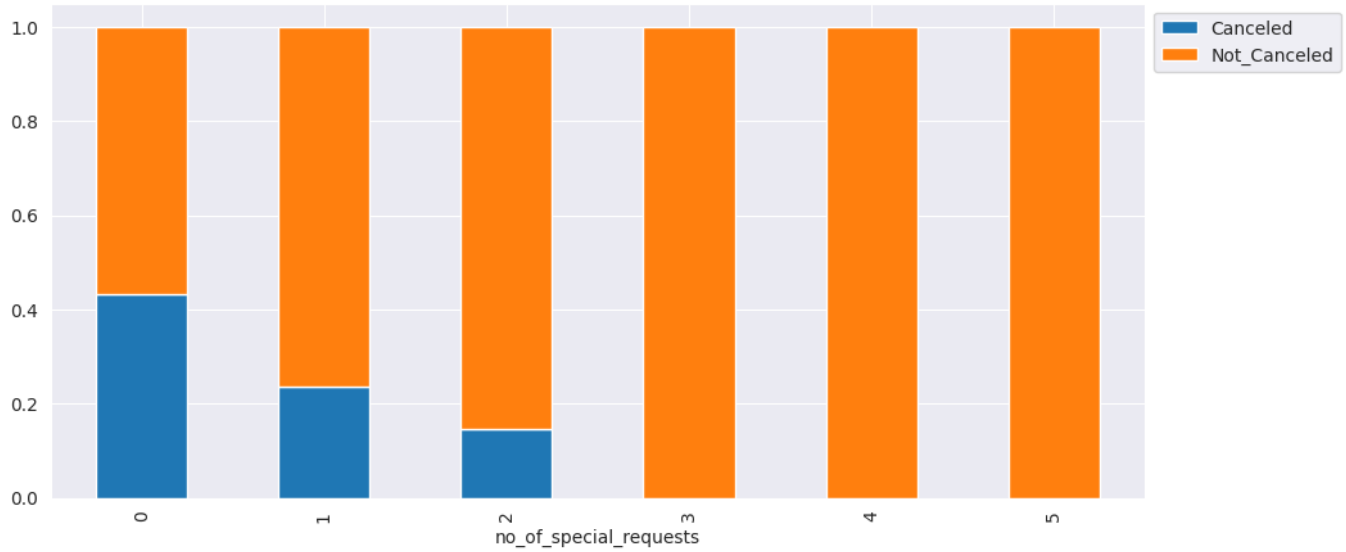
- Those visitors who do not required parking space are tend to cancel the booking.
- The cancelation is less when visitor demand to have a parking space than who dont demand.

- **Arrival date vs booking status :**



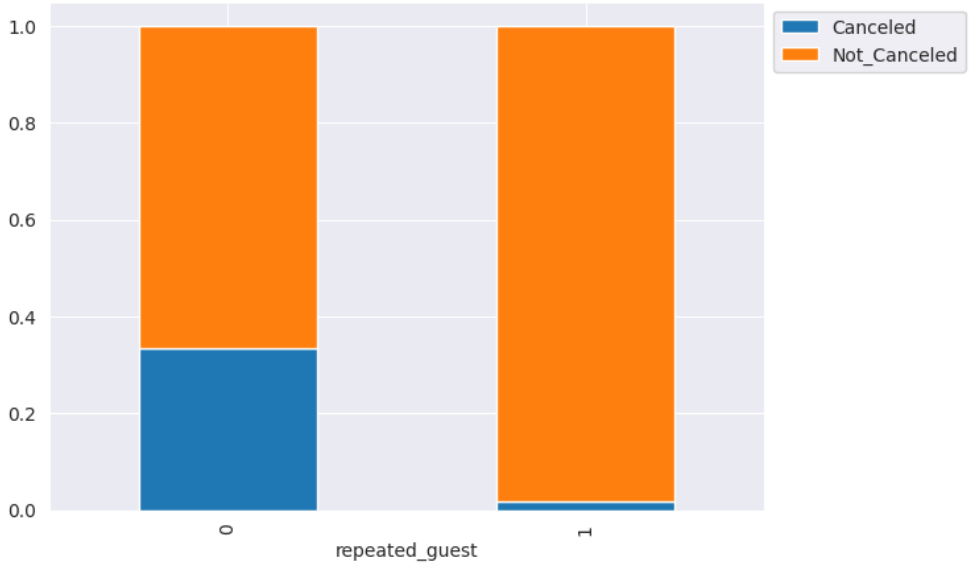
➢ cancelation is evenly spread for each day so its difficult to predict cancelation using arrival date.

- **No of special request and booking status :**



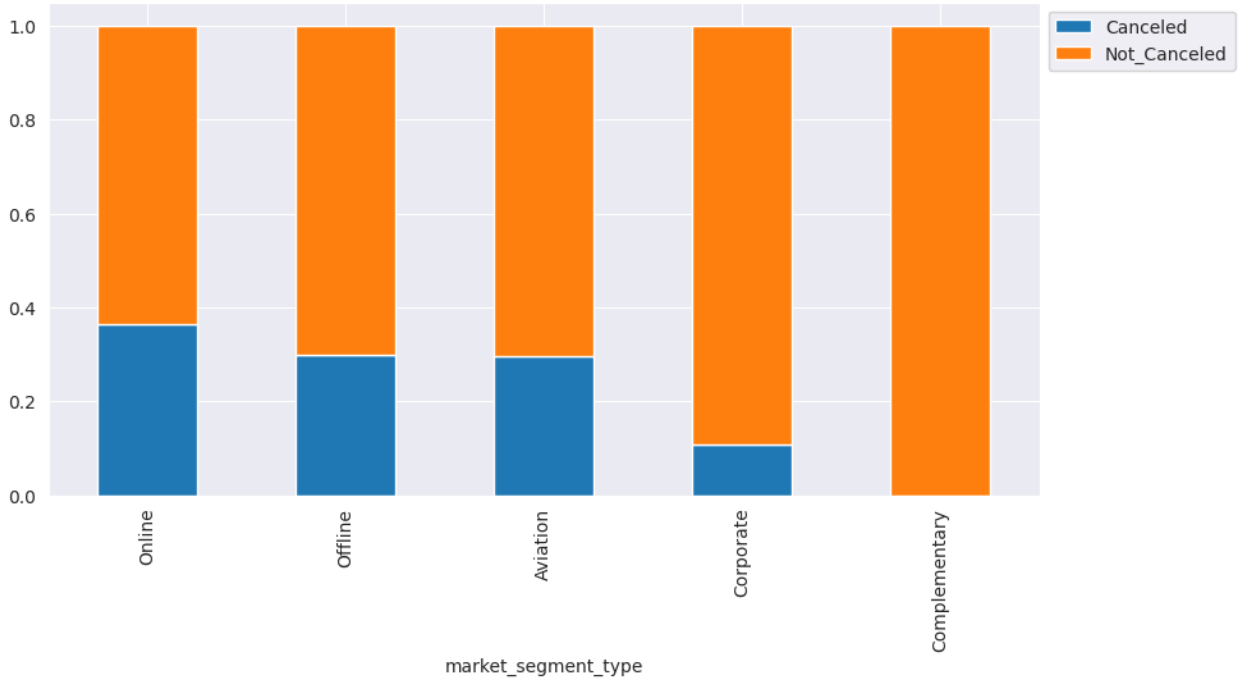
- Those who dont have any special request are tend to cancel the booking.
- As the special request increases number of cancelation decreases.
- As said above when special request increased to 1 and then increased to 2 cancelation rate decreased.

- **Repeated guest VS booking status :**



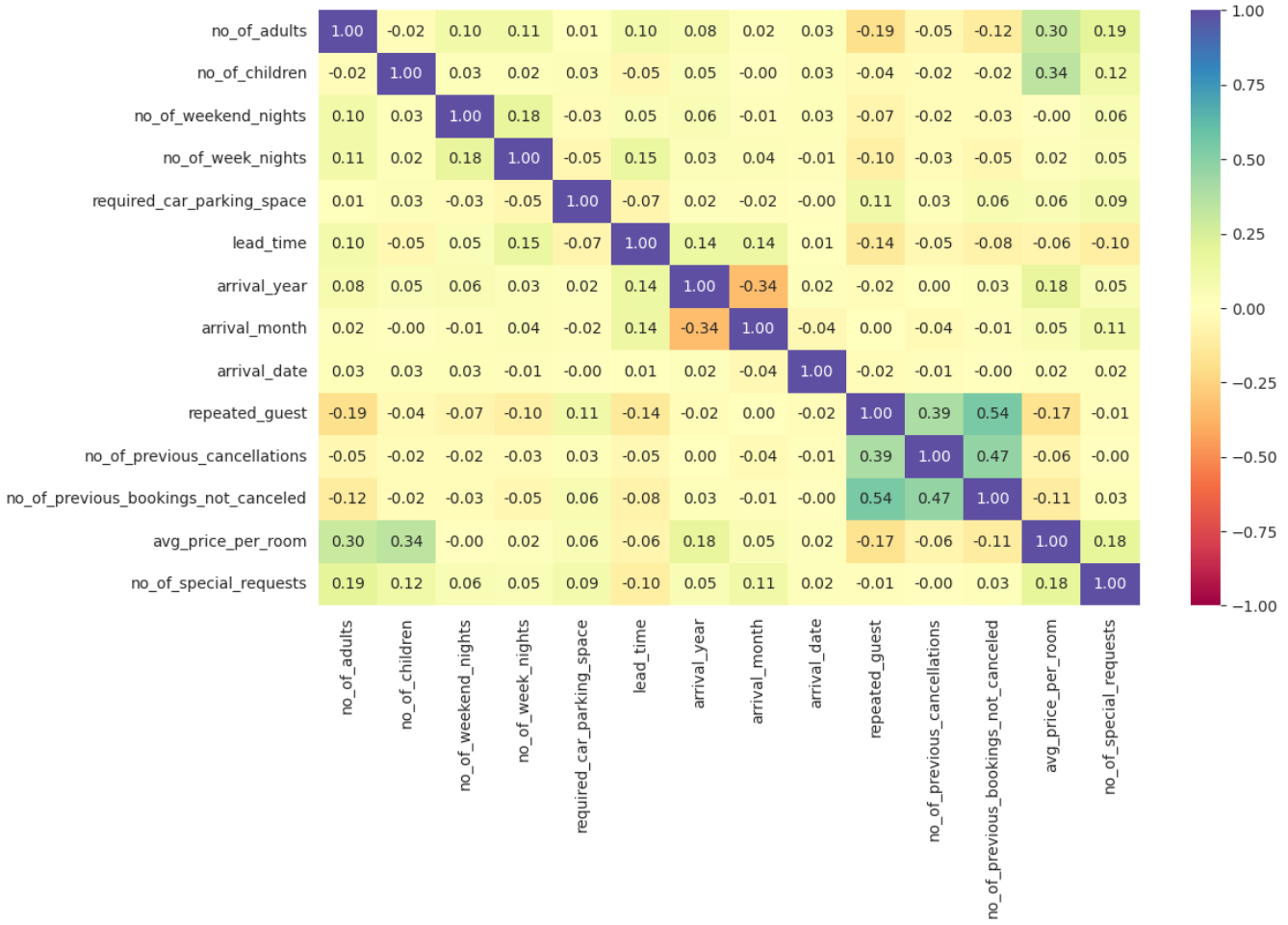
- Those who are not repeated customers tend to cancel the booking.
- Those who are repeated customers are not tend to cancel the booking.

- **Market segment VS Booking status :**



- From the above graph we can understand that the highest price per room is in online market segment type.
- The lowest price per room is in complementary market segment type.
- Aviation has second highest priced market segment.
- Offline has lil more price than corporate market segment.

- **Correlation matrix:**

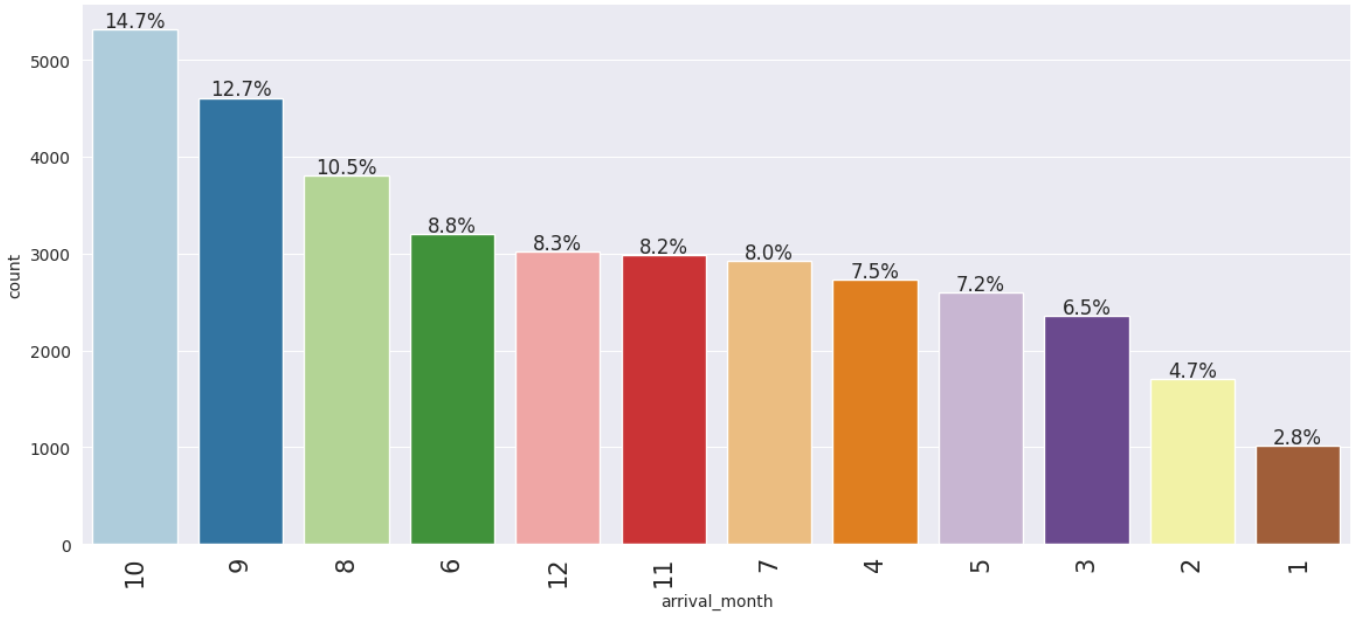


- The highest correlation is between no of previous bookings not canceled and the repeated customers which make sense too.
- There is lil less correlation between no of previous bookings canceled and repeated customers than the no of previous bookings not canceled and the repeated customers which is a good thing , which means people who visits hotel would like to stay in hotel.
- There is a correlation between no of previous bookings not canceled and no of previous bookings canceled.
- There is slightly high correlation between between avg price per room and no of children as compared to avg price per room and no of adults. There is negative correlation between lead time and repeated customers.

1.5. Answers to the key questions provided:

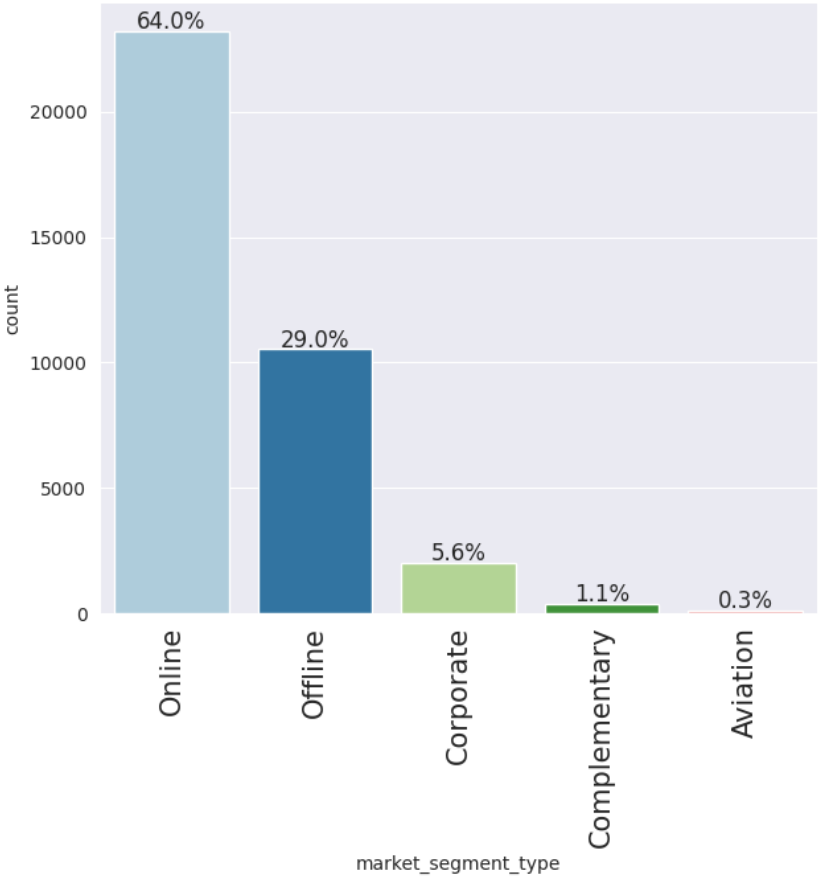
The following questions need to be answered:

1) What are the busiest month in the hotel?



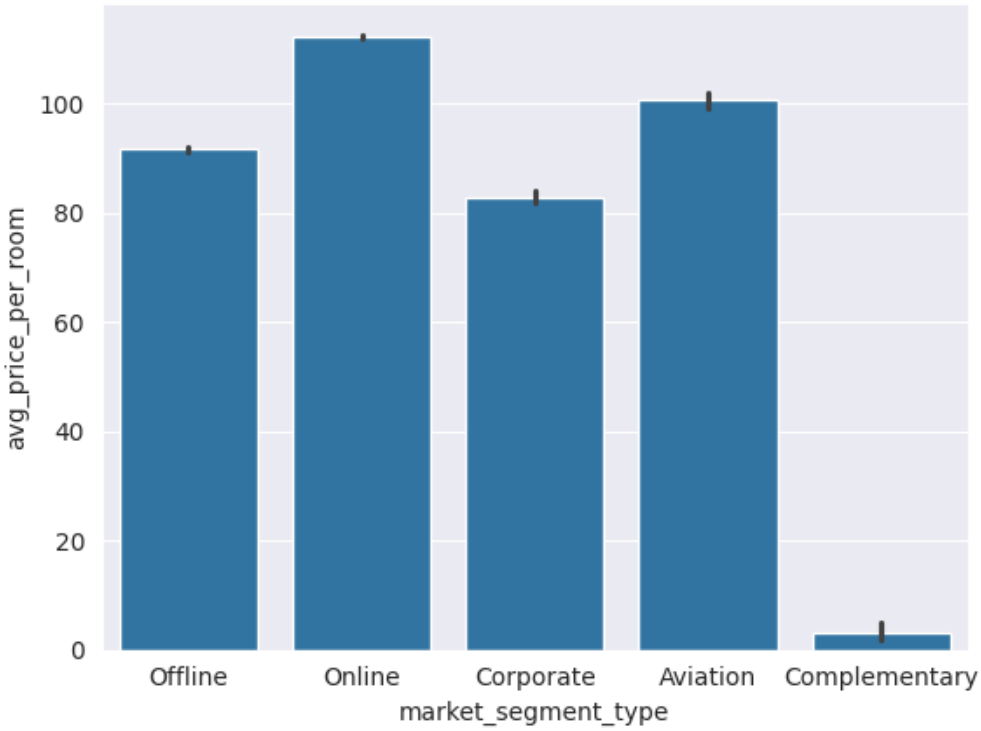
- During the month of october the visitors used to visit hotel more often, 14.7% of the visitors used to visit on october.
- 2nd most preferable month is september ,12.7% of the people visits hotel in this month.
- 3rd preferable month is august , where 10.5% of the people visits.
- **OCTOBER, SEPTEMBER, AUGUST ARE THE BUSIEST MONTHS IN THE HOTEL.**

2)What market segment do most of the guests come from?



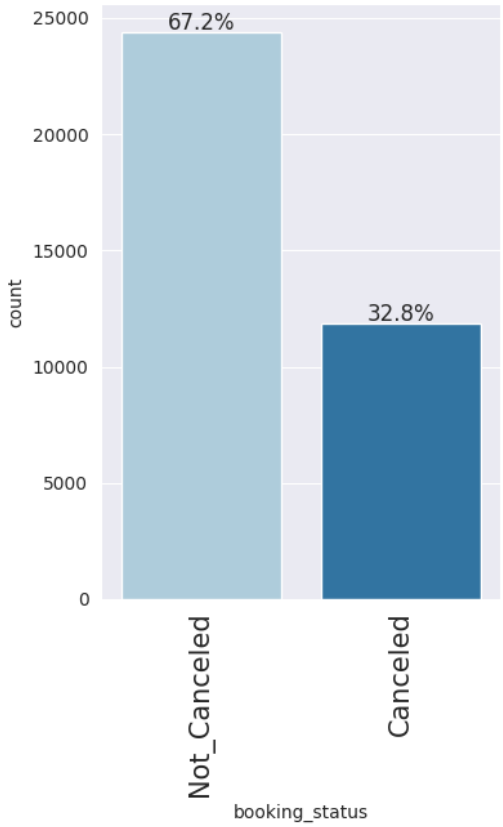
- Most used market segment is online , 64% of marketing is done through the online.
- Offline is second more preferable market segment type, 29% of marketing is done.
- 3rd most important one is corporate which is of 5.6%.
- **ONLINE IS THE MARKET SEGMENT DO MOST OF THE GUESTS COME FROM**

3) Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?



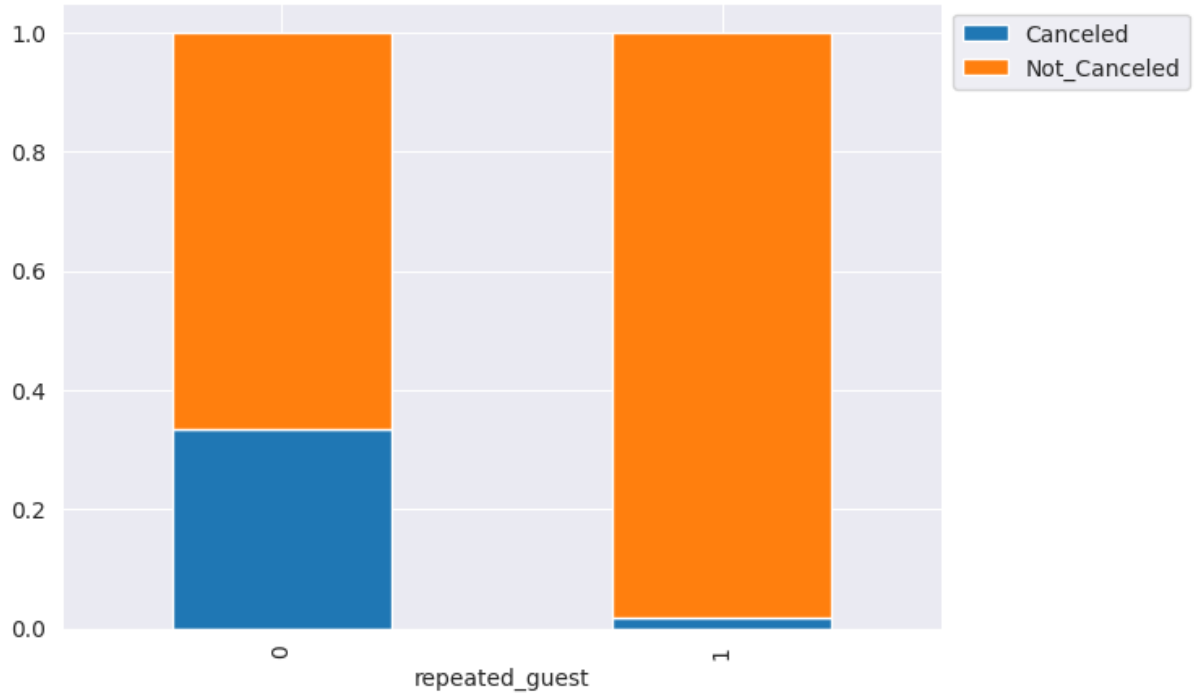
- From the above graph we can understand that the highest price per room is in online market segment type.
- The lowest price per room is in complementary market segment type.
- Aviation has second highest priced market segment.
- Offline has lil more price than corporate market segment.

4) What percentage of bookings are canceled?



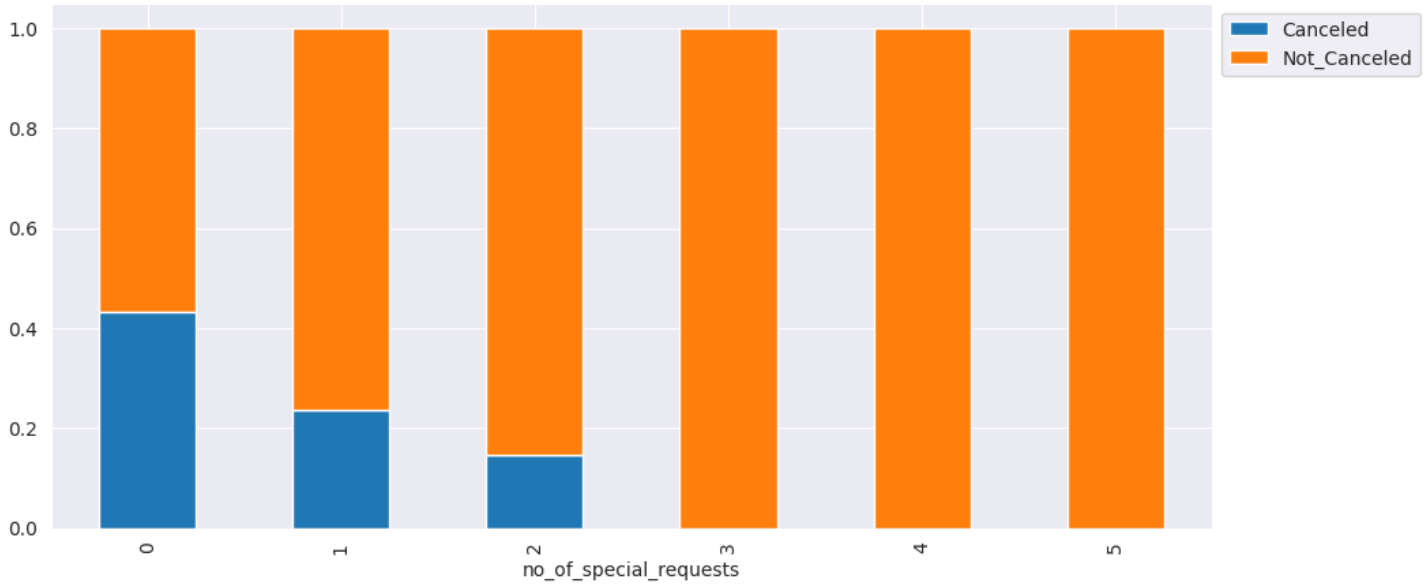
- 67.2% of booking is not cancelled
- 32.8% of booking is cancelled which has to predict using given variables.

5) Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?



- Those who are not repeated customers tend to cancel the booking.
- Those who are repeated customers are not tend to cancel the booking.

6) Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?



- Those who dont have any special request are tend to cancel the booking.
- As the special request increases number of cancelation decreases.
- As said above when special request increased to 1 and then increased to 2 cancelation rate decreased.

greatlearning
Learning for Life

2. DATA PREPROCESSING

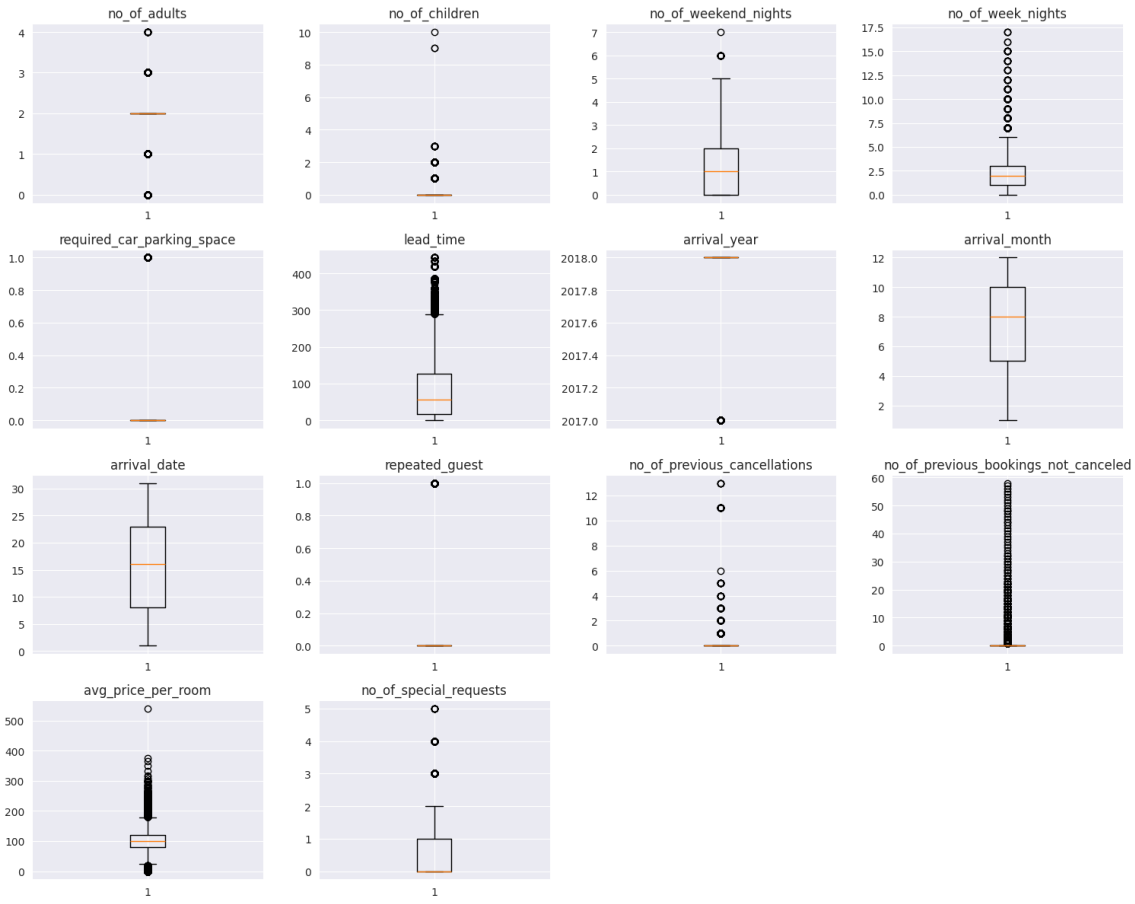
2.1. Checking for duplicate entries in the dataset

- There are no duplicate entries in the dataset.

Checking for missing values in the dataset.

- There are no null values in the dataset.

2.2. Outlier checking and treatment



Observations

- There are quite a few outliers in the data.
- However, we will treat some of them.
- And we not treat some as they are proper values genuine.
- Treating outliers which looks genuine will lead to loss in data.

2.3. Feature engineering

- In feature engineering we dropped unwanted columns.
- Some of the avg price were not correctly given, replaced those with nans and then filled those rows using mean of avg price of the hotel.

2.4. Data preparation for modeling

- We want to predict the booking status.
- Before we proceed to build a model, we'll have to encode categorical features
- We'll split the data into train and test to be able to evaluate the model that we build on the train data
- We will build a Logistics Regression model using the train data and then check it's performance

2.5. Data Scaling

Data scaling has done after splitting the data into test and train.

greatlearning
Learning for Life

3. Model building

3.1. Logistic Regression(stats models)

- We will now perform logistic regression using statsmodels, a Python module that provides functions for the estimation of many statistical models, as well as for conducting statistical tests, and statistical data exploration.
- Using statsmodels, we will be able to check the statistical validity of our model - identify the significant predictors from p-values that we get for each predictor variable.

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25369
Method:	MLE	Df Model:	22
Date:	Sun, 12 Jan 2025	Pseudo R-squ.:	0.3266
Time:	09:37:53	Log-Likelihood:	-10815.
converged:	False	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	
const	-2.6608	0.272	-9.772	0.000	
no_of_adults	0.0609	0.037	1.631	0.103	
no_of_children	0.0977	0.060	1.635	0.102	
no_of_weekend_nights	0.1450	0.020	7.355	0.000	
no_of_week_nights	0.0293	0.012	2.404	0.016	
required_car_parking_space	-1.6510	0.138	-12.007	0.000	
lead_time	0.0168	0.000	64.308	0.000	
arrival_month	-0.0656	0.006	-10.966	0.000	
repeated_guest	-1.8774	0.663	-2.834	0.005	
no_of_previous_cancellations	0.3469	0.102	3.403	0.001	
no_of_previous_bookings_not_canceled	-1.4184	0.919	-1.544	0.123	
avg_price_per_room	0.0202	0.001	26.941	0.000	
no_of_special_requests	-1.4713	0.030	-48.899	0.000	
room_type_reserved_Room_Type 2	-0.4585	0.132	-3.483	0.000	
room_type_reserved_Room_Type 3	1.0344	1.824	0.567	0.571	
room_type_reserved_Room_Type 4	-0.3213	0.052	-6.186	0.000	
room_type_reserved_Room_Type 5	-0.6417	0.213	-3.016	0.003	
room_type_reserved_Room_Type 6	-0.6635	0.148	-4.477	0.000	
room_type_reserved_Room_Type 7	-0.8518	0.283	-3.005	0.003	
market_segment_type_Complementary	-24.0810	2.36e+04	-0.001	0.999	-4
market_segment_type_Corporate	-1.0142	0.276	-3.674	0.000	
market_segment_type_Offline	-1.9357	0.264	-7.337	0.000	
market_segment_type_Online	-0.1235	0.261	-0.473	0.636	

Observations

- Negative values of the coefficient show that the probability of a person's cancellation will get negatively impacted from the corresponding attribute value.
- Positive values of the coefficient show that the probability of a person's cancellation will get positively impacted from the corresponding attribute value.
- p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant.

Model can make wrong predictions as:

1. Predicting a person can cancel the booking but in reality the person couldnt cancel the booking .
2. Predicting a person doesn't cancel the booking but in reality the person does cancel the booking.

Which case is more important?

- Both the cases are important as:
    - If we predict a person can cancel the booking but actually the person wouldnt cancel the booking will lead to bad service to the customer which can cause a bad image in the market.
    - If we predict a person doesn't cancel but actually the person cancel the booking then the hotel can incurr loss.

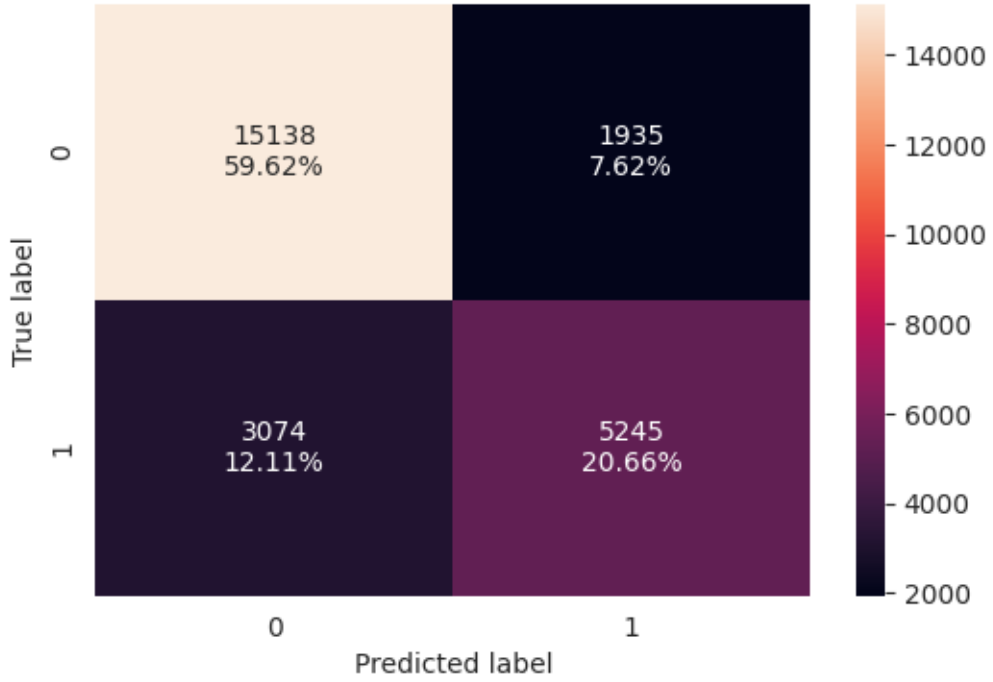
How to reduce this loss?

- We need to reduce both False Negatives and False Positives
- `f1_score` should be maximized as the greater the f1_score, the higher the chances of reducing both False Negatives and False Positives and identifying both the classes correctly
    - fi_score is computed as
    - f1_score=2∗Precision∗RecallPrecision+Recall

- **Model performance on training dataset :**

	Accuracy	Recall	Precision	F1
0	0.80273	0.63048	0.73050	0.67682

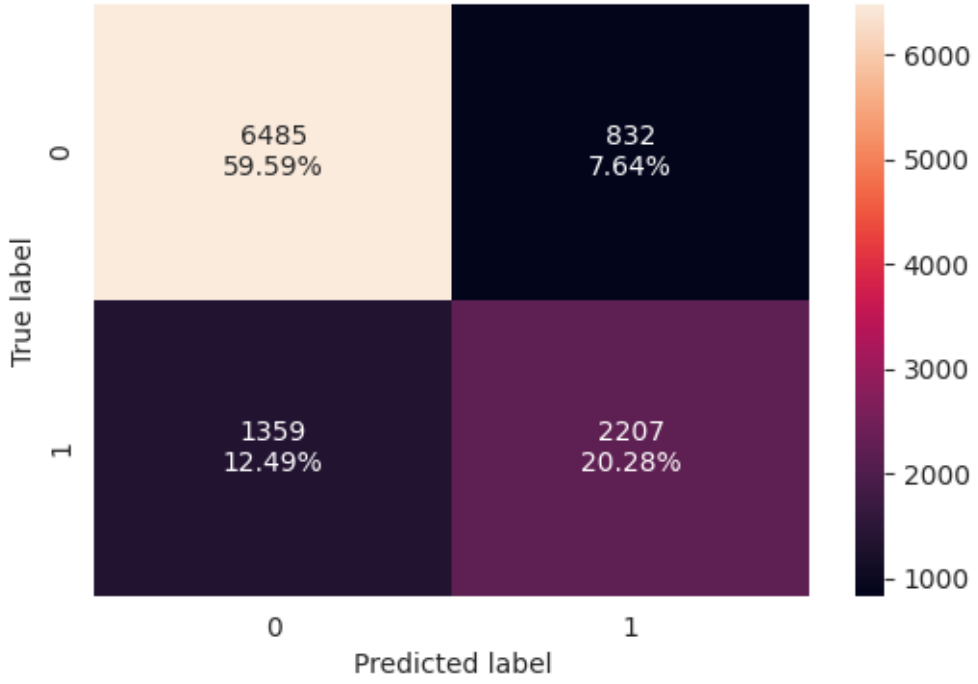
Confusion matrix:



- **Model performance on testing dataset :**

	Accuracy	Recall	Precision	F1
0	0.79868	0.61890	0.72623	0.66828

Confusion matrix:



Observations

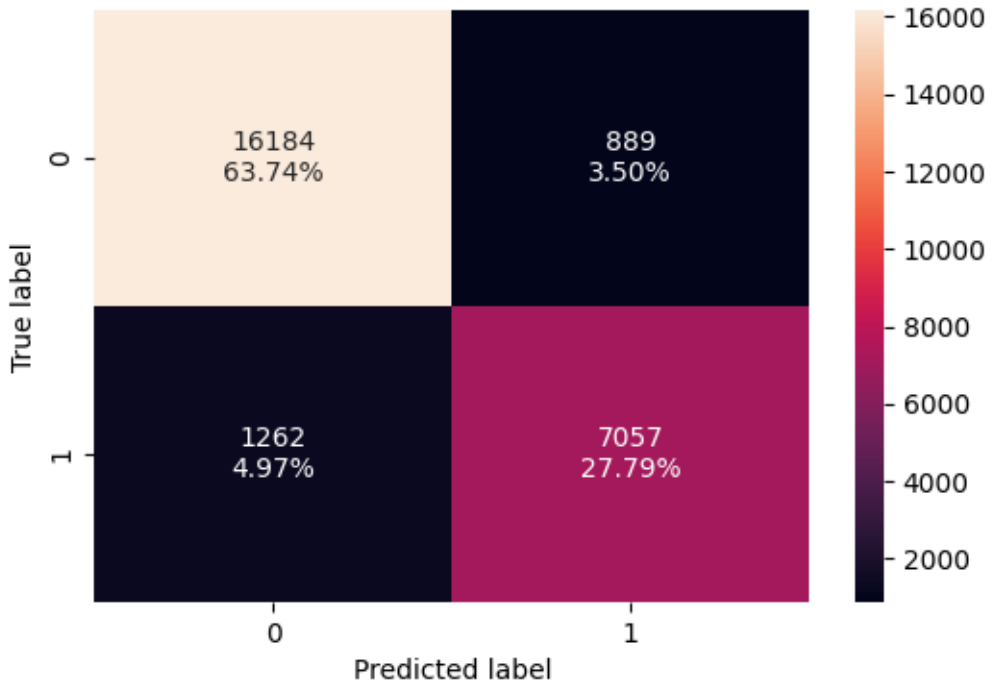
- The f1_score of the model is 0.67 and 0.66 for training set and testing set and we will try to maximize it further
- The variables used to build the model might contain multicollinearity, which will affect the p-values
- We will have to remove multicollinearity from the data to get reliable coefficients and p-values

3.2 k nearest neighbor(KNN)

In order to optimize our model, it's essential to experiment with different values of k to find the most suitable fit for our data. We can commence this process by setting k equal to 3 and gradually exploring other values to assess their impact on the model's performance.

- We'll only consider odd values of K as the classification will be done based on majority voting.

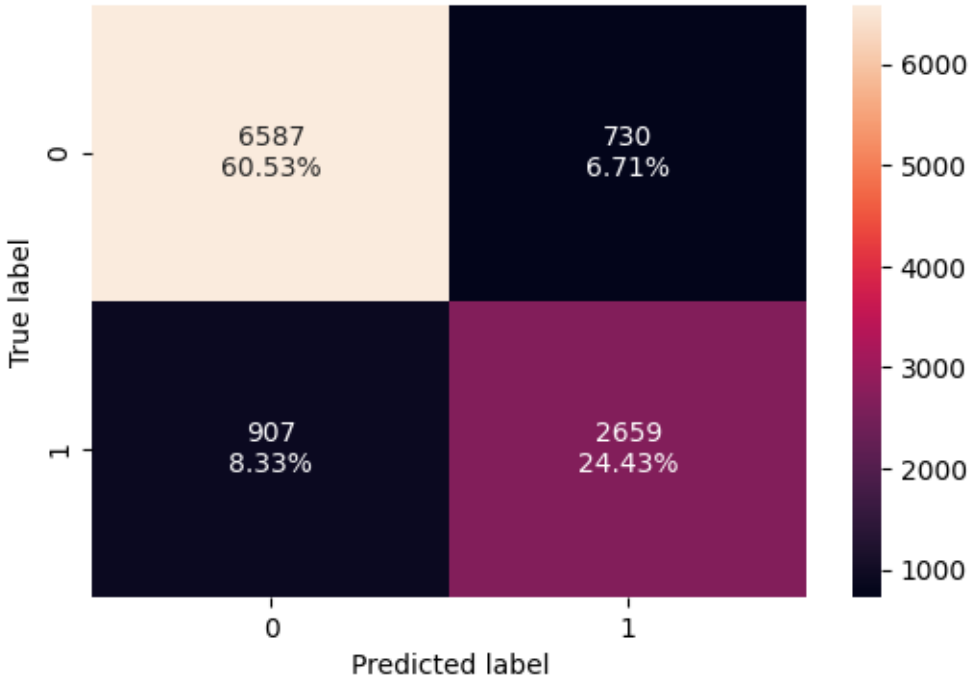
Confusion matrix when knn is 3 for training set:



Model performance on training set

	Accuracy	Recall	Precision	F1
0	0.915288	0.848299	0.88812	0.867753

- **Confusion matrix when knn is 3 for testing set:**



- **Model performance on testing set :**

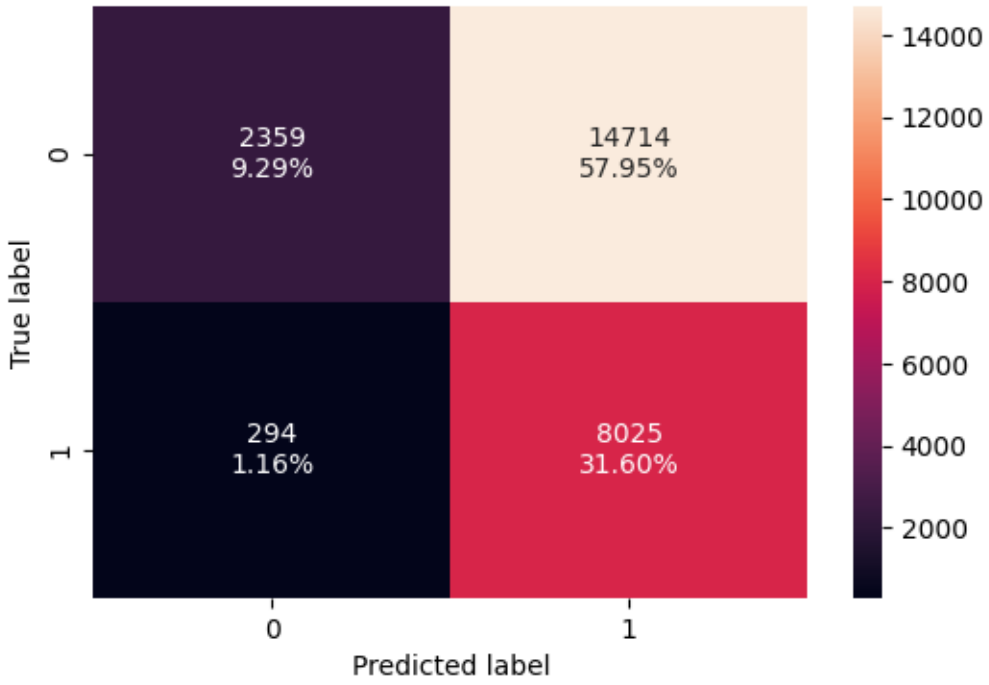
	Accuracy	Recall	Precision	F1
0	0.849582	0.745653	0.784597	0.76463

Observations:

- Knn model is overfitting, performing very well on the training data but not good enough on the testing data.
- There is a difference in accuracy, recall, precision, and f1 score between training and testing data.

3.3. Naïve Bayes:

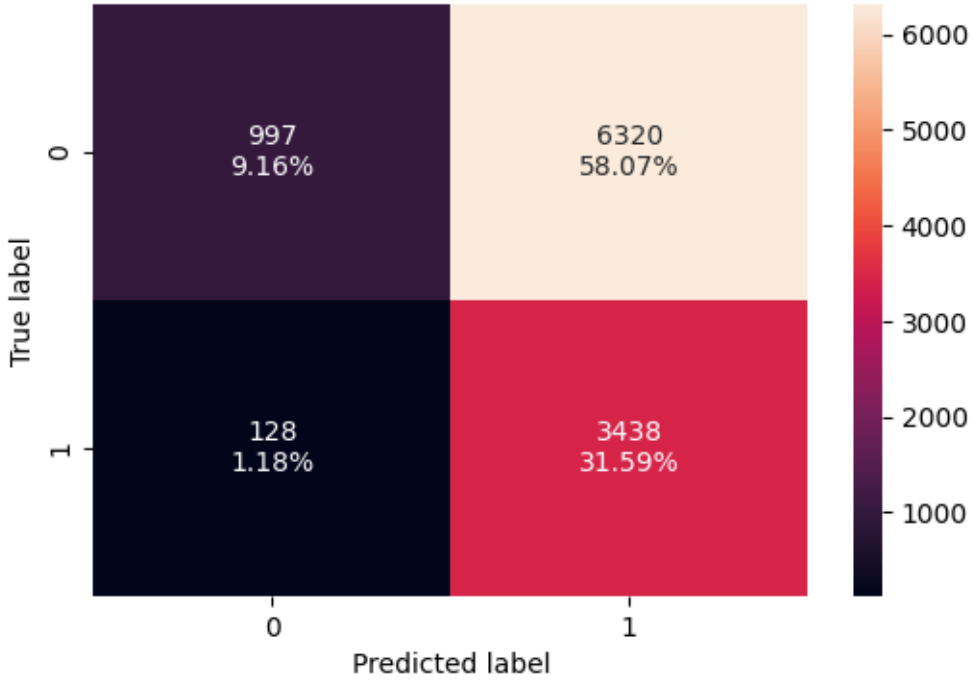
- **Confusion matrix on Naïve Bayes for training set:**



- **Model performance on training set :**

	Accuracy	Recall	Precision	F1
0	0.408948	0.964659	0.352918	0.516775

- **Confusion matrix on Naïve Bayes for testing set:**



- **Model performance on testing set :**

	Accuracy	Recall	Precision	F1
0	0.407516	0.964105	0.352326	0.516061

Observation:

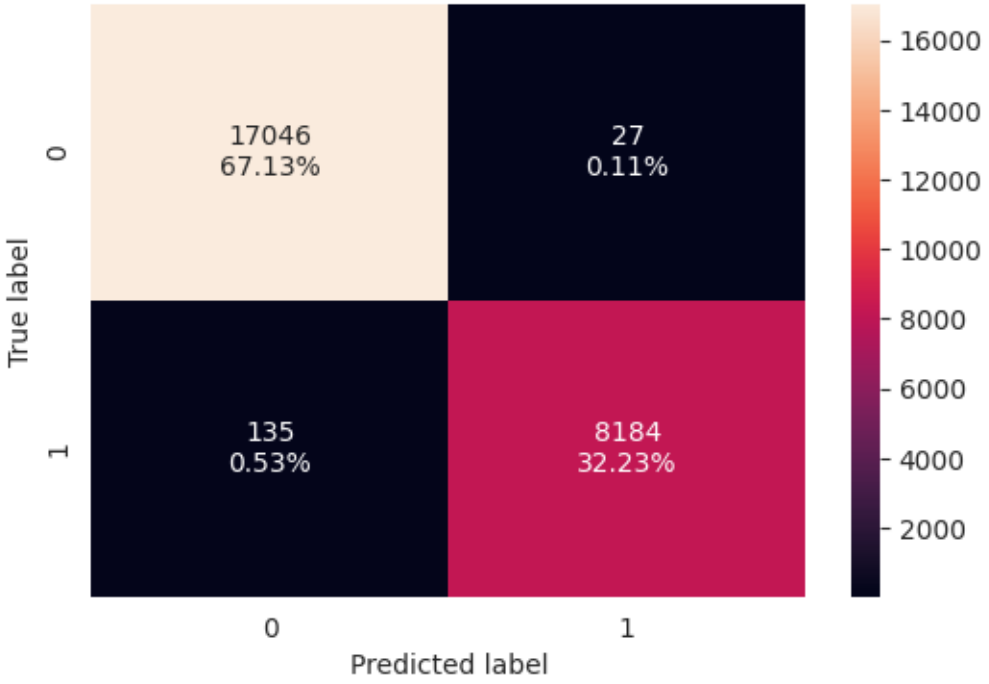
- The data is performing poor on test and train data
- The accuracy recall precision and f1 are similar in both the data

3.4. Decision tree classifier

- **Model performance on the training set:**

	Accuracy	Recall	Precision	F1
0	0.99362	0.98377	0.99671	0.99020

- **Confusion matrix for training set:**

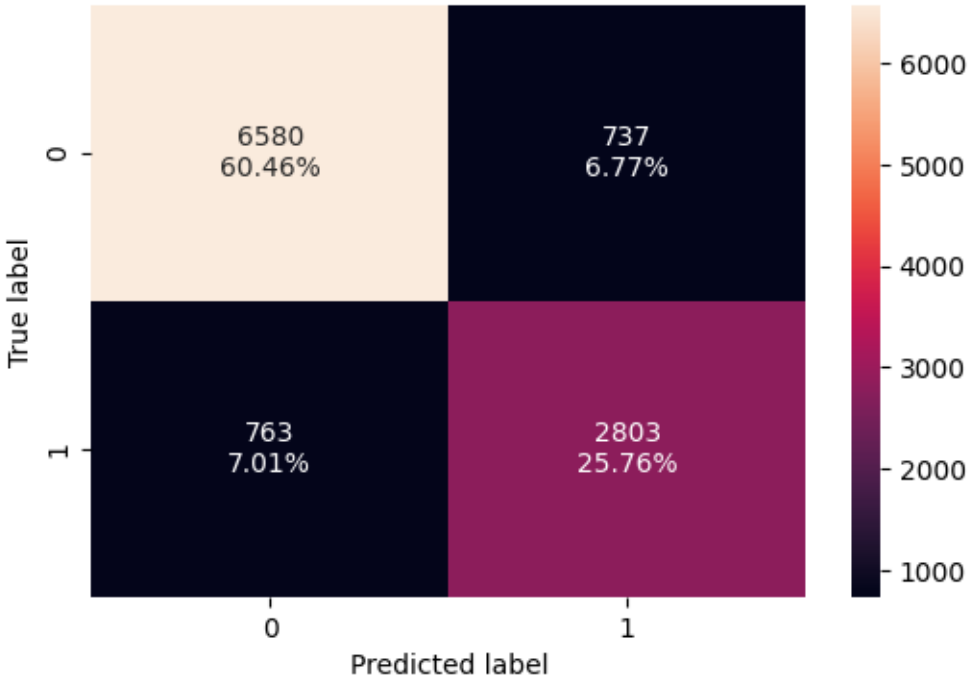


- Almost 0 errors on the training set, each sample has been classified correctly.
- Model has performed very well on the training set.
- As we know a decision tree will continue to grow and classify each data point correctly if no restrictions are applied as the trees will learn all the patterns in the training set.
- Let's check the performance on test data to see if the model is over fitting.

- **Model performance on the testing set:**

	Accuracy	Recall	Precision	F1
0	0.86217	0.786035	0.791808	0.788911

- **Confusion matrix for test data:**



- The decision tree model is over fitting the data as expected and not able to generalize well on the test set.
- We will have to prune the decision tree.

greatlearning
Learning for Life

4. Model performance improvement :

4.1. Tuning the model.

a) Logistic regression :

There are different ways of detecting (or testing for) multicollinearity. One such way is using the Variation Inflation Factor (VIF).

- **Variance Inflation factor**: Variance inflation factors measure the inflation in the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient βk is "inflated" by the existence of correlation among the predictor variables in the model.
- **General Rule of thumb**:
    - If VIF is 1 then there is no correlation among the kth predictor and the remaining predictor variables, and hence the variance of βk is not inflated at all
    - If VIF exceeds 5, we say there is moderate multicollinearity
    - If VIF is equal or exceeding 10, it shows signs of high multi-collinearity
- The purpose of the analysis should dictate which threshold to use

Model performance improvement on train data set:

Series before feature selection:

```
const                                 320.84294
no_of_adults                            1.33378
no_of_children                          1.98963
no_of_weekend_nights                    1.06672
no_of_week_nights                       1.09256
required_car_parking_space              1.03358
lead_time                               1.20158
arrival_month                           1.05130
repeated_guest                          1.74796
no_of_previous_cancellations            1.32106
no_of_previous_bookings_not_canceled    1.56548
avg_price_per_room                      1.62175
no_of_special_requests                  1.24340
room_type_reserved_Room_Type 2          1.08924
room_type_reserved_Room_Type 3          1.00359
room_type_reserved_Room_Type 4          1.29440
room_type_reserved_Room_Type 5          1.03043
room_type_reserved_Room_Type 6          1.94408
room_type_reserved_Room_Type 7          1.07430
market_segment_type_Complementary       4.15767
market_segment_type_Corporate          16.60421
market_segment_type_Offline            62.28683
market_segment_type_Online             69.27534
dtype: float64
```

- Some categorical levels of ‘market_segment’ exhibit high multi collinearity
- We see that some variables have high VIF
  - The high VIF indicate perfect correlation between variables
- We will drop ‘market_segment_type_Online.’

Series after feature selection:

```
const                                  39.87049
no_of_adults                            1.31674
no_of_children                          1.98901
no_of_weekend_nights                    1.06625
no_of_week_nights                       1.09187
required_car_parking_space              1.03350
lead_time                               1.19937
arrival_month                           1.05116
repeated_guest                          1.74465
no_of_previous_cancellations            1.32094
no_of_previous_bookings_not_canceled    1.56525
avg_price_per_room                      1.62160
no_of_special_requests                  1.23904
room_type_reserved_Room_Type 2          1.08910
room_type_reserved_Room_Type 3          1.00359
room_type_reserved_Room_Type 4          1.28832
room_type_reserved_Room_Type 5          1.03043
room_type_reserved_Room_Type 6          1.94384
room_type_reserved_Room_Type 7          1.07416
market_segment_type_Complementary       1.10881
market_segment_type_Corporate           1.48839
market_segment_type_Offline             1.37940
dtype: float64
```

- Dropping ‘market_segment_type_online’ fixes the multicollinearity in all column.

  Training performance:

	Accuracy	Recall	Precision	F1
0	0.802654	0.630484	0.730298	0.676731

- No significant change in the model performance.

Observations:

1. Dropping market_segment_type_online doesn't have a significant impact on the model performance.
2. We can choose any model to proceed to the next steps.
3. For other attributes present in the data, the p-values are high only for few dummy variables and since only one (or some) of the categorical levels have a high p-value we will drop them iteratively as sometimes p-values change after dropping a variable. So, we'll not drop all variables at once.
4. We are doing the above process manually by picking one variable at a time that has a high p-value, dropping it, and building a model again.
5. Removing room_type_reserved_Room_Type 3, no_of_previous_bookings_not_canceled , no_of_children, no_of_adults columns because of high p_values.

Model after removing high p_values and high vif values:

	Accuracy	Recall	Precision	F1
0	0.801355	0.625556	0.729567	0.67357

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25375
Method:	MLE	Df Model:	16
Date:	Thu, 16 Jan 2025	Pseudo R-squ.:	0.3238
Time:	06:52:32	Log-Likelihood:	-10859.
converged:	True	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

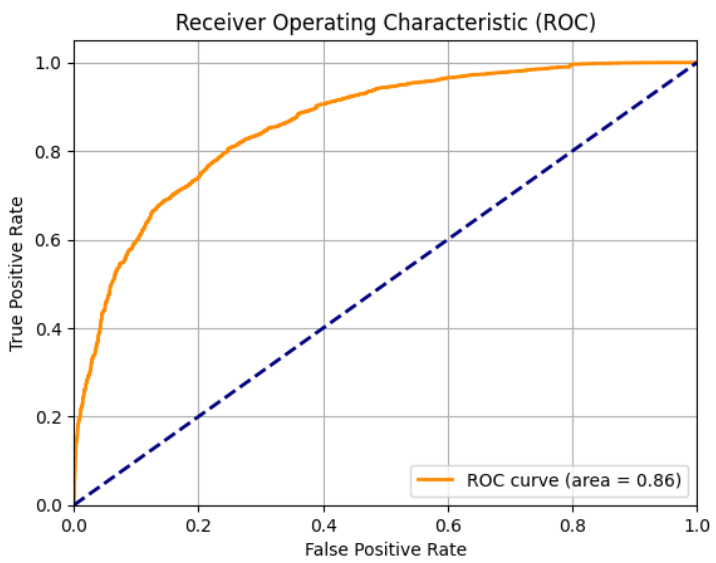
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7474	0.093	-29.561	0.000	-2.930	-2.565
no_of_weekend_nights	0.1516	0.020	7.708	0.000	0.113	0.190
no_of_week_nights	0.0324	0.012	2.662	0.008	0.009	0.056
required_car_parking_space	-1.6427	0.137	-11.959	0.000	-1.912	-1.373
lead_time	0.0170	0.000	65.306	0.000	0.016	0.017

```
arrival_month                       -0.0675      0.006    -11.311      0.000      -
0.079      -0.056
repeated_guest                      -2.9420      0.521     -5.644      0.000      -
3.964      -1.920
no_of_previous_cancellations         0.2796      0.074      3.760      0.000
0.134       0.425
avg_price_per_room                   0.0207      0.001     28.149      0.000
0.019       0.022
no_of_special_requests              -1.4598      0.030    -49.064      0.000      -
1.518      -1.402
room_type_reserved_Room_Type 2      -0.4215      0.127     -3.322      0.001      -
0.670      -0.173
room_type_reserved_Room_Type 4      -0.3028      0.050     -6.010      0.000      -
0.402      -0.204
room_type_reserved_Room_Type 5      -0.6431      0.211     -3.044      0.002      -
1.057      -0.229
room_type_reserved_Room_Type 6      -0.5089      0.114     -4.462      0.000      -
0.732      -0.285
room_type_reserved_Room_Type 7      -0.7638      0.274     -2.790      0.005      -
1.300      -0.227
market_segment_type_Corporate       -0.9073      0.102     -8.908      0.000      -
1.107      -0.708
market_segment_type_Offline         -1.8099      0.048    -37.828      0.000      -
1.904      -1.716
==============================================================================
```

Now no categorical feature has p-value greater than 0.05, so we'll consider the features in *x_train6* as the final ones and *LogisticReg_tuned* as final model.

Determining optimal threshold using ROC Curve:

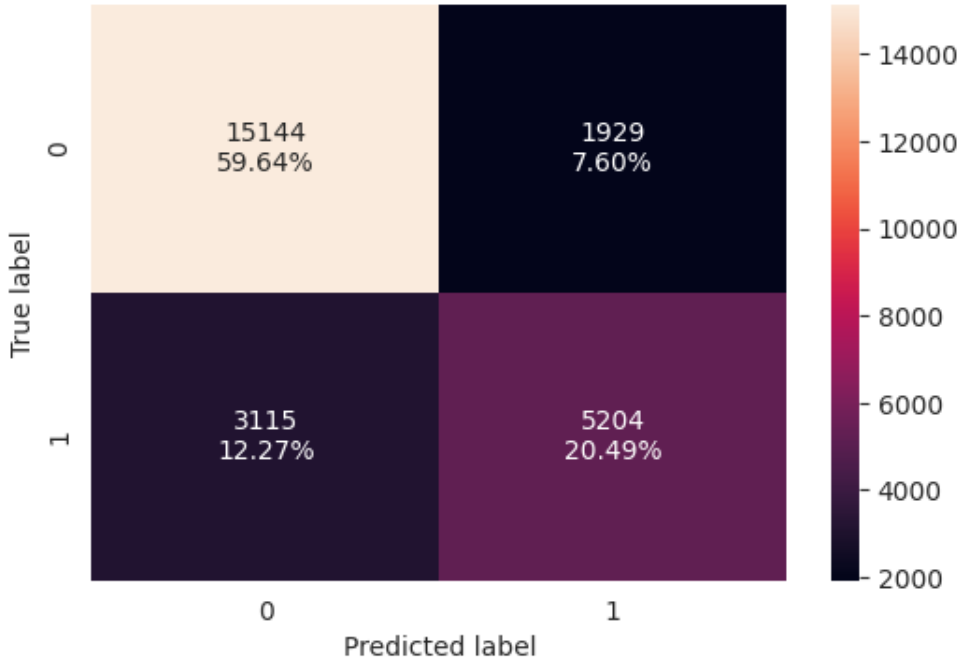
- Let's see if the f1_score can be improved further by changing the model threshold
- First, we will check the ROC curve, compute the area under the ROC curve (ROC-AUC), and then use it to find the optimal threshold
- Next, we will check the Precision-Recall curve to find the right balance between precision and recall as our metric of choice is f1_score



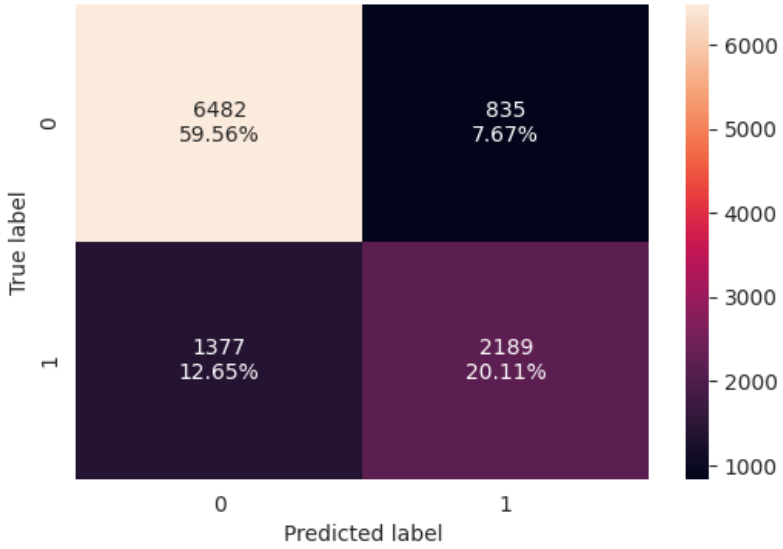
Optimal Threshold: 0.293

Final model on train data set:

	Accuracy	Recall	Precision	F1
0	0.770006	0.804904	0.61358	0.696339



Final model on test data:



	Accuracy	Recall	Precision	F1
0	0.767527	0.802299	0.610542	0.693408

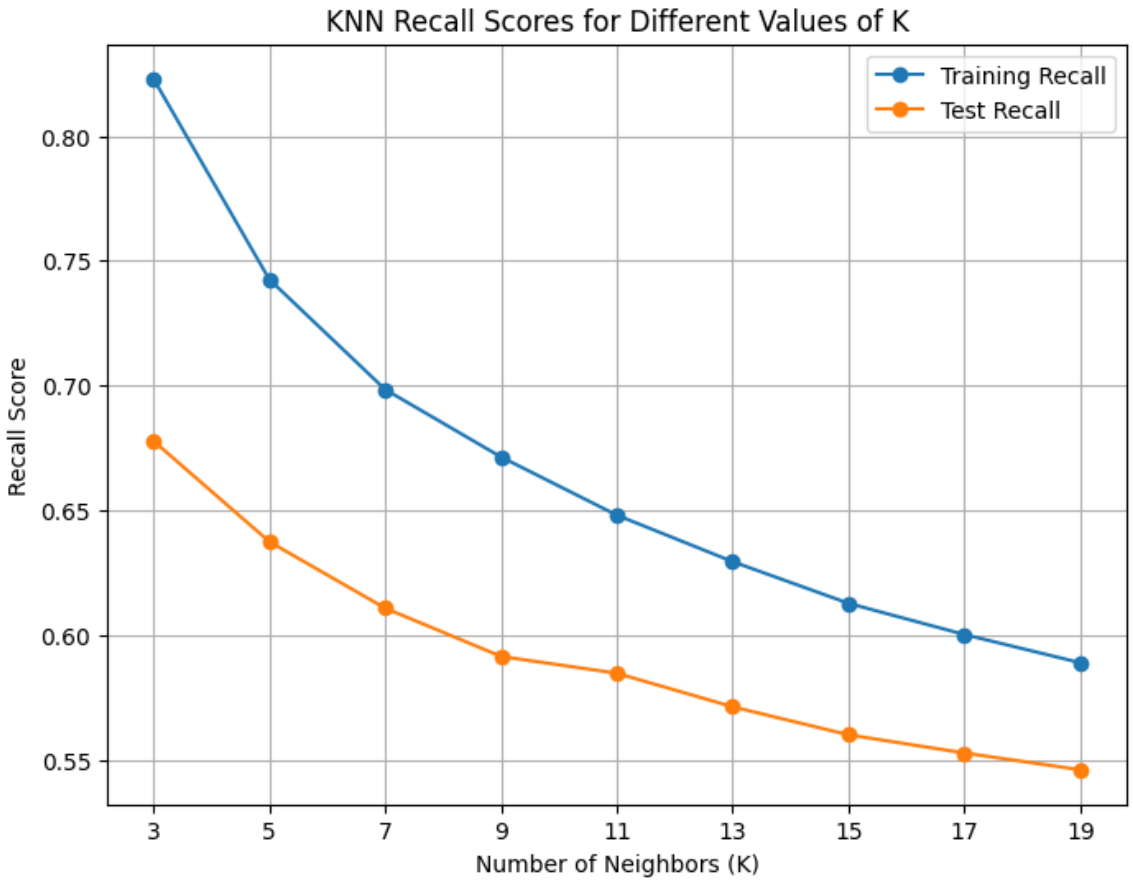
Comments:

- The model is giving a good f1_score of ~0.67 and ~0.66 on the train and test sets respectively
- As the train and test performances are comparable, the model is not over fitting
- Moving forward we will try to improve the performance of the model

greatlearning
Learning for Life

KNN

KNN recall scores for different values of k



Comments:

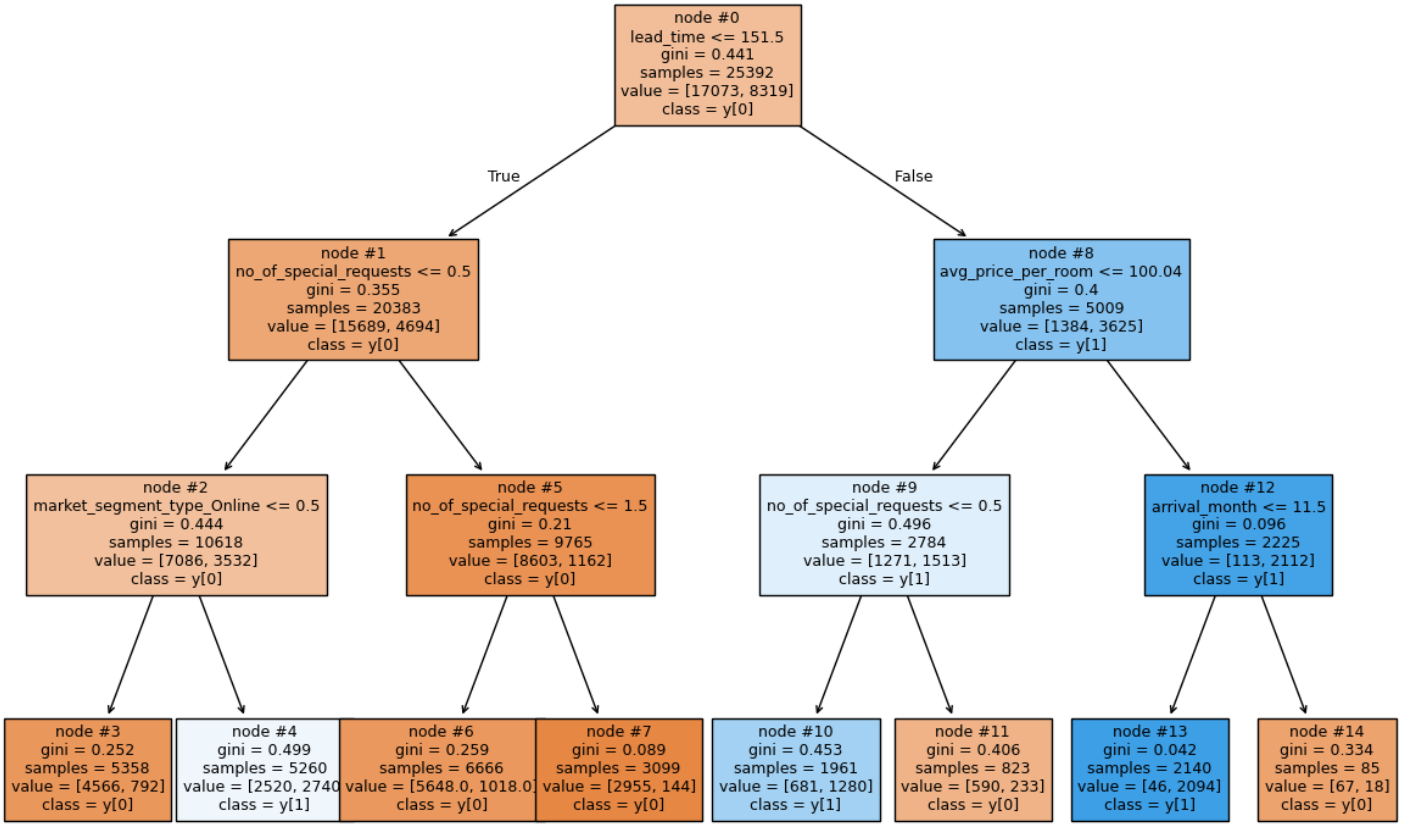
- The recall scores for both training and test sets are highest when k=3. This suggests that with k=3, the model is better at identifying positive instances in both the training and test data compared to other values of k.
- As the value of k increases beyond 3, the recall scores tend to decrease for both training and test sets. This indicates a potential risk of the model not being able to identify the underlying patterns in the data.
- Therefore, based on the provided recall scores, k=3 appears to be the most suitable choice for balancing model performance between capturing positive instances effectively and generalizing well to new data.

greatlearning
Learning for Life

Decision Tree Classifier

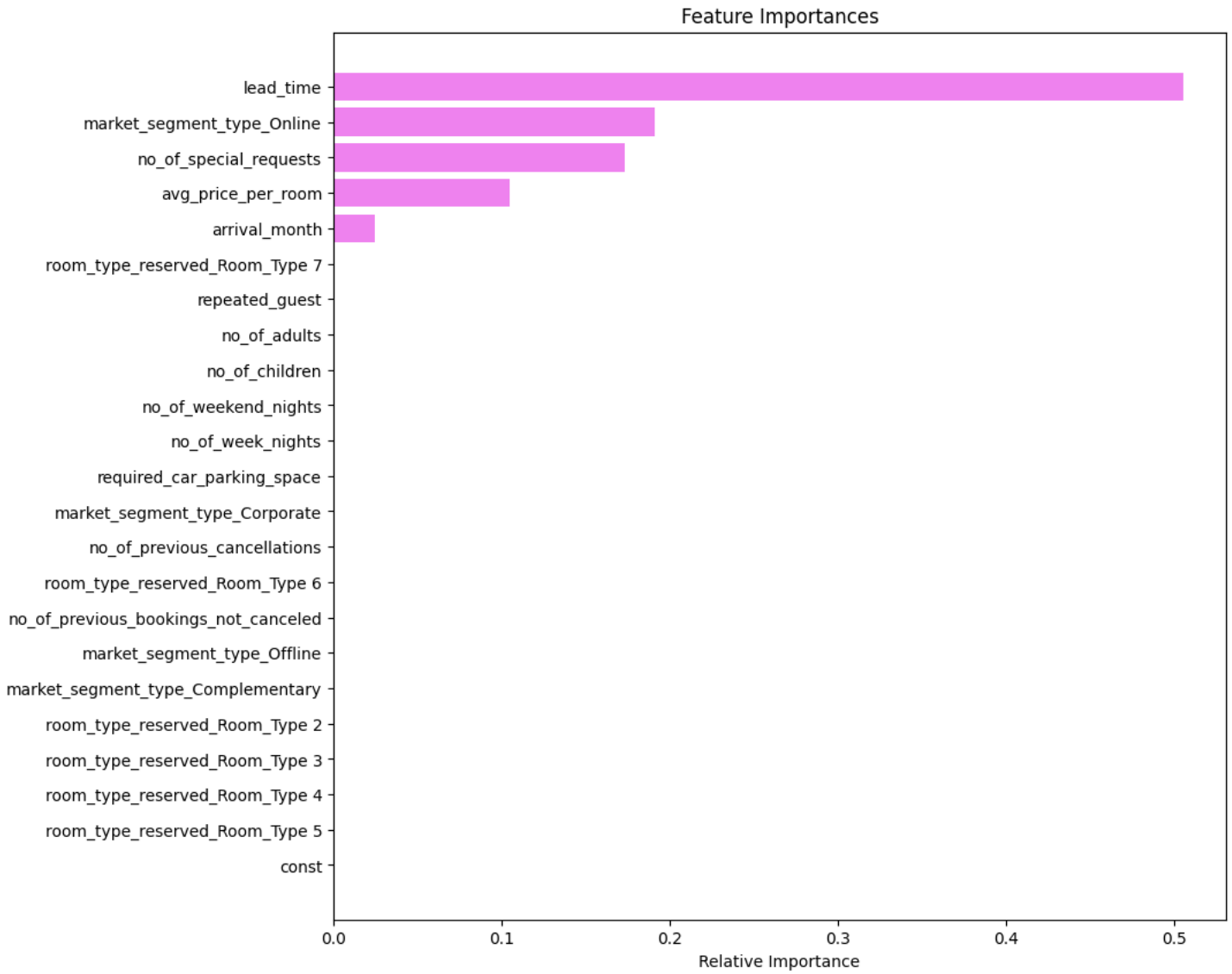
Pre-pruned decision tree:

- In general, the deeper you allow your tree to grow, the more complex your model will become because you will have more splits and it captures more information about the data and this is one of the root causes of over fitting
- Let's try Limiting the max depth of tree to 3



- The tree has become readable now

Feature engineering:



- You can see in important features of previous model and here lead time on top.
- That's why we will go for pre pruning using grid search, maybe setting max depth to 3 is not good enough
- It is bad to have a very low depth because your model will under fit
- Let's see how to find the best values

greatlearning
Learning for Life

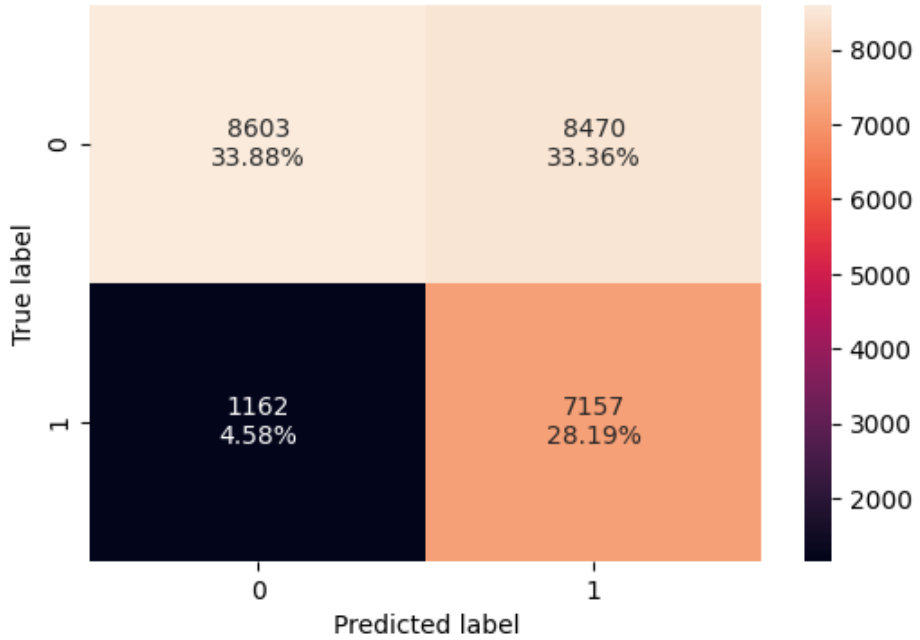
Using GridSearch for Hyperparameter tuning of our tree model

- Hyperparameter tuning is also tricky in the sense that there is no direct way to calculate how a change in the hyperparameter value will reduce the loss of your model, so we usually resort to experimentation. i.e we'll use Grid search
- Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters.
- It is an exhaustive search that is performed on the specific parameter values of a model.
- The parameters of the estimator/model used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

Model performance on training dataset:

	Accuracy	Recall	Precision	F1
0	0.620668	0.86032	0.457989	0.597762

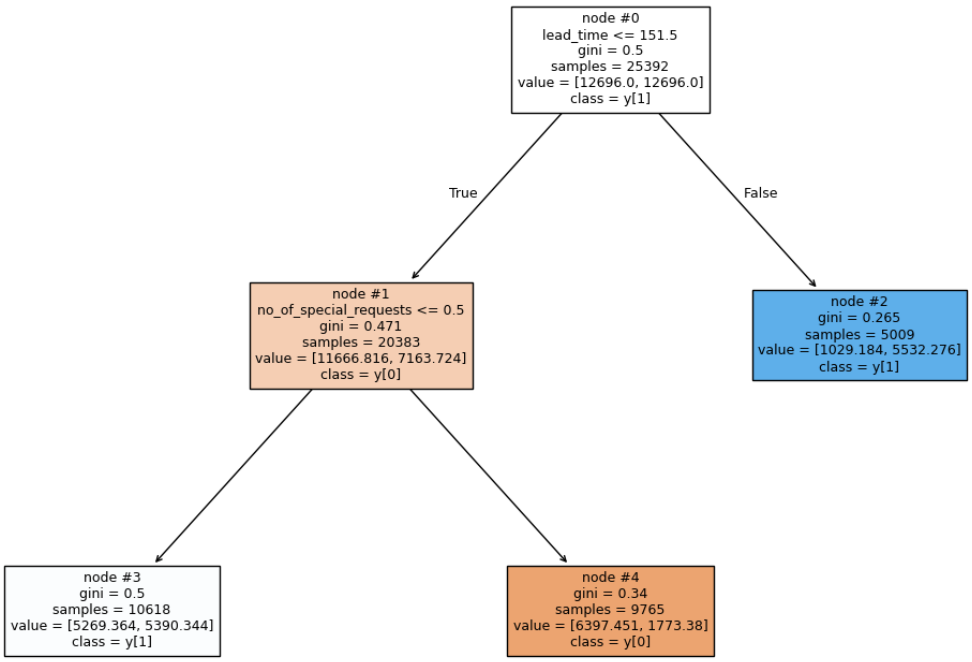
Confusion matrix on training dataset:



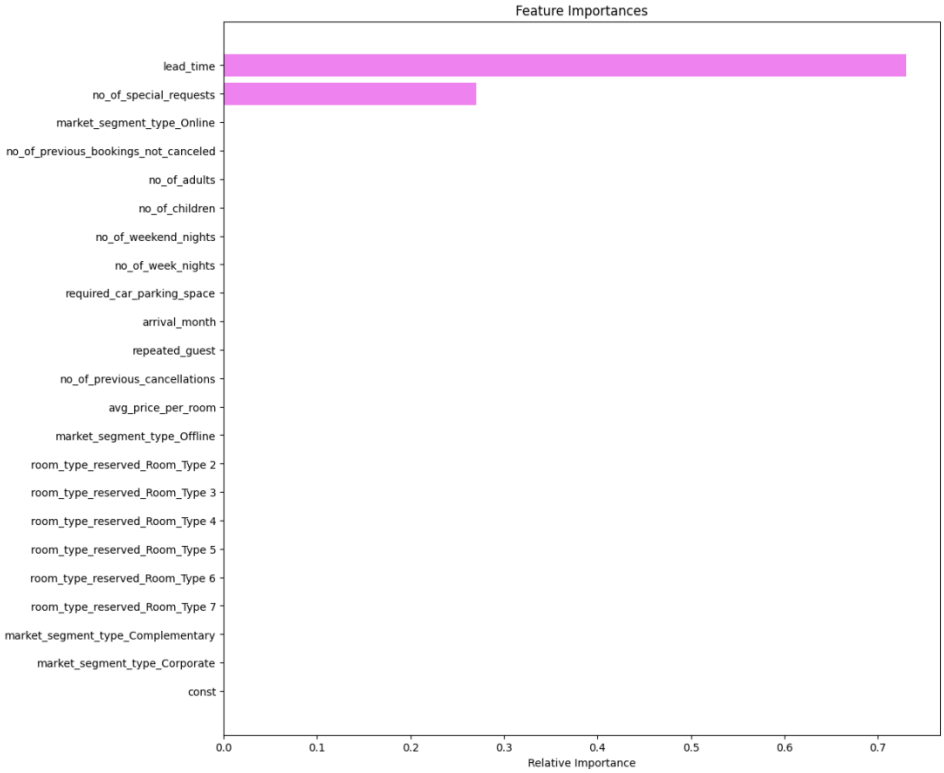
Model performance on testing dataset:

	Accuracy	Recall	Precision	F1
0	0.613434	0.853337	0.452356	0.591276

Visualizing the tree:



Feature importance:



- We can see in important features of previous model,Lead time was on the top and here lead time is on the top.
- But post pruning might give even better results, since there is quite a good possibility that we might neglect some hyperparameters, post pruning will take care of all that.

Post pruning:

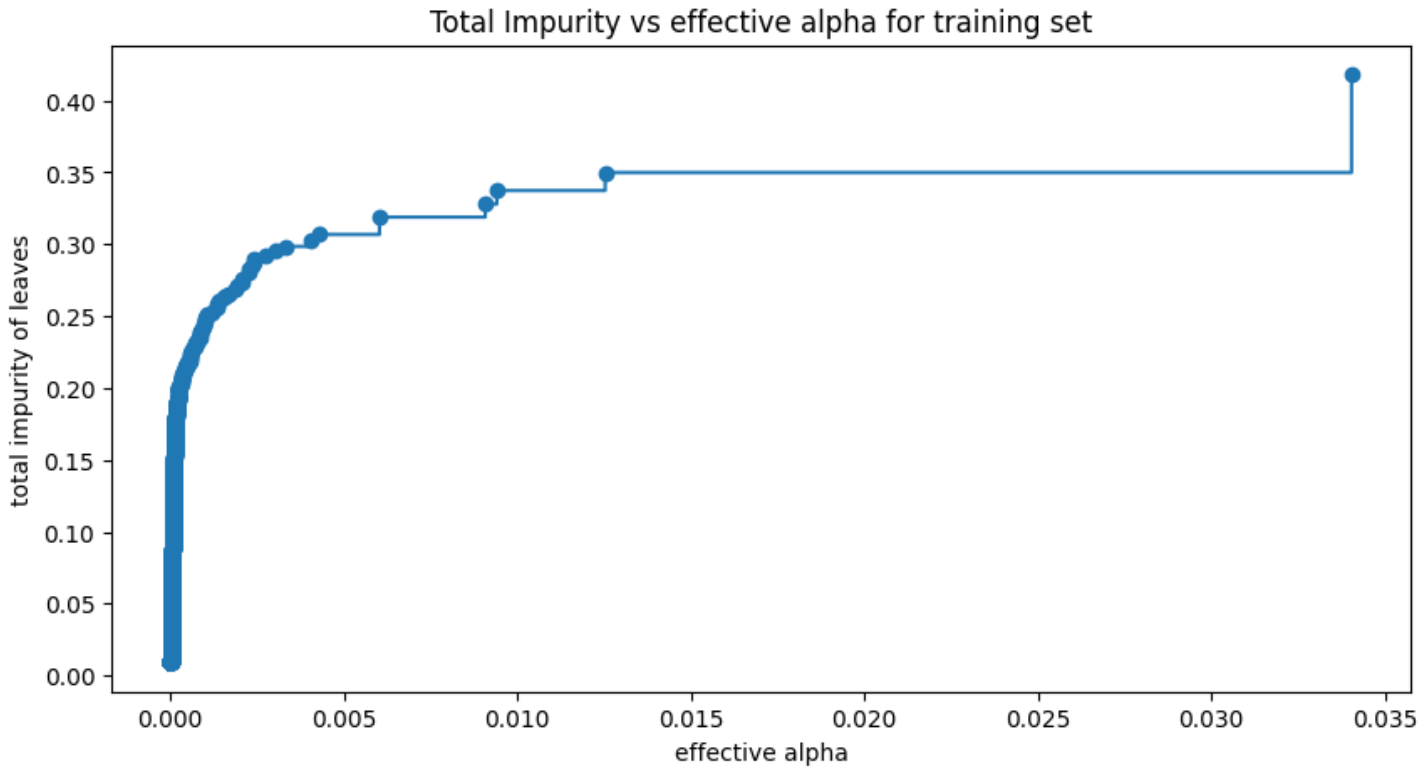
Cost complexity pruning:

The Decision Tree Classifier provides parameters such as min samples leaf and max depth to prevent a tree from over fitting. Cost complexity pruning provides another option to control the size of a tree. In Decision Tree Classifier, this pruning technique is parameterized by the cost complexity parameter, ccp alpha. Greater values of ccp alpha increase the number of nodes pruned. Here we only show the

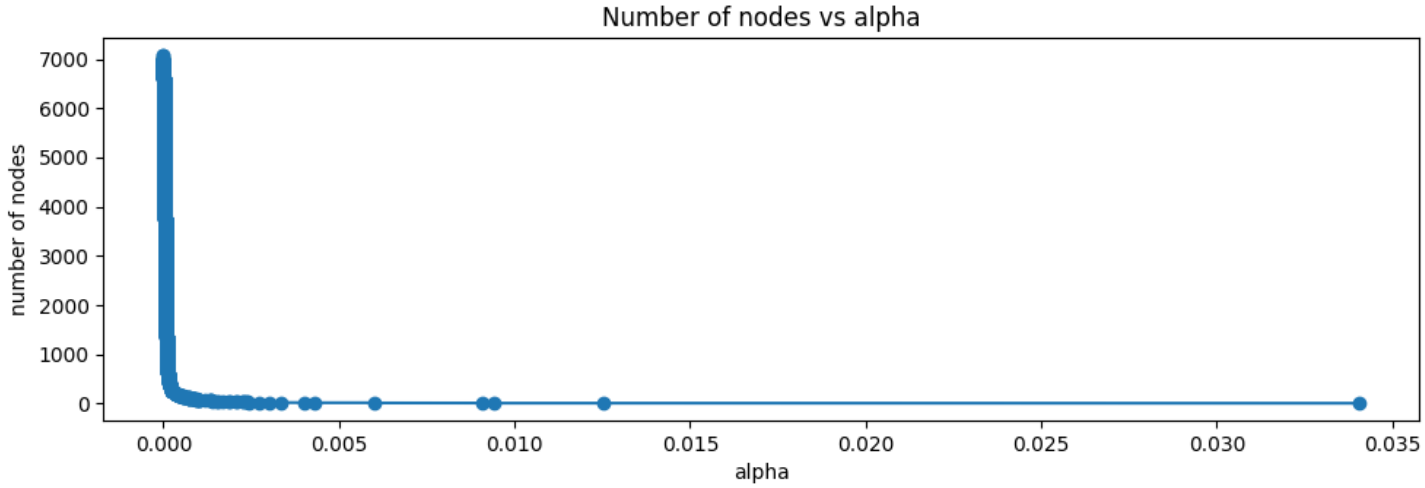
effect of ccp alpha on regularizing the trees and how to choose a ccp alpha based on validation scores.

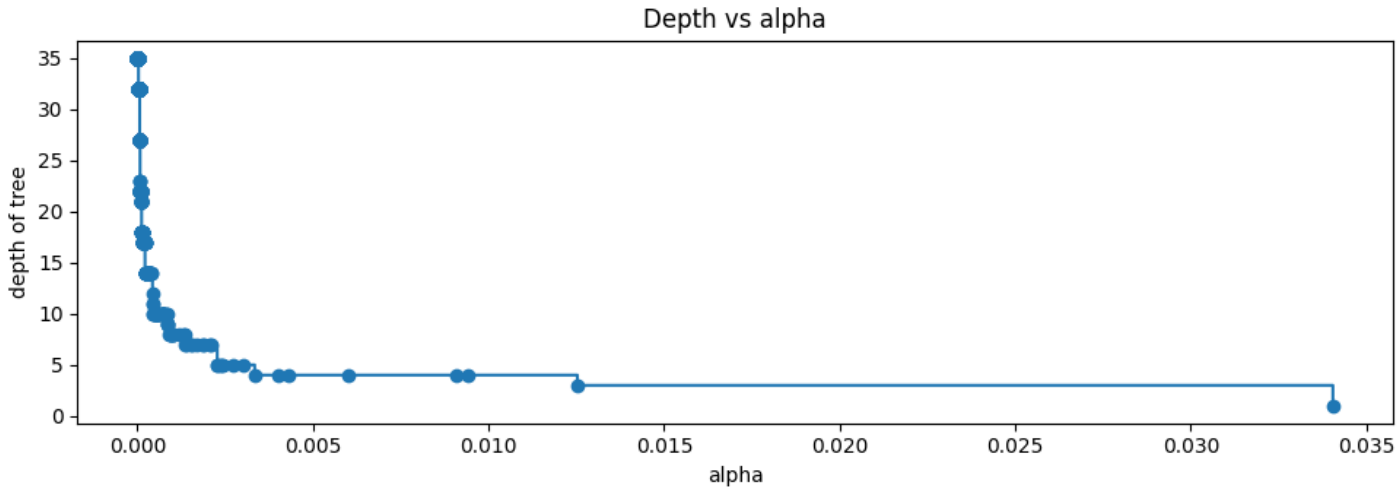
Total impurity of leaves vs effective alphas of pruned tree:

Minimal cost complexity pruning recursively finds the node with the "weakest link". The weakest link is characterized by an effective alpha, where the nodes with the smallest effective alpha are pruned first. To get an idea of what values of ccp alpha could be appropriate, scikit learn provides Decision Tree Classifier cost complexity pruning path that returns the effective alphas and the corresponding total leaf impurities at each step of the pruning process. As alpha increases, more of the tree is pruned, which increases the total impurity of its leaves.



Next, we train a decision tree using the effective alphas. The last value in ccp alphas is the alpha value that prunes the whole tree, leaving the tree, clfs[-1], with one node, because it is the trivial tree with only one node. Here we show that the number of nodes and tree depth decreases as alpha increases.

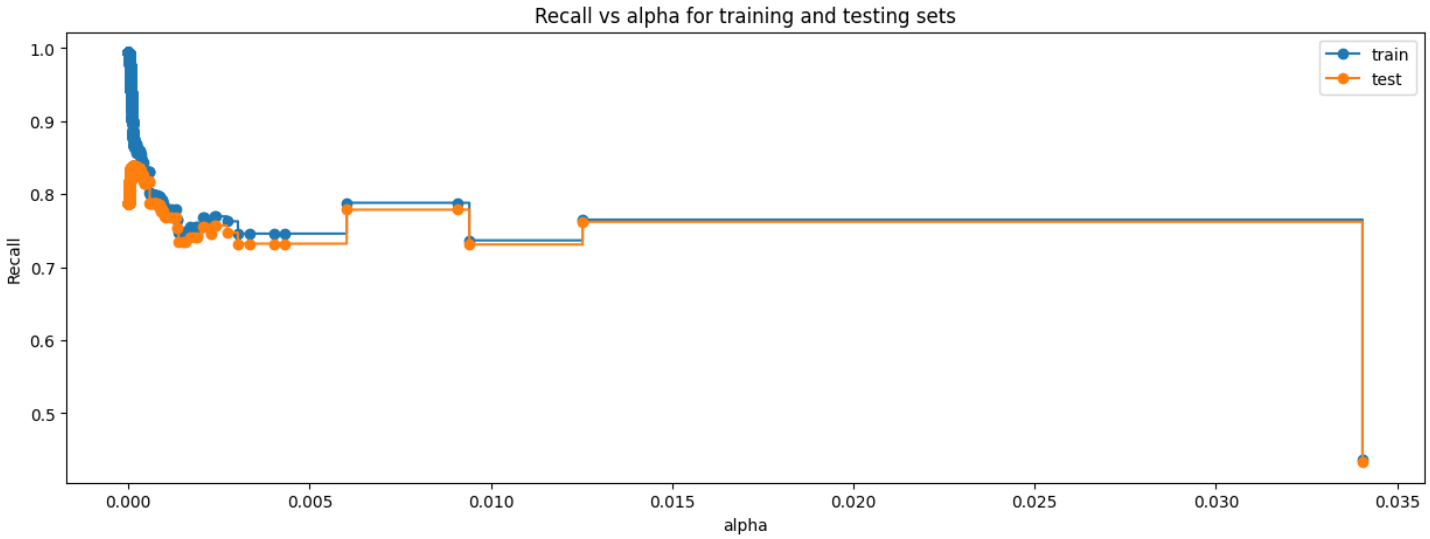




Training accuracy of best model: 0.9760160680529301
Test accuracy of best model: 0.862813562436828

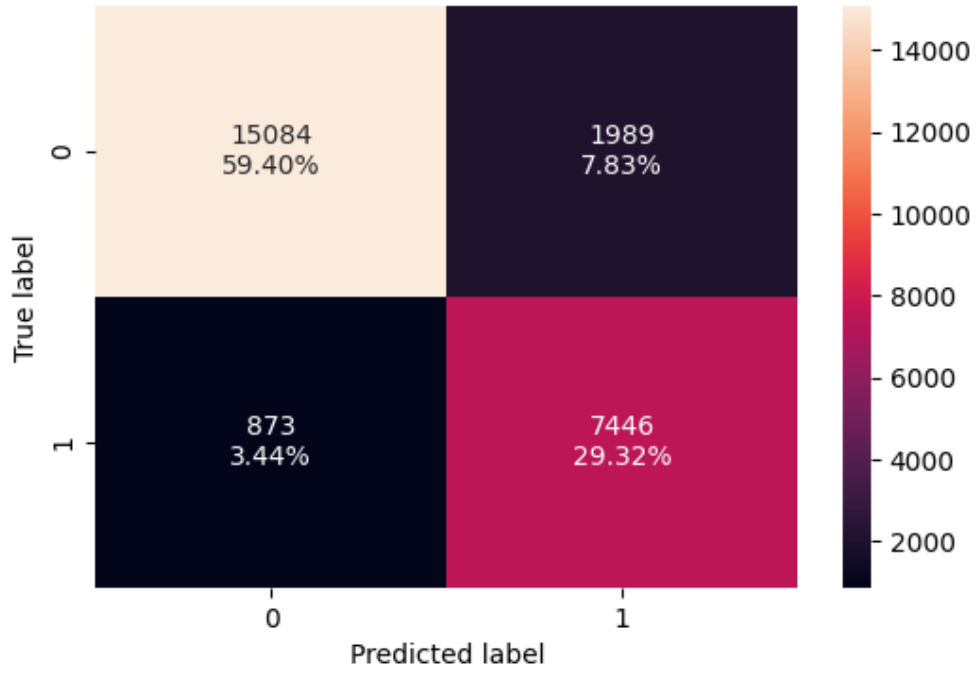
- The tree over fits, leading to a 97% training accuracy and 86% testing accuracy.

Since accuracy isn't the right metric for our data we would want high recall:



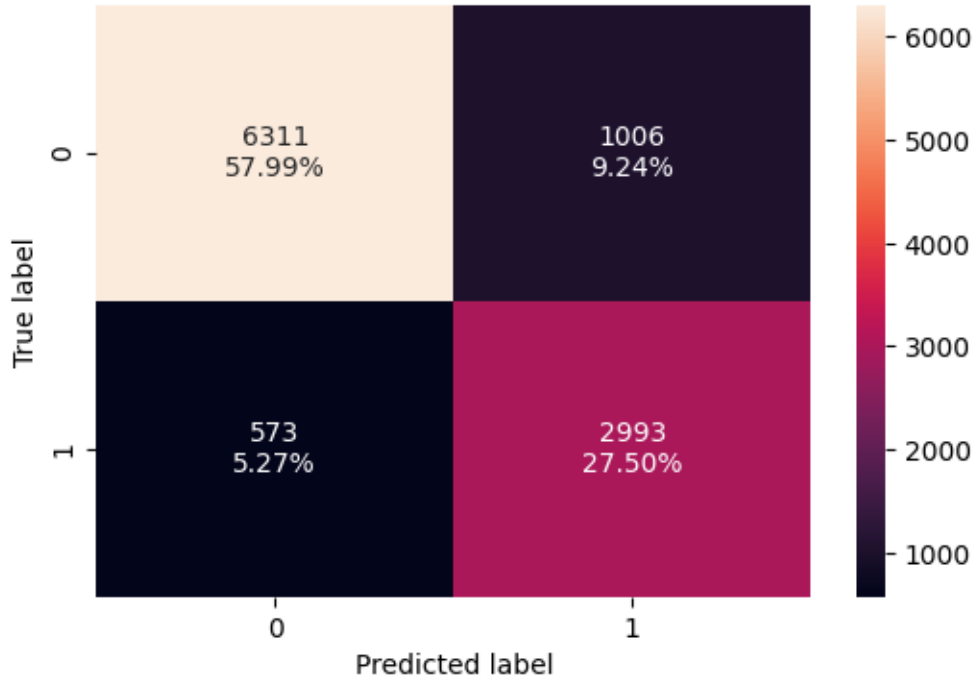
Post pruned Model Performance on the train dataset:

	Accuracy	Recall	Precision	F1
0	0.620668	0.86032	0.457989	0.597762



Post pruned Model Performance on the test dataset:

	Accuracy	Recall	Precision	F1
0	0.6206680	0.86032	0.457989	0.597762



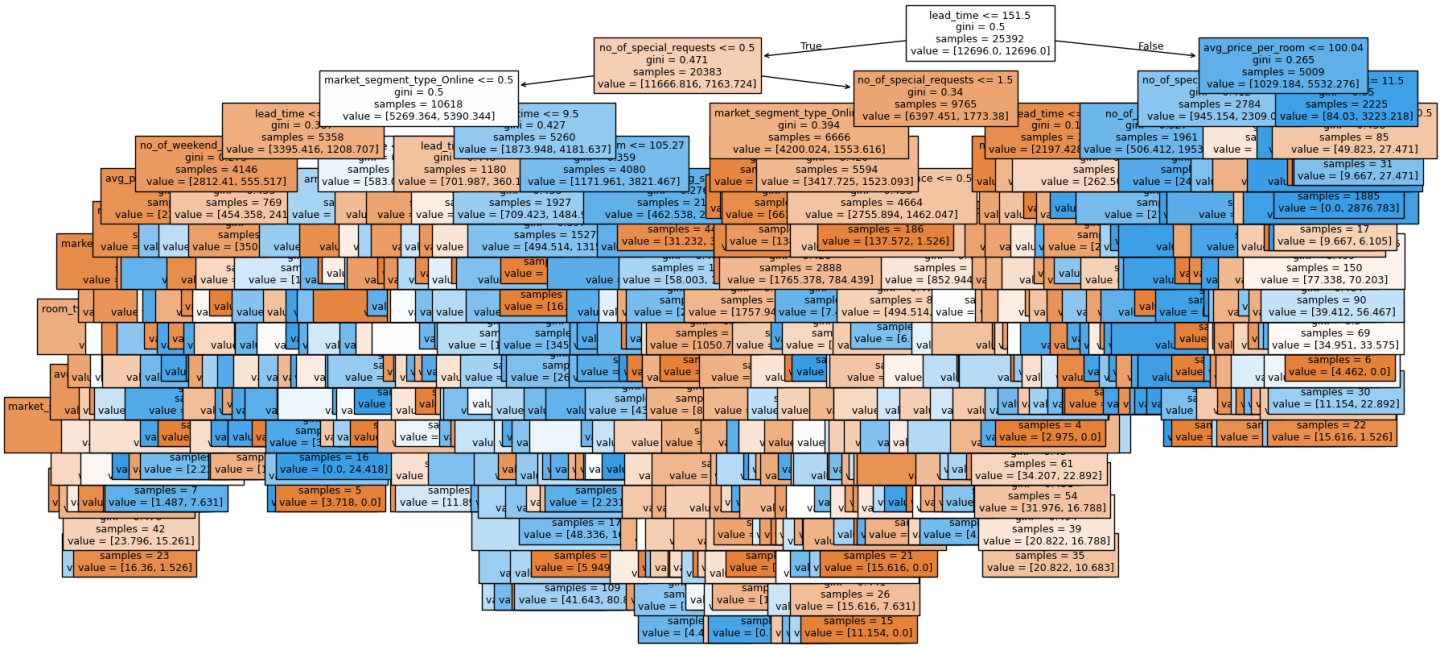
	Accuracy	Recall	Precision	F1
0	0.854911	0.839316	0.748437	0.791276

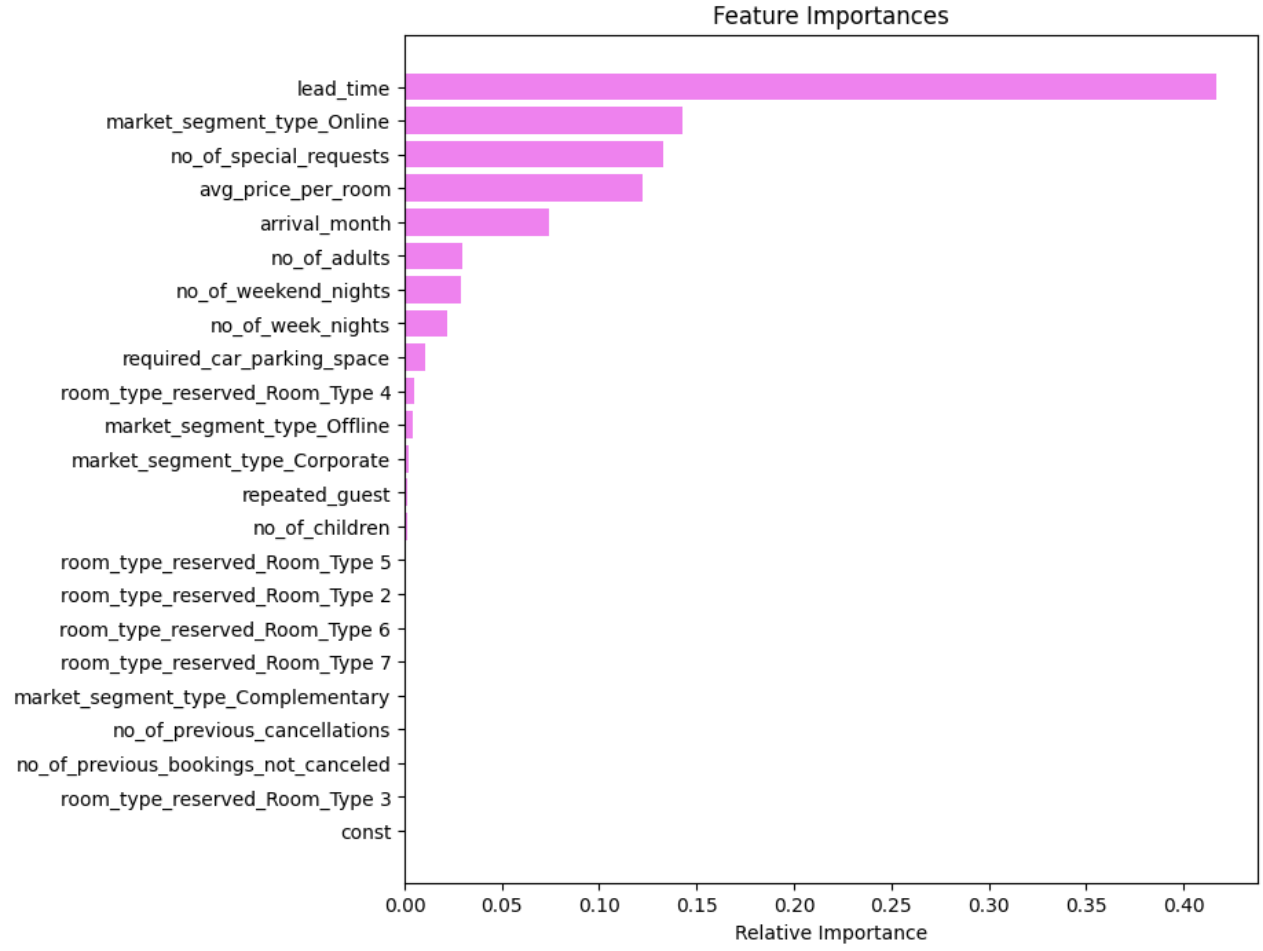
	Accuracy	Recall	Precision	F1
0	0.887287	0.89506	0.789189	0.838797

- With post-pruning we get the highest recall on both training and test set

greatlearning
Learning for Life

Post pruned decision tree:





5. Model performance comparison and Final model selection

Training performance comparison:

	Logistic Regression Base	Logistic Regression Improved	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Pre-Pruned	Decision Tree Post-Pruned
Accuracy	0.802733	0.770006	0.408948	0.915288	0.915288	0.993620	0.620668	0.887287
Recall	0.630484	0.804904	0.964659	0.848299	0.848299	0.983772	0.860320	0.895060
Precision	0.730501	0.613580	0.352918	0.888120	0.888120	0.996712	0.457989	0.789189
F1	0.676818	0.696339	0.516775	0.867753	0.867753	0.990200	0.597762	0.838797

Test set performance comparison:

	Logistic Regression Base	Logistic Regression Tuned	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Pre-Pruned	Decision Tree Post-Pruned
Accuracy	0.798677	0.767527	0.407516	0.849582	0.849582	0.862170	0.613434	0.854911
Recall	0.618901	0.802299	0.964105	0.745653	0.745653	0.786035	0.853337	0.839316
Precision	0.726226	0.610542	0.352326	0.784597	0.784597	0.791808	0.452356	0.748437
F1	0.668282	0.693408	0.516061	0.764630	0.764630	0.788911	0.591276	0.791276

Observations

- Logistic Regression tuned is much better than logistic regression base
- Naive bayes base is not overfitting but not performing well.
- KNN base is overfitting, performing well on train data and not performing well on test data. KNN tuned also is not performing well on test data.
- Decision tree model with default parameters is overfitting the training data and is not able to generalize well.
- Post pruned tree has performing well on train and test on accuracy, recall, precision, F1.
- But pre-pruned tree has performing well on recall compared to post pruned,Pre-pruned tree has given a generalized performance with the recall score of 0.86 and 0.85 on training and test set, respectively.
- The company can predict the interested leads better using the pre-pruned tree.

greatlearning
Learning for Life

6. Actionable insights and recommendations:

6.1. Actionable Insights

Target the right segment: The campaign should target customers online, where the highest bookings are made even after getting highest price on booking but the cancelation is also high in online segment.Corporate has less number of bookings and near to zero cancelation.

Target the right month: OCTOBER, SEPTEMBER, AUGUST are month with highest bookings.

Target the repeated customers: Those are repeated customers tend to cancel less.

Number of special request: As the special request increases the cancelation decreases.

Required parking space: Those who require a parking space are tend to cancel less.

Number of previous cancelation: Number of previous cancelation affect the cancelation.

6.2. Recommendations

Feature Engineering: Explore creating new features based on existing ones.

Focus on the most promising customer segments: Concentrate marketing efforts on customers with a higher special requests, repeatative customers.

Refine customer understanding: Dig deeper into customer behavior and preferences. This can help tailor marketing messages and product offerings for maximum impact.

Test and iterate: Continuously evaluate the model's performance and make adjustments as needed. Experiment with different targeting strategies and messaging to optimize results.

Invest in data quality: Ensure the data used to build the model is accurate, complete, and up-to-date. High-quality data will lead to more reliable predictions.
