# SLC DSBA

## INN Hotels and PGP-DSBA

22/08/2023

# Contents / Agenda

# Executive Summary

**Insights and Recommendations:-**

**Insights:-**

- There are the three most important variables in cancellation are lead time, how advance they have booked the room, special request for the stay, and average price of the room.

- It has been assumed that rooms booked over 151 days are more likely to cancel.

- It is also determined that price was a determine factor for those cancellations.

- Rooms booked in advance are less likely to cancel.

# Executive Summary

- **Recommendations:-**

  - Require a non-refundable deposit on all rooms in advance over 5 months.

  - Replace the full board option on the booking with a menu of special request available , instead of waiting for them to come in sell them even if they are no charge.

  - Seasonal high prices may peak to early OCT.

  - During online booking process it can offer additional customizations that would be helped in the likelihood of cancelled booking.

# Business Problem Overview and Solution Approach

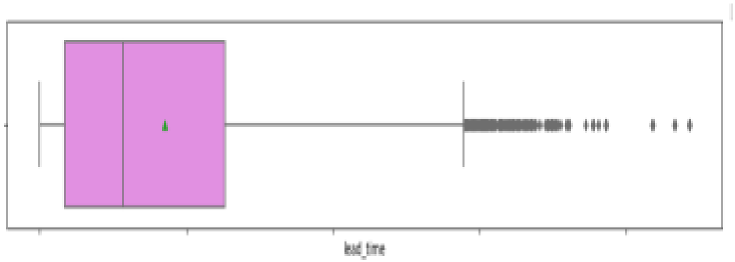
- A sizable portion of hotel reservations are canceled or missed owing to rescheduling issues, plan changes, etc.
- Although these cancellations are done for free or preferable at a minimal cost, which is advantageous to hotel customers, it is less desirable and a factor that reduces revenue for INN hotels. These losses are especially substantial with regard to last-minute cancellations.
- The use of online booking channels has undergone a significant change as a result of new technology.
- To analyze the information, discover which components have a tall impact on booking cancellations, construct a prescient show that can anticipate which booking is getting to be canceled in development, and offer assistance in defining productive arrangements for cancellations and discounts.

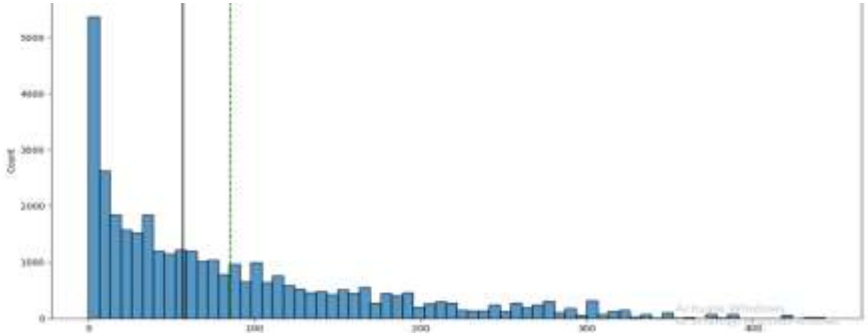
# EDA Results

- In this EDA results of the data set it have 36275 rows and 19 columns with no missing duplicate values. So from that we can believe that it is well managed data set.
- 5 columns are object data sets which is Booking_id, type_of_meal_plan, room_type observed, and markey_segment_type.
- The last is  booking status which is testing result and is Boolean but we do not have to deal with it.

❑ Univariate Analysis:-

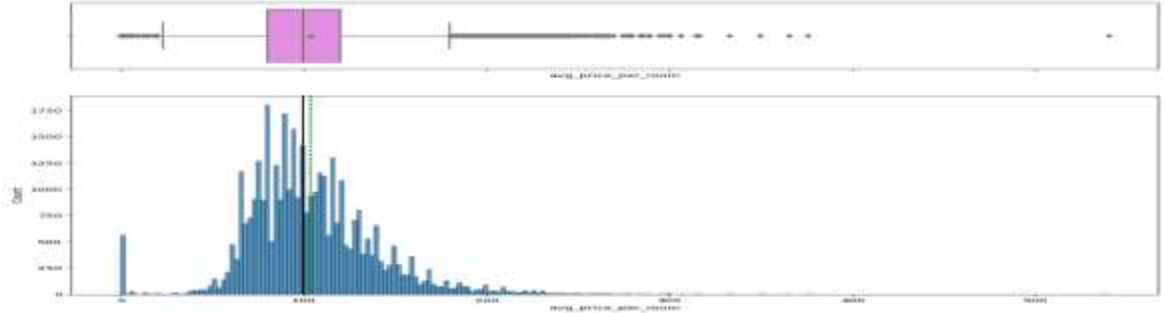
Observations and Lead time:



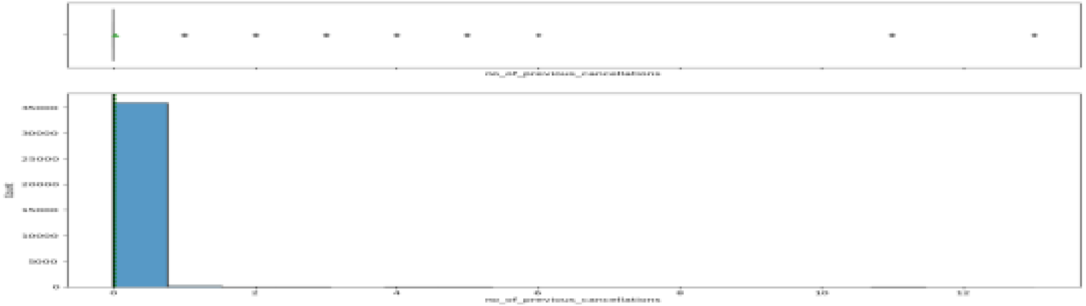



# EDA Results

- Observations on average price per room:-



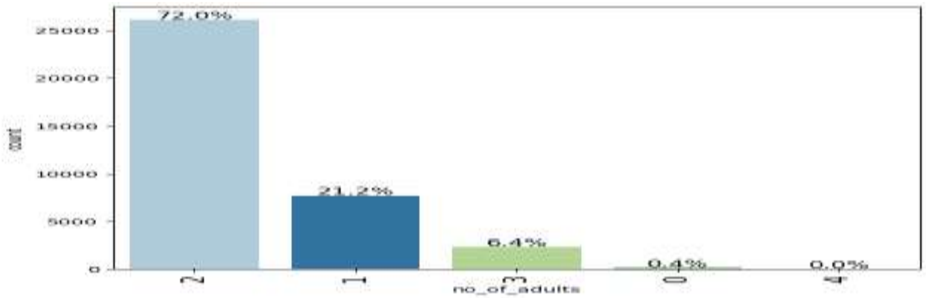
- Observations on number of previous booking cancellations



# EDA Results

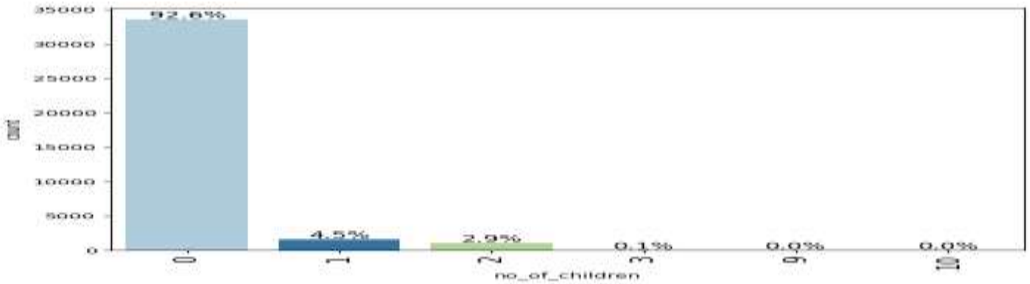
- Observations on number of previous booking not canceled:-

- Observations on number of adults

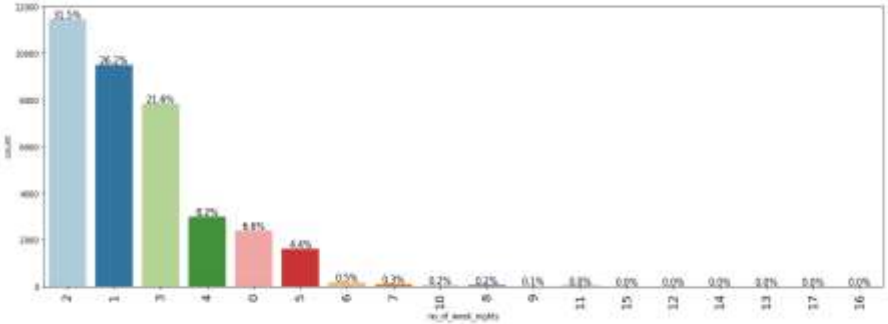


# EDA Results

- Observations on number of children:-

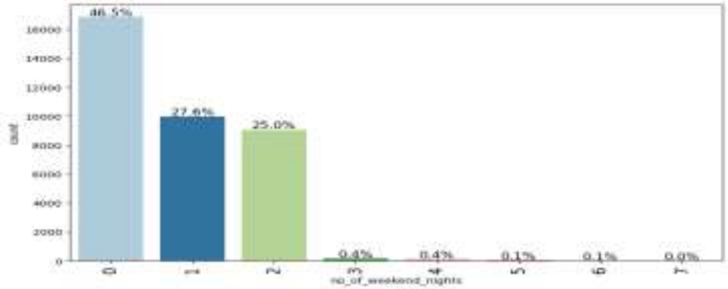



- Observations on number of week nights

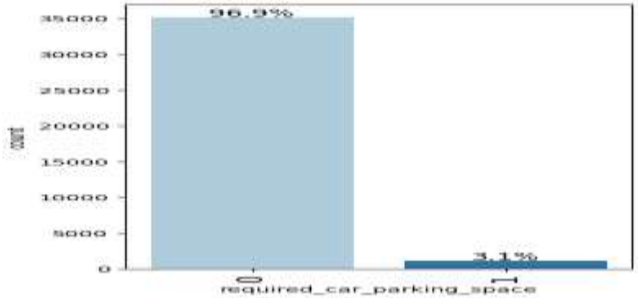


# EDA Results

- Observations on number of weekend nights:-

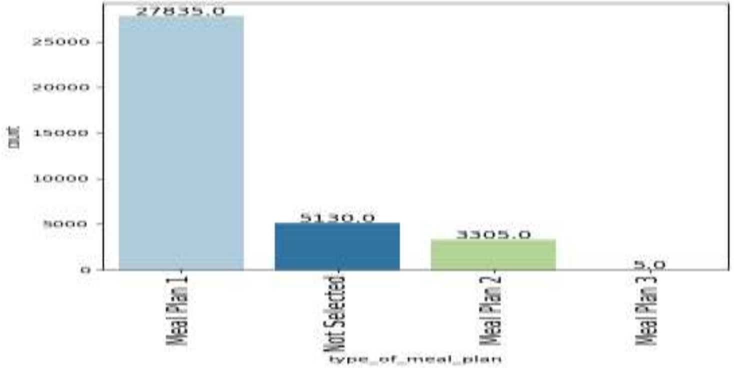



- Observations on required car parking space:-

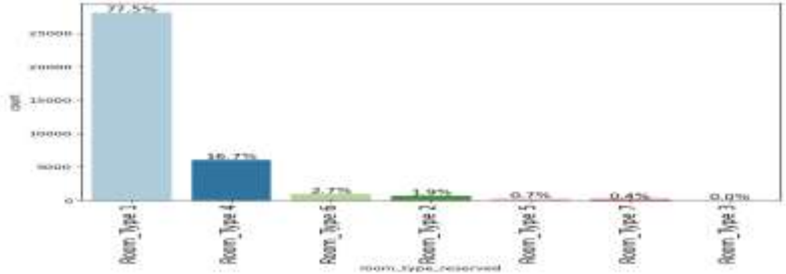


# EDA Results

- Observations on type of meal plan:-

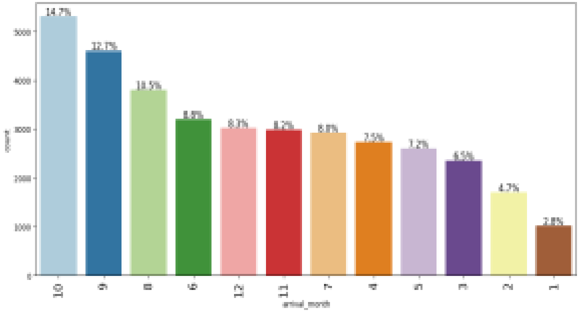



- Observations on room type reserved

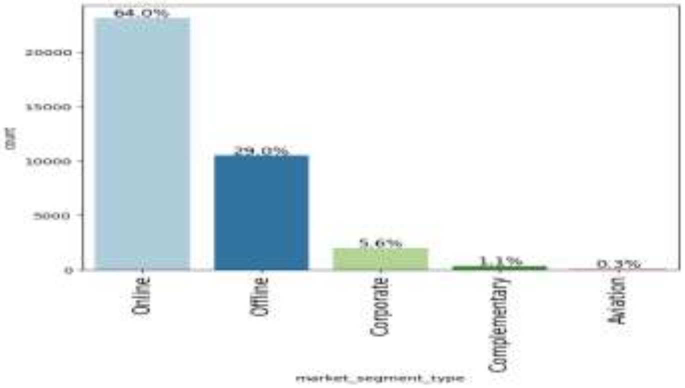


# EDA Results

- Observations on arrival month:-

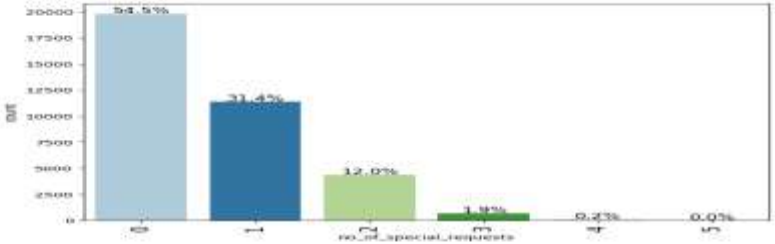


- Observations on market segment type

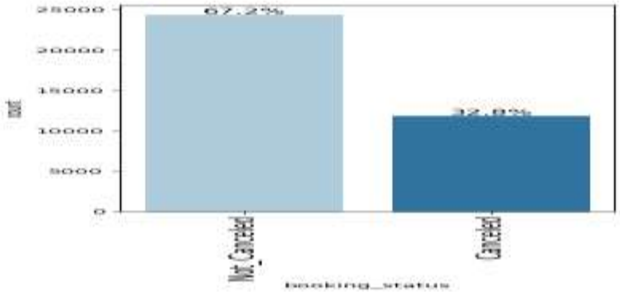


# EDA Results

- Observations on number of special requests:-

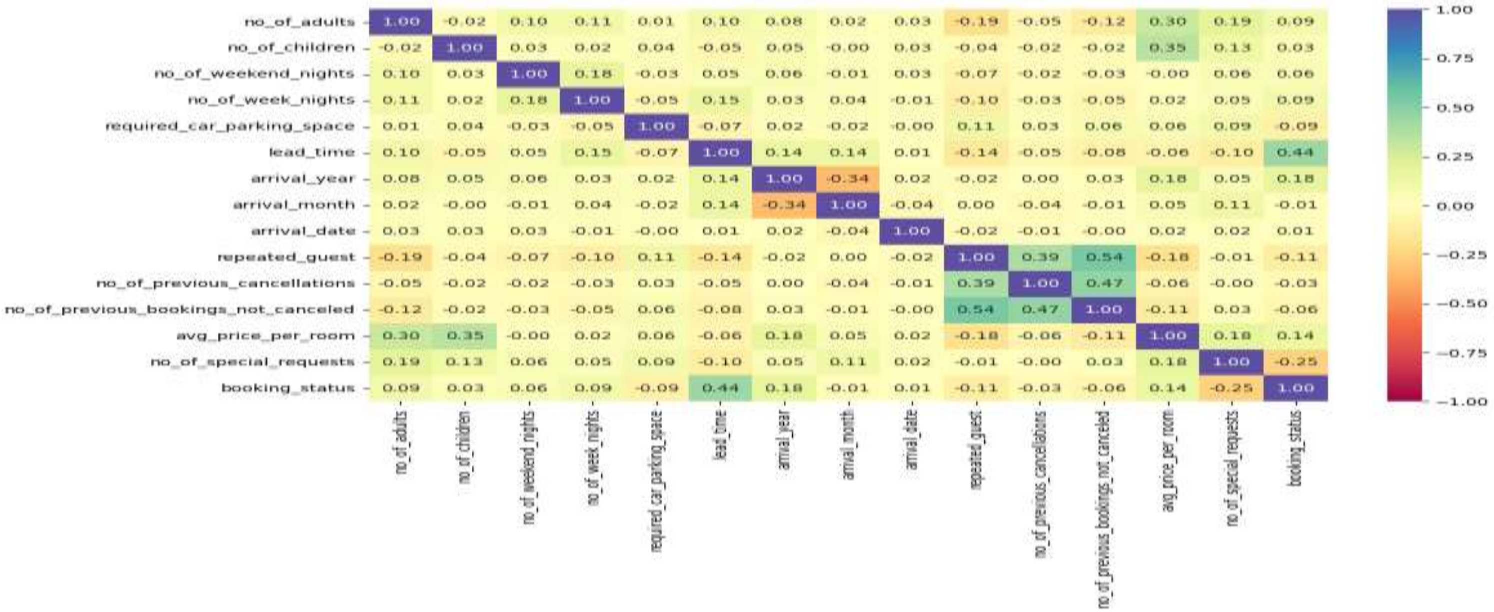



- Observations on booking status



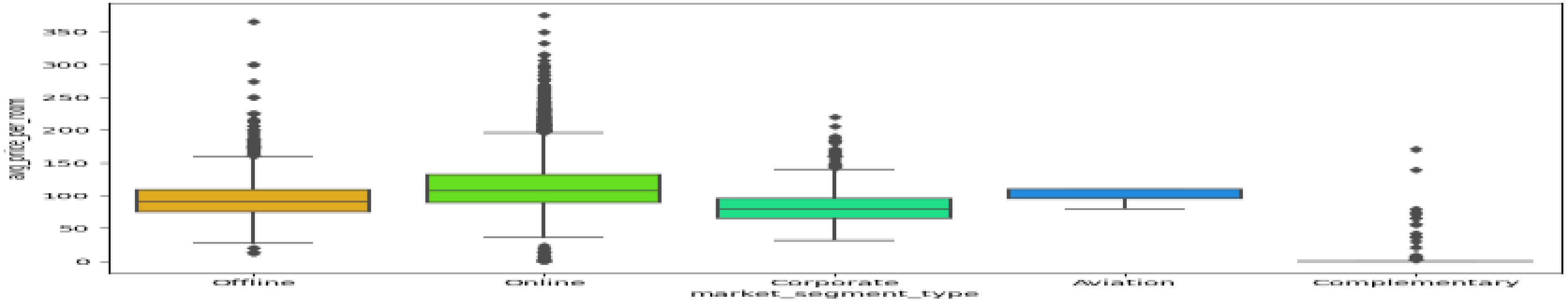
# EDA Results

- Bivariate analysis:-



# EDA Results

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

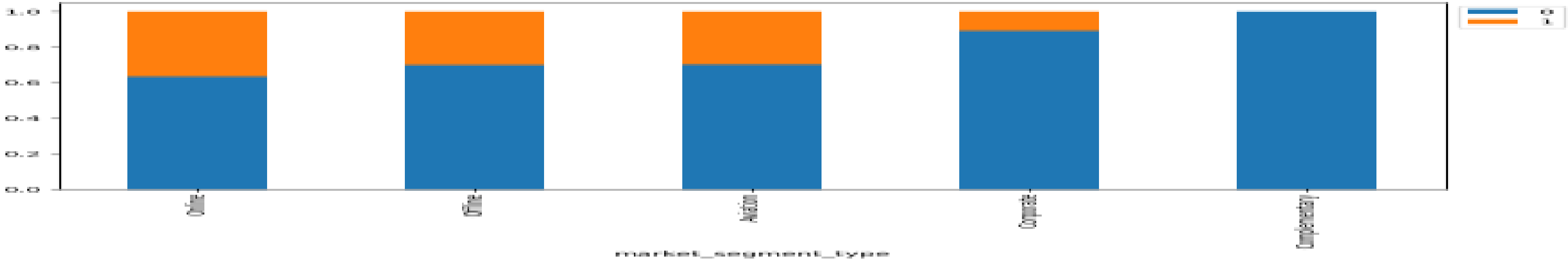



- **Observation:-**

Online booking are the highest despite also having the highest amount of free rooms (I suppose they are redeemed from online retailers points systems) Aviation, Offline, and Corporate are generally slightly lower priced with Corporate edging out for the lowest. Complimentary are of course free.

# EDA Results

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

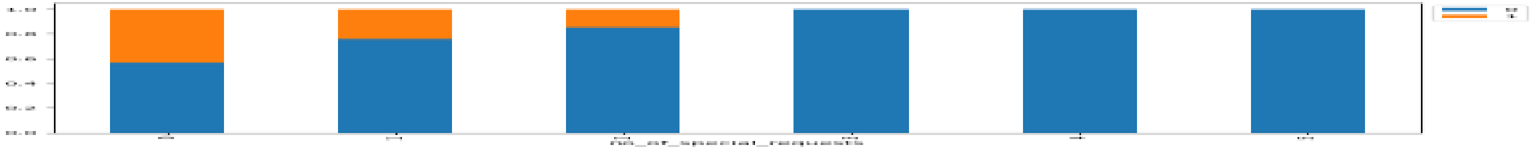


**Observation:-**

We can observe here that according to market segment and booking status it has online offline, aviation, corporate, and complimentary booking as per the average price per room.

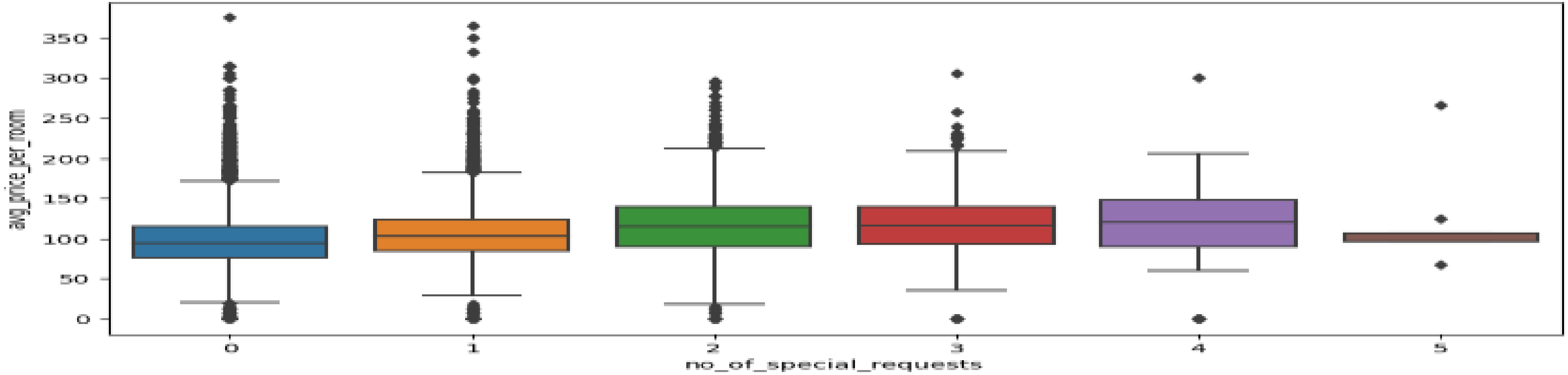
# EDA Results

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



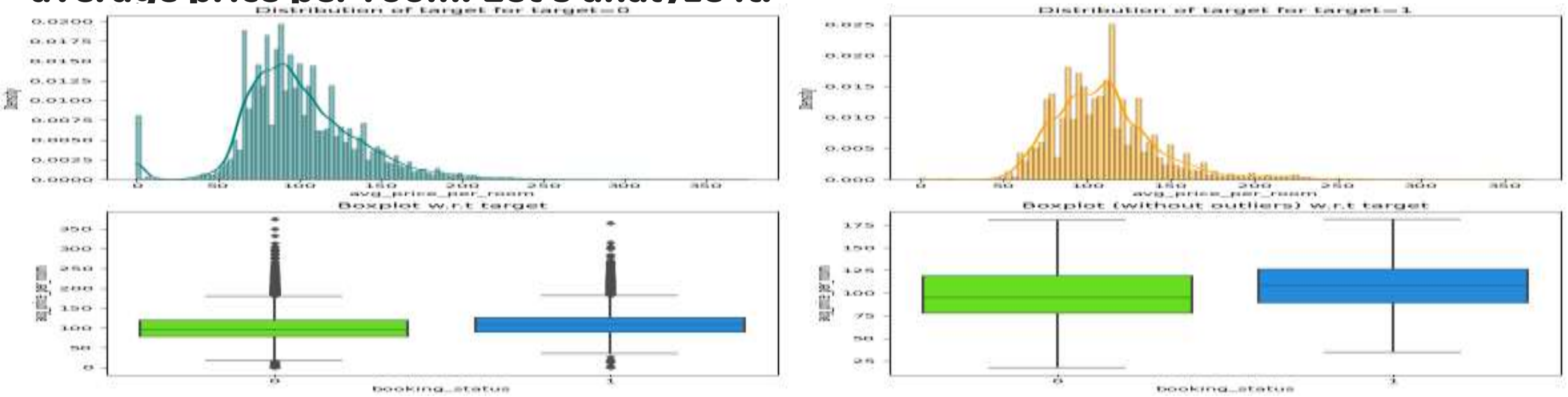
The absence of special request increases the likelihood of cancellation, the addition of special request begins to reduce the likelihood of cancellation at one and progressively reduces cancellation to Zero on the instance of a third request.

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



# EDA Results

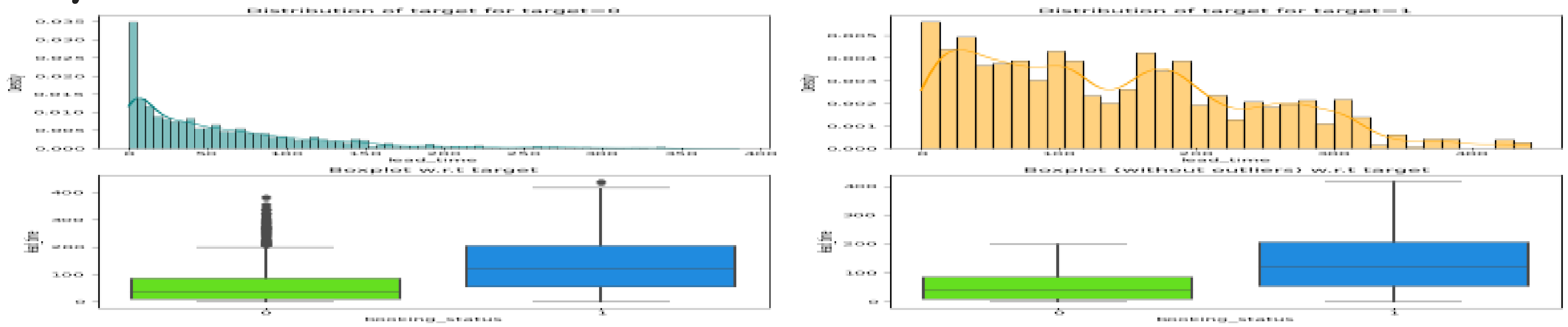
- **We saw earlier that there is a positive correlation between booking status and average price per room. Let's analyze it:-**




We can observe here that there is a positive correlation and even through a boxplot we can analyze as average price per room has also impacted on the booking status as we can see in the graph.

# EDA Results

- **There is a positive correlation between booking status and lead time also. Let's analyze it further.**



We have observed here that there is positive correlation as well as from the graphs we can conclude that distribution of lead time within the target groups shows the clear picture of the positive correlation.

# EDA Results

- **Let's do a similar analysis for the customer who stay for at least a day at the hotel.**

**Observations:-**

As the graph has lot of rows and columns it cannot fix in this slide but I have observed that it has 17094 rows and 18 columns.

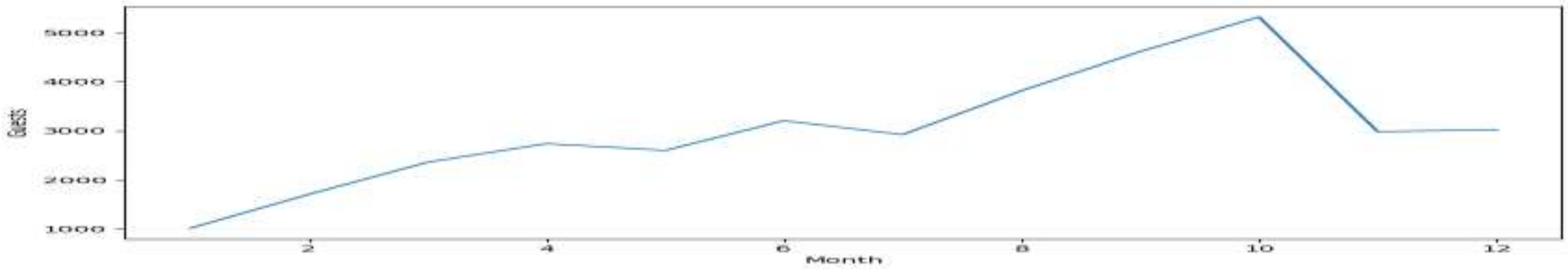
The stacked bar plot includes the lead time and booking status for the no of week nights as well a no of weekend nights.

- **Repeating guests are the guests who stay in the hotel often and are important to brand equity. Let's see what percentage of repeating guests cancel?**


We can see that there are repeating guests cancel more and that also impacted the hotel brand equity.

# EDA Results

- **Let's find out what are the busiest months in the hotel.**




**We have observed here that month 10 with 14.7%of the total booking of the year has the busiest.**

# EDA Results

- **More EDA insights:-**

  - Of 36275 rooms rentals 545 are free of charge over the course of survey.
  - Online booking rooms have the highest cost of booking.
  - Children are rate at hotel,as92.6% of booking don’t include children in rooms.
  - The hotel rarely has long stay guests.
  - Parking is not a factor almost for all of the guests.
  - Nearly 2/3 of bookings come from online.
  - Repeated guests rarely cancel so we can assume that the level of satisfaction through hotel is very high.

# Model Building - Logistic Regression

- **Building Logistic Regression Model:-**

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | booking_status | No. Observations: | 25392 |
| Model: | Logit | Df Residuals: | 25368 |
| Method: | MLE | Df Model: | 23 |
| Date: | Fri, 14 Jan 2022 | Pseudo R-squ.: | 0.2687 |
| Time: | 18:37:31 | Log-Likelihood: | -11767. |
| converged: | True | LL-Null: | -16091. |
| Covariance Type: | nonrobust | LLR p-value: | 0.000 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -3.6818 | 0.098 | -37.650 | 0.000 | -3.873 | -3.490 |
| no_of_adults | 0.2321 | 0.035 | 6.614 | 0.000 | 0.163 | 0.301 |
| required_car_parking_space | -1.4537 | 0.135 | -10.742 | 0.000 | -1.719 | -1.188 |
| arrival_month | -0.0668 | 0.006 | -11.685 | 0.000 | -0.078 | -0.056 |
| repeated_guest | -2.6424 | 0.630 | -4.193 | 0.000 | -3.878 | -1.407 |
| avg_price_per_room | 0.0229 | 0.001 | 33.788 | 0.000 | 0.022 | 0.024 |
| length_stay | 0.1088 | 0.009 | 11.946 | 0.000 | 0.091 | 0.127 |
| no_of_children_log | 0.5488 | 0.093 | 5.887 | 0.000 | 0.366 | 0.732 |
| no_of_previous_cancellations_log | 1.2323 | 0.490 | 2.515 | 0.012 | 0.272 | 2.193 |
| no_of_previous_bookings_not_canceled_log | -0.6731 | 0.477 | -1.411 | 0.158 | -1.608 | 0.262 |
| no_of_special_requests_log | -1.9180 | 0.044 | -43.892 | 0.000 | -2.004 | -1.832 |
| type_of_meal_plan_Meal Plan 2 | -0.3480 | 0.056 | -6.165 | 0.000 | -0.459 | -0.237 |
| type_of_meal_plan_Meal Plan 3 | 1.7182 | 2.912 | 0.590 | 0.555 | -3.989 | 7.425 |
| type_of_meal_plan_Not Selected | 0.8463 | 0.048 | 17.563 | 0.000 | 0.752 | 0.941 |

## Observation:-

- Through the above table we get the results of coefficient, std err, and high or low p-values. Further will check the multicollinearity and vif which would be more clear the reason of cancellations.

# Model Building - Logistic Regression

- In order to make statistical inferences from a logistic regression model, it is important to ensure that there is no multicollinearity present in the data.

| | |
|---|---|
| const | 326.141919 |
| no_of_adults | 1.346659 |
| no_of_week_nights | 100.277464 |
| required_car_parking_space | 1.041578 |
| arrival_month | 1.051511 |
| repeated_guest | 3.340040 |
| avg_price_per_room | 1.936037 |
| length_stay | 146.442538 |
| no_of_children_log | 1.666322 |
| no_of_weekend_nights_log | 34.428764 |
| no_of_previous_cancellations_log | 1.597137 |
| no_of_previous_bookings_not_canceled_log | 3.508907 |
| no_of_special_requests_log | 1.267959 |
| type_of_meal_plan_Meal Plan 2 | 1.217525 |
| type_of_meal_plan_Meal Plan 3 | 1.025316 |
| type_of_meal_plan_Not Selected | 1.236534 |
| room_type_reserved_Room_Type 2 | 1.098668 |
| room_type_reserved_Room_Type 3 | 1.003381 |
| room_type_reserved_Room_Type 4 | 1.364652 |
| room_type_reserved_Room_Type 5 | 1.020015 |
| room_type_reserved_Room_Type 6 | 1.058575 |
| room_type_reserved_Room_Type 7 | 1.111082 |
| market_segment_type_Complementary | 4.507245 |
| market_segment_type_Corporate | 16.930019 |
| market_segment_type_Offline | 64.016661 |
| market_segment_type_Online | 71.248287 |
| lead_time_y_short | 1.119269 |
| lead_time_y_med | 1.105320 |
| lead_time_y_long | 1.151828 |
| lead_time_y_advanced | 1.047117 |

| | const | no_of_adults | required_car_parking_space | arrival_month | repeated_guest | avg_price_per_room | length_stay | no_of_children_log | no_of_previous_cancellations_log | no_of_ |
|---|---|---|---|---|---|---|---|---|---|---|
| odds | 0.024636 | 1.271006 | 0.234679 | 0.935032 | 0.051575 | 1.023344 | 1.116331 | 1.749625 | 2.60091 | |

| | const | no_of_adults | required_car_parking_space | arrival_month | repeated_guest | avg_price_per_room | length_stay | no_of_children_log | no_of_previous_cancellations_l |
|---|---|---|---|---|---|---|---|---|---|
| change_odds% | -97.53636 | 27.100615 | -76.542091 | -6.496819 | -94.842473 | 2.33444 | 11.6331 | 74.962538 | 160.0910 |

**Observation:-**

- Here we can observe the VIF values where we can see that length of stay has the higher vif values and also const which is need to be check further.
- We can even see that there are the results obtained for converting to coefficient to odds and above we can see the results even in the percentage form.

# Model Building - Decision Tree

- The decision tree model can be made from 3 of the steps on which we have work.
  We want to predict which bookings will be canceled.
- 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.
- First, let's create functions to calculate different metrics and confusion matrix so that we don't have to use the same code repeatedly for each model.
- The model_performance_classification_sklearn function will be used to check the model performance of models.
- The confusion_matrix_sklearnfunction will be used to plot the confusion matrix.

# Model Building - Decision Tree

```python
# defining a function to compute different metrics to check performance of a classification model built using sklearn
def model_performance_classification_sklearn(model, predictors, target):
    """
    Function to compute different metrics to check classification model performance

    model: classifier
    predictors: independent variables
    target: dependent variable
    """

    # predicting using the independent variables
    pred = model.predict(predictors)

    acc = accuracy_score(target, pred)  # to compute Accuracy
    recall = recall_score(target, pred)  # to compute Recall
    precision = precision_score(target, pred)  # to compute Precision
    f1 = f1_score(target, pred)  # to compute F1-score

    # creating a dataframe of metrics
    df_perf = pd.DataFrame(
        {"Accuracy": acc, "Recall": recall, "Precision": precision, "F1": f1,},
        index=[0],
    )

    return df_perf
```

```python
def confusion_matrix_sklearn(model, predictors, target):
    """
    To plot the confusion_matrix with percentages

    model: classifier
    predictors: independent variables
    target: dependent variable
    """
    y_pred = model.predict(predictors)
    cm = confusion_matrix(target, y_pred)
    labels = np.asarray(
        [
            ["{0:0.0f}".format(item) + "\n{0:.2%}".format(item / cm.flatten().sum())]
            for item in cm.flatten()
        ]
    ).reshape(2, 2)

    plt.figure(figsize=(6, 4))
    sns.heatmap(cm, annot=labels, fmt="")
    plt.ylabel("True label")
    plt.xlabel("Predicted label")
```

- Through this codes we can build the model of decision tree.

# Model Building - Decision Tree

- Accuracy on training set : 0.9924385633270322
- Accuracy on test set : 0.8585867867315997

**Observation:-**

- Here we can observe that tree scores are at a perfect accuracy level and has most of the data,.
- It has almost 11885 predictions of cancellations and actual was about 11989 so that can not be a good model.
- To avoid the cancellations we need to calculate different metrics and confusion matrix so that we don't have to use the same model repedeatly for each model.

# Model Building - Decision Tree

- **Checking model performance on training and testing set through confusion matrix:-**

Training Performance:

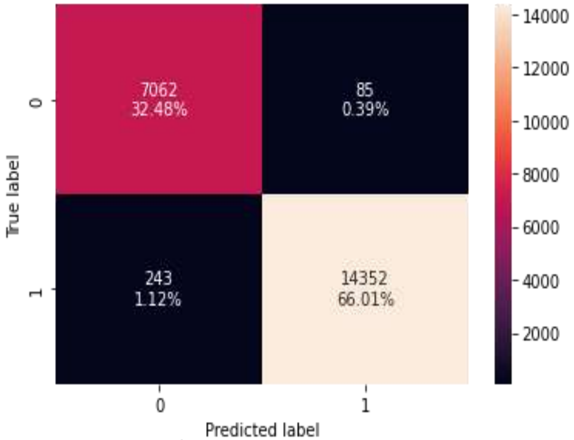
| | |
|---|---|
| **Accuracy** | **0.992438 5** |
| Precision | 0.994112 |
| Recall | 0.98335 |
| F1 | 0.988702 |

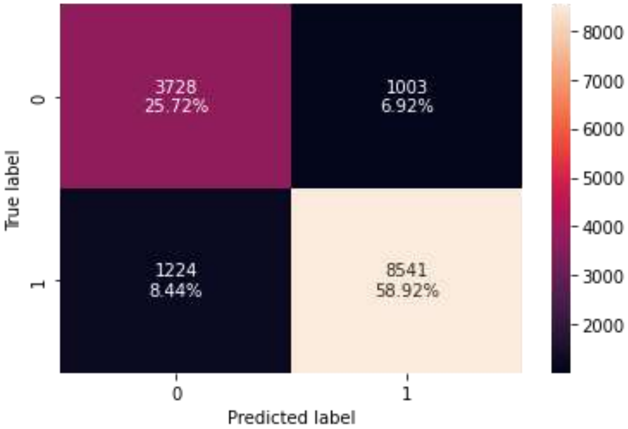
Testing Performance:-

| | |
|---|---|
| **Accuracy** | **0.858586** |
| Precision | 0.894908 |
| Recall | 0.874654 |
| F1 | 0.884665 |

# Model Building - Decision Tree

- **Confusion Matrix for training and test performances**:-



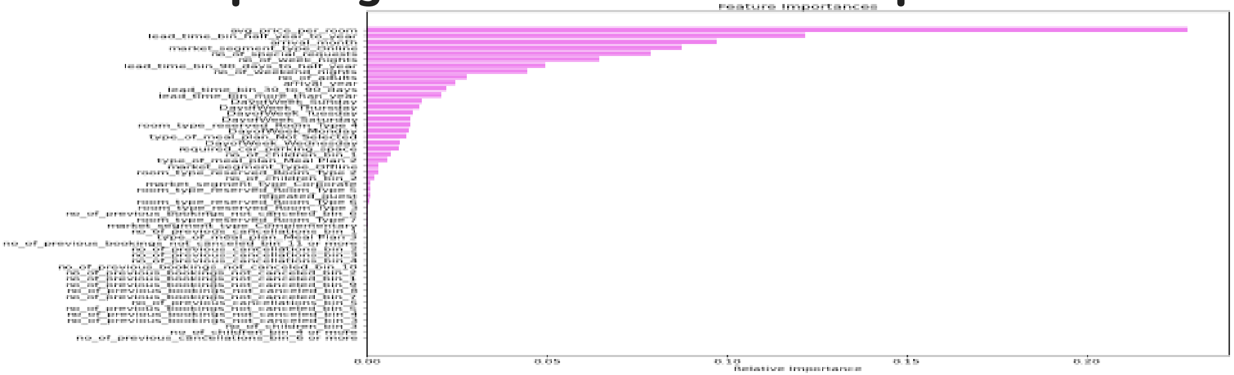




**Observations:-**

- The training data's f1 is very close to 1 and the testing data is near to 0.90.While we can say that the model captures the more amount of testing data so from this matrix we can assume that without tuning this model is over fitting.

# Model Building - Decision Tree

- **Before pruning the tree let's check the important features.**




**Observation:-**

The average price per room , the time between 6 to 12 months, the online booking, and the number of special request have the high number of ways that determines that the guest will cancel.

# Model Building - Decision Tree

- **Pre-Prunning**

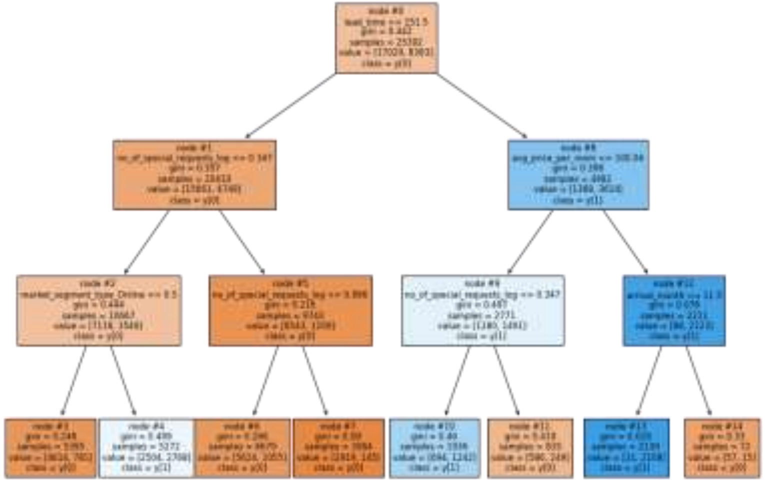
Accuracy on training set : 0.7844202898550725
Accuracy on test set : 0.7913259211614444
Recall on training set : 0.7315556618438359
Recall on test set : 0.7385008517887564

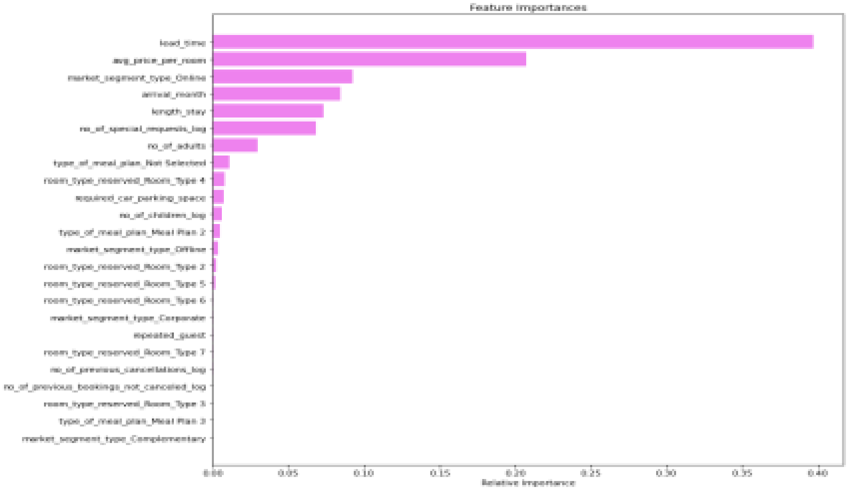
**Observations:-**

- Due to over fitting  the decision tree branches are very complex so we have to limit depth and post pruning.
- By looking at the accuracy rate compare to previous one we have eliminated the closeness of training and testing accuracy from the model.
- Having accuracy up to 79% is a good improvement.
- Also we can see that recall matrix is also much better than the previous model.

# Model Building - Decision Tree

- **Visualizing the Decision Tree:-**





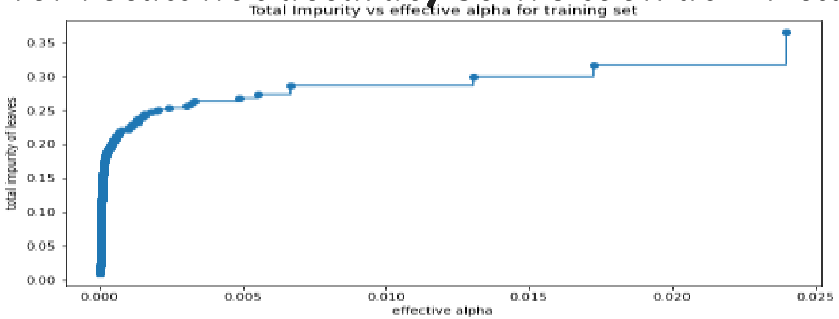
# Model Building - Decision Tree

- **Cost Complexity Pruning:-**

**In this pruning we are still looking for recall not accuracy so we look at DT classifier:-**

| | ccp_alphas | impurities |
|---|---|---|
| 0 | 0.000000e+00 | 0.009478 |
| 1 | 0.000000e+00 | 0.009478 |
| 2 | 0.000000e+00 | 0.009478 |
| 3 | 4.600391e-07 | 0.009478 |
| 4 | 5.329960e-07 | 0.009479 |
| ... | ... | ... |
| 1508 | 6.665684e-03 | 0.286897 |
| 1509 | 1.304180e-02 | 0.299942 |
| 1510 | 1.725993e-02 | 0.317202 |
| 1511 | 2.399048e-02 | 0.365183 |
| 1512 | 7.657789e-02 | 0.441761 |

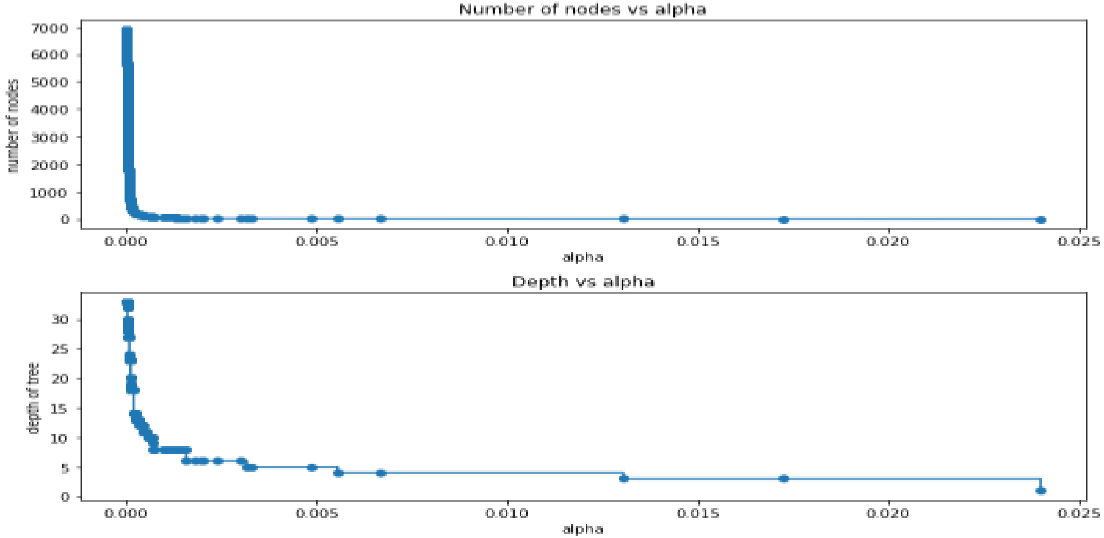
1513 rows × 2 columns




Number of nodes in the last tree is: 1 with ccp_alpha: 0.0765778947737134

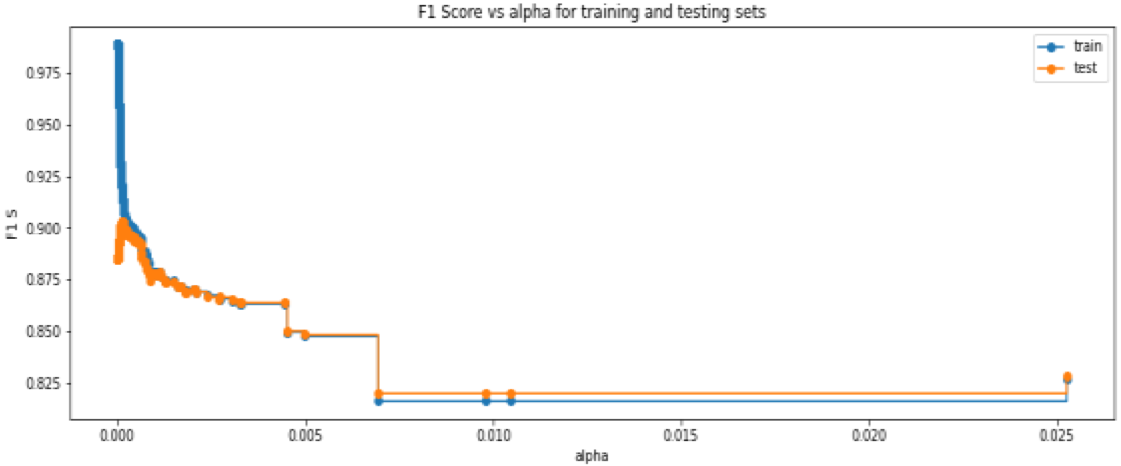
# Model Building - Decision Tree

- DT Classifier for every alpha:-



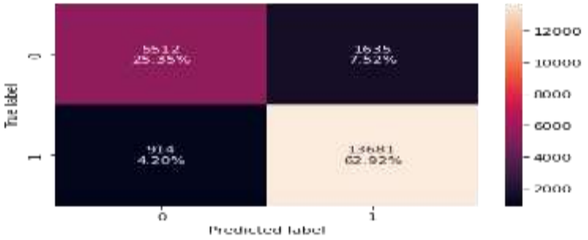
# Model Building - Decision Tree

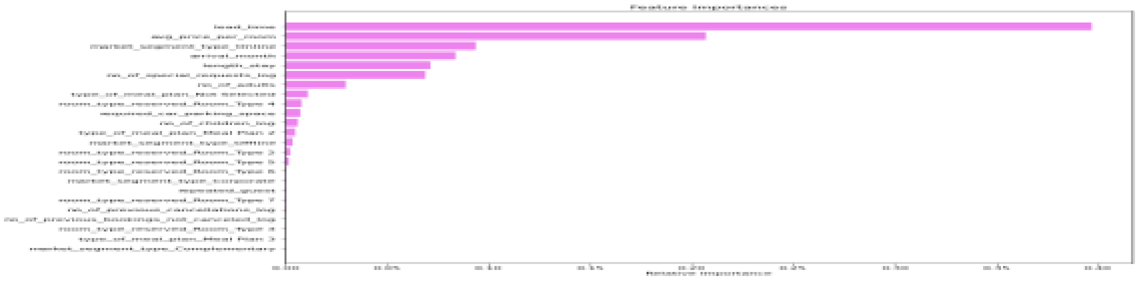
- F1 Score vs alpha for training and testing sets:



# Model Building - Decision Tree

- Checking performance on training and test set through a confusion matrix:-



**Observations:-**

- Like the original model the same features are important but a notice change to a order. 6 months to a year lead time has overtaken the average price of room while online bookings move up.
- It is assumed that a lead time play a vital role in a person will cancel their booking.

# Model Performance Evaluation and Improvement - Decision Tree

## Comparing Decision Tree models:-

| | Model | Train_Recall | Test_Recall |
|---|---|---|---|
| 0 | Initial decision tree model | 0.981 | 0.792 |
| 1 | Decision tree with restricted maximum depth | 0.732 | 0.739 |
| 2 | Decision treee with hyperparameter tuning | 0.732 | 0.739 |
| 3 | Decision tree with post-pruning | 0.979 | 0.794 |

## Observation:-

- The post-pruning model known as the best model shows the highest f1 score for testing data.
- The tress with maximum tuning and it performed while reducing over fitting. I would recommend this model to use for future predictions.

# Happy Learning !