Star Hotels

- Holidays begin here

Data based prediction model : Predict if a booking will be cancelled ?

Background

1. 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.

2. Cancellations are less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

3. The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior.

Business Problem Overview and Solution Approach

- Financial implications: Loss of resources (revenue) when the hotel cannot resell the room.
  The cancellation of bookings impact a hotel on various fronts:

  - Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
  - Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
  - Human resources to make arrangements for the guests.

- **Objective**

- The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled.
- Find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

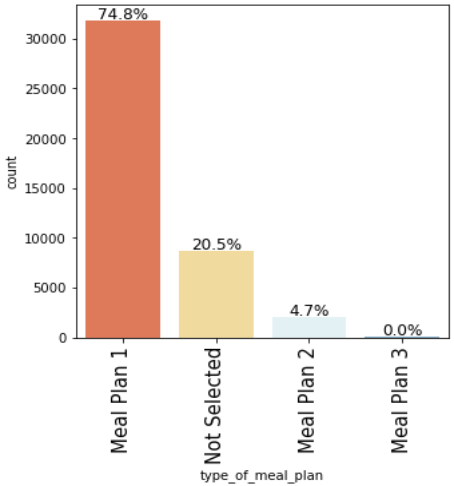
Data Overview

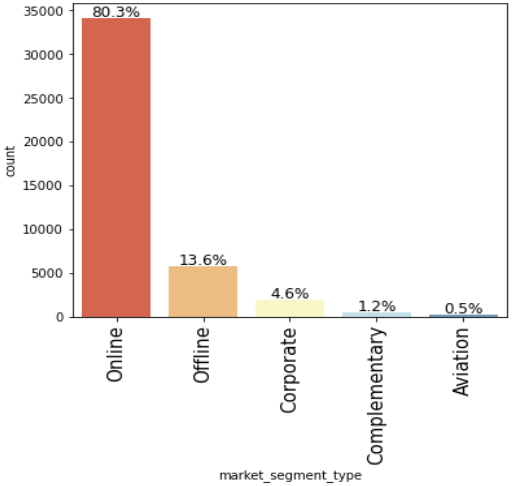
Variable	Description
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked
type_of_meal_plan	Type of meal plan booked by the customer:
required_car_parking_space	Does the customer require a car parking space
room_type_reserved	Type of room reserved by the customer
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?

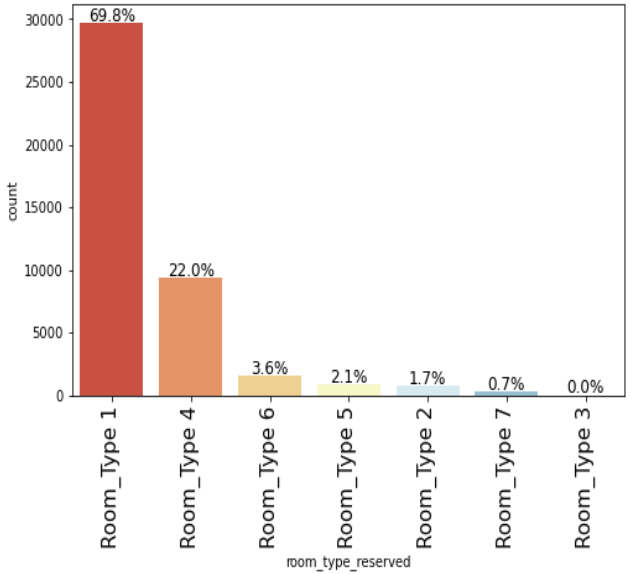
Variable	Description
no_of_previous_cancellations	Previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled	Previous bookings not canceled by the customer prior to the current booking
avg_price_per_room	Average price per day of the reservation
no_of_special_requests	Number of special requests
booking_status	Flag indicating if the booking was canceled or not.

Observations	Variables
56926	18

Exploratory Data Analysis





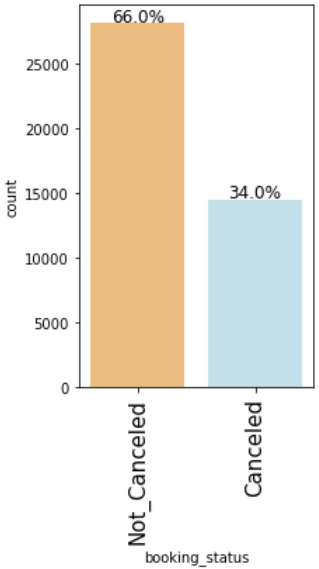


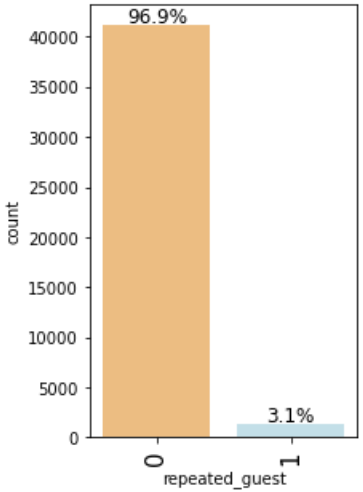
- Meal Plan 1 (Breakfast) accounts for almost 75% of the bookings

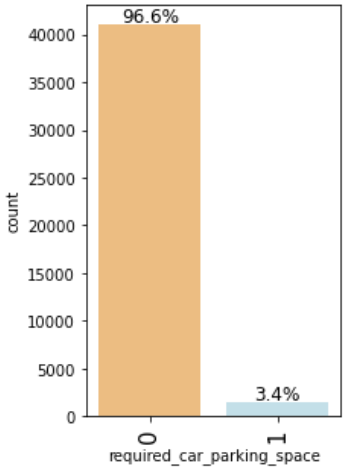
- Online market segment is the largest. Offline is the second most booked segment

- Room Type 1 , accounts for almost 70% of the bookings
- Room Type 1 and Room Type 4 together make up over 90% of the bookings

EDA

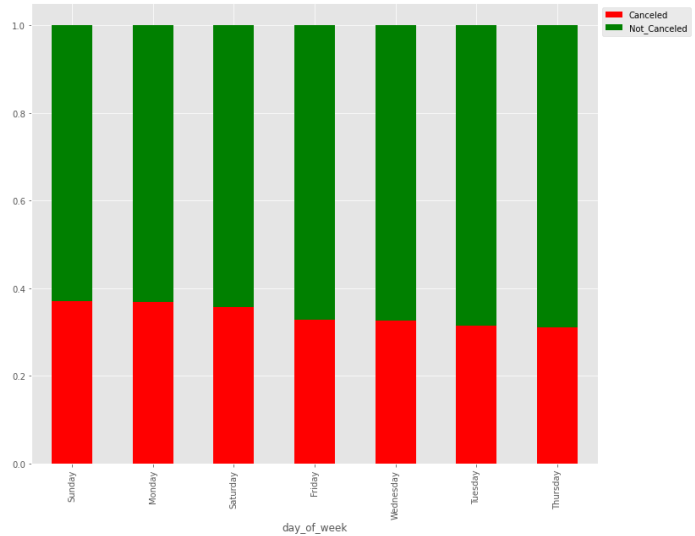




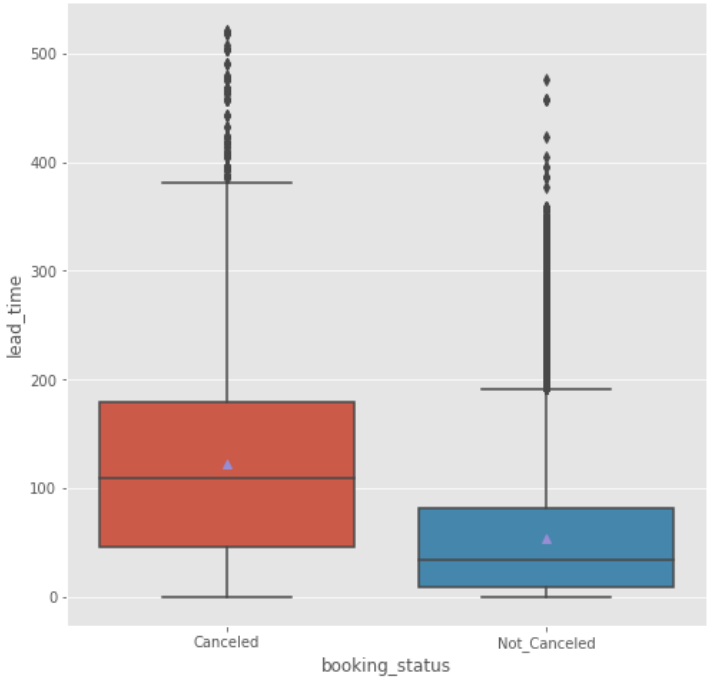


- 66% of the bookings are not canceled
- About 3% of the booking are from repeated guests
- Most bookings did not need a car parking space

EDA

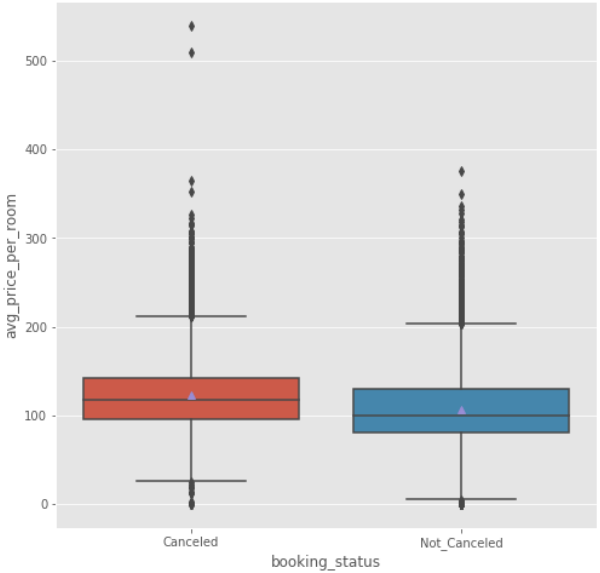


Almost all days of the week have similar possibility of cancellation

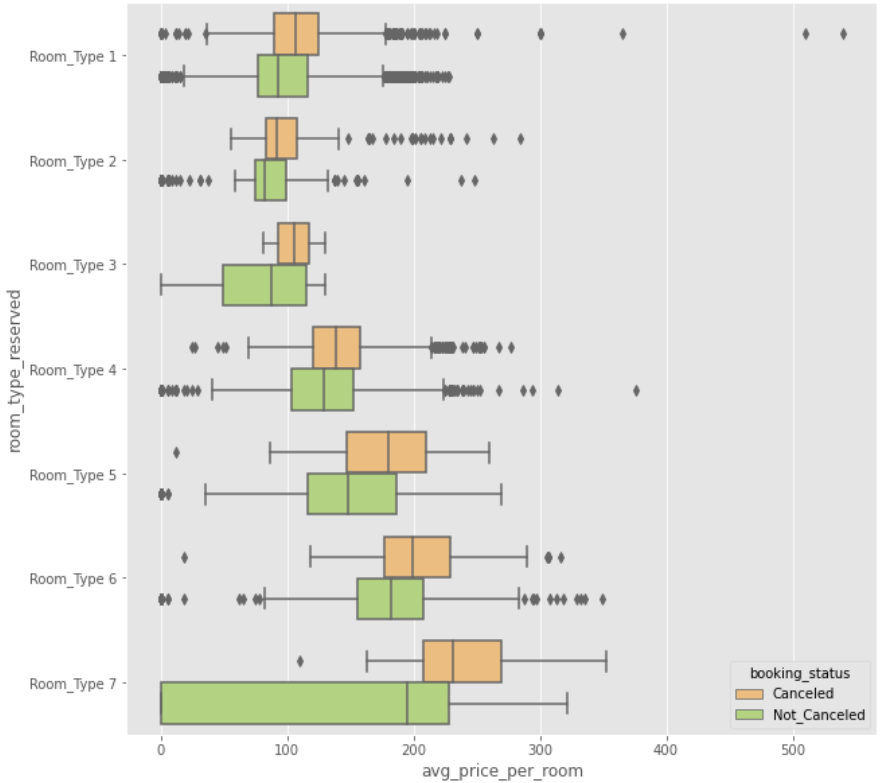


Higher lead_time bookings have risk of cancellation

EDA

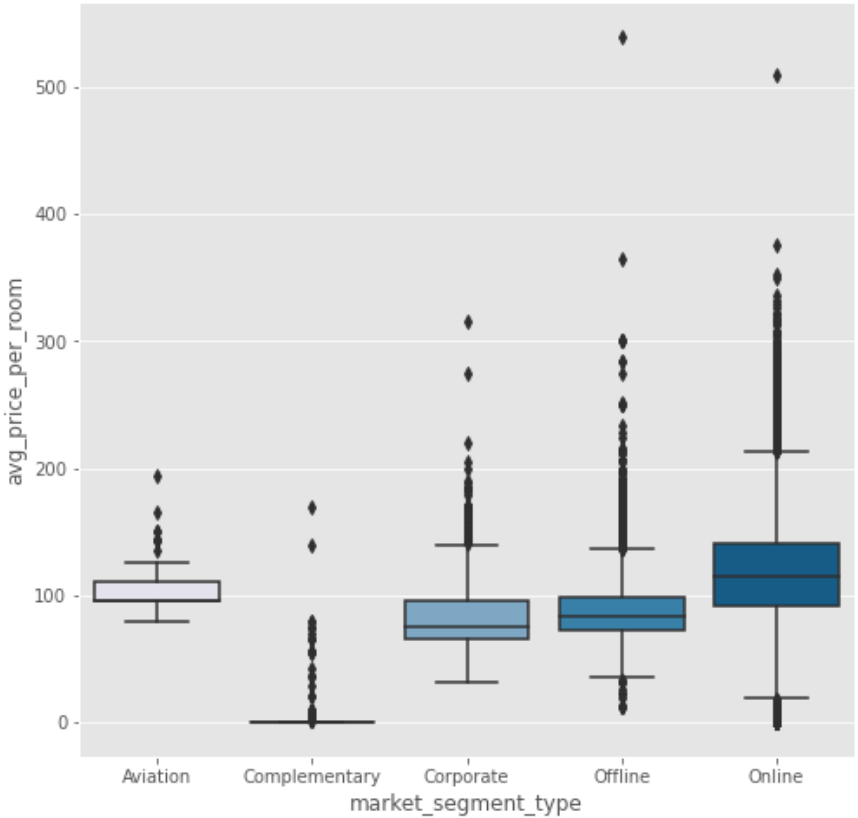


• No particualr distinction on avg_booking_price wrt to booking_status



• room_type 7, with avg_price_per_room greater than 200 Euros is more likely to be canceled.

EDA – Contd.

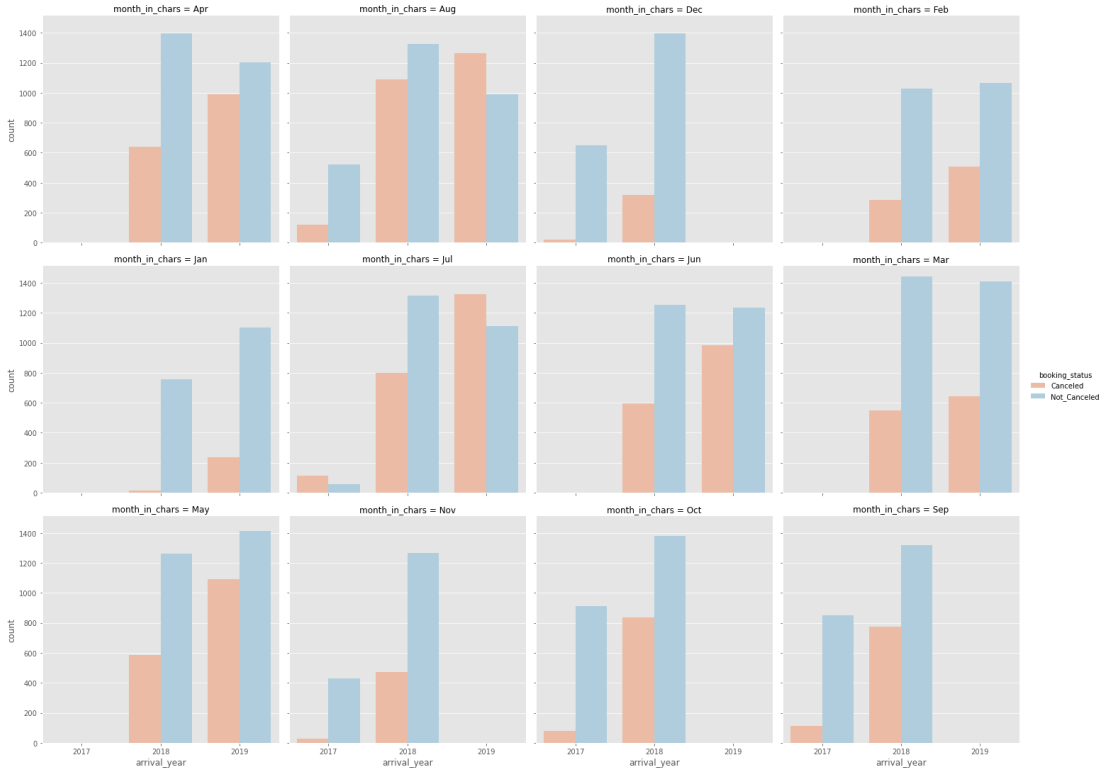


•Aviation segment: Although the price range is less, the avg price is higher then most segments
•Corporate and Offline segments have almost similar average price ranges
•Online has the largest range on prices

Data Overview

•Observations
•As the no of week nights increases, the cancellation probability increases.
•The more the number of weekend nights , more likely to be cancelled.
•Repeated_guest are least likely to cancel.
•Higher lead_time bookings have risk of cancellation.

EDA – Contd.



- Data is available for Jul 2017 to Aug 2019
- So it will not be relevant to include arrival_year, arrival_month and arrival_date in the model

```
print (data['full_date'].min())
print (data['full_date'].max())
```

```
2017-07-01 00:00:00
2019-08-31 00:00:00
```

Key Questions:

1. What are the busiest months in the hotel?

- March ,it has the highest non cancellations. Aug, May and Apr as well.

2. Which market segment do most of the guests come from?

- Online

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

- Aviation : Mostly between 90 and 110.
- Corporate : 70 to 100
- Offline : 80 to 100, with many outliers and max reaching above 500
- Online : 90 to 130, with many outliers and max reaching above 500

4. What percentage of bookings are canceled?

- At about 34% of booking are canceled.

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

- Less than 1% of repeated_guest cancel the booking

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

- The more special requirements, the less likely they are to cancel

Model evaluation criterion

- **Model can make wrong predictions as:**

  1. Predicting a booking is *not* going to be Canceled but in reality is Canceled - Loss of resources
  2. Predicting a booking is going to Cancel but in reality will *not* Canceled -Wrong flagging

- **Which Loss is greater ?**

  - The cancellation of bookings impact a hotel on various fronts.
  - Loss of resources will be the greater loss as the hotel will lose out on revenue and other additional expenditure the cancel of bookings will cause.

- **How to reduce this loss i.e need to reduce False Negatives ?**

  - Model should reduce false negatives, this can be done by maximizing the Recall. Greater the recall lesser the chances of false negatives.

- **Positive event and Negative Event**

  - Positive event is cancellation
  - Negative event is Non cancellation

Logistic Regression

Training performance comparison:

	Logistic Regression sklearn	Logistic Regression-0.28 Threshold	Logistic Regression-0.45 Threshold
Accuracy	0.741536	0.680703	0.731235
Recall	0.553807	0.826849	0.617266
Precision	0.636405	0.518146	0.600732
F1	0.592240	0.637071	0.608887

Model performance evaluation

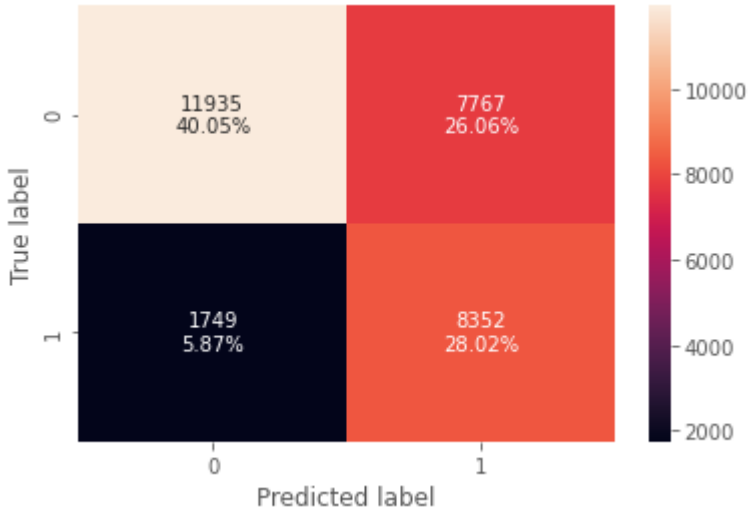
•All the models are giving a generalized performance on training and test set.
•The highest recall is 82% on the training set.
•Using the model with 0.28 threshold the model will give a high recall. This model will help the hotel identify potential bookings that will be cancelled

Final Model Summary

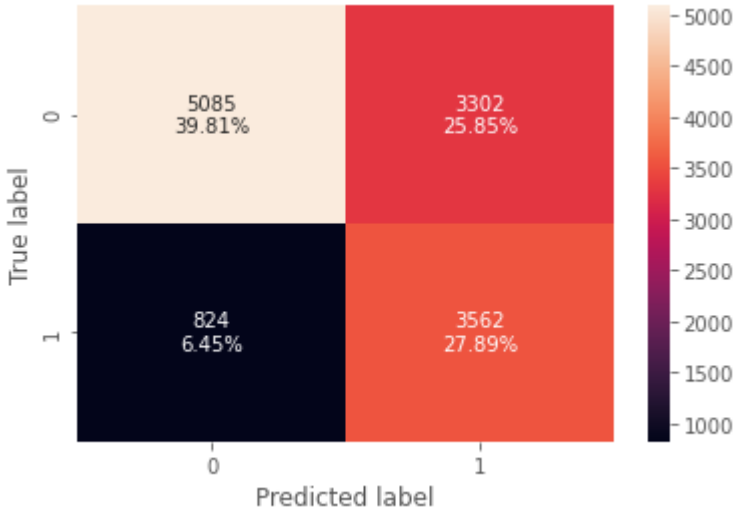
•From our logistic regression model we identified that lead_time is a significant predictor of a booking being canceled.
•Bookings that have no children are more likely to be cancelled
•Online bookings are more likely to be cancelled
•repeated_guest are least likely to cancel
•The more special requirements, the less likely they are to cancel

Logistic Regression

Training :optimal_threshold_auc_roc

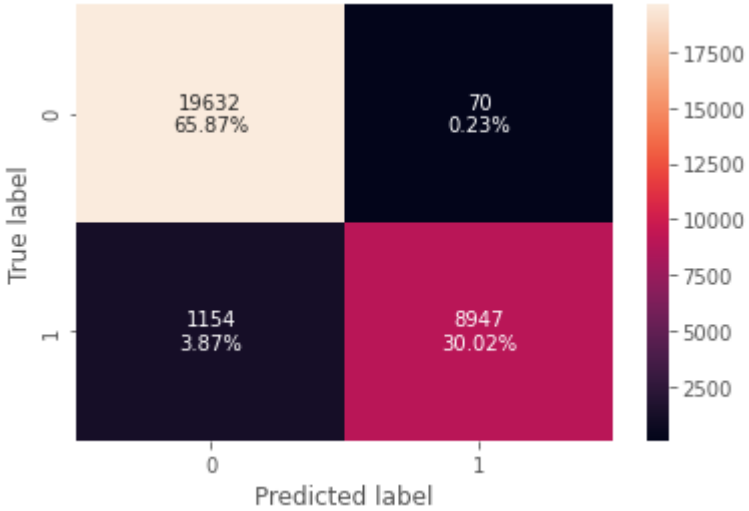


Test :optimal_threshold_auc_roc

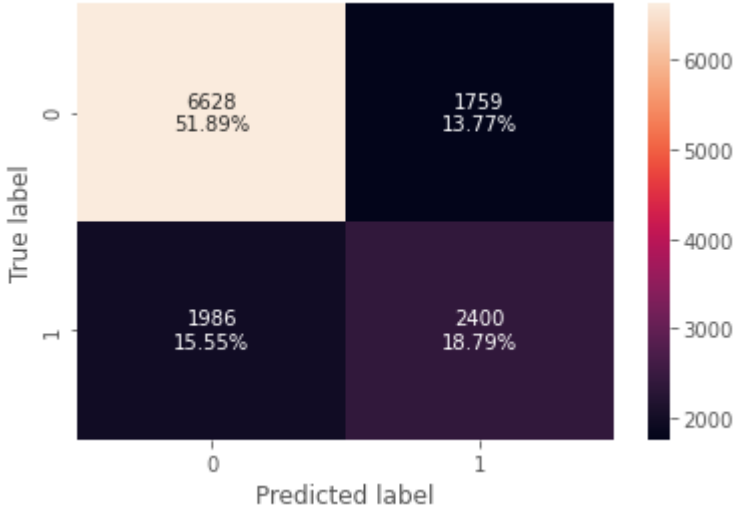


Decision Tree model

Training :
Recall Score: 0.8857538857538858



Test:
Recall Score: 0.5471956224350205



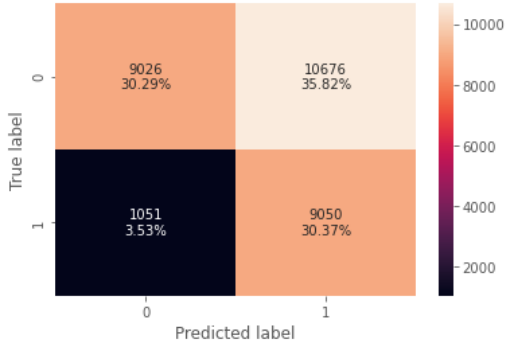
There is a huge disparity in performance of model on training set and test set, which suggests that the model is over fiiting.

Decision Tree model - Reducing over fitting with Grid Search

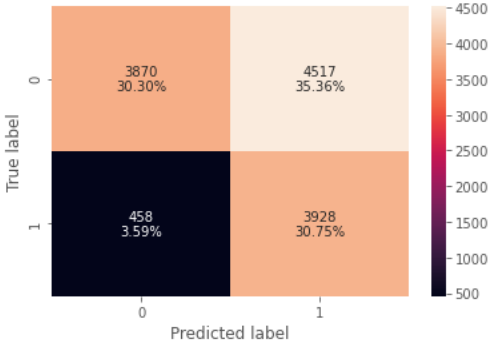
Pre-Pruning

```
DecisionTreeClassifier(random_state=1,class_weight=
{0: 0.34, 1: 0.66})


# Grid of parameters to choose from

parameters = {'max_depth': np.arange(1,10),
              'min_samples_leaf': [1, 2, 5, 7, 10,15,20],
              'max_leaf_nodes' : [2, 3, 5, 10],
              'min_impurity_decrease': [0.001,0.01,0.1]
             }
```

Training



Test



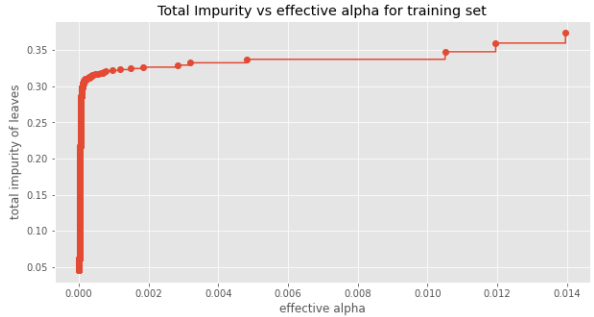
Training :
Recall Score: 0.8959508959508959

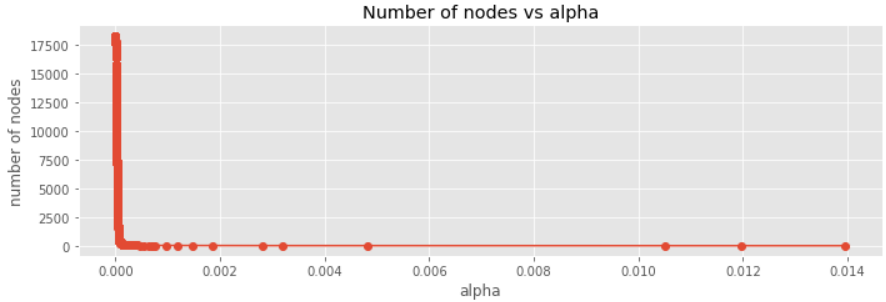
Test :
Recall Score: 0.895576835385317

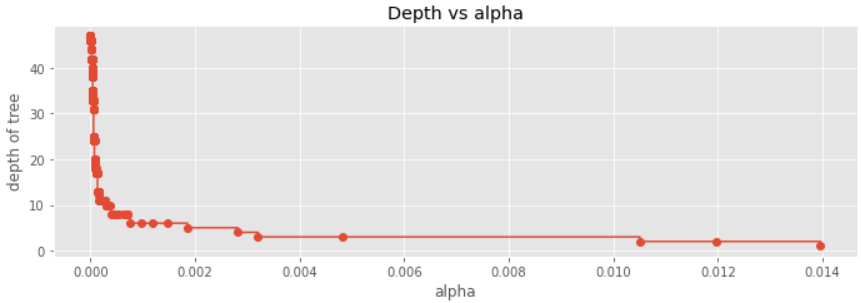
- The model is giving a generalized result now.

Decision Tree model - Cost Complexity Pruning (post-Pruning)

Next, we trained 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.



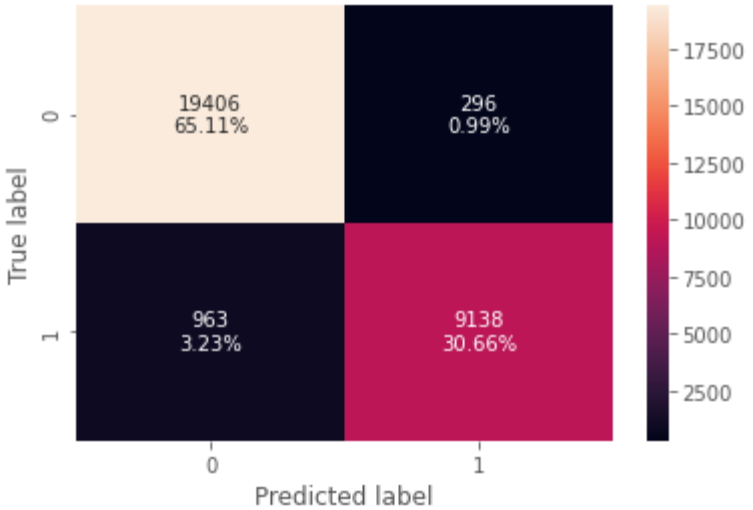




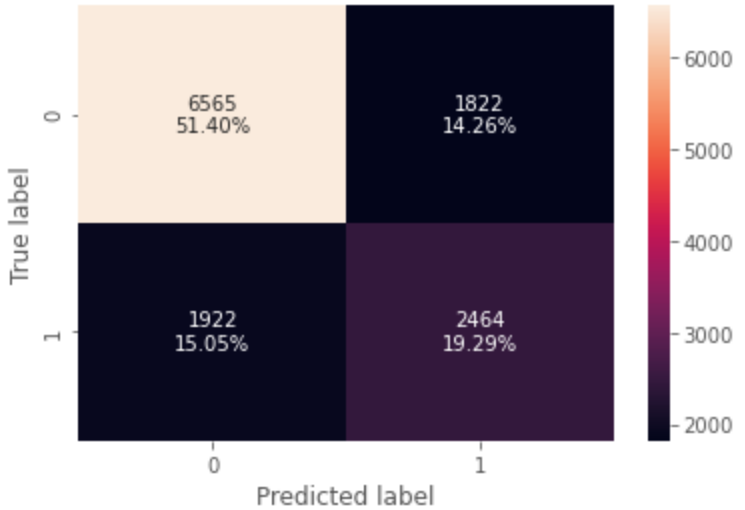
Number of nodes in the last tree is: 1 with ccp_alpha: 0.07447795787334116
For the remainder, we remove the last element in clfs and ccp_alphas, because it is the trivial tree with only one node.

Decision Tree model – CCP – post-pruning

Training :
Recall Score: 0.8857538857538858



Test:
Recall Score: 0.5617875056999544



There is a huge disparity in performance of model on training set and test set. The pre-pruned model gives a better performance.

Decision Tree – Model Performance Comparison and Conclusions

Training performance comparison:

	Recall on training set
0	0.885754
1	0.895951
2	0.904663
3	0.885754

Test performance comparison:

	Recall on testing set
0	0.547196
1	0.895577
2	0.561788
3	0.547196

0. Default Decision Tree
1. Pre-pruned (with hyper parameters set)
2. Post-pruned(CCP with best fit)
3. Post-pruned (CCP with **0.000000001 ccp_alpha)**

Decision Tree – Conclusion

•Decision tree model with pre-pruning has given the best recall score on training data and test data.

•The pre-pruned and the post-pruned models have reduced overfitting and the model is giving a generalized performance.

•The Decision tree with these parametes gives the best model

DecisionTreeClassifier(class_weight={0: 0.34, 1: 0.66}, max_depth=3, max_leaf_nodes=5, min_impurity_decrease=0.001, random_state=1)

Conclusions

•With the logistic regression model the best performance of recall metric was 0.826849 on training and 0.81213 on test.

•The logistic regression model had a optimal_threshold_auc_roc of 0.28767.

•With Decision tree and pre-pruning the best performance was 0.89595 on training and 0.89557 on the test.

•The Decision tree model had the pre-pruning parameters of class_weight={0: 0.34, 1: 0.66}, max_depth=3, max_leaf_nodes=5, min_impurity_decrease=0.001, random_state=1)

•As with Decision tree model the important variable for determining the booking_status is lead_time and total_no_of_days the booking is done for.

•For the logistic regression model repeated_guest and avg_price_per_room form the important variables.

Business Insights and Recommendations

- **What profitable policies for cancellations and refunds can the hotel adopt?**

- Higher lead_time bookings have risk of cancellation, a maximum of 90 days thershold can be set.
- Higher cancellation charges can be applied if the customer booked with a lead time of more thatn 60 days.
- When the booking is done for longer no of nights, the possibility of canellation is high.
- Repeat guest are less likely to cancel. Hotel can consider to offer loyalty programs.
- Hotel can apply additional surcharge when the booking exceeds total_no_of_day as 7.
- Bookings with Sunday as the day of arrival, have a slightly higher changes of being cancelled, so hotel can charge a premium for sundays as arrival_day
- Booking that are for only weekdays have higher change of being cancelled. If the booking is only for weekday cancellations policy can be adjusted to make it less attractive for cancellation.

Business Insights and Recommendations. Contd..

- **What profitable policies for cancellations and refunds can the hotel adopt?**

- Booking with children (young family) are likely to be cancelled more often.
- Online booking form the bulk of booking and also have the higher cancellation.
- Bookings that have Meal Plan 3 as their choice are less likey to cancel.
- Booking that have special_requests are less likely to cancel.
- Bookings with room_type_reserved as 1 and 3 are less likey to cancel.

Business Insights and Recommendations

- **What other recommendations would you suggest to the hotel?**

- Repeat guest are less likely to cancel. Hotel can consider to offer loyalty programs.
- Hotel can adopt dynamic pricing as the date draws closer as customer booking close to the arrival date are less likely to cancel.
- Hotel can offer special rewards to coporate and aviation as the cancellations are lesser.
- Bookings with room_type_reserved as 1 and 3 are less likely to cancel. Hotel can charge additional cancellation fee for other room types
- For online segment, hotel can enforce a holding amount which can be refunded based on the cancellation date and arrival date.

greatlearning
Power Ahead

Happy Learning !