STAR HOTELS PROJECT

BOOKING CANCELLATION PREDICTION

BY: SYEDA AMBREEN KARIM BOKHARI

CONTENTS

CORE BUSINESS IDEA:

Star Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations due to various reasons and have reached out to our firm for data-driven solutions.

BUSINESS PROBLEM OVERVIEW AND SOLUTION APPROACH

- **Problem to tackle**
  - A significant number of hotel bookings are called-off due to cancellations or no-shows.
  - Free of charge cancellations or cancellations at a low cost 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 adds to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.
  - Machine Learning based solution is needed that can help in predicting which booking is likely to be cancelled..

BUSINESS PROBLEM OVERVIEW AND SOLUTION APPROACH

- **Financial implications**

  The cancellation of bookings impact a hotel on various fronts:

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

BUSINESS PROBLEM OVERVIEW AND SOLUTION APPROACH

- **How to use ML model to solve the problem**

  We need to analyse the provided data and build a Classification model to predict booking status.

  The logistic regression model is trained and tested on the available data and can be used to predict with ~95% accuracy, the future price of a used phone and identify factors that significantly influence it.

Data Overview: The dataset file contains the used phone data with following specifications:

Variable name	Data types	Description	Unique values
1. no_of_adults:	numeric	Number of adults	5
2. no_of_children:	numeric	Number of Children	6
3. no_of_weekend_nights:	numeric	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	9
4. no_of_week_nights:	numeric	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	18
5. type_of_meal_plan:	categorical	Type of meal plan booked by the customer:	4
6. required_car_parking_space:	numeric	Does the customer require a car parking space?	2
7. room_type_reserved:	categorical	Type of room reserved by the customer.	7
8. lead_time:	numeric	Number of days between the date of booking and the arrival date	397
9. market_segment_type:	categorical	Market segment designation.	5
10. repeated_guest:	numeric	Is the customer a repeated guest?	2

Data Overview: The dataset file contains the used phone data with following specifications:

Variable name	Data types	Description	Unique values
11. arrival_year:	numeric	Year of arrival date	3
12. arrival_month:	numeric	Month of arrival date	12
13. arrival_date:	numeric	Date of the month	31
14. no_of_previous_cancellations:	numeric	Number of previous bookings that were cancelled by the customer prior to the current booking	9
15. no_of_previous_bookings_not_canceled:	numeric	Number of previous bookings not cancelled by the customer prior to the current booking	73
16. avg_price_per_room:	numeric	Average price per day of the reservation (in Euros)	4939
17. no_of_special_requests:	numeric	Total number of special requests made by the customer	6
18. booking_status:	categorical	Flag indicating if the booking was cancelled or not.	2

DATA OVERVIEW: FEATURE ENGINEERING

- Brief description of significant manipulations made to raw data

Observations	Variables	Missing	Duplicates	Dependent variable
56926	18	0	14350 Duplicate records were dropped	booking_status:

Variable name	Data description	Treatment
booking_status	It is the target variable. It was categorical so needed to be converted to number.	It was converted to boolean integer: 1:Canceled and 0: Not_Canceled
arrival_date, arrival_month, arrival_year	These three variables were combined to arrival_date_full to check dates and reduce number of columns. There were 35 records where date was wrongly entered : 29/02/2018 as 2018 was not a leap year.	These records with wrong dates were dropped. Dates were used to get weekdays. Dates were then converted to yearly quarters for 2017 to 2019 so it can be hot encoded for the model.

DATA OVERVIEW

- Brief description of significant manipulations made to raw data

Observations	Variables	Missing	Dependent variable
42576 After dropping duplicates	18	0	booking_status:

Categorical Variables: Encoding into numeric.	
Variable name	**Treatment**
type_of_meal_plan room_type_reserved market_segment_type	One Hot Encoding was used as there were fewer unique values.

DATA OVERVIEW: OUTLIERS TREATMENT & FEATURE ENGINEERING

- Brief description of significant manipulations made to raw data

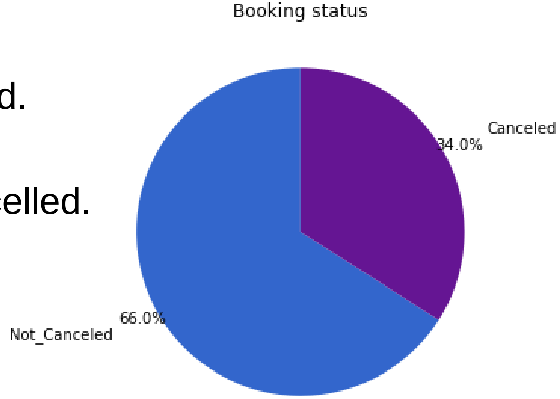
Variable name	Outliers detection and treatment
• lead_time has right skewed distribution. • avg_price_per_room has right skewed distribution. • no_of_special_requests has right skewed distribution. • no_of_week_nights has right skewed distribution • no_of_previous_cancellations:has right skewed distribution • no_of_previous_bookings_not_canceled: has right skewed distribution	• IQR was used to detect outliers in all the numeric fields. • Outliers in the data were treated by flooring and capping.

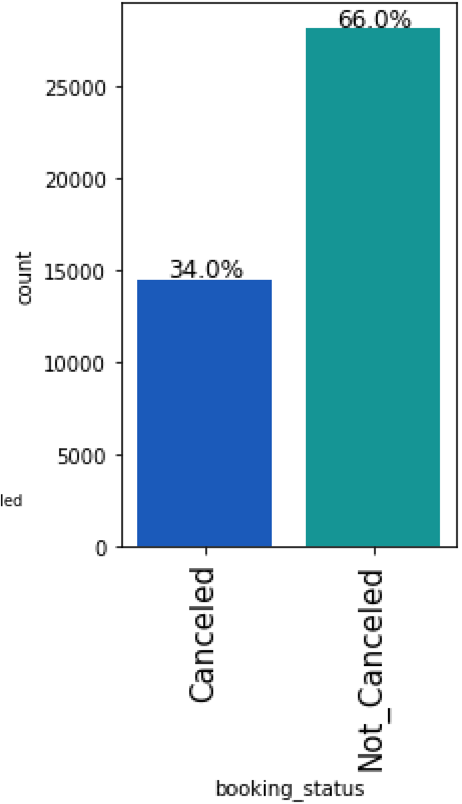
EXPLORATORY DATA ANALYSIS

- Graphs and observation about the target attribute:

Observations

- booking_status is the target variable.
- It is a categorical variable with two values: Cancelled; 1 and Not_cancelled: 0
- 34% of the data has booking status cancelled.
- 66% of the data does not have booking cancelled.





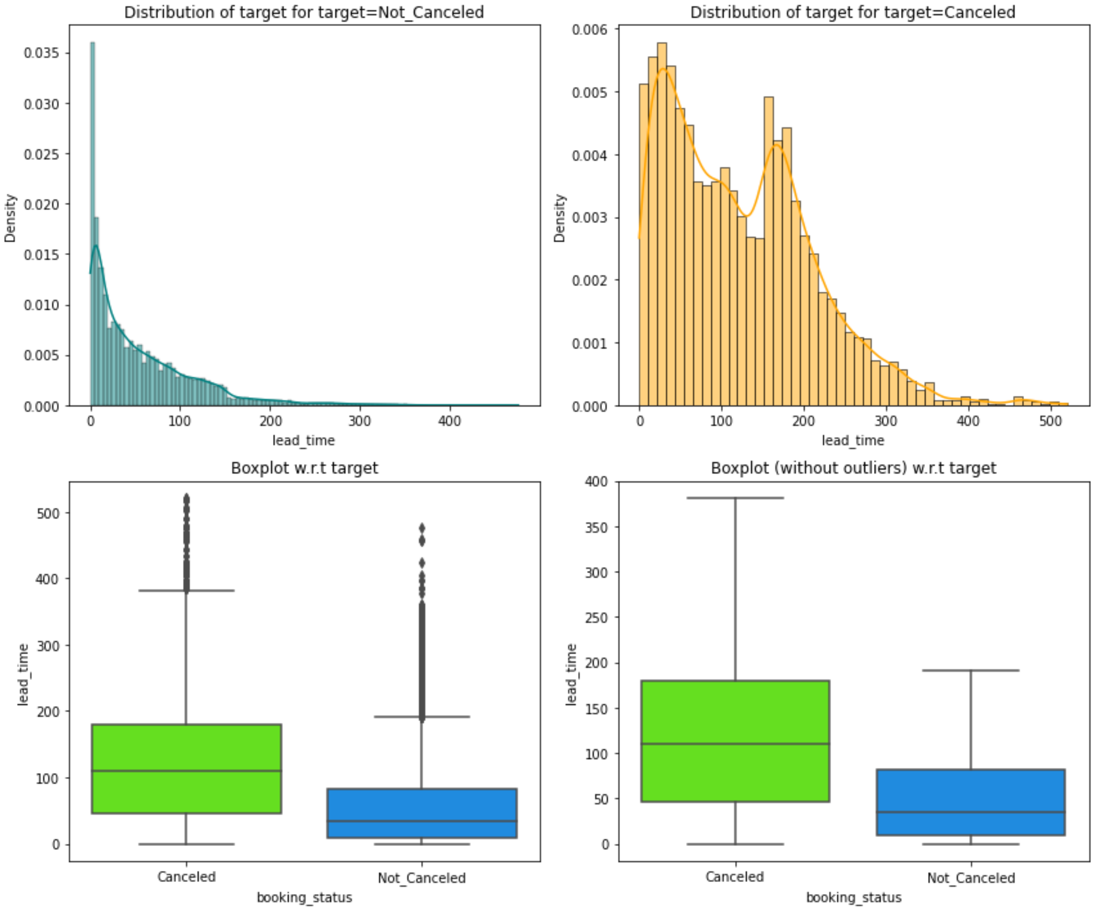
EDA

- Graphs showing the factors most heavily impacting the target attribute

Observations:

- 50% of cancelled bookings have Lead-time between 50 - ~180 days.

- 50% of not cancelled bookings are within 100 days lead-time. So there seems some association between lead-time and booking status.



EDA

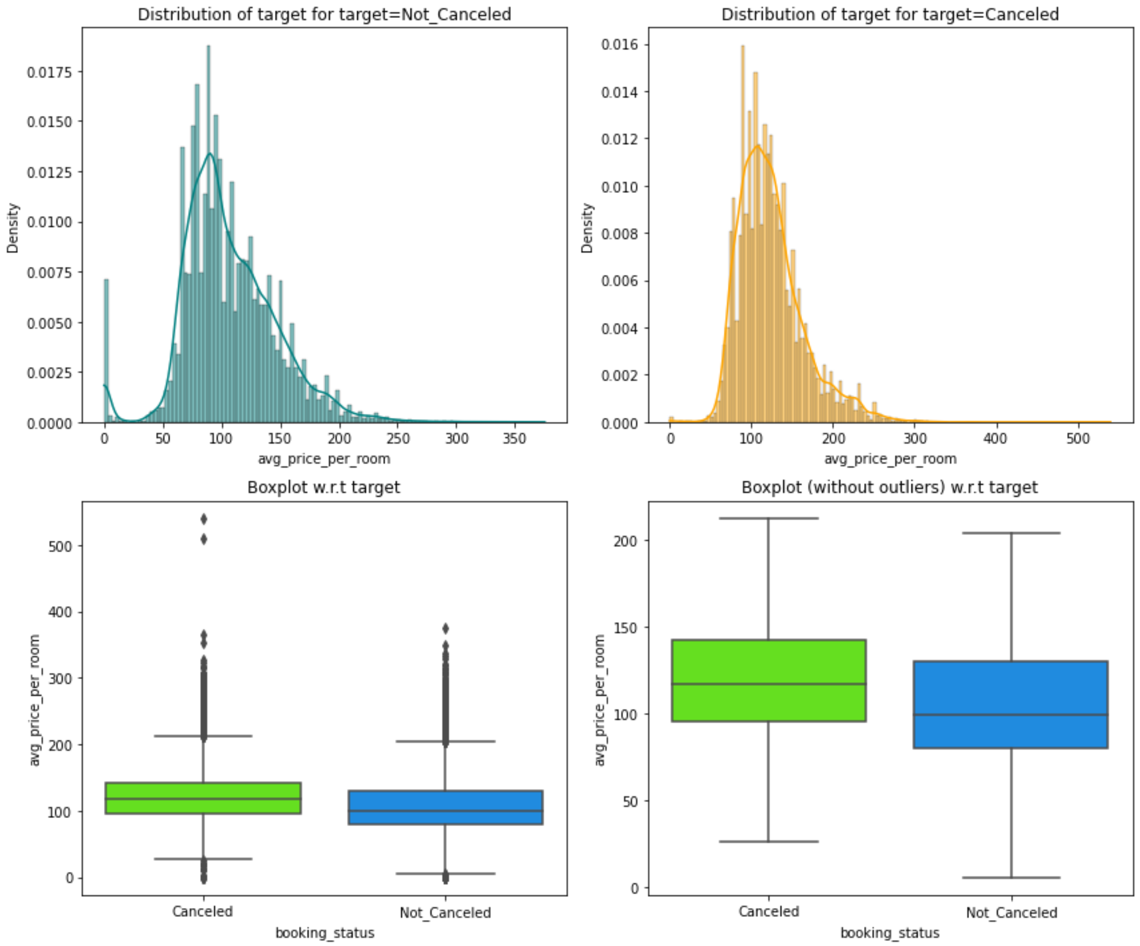
- Graphs showing the factors most heavily impacting the target attribute

Observations:

- average room price between 100-150 Euros has a little bit higher cancellation rate than those rooms with 90 - 110 Euros price range. Shows a little association.

Observations on other variables:

- number of adults , number of children, car parking, number of special request and repeated guests with no cancellation and those who cancelled, also have similar plots after outlier removal.

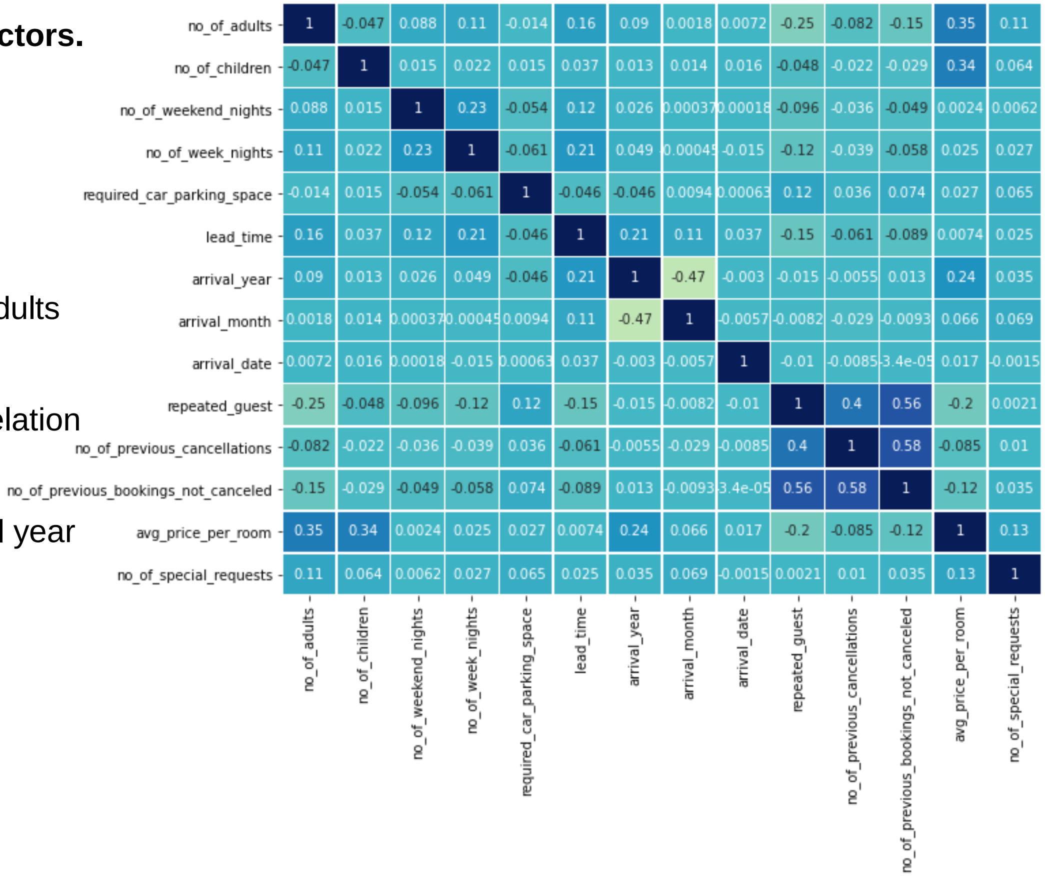


EDA

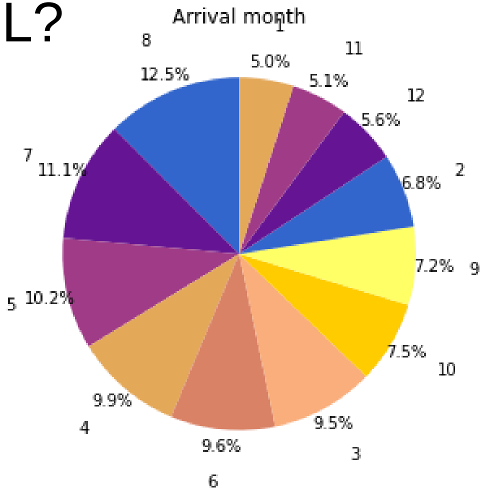
Correlation map showing association between predictors.

<u>observations:</u>

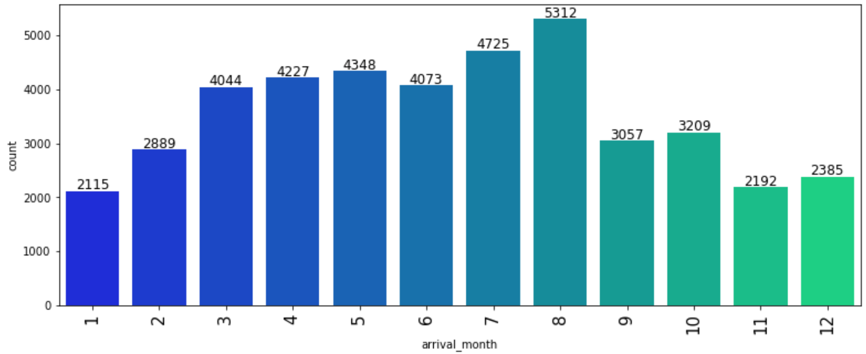
- repeated guests have 0.4 correlation with no_of_previous_cancellations and 0.6 correlation with no_of_previous_bookings_not_canceled
- avg_price_per_room has 0.4 correlation with no_of_adults and no_of_children
- no_of_previous_bookings_not_canceled has 0.6 correlation with no_of_previous_cancellations
- arrival_month has -0.5 negative correlation with arrival year

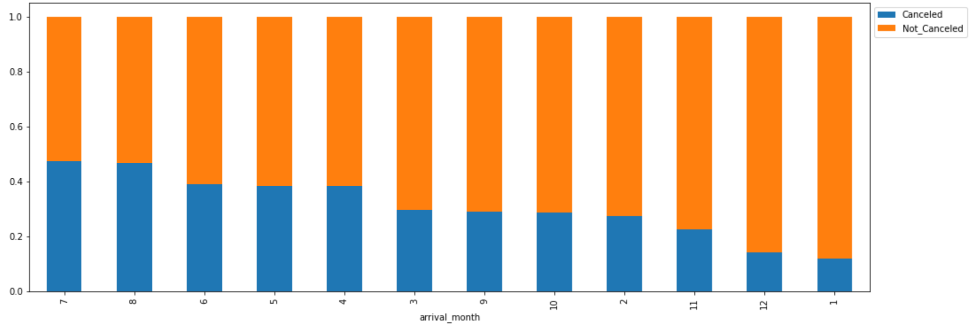


WHAT ARE THE BUSIEST MONTHS IN THE HOTEL?



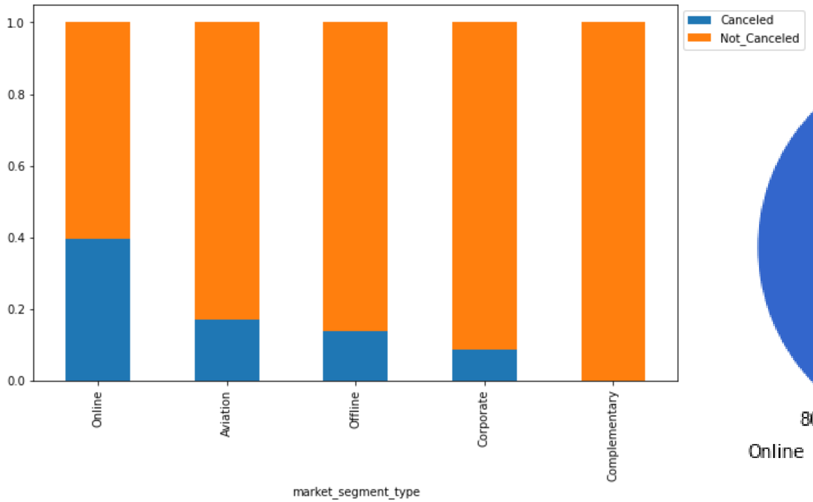
- August is the busiest month with 5312 entries
- July is 2nd busiest with 4725 entries
- January is the least busiest month.

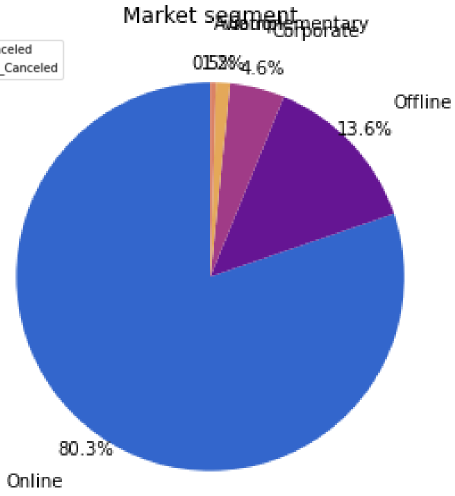


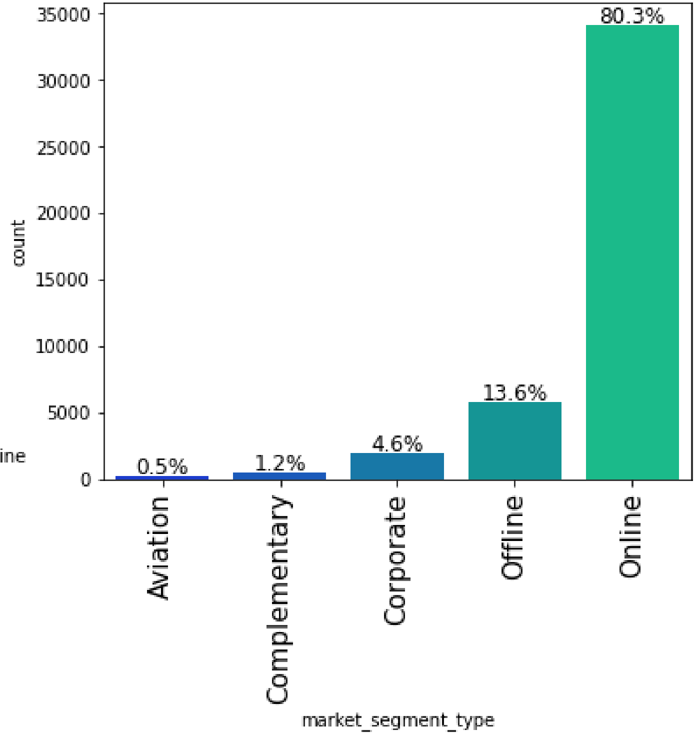


WHICH MARKET SEGMENT DO MOST OF THE GUESTS COME FROM?

- Most of the guest , 80.3% come from online market segment.
- 13.6% come from offline market segment
- Only 4.5% from corporate, 1.2% from complementary and 0.5% come from Aviation market segment.

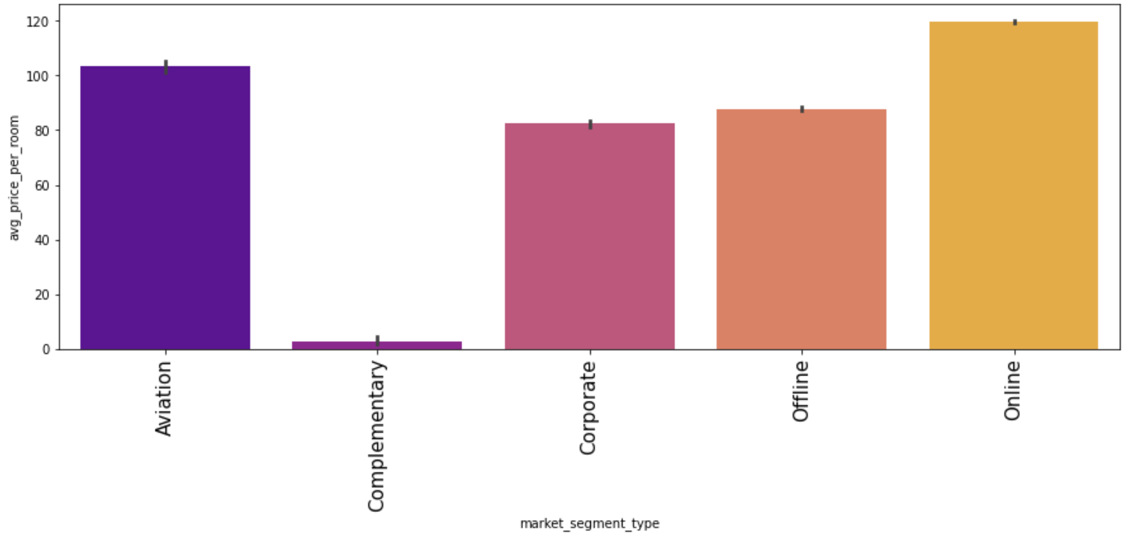


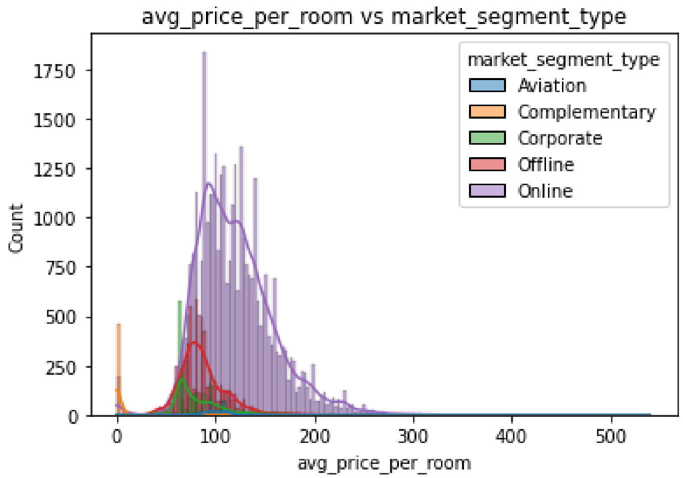




WHAT ARE THE DIFFERENCES IN ROOM PRICES IN DIFFERENT MARKET SEGMENTS?

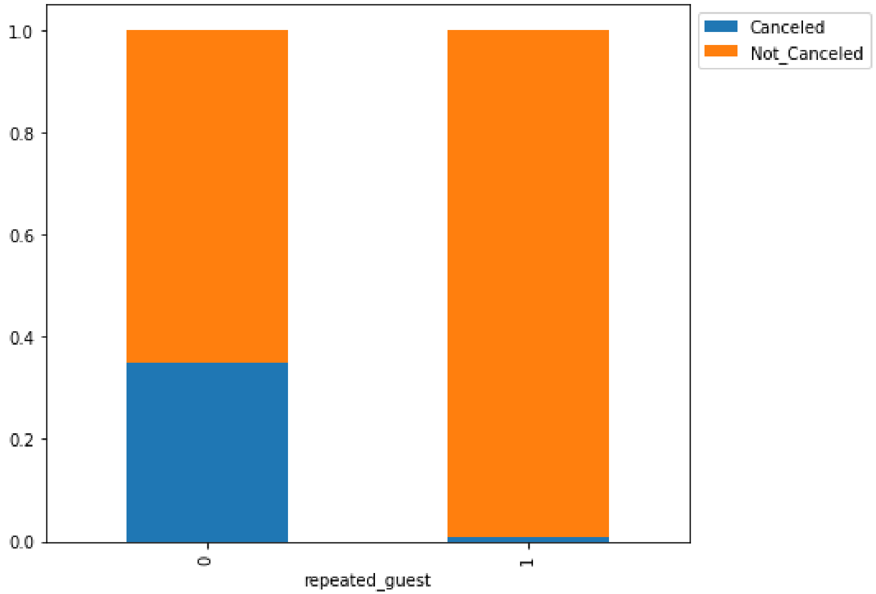
- Online market has the highest average room price.
- Aviation has second highest average room price.
- Complimentary has the lowest average room price, which is understandable as its complimentary.

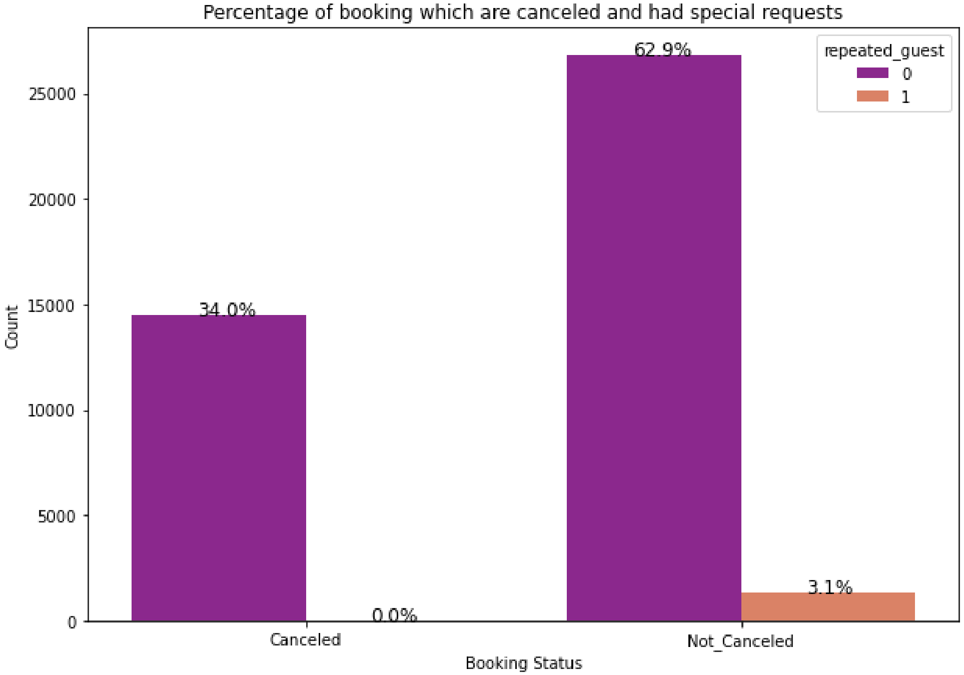




WHAT PERCENTAGE OF REPEATING GUESTS CANCEL?

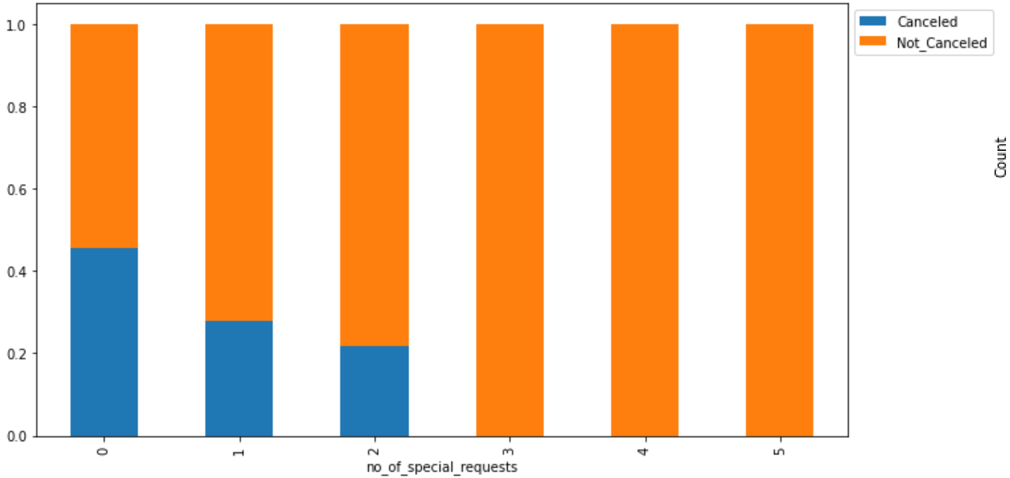
- All 34% bookings which were cancelled were not repeated guests.
- Majority bookings, 62.9% which were not cancelled were also not repeated guests.
- Only 3.1% of guests who did not cancel booking were repeated guests.

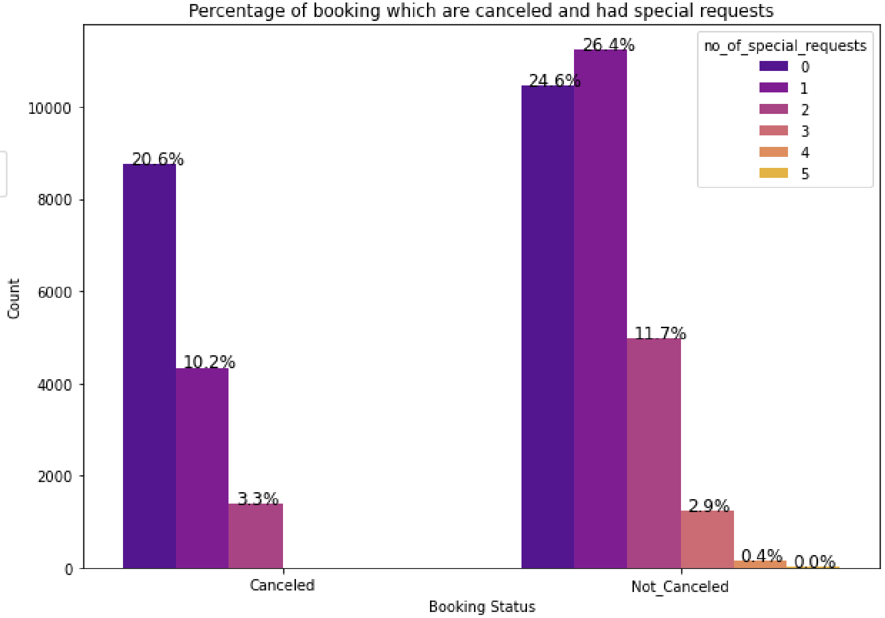




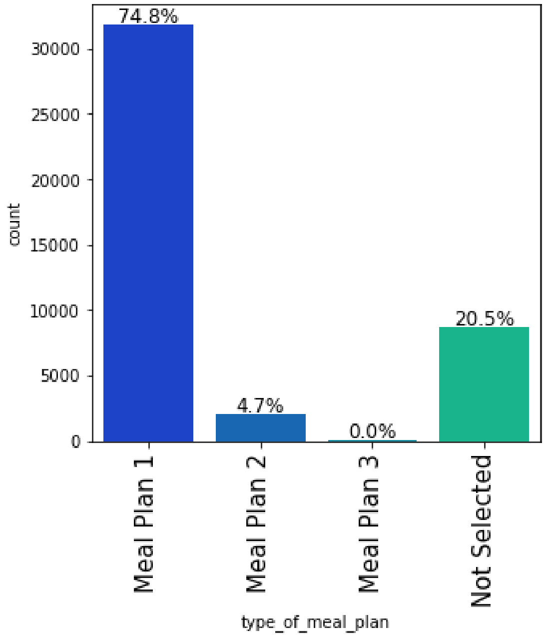
DO THESE SPECIAL REQUIREMENTS REQUESTS AFFECT BOOKING CANCELLATION?

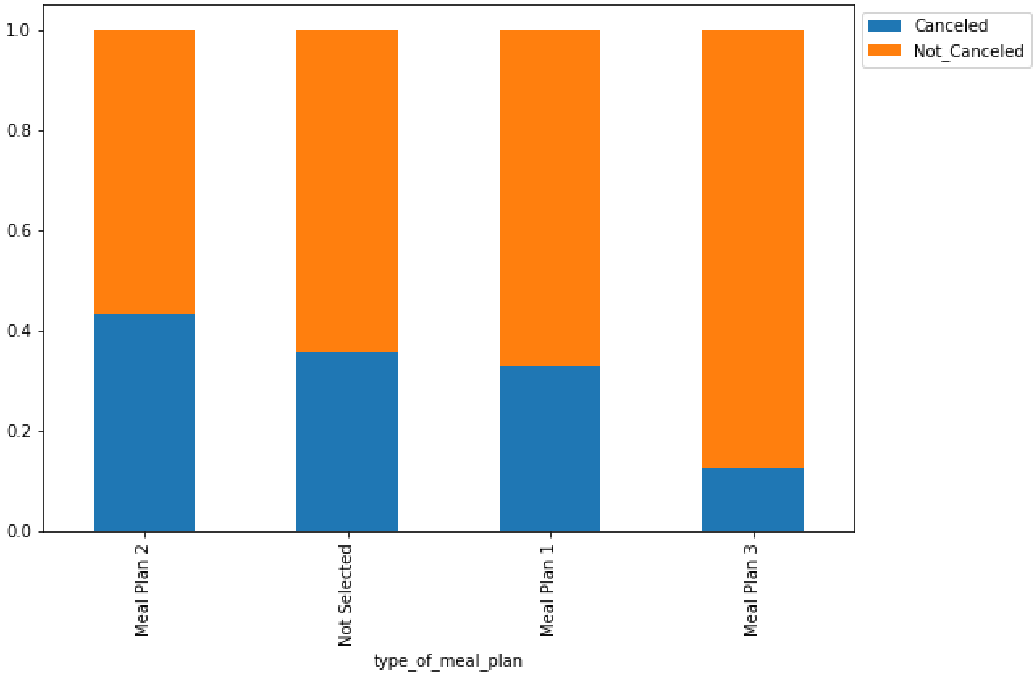
- Majority of guests, 20.6%, who cancelled bookings do not have a special request.
- Only 10.2% of guests, who cancelled bookings have a special request number 1.
- Majority of guests, 26.4%, who did not cancelled bookings have a special request number 1.
- 24.6%, guests who did not cancel booking have no special request.





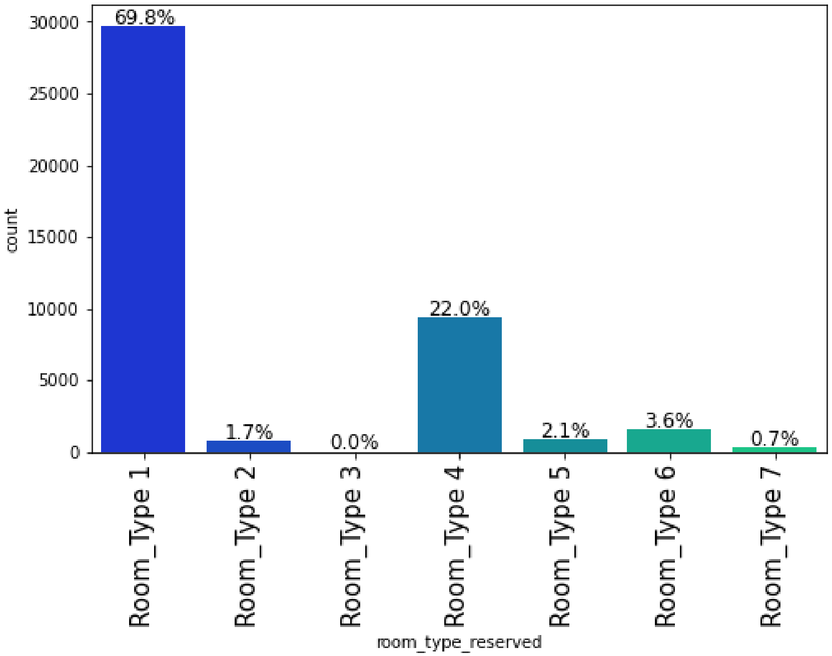
DOES MEAL PLAN HAVE ANY ASSOCIATION WITH BOOKING CANCELLATION?

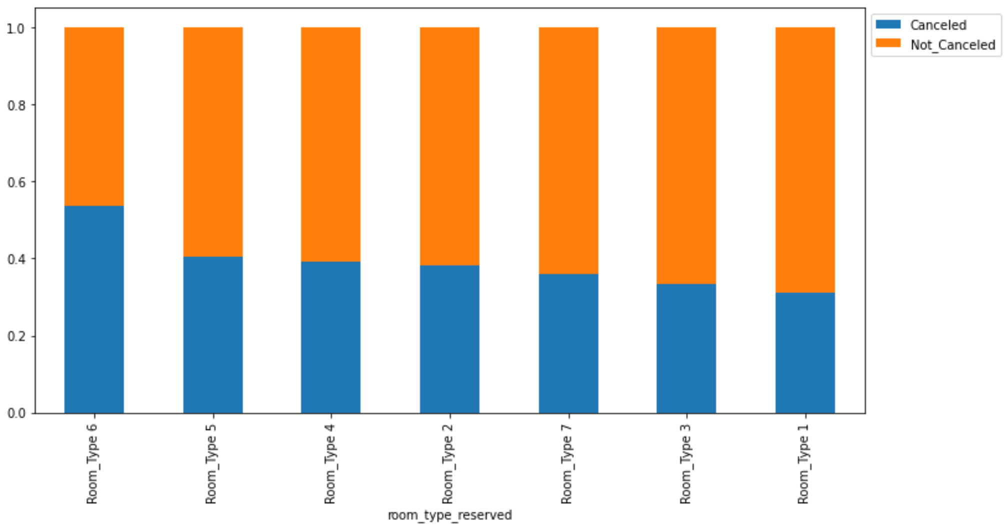




- Less cancellations on meal plan 3
- meal plan 2 has most cancellations.

DOES ROOM TYPE HAVE ANY ASSOCIATION WITH BOOKING CANCELLATION?

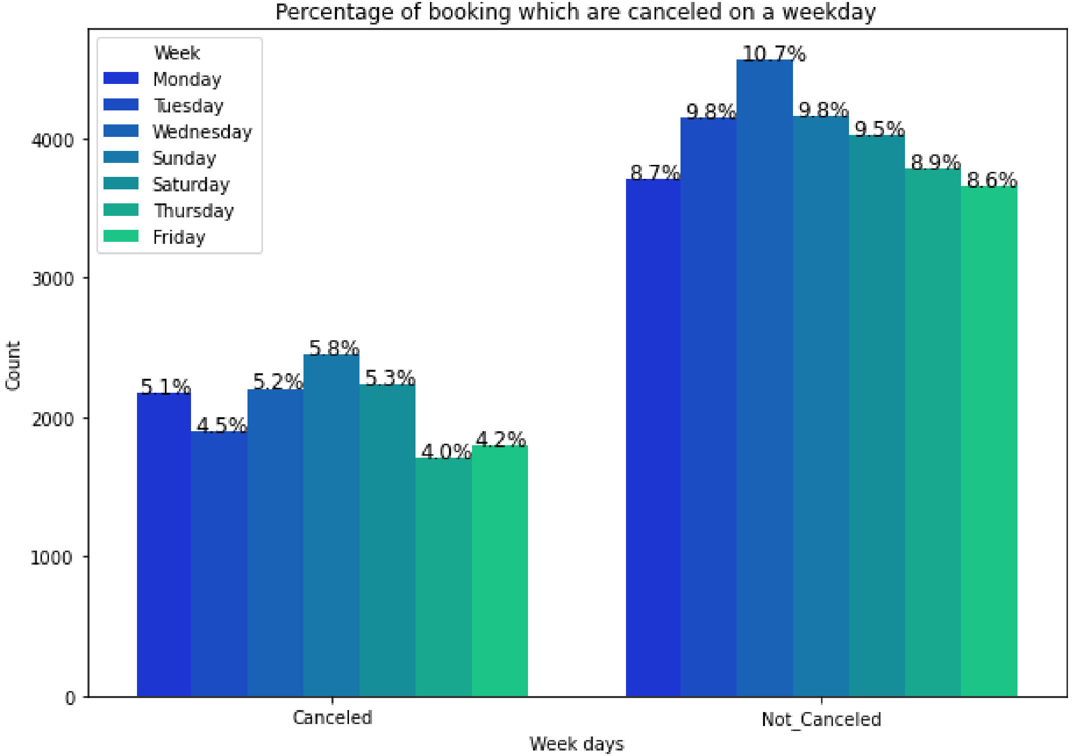




- Room type 6 has most cancellations
- Room type 1, 3 , 7 have less cancellations as compared to others.

DOES WEEK DAY HAS ANY ASSOCIATION WITH BOOKING CANCELLATION?

- Highest cancellation rate is for Sunday: 5.8%. Lowest cancellation was on Thursday 4%
- Highest non cancellation rate is for Wednesday: 10.7%.Lowest non cancellation was on Friday 8.6%



MODEL PERFORMANCE SUMMARY

Overview of ML model and its parameters:

- Multiple Linear Regression model was built to
  - find dependency of target variable: booking_status on predictors and
  - Predict fitted values and compare them to actual values
- **Total rows and columns** after data preprocessing: 42541 rows × 17 columns
- **Target variable:** booking_status
- **Predictors:** no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, type_of_meal_plan , required_car_parking_space, room_type_reserved, lead_time, market_segment_type, repeated_guest, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests, booking_status, Day of the Week, yearly_MONTHS

MODEL PERFORMANCE SUMMARY

Overview of ML model and its parameters:

- Data was divided into Train and Test at 60:40 ratio.
- **Number of rows in train data** = 25524,
- **Number of rows in test data** = 17017
- Original Canceled True Values : 14480 (34.04%)
- Original Canceled False Values : 28061 (65.96%)

- Training Canceled True Values : 8683 (34.02%)
- Training Canceled False Values : 16841 (65.98%)

- Test Canceled True Values : 5797 (34.07%)
- Test Canceled False Values : 11220 (65.93%)

MODEL EVALUATION CRITERION

Model can make wrong predictions as:

- Predicting a customer will not cancel the booking but in reality the customer would cancel.
- Predicting a customer will cancel booking but in reality the customer would not cancel.
- Which case is more important?
- If we predict a non-cancelling customer as a cancelling customer hotel would lose opportunity.
  - How to reduce this loss i.e need to reduce False Positives?
- If we predict a cancelling customer as a non-cancelling customer hotel would lose revenue.
  - How to reduce this loss i.e need to reduce False Negatives? recall should be maximized,
- the greater the recall higher the chances of minimizing the false negatives.

LOGISTIC REGRESSION MODEL USING SKLEARN LIBRARY

The first ML model was tested using the following performance matrices:

```
Training set performance:
Accuracy: 0.7937235543018336
Precision: 0.7287817938420348
Recall: 0.6269722446159162
F1: 0.6740543552281311
```

```
Test set performance:
Accuracy: 0.7966739143209731
Precision: 0.7344972907886815
Recall: 0.6313610488183543
F1: 0.6790352504638218
```

- Logistic Regression model is giving generalized results on both training and testing dataset.
- Recall , Precision of both sets is comparable.
- Recall needs to be improved to decrease False Negatives for minimum financial loss.
- Precision should not be too minimized to lose potential customer.
- There should be a balance between Precision and Recall.

LOGISTIC REGRESSION MODEL USING STATS LIBRARY

No feature has p-value greater than 0.05, so we'll consider the features in *X_train3* as the final ones and *lg2* as final model.

Coefficient interpretations

Coefficient of

'lead_time',

market_segment_type_Online,

'avg_price_per_room',

'no_of_weekend_nights',

'no_of_week_nights'

'type_of_meal_plan_Not Selected',

are positive; an increase in these will lead to an increase in chances of a customer cancelling the booking.

Dep. Variable:	booking_status	No. Observations:	29778
Model:	Logit	Df Residuals:	29738
Method:	MLE	Df Model:	39
Date:	Fri, 17 Sep 2021	Pseudo R-squ.:	0.3375
Time:	19:35:03	Log-Likelihood:	-12658.
converged:	True	LL-Null:	-19107.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9526	0.263	-7.434	0.000	-2.467	-1.438
no_of_weekend_nights	0.0623	0.018	3.449	0.001	0.027	0.098
no_of_week_nights	0.0975	0.011	8.619	0.000	0.075	0.120
lead_time	0.0170	0.000	58.649	0.000	0.016	0.018
avg_price_per_room	0.0189	0.001	25.580	0.000	0.017	0.020
no_of_special_requests	-1.3607	0.024	-56.643	0.000	-1.408	-1.314
type_of_meal_plan_Meal Plan 2	-0.2042	0.081	-2.527	0.012	-0.363	-0.046
type_of_meal_plan_Not Selected	0.3432	0.044	7.874	0.000	0.258	0.429
room_type_reserved_Room_Type 2	-0.2246	0.129	-1.735	0.083	-0.478	0.029
room_type_reserved_Room_Type 4	-0.1808	0.044	-4.135	0.000	-0.266	-0.095
room_type_reserved_Room_Type 5	-0.3984	0.113	-3.516	0.000	-0.620	-0.176
room_type_reserved_Room_Type 6	-0.5064	0.101	-5.023	0.000	-0.704	-0.309
room_type_reserved_Room_Type 7	-0.9387	0.203	-4.626	0.000	-1.336	-0.541
market_segment_type_Offline	-1.3566	0.115	-11.795	0.000	-1.582	-1.131
market_segment_type_Online	0.8475	0.102	8.302	0.000	0.647	1.048
yearly_MONTHS_2017-08	-1.8344	0.277	-6.623	0.000	-2.377	-1.292
yearly_MONTHS_2017-09	-2.6762	0.275	-9.721	0.000	-3.216	-2.137
yearly_MONTHS_2017-10	-3.1336	0.289	-10.833	0.000	-3.701	-2.567
yearly_MONTHS_2017-11	-2.7543	0.361	-7.631	0.000	-3.462	-2.047
yearly_MONTHS_2017-12	-3.7077	0.355	-10.431	0.000	-4.404	-3.011
yearly_MONTHS_2018-01	-4.1715	0.403	-10.355	0.000	-4.961	-3.382
yearly_MONTHS_2018-02	-1.5532	0.257	-6.045	0.000	-2.057	-1.050
yearly_MONTHS_2018-03	-1.7679	0.250	-7.065	0.000	-2.258	-1.277
yearly_MONTHS_2018-04	-2.1245	0.250	-8.509	0.000	-2.614	-1.635
yearly_MONTHS_2018-05	-2.4278	0.252	-9.621	0.000	-2.922	-1.933
yearly_MONTHS_2018-06	-2.2475	0.252	-8.904	0.000	-2.742	-1.753
yearly_MONTHS_2018-07	-2.4666	0.251	-9.831	0.000	-2.958	-1.975
yearly_MONTHS_2018-08	-2.3419	0.250	-9.353	0.000	-2.833	-1.851
yearly_MONTHS_2018-09	-2.0947	0.253	-8.291	0.000	-2.590	-1.600
yearly_MONTHS_2018-10	-2.0602	0.252	-8.187	0.000	-2.553	-1.567
yearly_MONTHS_2018-11	-1.7170	0.253	-6.781	0.000	-2.213	-1.221
yearly_MONTHS_2018-12	-3.1144	0.260	-11.984	0.000	-3.624	-2.605
yearly_MONTHS_2019-01	-2.5793	0.261	-9.865	0.000	-3.092	-2.067
yearly_MONTHS_2019-02	-1.5461	0.253	-6.121	0.000	-2.041	-1.051
yearly_MONTHS_2019-03	-1.9576	0.250	-7.824	0.000	-2.448	-1.467
yearly_MONTHS_2019-04	-2.1908	0.252	-8.709	0.000	-2.684	-1.698
yearly_MONTHS_2019-05	-2.4297	0.252	-9.633	0.000	-2.924	-1.935
yearly_MONTHS_2019-06	-2.5605	0.253	-10.117	0.000	-3.057	-2.064

LOGISTIC REGRESSION MODEL USING STATS LIBRARY

Coefficient interpretations

Coefficients of 'no_of_special_requests','type_of_meal_plan_Meal Plan 2', 'room_type_reserved_Room_Type 2', room_type_reserved_Room_Type4', 'room_type_reserved_Room_Type 5', 'room_type_reserved_Room_Type6', 'room_type_reserved_Room_Type 7', 'market_segment_type_Offline','yearly_MONTHS_2017-08', 'yearly_MONTHS_2017-09', 'yearly_MONTHS_2017-10', 'yearly_MONTHS_2017-11', 'yearly_MONTHS_2017-12', 'yearly_MONTHS_2018-01', 'yearly_MONTHS_2018-02', 'yearly_MONTHS_2018-03', 'yearly_MONTHS_2018-04', 'yearly_MONTHS_2018-05', 'yearly_MONTHS_2018-06', 'yearly_MONTHS_2018-07', 'yearly_MONTHS_2018-08', 'yearly_MONTHS_2018-09', 'yearly_MONTHS_2018-10', 'yearly_MONTHS_2018-11', 'yearly_MONTHS_2018-12', 'yearly_MONTHS_2019-01', 'yearly_MONTHS_2019-02', 'yearly_MONTHS_2019-03', 'yearly_MONTHS_2019-04', 'yearly_MONTHS_2019-05', 'yearly_MONTHS_2019-06', 'yearly_MONTHS_2019-07', 'yearly_MONTHS_2019-08'', are **negative**; an increase in these will lead to a decrease in chances of a customer cancelling a booking.

Dep. Variable:	booking_status	No. Observations:	29778
Model:	Logit	Df Residuals:	29738
Method:	MLE	Df Model:	39
Date:	Fri, 17 Sep 2021	Pseudo R-squ.:	0.3375
Time:	19:35:03	Log-Likelihood:	-12658.
converged:	True	LL-Null:	-19107.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9526	0.263	-7.434	0.000	-2.467	-1.438
no_of_weekend_nights	0.0623	0.018	3.449	0.001	0.027	0.098
no_of_week_nights	0.0975	0.011	8.619	0.000	0.075	0.120
lead_time	0.0170	0.000	58.649	0.000	0.016	0.018
avg_price_per_room	0.0189	0.001	25.580	0.000	0.017	0.020
no_of_special_requests	-1.3607	0.024	-56.643	0.000	-1.408	-1.314
type_of_meal_plan_Meal Plan 2	-0.2042	0.081	-2.527	0.012	-0.363	-0.046
type_of_meal_plan_Not Selected	0.3432	0.044	7.874	0.000	0.258	0.429
room_type_reserved_Room_Type 2	-0.2246	0.129	-1.735	0.083	-0.478	0.029
room_type_reserved_Room_Type 4	-0.1808	0.044	-4.135	0.000	-0.266	-0.095
room_type_reserved_Room_Type 5	-0.3984	0.113	-3.516	0.000	-0.620	-0.176
room_type_reserved_Room_Type 6	-0.5064	0.101	-5.023	0.000	-0.704	-0.309
room_type_reserved_Room_Type 7	-0.9387	0.203	-4.626	0.000	-1.336	-0.541
market_segment_type_Offline	-1.3566	0.115	-11.795	0.000	-1.582	-1.131
market_segment_type_Online	0.8475	0.102	8.302	0.000	0.647	1.048
yearly_MONTHS_2017-08	-1.8344	0.277	-6.623	0.000	-2.377	-1.292
yearly_MONTHS_2017-09	-2.6762	0.275	-9.721	0.000	-3.216	-2.137
yearly_MONTHS_2017-10	-3.1336	0.289	-10.833	0.000	-3.701	-2.567
yearly_MONTHS_2017-11	-2.7543	0.361	-7.631	0.000	-3.462	-2.047
yearly_MONTHS_2017-12	-3.7077	0.355	-10.431	0.000	-4.404	-3.011
yearly_MONTHS_2018-01	-4.1715	0.403	-10.355	0.000	-4.961	-3.382
yearly_MONTHS_2018-02	-1.5532	0.257	-6.045	0.000	-2.057	-1.050
yearly_MONTHS_2018-03	-1.7679	0.250	-7.065	0.000	-2.258	-1.277
yearly_MONTHS_2018-04	-2.1245	0.250	-8.509	0.000	-2.614	-1.635
yearly_MONTHS_2018-05	-2.4278	0.252	-9.621	0.000	-2.922	-1.933
yearly_MONTHS_2018-06	-2.2475	0.252	-8.904	0.000	-2.742	-1.753
yearly_MONTHS_2018-07	-2.4666	0.251	-9.831	0.000	-2.958	-1.975
yearly_MONTHS_2018-08	-2.3419	0.250	-9.353	0.000	-2.833	-1.851
yearly_MONTHS_2018-09	-2.0947	0.253	-8.291	0.000	-2.590	-1.600
yearly_MONTHS_2018-10	-2.0602	0.252	-8.187	0.000	-2.553	-1.567
yearly_MONTHS_2018-11	-1.7170	0.253	-6.781	0.000	-2.213	-1.221
yearly_MONTHS_2018-12	-3.1144	0.260	-11.984	0.000	-3.624	-2.605
yearly_MONTHS_2019-01	-2.5793	0.261	-9.865	0.000	-3.092	-2.067
yearly_MONTHS_2019-02	-1.5461	0.253	-6.121	0.000	-2.041	-1.051
yearly_MONTHS_2019-03	-1.9576	0.250	-7.824	0.000	-2.448	-1.467
yearly_MONTHS_2019-04	-2.1908	0.252	-8.709	0.000	-2.684	-1.698
yearly_MONTHS_2019-05	-2.4297	0.252	-9.633	0.000	-2.924	-1.935
yearly_MONTHS_2019-06	-2.5605	0.253	-10.117	0.000	-3.057	-2.064

ASSUMPTIONS OF LOGISTIC REGRESSION

We will be checking ML model on Linear Regression assumptions:
No Multicollinearity: final ML model features did not have VIF> 5

no_of_weekend_nights	2.733855
no_of_week_nights	4.224263
lead_time	2.916432
avg_price_per_room	26.353510
no_of_special_requests	2.024628
Day of the Week	4.273069
type_of_meal_plan_Meal Plan 2	1.170414
type_of_meal_plan_Meal Plan 3	1.021418
type_of_meal_plan_Not Selected	1.683376
room_type_reserved_Room_Type 2	1.057852
room_type_reserved_Room_Type 3	1.001528
room_type_reserved_Room_Type 4	1.732824
room_type_reserved_Room_Type 5	1.147140
room_type_reserved_Room_Type 6	1.489206
room_type_reserved_Room_Type 7	1.142764
market_segment_type_Complementary	1.359042
market_segment_type_Corporate	1.391719
market_segment_type_Online	8.782936
yearly_MONTHS_2017-08	1.329295
yearly_MONTHS_2017-09	1.565355
yearly_MONTHS_2017-10	1.492861
yearly_MONTHS_2017-11	1.176909
yearly_MONTHS_2017-12	1.303241
yearly_MONTHS_2018-01	1.313814
yearly_MONTHS_2018-02	1.541629
yearly_MONTHS_2018-03	2.039171
yearly_MONTHS_2018-04	2.204071
yearly_MONTHS_2018-05	2.270642
yearly_MONTHS_2018-06	2.242922
yearly_MONTHS_2018-07	2.574963
yearly_MONTHS_2018-08	2.923821
yearly_MONTHS_2018-09	2.682193
yearly_MONTHS_2018-10	2.585414
yearly_MONTHS_2018-11	2.050693
yearly_MONTHS_2018-12	2.047680
yearly_MONTHS_2019-01	1.736283
yearly_MONTHS_2019-02	1.869963
yearly_MONTHS_2019-03	2.167479
yearly_MONTHS_2019-04	2.873150
yearly_MONTHS_2019-05	3.343459

No significant Multicollinearity is present among features other than average price.

COEFFICIENT INTERPRETATIONS OF SOME IMPORTANT VARIABLES

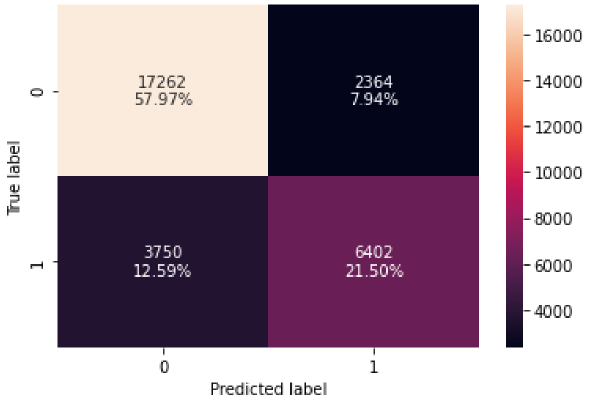
- **lead_time**: The odds of a customer who has a lead time in days, cancelling booking is 0.1.02 times or 1.72% more odds.
- **no_of_special_requests**: Holding all other features constant a unit change in no_of_special_requests will decrease the odds of a customer cancelling booking by 0.26 times or a 74.35% decrease in odds.
- **avg_price_per_room:** Holding all other features constant a unit change in avg_price_per_room will increase the odds of a customer cancelling booking by 1.02 times or a 1.91% increase in odds.
- **market_segment_type_Online:** Holding all other features constant a unit change in market_segment_type_Online will increase the odds of a customer cancelling booking by 2.33 times or a 133.38% increase in odds.
- **market_segment_type_Offline:** Holding all other features constant a unit change in market_segment_type_Offline will decrease the odds of a customer cancelling booking by 0.26 times or a 73.3% decrease in odds.
- **no_of_weekend_nights:** Holding all other features constant a unit change in no_of_weekend_nights will increase the odds of a customer cancelling booking by 1.06 times or a 6.4% increase in the odds.
- **no_of_week_nights:** Holding all other features constant a unit change in no_of_weekend_nights will increase the odds of a customer cancelling booking by 1.1 times or a 10.2% increase in the odds.

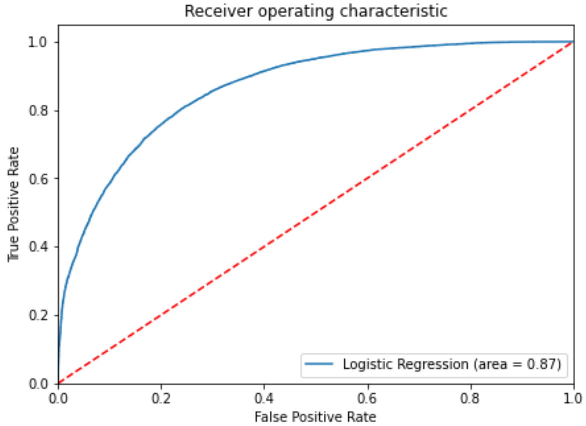
MODEL PERFORMANCE SUMMARY

<u>The first ML stats model performance and The confusion matrix:</u>

Training performance:

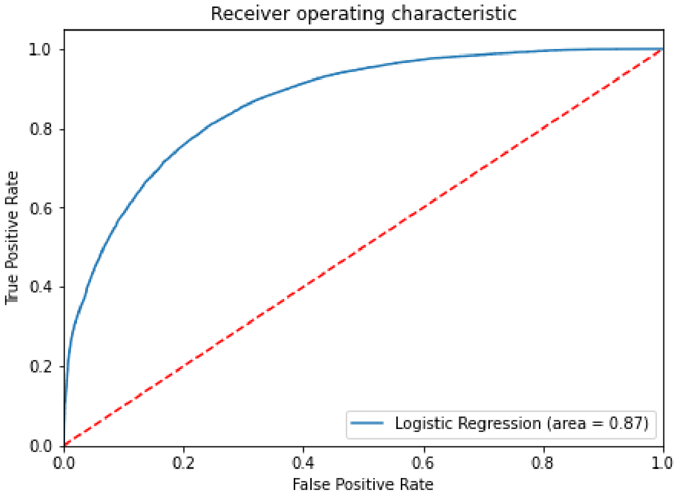
	Accuracy	Recall	Precision	F1
0	0.794681	0.630615	0.730322	0.676816

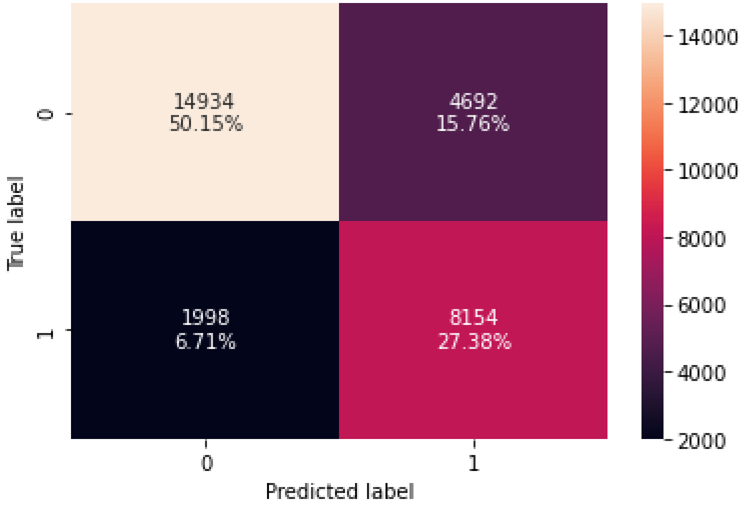




- True Positives (TP): we correctly predicted that they will cancel the booking and they actually cancelled are 6402 or 21.50%
- True Negatives (TN): we correctly predicted that they will not cancel the booking and they did not cancel are 17262 or 57.97%
- False Positives (FP): we incorrectly predicted that they they will cancel the booking and they actually did not cancelled are (a "Type I error") 2364 or 7.94% Falsely predict positive Type I error
- False Negatives (FN): we incorrectly predicted that they will not cancel the booking and they actually cancel (a "Type II error") 3750 or 12.59% Falsely predict negative Type II error
- Logistic Regression model is an ok recall and ROC-AUC score.

OPTIMAL THRESHOLD:0.323, USING AUC-ROC CURVE





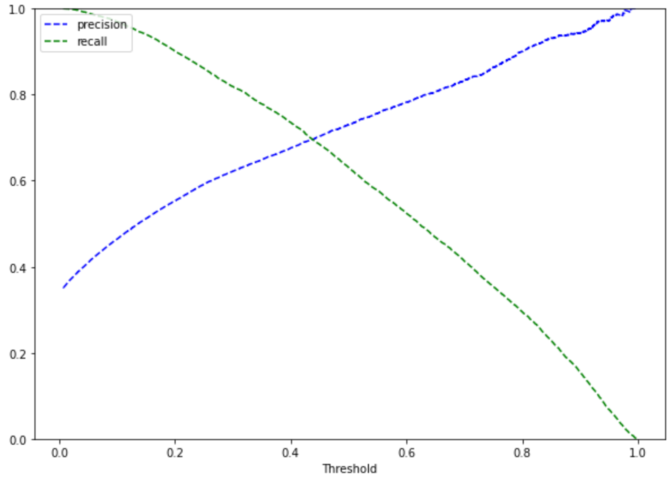
Training performance:

	Accuracy	Recall	Precision	F1
0	0.775337	0.803191	0.63475	0.709105

- Model performance has improved significantly.
- Model is giving a recall of 0.803 as compared to initial model which was giving a recall of 0.63.
- Precision has decreased from 0.73 to 0.64.

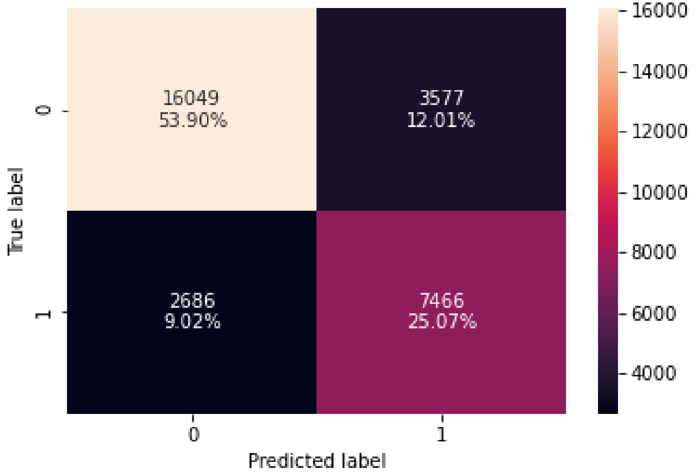
USING OPTIMAL THRESHOLD CURVE = 0.40

- Using Precision-Recall curve to find a better threshold
- At threshold around 0.42 we will get equal precision and recall but taking a step back and selecting value around 0.40 will provide a higher recall and a good precision.



Training performance:

	Accuracy	Recall	Precision	F1
0	0.789677	0.735422	0.676084	0.704506



- Recall has decreased as compared to the initial model.
- Model is giving a better performance with 0.323 threshold found using AUC-ROC curve.

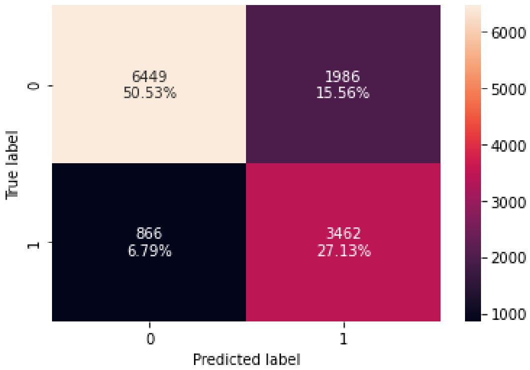
MODEL PERFORMANCE SUMMARY

Training performance comparison:

	Logistic Regression sklearn	Logistic Regression-0.323 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.794681	0.775337	0.789677
Recall	0.630615	0.803191	0.735422
Precision	0.730322	0.634750	0.676084
F1	0.676816	0.709105	0.704506

Test set performance comparison:

	Logistic Regression sklearn	Logistic Regression-0.323 Threshold	Logistic Regression-0.40 Threshold
Accuracy	0.799655	0.776542	0.793074
Recall	0.635397	0.799908	0.732440
Precision	0.737463	0.635463	0.681281
F1	0.682636	0.708265	0.705935



- **<u>Observations:</u>**
- The training and testing best recall are 80.31% and 79.99% respectively.
- Recall on the train and test sets are comparable.
- This shows that the model is giving a generalised result.
- **<u>Final Model Summary</u>**
- We'll consider the features in X_train3 as the final ones and lg2 as final model and threshold of 0.323 as final

MODEL PERFORMANCE EVALUATION

Conclusion

- All the models are giving a generalized performance on training and test set.
- The highest recall is 80.03% on the training set.
- Using the model with default threshold the model will give a low recall and good precision scores - - This model will help the hotel save resources but lose on potential customers.
- Using the model with 0.323 threshold the model will give a a balance recall and precision score - - This model will help the bank to maintain a balance in identifying potential customer and the cost of resources.
- Using the model with 0.40 threshold the model will give a a low recall and good precision scores - -- This model will help the hotel save resources but may lead to loss of potential customers.

RECOMMENDATIONS

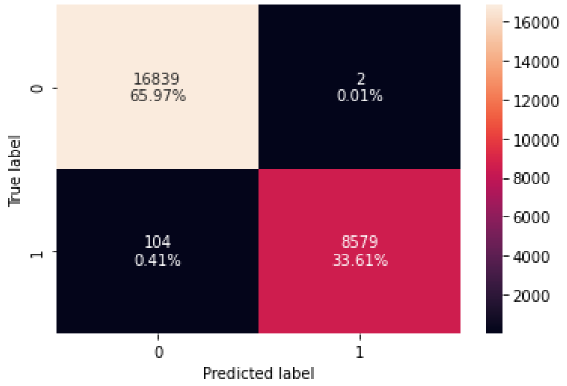
From our logistic regression model we identified that

- lead_time: The odds of a customer who has a more days in lead time, cancelling booking is greater.
- avg_price_per_room: change in avg_price_per_room will increase the odds of a customer cancelling booking.
- no_of_special_requests: A customer with no_of_special_requests is less likely to cancelling booking.
- market_segment_type_Online: A customer who booked online is more likely to cancel booking by.
- market_segment_type_Offline: A customer who booked offline is less likely to cancel the booking.
- Bookings done for yearly quarter 3 and 4 are less likely to be cancelled.
- Bookings done for room type 2,4,5, 6,7 are less likely to be cancelled.

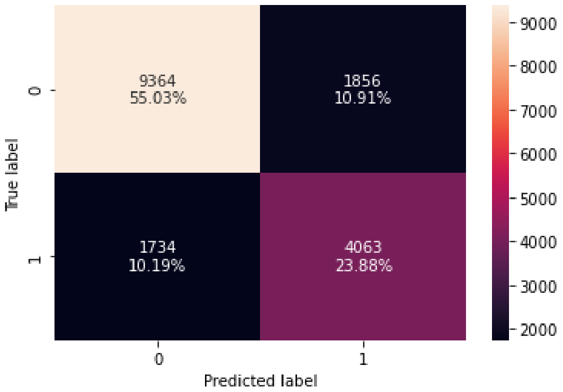
DECISION TREE MODEL

- We will build our model using the DecisionTreeClassifier function.
- Using default 'gini' criteria to split
- Model is able to perfectly classify all the data points on the training set.
- 99% recall on the training set, each sample has been classified correctly.
- 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.
- This generally leads to overfitting of the model as Decision Tree will perform well on the training set but will fail to replicate the performance on the test set.

Training Recall Score: 0.988



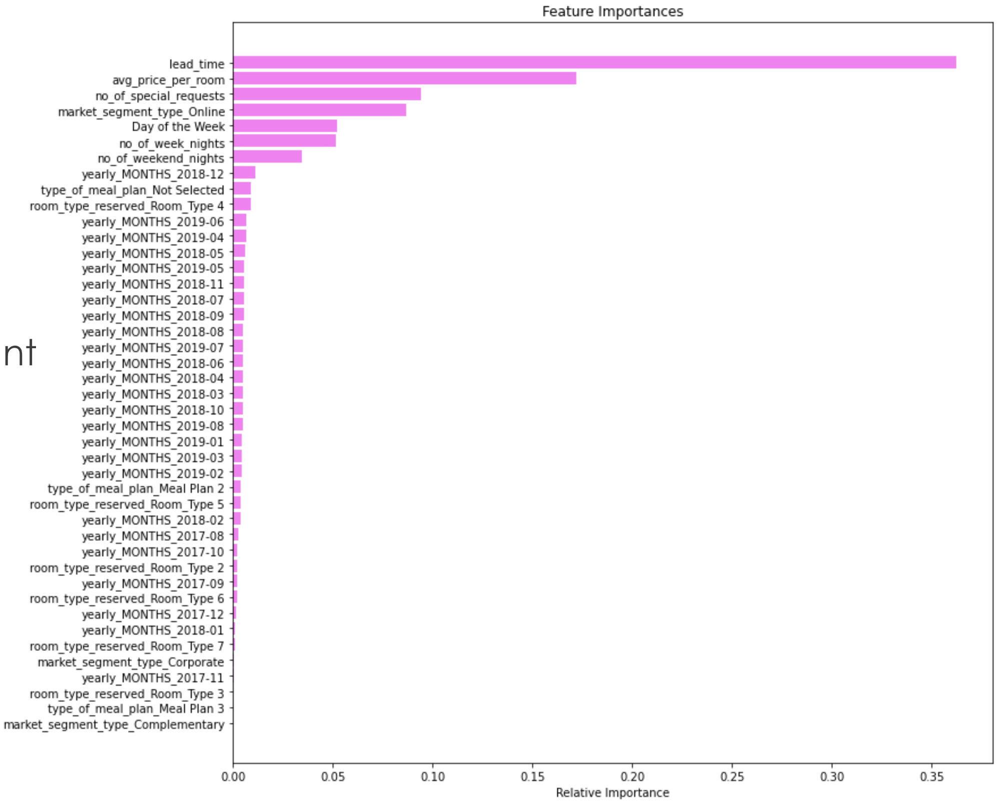
Testing Recall Score: 0.701



FINAL MODEL PERFORMANCE SUMMARY

According to the decision tree model,

- lead_time is the most important variable for predicting the booking_status.
- avg_price_per_room is second most important
- no_of_special_requests is third most special request.



REDUCING OVERFITTING OF THE DECISION TREE USING CCP-ALPHA: 0.071

Reducing over fitting

Using GridSearch for Hyperparameter tuning of our tree model

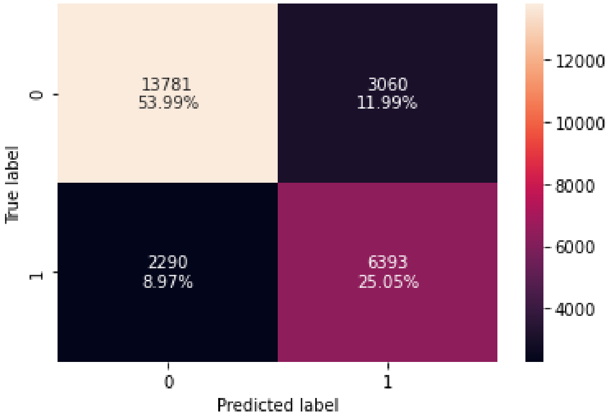
Hyperparameter tuning is also tricky in the sense that there is no direct way to calculate how a change in the hyperparameter value will reduce the loss of your model, we'll use Grid search to compute the optimum values of hyperparameters.

The parameters of the estimator/model used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

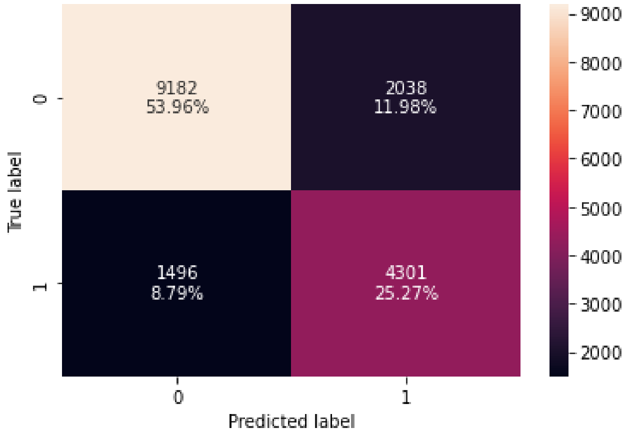
```
DecisionTreeClassifier(criterion='entropy', max_depth=5,
                       min_impurity_decrease=0.01, random_state=1)
```

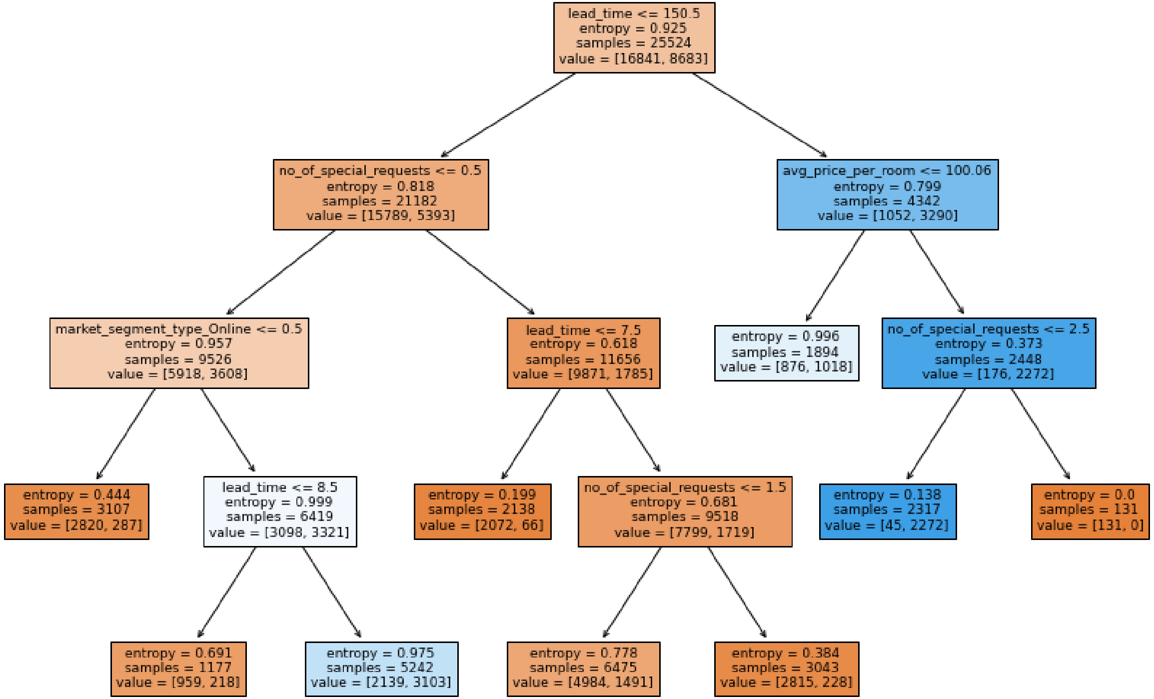
THE DECISION TREE USING CCP-ALPHA: 0.071

Training Set Recall Score: 0.736



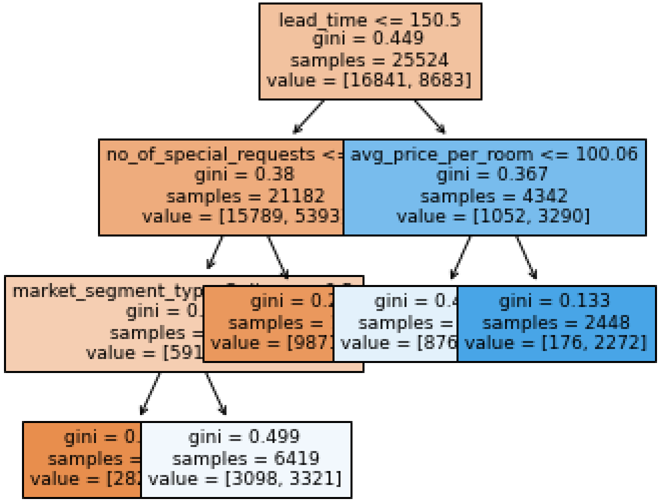
Testing Set Recall Score: 0.742





THE DECISION TREE USING CCP-ALPHA: 0.012, BEST MODEL FIT

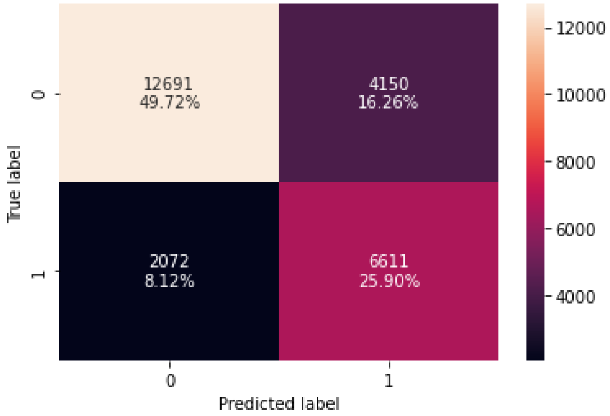
- This decision tree is too simple so a business would not be able to use it to actually predict the booking status.



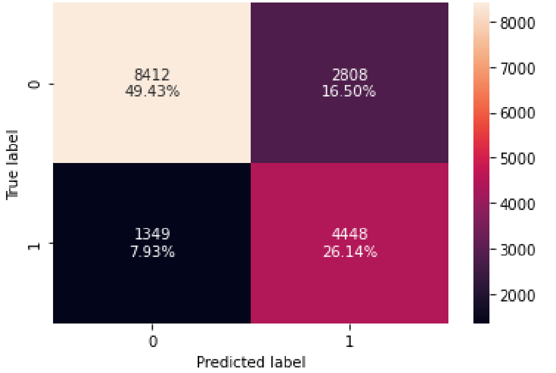

```
best_model.fit(X_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.012484589094136037, random_state=1)
```

Training set Recall Score: 0.761

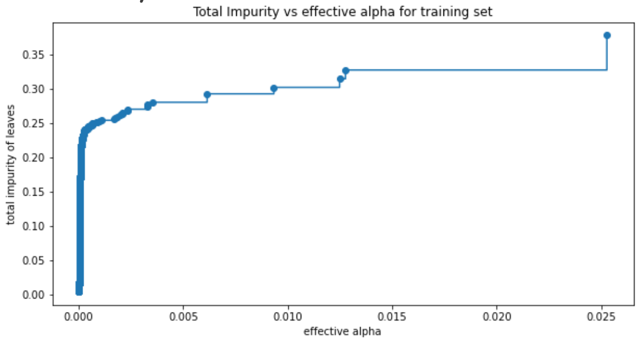


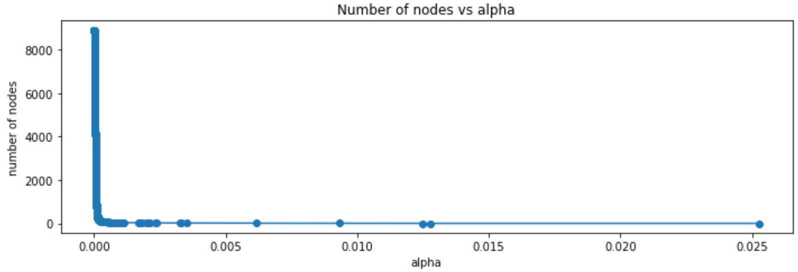
Testing set Recall Score: 0.767

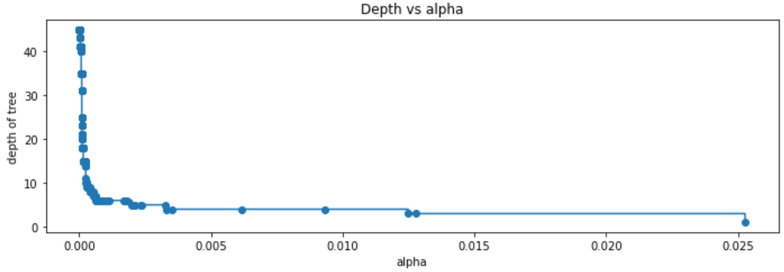


PRUNING THE DECISION TREE USING CCP-ALPHA

- For the remainder, we remove the last element in clfs and ccp_alphas, 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.



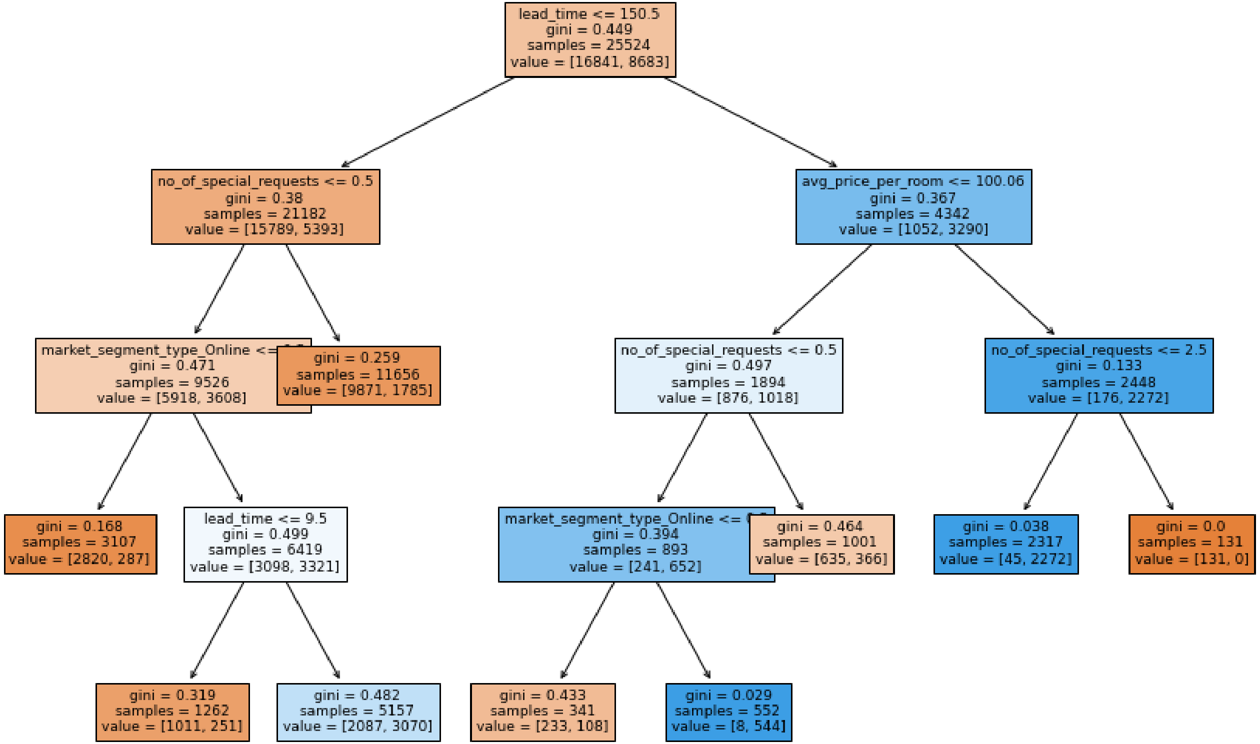




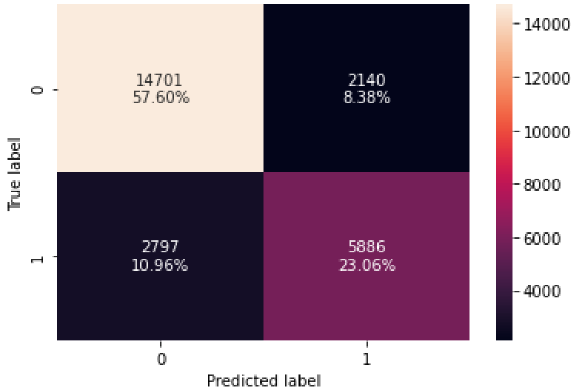
Maximum value of Recall is at 0.025 alpha, but if we choose decision tree will only have a root node and we would lose the buisness rules, instead we can choose alpha 0.004 retaining information and getting higher recall.

THE DECISION TREE USING CCP-ALPHA: 0.004

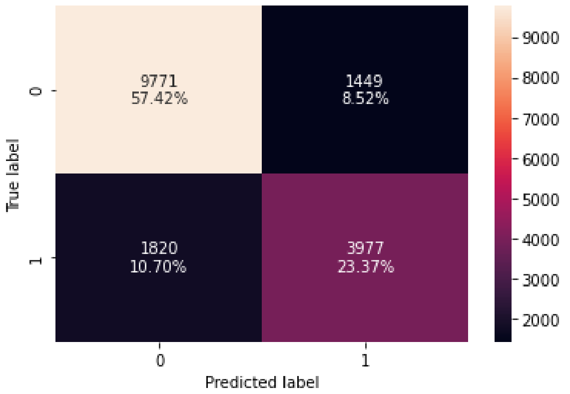
- The results have improved from the initial model.
- The performance is comparable to the hyperparameter tuned model.
- The model is giving a generalized performance on training and test set.



Training Set Recall Score: 0.68



Testing Set Recall Score: 0.68



MODEL PERFORMANCE COMPARISON AND CONCLUSIONS

Training Set Recall Score:

Training performance comparison:

	Recall on training set
0	0.988023
1	0.736266
2	0.677876

Testing Set Recall Score:

Test performance comparison:

	Recall on testing set
0	0.700880
1	0.741935
2	0.686045

- Decision tree model with pre-pruning has given the best recall score on training data.
- The pre-pruned and the post-pruned models have reduced overfitting and the model is giving a generalized performance.
- Last Model with ccp-alpha 0.004 is my final model for decision tree.

LOGISTIC REGRESSION MODEL VS DECISION TREE MODEL

From our logistic regression model we identified that following are the most important to predict booking cancellation by customer.

- lead_time:
- avg_price_per_room:
- no_of_special_requests:
- market_segment_type_Online:
- market_segment_type_Offline:

Training performance comparison:

	Logistic Regression sklearn	Logistic Regression-0.323 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.794681	0.775337	0.789677
Recall	0.630615	0.803191	0.735422
Precision	0.730322	0.634750	0.676084
F1	0.676816	0.709105	0.704506

Decision tree model

Training performance comparison:

	Recall on training set
0	0.988023
1	0.736266
2	0.677876

Both the model results are comparable.

BUSINESS INSIGHTS AND RECOMMENDATIONS

Conclusions

- We analyzed the "Booking cancellation status" 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. Easy interpretation is one of the key benefits of Decision Trees.
- We verified the fact that how much less data preparation is needed for Decision Trees and such a simple model gave good results even with outliers and imbalanced classes which shows the robustness of Decision Trees.
- lead_time ,no_of_special_requests ,market_segment_type_Online, avg_price_per_room are the most important variable in predicting the customers that will cancel the booking or not.
- We established the importance of hyper-parameters/ pruning to reduce overfitting.

BUSINESS INSIGHTS AND RECOMMENDATIONS

- According to the decision tree model -
- a) If a customer books with lead time less than 150 days and number of special requests is less than ~5 with the market segment online, then there is a very high chance that the customer is going to cancel the booking. b) If the room price >100 and number of special requests is greater than 2.5 then customer is less likely to cancel the booking.
- Potential Customers - Employ the predictive model to predict potential customers (customers who can book the room), Offer limited-time coupons/discounts on a real-time basis only to those customers. This can also be employed for the customers in months like July, August, April May, as in those months, the traffic is higher so these months have potential confirming customers.
- It is observed that less cancellations are seen on the Wednesday, Tuesday and Sunday, - the hotel should initiate schemes/offers on the special days with minimum lead time to attract more customers on such days.
- December and January were the months where the hotels saw the lowest booking cancellations, with further data it should be investigated what portfolios were running in those months and an inspiration to create more such portfolios can be drawn and implemented.

BUSINESS INSIGHTS AND RECOMMENDATIONS

- Customer retention - Member Loyalty programs initiatives like special discounts, coupons, etc can be provided.

- Better resource management - Tuesday, Wednesday and Sunday is when the hotel sees the most traffic, resources such as customer care services can be allocated more for these days.

- Hotel should make more complementary packages for repeated guests who come more frequently. Like give them complementary spa work.

THE END

BY: SYEDA AMBREEN KARIM BUKHARI