Project – 4
INN Hotels

Supervised Learning - Classification

Date: 06-Jan-2023

By Ann Mariya Jomon

Content / Agenda

1. Executive Summary
2. Business problem overview & Solution approach
3. EDA Results
4. Data preprocessing
5. Model performance summary
6. Appendix

1 Executive Summary

Conclusions

Logistic Regression

- The model with F1 score of 0.70 (i.e., having a threshold of 0.37) can be used to predict whether a booking will get cancelled or not.
- An increase in required car parking space, arrival month, repeated guest, number of special requests, certain room types, and corporate and offline market segment will reduce the chances of a booking being cancelled.
- An increase in number of adults, children, week nights, weekend nights, lead time, arrival year, number of previous cancellations, average price per room, meal plan 2, and meal plan not selected for a booking increases the chance of the booking being cancelled.

Decision Tree

- The post-pruned decision tree has the highest F1 score and hence this model will be used to predict the booking cancellations.
- Lead time, online market segment, average price per room, and number of special requests are the most important features in the prediction.

Recommendations

- The hotel will need to focus on the lead times and bookings made online since these are the key factors that determine the chances of a booking being cancelled.
- Since most of the bookings were made online, there are greater chances of customers comparing the prices of other hotels with INN Hotel's prices. This may or may not lead to cancellations. Therefore, INN Hotel must conduct market research and determine a competitive price to retain their guests.
- The hotel can also try to provide more complimentary services to their guests. This reduces the likelihood of the booking being cancelled.
- They can also provide more special requests that are more personalized such as special floral arrangement or something for special occasions like birthdays, anniversaries, etc.
- The hotel can try to retain many of their existing and new guests by introducing customer loyalty programs and reward schemes that can provide incentives for the customers to book a room at INN Hotels and reduces the chance of the booking being cancelled.

2

Business problem overview & Solution approach

Business problem overview

INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations.

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled.

Solution approach

1. **Exploratory data analysis**
   (univariate/bivariate analysis)
2. **Data preprocessing**
   (missing values/duplicates/feature engineering/outlier check/data preparation for modeling)
3. **Logistic Regression**
   (building model/multicollinearity check/model performance check with different thresholds)
4. **Decision Tree**
   (building model/pre-pruning/post-pruning/important features)
5. **Insights & Recommendations**

3

EDA Results

Link to appendix

Univariate Analysis – Key Results

LEAD TIME

The lead time ranges from **0 to 443 days** and the distribution is skewed to the right.

There are **243** bookings with lead times crossing a year.

Most of the bookings have lead times of **0-1 day**.

AVG. PRICE PER ROOM

Average price per room displays a somewhat **normal distribution** with majority of the prices clustered around **100 euros**.

1.5% (545) of total bookings were provided at free of cost.

Out of this, **65%** is complementary and **35%** is online.

ROOM TYPE

78% of the bookings reserved for Room_Type 1.

Room_Type 1, Room_Type 4, and Room_Type 6 are the top three room types reserved by the customers.

SPECIAL REQUEST

Majority of the bookings **(55%)** do not have any special requests.

31% of the bookings have one special request

Link to appendix

Univariate Analysis – Key Results

ADULTS & CHILDREN

72% of the total bookings have 2 adults

93% of the bookings have no children

WEEKEND/WEEK NIGHTS

Majority of the bookings (**32%**) are for 2 week nights.

47% of the bookings, do not have any weekend nights.

ARRIVAL MONTH

Most of the arrivals are during **October, September, and August.**

15% of the arrivals are during the month of October which is the highest.

The least number of arrivals are during **January**.

CAR PARKING

97% of the bookings require car parking space

Link to appendix

Univariate Analysis – Key Results

MEAL PLAN

77% of the bookings included **Meal Plan 1** (Breakfast).

14% of the bookings did not have any meal plan selected.

MARKET SEGMENT

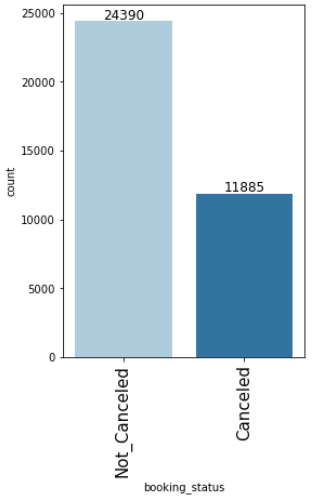
64% of the bookings were done **online** and is the most common market segment for INN Hotels.

29% of bookings were made **offline**.

BOOKING STATUS

67% of total bookings were not canceled.

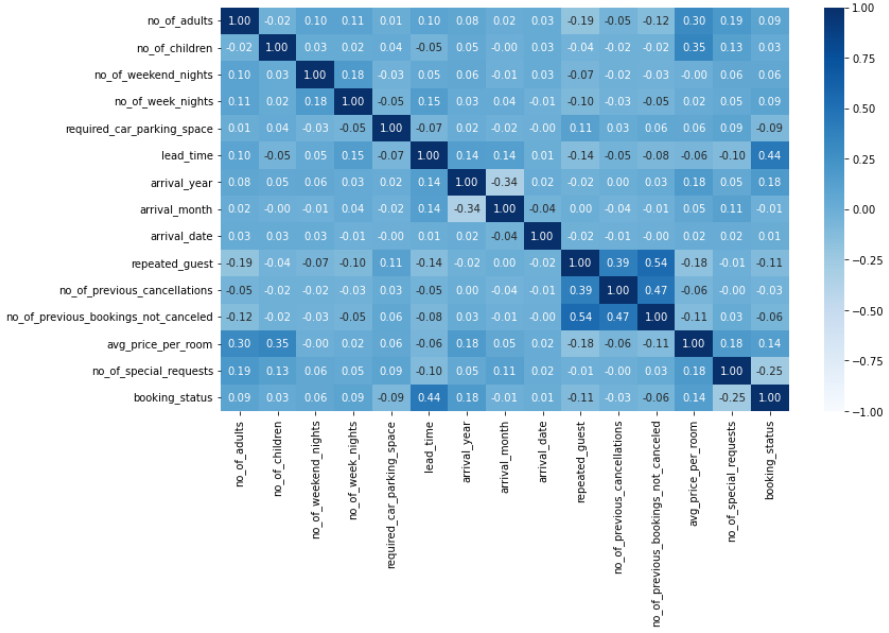
33% were canceled.



Bivariate Analysis – Key Results

Correlation between variables

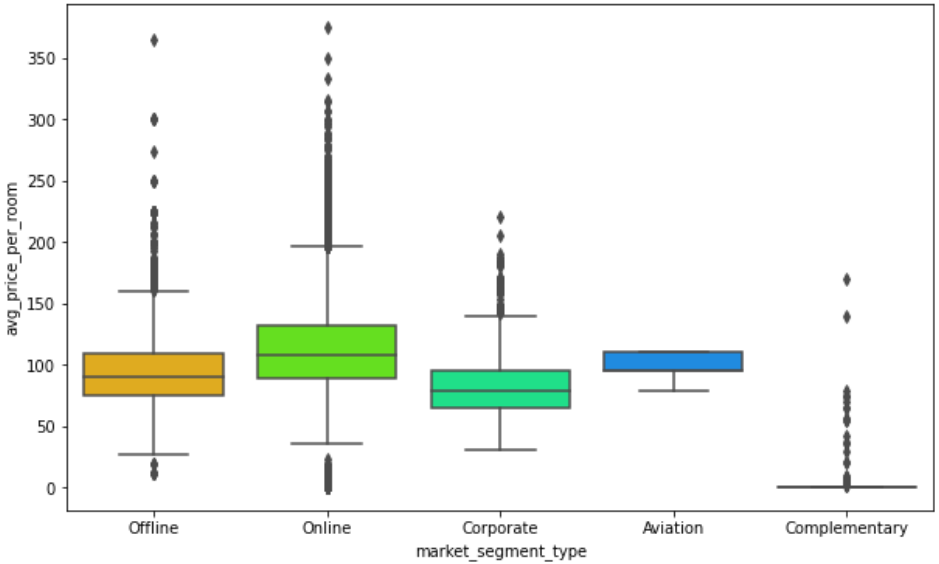
- None of the variables have a high correlation with each other.
- There is a moderate positive correlation between no_of_previous_bookings_not_canceled and repeated_guest



Bivariate Analysis – Key Results

Prices vs market segments

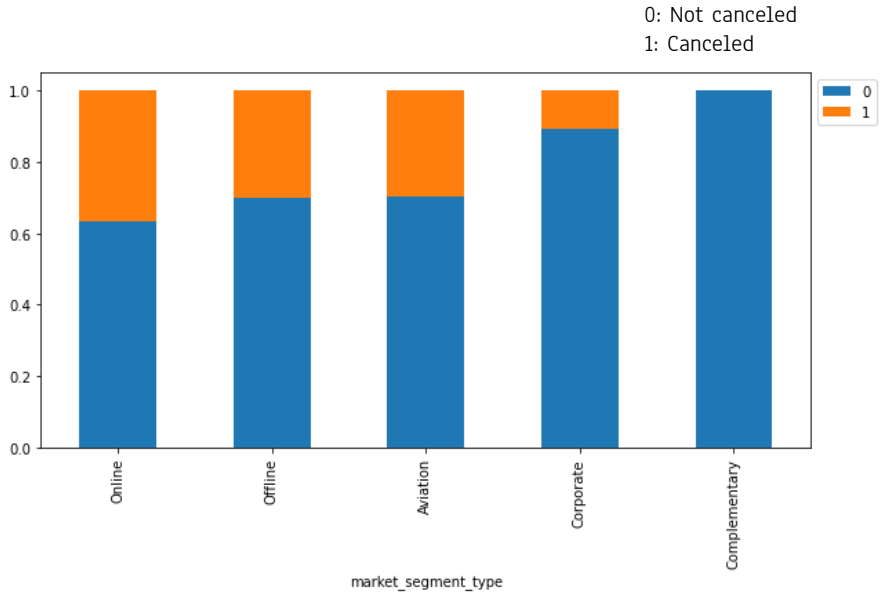
- The prices are higher for online market segments.
- The prices are lower for complementary bookings.
- 75% of the online bookings have prices ranging approx. between 30 - 130 euros.
- The median price of all market segments except for complementary is around 100 euros.
- The prices have some variations between different market segments.



Bivariate Analysis – Key Results

Booking status vs market segments

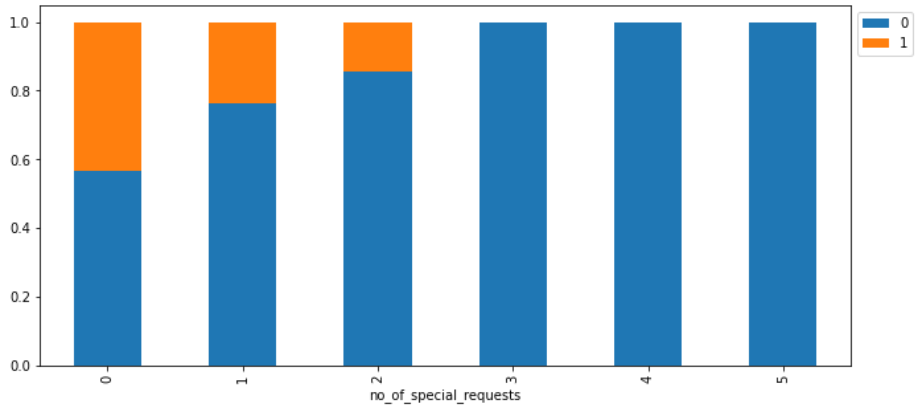
- None of the complementary bookings have been canceled.
- The highest cancellations are for online bookings.
- Aviation and offline have an equal number of cancellations



Bivariate Analysis – Key Results

Special requests vs market segments

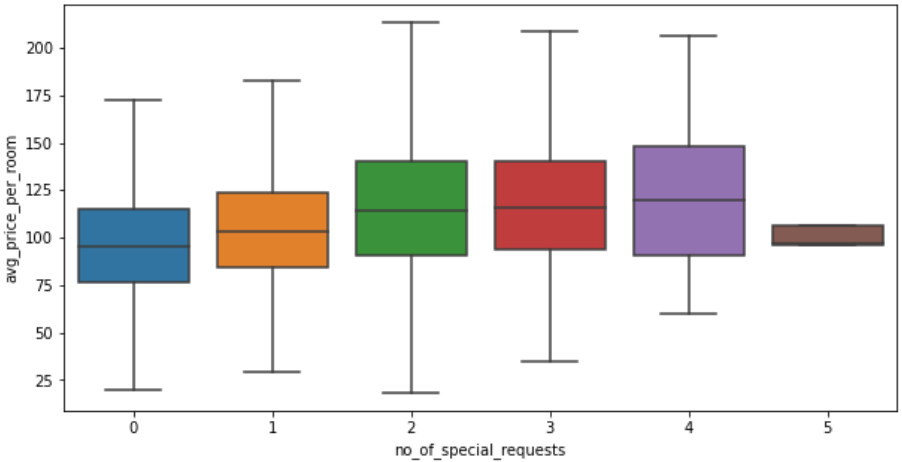
- Bookings having more special requests have lesser or zero cancelations.
- Bookings with 3, 4, and 5 special requests have no cancellations.
- Bookings with 0 special requests have the highest number of cancellations.



Bivariate Analysis – Key Results

Special requests vs Prices

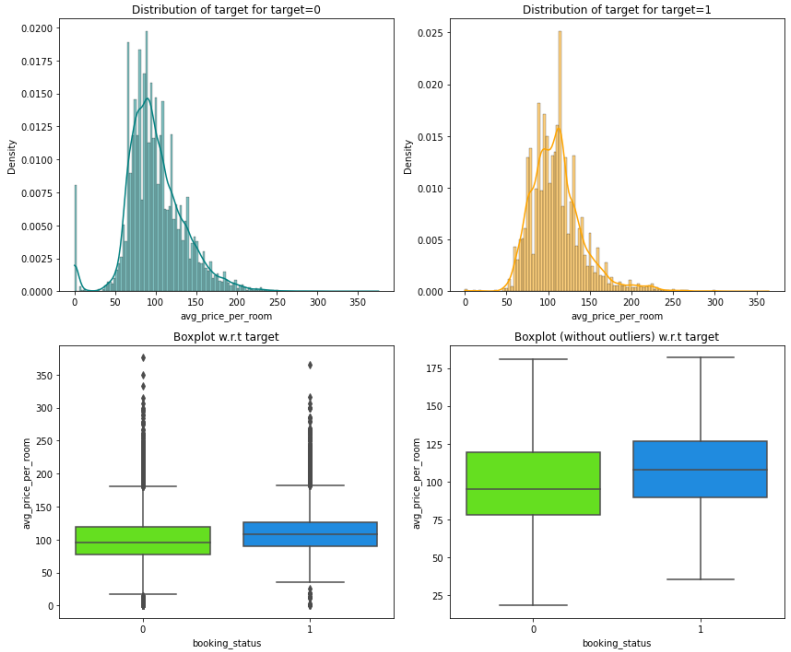
- Price per room is lower for bookings with lesser special requests.
- The price range shows a gradual increase when moving from 0-4 special requests.
- However, the median price is lowest for bookings with 5 special requests.
- The range of avg price per room is higher for bookings with 2 special requests.



Bivariate Analysis – Key Results

Booking status vs Prices

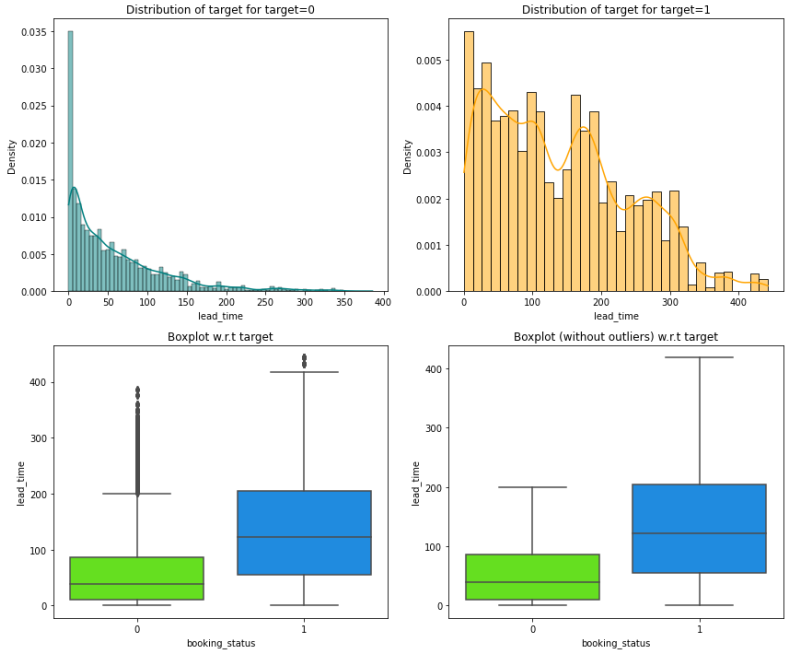
- The prices tend to be higher for bookings that were canceled



Bivariate Analysis – Key Results

Booking status vs lead time

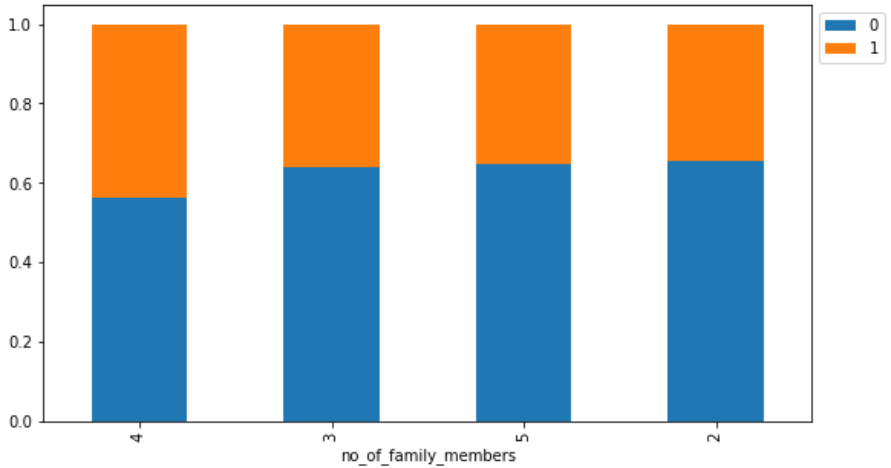
- Lead time is higher for bookings that were canceled.



Bivariate Analysis – Key Results

Booking status vs Family members

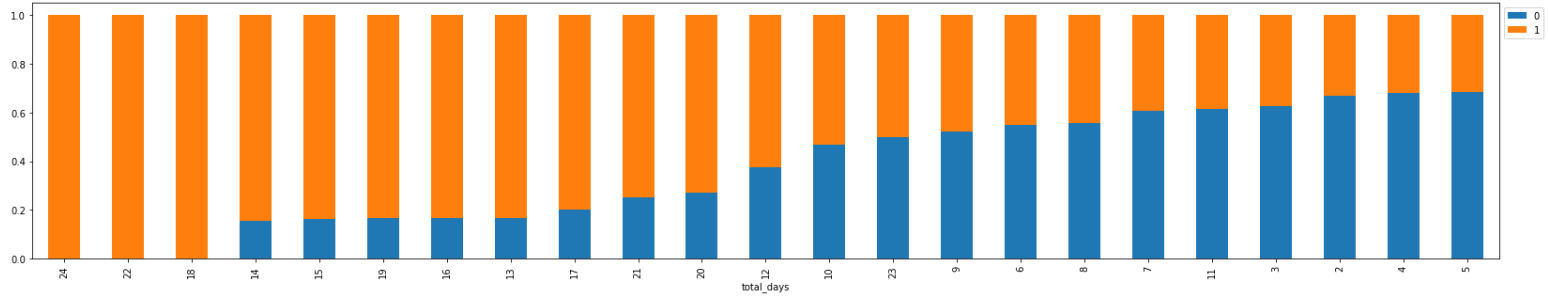
- There is higher number of cancellations for bookings with 4 member families.
- Atleast 30%-40% of the bookings have been canceled from families with 2 or more members.



Bivariate Analysis – Key Results

Booking status vs Family members

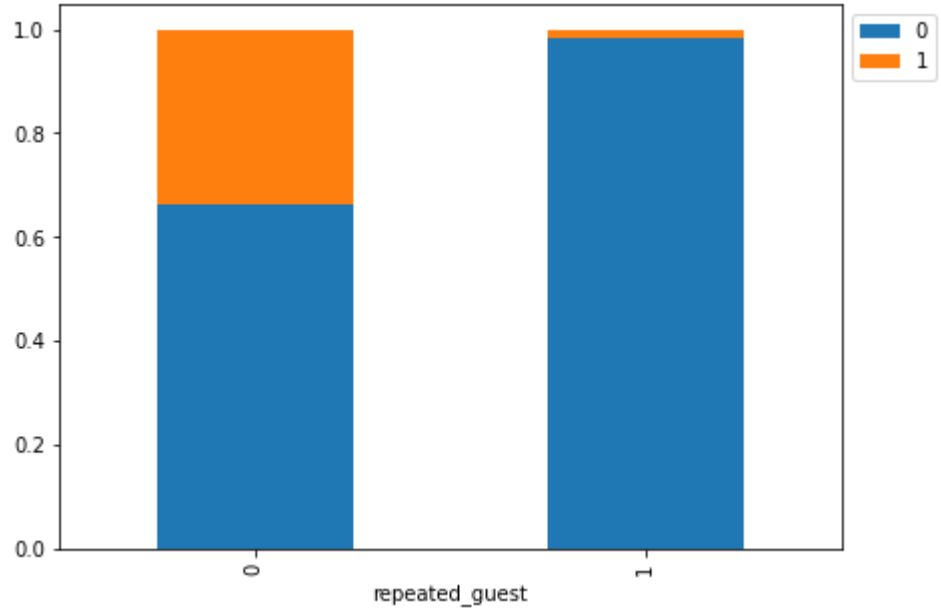
- The more days a customer booked to stay, the higher the chances of the booking being canceled.
- Bookings with more than 20 days of stay have more than 50% chances of being canceled.
- Bookings with 2-7 days of stay have around 30%-40% chances of being canceled.



Bivariate Analysis – Key Results

Booking status vs Repeated guest

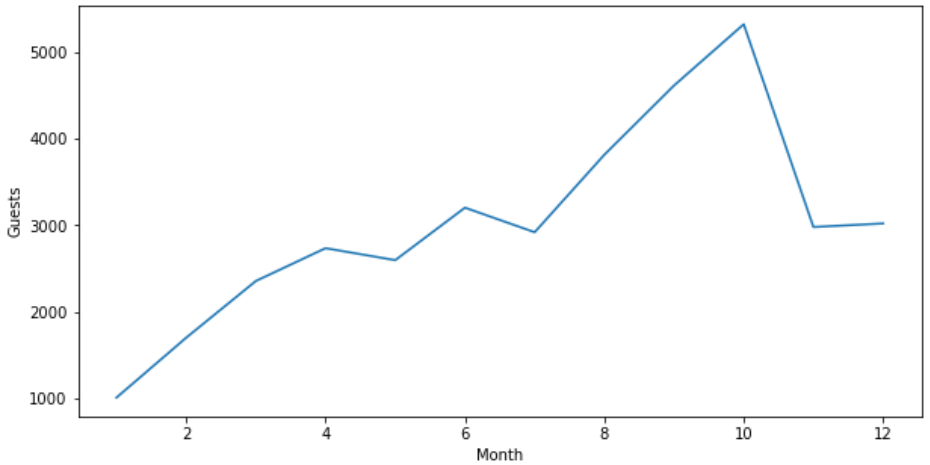
- Bookings from repeating guests are less likely to be canceled.



Bivariate Analysis – Key Results

Guests vs Month

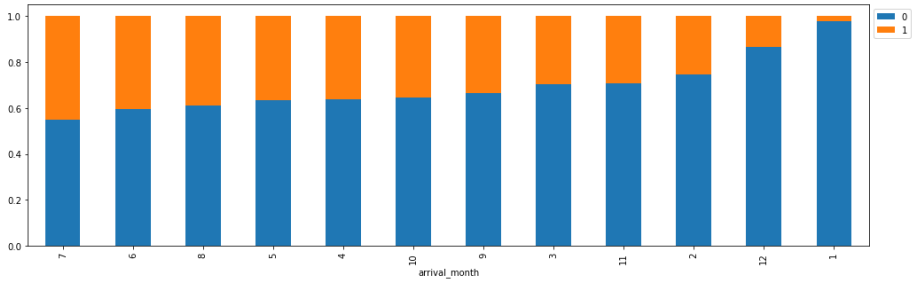
- The hotel gets busy during the month of October.
- There is a gradual increase in the number of guests from August to October.
- The number of guests is the lowest during January.



Bivariate Analysis – Key Results

Booking status vs Month

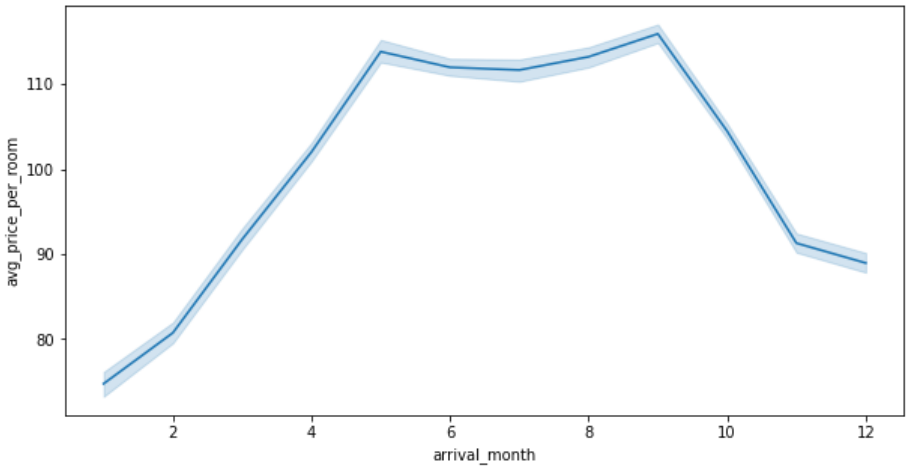
- More number of bookings were canceled during July.
- Lesser bookings were canceled during months of January and December.
- Around 40% of October bookings were canceled.



Bivariate Analysis – Key Results

Prices vs Month

- The prices are highest between May to September.
- The prices are lowest during January.



4

Data preprocessing

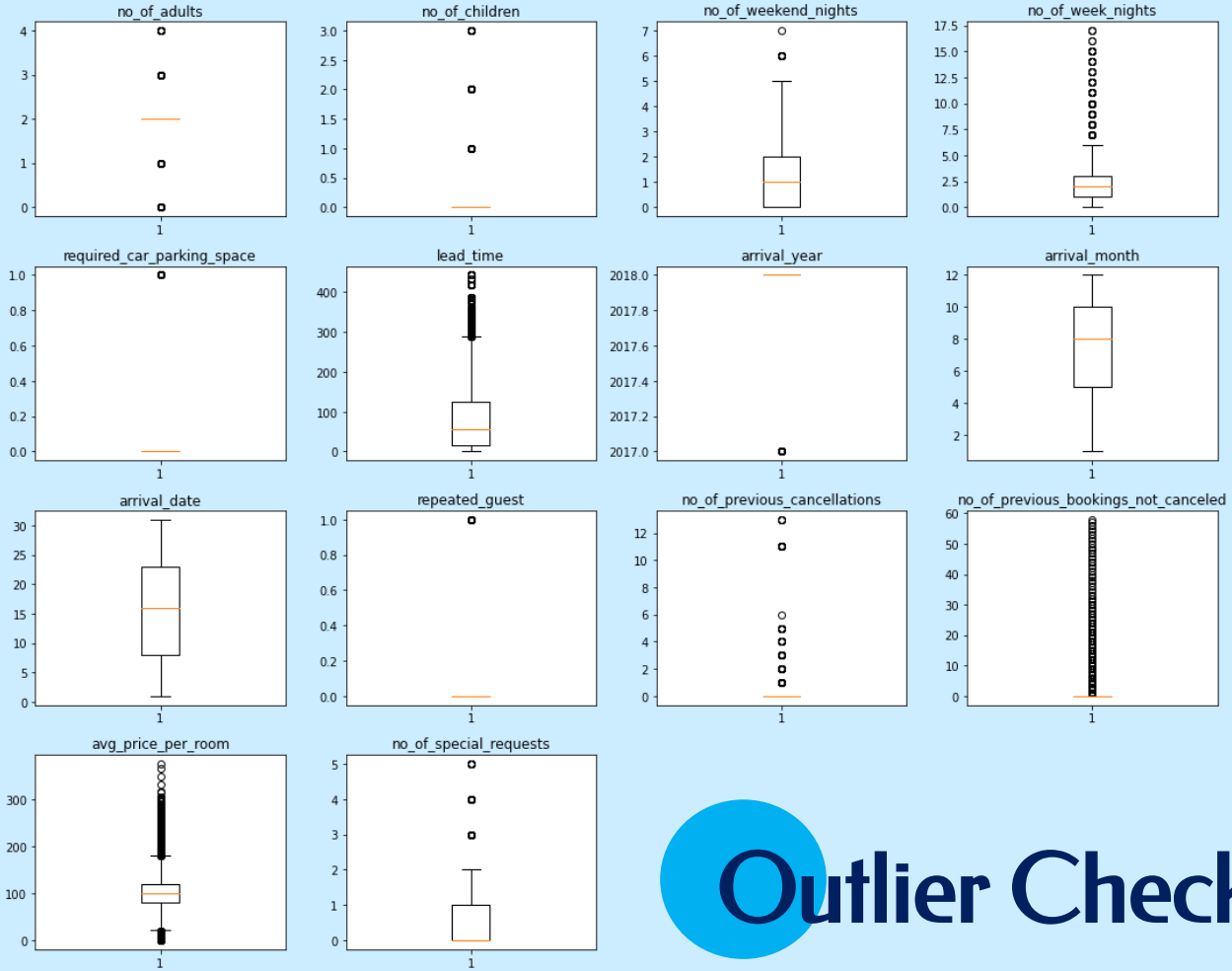
0 MISSING VALUES

```
In [126]:  1 data.isnull().sum()

Out[126]: no_of_adults                        0
          no_of_children                      0
          no_of_weekend_nights                0
          no_of_week_nights                   0
          type_of_meal_plan                   0
          required_car_parking_space          0
          room_type_reserved                  0
          lead_time                           0
          arrival_year                        0
          arrival_month                       0
          arrival_date                        0
          market_segment_type                 0
          repeated_guest                      0
          no_of_previous_cancellations        0
          no_of_previous_bookings_not_canceled 0
          avg_price_per_room                  0
          no_of_special_requests              0
          booking_status                      0
          dtype: int64
```

0 DUPLICATES

```
In [8]:  1 # checking for duplicate values
         2 data.duplicated().sum() ##  Comp

Out[8]: 0
```

- There are quite a few outliers.
- However, there is no need to treat them since the values are proper.



Outlier Check

Data preparation for modeling

In [62]:

```
X = data.drop(["booking_status"], axis=1)
Y = data["booking_status"]

# adding constant
X = sm.add_constant(X) ## Complete the code to add constant to X

X = pd.get_dummies(X, drop_first=True) ## Complete the code to create dummies for X

# Splitting data in train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.30, random_state=1)
```

In [63]:

```
print("Shape of Training set : ", X_train.shape)
print("Shape of test set : ", X_test.shape)
print("Percentage of classes in training set:")
print(y_train.value_counts(normalize=True))
print("Percentage of classes in test set:")
print(y_test.value_counts(normalize=True))
```

```
Shape of Training set :  (25392, 28)
Shape of test set :  (10883, 28)
Percentage of classes in training set:
0    0.67064
1    0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0    0.67638
1    0.32362
Name: booking_status, dtype: float64
```

Logistic Regression

- It was observed from the EDA results that 67% of the bookings were not canceled and 33% were canceled.
- After splitting the data into train and test, the same percentages have been maintained in both train and test data.

Data preparation for modeling

In [92]:
```
X = data.drop(["booking_status"], axis=1)
Y = data["booking_status"]

X = pd.get_dummies(X, drop_first=True) ## Complete the code to create dummies for X

# Splitting data in train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.30, random_state=1)
```

In [93]:
```
print("Shape of Training set : ", X_train.shape)
print("Shape of test set : ", X_test.shape)
print("Percentage of classes in training set:")
print(y_train.value_counts(normalize=True))
print("Percentage of classes in test set:")
print(y_test.value_counts(normalize=True))
```

```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set:
0    0.67064
1    0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0    0.67638
1    0.32362
Name: booking_status, dtype: float64
```

Decision Tree

- It was observed from the EDA results that 67% of the bookings were not canceled and 33% were canceled.
- After splitting the data into train and test, the same percentages have been maintained in both train and test data.

5

Model Performance Summary

Model Overview – Logistic Regression

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25370
Method:	MLE	Df Model:	21
Date:	Mon, 02 Jan 2023	Pseudo R-squ.:	0.3282
Time:	16:17:21	Log-Likelihood:	-10810.
converged:	True	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-915.6391	120.471	-7.600	0.000	-1151.758	-679.520
no_of_adults	0.1088	0.037	2.914	0.004	0.036	0.182
no_of_children	0.1531	0.062	2.470	0.014	0.032	0.275
no_of_weekend_nights	0.1086	0.020	5.498	0.000	0.070	0.147
no_of_week_nights	0.0417	0.012	3.399	0.001	0.018	0.066
required_car_parking_space	-1.5947	0.138	-11.564	0.000	-1.865	-1.324
lead_time	0.0157	0.000	59.213	0.000	0.015	0.016
arrival_year	0.4523	0.060	7.576	0.000	0.335	0.569
arrival_month	-0.0425	0.006	-6.591	0.000	-0.055	-0.030
repeated_guest	-2.7367	0.557	-4.916	0.000	-3.828	-1.646
no_of_previous_cancellations	0.2288	0.077	2.983	0.003	0.078	0.379
avg_price_per_room	0.0192	0.001	26.336	0.000	0.018	0.021
no_of_special_requests	-1.4698	0.030	-48.884	0.000	-1.529	-1.411
type_of_meal_plan_Meal Plan 2	0.1642	0.067	2.469	0.014	0.034	0.295
type_of_meal_plan_Not Selected	0.2860	0.053	5.406	0.000	0.182	0.390
room_type_reserved_Room_Type 2	-0.3552	0.131	-2.709	0.007	-0.612	-0.098
room_type_reserved_Room_Type 4	-0.2828	0.053	-5.330	0.000	-0.387	-0.179
room_type_reserved_Room_Type 5	-0.7364	0.208	-3.535	0.000	-1.145	-0.328
room_type_reserved_Room_Type 6	-0.9682	0.151	-6.403	0.000	-1.265	-0.672
room_type_reserved_Room_Type 7	-1.4343	0.293	-4.892	0.000	-2.009	-0.860
market_segment_type_Corporate	-0.7913	0.103	-7.692	0.000	-0.993	-0.590
market_segment_type_Offline	-1.7854	0.052	-34.363	0.000	-1.887	-1.684

- Negative values of the coefficient show that the probability of a booking being canceled decreases with the increase of the corresponding attribute value.
- Positive values of the coefficient show that the probability of a booking being canceled increases with the increase of 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.
- Since all variables with p-values > 0.05 have been dropped, the existing variables in the model are all considered significant.

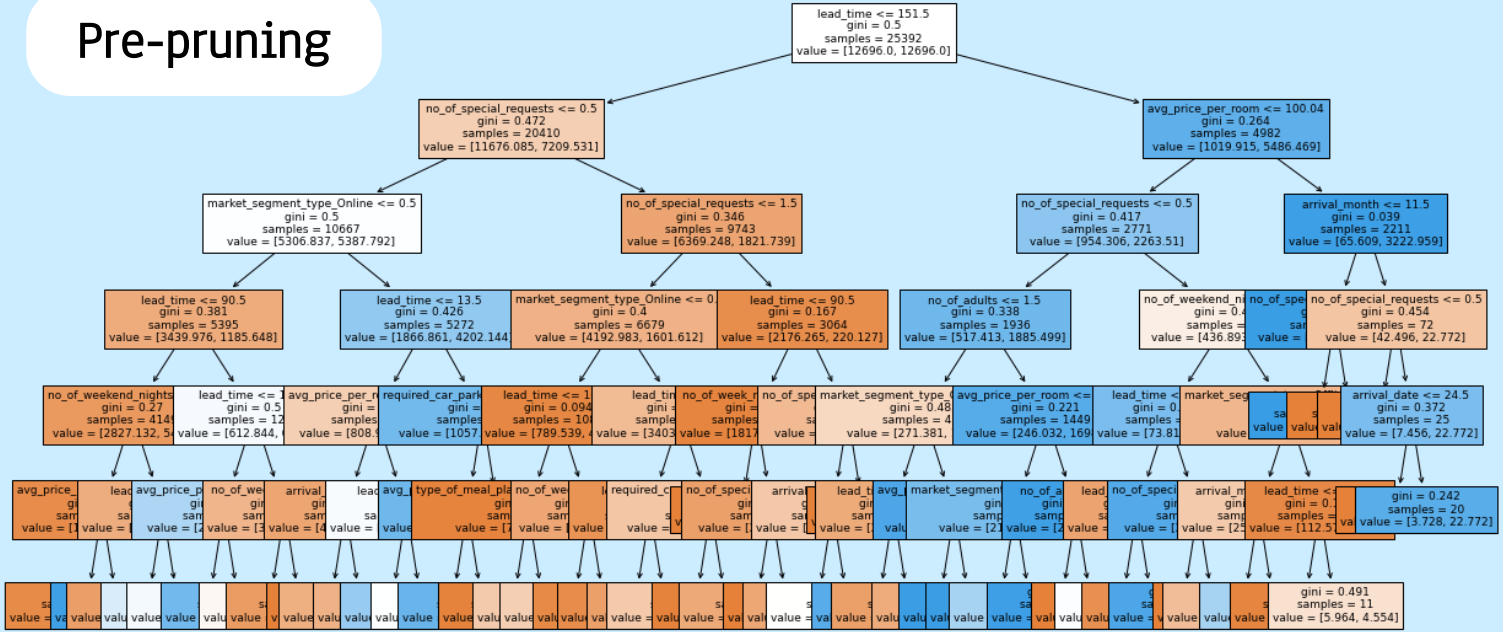
Model Overview – Logistic Regression

- Holding all other features constant, a one unit change in the no. of children will increase the odds of the booking being canceled by ~1.16 times or a ~16.54% increase in the odds of a booking being canceled.
- Holding all other features constant, a one unit change in the lead time will increase the odds of the booking being canceled by ~1.11 times or a ~11.49% increase in the odds of a booking being canceled.
- Holding all other features constant, a one unit change in the repeated guest will decrease the odds of the booking being canceled by ~0.06 times or a ~93.52% decrease in the odds of a booking being canceled.
- The odds of a booking with Room type 2 being cancelled is ~0.70 times less than a booking with Room type 1 or Room type 3, or ~29.9% fewer odds of the booking being cancelled.
- The odds of a booking with Meal plan 2 being cancelled is ~1.18 times more than a booking with Meal plan 1 or Meal plan 3, or ~17.85% more odds of the booking being cancelled.
- The odds of a booking from Corporate market segment being cancelled is ~0.45 times less than a booking from market segments online, complementary, or aviation, or ~54.67% less odds of the booking being cancelled.

	Odds	Change_odd%
const	0.00000	-100.00000
no_of_adults	1.11491	11.49096
no_of_children	1.16546	16.54593
no_of_weekend_nights	1.11470	11.46966
no_of_week_nights	1.04258	4.25841
required_car_parking_space	0.20296	-79.70395
lead_time	1.01583	1.58331
arrival_year	1.57195	57.19508
arrival_month	0.95839	-4.16120
repeated_guest	0.06478	-93.52180
no_of_previous_cancellations	1.25712	25.71181
avg_price_per_room	1.01937	1.93684
no_of_special_requests	0.22996	-77.00374
type_of_meal_plan_Meal Plan 2	1.17846	17.84641
type_of_meal_plan_Not Selected	1.33109	33.10947
room_type_reserved_Room_Type 2	0.70104	-29.89588
room_type_reserved_Room_Type 4	0.75364	-24.63551
room_type_reserved_Room_Type 5	0.47885	-52.11548
room_type_reserved_Room_Type 6	0.37977	-62.02290
room_type_reserved_Room_Type 7	0.23827	-76.17294
market_segment_type_Corporate	0.45326	-54.67373
market_segment_type_Offline	0.16773	-83.22724

Model Overview – Decision Tree

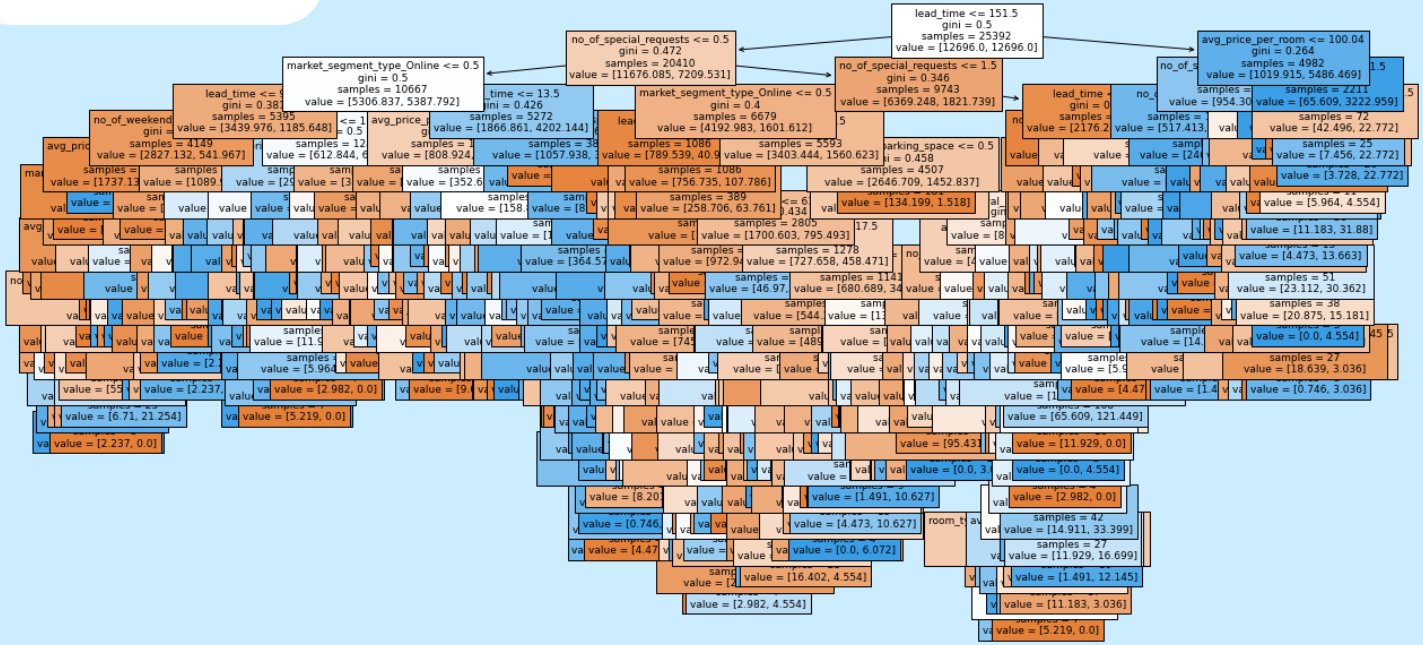
Pre-pruning



If the lead time is less than or equal to 151.50, the no. of special requests is less than or equal to 0.50, the market segment type online is less than or equal to 0.50, lead time is less than or equal to 90.50, no. of weekend nights are less than or equal to 0.50, and average price per room is greater than 196.50, then the booking is most likely to be cancelled.

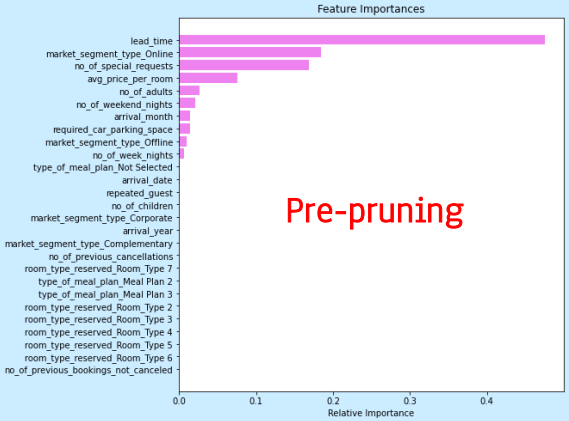
Model Overview – Decision Tree

Post-pruning

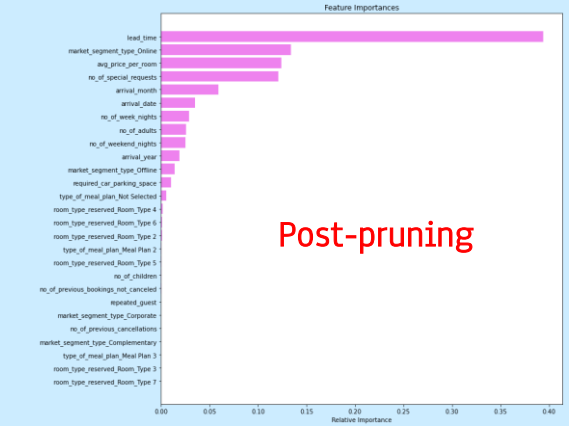


The post-pruned tree also exhibits the same decision tree rules as that of the pre-pruned tree.

Most important features



Pre-pruning



Post-pruning

- Lead time
- Market segment type: Online
- No. of special requests
- Avg price per room

Performance metrics

Model Evaluation Criterion:

Model can make wrong predictions as:

- Predicting a customer will not cancel their booking but in reality, the customer will cancel their booking.
- Predicting a customer will cancel their booking but in reality, the customer will not cancel their booking.

Both the cases are important as:

- If we predict that a booking will not be canceled and the booking gets canceled then the hotel will lose resources and will have to bear additional costs of distribution channels.
- If we predict that a booking will get canceled and the booking doesn't get canceled the hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be canceled. This might damage the brand equity.

Hotel would want F1 Score to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Performance metrics – Logistic Regression

- The model with 0.37 threshold is giving the highest F1 score in both the train set and test set.
- There doesn't seem to be a problem of overfitting in both the cases as well.
- Therefore, the final model will be the one with 0.37 threshold.

Train data

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Test data

Test set performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

Performance metrics – Decision Tree

- Decision Tree with Post-Pruning is giving a higher F1 score for both train and test data.
- Therefore, we will choose post pruned tree as the best model.

Train data

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89954
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81274
F1	0.99117	0.75390	0.85551

Test data

Testing performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86879
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

6

Appendix

Data Background & Contents

Data Dictionary

The data contains the different attributes of customers' booking details.

SL. No.	Variable	Description
1	Booking_ID	Unique identifier of each booking
2	no_of_adults	Number of adults
3	no_of_children	Number of Children
4	no_of_weekend_nights	Number of weekend nights (Saturday/Sunday) the guest stayed or booked to stay at the hotel
5	no_of_week_nights	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
6	type_of_meal_plan	Type of meal plan booked by the customer (Not selected/Meal Plan 1/Meal Plan 2/Meal Plan 3)
7	required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)
8	room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
9	lead_time	Number of days between the date of booking and the arrival date
10	arrival_year	Year of arrival date
11	arrival_month	Month of arrival date
12	arrival_date	Date of the month
13	market_segment_type	Market segment designation.
14	repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)
15	no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
16	no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
17	avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
18	no_of_special_requests	Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
19	booking_status	Flag indicating if the booking was canceled or not.

Data Background & Contents

Data Overview

Shape of dataset

Rows	36275
Columns	19

Datatypes

integer	13
float	1
Object	5

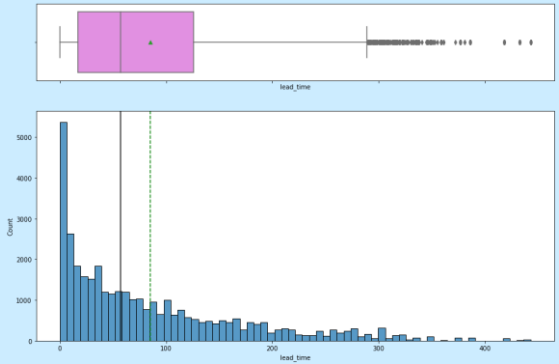
Statistical summary of numerical variables

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	0.10476	0.39466	0.00000	0.00000	0.00000	0.00000	3.00000
no_of_weekend_nights	36275.00000	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
required_car_parking_space	36275.00000	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
lead_time	36275.00000	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
repeated_guest	36275.00000	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	103.41360	35.01675	0.00000	80.30000	99.45000	120.00000	375.50000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000
booking_status	36275.00000	0.32764	0.46936	0.00000	0.00000	0.00000	1.00000	1.00000

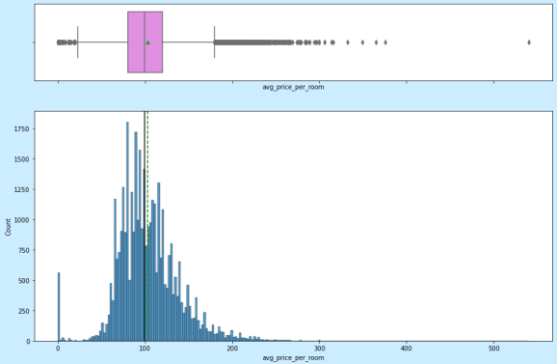
Data Background & Contents

EDA – Univariate Analysis charts

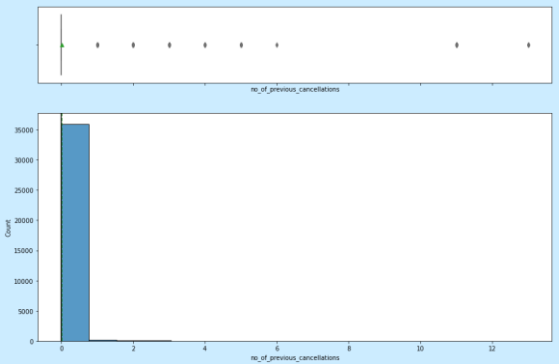
Lead time



Avg price per room



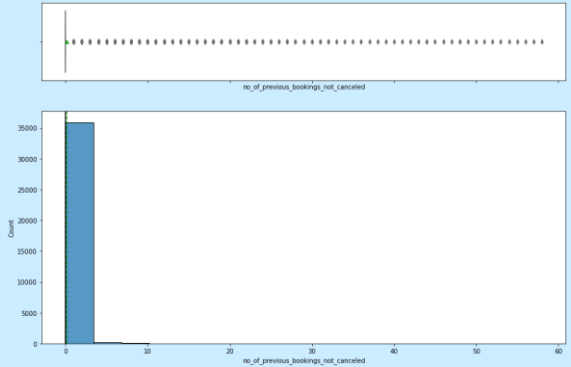
No. of previous cancellations



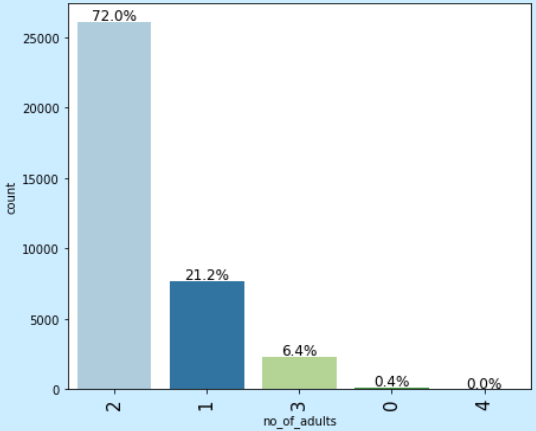
Data Background & Contents

EDA – Univariate Analysis charts

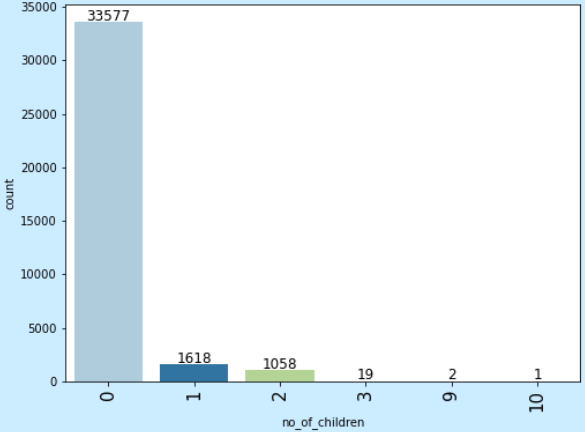
No of previous bookings not canceled



No. of adults



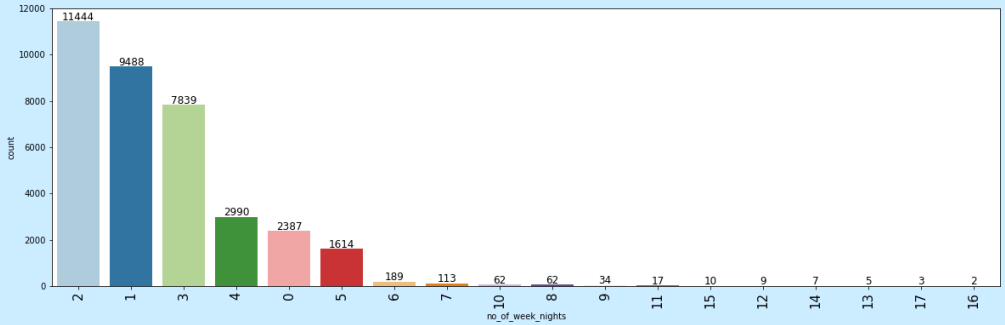
No. of children



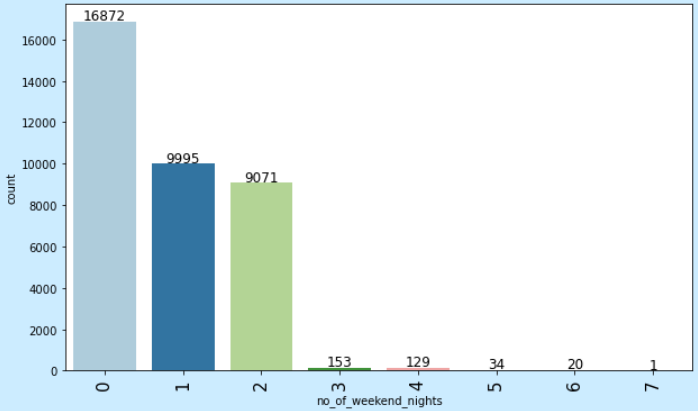
Data Background & Contents

EDA – Univariate Analysis charts

No of week nights



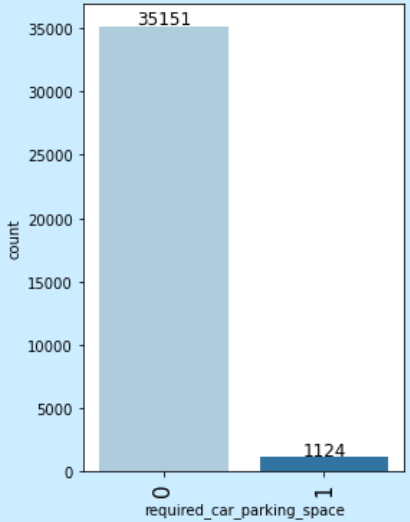
No. of weekend nights



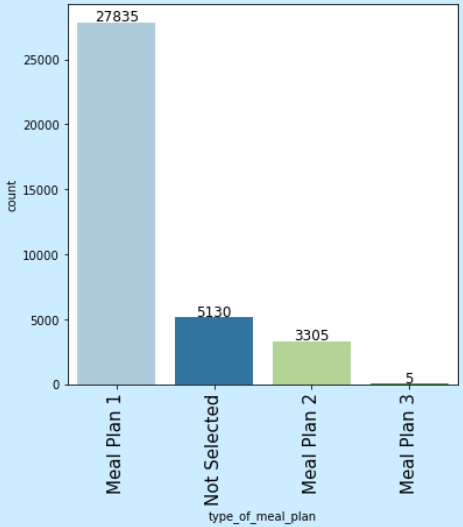
Data Background & Contents

EDA – Univariate Analysis charts

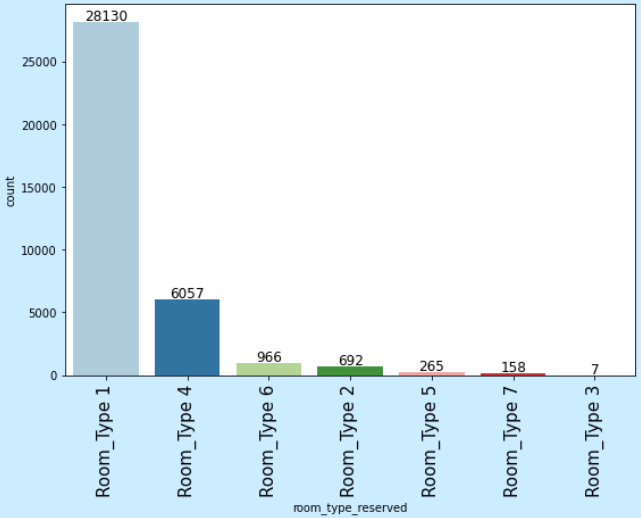
Required car parking space



Type of meal plan



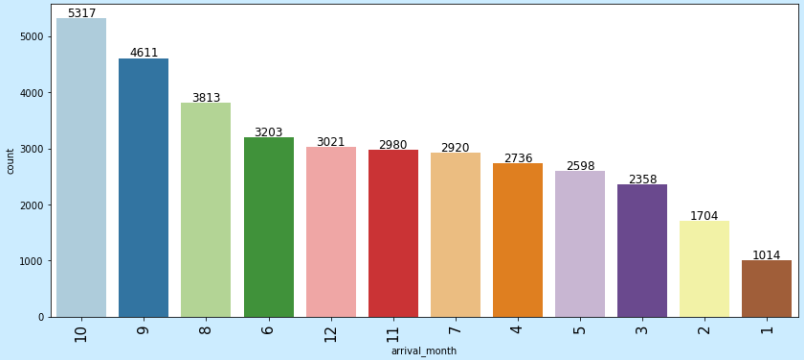
Room type reserved



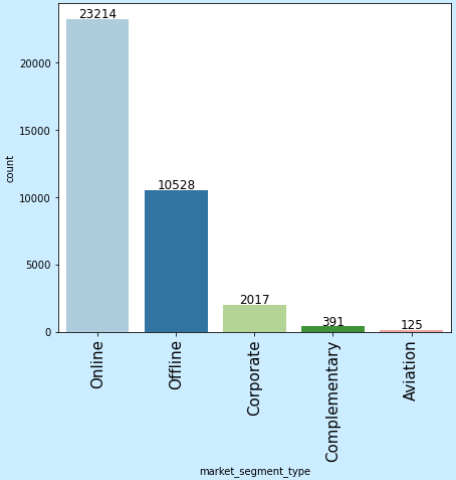
Data Background & Contents

EDA – Univariate Analysis charts

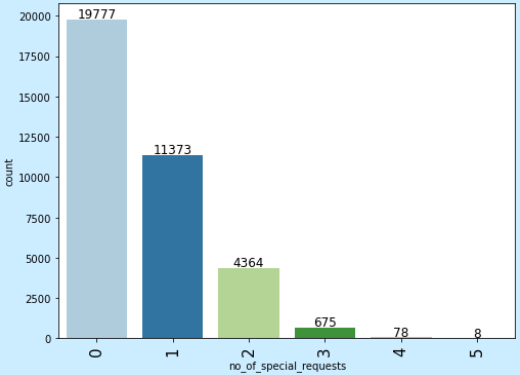
Arrival month



Market segment type



No. of special requests



Model Building – Logistic Regression

Multicollinearity Test

1 Checking VIF

Since none of the variables have high VIFs (> 5), there only exists low multicollinearity between the variables.

	feature	VIF
0	const	39497686.20788
1	no_of_adults	1.35113
2	no_of_children	2.09358
3	no_of_weekend_nights	1.06948
4	no_of_week_nights	1.09571
5	required_car_parking_space	1.03997
6	lead_time	1.39517
7	arrival_year	1.43190
8	arrival_month	1.27633
9	arrival_date	1.00679
10	repeated_guest	1.78358
11	no_of_previous_cancellations	1.39569
12	no_of_previous_bookings_not_canceled	1.65200
13	avg_price_per_room	2.06860
14	no_of_special_requests	1.24798
15	type_of_meal_plan_Meal Plan 2	1.27328
16	type_of_meal_plan_Meal Plan 3	1.02526
17	type_of_meal_plan_Not Selected	1.27306
18	room_type_reserved_Room_Type 2	1.10595
19	room_type_reserved_Room_Type 3	1.00330
20	room_type_reserved_Room_Type 4	1.36361
21	room_type_reserved_Room_Type 5	1.02800
22	room_type_reserved_Room_Type 6	2.05614
23	room_type_reserved_Room_Type 7	1.11816
24	market_segment_type_Complementary	4.50276
25	market_segment_type_Corporate	16.92829
26	market_segment_type_Offline	64.11564
27	market_segment_type_Online	71.18026

2 Checking p-values > 0.05

Following variables were dropped due to p-values > 0.05:

- arrival_date
- no_of_previous_bookings_not_canceled
- type_of_meal_plan_Meal Plan 3
- room_type_reserved_Room_Type_3
- market_segment_type_Complementary
- market_segment_type_Offline

Model Performance Evaluation & Improvement

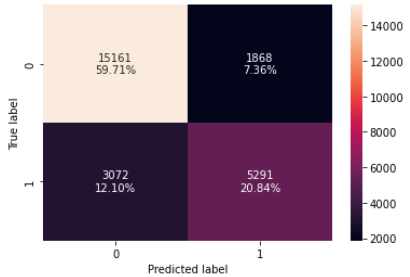
Logistic Regression

1 Checking initial model performance

Default threshold : 0.5

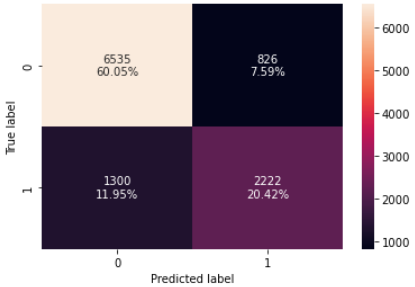
- The model is giving a good f1_score of ~0.681 and ~0.676 on the train and test sets respectively
- As the train and test performances are comparable, the model is not overfitting
- Moving forward we will try to improve the performance of the model

Train set



	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

Test set



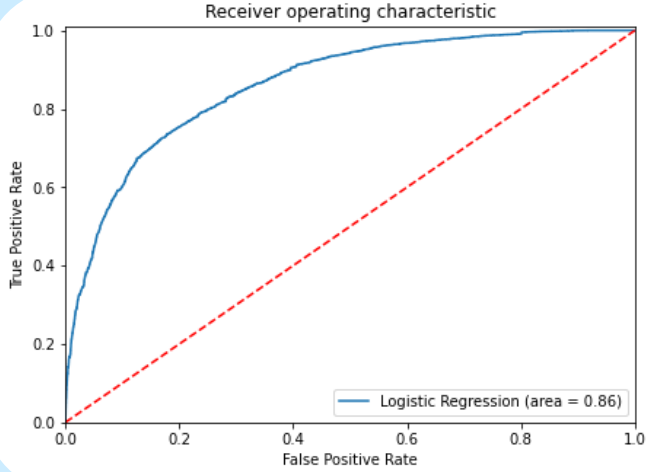
	Accuracy	Recall	Precision	F1
0	0.80465	0.63089	0.72900	0.67641

Model Performance Evaluation & Improvement

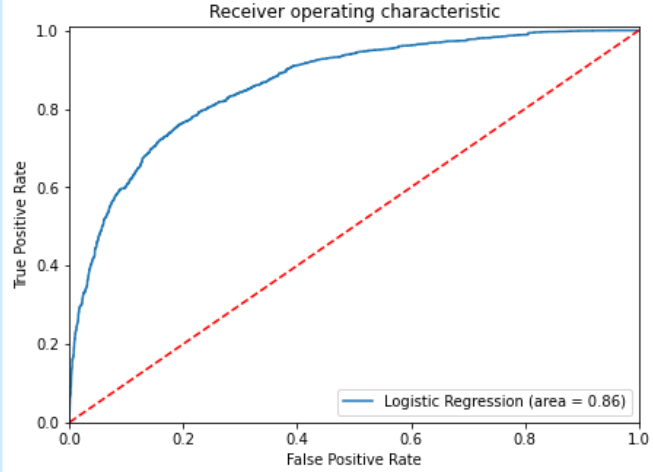
Logistic Regression

2 Optimal threshold using ROC-AUC

Train set



Test set



Optimal Threshold : 0.37

Area under ROC curve is greater than 50% → Good model

Model Performance Evaluation & Improvement

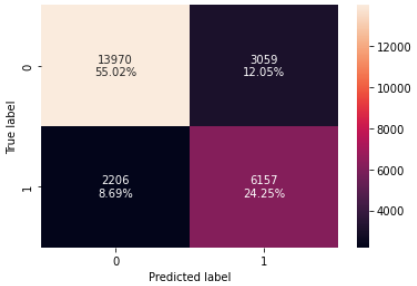
Logistic Regression

3 Checking model performance (ROC-AUC)

Optimal threshold (AUC-ROC) : 0.37

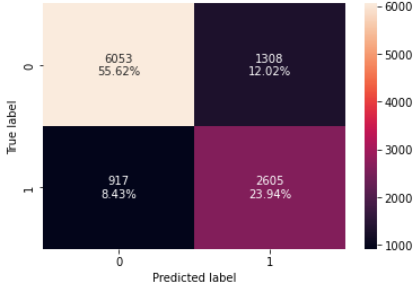
- The F1 score and Recall has increased in both test and train set.
- As the train and test performances are comparable, the model is not overfitting

Train set



	Accuracy	Recall	Precision	F1
0	0.79265	0.73622	0.66808	0.70049

Test set



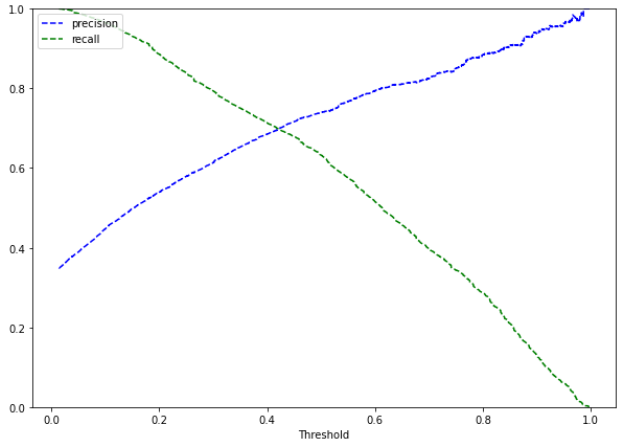
	Accuracy	Recall	Precision	F1
0	0.79555	0.73964	0.66573	0.70074

Model Performance Evaluation & Improvement

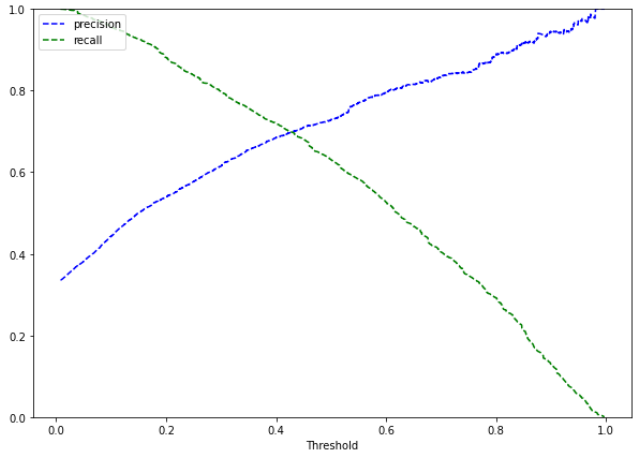
Logistic Regression

4 Optimal threshold using Precision-Recall Curve

Train set



Test set



Optimal Threshold : 0.42

Model Performance Evaluation & Improvement

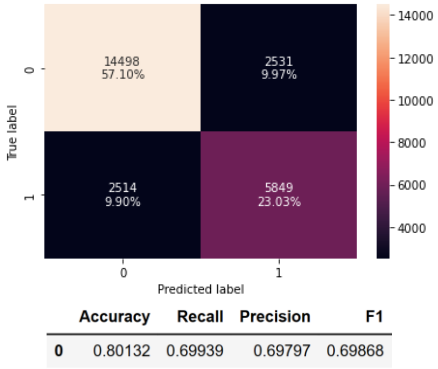
Logistic Regression

5 Checking model performance (Precision-Recall curve)

Optimal threshold (Precision-Recall) : 0.42

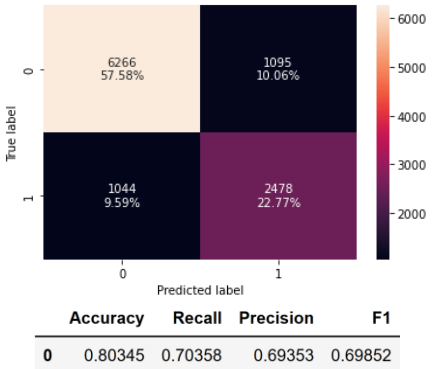
- Model is performing well but there is no improvement in the F1 score.
- As the train and test performances are comparable, the model is not overfitting

Train set



	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

Test set

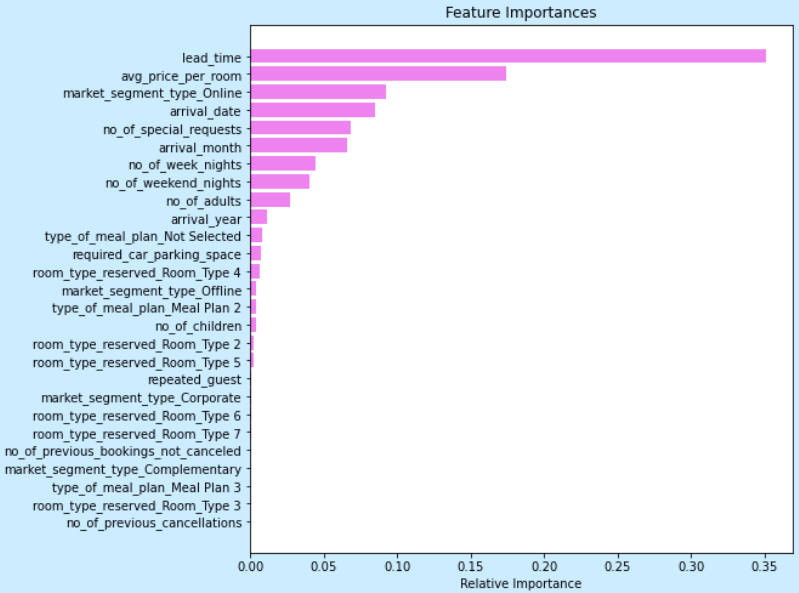


	Accuracy	Recall	Precision	F1
0	0.80345	0.70358	0.69353	0.69852

Model Building – Decision Tree

```
In [96]:
1  model = DecisionTreeClassifier(random_state=1)
2  model.fit(X_train, y_train) ## Complete the coc

Out[96]: DecisionTreeClassifier(random_state=1)
```

Fitting the decision on train data



Important features before pruning

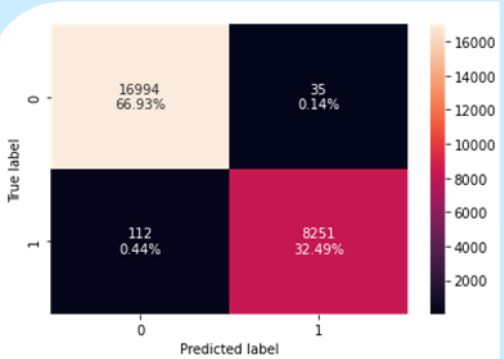
Model Performance Evaluation & Improvement

Decision Tree

1 Checking initial model performance

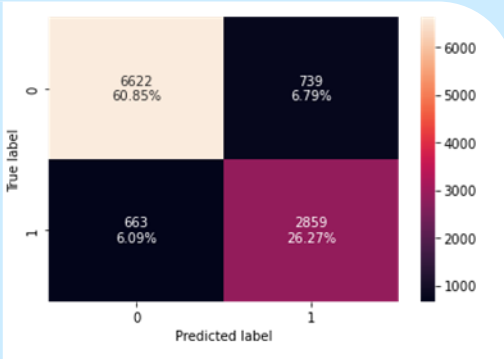
- There is high disparity between the model performance of train and test.
- This indicates signs of overfitting which does not show a good model.

Train set



	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Test set



	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Model Performance Evaluation & Improvement

Decision Tree

PRE-PRUNING

In [102]:

```
# Choose the type of classifier.
estimator = DecisionTreeClassifier(random_state=1, class_weight="balanced")

# Grid of parameters to choose from
parameters = {
    "max_depth": np.arange(2, 7, 2),
    "max_leaf_nodes": [50, 75, 150, 250],
    "min_samples_split": [10, 30, 50, 70],
}

# Type of scoring used to compare parameter combinations
acc_scorer = make_scorer(f1_score)

# Run the grid search
grid_obj = GridSearchCV(estimator, parameters, scoring=acc_scorer, cv=5)
grid_obj = grid_obj.fit(X_train, y_train)

# Set the clf to the best combination of parameters
estimator = grid_obj.best_estimator_

# Fit the best algorithm to the data.
estimator.fit(X_train, y_train)
```

Out[102]:

```
DecisionTreeClassifier(class_weight='balanced', max_depth=6, max_leaf_nodes=50,
                       min_samples_split=10, random_state=1)
```

Best combination of parameters

max_depth : 6

max_leaf_nodes : 50

min_samples_split : 10

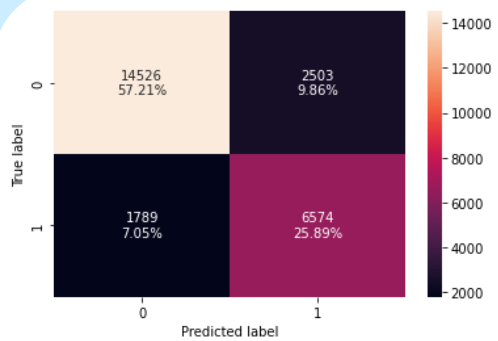
Model Performance Evaluation & Improvement

Decision Tree

2 Pre-pruning model performance

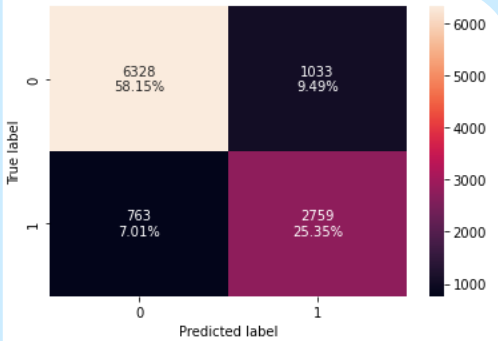
- There is an improvement since the train and test set are comparable.
- The F1 score is coming around 0.75 for both train and test data.
- Hence, there is no overfitting.

Train set



	Accuracy	Recall	Precision	F1
0	0.83097	0.78608	0.72425	0.75390

Test set

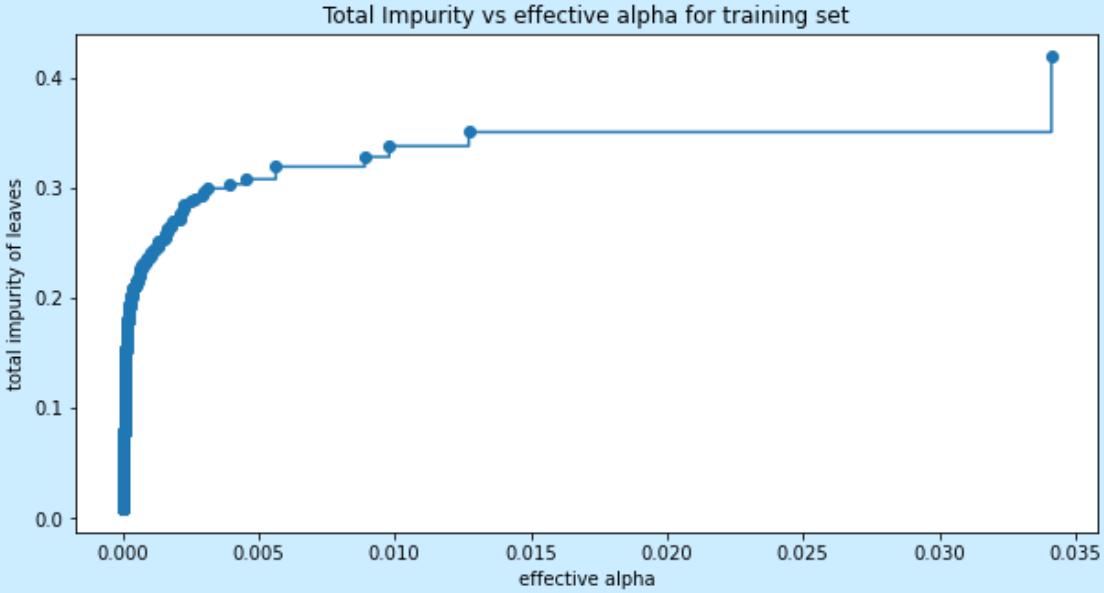


	Accuracy	Recall	Precision	F1
0	0.83497	0.78336	0.72758	0.75444

Model Performance Evaluation & Improvement

Decision Tree

POST-PRUNING

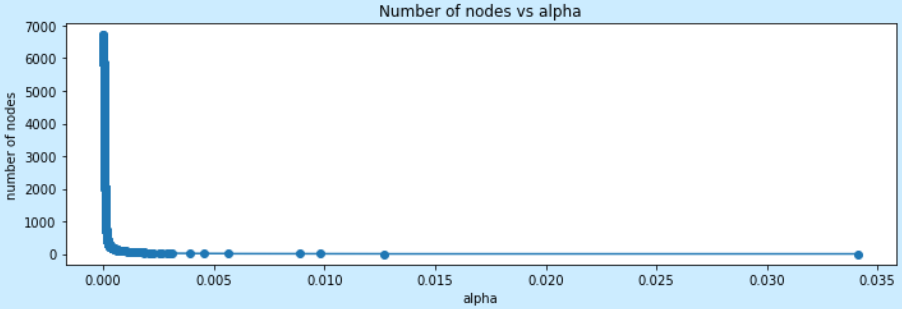


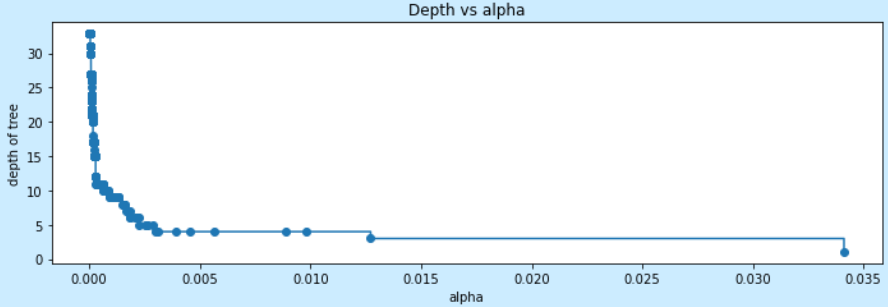
- The total impurity increases as the alpha increases.

Model Performance Evaluation & Improvement

Decision Tree

POST-PRUNING





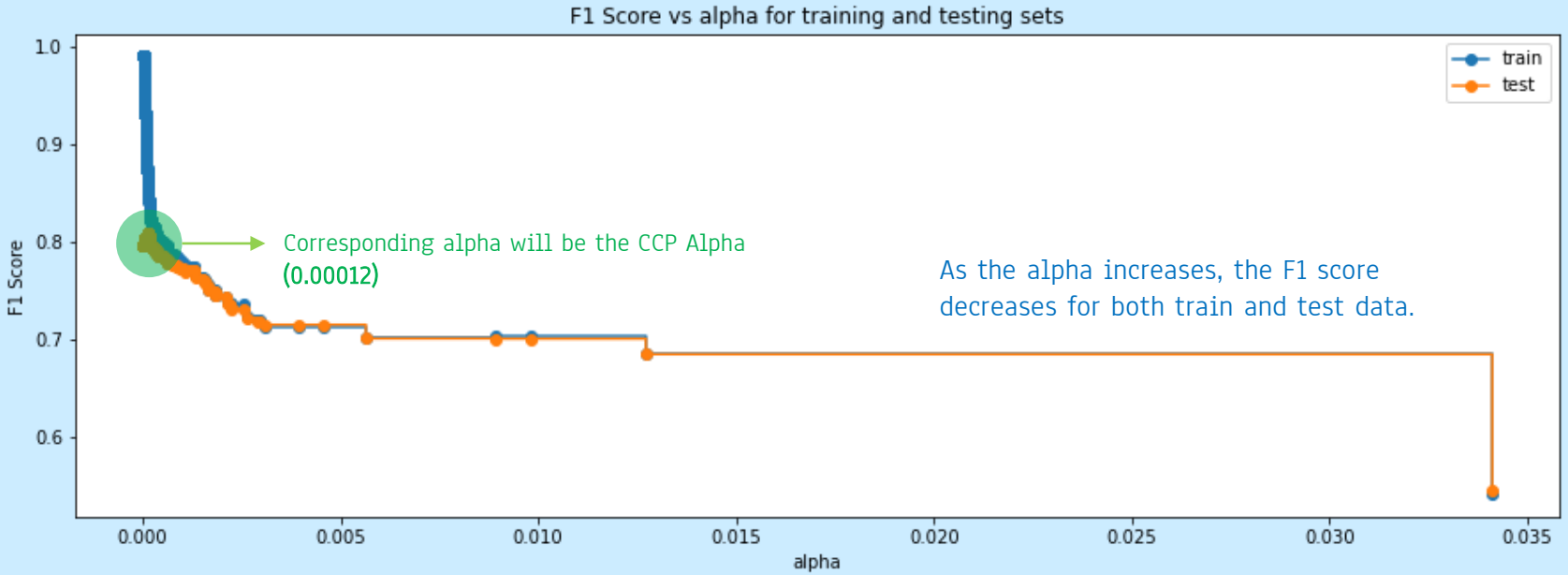
- The number of nodes decrease as alpha increases.

- The depth of the tree decreases as alpha increases

Model Performance Evaluation & Improvement

Decision Tree

POST-PRUNING



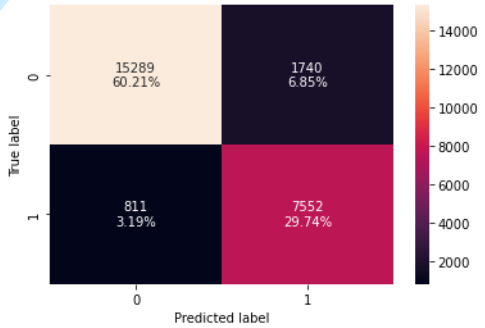
Model Performance Evaluation & Improvement

Decision Tree

3 Post-pruning model performance

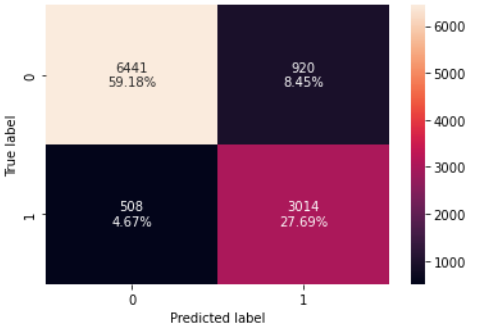
- There is improvement in the F1 score for both train and test set.

Train set



	Accuracy	Recall	Precision	F1
0	0.89954	0.90303	0.81274	0.85551

Test set



	Accuracy	Recall	Precision	F1
0	0.86879	0.85576	0.76614	0.80848