Predicting Chances of booking cancellation using Logistics Regression

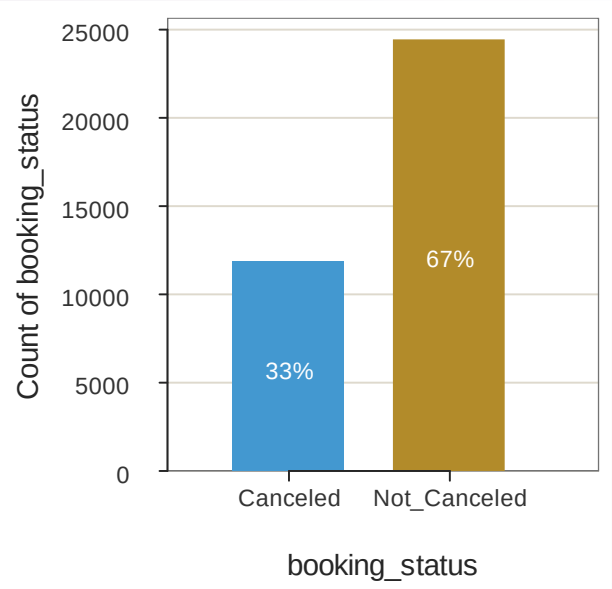
By-

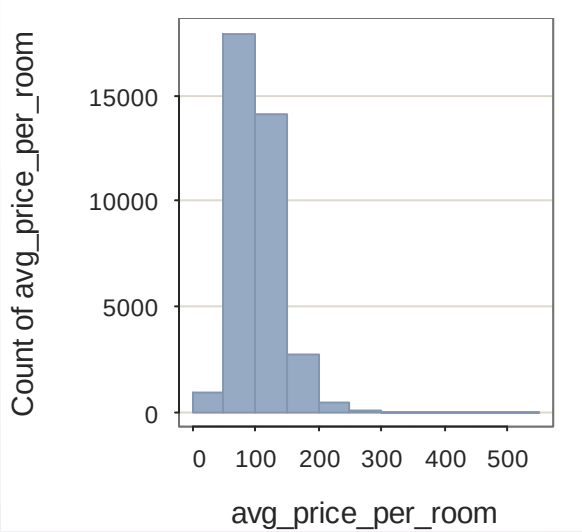
Deepshi Sachan

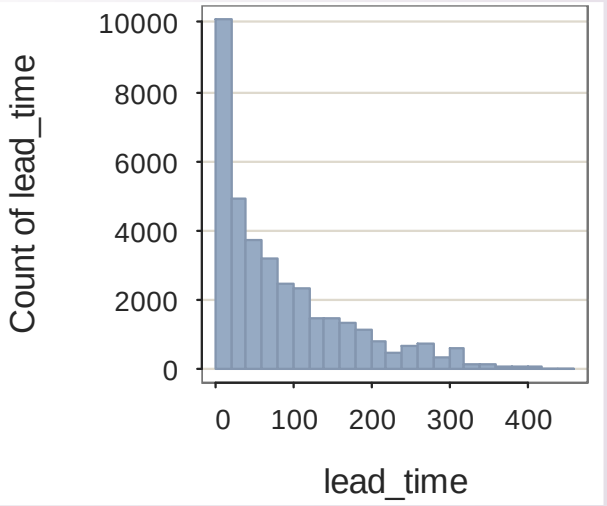
ID: 4213188

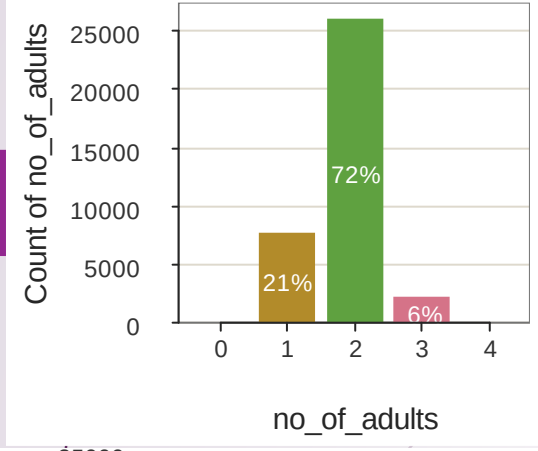
ABOUT DATASETS

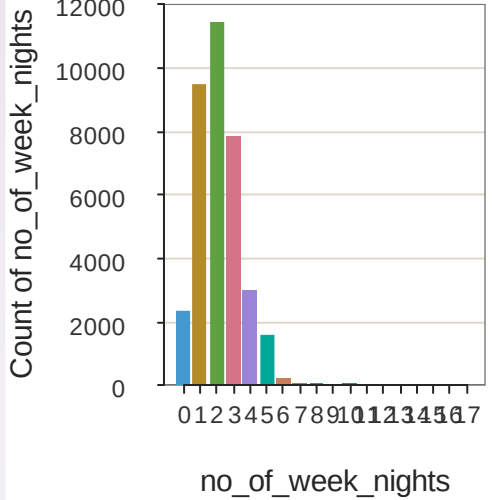
- Source of this dataset is Kaggle.com website. (Hotel_reservations.csv)
- This dataset contains 36725 observations and 19 variables.
- Contains different attributes of hotel reservations like type of room reserved by the customer, Number of weekend nights, type of meal plan booked, etc.
- Used Logistic Regression model to forecast the chances of booking cancellation.
- Transformed Categorical variable (booking status) to Numerical. (0- Not Cancelled, 1- Cancelled)

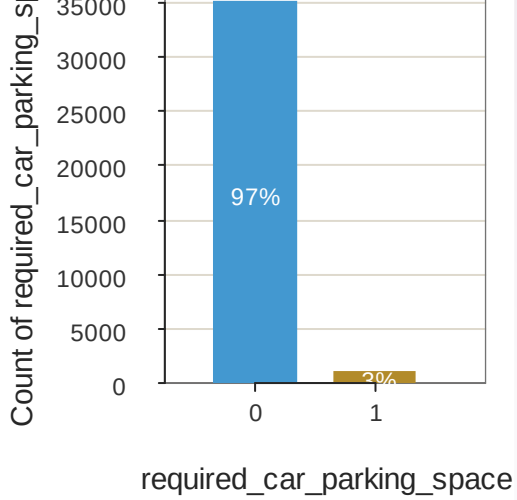


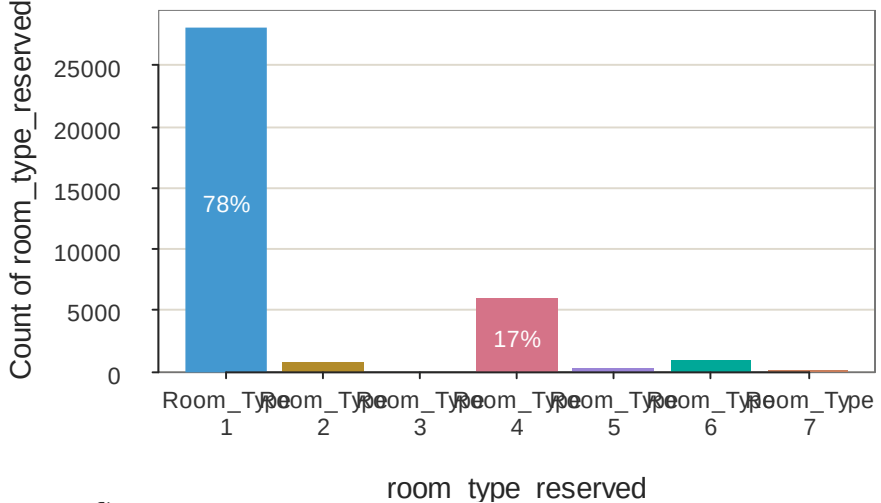


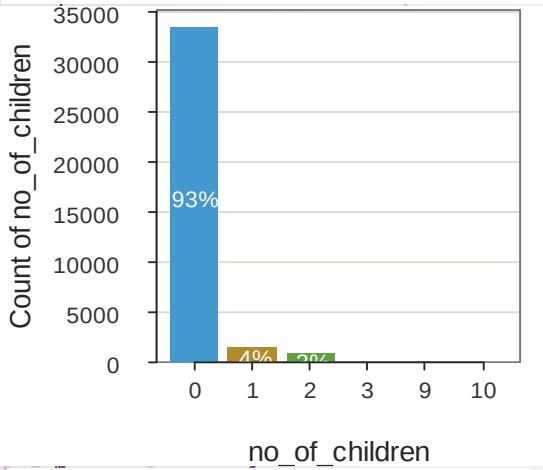


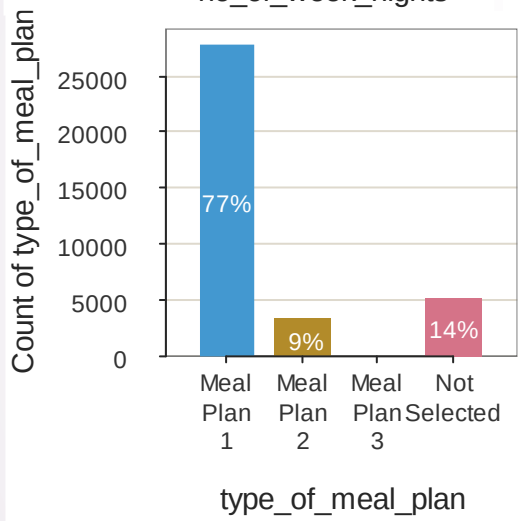

Count of no_of_adults
25000
20000
15000
10000
5000
0
72%
21%
6%
0
1
2
3
4
no_of_adults

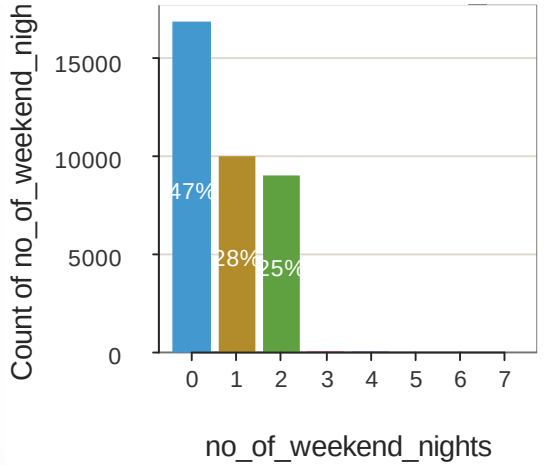

Count of no_of_week_nights
12000
10000
8000
6000
4000
2000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
no_of_week_nights

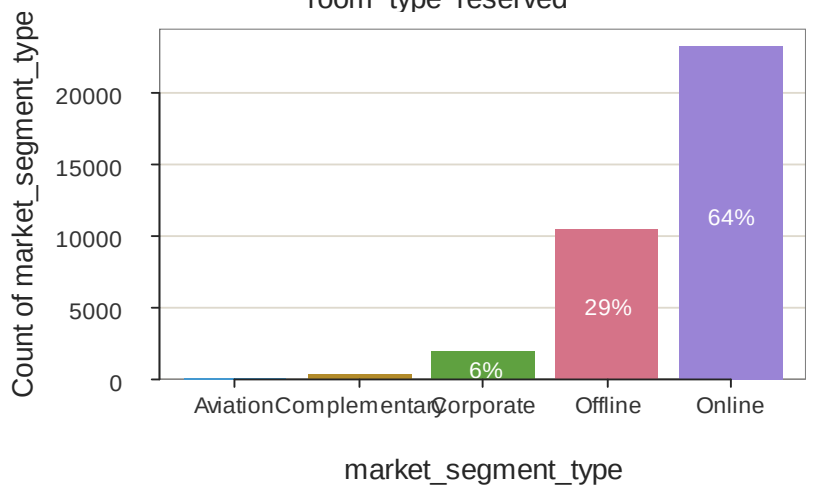

Count of required_car_parking_s
35000
30000
25000
20000
15000
10000
5000
0
97%
3%
0
1
required_car_parking_space

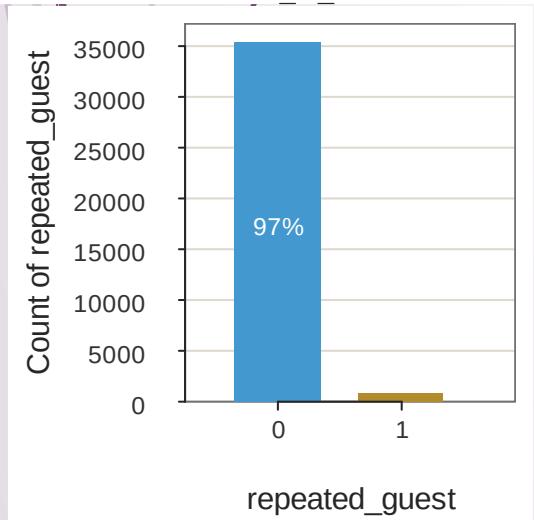

Count of room_type_reserved
25000
20000
15000
10000
5000
0
78%
17%
Room_Type 1
Room_Type 2
Room_Type 3
Room_Type 4
Room_Type 5
Room_Type 6
Room_Type 7
room_type_reserved

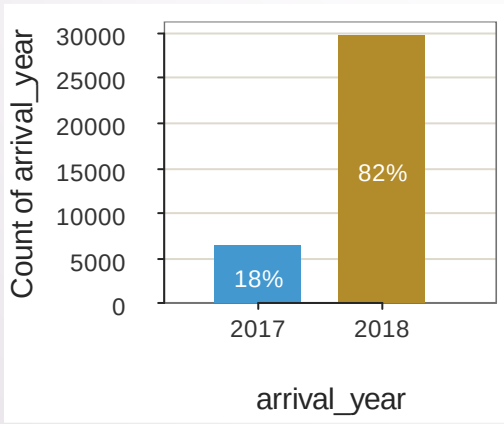

Count of no_of_children
35000
30000
25000
20000
15000
10000
5000
0
93%
4%
2%
0
1
2
3
9
10
no_of_children

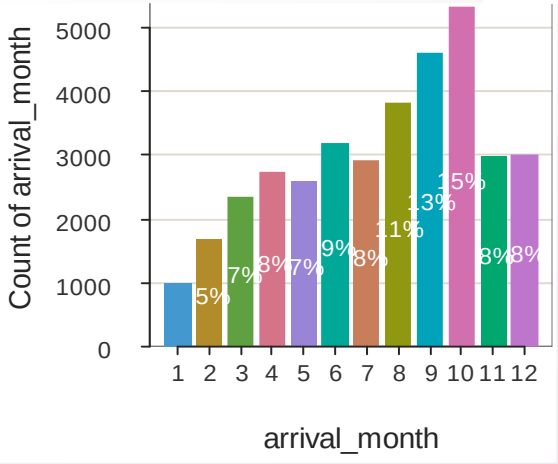

Count of type_of_meal_plan
25000
20000
15000
10000
5000
0
77%
9%
14%
Meal Plan 1
Meal Plan 2
Meal Plan 3
Not Selected
type_of_meal_plan

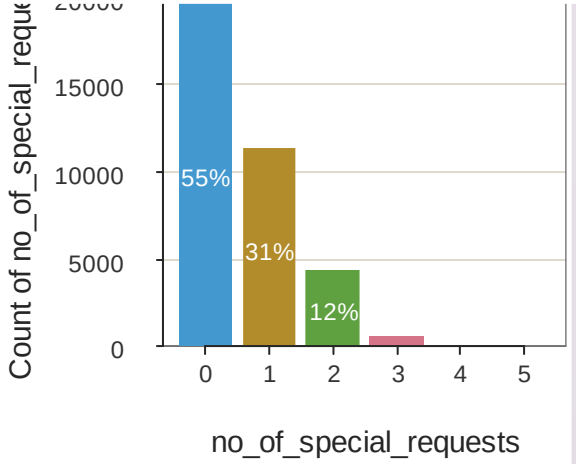

Count of no_of_weekend_nigh
15000
10000
5000
0
47%
28%
25%
0
1
2
3
4
5
6
7
no_of_weekend_nights


Count of market_segment_type
20000
15000
10000
5000
0
64%
29%
6%
Aviation
Complementary
Corporate
Offline
Online
market_segment_type


Count of repeated_guest
35000
30000
25000
20000
15000
10000
5000
0
97%
0
1
repeated_guest

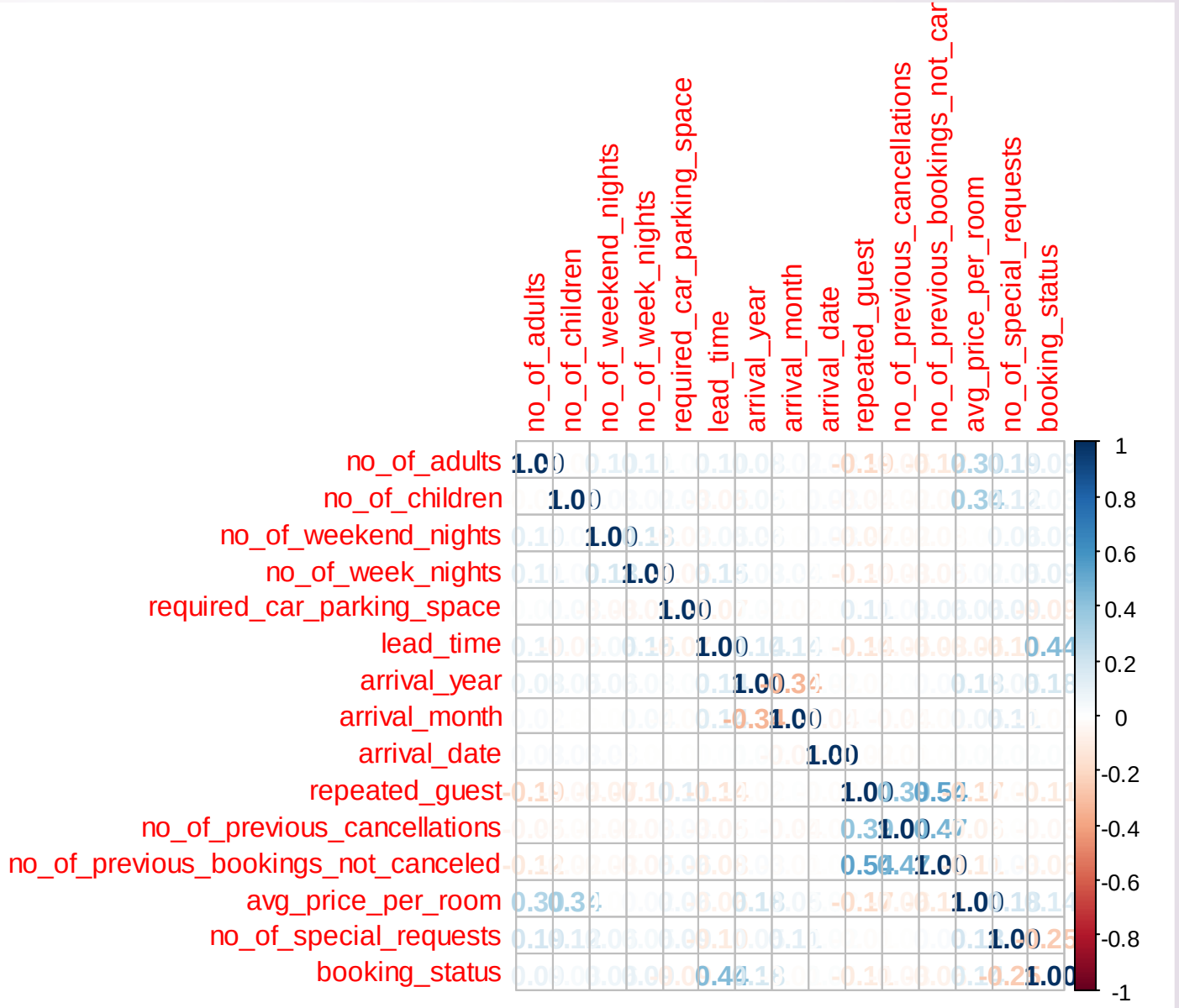

Count of arrival_year
30000
25000
20000
15000
10000
5000
0
82%
18%
2017
2018
arrival_year


Count of arrival_month
5000
4000
3000
2000
1000
0
5%
7%
8%
7%
9%
8%
11%
13%
15%
8%
8%
1 2 3 4 5 6 7 8 9 10 11 12
arrival_month


Count of no_of_special_requ
15000
10000
5000
0
55%
31%
12%
0
1
2
3
4
5
no_of_special_requests

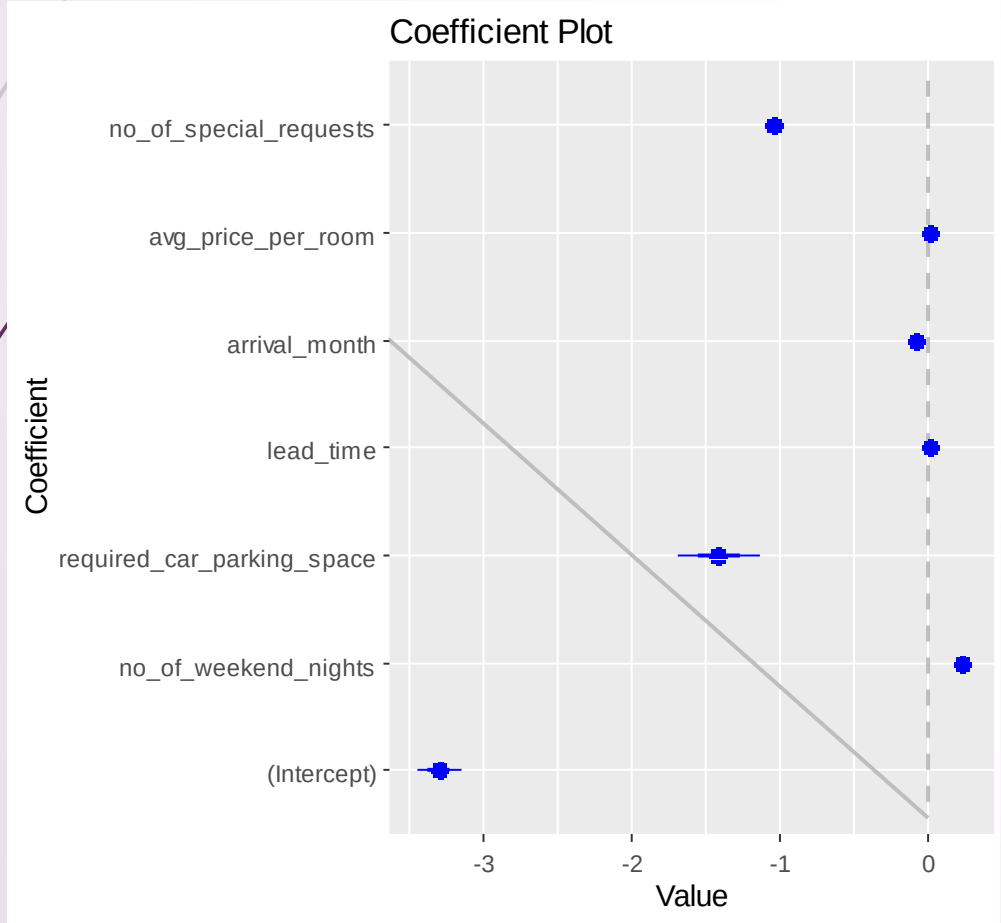
Verifying the correlation between our variables.

- No Strong Correlation



Logistic Regression Model

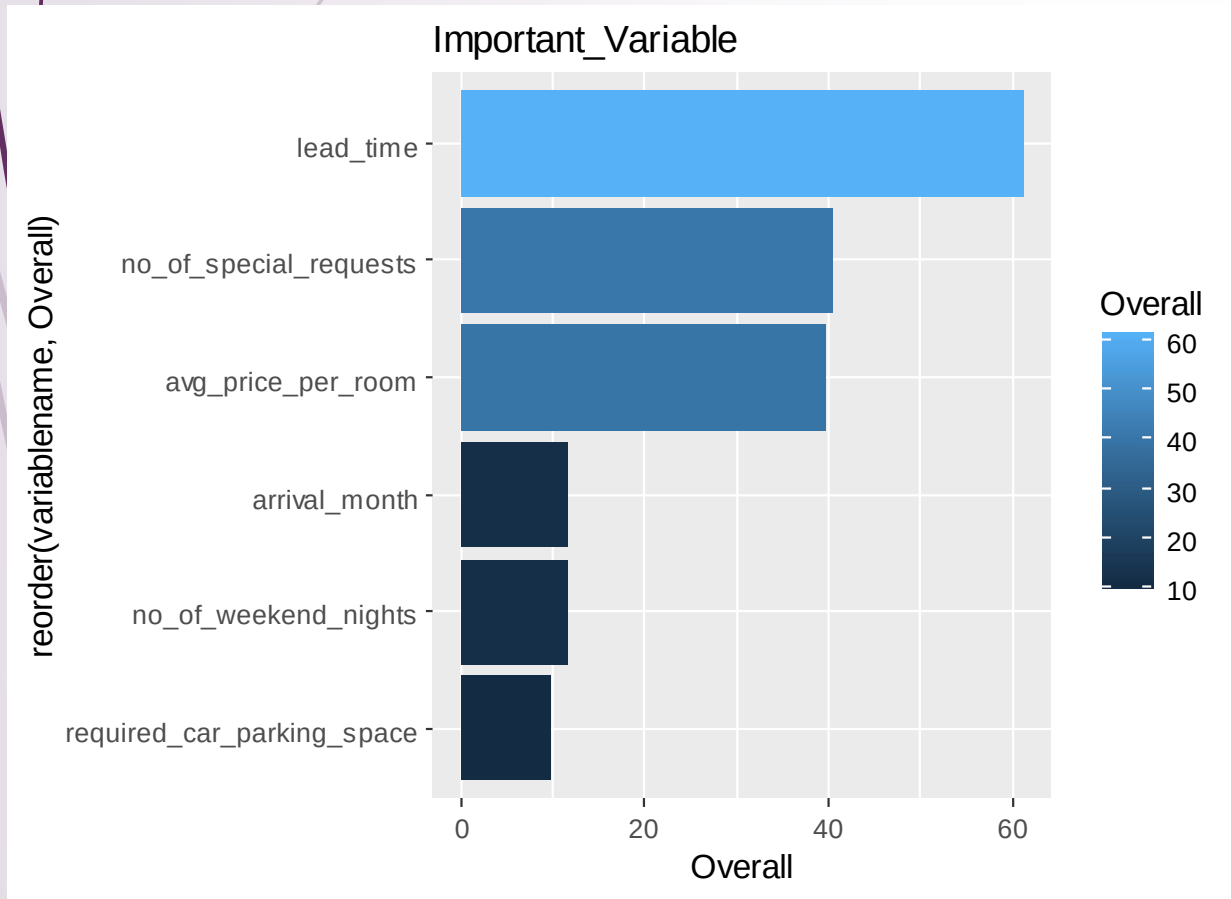
- Ratio of training to test dataset is 7:3.
- Model trained on the train dataset.
- Model tested on test dataset and results are shown below.

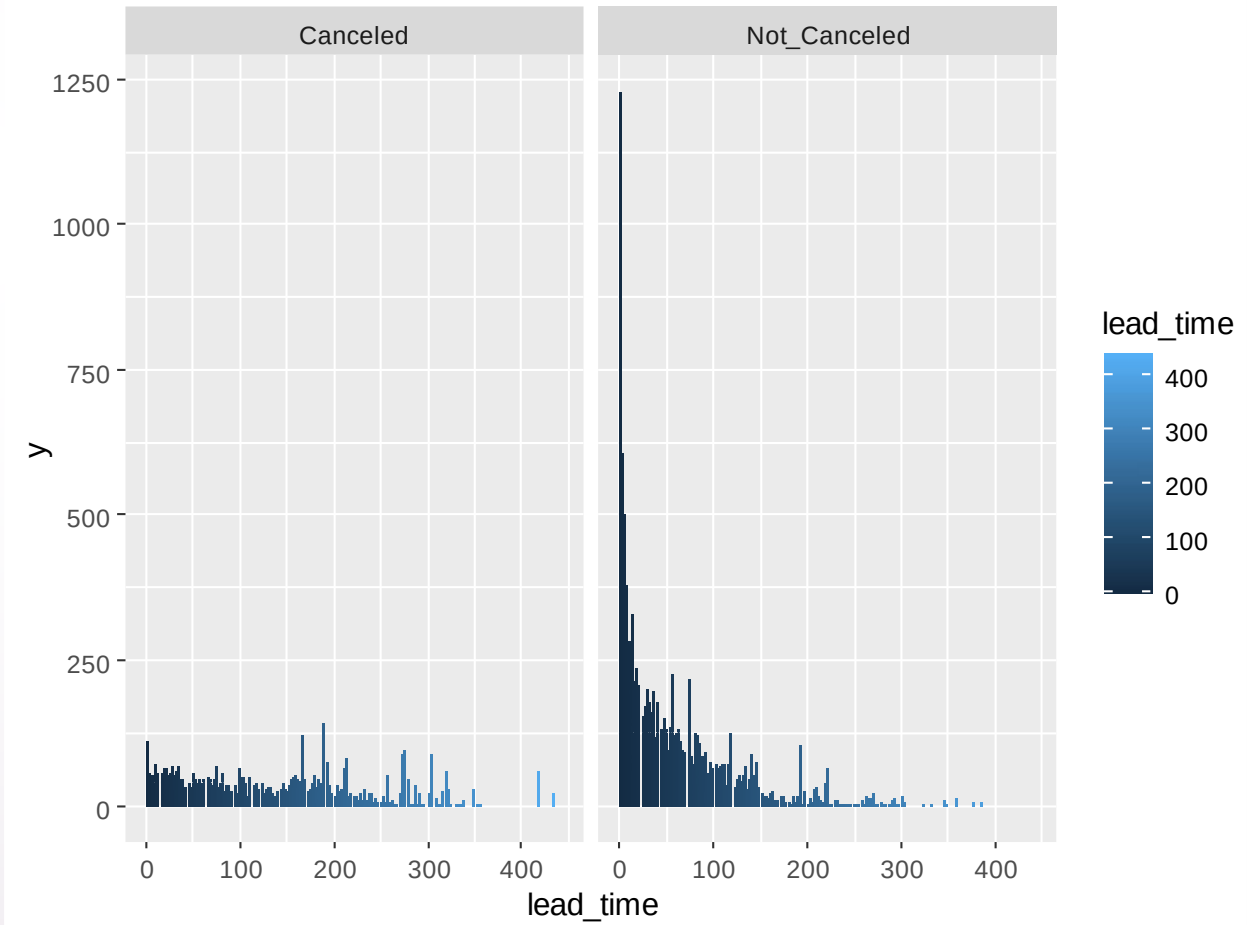



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6554  779
         1 1405 2200

               Accuracy : 0.8003
                 95% CI : (0.7927, 0.8078)
    No Information Rate : 0.7276
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5273

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.8235
            Specificity : 0.7385
         Pos Pred Value : 0.8938
         Neg Pred Value : 0.6103
             Prevalence : 0.7276
         Detection Rate : 0.5992
   Detection Prevalence : 0.6704
      Balanced Accuracy : 0.7810

       'Positive' Class : 0
```

Checking the most important variables of the model

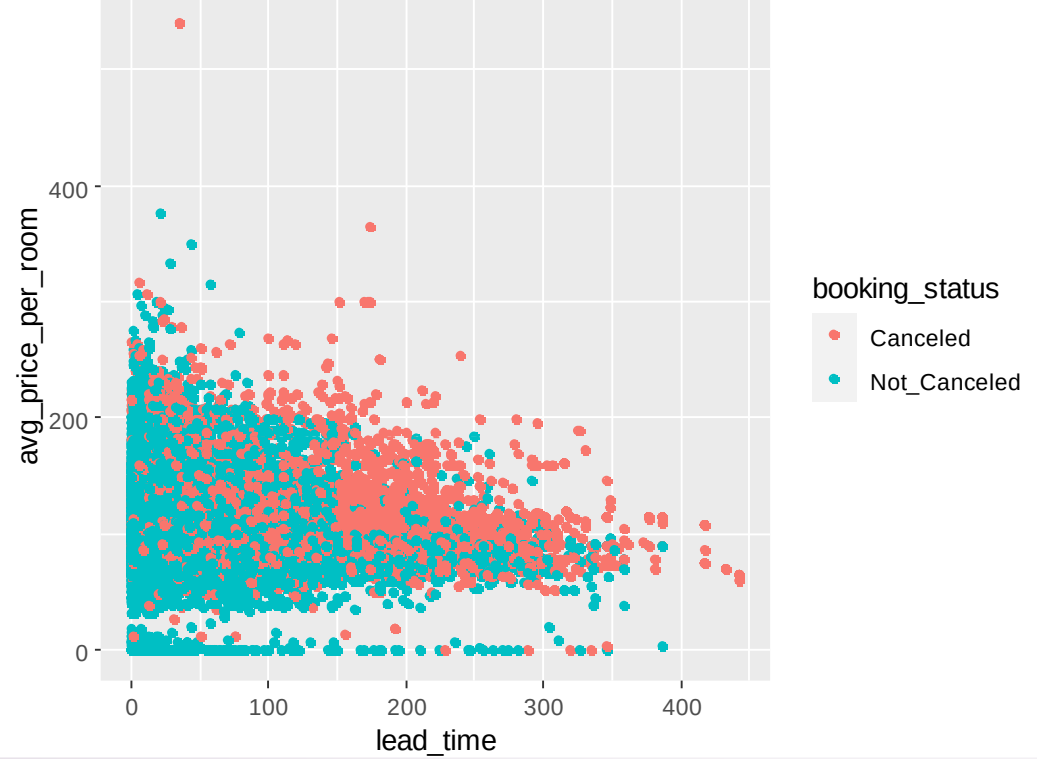
- Lead_time is by far the most important predictor variable, followed by no_of_special_request, and average price per room.



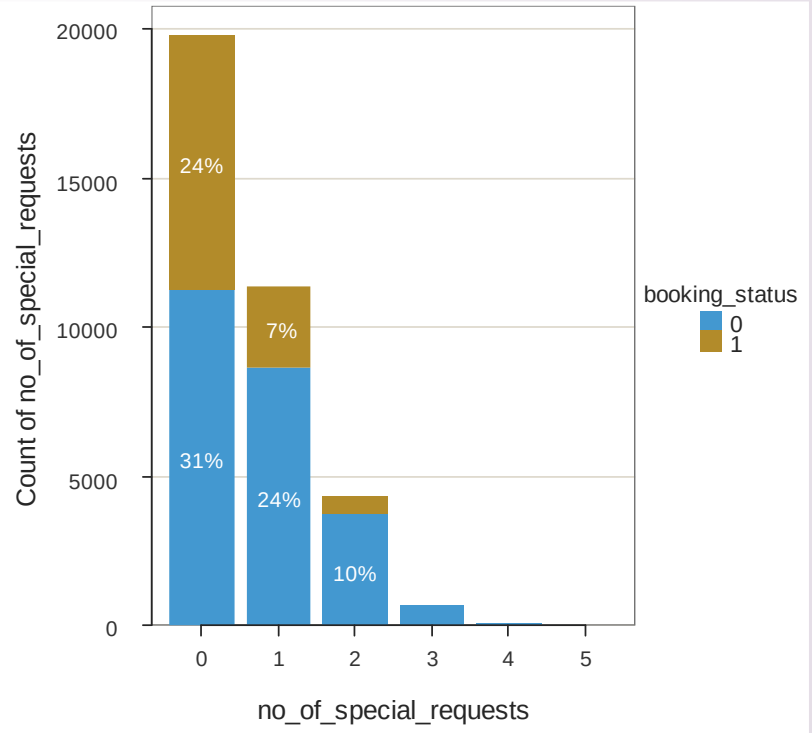


CONCLUSION

- Model Accuracy is 80.03%.
- Most important variable for predicting is Lead_time, followed by Average price per room and no. of special guest.



(0 = Not Cancelled, 1 = Cancelled)



Thank You

APPENDIX

Data Dictionary

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