★★★★★

Hotel Resevation

Mehdi Ghasemi
Ali Ziaei Jazi

I

Introduction

- ***Introduction***

HOTEL & SPA

★★★★★

What is Problem?

Customer behavior and booking possibilities have been radically changed by online hotel reservation channels. Cancellations or no-shows cause a significant number of hotel reservations to be canceled. Cancellations can be caused by a variety of factors, such as scheduling conflicts, changes in plans, etc. In many cases, this is made easier by the possibility of doing so free or at a low cost, which is beneficial for hotel guests but less desirable and possibly revenue-diminishing for hotels.

Introduction

What is the data mining question asked to solve the problem?

As a Data Scientist, your job is to build a Machine Learning model to help the Hotel Owners better understand if the customer is going to honor the reservation or cancel it ?

DataSet

The file contains the different attributes of customers' reservation details.

36275 Records
19 Colums

- *Introduction*

Attributes

1

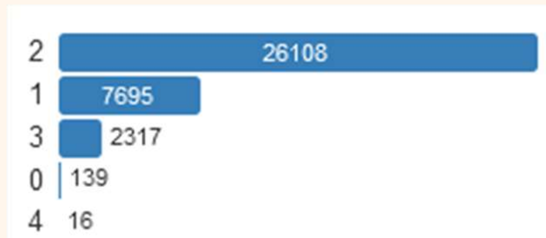
Booking_ID

unique identifier of each booking

2

no_of_adults

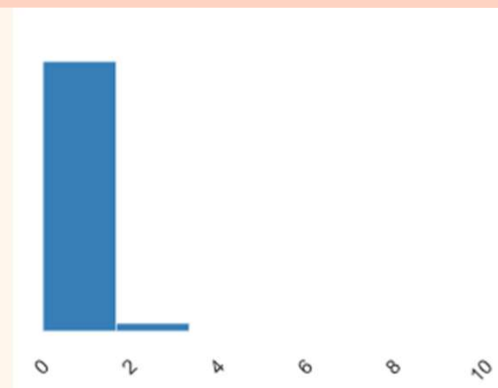
Number of adults



3

no_of_children

Number of Children

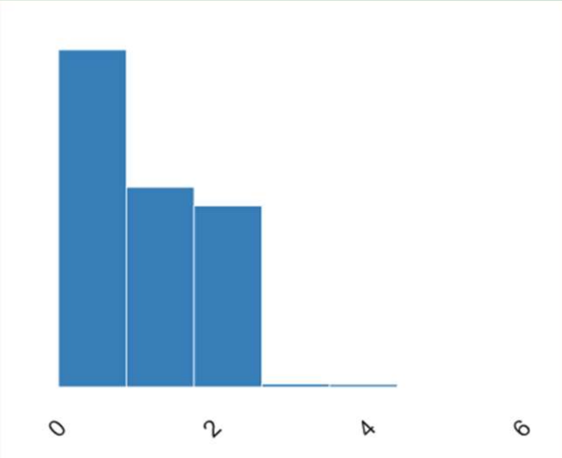


★★★★★

4

no_of_weekend_nights

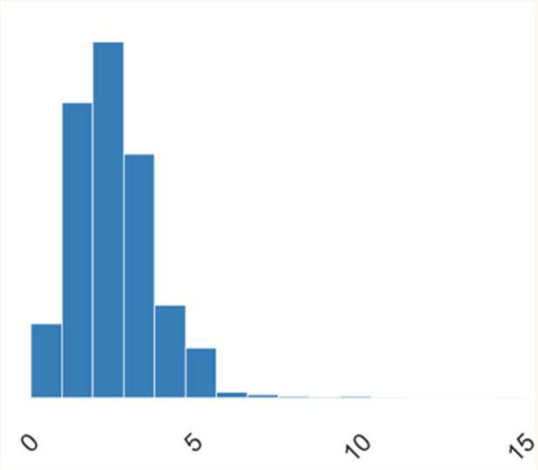
Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel



5

no_of_week_nights

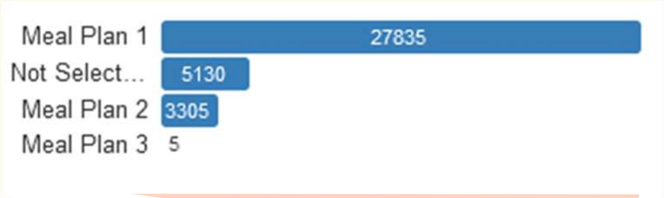
Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel



6

type_of_meal_plan

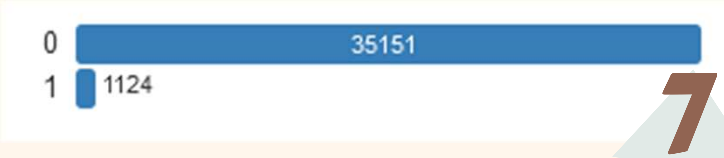
Type of meal plan booked by the customer



7

required_car_parking_space

Does the customer require a car parking space? (0 - No, 1- Yes)



Attributes

7

- ***Introduction***

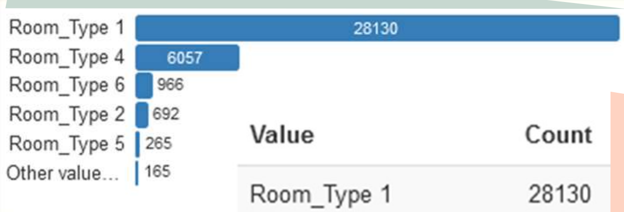
Attributes

★★★★★

8

room_type_reserved

Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.

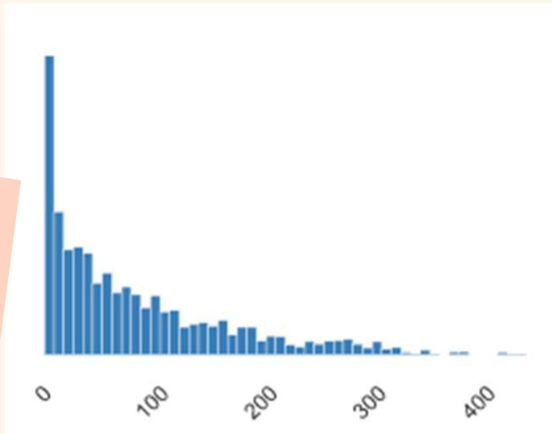


Value	Count
Room_Type 1	28130
Room_Type 4	6057
Room_Type 6	966
Room_Type 2	692
Room_Type 5	265
Room_Type 7	158
Room_Type 3	7

9

lead_time

Number of days between the date of booking and the arrival date



10

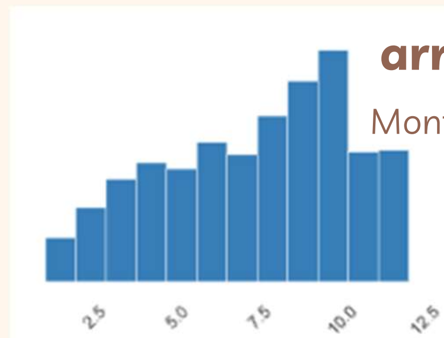
arrival_year

Year of arrival date

11

arrival_month

Month of arrival date

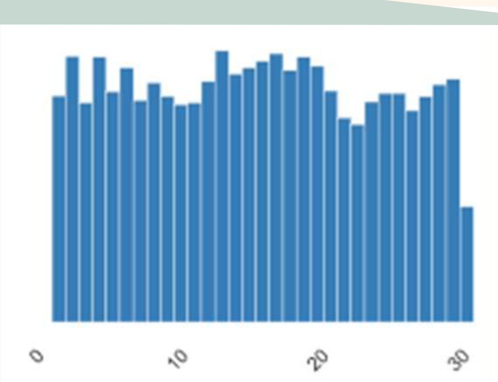


Introduction

12

arrival_date

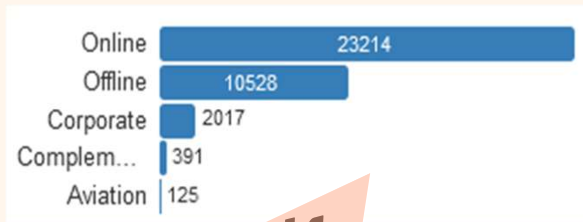
Date of the month



13

market_segment_type

Market segment designation.



14

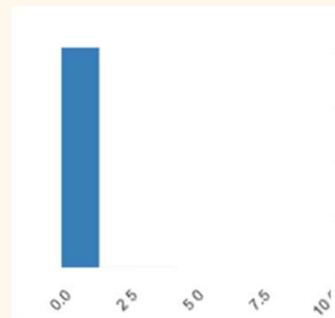
repeated_guest

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



0	35937
1	198
2	46
3	43
11	25
5	11
4	10
13	4
6	1

★★★★★

Attributes

- ## *Introduction*

Attributes

16

no_of_previous_bookings_not_canceled

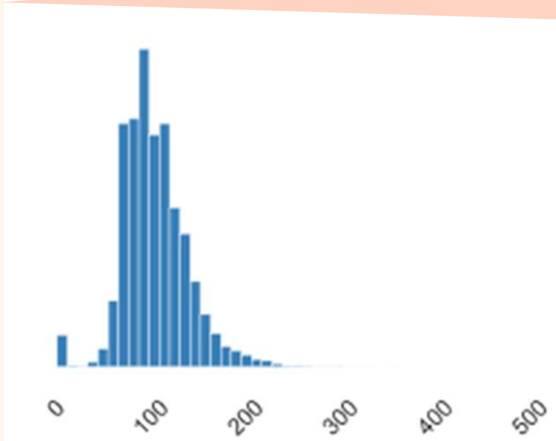
Number of previous bookings not canceled by the customer prior to the current booking

Value	Count
0	35463
1	228
2	112
3	80
4	65
5	60
6	36
7	24
8	23
10	19
Other values (...	165

17

avg_price_per_room

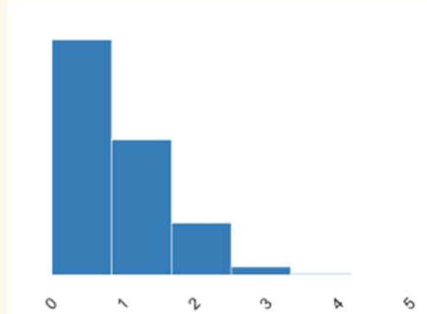
Average price per day of the reservation; prices of the rooms are dynamic. (in euros)



18

no_of_special_requests

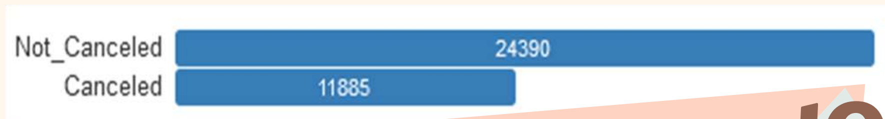
Total number of special requests made by the customer (e.g. high floor, view from the room, etc)



19

booking_status

Flag indicating if the booking was canceled or not.



- ***Introduction***

How to solve the problem

- Analysis DataSet using EDA
- Perform the necessary actions according to each feature, such as conversion, deletion and combination
- Using techniques such as oversampling and clustering and remove outlier data
- Testing different models such as KNN, SVC , RandomForest , Bagging , LightGBM with setting hyperparameters to achieve the best accuracy

- ***Introduction***

To evaluate the goodness of the model, we use metrics such as acuuracy_score , F1 score , precision score , recall score

Paying attention: because of data imbalance, except accuracy_score , we should also use another metrics.

How to evaluate

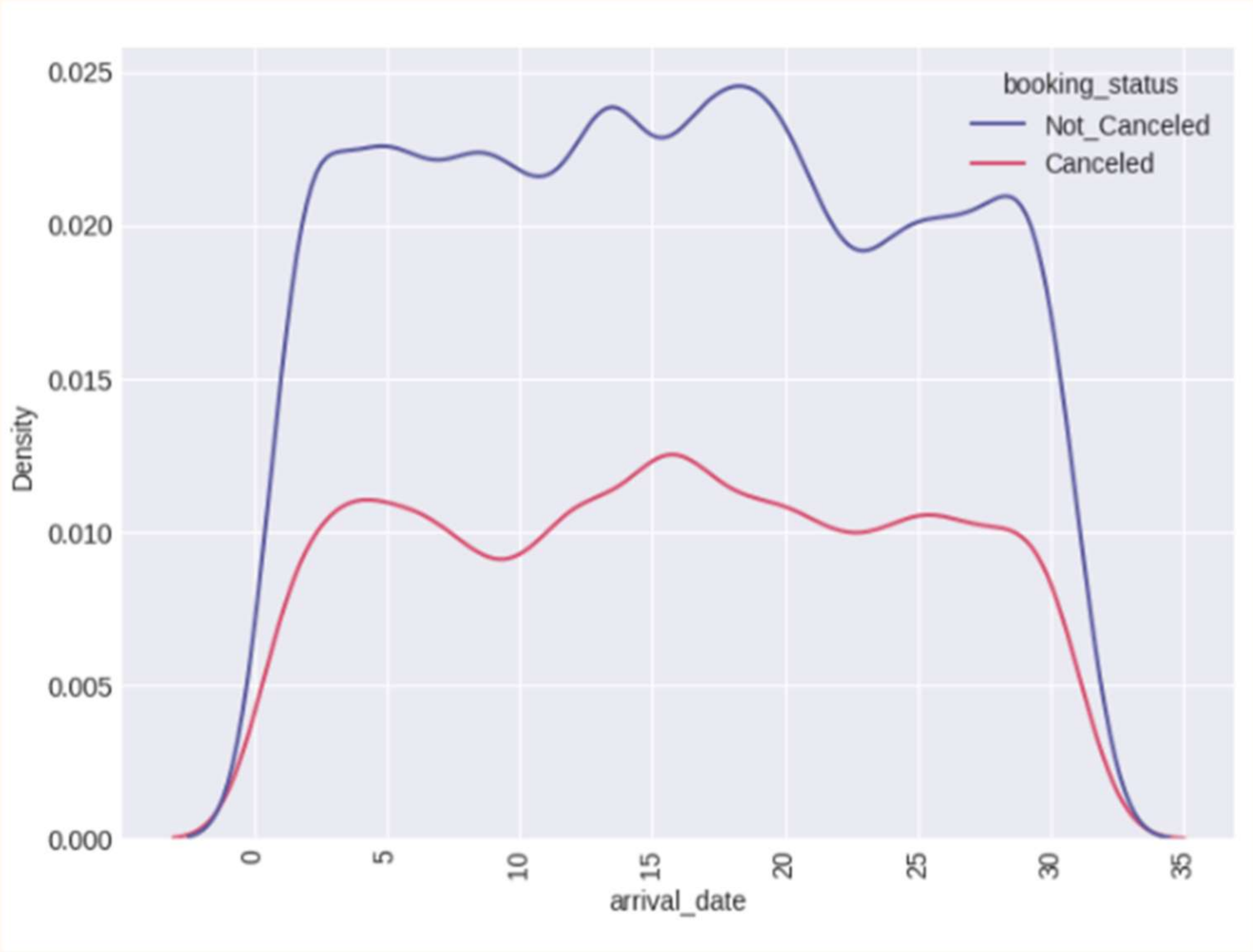
★★★★★

02

Experiments

• ***Experiments : Data***

Arrival_date

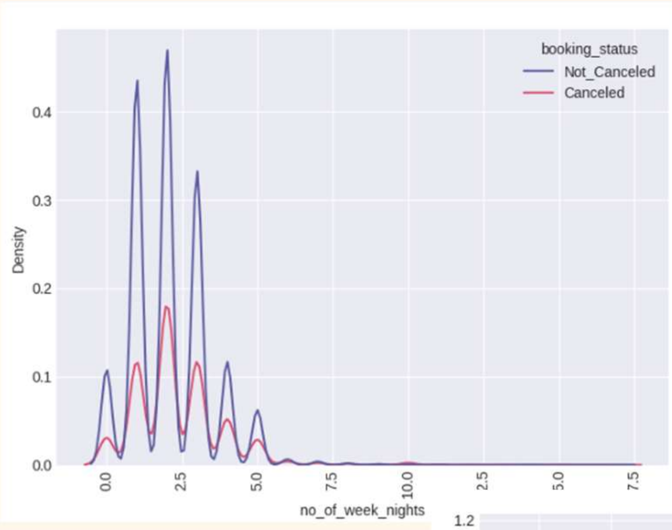


★★★★★

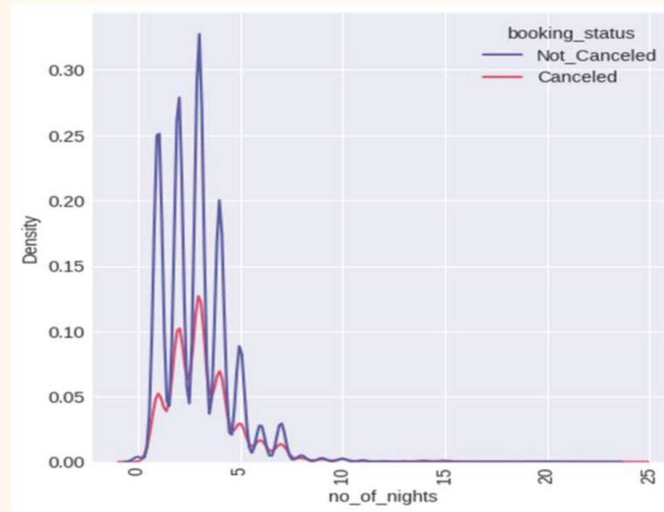
- According to the graph drawn and its relationship with the cancellation situation, because the graph is normal, it does not give us any specific information. So we need to remove this feature

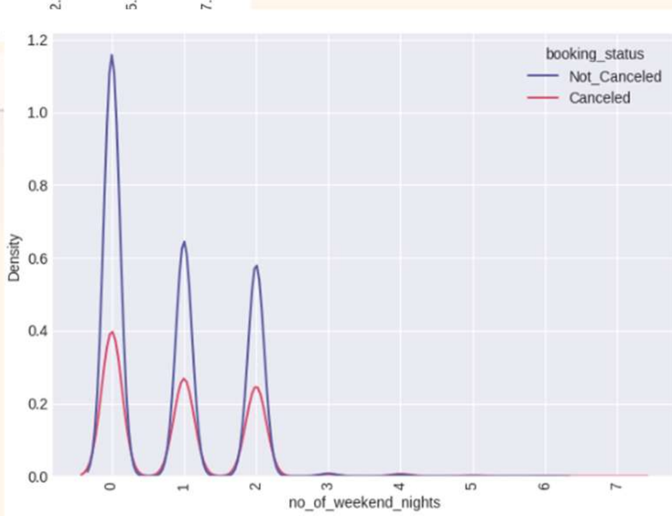
• *Experiments : Data*

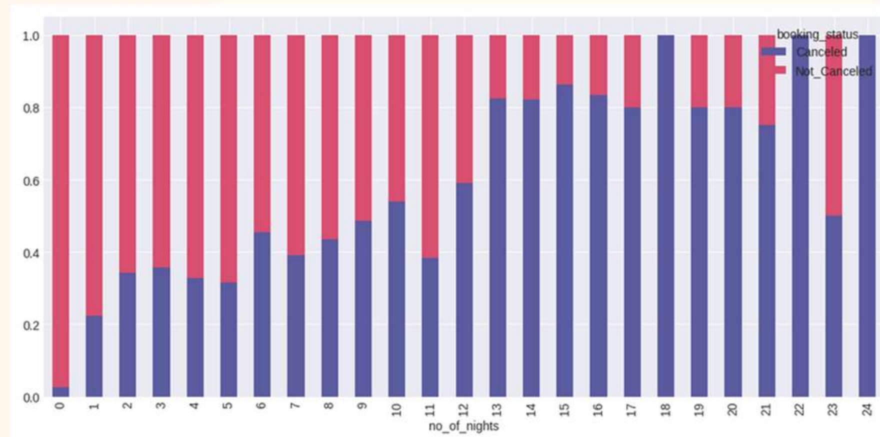
No_of_week_and_weekend_nights



- Because these two graphs are normal. We are going to get more new information by combining these two features and making a new feature.
- If the nights of stay are higher than a certain limit, the possibility of cancellation of the reservation is also higher

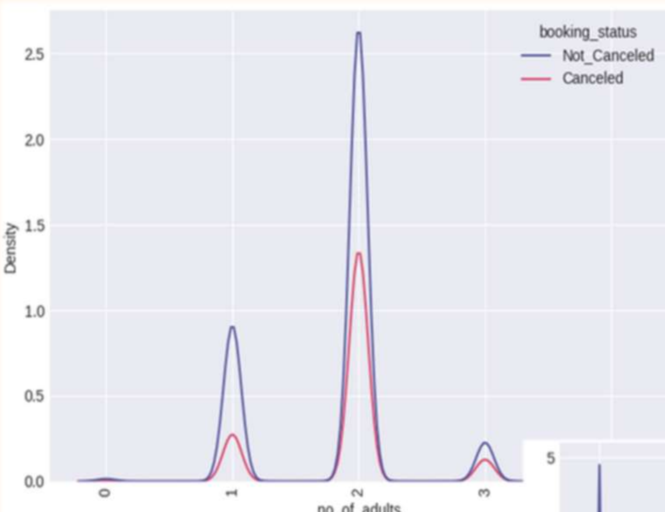




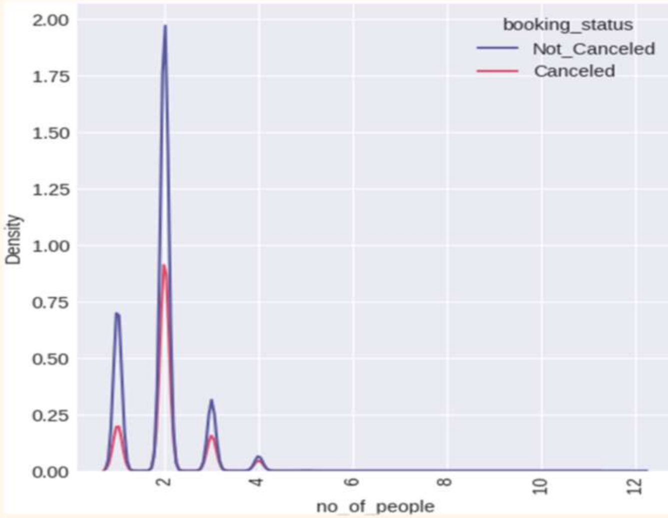


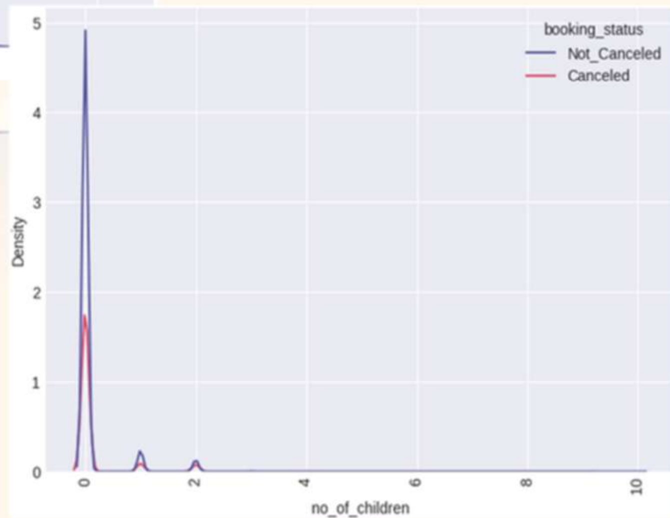
- ***Experiments : Data***

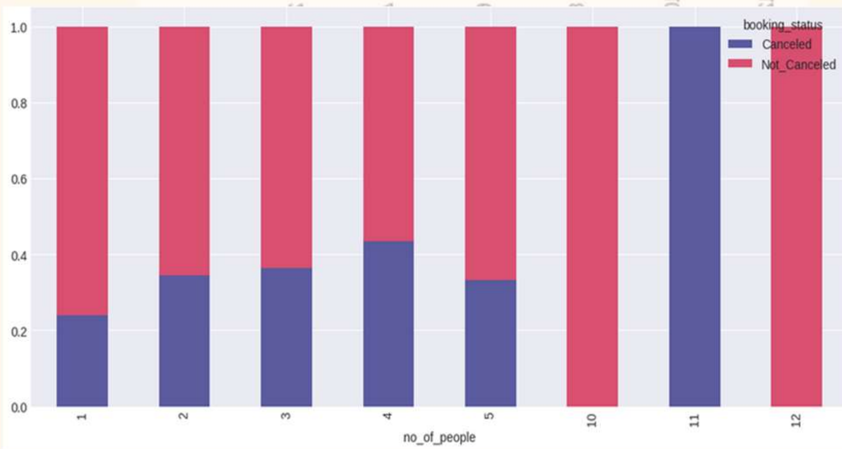
No_of_adults_and_children



- Because these two graphs are normal. We are going to get more new information by combining these two features and making a new feature.
- It is not necessary to examine each separately. Therefore, it is better to combine the two

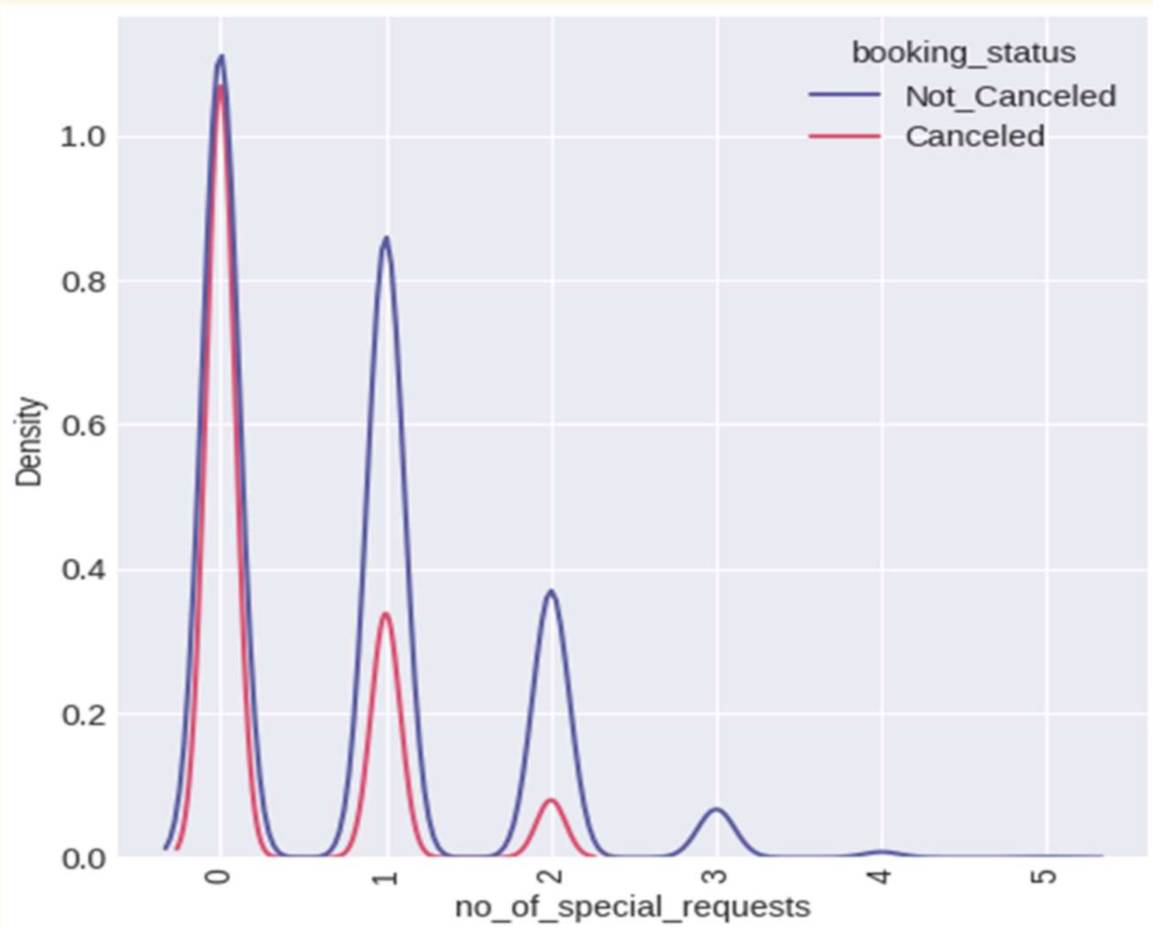






• *Experiments : Data*

No_of_special_request



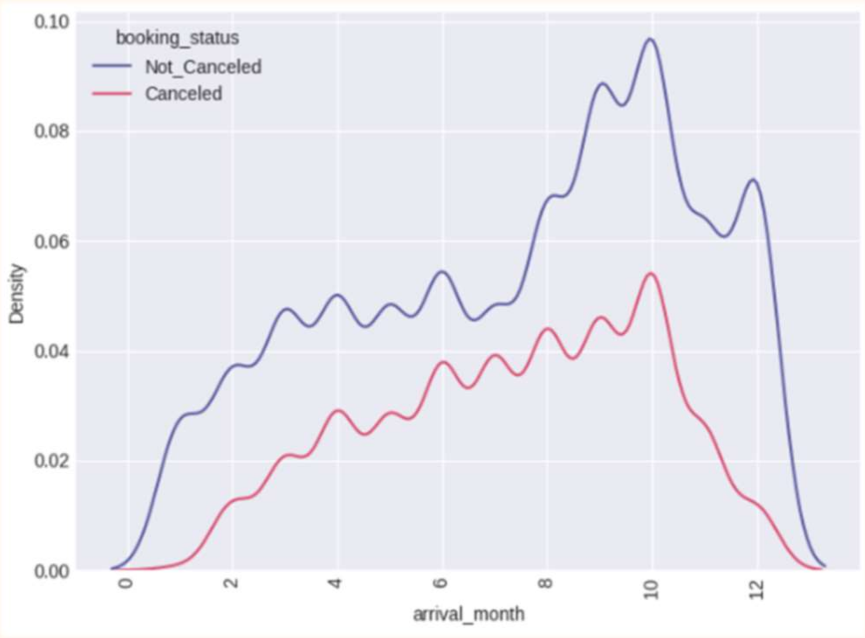
- according to the chart, because at point 0, the reservation and cancellation charts are similar, and at other points, the reservation and cancellation percentages are also similar, we can get better information by converting this chart to having a specific reservation.
- Convert this attributes to have_special_requests

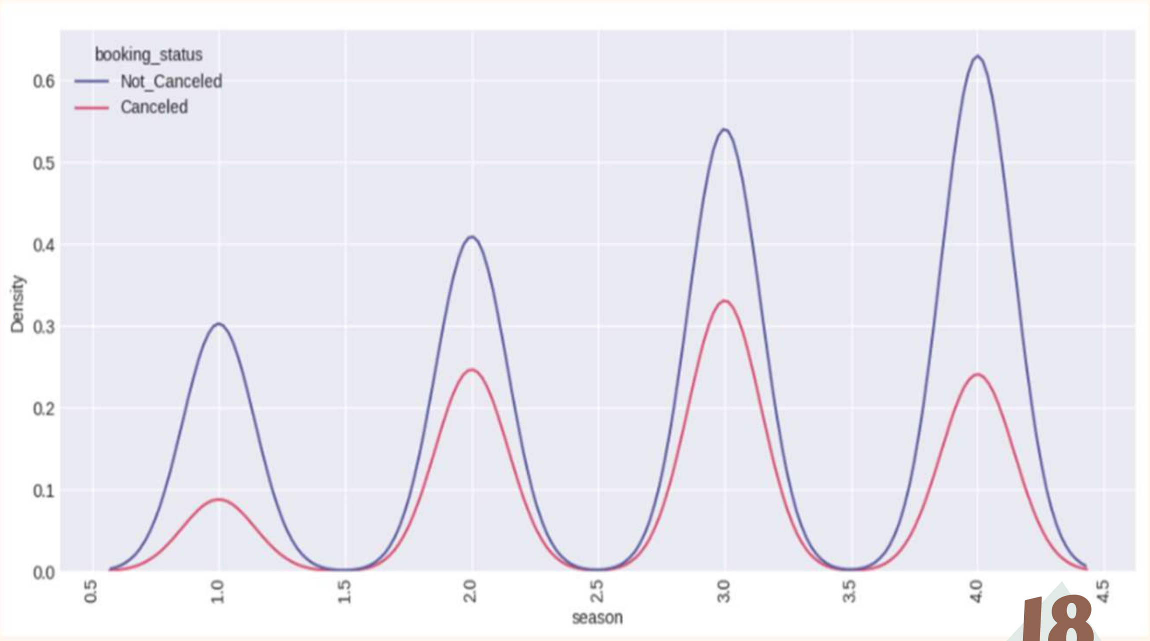
- ***Experiments* : *Data***

Arrival_month

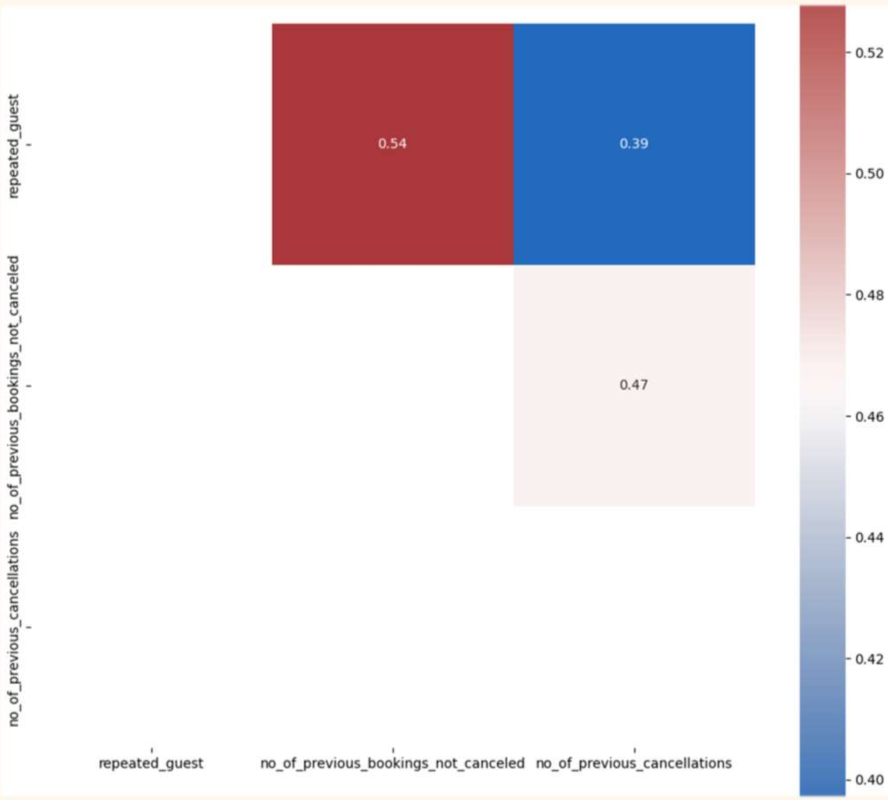
★★★★★

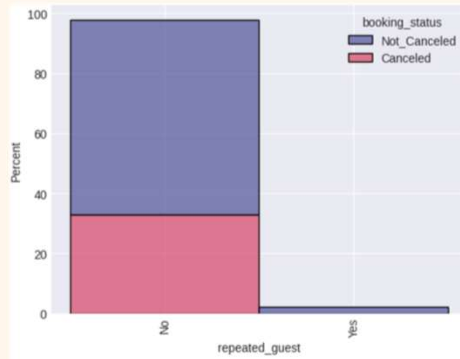
- It seems that this chart is also a normal chart, but it seems that in the last months of the year, the reservation status and non-cancellation is better. For further analysis, we convert the month chart to the season chart
- It seems that the number of reservations in the last months of the year was more than other months. Also, the number of canceled reservations in the last months is much less compared to other months

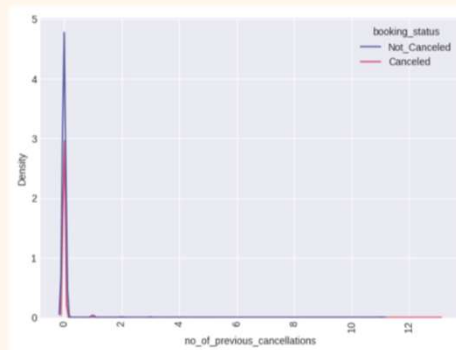


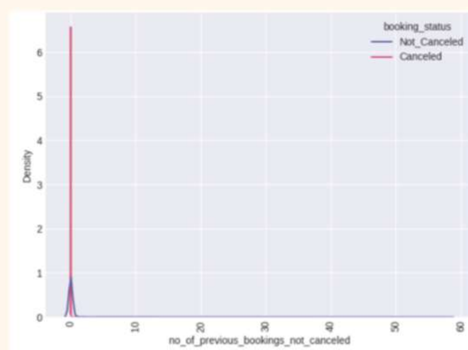


- ***Experiments : Data***



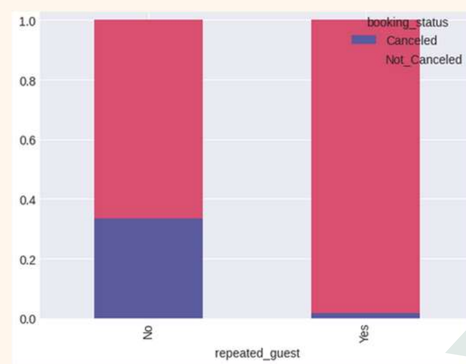






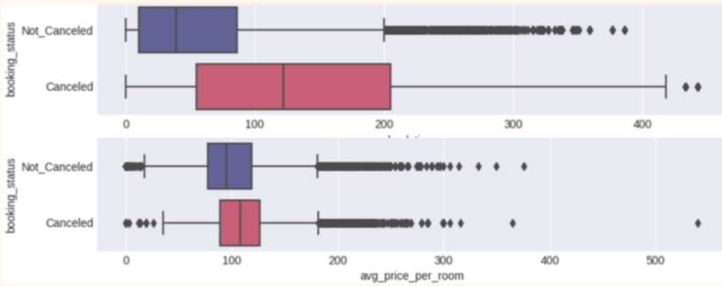
Repeated_guest

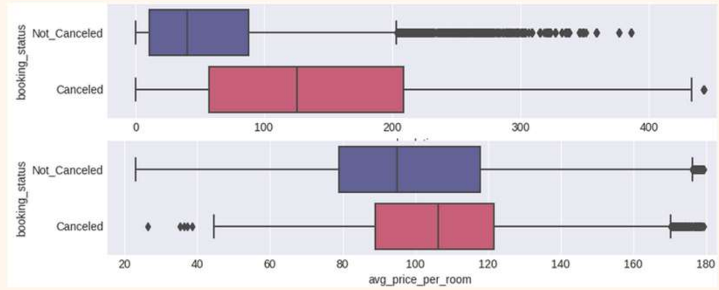
- We have an additional feature called repeated_guest. Due to the existence of the following two features, we can get the value of the repeated_guest feature by combining these two. So we can remove the repeated_guest feature from the data
- According correlation chart between these three features and observing their relationship, combining these two features is not a mistake.

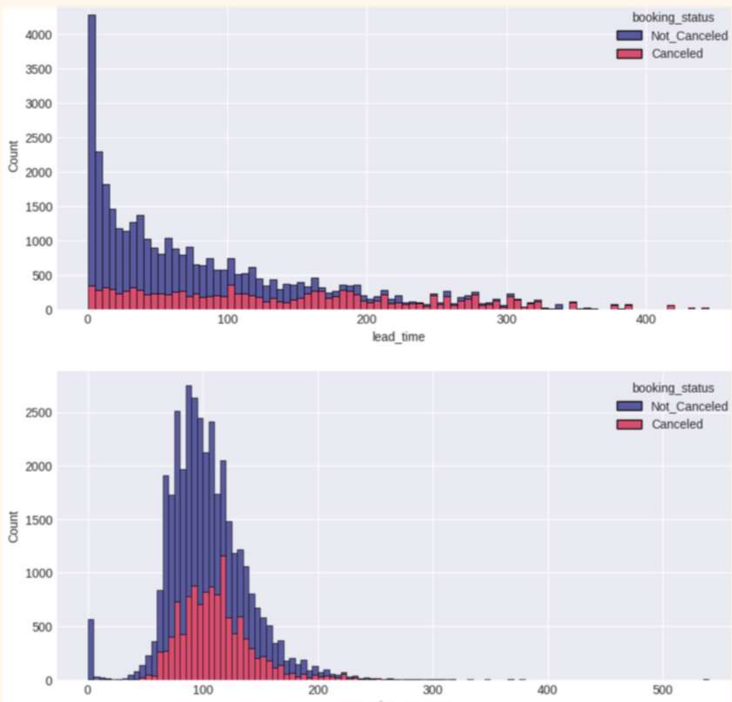


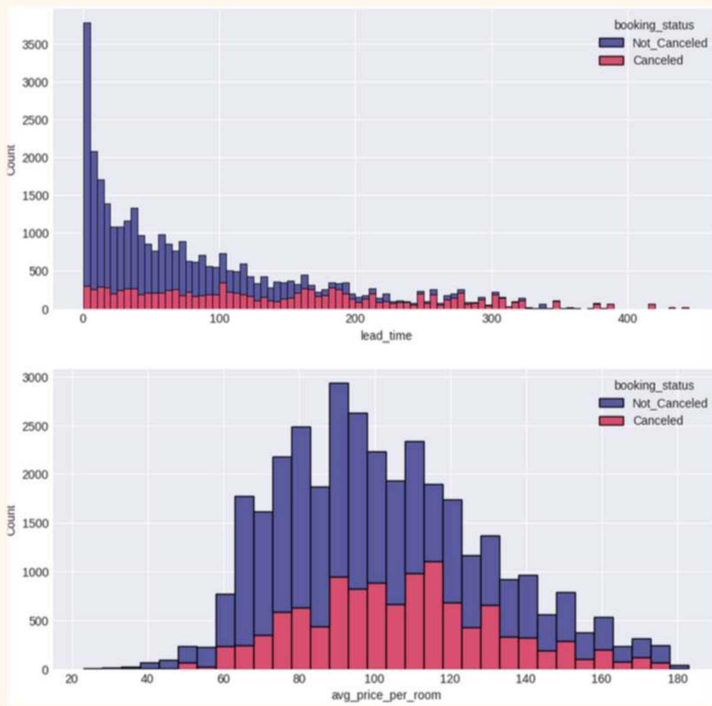
• *Experiments : Data*

Noise and outliers





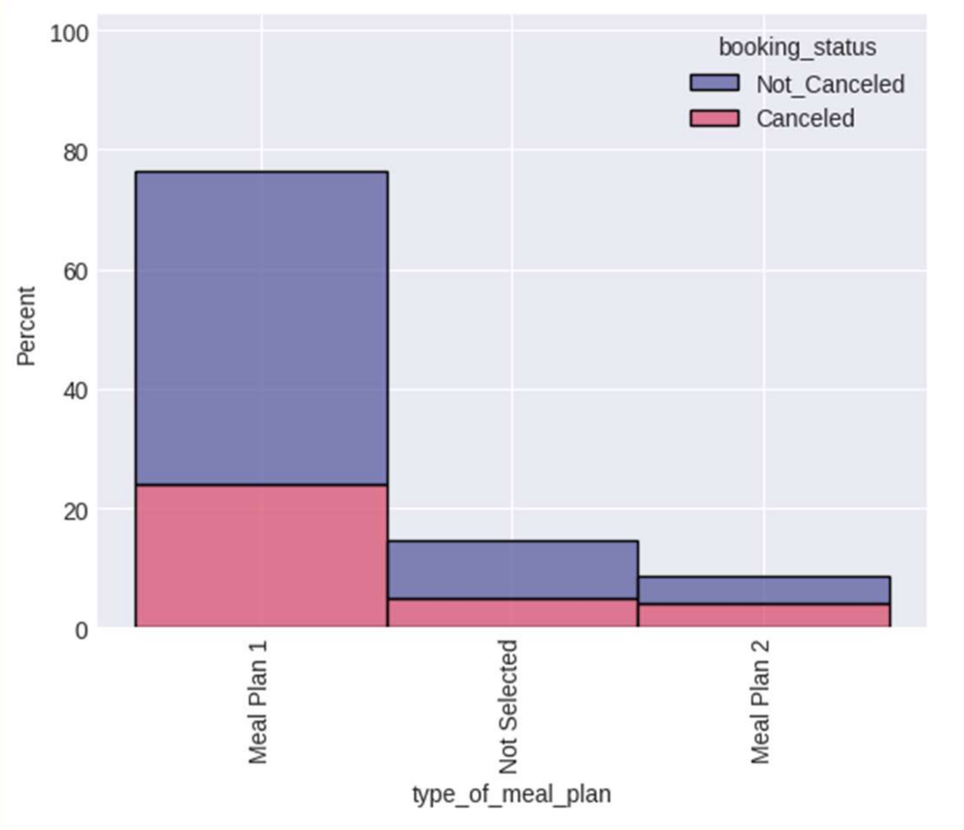




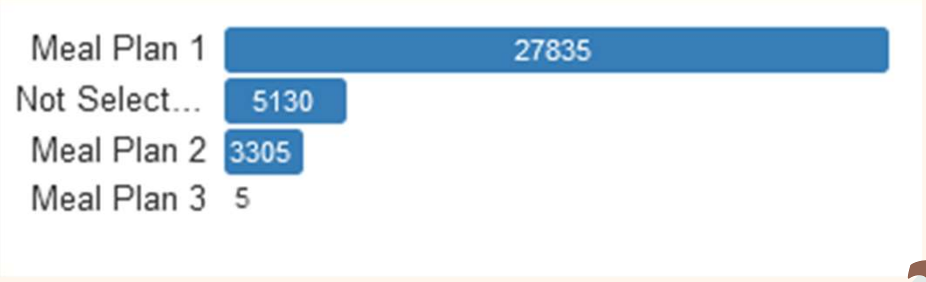
- We have two continuous variables in the data. According to the histogram and boxplot diagram,
- Data hase outlier data

- ***Experiments* : *Data***

Type_of_meal_plan



- We have an attribute called type_of_meal_plan that contains 4 unique values. Considering the small number of meal_plan_3, it can be considered economical, and by the way, after removing the outlier data, these values were removed from the dataset.

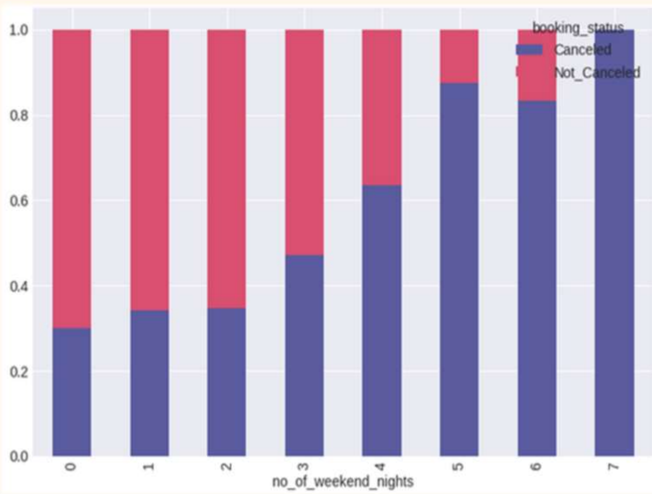


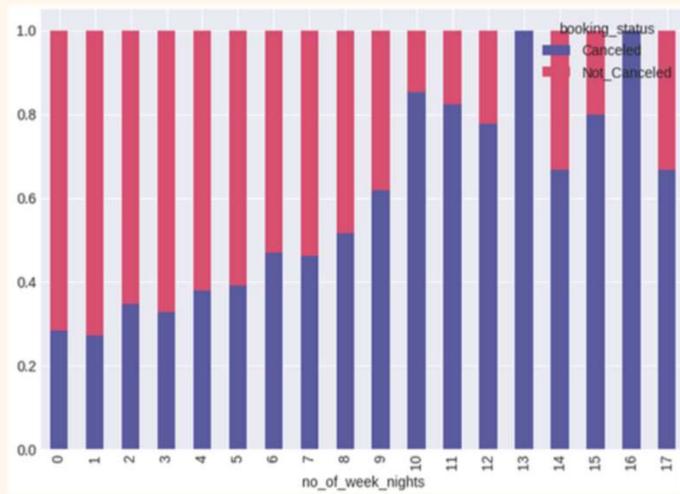
- ***Experiments* : *Data***

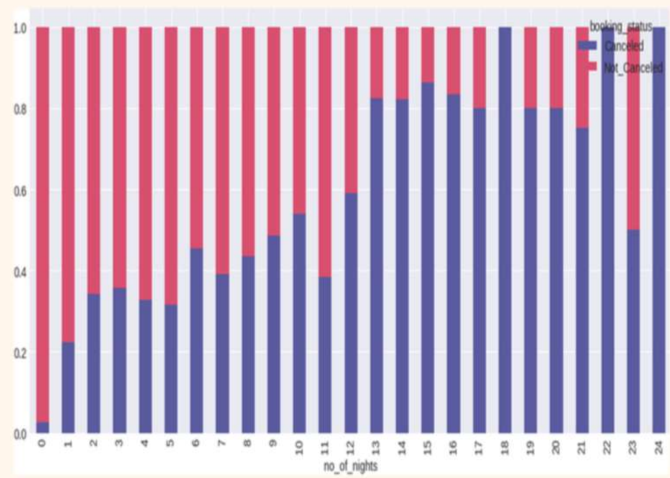
No_of_nights

★★★★★

- By drawing graphs as a percentage, you can understand that the more the number of days of stay exceeds a certain limit, the higher the probability of cancellation of that reservation. Therefore, the number of days of stay has a direct relationship with the increase in the probability of cancellation.
- Before combining these two features, the accuracy was checked because of , accuracy did not differ much, we combined these two features





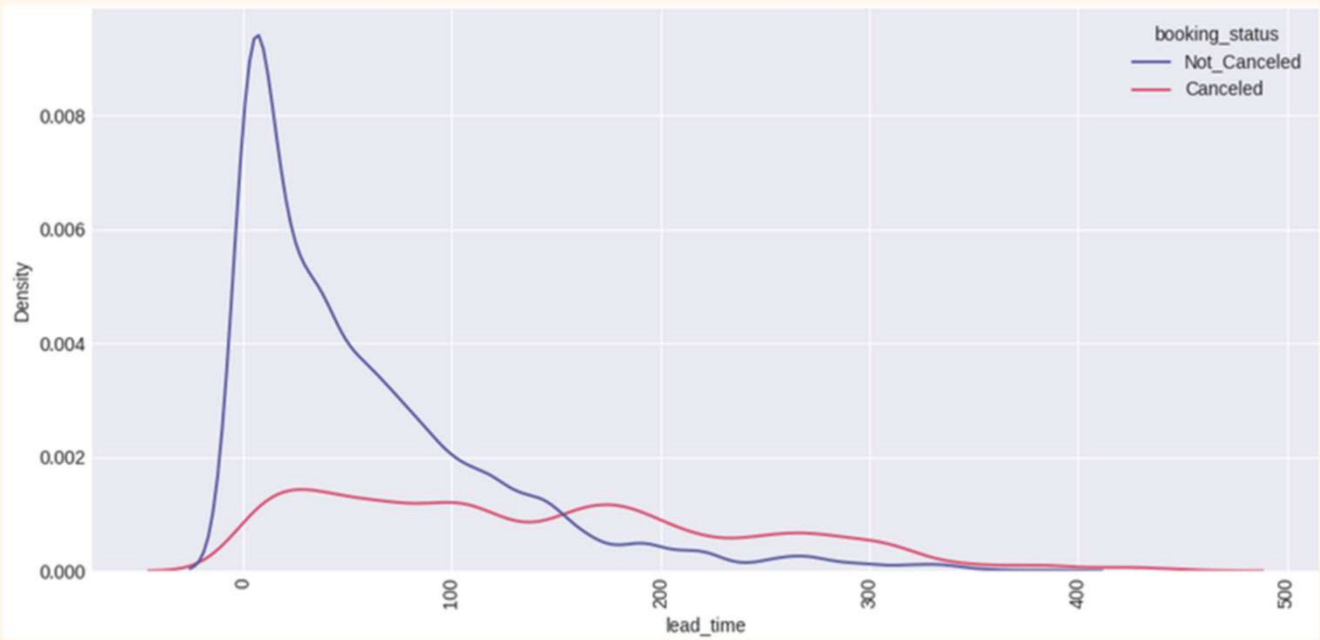


• *Experiments : Data*

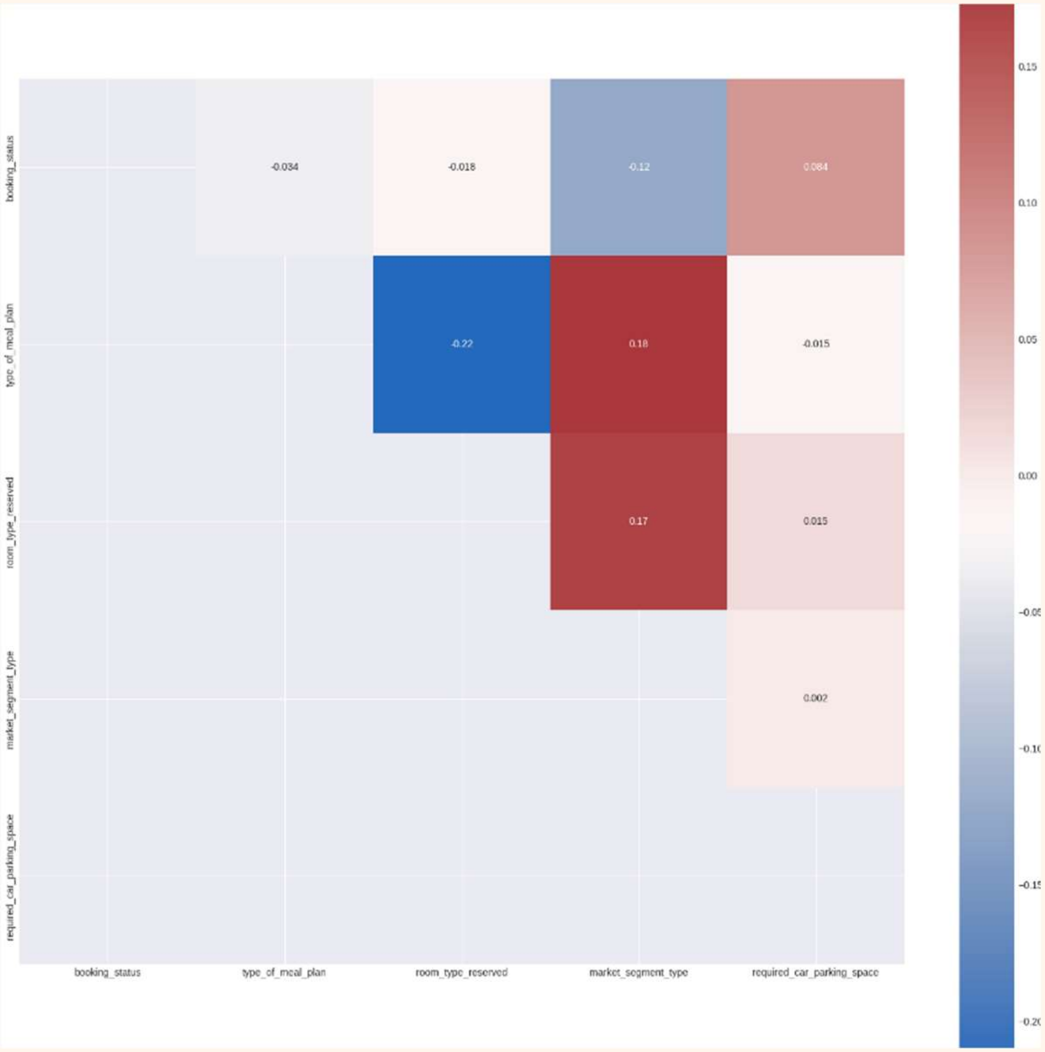
Lead_time

★★★★★

- According to the diagram, if the time to arrive the hotel (lead_time) is higher than a certain limit, the possibility of cancellation of the reservation will increase.



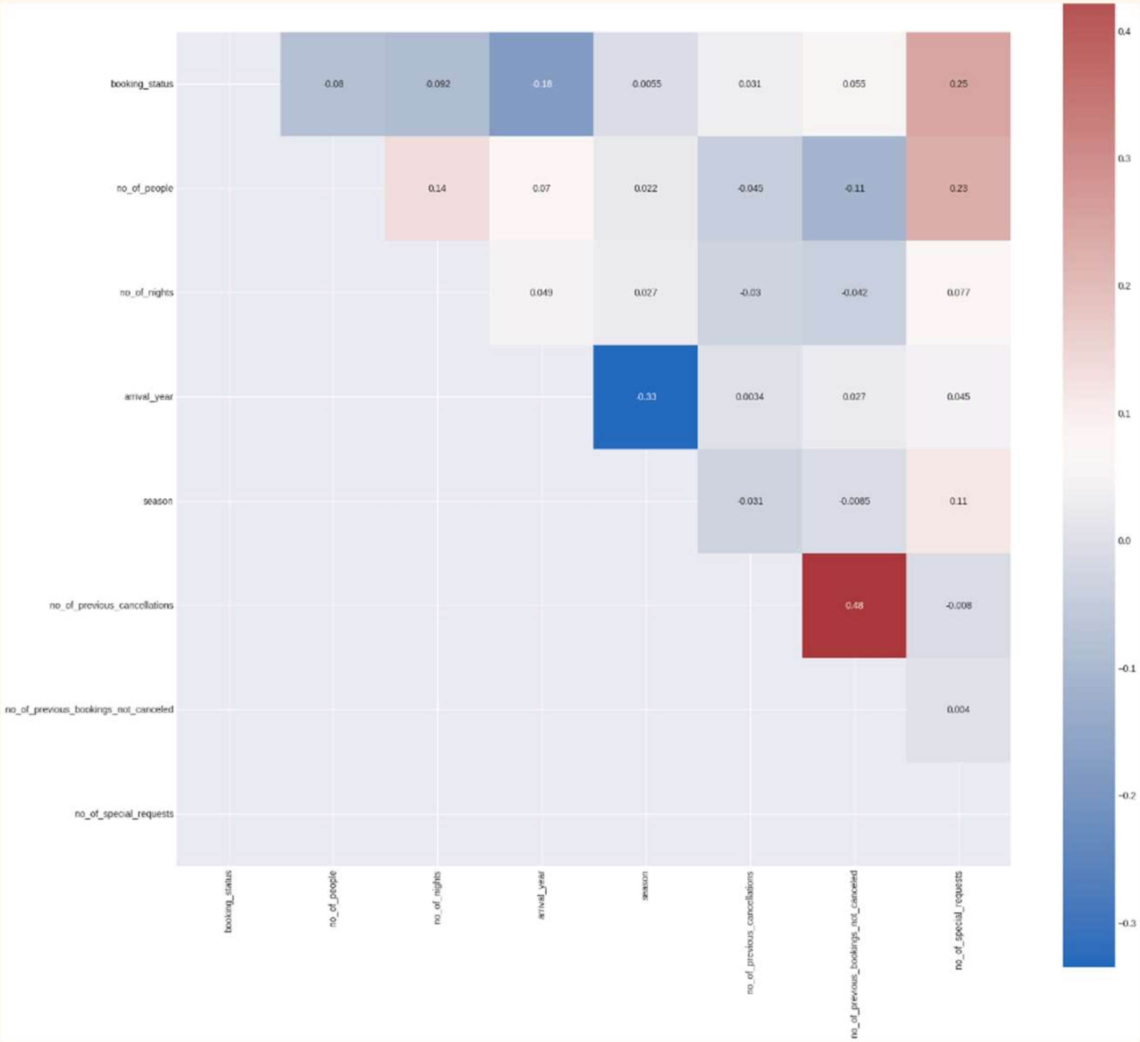
• Experiments : Data



★★★★★

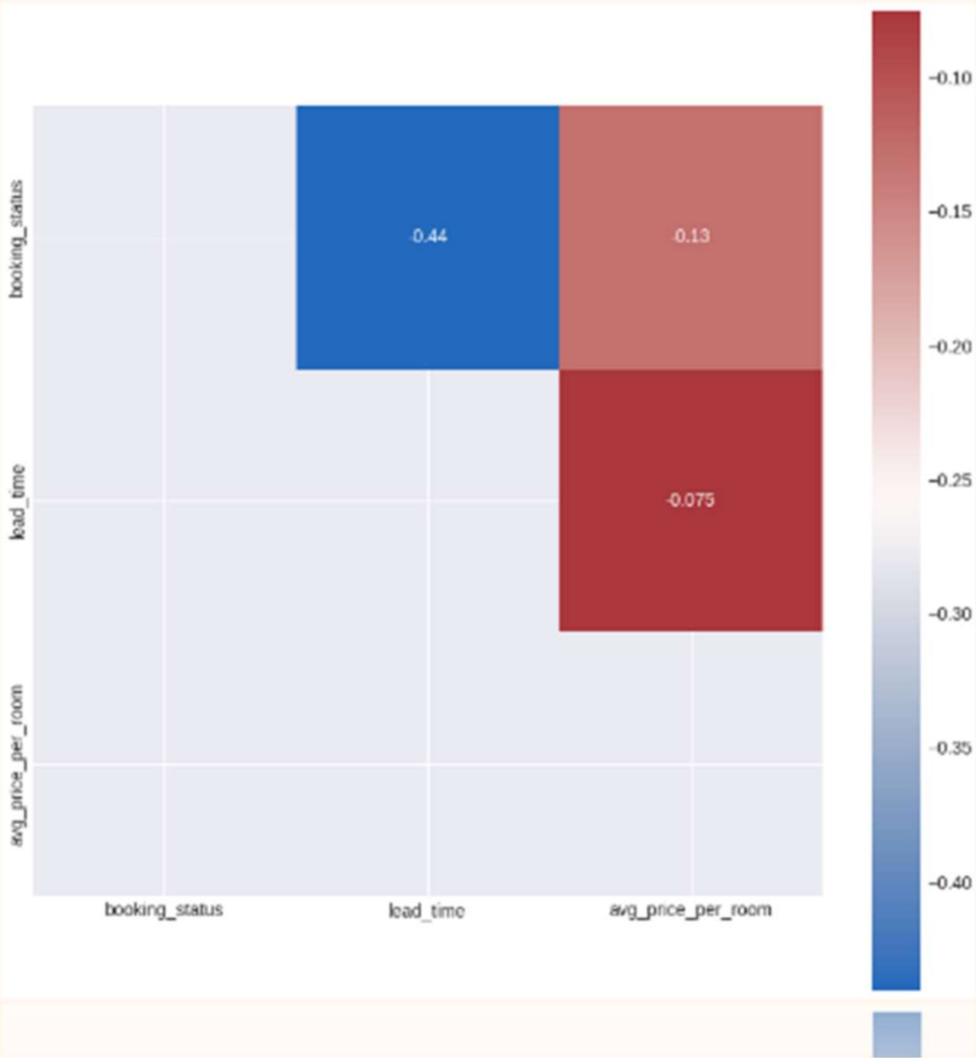
- Now we use correlation functions to check and select features. The diagram below shows the relationship between the nominal variables and the target variable. As it is clear, the market_segment_type feature has the most impact.
- It seems that better information cannot be obtained from this chart. Therefore, we also check other variables.

• Experiments : Data



- The diagram below shows the relationship between non-nominal variables and the target variable. As it is known, the features no_of_special_request and arrival_year have the most impact.
- We checked the arrival_year variable, which did not give us specific results.

• *Experiments : Data*



- The diagram below shows the relationship between continues variables and the target variable.
- these two variables have a direct effect on the target variable

- # Experiments : Modeling

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
RandomForestClassifier	0.89	0.87	0.87	0.89	1.77
XGBClassifier	0.89	0.86	0.86	0.89	2.23
BaggingClassifier	0.88	0.86	0.86	0.88	0.46
ExtraTreesClassifier	0.88	0.86	0.86	0.88	1.51
LGBMClassifier	0.88	0.86	0.86	0.88	0.32
DecisionTreeClassifier	0.86	0.84	0.84	0.86	0.09
LabelPropagation	0.85	0.83	0.83	0.85	21.54
LabelSpreading	0.85	0.83	0.83	0.85	42.55
ExtraTreeClassifier	0.85	0.83	0.83	0.85	0.04
KNeighborsClassifier	0.85	0.83	0.83	0.85	1.68
SVC	0.84	0.79	0.79	0.84	23.76
AdaBoostClassifier	0.82	0.78	0.78	0.81	0.67
LogisticRegression	0.81	0.76	0.76	0.80	0.09
NearestCentroid	0.76	0.75	0.75	0.77	0.03
NuSVC	0.81	0.75	0.75	0.80	35.28
CalibratedClassifierCV	0.80	0.75	0.75	0.80	4.78
LinearSVC	0.80	0.75	0.75	0.80	1.27
LinearDiscriminantAnalysis	0.80	0.74	0.74	0.79	0.18
RidgeClassifier	0.80	0.73	0.73	0.79	0.06
RidgeClassifierCV	0.80	0.73	0.73	0.79	0.07
Perceptron	0.73	0.72	0.72	0.73	0.06
SGDClassifier	0.78	0.70	0.70	0.77	0.10
BernoulliNB	0.76	0.70	0.70	0.75	0.12
QuadraticDiscriminantAnalysis	0.58	0.67	0.67	0.58	0.05
PassiveAggressiveClassifier	0.77	0.67	0.67	0.74	0.13
GaussianNB	0.53	0.64	0.64	0.52	0.04
DummyClassifier	0.68	0.50	0.50	0.54	0.04

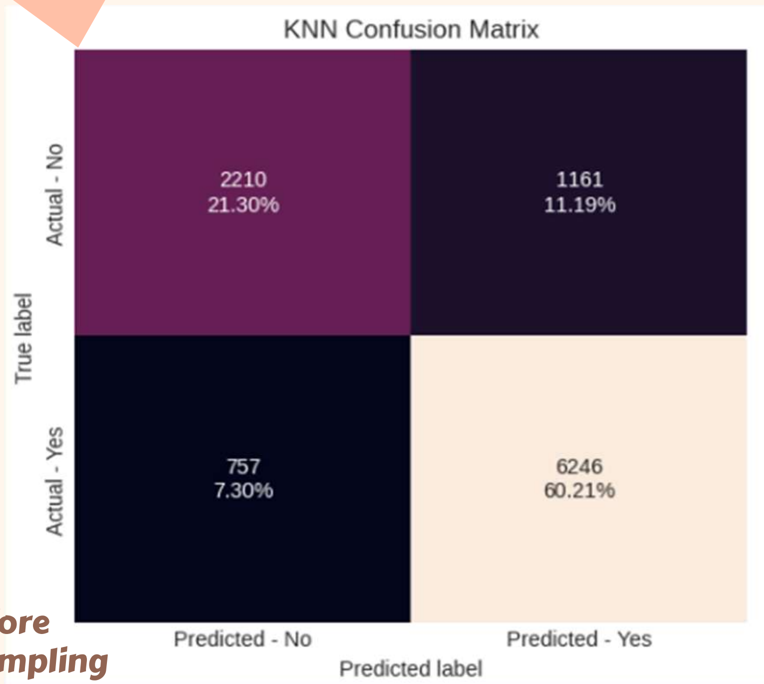
★★★★★

- Finding the best classifiers using LazyPredict package
- Tree-Based models perform well on the dataset
- In Contrast models like Naïve Bayse tha are less resistance to the imbalance of datasets, tend to have more errors.

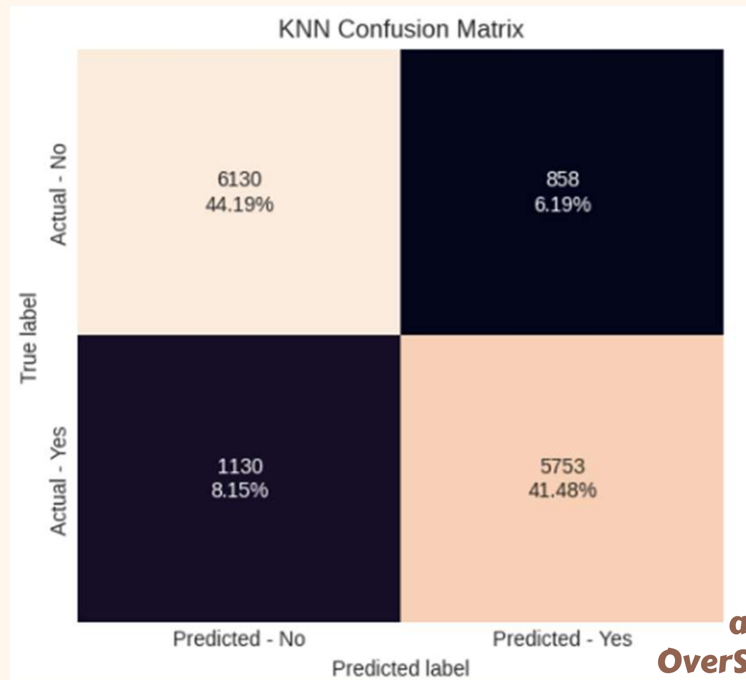
• Experiments : Modeling

KNN

	Precision	Recall	F1 Score	Accuracy Score
Before OverSampling	0/84	0/89	0/87	0/82
after OverSampling	0/87	0/84	0/85	0/86



Before OverSampling



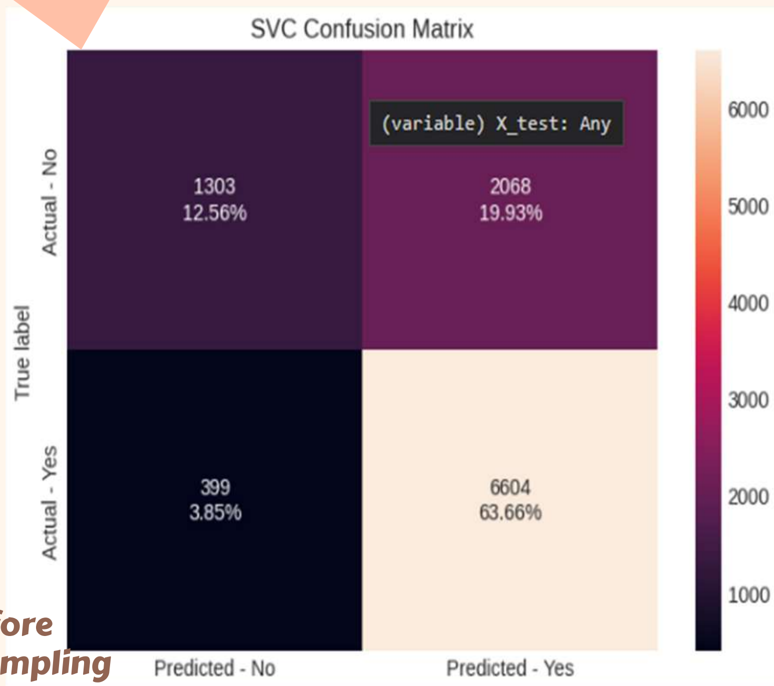
after OverSampling

• Experiments : Modeling

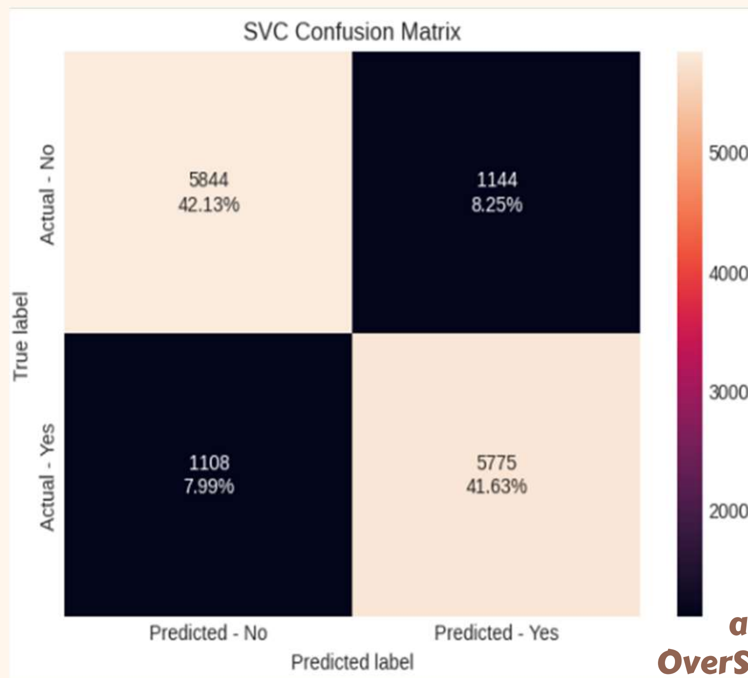
SVM

	Precision	Recall	F1 Score	Accuracy Score
Before OverSampling	0/76	0/94	0/84	0/76
after OverSampling	0/83	0/84	0/84	0/84

- As it is clear in the figure, Support Vector Machine performs poorly in imbalance datasets, so that 84% of the data is predicted by the majority class (explanation).



Before OverSampling

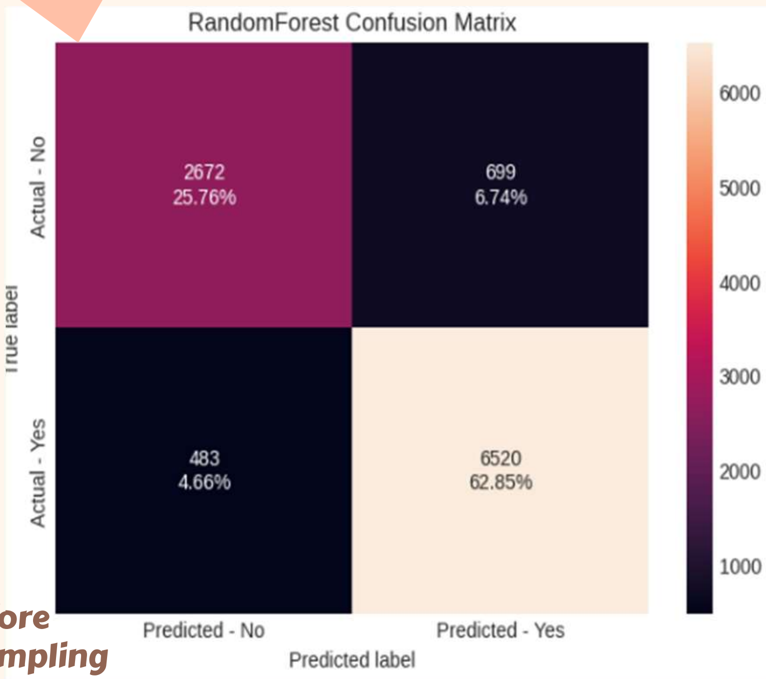


after OverSampling

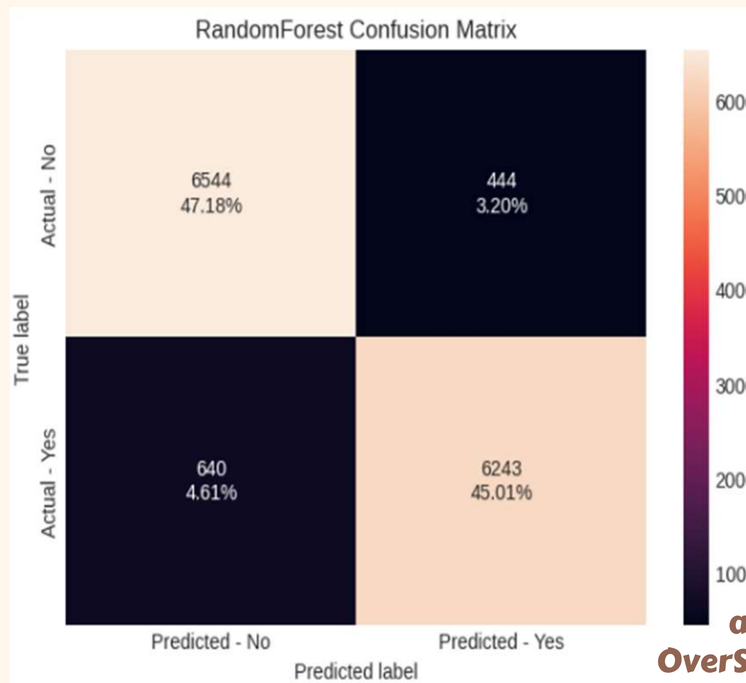
• Experiments : Modeling

Random Forest

	Precision	Recall	F1 Score	Accuracy Score
Before OverSampling	0/90	0/93	0/92	0/89
after OverSampling	0/93	0/91	0/92	0/92



Before OverSampling



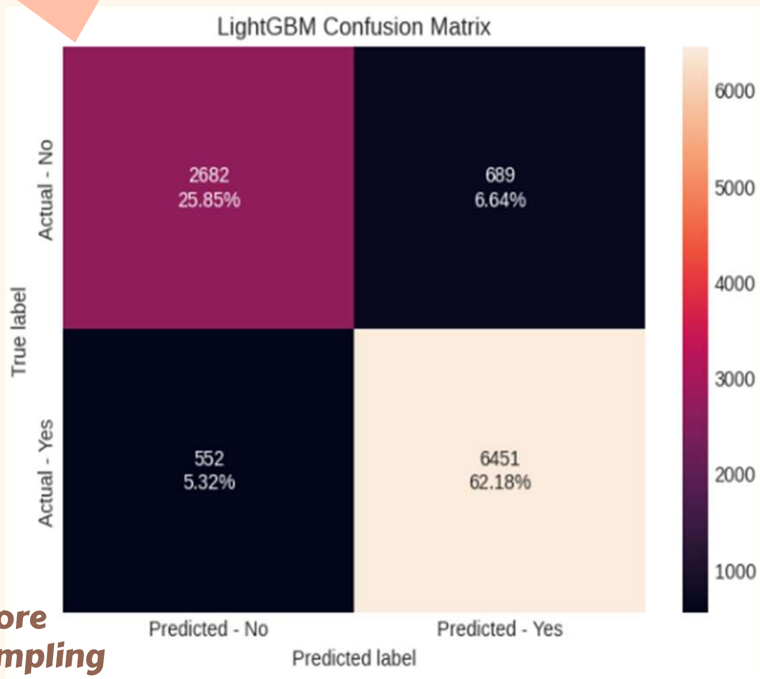
after OverSampling

• Experiments : Modeling

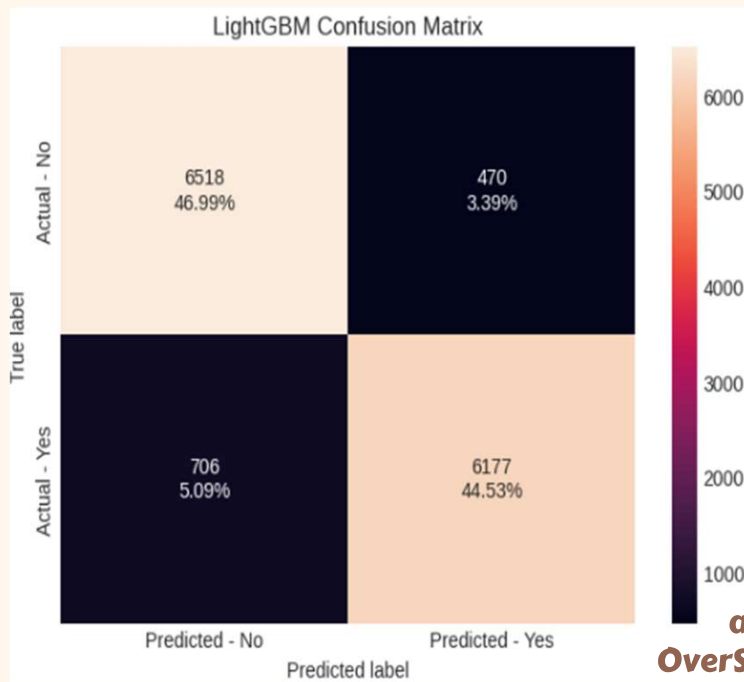
LGBM

	Precision	Recall	F1 Score	Accuracy Score
Before OverSampling	0/90	0/92	0/91	0/88
after OverSampling	0/93	0/90	0/91	0/92

★★★★★



Before OverSampling



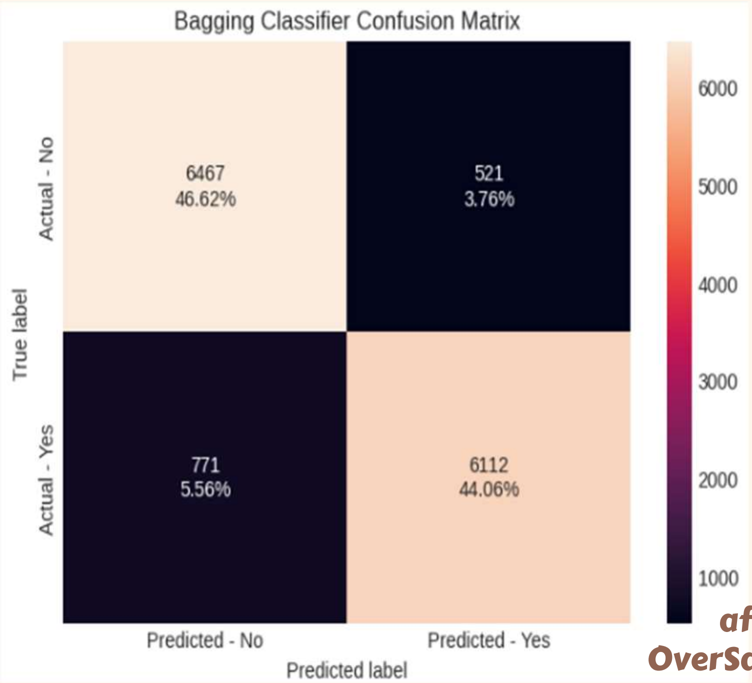
after OverSampling

• Experiments : Modeling

Bagging

	Precision	Recall	F1 Score	Accuracy Score
Before OverSampling	0/91	0/93	0/92	0/89
after OverSampling	0/92	0/89	0/90	0/91

Before OverSampling



after OverSampling

03

Results

Results

KNN

Precision : 0/87
Recall: 0/84
F1 Score: 0/85
Accuracy Score: 0/86

SVM

Precision : 0/83
Recall: 0/84
F1 Score: 0/84
Accuracy Score: 0/84

Random Forest

Precision : 0/93
Recall: 0/91
F1 Score: 0/92
Accuracy Score: 0/92

LGBM

Precision : 0/93
Recall: 0/90
F1 Score: 0/91
Accuracy Score: 0/92

Bagging

Precision : 0/92
Recall: 0/89
F1 Score: 0/90
Accuracy Score: 0/91

Suggestion

04

Suggestion

- Deducting an amount from the room rent as a penalty for canceling the reservation.
- Failure to provide services to those who have canceled more than a certain number of reservations
- Not allowing hotel reservations for the next 100 days. If such a reservation is made, it is not possible to cancel it
- Talk to customers who are about to cancel their reservations and give them better offers.
- Offering discounts to loyal customers
- Creating a national reservation network that hotels can introduce these customers to that hotel if the customers want to change the hotel.

05

Reference

Reference

- https://www.kaggle.com/code/eslamfouad/hotel-reservations-dataset-customers-statistics/
- https://www.kaggle.com/code/gabrielcord/cancellation-prediction-model-99-accuracy
- https://www.kaggle.com/code/kylegraupe/how-to-test-27-binary-classifiers-take-a-look/notebook
- https://www.kaggle.com/code/jcaliz/ps-s03e07-a-complete-eda/notebook
- https://www.kaggle.com/code/christophertimmons/random-forest-97-accuracy-score
- https://www.kaggle.com/code/the314arham/eda-bunch-of-models-optuna-lgbm-92-acc

Thank you for your attention

Do you have any questions?

Mehdi Ghasemi
Ali Ziaei Jazi
Spring 1402