Hotel Reservation Cancellation Prediction

V. Sruthi
119cs0048

Contents

Problem Statement

Given a dataset containing data of reservations made by customers in different hotels, build a machine learning model to predict whether the customer cancels his/her hotel reservation.

About the Data set

No. of rows = 119390
No. of attributes = 32
Target variable = is_canceled
No. of independent variables = 31
No. of numeric variables = 12
No. of object variables = 19

Independent variables in the data set

1) Hotel
2) Lead_time
3) Arrival_date_year
4) Arrival_date_month
5) Arrival_date_week_number
6) Arrival_date_day_of_month
7) Stays_in_weekend_nights
8) Stays_in_week_nights
9) Adults
10) Children
11) Babies

Independent variables in the data set

12)Meal

13)Country

14)Market_segment

15)distribution_channel

16)is_repeated_guest

17)previous_cancellations

18)previous_bookings_not_cancelled

19)reserved_room_type

Independent variables in the data set

20)assigned_room_type

21)booking_changes

22)deposit_type

23)agent

24)company

25)days_in_waiting_list

26)customer_type

27)adr

Independent variables in the data set

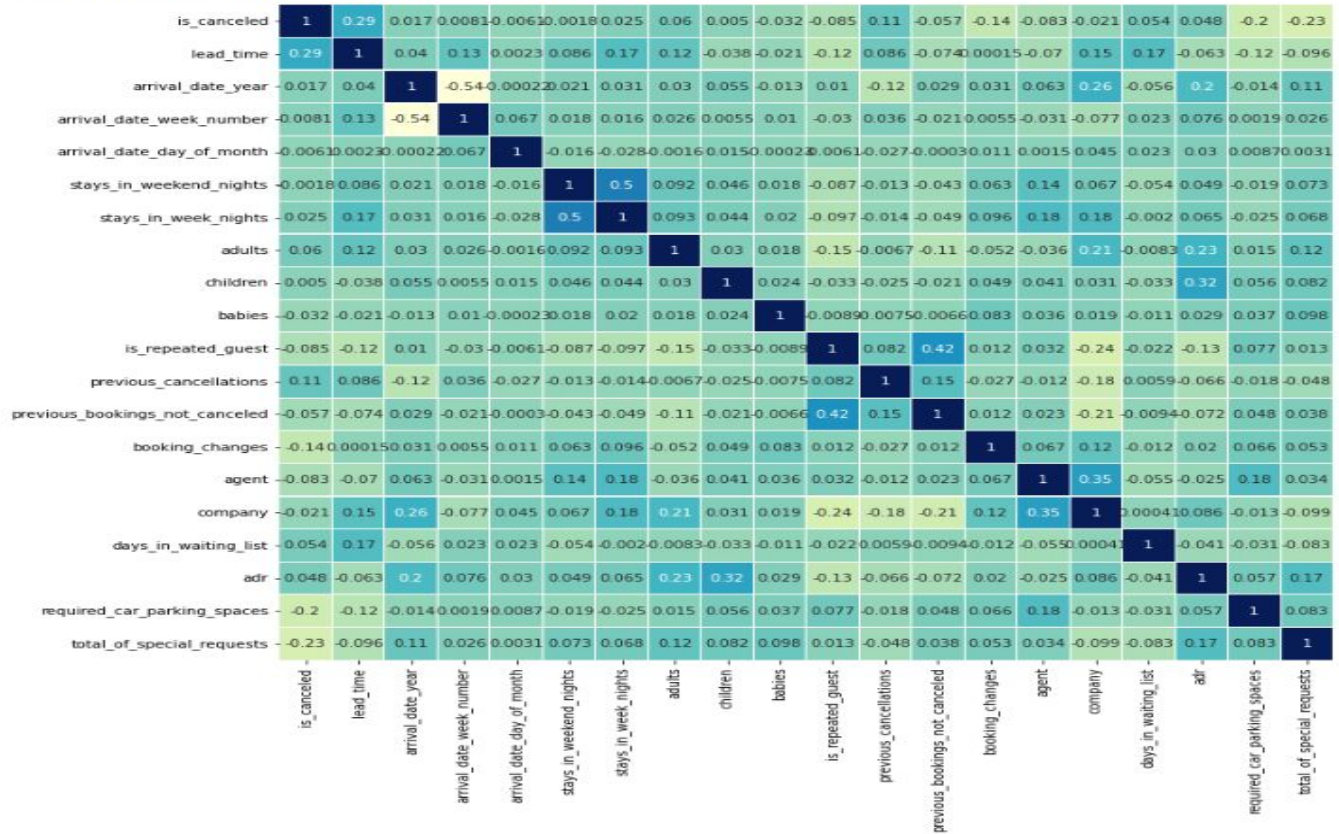
28)required_car_parking_spaces

29)total_of_special_requests

30)reservation_status

31)reservation_status_date

Correlation matrix



Data Preprocessing: Removing the null values in the data set

No. of null values in [children] = 4

No. of null values in [country] = 488

No. of null values in [agent] = 16340

No. of null values in [company] = 112593

Agent, Company and children are numerical variables. So, we fill the null values in these variables using their respective median.

Country is a categorical variable. So, we fill the null values in this variable using mode of the variable.

Data Cleaning: Removing the duplicate values in the data set

No. of duplicate values in the data set = 32013

After removing the duplicate values in the data set,

No. of rows in the dataset = 87377

Data Preprocessing: Encoding the categorical variables

Categorical data is converted into integer format to train the machine learning model. Categorical variables in our data set:

1) Hotel

2) arrival_date_month

3)meal

4)country

5)Market_segment

6)distribution_channel

Data Preprocessing: Encoding the categorical variables

7)reserved_room_type

8)assigned_room_type

9)deposit_type

10)customer_type

11)reservation_status

12)reservation_status_date

Training the model

Size of training data = 75%

Size of testing data = 25%

Accuracy using various algorithms

Logistic Regression

Training Accuracy : 0.9882194958188366

Testing Accuracy : 0.9888303959716183

KNN

Training Accuracy : 0.9882500152597204

Testing Accuracy : 0.983611810482948

Accuracy using various algorithms

Decision Tree classifier

Training Accuracy : 0.9882347555392785

Testing Accuracy : 0.9888761730373083

Bagging

Training Accuracy : 1.0

Testing Accuracy : 1.0

Conclusion

The highest accuracy in this problem is obtained using the bagging classifier.

Highest accuracy = 100%