

Hotel Reservation Cancellation Prediction

V. Sruthi
119cs0048

Contents

- Problem Statement
- About the Dataset
- Variables in Dataset
- Correlation matrix
- Data Preprocessing
- Accuracy using various algorithms
- Conclusion

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

is_canceled	1	0.29	0.017	0.0081	-0.0061	0.0018	0.025	0.06	0.005	-0.032	-0.085	0.11	-0.057	-0.14	-0.083	-0.021	0.054	0.048	-0.2	-0.23
lead_time	0.29	1	0.04	0.13	0.0023	0.086	0.17	0.12	-0.038	-0.021	-0.12	0.086	-0.074	0.00015	-0.07	0.15	0.17	-0.063	-0.12	-0.096
arrival_date_year	-0.017	0.04	1	-0.54	-0.00022	0.021	0.031	0.03	0.055	-0.013	0.01	-0.12	0.029	0.031	0.063	0.26	-0.056	0.2	-0.014	0.11
arrival_date_week_number	-0.0081	0.13	-0.54	1	0.067	0.018	0.016	0.026	0.0055	0.01	-0.03	0.036	-0.021	0.0055	-0.031	-0.077	0.023	0.076	0.0019	0.026
arrival_date_day_of_month	-0.0061	0.0023	0.00022	0.067	1	-0.016	-0.028	0.0016	0.015	0.00022	0.0061	-0.027	-0.0003	0.011	0.0015	0.045	0.023	0.03	0.0087	0.0031
stays_in_weekend_nights	-0.0018	0.086	0.021	0.018	-0.016	1	0.5	0.092	0.046	0.018	-0.087	-0.013	-0.043	0.063	0.14	0.067	-0.054	0.049	-0.019	0.073
stays_in_week_nights	-0.025	0.17	0.031	0.016	-0.028	0.5	1	0.093	0.044	0.02	-0.097	-0.014	-0.049	0.096	0.18	0.18	-0.002	0.065	-0.025	0.068
adults	0.06	0.12	0.03	0.026	-0.0016	0.092	0.093	1	0.03	0.018	-0.15	-0.0067	-0.11	-0.052	-0.036	0.21	-0.0083	0.23	0.015	0.12
children	-0.005	-0.038	0.055	0.0055	0.015	0.046	0.044	0.03	1	0.024	-0.033	-0.025	-0.021	0.049	0.041	0.031	-0.033	0.32	0.056	0.082
babies	-0.032	-0.021	-0.013	0.01	-0.00023	0.018	0.02	0.018	0.024	1	-0.0089	0.0075	0.0066	0.083	0.036	0.019	-0.011	0.029	0.037	0.098
is_repeated_guest	-0.085	-0.12	0.01	-0.03	-0.0061	-0.087	-0.097	-0.15	-0.033	-0.0089	1	0.082	0.42	0.012	0.032	-0.24	-0.022	-0.13	0.077	0.013
previous_cancellations	0.11	0.086	-0.12	0.036	-0.027	-0.013	-0.014	-0.0067	-0.025	-0.0075	0.082	1	0.15	-0.027	-0.012	-0.18	0.0059	-0.066	-0.018	-0.048
previous_bookings_not_canceled	-0.057	-0.074	0.029	-0.021	-0.0003	-0.043	-0.049	-0.11	-0.021	-0.0066	0.42	0.15	1	0.012	0.023	-0.21	-0.0094	0.072	0.048	0.038
booking_changes	-0.14	0.00015	0.031	0.0055	0.011	0.063	0.096	-0.052	0.049	0.083	0.012	-0.027	0.012	1	0.067	0.12	-0.012	0.02	0.066	0.053
agent	-0.083	-0.07	0.063	-0.031	0.0015	0.14	0.18	-0.036	0.041	0.036	0.032	-0.012	0.023	0.067	1	0.35	-0.055	-0.025	0.18	0.034
company	-0.021	0.15	0.26	-0.077	0.045	0.067	0.18	0.21	0.031	0.019	-0.24	-0.18	-0.21	0.12	0.35	1	0.00041	0.086	-0.013	-0.099
days_in_waiting_list	-0.054	0.17	-0.056	0.023	0.023	-0.054	-0.002	-0.0083	-0.033	-0.011	-0.022	0.0059	-0.0094	-0.012	-0.055	0.00041	1	-0.041	-0.031	-0.083
adr	0.048	-0.063	0.2	0.076	0.03	0.049	0.065	0.23	0.32	0.029	-0.13	-0.066	-0.072	0.02	-0.025	0.086	-0.041	1	0.057	0.17
required_car_parking_spaces	-0.2	-0.12	-0.014	-0.0019	0.0087	-0.019	-0.025	0.015	0.056	0.037	0.077	-0.018	0.048	0.066	0.18	-0.013	-0.031	0.057	1	0.083
total_of_special_requests	-0.23	-0.096	0.11	0.026	0.0031	0.073	0.068	0.12	0.082	0.098	0.013	-0.048	0.038	0.053	0.034	-0.099	-0.083	0.17	0.083	1
is_canceled		lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	agent	company	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests

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%