Report

Xinjie Qian

10/28/2022

Introduction

This project aims on hotel demand data, which is a dataset found on Kaggle and here is its link: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand. The dataset contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and the number of available parking spaces, among other things. It has 32 variables and approximately 119000 observations.

In this project, I will do analysis on predicting whether or not a booking is cancelled.

Data preprocessing and exploratory data analysis

I first examined the presence of missing data and non-standard expression of data. Variables which had more than 10% missing value were deleted.

Next I did some exploratory data analysis. Here are two plots for the City hotel vs Resort hotel:

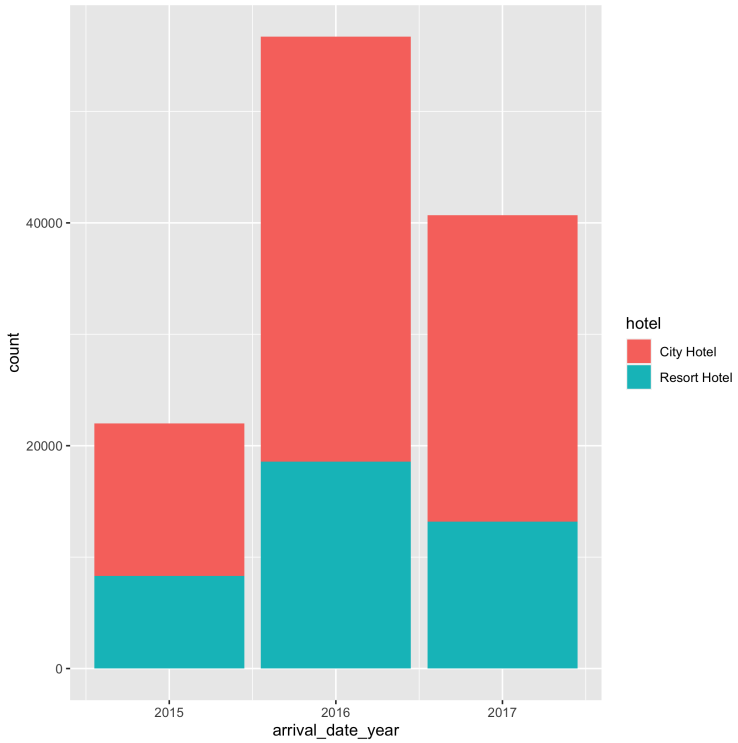

Figure 1: city hotel vs resort hotel for different years

From figure 1, we can see that the cancellation counts of city hotel are greater than resort hotel in all 3 years.

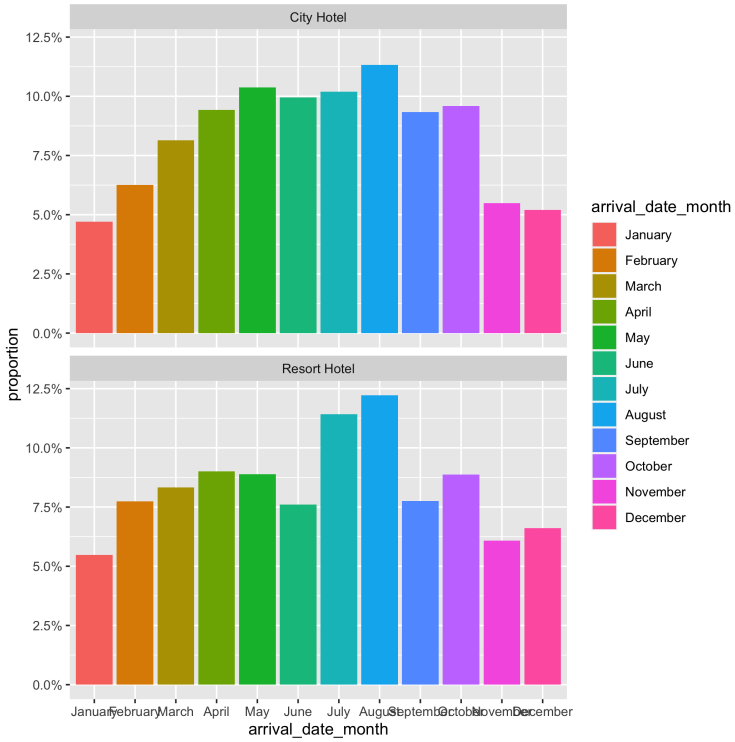

Figure 2: city hotel vs resort hotel for different month

From figure 2, we can see that July and August have high proportion of cancellation in both city hotel and resort hotel.

Since variable "country" had too many categories (more than 100), it's better to delete it. Also, some variables were correlated, for example, "arrival_date_month" and "arrival_date_week_number". It's enough to pick only one of them. Those variables which were not relevant to our aim were also deleted. Finally, we would use 17 variables to do the analysis.

Method

I used 2 methods to do the analysis, logistic regression and decision tree.

1.Logistic regression

The logistic regression is of the form:

$$p(x) = \frac{1}{1 + e^{-\beta X}}$$

where $p(x)$ is the probability of cancellation and β is the coefficient matrix and X is the covariate matrix.

I randomly selected 70% of the data as training set and the rest 30% of the data as testing set.

From table 1, we can see the main result of the logistic regression. Most of the variables are significant.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.814e+00	4.229e-02	-42.889	<2e-16
hotelResort Hotel	-1.077e-01	2.017e-02	-5.341	9.23e-08
lead_time	2.745e-03	9.856e-05	27.852	<2e-16
arrival_date_week_number	-5.592e-03	6.937e-04	-8.061	7.59e-16
stays_in_weekend_nights	5.348e-02	1.022e-02	5.233	1.67e-07
stays_in_week_nights	4.328e-02	5.354e-03	8.084	6.26e-16
adults	1.928e-01	1.938e-02	9.947	<2e-16
children	1.873e-01	2.165e-02	8.649	<2e-16
is_repeated_guest	-7.337e-01	9.514e-02	-7.713	1.23e-14
previous_cancellations	2.564e+00	7.099e-02	36.114	<2e-16
previous_bookings_not_canceled	-5.223e-01	3.189e-02	-16.377	<2e-16
booking_changes	-4.752e-01	1.855e-02	-25.619	<2e-16
deposit_typeNon Refund	5.320e+00	1.322e-01	40.229	<2e-16
deposit_typeRefundable	-1.107e-01	2.314e-01	-0.478	0.632
days_in_waiting_list	-4.404e-03	5.904e-04	-7.459	8.69e-14
adr	6.555e-03	2.125e-04	30.847	<2e-16
required_car_parking_spaces	-2.648e+03	9.223e+05	-0.003	0.998
total_of_special_requests	-4.601e-01	1.237e-02	-37.202	<2e-16

Table 1: Result of logistic regeression

Table 2 is the confusion matrix of the logistic regression. We can get that the accuracy of this model is 0.7766. The sensitivity is 0.9663 and the specificity is 0.4600. Although the accuracy is not bad in some degree, the false positive rate is too high.

		reference	
		0	1
prediction	0	21644	7245
	1	756	6171

Table 2: Confusion matrix of logistics regression

Table 3 is the variable importance of logistic regression. As shown in table 3, "lead_time", "previous_cancellations", "booking_changes", "deposit_typeNon Refund", "adr" (average daily rate) and "total_of_special_requests" are much more important than other variables. This result is consistent with what I get in the method of decision tree (we will see it later).

Table 3: Variable improtance of logistic regression

X	Overall
hotelResort Hotel	5.3413343
lead_time	27.8516665
arrival_date_week_number	8.0606785
stays_in_weekend_nights	5.2329031
stays_in_week_nights	8.0840675
adults	9.9470932
children	8.6494101
is_repeated_guest	7.7126270
previous_cancellations	36.1140045
previous_bookings_not_canceled	16.3772802
booking_changes	25.6188766
deposit_typeNon Refund	40.2288118
deposit_typeRefundable	0.4784540
days_in_waiting_list	7.4594732
adr	30.8467330
required_car_parking_spaces	0.0028713
total_of_special_requests	37.2016660

2. Decision tree

The second method I use is decision tree. As shown in Figure 3, for non-refundable deposit, the predicted result of cancellation is yes. For no deposit and refundable deposit, when the required car parking space is 0, the predicted result of cancellation is not. When the required car parking space is greater or equal to 1, then we need to check leadtime and previous cancellation.

From the deposit, leadtime and previous cancellations aspects I want to give the hotels two recommendations. The first is manage and hold long leadtime bookings. Hotel can send email to remind the customers who book the hotel much earlier before check-in and update hotel news with them to make the customers feel that they are valued and strengthen their impression of the hotel. Also, hotel can give earlier customers cash coupon for the next time booking after they actually check-out from the hotel to increase customer stickiness and decrease cancellation rate.

In terms of deposit and refund policy, hotel can set higher deposit and non-refund policy for last-minute bookings because if they choose to cancel the bookings, the time for the hotel to release vacancy rooms and gain new bookings is too short and the rooms may be wasted. Set non-refund deposit for customers during peak demanding seasons. No deposit for loyal customers who didn't cancel the reservation previously because they may plan their trips in relatively earlier time and keep stickiness and loyalty for the hotel. The probability of canceling the booking is lower for these customers so the hotel may not have loss.

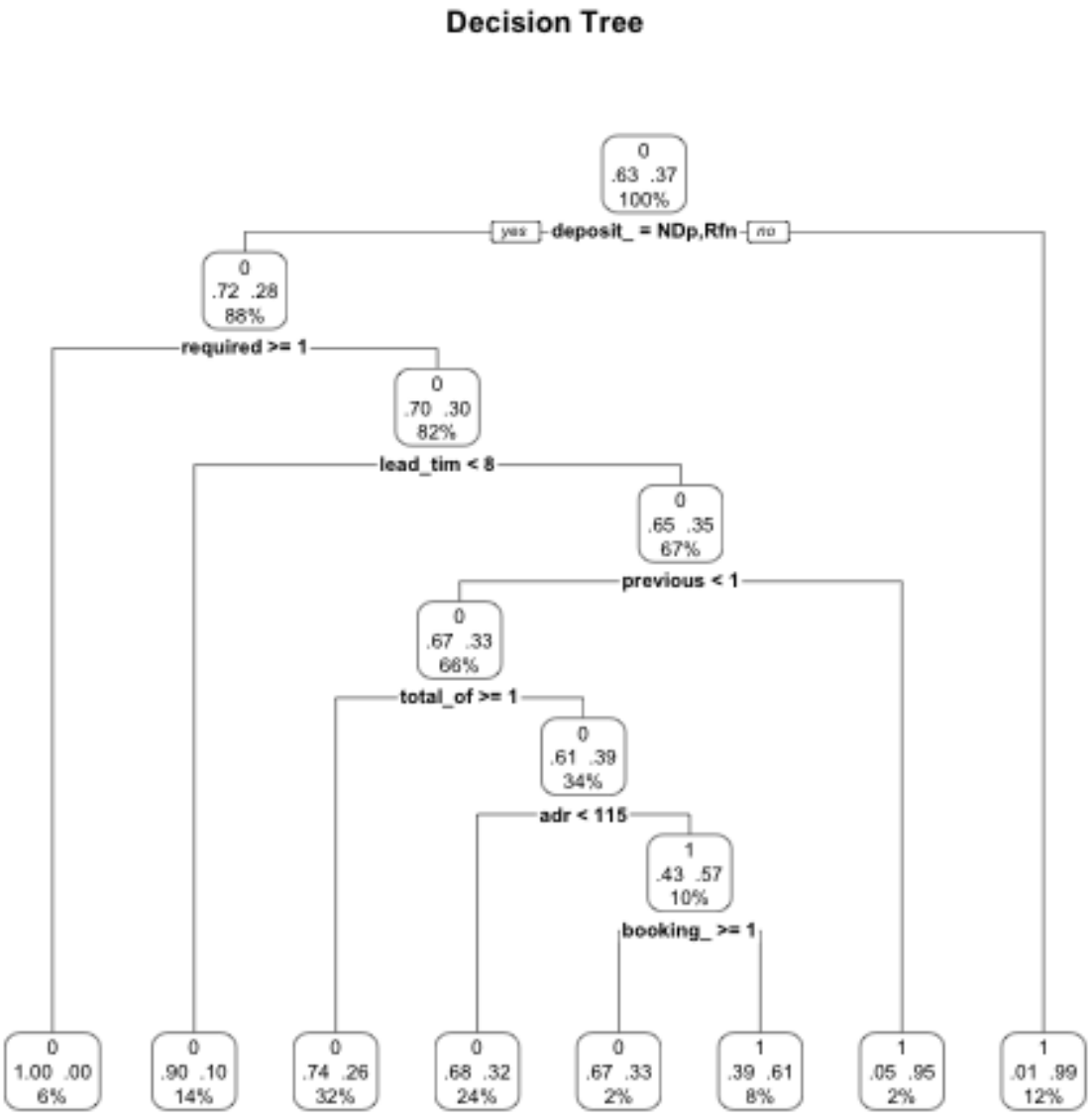

Figure 3: Decision tree

Table 4 is the confusion matrix of decision tree. From this table, we can get that the accuracy of this model is 0.7861. The false positive rate is still very high, which is very similar to that of the logistic regression model.

		actual	
		0	1
prediction	0	21187	6449
	1	1213	6967

Table 4: Confusion matrix of decision tree

Conclusion

I used logistic regression and decision model to do the analysis. Both of the models have the accuracy greater than 0.75, but the false positive rate is a bit high. Also according to the decision tree model, I give the hotels some recommendations that can decrease the cancellation rate.

Future work

Next, I plan to deal with the problem of high false positive rate. I will try some neural network methods and other classification methods.