MA678-Final Project

ChenxunLi

11/23/2020

Abstract

This paper explored the relevant data of Resort Hotel and City Hotel from 2015 to 2019, established a Multilevel logistic model to study the influencing factors of the cancellation rate with a prediction accuracy of 78% and AUC of 0.86.

Introduction

Background

The data set contains the information about 119k booking records for a city hotel and a resort hotel. I will use this data set to predict the cancellation rate in Multilevel logistic model.

Main variables Introduction

'is_canceled': represent whether the booking was canceled; 0 represents 'not canceled', 1 represents 'canceled'

'hotel': 'Resort Hotel' and 'City Hotel'

'lead_time': Number of days booked in advance (the arrival date - the booking date)

'adr': average Daily Rate (the sum of all lodging transactions / the total number of staying nights)

'adults': numbers of adults

'children': numbers of children

'babies': numbers of babies

'is_repeated_guest': represent whether the guest booked the this hotel before (1) or not(0)

'previous_cancellations': number of previous bookings cancelled by the guest before the current booking

'previous_bookings_not_canceled': number of previous bookings not cancelled by the guest before the current booking

'market_segment': In detail, "TA" means "Travel Agents" and "TO" means "Tour Operators"

'arrival_date_month': month of arrival date

'meal': type of meal booked

'country': country of origin.

'reserved_room_type': room type when guest reserved

'deposit_type': the type that customer made a deposit to guarantee the booking

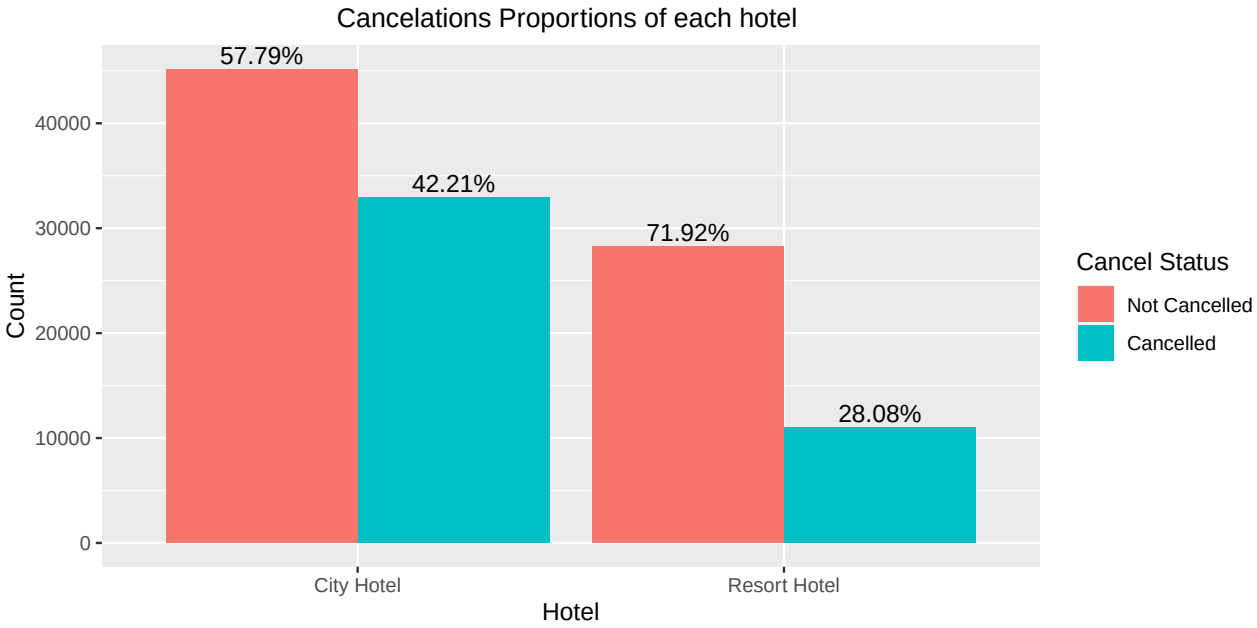
'customer_type': type of booking divided in to three categories

Data Processing

At first, I need to process the data because there are many NA and outliers in the data set. For example, there are 'undefined' in 'meal' variable, and 'undefined' and 'SC' all represent no meal, so I modify the "Undefined" to "SC". Besides, in 'adults','children' and 'babies', there are some outliers like 20 adults, 30 children, I remove these value.

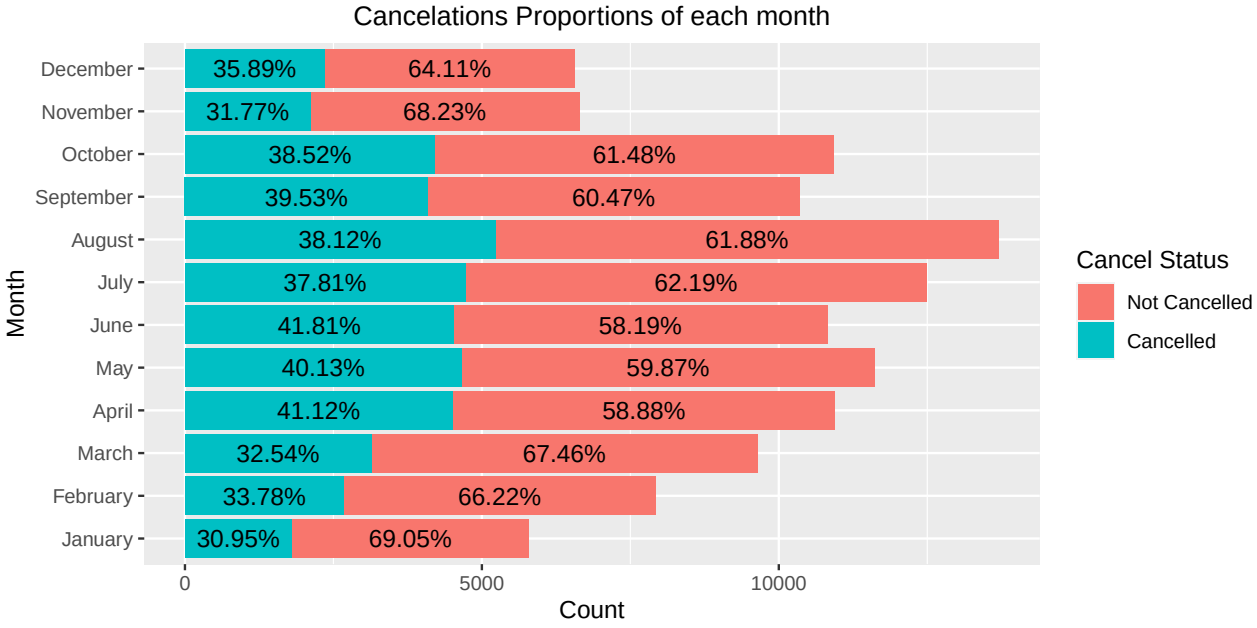
EDA

1.Which hotel have higher cancelations?



From the bar plot, we can obviously find that city hotel have a higher cancellation rate than the resort hotel.

2. Which month have the highest number of cancelations?



From the plot, we can find that the most number of booking is in August, and the booking is most possibly

canceled when the arrive date is in June.

Method

Choose the method

I would like to study the factors of cancellation, so my dependent variable is 'is_canceled'. 'is_canceled' is a binomial variable that 1 represents 'canceled' and 0 represents 'not canceled', so my first model is logistic model. However, after the check of the model, I thought the accuracy of prediction is so low and I found that many of the independent variables are structured by group. So, I choose multilevel logistic model at last.

After many attempts, I use this formula for model:

```
multilevel_logistic <- glmer(data=log_data,is_canceled~lead_time+adr+
 adults+children+babies+is_repeated_guest+previous_cancellations+
 previous_bookings_not_canceled+(1|hotel)+(1|market_segment)+
 (1|arrival_date_month)+(1|meal)+(1|country)+(1|reserved_room_type)+
 (1|deposit_type)+(1|customer_type), family = binomial)
```

What is Multilevel logistc Model

Multilevel logistic Model is a Multilevel Model applies to logistic regression and its coefficients are grouped into batches and a probability distribution is assigned to each batch.

Result

1.Fixed variable Interpreting

effect	term	estimate	std.error	statistic	p.value
fixed	(Intercept)	-1.4979228	1.3797038	-1.085684	0.2776187
fixed	lead_time	0.0059109	0.0001008	58.634196	0.0000000
fixed	adr	0.0033270	0.0002420	13.749482	0.0000000
fixed	adults	0.0866570	0.0181728	4.768512	0.0000019
fixed	children	0.0921184	0.0228010	4.040107	0.0000534
fixed	babies	-0.7470744	0.0932031	-8.015553	0.0000000
fixed	is_repeated_guest	-0.9499099	0.0844067	-11.253964	0.0000000
fixed	previous_cancellations	2.2323636	0.0624061	35.771578	0.0000000
fixed	previous_bookings_not_canceled	-0.4995697	0.0315958	-15.811278	0.0000000

Table 1:
The table above is the fixed effect of the Multilevel logistc Model.

From the fixed variable:

First of all, the each variable's p-value is so small except intercept, so the outcome is significant mostly.

Beside,I find that the effect of 'previous_cancellations' seem to be most important(2.23 increase with each increase of 1)

I also find that the effect of 'is_repeated_guest' seem to be important, that means if a guest is a repeated guest(is_repeated_guest=1), he/she will have 0.95 less than a 'not repeated guest'(is_repeated_guest=0).

'Babies' seems to be significant as well (0.75 decrease with each increase of 1), same with the 'previous_bookings_not_canceled'(0.5 decrease with each increase of 1)

2.Ramdom effects

Groups	Name	Variance	Std.Dev.
country	(Intercept)	0.874098	0.93493
arrival_date_month	(Intercept)	0.017313	0.13158
reserved_room_type	(Intercept)	0.009538	0.09766
market_segment	(Intercept)	0.332734	0.57683
customer_type	(Intercept)	0.254195	0.50418
meal	(Intercept)	0.061044	0.24707
deposit_type	(Intercept)	5.125269	2.26391
hotel	(Intercept)	0.086402	0.29394

Table 2:

The table above is the random effect of Multilevel logistc Model.

From the random effect, we can obviously find that the variance of 'deposit_type' is biggest, and that means the different type of deposit influence whether cancel or not most than other groups.

Besides, we can also find that the variance of 'reserved_room_type' is smallest, that means the reserved room type nearly has no effect on the cancellation.

So we check that by viewing the coefficient of 'deposit_type' and 'reserved_room_type'.

	(Intercept)	lead_time	adr	adults	children	babies
No Deposit	-3.127562	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
Non Refund	1.698615	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
Refundable	-2.966026	0.0059109	0.003327	0.086657	0.0921184	-0.7470744

Table 3:

The table above is the part coefficient of each level in group 'deposit_type'.

From the table, we can find that the intercept varies big, the intercept of 'Non Refund' is biggest and the intercept of 'No Deposit' is smallest.

Besides, we can see that the coefficient of fixed variable is the same.

	(Intercept)	lead_time	adr	adults	children	babies
A	-1.581806	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
B	-1.531106	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
C	-1.473641	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
D	-1.444413	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
E	-1.403917	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
F	-1.654307	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
G	-1.482127	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
H	-1.419178	0.0059109	0.003327	0.086657	0.0921184	-0.7470744
L	-1.490626	0.0059109	0.003327	0.086657	0.0921184	-0.7470744

Table 4:

The table above is the part coefficient of each level in group 'reserved_room_type'.

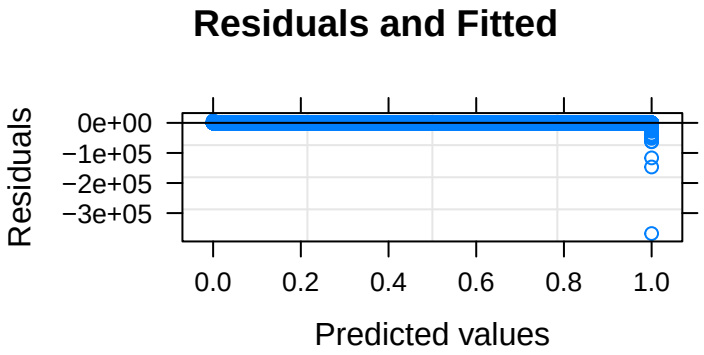
From this table, we can see few difference between each room type, so reserved_room_type has few effects on cancellation.

Besides, we can see that the coefficient of fixed variable is the same.

Discussion

Check the model

1.Residual Plot



From the plot, we can see that the residuals are all nearly zero except fitted = 1, so the model still need to be modified.

2.The accuracy of prediction

Beside, I also want to know whether this model can predict the cancellation, so I use 'predict()' and check the accuracy among these 119k rows. And the out come is below

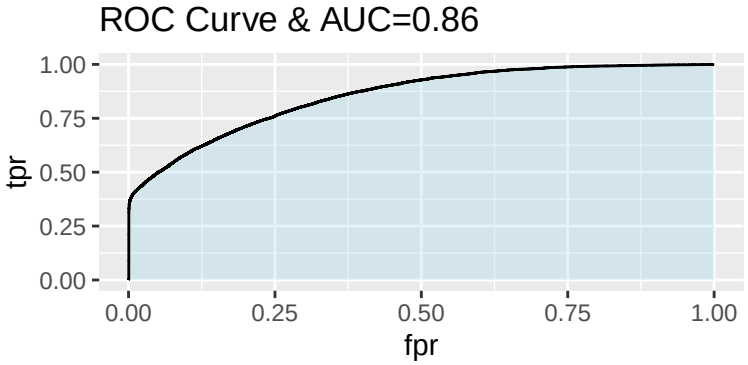
```
## [1] "The accuracy of prediction is  78.26%"
```

3.ROC curve and AUC

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

AUC (Area Under Curve) is defined as the area bounded by the coordinate axis under the ROC Curve. Obviously, the value of this Area will not be greater than 1. Since ROC curve is generally above the line y=x, AUC value ranges between 0.5 and 1. The closer the AUC is to 1.0, the higher the authenticity of the detection method is. When it is equal to 0.5, it has the lowest authenticity and no application value.

And I will plot the ROC curve and compute the AUC.



The AUC is 0.86, and plus the accuracy is 78.26%, so this model can basically be a binary classifier to predict cancellation.

Appendix

Classical logistic regression VS Multilevel logistc Model

1.Result comparason

```
model_glm <- glm(data=log_data,is_canceled~lead_time+adr+adults+
  children+babies+is_repeated_guest+previous_cancellations+
    previous_bookings_not_canceled, family = binomial)
summary(model_glm)
```

```
##
## Call:
## glm(formula = is_canceled ~ lead_time + adr + adults + children +
##     babies + is_repeated_guest + previous_cancellations + previous_bookings_not_canceled,
##     family = binomial, data = log_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.8826  -0.7354   1.2154   6.7282
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.518e+00  2.864e-02 -53.013   <2e-16 ***
## lead_time                       4.987e-03  6.454e-05  77.274   <2e-16 ***
## adr                             3.804e-03  1.554e-04  24.481   <2e-16 ***
## adults                         -7.370e-03  1.440e-02  -0.512    0.609
## children                       -2.583e-02  1.679e-02  -1.538    0.124
## babies                         -8.868e-01  8.784e-02 -10.096   <2e-16 ***
## is_repeated_guest              -1.103e+00  8.520e-02 -12.941   <2e-16 ***
## previous_cancellations          3.114e+00  5.496e-02  56.657   <2e-16 ***
## previous_bookings_not_canceled -6.720e-01  3.406e-02 -19.731   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155341  on 117424  degrees of freedom
## Residual deviance: 137447  on 117416  degrees of freedom
## AIC: 137465
##
## Number of Fisher Scoring iterations: 8
```

According to the classic logistic model, the 'children' and 'babies' are not significant.

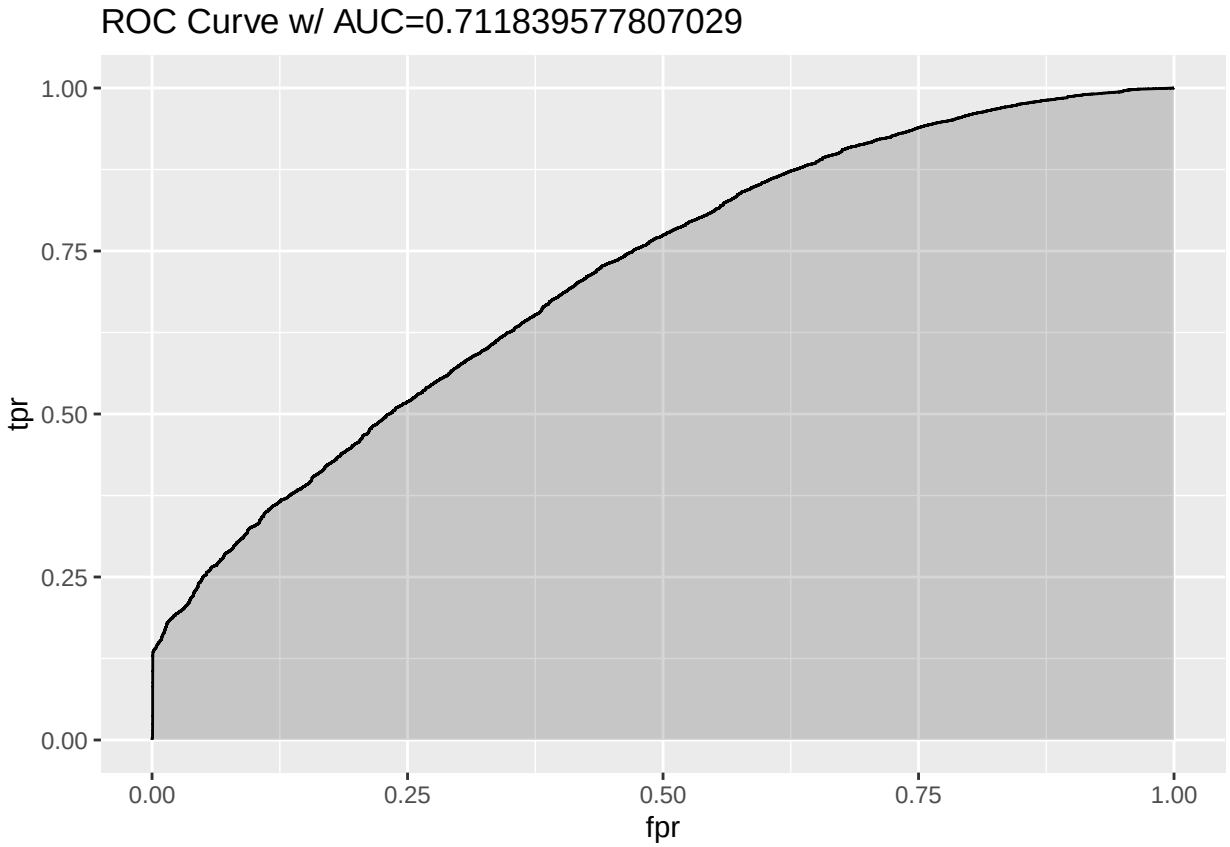
2.Predictions comparison

```
log_data$pred <- predict(model_glm,log_data,type="response")
log_data$pred_canceled <- ifelse(log_data$pred>=0.5,1,0)
accuracy <- percent(mean(log_data$is_canceled==log_data$pred_canceled),accuracy = 0.01)
print(paste("The accuracy of prediction is ",accuracy))
```

```
## [1] "The accuracy of prediction is  68.63%"
```

Classic logistic model can only have the accuracy of 68.6%, which is 9.4% lower than multilevel logistic model

3. AUC comparison



The AUC is 0.71, which is 0.15 lower than the multilevel logistic model.

4.AIC and BIC

```
model <- c("multilevel logistic model","classic logistic model")
AIC <- c(AIC(multilevel_logistic),AIC(model_glm))
BIC <- c(BIC(multilevel_logistic),BIC(model_glm))
data_frame(model,AIC,BIC)
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## # A tibble: 2 x 3
##   model                          AIC     BIC
##   <chr>                        <dbl>   <dbl>
## 1 multilevel logistic model 102318. 102483.
## 2 classic logistic model    137465. 137552.
```

We can see that multilevel_logistic has lower AIC and BIC.

Reference

1.kaggle notebooks

2.WikiPedia

3.R Documentation