Order now or later?
– Predicting the popularity of a listing on Airbnb in NYC

Mengyao Liu (ml7hc)

Oct. 27th, 2019

Keywords: business — internet — classification analysis — New York City — hotels and accommodations

1 Introduction

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. In a big city like New York City (NYC), which attracts millions of tourists every year[1], there are over 10,000 hosts posting over 45,000 house listings on Airbnb in 2019 alone[2]. Suppose you are a tourist who is looking for accommodation on Airbnb. You have several options in your wishlist but you need more time for further comparison or deeper investigation. Compared with an imperfect choice, you are more worried about your favorite being taken by someone else before you make up your mind. Thus you may want to know the popularity of this listing to decide whether to place the order soon or you could wait until later. Fortunately, thanks to Internet, we can find public dataset from Airbnb and help make a wise decision by learning from the history.

This report aims to predict whether a room/house on Airbnb is popular or not.

We use the dataset describing the listing activity and metrics in NYC, NY for 2019 from Kaggle[2]. Essentially the data file is a `csv` table. This table includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. This public dataset is part of Airbnb, and the original source can be found on this website.

The predictors we will use are `neighborhood_group`, `room_type`, `price`, `minimum_nights`, `number_of_reviews`, and `number_of_reviews_per_month`. `neighborhood_group` is a string of the neighborhood location name. `room_type` is a string of space type. The string will be one of the following three: Entire home/apt, Private room, Shared room. `price` is the price in dollars. `minimum_nights` is a number specifying the amount of nights minimum. The response will be two classes of popularity. We define "Popular" as `availability_365` < 122, where `availability_365` is the number of days when listing is available for booking. We define "Unpopular" as `availability_365` $\geqslant 122$.

2 Exploratory Data Analysis

2.1 Data Cleaning

There are 48,895 observations in total and there are 16 variables. First we check the missing data. It seems the columns that miss data are `name`, `host_name`, `last_review` and `review_per_month` as shown in Figure 1. `name` and `host_name` are not useful for the problem so we do not need to worry

about them. As to the 10,052 observations with missing `last_review` and `review_per_month`, after scrutinization it turns out that they all have `number_of_reviews = 0`. This means our dataset is quite clean and we do not have the problem of missing data.

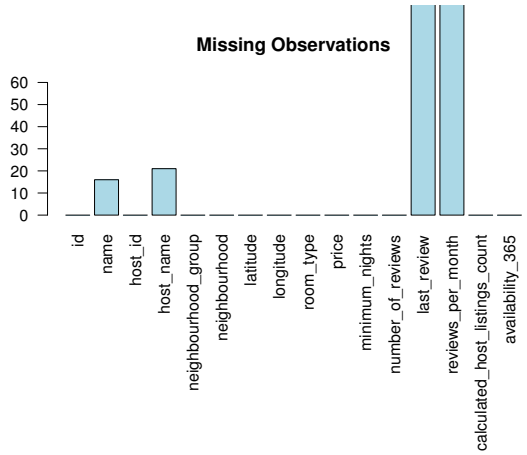

Figure 1

We have also found that there are 11 observations that have `price = 0`. This is unusual and we do not aim to investigate the reasons here. Thus we remove those 11 observations.

2.2 Exploratory Summary

In the sample of 48,884 observations, we have 30,632 observations in class "Popular" and 18,252 observations in class "Unpopular". To investigate the influence of different predictors on popularity, we plot the distribution of `price` (see Figure 2), `minimum_nights` (see Figure 3), `number_of_reviews` (see Figure 4), and `number_of_reviews_per_month` (see Figure 5) by `popularity`, respectively, where the right panel is a zoom-in plot of the left panel. We also plot the distribution of `availability_365` by `neighborhood_group` (see Figure 6) and `room_type` (see Figure 7). We made these plots to gain insight on the significance of each predictor to help determine whether to reduce any of them later as well as insight on the signs of the coefficients.

From those plots, we can tell that a room/house tends to be popular if it has lower `price`, `minimum_nights`, `number_of_reviews`, and `number_of_reviews_per_month`. Also there seems to be a difference in `popularity` between different `neighborhood_group` and `room_type`: listings in Brooklyn, Manhattan, and Queens are more popular than those in Bronx and Staten Island; entire homes/apts and private rooms are more popular than shared rooms.

3 Classification Model Building

Since the goal is to help decide whether a guest should place his order sooner than later, and we always hope to have more time to compare, we make an conservative assumption that a room/house is popular if it is not available more than two thirds of a year.

For this problem, we will adopt both logistic regression and linear discriminant analysis (LDA) with the binary response variable. In order to compare the two models, first we use the same predictors. Since discriminant analysis assumes the predictors come from a multivariate normal distribution and categorical predictors should not be used in discriminant analysis, we only use `price`, `minimum_nights`, `number_of_reviews`, and `number_of_reviews_per_month` as predictors for both models to begin with.

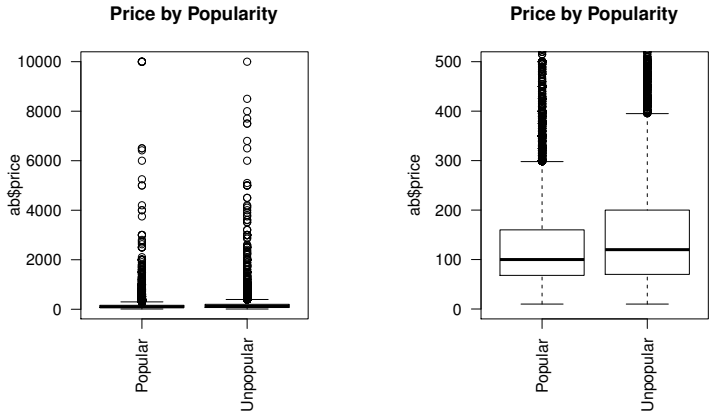

Figure 2

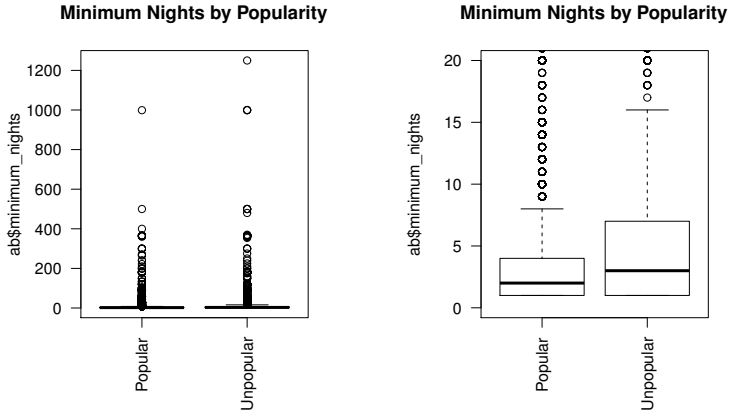

Figure 3

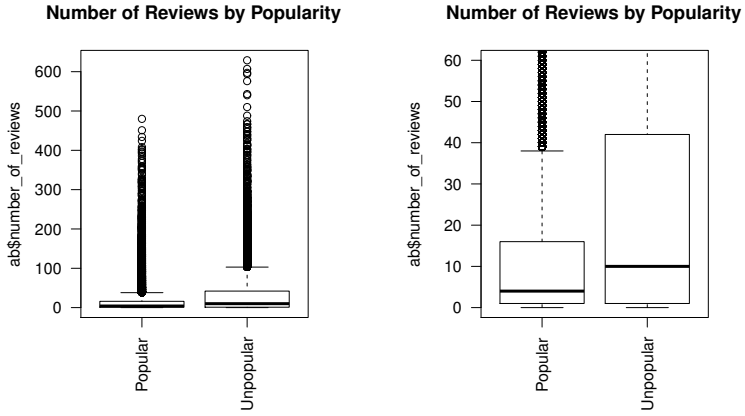

Figure 4

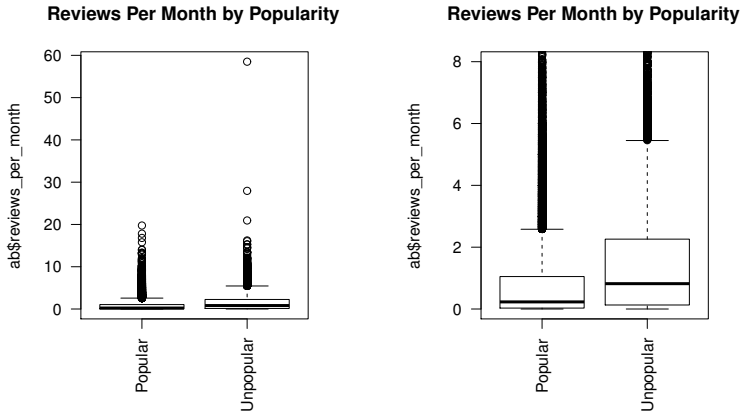

Figure 5

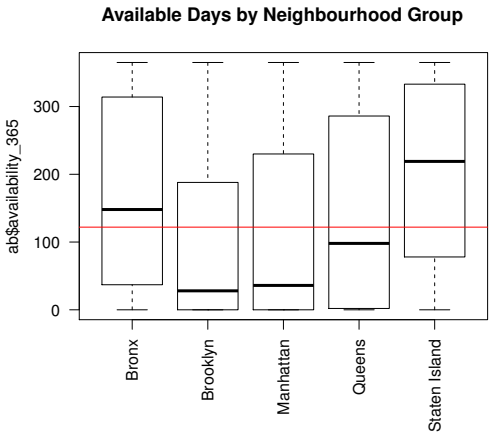

Figure 6

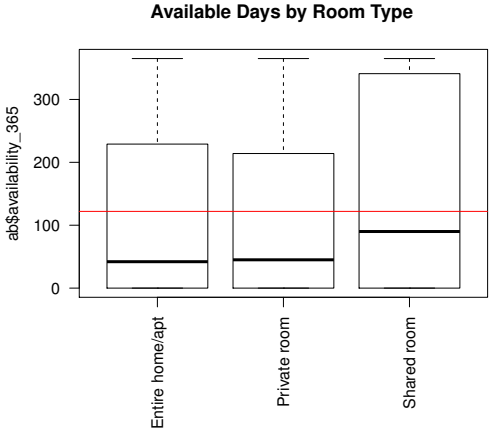

Figure 7

We randomly split the data into a training set (60% of the observations) and a test set (40% of the observations). Additionally, we set `number_of_reviews_per_month` of those observations with missing `number_of_reviews_per_month` to zero since they have `number_of_reviews = 0`. Note that in the models "Unpopular" is the positive class while "Popular" is the negative class.

3.1 Logistic Regression

We perform logistic regression on the training set and the results are shown in Figure 8. We see all the predictors are very significant. The coefficients are all positive, indicating that a room/house tends to be popular if it has lower `price`, `minimum_nights`, `number_of_reviews`, and `number_of_reviews_per_month`, which agrees with the plots in §2.2.

We test the model on the test set. The ROC curve is shown in Figure 9. The AUC is 0.712.

The test error rate using k fold cross-validation is 0.2076 with k = 5 and 0.2077 with k = 10.

```
Call:
glm(formula = popularity ~ price + minimum_nights + number_of_reviews +
    reviews_per_month, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.8642  -0.7488   1.1451   1.7819

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.4210079  0.0243910  -58.26   <2e-16 ***
price             0.0014972  0.0000900   16.64   <2e-16 ***
minimum_nights    0.0456363  0.0014042   32.50   <2e-16 ***
number_of_reviews 0.0071296  0.0003812   18.70   <2e-16 ***
reviews_per_month 0.1934322  0.0101932   18.98   <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: 38812  on 29329  degrees of freedom
Residual deviance: 35849  on 29325  degrees of freedom
AIC: 35859

Number of Fisher Scoring iterations: 6
```

Figure 8

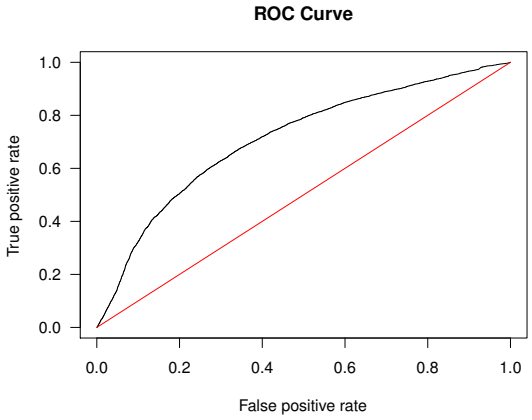

Figure 9

3.2 Linear Discriminant Analysis

We perform LDA on the training set and the results are shown in Figure 10. The coefficients are all positive, indicating that a room/house tends to be popular if it has lower `price`, `minimum_nights`, `number_of_reviews`, and `number_of_reviews_per_month`, which agrees with the plots in §2.2 and §3.1.

We test the model on the test set. The ROC curve is shown in Figure 11. The AUC is 0.708.

The test error rate using k fold cross-validation is 0.3530 with k = 5 and 0.3529 with k = 10.

```
Call:
lda(popularity ~ price + minimum_nights + number_of_reviews +
    reviews_per_month, data = train)

Prior probabilities of groups:
  Popular Unpopular
0.6248551 0.3751449

Group means:
             price minimum_nights number_of_reviews reviews_per_month
Popular   136.6309       4.930758          16.93190         0.8811038
Unpopular 179.1750      10.529855          33.67154         1.4366264

Coefficients of linear discriminants:
                           LD1
price              0.001593107
minimum_nights     0.026897001
number_of_reviews  0.012042442
reviews_per_month  0.265451408
```

Figure 10

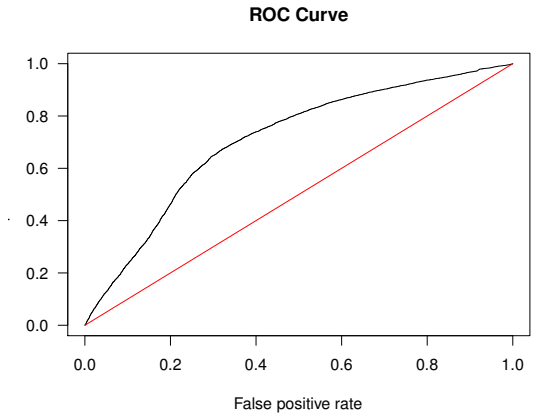

Figure 11

3.3 Improved Logistic Regression

In order to improve the logistic regression model, we add two categorical predictors `neighborhood_group` and `room_type`. We perform the improved logistic regression on the training set and the results are shown in Figure 12. We see all the predictors are very significant. The coefficients for `price`, `minimum_nights`, `number_of_reviews`, and `number_of_reviews_per_month` are all positive, which agrees with the results in §3.1. The coefficients for `neighborhood_group` of Brooklyn, Manhattan, and Queens are negative, indicating the rooms/houses in those three neighborhoods are more popular than those in Bronx and Staten Island. The coefficients for `room_type` of "private room" and "shared room" are positive, indicating "entire homes/apts" are more popular. The coefficient for

"shared room" is even larger than that of "private room", indicating "shared room" is the most unpopular room type. These results all agree with the plots in §2.2.

We test the model on the test set. The ROC curve is shown in Figure 13. The AUC is 0.738, which is better than that in §3.1.

The test error rate using k fold cross-validation is 0.2032 with k = 5 and 0.2032 with k = 10, also slightly better than those in §3.1.

```
Call:
glm(formula = popularity ~ price + minimum_nights + number_of_reviews +
    reviews_per_month + neighbourhood_group + room_type, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.8599  -0.7116   1.1377   1.8741

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -0.8001472  0.0864650  -9.254  < 2e-16 ***
price                           0.0021607  0.0001136  19.027  < 2e-16 ***
minimum_nights                  0.0490055  0.0014520  33.751  < 2e-16 ***
number_of_reviews               0.0077555  0.0003859  20.095  < 2e-16 ***
reviews_per_month               0.1768877  0.0103267  17.129  < 2e-16 ***
neighbourhood_groupBrooklyn    -0.9623660  0.0837381 -11.493  < 2e-16 ***
neighbourhood_groupManhattan   -0.9366377  0.0841933 -11.125  < 2e-16 ***
neighbourhood_groupQueens      -0.3442836  0.0886613  -3.883 0.000103 ***
neighbourhood_groupStaten Island  0.3206833  0.1661349   1.930 0.053575 .
room_typePrivate room           0.1943730  0.0294764   6.594 4.28e-11 ***
room_typeShared room            0.7853235  0.0830221   9.459  < 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: 38812  on 29329  degrees of freedom
Residual deviance: 35304  on 29319  degrees of freedom
AIC: 35326

Number of Fisher Scoring iterations: 6
```

Figure 12

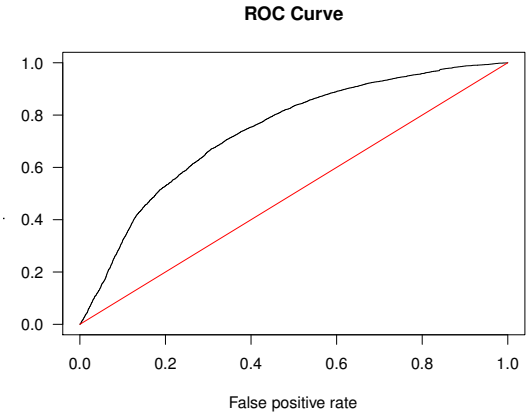

Figure 13

4 Conclusions

We have built two models using logistic regression and LDA respectively to predict the popularity of a listing on Airbnb in NYC given its `neighborhood_group`, `room_type`, `price`, `minimum_nights`,

number_of_reviews, and number_of_reviews_per_month. The ROCs derived from the validation set method are pretty good with an AUC of ∼0.7. The test error rates derived from k-fold cross-validation are also acceptable, with ∼20% for logistic regression and ∼35% for LDA. When adding categorical predictors, the logistic regression model gets improved and the test error rates derived from k-fold cross-validation drop for ∼0.4%.

Both models return results that basically agree with our intuition – a cheaper (lower price), more private (entire home/apt and private room rather than shared room) room/house requiring fewer minimum nights (lower minimum_nights) tends to be mote popular and a guest should not hesitate too long before placing the order. An interesting result is that unpopular rooms/houses tend to have more reviews (higher number_of_reviews and number_of_reviews_per_month), which implies that people may be more keen to post complaints than compliments.

We had some bugs when we first tried to set "Unpopular" as the reference class for the dummy coding. It resulted in a ROC below the random guess. We picked a random threshold value and calculated the true positive rate and the false positive rate, respectively, and found the derived data point did not lie on the ROC. This indicated somehow the ROC plotted was flipped and using relevel introduced some error. Thus to avoid introducing underlying errors, we use the default classification of popularity, i.e., "Unpopular" being the positive class and "Popular" being the negative class. Comparing the results with and without releveling the classes, we found all the coefficients have the same absolute values with only the opposite signs and the two AUCs with and without releveling summed to 1, which means our suspect is correct and the results are consistent.

Another challenge is that we got a warning message saying glm.fit: fitted probabilities numerically 0 or 1 occurred when building the logistic regression model. We checked the fitted probability of the model and found the model returned a probability of 1 for 8 observations in the training set. In order to understand how it occurred, we further examined the 8 observations and found one of them has very high price, one has very high minimum_nights, and two have very high number_of_reviews. While the 4-dimensional predictor space leads to some difficulty in visualization, we tried plotting a 3d image with price, minimum_nights and number_of_reviews. We saw there was some space with very high minimum_nights and number_of_reviews where there are only "Unpopular" observations. Thus one possibility is that in a certain region in the 4-dimensional predictor space the two classes are close to completely separated.

Reference

[1] https://nycfuture.org/research/destination-new-york
[2] https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data