Midterm Report

Zhengyu Long(zl478), Han Dong(hd324)

1.Introduction

Airbnb is an online marketplace that enables people to list, find, and then rent vacation homes for a processing fee. With the data including all the listings information of Airbnb at San Francisco, we are trying to answer the question of “what price would a Airbnb host set to maximize his profit”. In this project, we aim to develop an price recommendation tool to maximize his/her profit with accommodations provided.

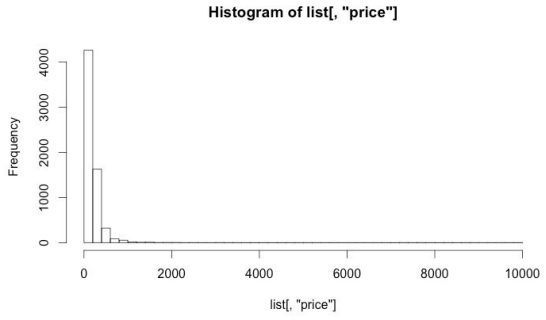
2.Data cleaning

For simplification purposes, we take a look at the simplified version of the airbnb data which includes 16 columns and 8606 rows. The main purpose of this analysis is determine the best price for the owner and thus maximize his/her total revenue. In order to achieve this, we first analyze features that may have certain impact on price. Some features, however, can be deemed unrelated to the price by intuition and some features are highly correlated with other features. In order to better understanding the model we build, we need to clean the data before fitting any models.

3.Descriptive Statistics and feature selection

3.1 Price and models

One of the key response for our pricing model is the price provided by host. To discover how current host are pricing their accommodations, our study shown as follows:




```
summary(list[,"price"])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.0   100.0   150.0   201.7   230.0 10000.0
```

As shown in Figure 3.1 The price of airbnb ranges from 1$ to 1000$ per night, with a median of 150. We may see that some of the hosts set their price quite randomly, giving the price range a wide standard deviation. In order to identify key features for pricing recommendation, we may need to pick the price that are set “reasonably” by the hosts as a response for us to identify the key impact.

To do that, we consider only the following two “meaningful” scenarios:

Model 1 - consider the total annual revenue ($price set by host × days rented in a 365-day period) as the response, in order to exclude those accommodations that are not sold to customers

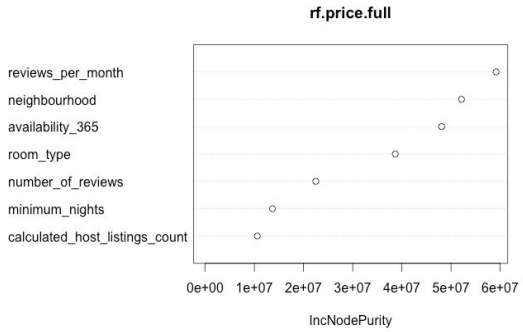
Model 2 - consider the price as “weighted”. As the response, the price of accommodation that has been rented more (having less availabilities) is considered more “reasonable” and weighs more.

3.2 Feature selection

After data cleanup, currently, the features includes: *"neighbourhood", "latitude", "longitude", "room_type", "minimum_nights", "number_of_reviews","last_review", "reviews_per_month", "calculated_host_listings_count", "availability_365".*

Intuitively, the feature *neighborhood* is redundant with “*latitude*” and “*longitude*” information, so for simplicity, we just use *"neighbourhood" in the feature.*

Some preliminary analysis are carried out to better understanding the features. Since the features contains both categorical and numeric values, we decide to use random forest model first, result as follows:



The importance plot shows how much each feature will help to influence node purity. As shown. Number of reviews, the location (neighbourhood or longitude and latitude), days rented (availabilities) and room-type are the four most important feature for response.

A different perspective from subset selection method, where all exhaustive, forward and backward selection showed room_type as the most important feature for price. The p-value of General linear regression methods also confirms our implication, saying mini_nights not significantly-important and longitude a less an important feature than latitude.

4.Overfit

(1)Training and test set

We split the data into training set and test set in order to test how well our current model does on out-of -sample data and also to do cross validation.

(2)Lasso and ridge

To further prevent overfitting, we fit two regularized regression models using ridge regression and Lasso. The best regularization parameters (lambda) are found using cross validation, and our result shows that the LASSO regression gives a smaller Least square error.

Since Lasso is also a feature selection method (sparse-tendency) and from the result of Lasso, we can also get important features from LASSO, which confirms our previous conclusion.

5. Conclusion

Our work so far have successfully narrowed down the scope of features we want to put in the pricing model, and identified that *room_type*, *location*, *availability* and *review_number* are the most important feature that influence a “reasonable” accommodation price.

6. Future work

6.1 Further discussion on room_type

As we found *room_type* is quite important by selection methods,we want to further see the details in it, including balcony, bathroom, bedroom and so on, to see each of their importance correspondingly on the price. These information can be found in the more detailed *listing.csv* airbnb data, which is more messy and provides more detailed explanation.

6.2 Further discussion on revenue

As of now, we are clear about the features that may affect price. While these prices are set by hosts themselves, we still aim to discover what price is “the best price” for them. It is suitable here to define “best price” to prevent ambiguity as the price that maximizes total revenue. Higher price may lead to less total nights booked while lower price may increase the total night booked. In order to adjust for this trade-off, we will fit a model with number of nights booked as the response variable and define the total revenue as price $\times$ total nights booked. The price corresponding to the maximum revenue is the best price.

6.3 Further discussion on availabilities

Another interesting topic to work is the availabilities. If time allowed, with the pricing recommendation model built, we will also try to make the model more flexible: when given a price set by host, he/she can figure out how many days will the accommodation be rented. We hope this model could be a dynamic decision-making tool for these hosts.