Prediction of Airbnb Listings with High Booking Rate

UNIVERSITY OF
MARYLAND
ROBERT H. SMITH
SCHOOL OF BUSINESS

BUDT758T-Final Project Report

Group 8

Xinyue Shi
Lei Guo
Tianhang Xu
Xin Tan

Executive Summary

Background

In recent years, Airbnb has been more and more established as a platform which operates an online marketplace and hospitality service for people to lease or rent short-term lodging. However, currently there are no clear indicators for house owners to improve their listing in order to get a high booking rate. Here we serve to analyze the data we collect from Airbnb and provide potential significant factors for the users in terms of the relationship between their listings and the booking rate. We went out and examine a dataset with 69 variables, some of the potential key factors are price, cleaning fee and availability etc., and we seek to predict a listing to determine if it would receive a high booking rate or not.

Analysis Result and Recommendations

This report was commissioned to use collected data to examine significant factors that influence ratings of the Airbnb listing and recommend house owners registered on the platform for raising their booking rate on Airbnb.

After analyzing the data through several models including logistic regression, ridge regression, LASSO regression, neural networks and random forest model. Our best result was achieved by random forest model with a prediction accuracy of 84.78047%, from the importance report we achieved from our model, we found out several variables that were more significant variables in predicting whether a listing can receive a high booking rate or not. They are price, minimum nights, house rules and host response rate. Our model is to predict whether a listing could receive a high booking rate or not for both Airbnb and the user.
Therefore, the function of our model is that:

1. Whenever there is a new listing on the Airbnb, our model can predict if that listing could be popular and receives a high booking rate. If one listing is predicted that it could not receive a high booking rate, we can then send out a notification to the user suggesting him or her to revise the information on the listing, mainly focusing on the price, cleaning fee and availability in order to become more popular.
2. Whenever there is a new listing on the Airbnb, after our model make a prediction on whether the listing could receive a high booking rate or not, the result would be sent to Airbnb so that they would have an idea to promote certain listings that we predicted to be successful. This would help Airbnb to promote their products and provide more options for the customers as well.
3. Although our model could not provide details on the relationship between the factors and the target variable in terms of how to improve a listing or what actions could be done to increase the booking rate of a listing, we could at least identify some factors that are more significant so we can advise the users to pay more attention to these factors when they post a listing, giving them a rough starting guide.

Technical Summary

The goal of our project was to predict Airbnb listings with high booking rate. In order to achieve this goal, we separated our task into four main parts: data cleaning, feature engineering, feature selection, and modeling.

Data Cleaning

The main tasks we did in our data cleaning part was to deal with missing values and wrong values. Before we started our job, we eliminated all the text variables because the missing values or wrong values of this type of variable was too hard to deal with.
In terms of the processing of missing values, we divided all the other variables into two groups: numerical variables and categorical variables. To deal with numerical missing values, we generated histograms for each variable to identify how each numerical value was distributed within that specific variable. If the shape of the plot is similar to a normal distribution, we replaced missing values with mean value. If the shape is irregular, we replaced missing values with median. If one value is very outstanding in the plot, we replaced missing values with this particular value. To deal with categorical variables, we generally replaced missing values with the most frequent group. For a couple of variables such as "host_response_rate", we identified that there were too many missing values (up to 70% or more) and therefore we just removed these variables from our analysis. For some specific variables, we replaced missing values with right values that we extrapolated by directly eyeballing their highly correlated variables. For example, to deal with missing values in the "state", we just eyeballed their corresponding street address information and found out that most missing fields could be replaced by "CA". Lastly, for some categorical variables, we considered missing values as a single group because these missing values took over a great proportion of observations. For example, for "host response time", we considered missing values as a single group and named it "Not Given".
In terms of the processing of wrong values, we identified that for some specific rows, the order of each feature was completely messed up so we just dropped these 12 rows. We also considered some outliers as wrong values. For example, in the "country" column, there were only two fields not having the value "US" so we just completely dropped these two rows. For the other numerical outliers, we usually replaced these outliers with the most reasonable values which are near to them.

Feature Engineering

In order to extract more information from existing data, we both did feature transformation and feature creation. In terms of feature transformation, we converted two date type variables into numerical variables by calculating the number of days from that date to current date. For some categorical variables, we incorporate some groups into a larger group to make sure that each group accounts for a decent proportion of observations. For example, in the processing of "cancellation policy" column, we binned three groups into a larger "strict" group because these three groups have very few observations.

In terms of feature creation, we created five new variables based on the existing data: availability_30_to_60, availability_60_to_90, availability_90_to_365, beds_per_room, and period difference. For the first three new variables, we subtracted one available days by the other to get the available days within a time period. We derived beds_per_room by simply divide “beds” by “bedsrooms+1” because we thought this was also an important factor that could be considered by the customers. For the last variable, we subtracted host_since by first review and therefore we got the period between two dates.

Despite those numerical variables and categorical variables, we also spot numerous meaningful text variables including access, description and transit etc. In order to excavate the potential value out of these variables, we decided to perform several text mining techniques to utilize this type of information. However, after our investigation of the text data, we have the assumption that these data might not be influential in our predictions since most of them are house holder’s own description or opinion of the houses they are listing, which is unlikely to contain negative information, and the topics would be very similar with each other since they are essentially talking about the same thing. Our assumption was proved after using sentiment analysis and topic modeling techniques. After necessary text preprocessing procedures were performed, the result of the sentiment analysis shows that majority of the text data in description is labeled while other are neutral. Our LDA method was not effective as well, we set the k = 2, meaning that we want the model to divide our text data into two topics, but the result of the words selected for two topics shows no strong support to distinguish these two topics. We applied both text mining methods to all our text variables and used these results as additional variables to rerun the previous models, the result had little to no improvement.

Our last attempt to utilize these text data was to find out the top 10 most appeared words in our text and investigate if these words’ appearances is related to the high booking rate. We generate a new categorical variable that is 1 if a line has any of the top most appeared words and 0 if it doesn’t have any of them and used it as an additional factor. However, it worsened our prediction accuracy of our models.

Feature Selection

Feature selection is a challenging and crucial task in the statistics modeling field, it can help to optimize model process. Before running our model, we used stepwise selection and LASSO regression to select the subset of predictors that do the best at maximizing accuracy.

First, we selected all numeric variables and converted high_booking_rate to numeric, then we generated LASSO plot. (**Appendix, Exhibit A**)

Based on this plot, we discovered that beds_per_room,accommodates, amenities, guests_included,host_verifications,host_response_rate,availability_30, beds, bathrooms, bedrooms, were potentially important predictors after running LASSO regression.

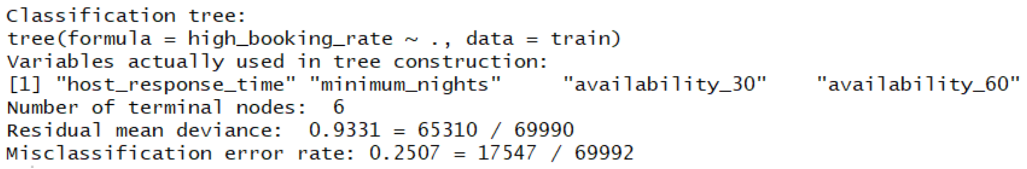
Variables selected through stepwise procedure are: host_response_time, minimum_nights, host_is_superhost, city_name, instant_bookable, cleaning_fee, amenities, room_type, price, host_since, house_rules, host_listings_count, cancellation_policy, availability_30, availability_90, is_business_travel_ready, guests_included, first_review, property_type, availability_365, monthly_price, extra_people, bathrooms, accommodates, bedrooms, weekly_price, requires_license, bed_type, host_verifications, require_guest_phone_varification and beds.

After comparing the results, we can see that both two methods selected accommodates, amenities, guests_included, host_verifications, availability_30, beds, bathrooms, bedrooms as their important variables. Besides eliminating data that are less related, these variables also provided us with insights concerning what Airbnb users care about most when they are looking for a residency.

Modeling

Classification Tree Model

Our goal is to predict a factor variable, high_booking_rate, and classification tree is a powerful tool for classification, after running classification tree model with the full variables to predict booking rate in testing dataset, our tree proved to be:



```
Classification tree:
tree(formula = high_booking_rate ~ ., data = train)
Variables actually used in tree construction:
[1] "host_response_time" "minimum_nights"     "availability_30"    "availability_60"
Number of terminal nodes:  6
Residual mean deviance:  0.9331 = 65310 / 69990
Misclassification error rate: 0.2507 = 17547 / 69992
```

As the output above shows, the tree has 6 terminal nodes, in the tree, we have variables such as host_response_time, minimum_nights, availability_30, and availability_60. And then we tried value, 2,3,4,5, as well as the full tree on our validation dataset, and found out that tree with 6 terminal nodes is our best tree. Then we applied final tree to testing dataset and get cutoff of 0.51. Finally, we ended up with getting accuracy of 0.7514 for testing dataset.

Logistic Regression Model

We also explored a traditional classification model: logistic regression. In order to get the highest accuracy, we ran logistic regression twice, once with all variables and another time with the variables selected from feature selection process. Because the selection of cutoffs can also influence the accuracy of logistic regression model, to eliminate this effect, we plotted line charts indicating the relationship between cutoff and accuracy. The results for these two logistic models indicated that the model will all variables obtains a higher accuracy which is 79.74%
According to the outputs, except variable beds, other variables selected through feature selection process all appeared to be significant for the best logistic model. However, some significant variables such as host_response_time and cancellation_policy were ignored by LASSO. Although the accuracy of the best logistic model is higher than the baseline which is approximately 75%, the accuracy was much lower than other models we tested. Therefore, we do not recommend applying a logistic regression towards this problem. (**For model details, see Appendix, Exhibit B**)

kNN Model

We also performed kNN approach on our numerical variables in our dataset since kNN is another primarily used model for classification problems. In order to achieve the best result from this model, we selected 8 different k values (3,5,7,9,11,13,15,17) and compare the accuracy for each of the k. Our results show that the model has the best accuracy when k = 15, with an accuracy of 73.58%. The result is acceptable but not satisfying, so we decided to move on to other models for this task.

Random Forest Model

After running some classic models that we learned from class, we decided to use an ensemble learning method to improve our classification accuracy so we chose to build random forest model. Instead of putting our selected variables into the model, we chose to put all the variables that we had into the model to let the tree have more options to do the splitting. After splitting the training data to 70% and 30%, the model built with 70% data gave an accuracy of 84.40% on the 30% validation data which was the best among all the results we had gained.
To tune the parameters for this model, we wrote a for loop in R and tried different value of mtry. Finally, we discovered that when mrty=12, the performance of the model reached the highest point. When we were predicting the testing data labels, we tried to use 100% data from training data with the best mtry to build the model in order to check whether the model built with full data could be more solid in the purpose of prediction. The prediction accuracy for the testing data turned out to be 84.78047% which fully met our expectation.

Appendix

Exhibit A:

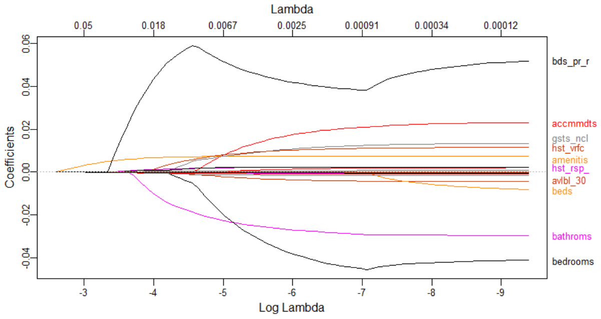
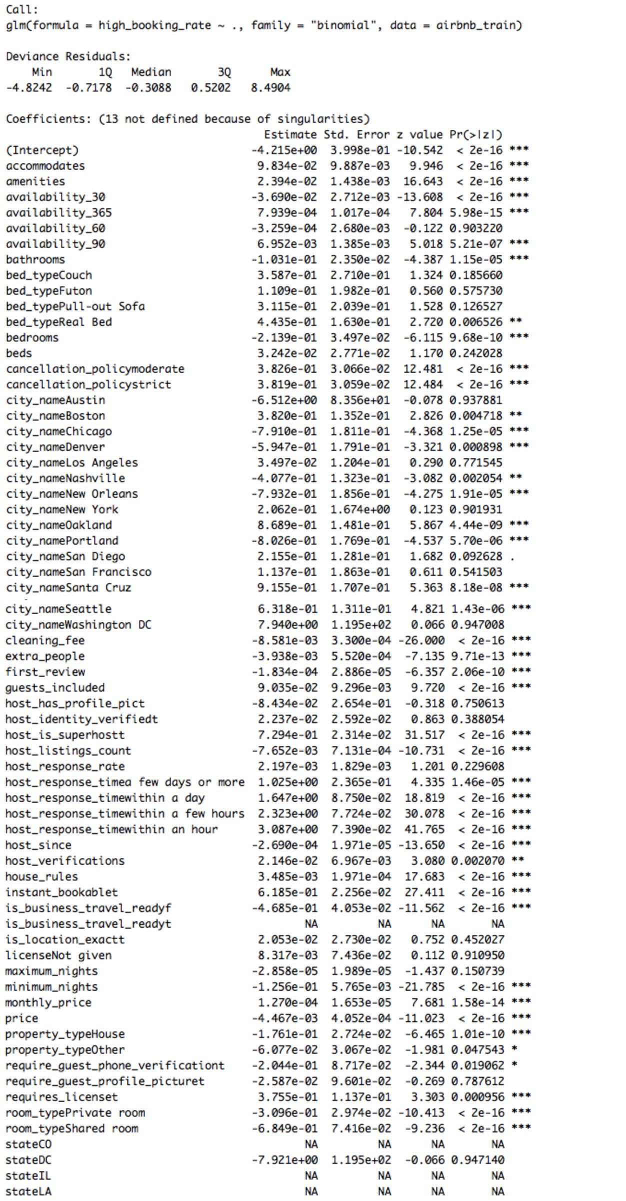


Exhibit B:

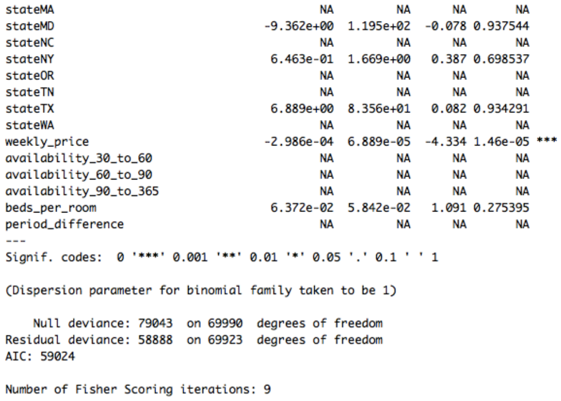


```
Call:
glm(formula = high_booking_rate ~ ., family = "binomial", data = airbnb_train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.8242  -0.7178  -0.3088   0.5202   8.4904  

Coefficients: (13 not defined because of singularities)
                                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)                           -4.215e+00  3.998e-01 -10.542  < 2e-16 ***
accommodates                           9.834e-02  9.887e-03   9.946  < 2e-16 ***
amenities                              2.394e-02  1.438e-03  16.643  < 2e-16 ***
availability_30                       -3.690e-02  2.712e-03 -13.608  < 2e-16 ***
availability_365                       7.939e-04  1.017e-04   7.804 5.98e-15 ***
availability_60                       -3.259e-04  2.680e-03  -0.122 0.903220    
availability_90                        6.952e-03  1.385e-03   5.018 5.21e-07 ***
bathrooms                             -1.031e-01  2.350e-02  -4.387 1.15e-05 ***
bed_typeCouch                          3.587e-01  2.710e-01   1.324 0.185660    
bed_typeFuton                          1.109e-01  1.982e-01   0.560 0.575730    
bed_typePull-out Sofa                  3.115e-01  2.039e-01   1.528 0.126527    
bed_typeReal Bed                       4.435e-01  1.630e-01   2.720 0.006526 ** 
bedrooms                              -2.139e-01  3.497e-02  -6.115 9.68e-10 ***
beds                                   3.242e-02  2.771e-02   1.170 0.242028    
cancellation_policymoderate            3.826e-01  3.066e-02  12.481  < 2e-16 ***
cancellation_policystrict              3.819e-01  3.059e-02  12.484  < 2e-16 ***
city_nameAustin                       -6.512e+00  8.356e+01  -0.078 0.937881    
city_nameBoston                        3.820e-01  1.352e-01   2.826 0.004718 ** 
city_nameChicago                      -7.910e-01  1.811e-01  -4.368 1.25e-05 ***
city_nameDenver                       -5.947e-01  1.791e-01  -3.321 0.000898 ***
city_nameLos Angeles                   3.497e-02  1.204e-01   0.290 0.771545    
city_nameNashville                    -4.077e-01  1.323e-01  -3.082 0.002054 ** 
city_nameNew Orleans                  -7.932e-01  1.856e-01  -4.275 1.91e-05 ***
city_nameNew York                      2.062e-01  1.674e+00   0.123 0.901931    
city_nameOakland                       8.689e-01  1.481e-01   5.867 4.44e-09 ***
city_namePortland                     -8.026e-01  1.769e-01  -4.537 5.70e-06 ***
city_nameSan Diego                     2.155e-01  1.281e-01   1.682 0.092628 .  
city_nameSan Francisco                 1.137e-01  1.863e-01   0.611 0.541503    
city_nameSanta Cruz                    9.155e-01  1.707e-01   5.363 8.18e-08 ***

city_nameSeattle                       6.318e-01  1.311e-01   4.821 1.43e-06 ***
city_nameWashington DC                 7.940e+00  1.195e+02   0.066 0.947008    
cleaning_fee                          -8.581e-03  3.300e-04 -26.000  < 2e-16 ***
extra_people                          -3.938e-03  5.520e-04  -7.135 9.71e-13 ***
first_review                          -1.834e-04  2.886e-05  -6.357 2.06e-10 ***
guests_included                        9.035e-02  9.296e-03   9.720  < 2e-16 ***
host_has_profile_pict                 -8.434e-02  2.654e-01  -0.318 0.750613    
host_identity_verifiedt                2.237e-02  2.592e-02   0.863 0.388054    
host_is_superhostt                     7.294e-01  2.314e-02  31.517  < 2e-16 ***
host_listings_count                   -7.652e-03  7.131e-04 -10.731  < 2e-16 ***
host_response_rate                     2.197e-03  1.829e-03   1.201 0.229608    
host_response_timea few days or more  1.025e+00  2.365e-01   4.335 1.46e-05 ***
host_response_timewithin a day         1.647e+00  8.750e-02  18.819  < 2e-16 ***
host_response_timewithin a few hours  2.323e+00  7.724e-02  30.078  < 2e-16 ***
host_response_timewithin an hour       3.087e+00  7.390e-02  41.765  < 2e-16 ***
host_since                            -2.690e-04  1.971e-05 -13.650  < 2e-16 ***
host_verifications                     2.146e-02  6.967e-03   3.080 0.002070 ** 
house_rules                            3.485e-03  1.971e-04  17.683  < 2e-16 ***
instant_bookablet                      6.185e-01  2.256e-02  27.411  < 2e-16 ***
is_business_travel_readyf             -4.685e-01  4.053e-02 -11.562  < 2e-16 ***
is_business_travel_readyt                     NA         NA      NA       NA    
is_location_exactt                     2.053e-02  2.730e-02   0.752 0.452027    
licenseNot given                       8.317e-03  7.436e-02   0.112 0.910950    
maximum_nights                        -2.858e-05  1.989e-05  -1.437 0.150739    
minimum_nights                        -1.256e-01  5.765e-03 -21.785  < 2e-16 ***
monthly_price                          1.270e-04  1.653e-05   7.681 1.58e-14 ***
price                                 -4.467e-03  4.052e-04 -11.023  < 2e-16 ***
property_typeHouse                    -1.761e-01  2.724e-02  -6.465 1.01e-10 ***
property_typeOther                    -6.077e-02  3.067e-02  -1.981 0.047543 *  
require_guest_phone_verificationt     -2.044e-01  8.717e-02  -2.344 0.019062 *  
require_guest_profile_picturet        -2.587e-02  9.601e-02  -0.269 0.787612    
requires_licenset                      3.755e-01  1.137e-01   3.303 0.000956 ***
room_typePrivate room                 -3.096e-01  2.974e-02 -10.413  < 2e-16 ***
room_typeShared room                  -6.849e-01  7.416e-02  -9.236  < 2e-16 ***
stateCO                                       NA         NA      NA       NA    
stateDC                               -7.921e+00  1.195e+02  -0.066 0.947140    
stateIL                                       NA         NA      NA       NA    
stateLA                                       NA         NA      NA       NA    
```



```
stateMA                                NA         NA      NA       NA
stateMD                        -9.362e+00  1.195e+02  -0.078 0.937544
stateNC                                NA         NA      NA       NA
stateNY                         6.463e-01  1.669e+00   0.387 0.698537
stateOR                                NA         NA      NA       NA
stateTN                                NA         NA      NA       NA
stateTX                         6.889e+00  8.356e+01   0.082 0.934291
stateWA                                NA         NA      NA       NA
weekly_price                   -2.986e-04  6.889e-05  -4.334 1.46e-05 ***
availability_30_to_60                  NA         NA      NA       NA
availability_60_to_90                  NA         NA      NA       NA
availability_90_to_365                 NA         NA      NA       NA
beds_per_room                   6.372e-02  5.842e-02   1.091 0.275395
period_difference                      NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 79043  on 69990  degrees of freedom
Residual deviance: 58888  on 69923  degrees of freedom
AIC: 59024

Number of Fisher Scoring iterations: 9
```

Group member roles

- **Xinyue Shi**: classification tree model (page 5), LASSO regression (page 4)
- **Lei Guo**: logistic regression model (page 5), stepwise procedure (page 5), feature creation (page 3)
- **Tianhang Xu**: random forest model (page 6), feature transformation (page 3), feature creation (page 3)
- **Xin Tan**: text mining (page 4), kNN model (page 6)
- **Entire team**: data cleaning (page 3)