Competition Project Report

Professor Courtney Paulson

BUDT758T: Data Mining and Predictive Analytics

By:

Jiaying Lu / Junyuan Ma / Yuran Wang / Yifan Dang

5/11/2018

Executive Summary

This report is prepared to use collected data on the booking rate to predict which listings have the great potential to be identified with high booking rate. This information is designed to help Airbnb increase its revenues by improving their overall booking rate.

We are trying to help Airbnb in two ways. One is figuring out the optimal prediction model with the highest accuracy that can be used to predict which Airbnb will be most likely booked by customers, the other is to evaluate the importance and determine the correlation between high booking rate and each variable through an interpretation model. Both information will lead to two purposes. The first one is to increase the exposure rate of potential high-booking-rate homes. Those highly popular homes will be on the hot searching listings on Airbnb's official website. Additionally, Airbnb can recommend these appealing listings to customers by putting them in a front position of the customers' searching pages which can increase the matching ratio between customers and hosts. The second purpose is to find out the listings with low booking rate and then make use of other information to give these hosts some suggestions to improve their listings' booking rate.

In order to understand the relationship between factors and booking rate, we have to find a model which is good for interpretation. After understanding the relationships between factors and booking rate, we can provide feasible solutions to hosts on how to improve booking condition, and move forward to boost the overall booking rate in Airbnb.

Overall, we run several models which include logistic regression, Ridge, Lasso, classification tree, SVM, random forests and boosting. It turns out that the optimal prediction model is boosting model with the highest accuracy of 85.026% for the testing dataset. And we chose logistic regression as our interpretation model since it explains the information we need more detailed.

After analyzing this data through a logistic model, we found quite a few factors which are significant variables related to high booking rate. In general, we found that listings with the lower price, lower cleaning fee, lower minimum_nights are more likely to have high booking rates. Listings in big cities such as Boston, New York City and Washington, DC are more highly booked than other cities on average. In order to increase booking rate, hosts may better become a superhost, reply to potential clients as much as they can and collect more verified information. Listings with amenities which bring guests more convenience and provide more comfort utilities may also help to increase the booking rate. For example, heater, air conditioning, and hot water make guests comfortable during their stay. Self-check and shampoo give travelers convenience.

Therefore, our recommendation is as follows:

- Use logistic regression to detect the relation between each variable and high_booking_rate. Give suggestions according to the relationship between high book rate and factors to hosts who do not have high booking rate.
- Use the boosting model to predict which listing would be a high booking rate home and rearrange the listing's priority in search results.
- If two models have any inconsistent prediction, use the boosting model as the final decision. Meanwhile, check the variables in the logistic model to see if there are enough variables to explain the book rate and whether the variables are unbiased.

Exploring Data and Feature Selection

A total of 69 variables were included in the training dataset, some of which were highly correlated. For example, the content of variable “city” and “city_name” are quite close except that “city” sometimes referred to a more specific district and more customized. There were also several variables that were describing locations. In these cases, we only chose variables that were more complete and more standardized in formats among similar variables.

The dataset also contained several text variables such as “amenities” and “transit” and “description”. Since customers were open to enter everything they like into these variables, it is tricky to handle these variables. We believed that when customers are searching for rooms, they must care about amenities. As “amenities” came in a relatively standardized format, we constructed a document-term matrix which included amenities with sparsity over 5%. Any observations with a non-blank “host_verifications” were regarded as having a verification. The same method was also applied to “house_rules”. We divided the variable “access” into three levels “Entire”, “Partial” and “Unknown”. If the text contains certain words, we assign it to “Entire” or “Partial”, otherwise it is “Unknown”. Text variables other than the above four were deleted.

For the “host_response_rate”, we filled with the mean of the existed numbers which is 0.96. And we inserted the blank of “host_response_time” with the mode “within an hour”.

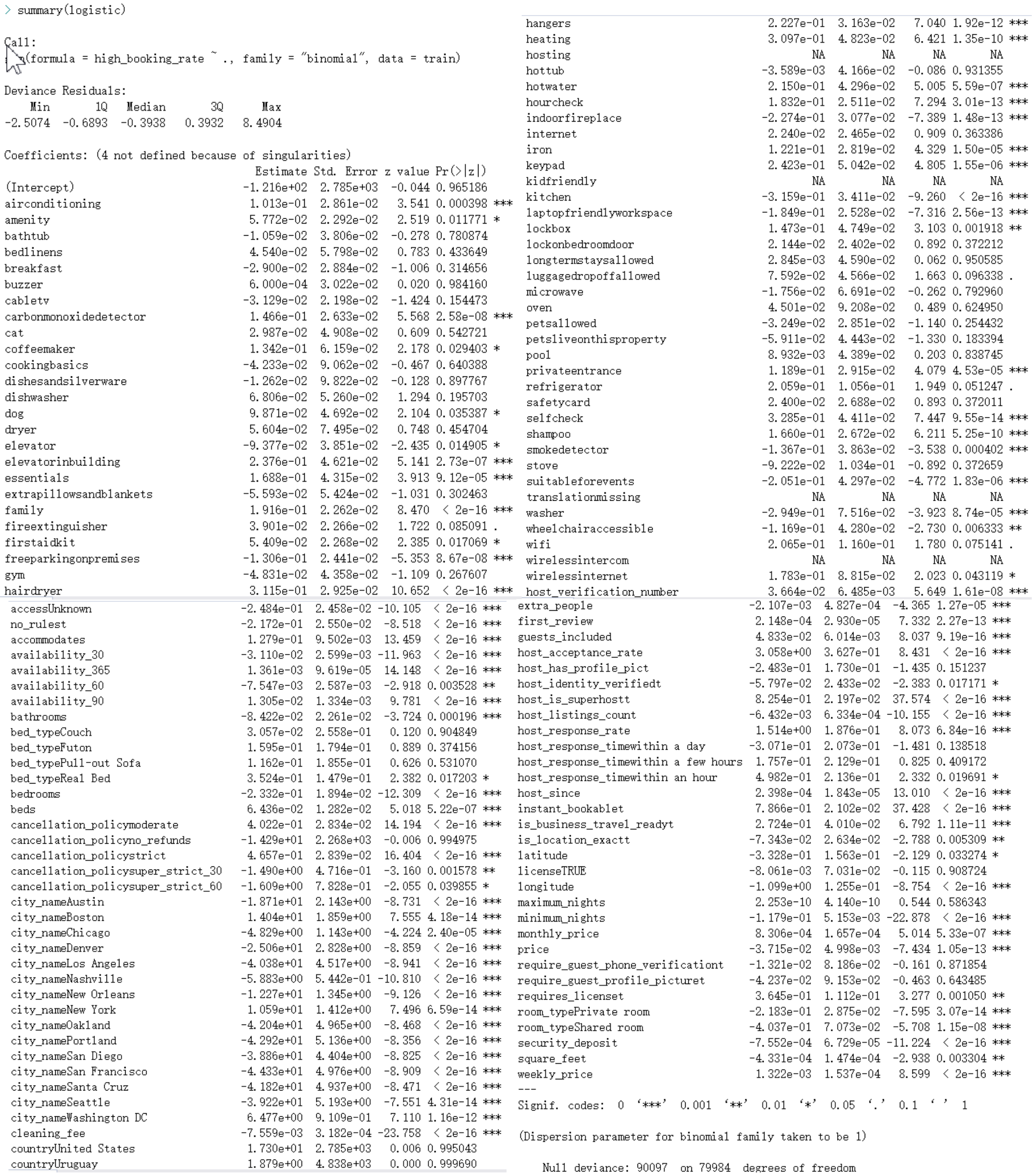
Model Evaluation

The goal is to choose the best classification method to predict whether or not an Airbnb listing will be a high booking rate (1 represents high_booking_rate) and to find out an interpretation model that could help us give suggestions to hosts. We combined the given train independent variables with given dependent variable. We set the seed to 12345 and randomly partitioned the given train data into 30% validation data and 70% remaining(training) data. The baseline we used was 75.11%, which is the percentage of the majority group (Not High Booking Rate) in the validation data. We used a cutoff of 0.5 for prediction classification and calculated the accuracy of all models on the validation data and compare them with the baseline. If the model's accuracy was higher than the baseline, it would a good model for our further model selection. Among those models with a higher accuracy than baseline, we chose the one with the highest accuracy on the validation data for interpretation purpose. Due to the factor that some models are hard to understand and interpret, we also chose one method with more detailed information, which is able to best interpret the relationships between features and booking rate. That is, we can get the exact coefficient not only for numerical variables but also for each level of categorical factors.

Model Details

- Logistic

Our first model is the logistic regression. Our initial attempt on logistic regression did not include any variable derived from text mining. The model reached an accuracy of 78.6% for the validation data. After we added the matrix of amenities, the accuracy of predicting the validation data increased to 79.6%. This confirmed our expectation that amenities is a major concern when people book their rooms. However as we added other variables derived from text mining, we didn’t see an increase in accuracy.



```
> summary(logistic)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5074  -0.6893  -0.3938   0.3932   8.4904

Coefficients: (4 not defined because of singularities)
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.216e+02  2.785e+03  -0.044 0.965186
airconditioning                   1.013e-01  2.861e-02   3.541 0.000398 ***
amenity                           5.772e-02  2.292e-02   2.519 0.011771 *
bathtub                          -1.059e-02  3.806e-02  -0.278 0.780874
bedlinens                         4.540e-02  5.798e-02   0.783 0.433649
breakfast                        -2.900e-02  2.884e-02  -1.006 0.314656
buzzer                            6.000e-04  3.022e-02   0.020 0.984160
cabletv                          -3.129e-02  2.198e-02  -1.424 0.154473
carbonmonoxidedetector            1.466e-01  2.633e-02   5.568 2.58e-08 ***
cat                               2.987e-02  4.908e-02   0.609 0.542721
coffeemaker                       1.342e-01  6.159e-02   2.178 0.029403 *
cookingbasics                    -4.233e-02  9.062e-02  -0.467 0.640388
dishesandsilverware              -1.262e-02  9.822e-02  -0.128 0.897767
dishwasher                        6.806e-02  5.260e-02   1.294 0.195703
dog                               9.871e-02  4.692e-02   2.104 0.035387 *
dryer                             5.604e-02  7.495e-02   0.748 0.454704
elevator                         -9.377e-02  3.851e-02  -2.435 0.014905 *
elevatorinbuilding                2.376e-01  4.621e-02   5.141 2.73e-07 ***
essentials                        1.688e-01  4.315e-02   3.913 9.12e-05 ***
extrapillowsandblankets          -5.593e-02  5.424e-02  -1.031 0.302463
family                            1.916e-01  2.262e-02   8.470  < 2e-16 ***
fireextinguisher                  3.901e-02  2.266e-02   1.722 0.085091 .
firstaidkit                       5.409e-02  2.268e-02   2.385 0.017069 *
freeparkingonpremises            -1.306e-01  2.441e-02  -5.353 8.67e-08 ***
gym                              -4.831e-02  4.358e-02  -1.109 0.267607
hairdryer                         3.115e-01  2.925e-02  10.652  < 2e-16 ***
```

```
hangers                           2.227e-01  3.163e-02   7.040 1.92e-12 ***
heating                           3.097e-01  4.823e-02   6.421 1.35e-10 ***
hosting                                  NA         NA      NA       NA
hottub                           -3.589e-03  4.166e-02  -0.086 0.931355
hotwater                          2.150e-01  4.296e-02   5.005 5.59e-07 ***
hourcheck                         1.832e-01  2.511e-02   7.294 3.01e-13 ***
indoorfireplace                  -2.274e-01  3.077e-02  -7.389 1.48e-13 ***
internet                          2.240e-02  2.465e-02   0.909 0.363386
iron                              1.221e-01  2.819e-02   4.329 1.50e-05 ***
keypad                            2.423e-01  5.042e-02   4.805 1.55e-06 ***
kidfriendly                              NA         NA      NA       NA
kitchen                          -3.159e-01  3.411e-02  -9.260  < 2e-16 ***
laptopfriendlyworkspace          -1.849e-01  2.528e-02  -7.316 2.56e-13 ***
lockbox                           1.473e-01  4.749e-02   3.103 0.001918 **
lockonbedroomdoor                 2.144e-02  2.402e-02   0.892 0.372212
longtermstaysallowed              2.845e-03  4.590e-02   0.062 0.950585
luggagedropoffallowed             7.592e-02  4.566e-02   1.663 0.096338 .
microwave                        -1.756e-02  6.691e-02  -0.262 0.792960
oven                              4.501e-02  9.208e-02   0.489 0.624950
petsallowed                      -3.249e-02  2.851e-02  -1.140 0.254432
petsliveonthisproperty           -5.911e-02  4.443e-02  -1.330 0.183394
pool                              8.932e-03  4.389e-02   0.203 0.838745
privateentrance                   1.189e-01  2.915e-02   4.079 4.53e-05 ***
refrigerator                      2.059e-01  1.056e-01   1.949 0.051247 .
safetycard                        2.400e-02  2.688e-02   0.893 0.372011
selfcheck                         3.285e-01  4.411e-02   7.447 9.55e-14 ***
shampoo                           1.660e-01  2.672e-02   6.211 5.25e-10 ***
smokedetector                    -1.367e-01  3.863e-02  -3.538 0.000402 ***
stove                            -9.222e-02  1.034e-01  -0.892 0.372659
suitableforevents                -2.051e-01  4.297e-02  -4.772 1.83e-06 ***
translationmissing                       NA         NA      NA       NA
washer                           -2.949e-01  7.516e-02  -3.923 8.74e-05 ***
wheelchairaccessible             -1.169e-01  4.280e-02  -2.730 0.006333 **
wifi                              2.065e-01  1.160e-01   1.780 0.075141 .
wirelessintercom                         NA         NA      NA       NA
wirelessinternet                  1.783e-01  8.815e-02   2.023 0.043119 *
host_verification_number          3.664e-02  6.485e-03   5.649 1.61e-08 ***
```

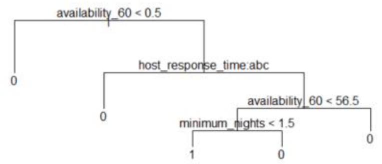
```
accessUnknown                    -2.484e-01  2.458e-02 -10.105  < 2e-16 ***
no_rulest                        -2.172e-01  2.550e-02  -8.518  < 2e-16 ***
accommodates                      1.279e-01  9.502e-03  13.459  < 2e-16 ***
availability_30                  -3.110e-02  2.599e-03 -11.963  < 2e-16 ***
availability_365                  1.361e-03  9.619e-05  14.148  < 2e-16 ***
availability_60                  -7.547e-03  2.587e-03  -2.918 0.003528 **
availability_90                   1.305e-02  1.334e-03   9.781  < 2e-16 ***
bathrooms                        -8.422e-02  2.261e-02  -3.724 0.000196 ***
bed_typeCouch                     3.057e-02  2.558e-01   0.120 0.904849
bed_typeFuton                     1.595e-01  1.794e-01   0.889 0.374156
bed_typePull-out Sofa             1.162e-01  1.855e-01   0.626 0.531070
bed_typeReal Bed                  3.524e-01  1.479e-01   2.382 0.017203 *
bedrooms                         -2.332e-01  1.894e-02 -12.309  < 2e-16 ***
beds                              6.436e-02  1.282e-02   5.018 5.22e-07 ***
cancellation_policymoderate       4.022e-01  2.834e-02  14.194  < 2e-16 ***
cancellation_policyno_refunds    -1.429e+01  2.268e+03  -0.006 0.994975
cancellation_policystrict         4.657e-01  2.839e-02  16.404  < 2e-16 ***
cancellation_policysuper_strict_30 -1.490e+00 4.716e-01 -3.160 0.001578 **
cancellation_policysuper_strict_60 -1.609e+00 7.828e-01 -2.055 0.039855 *
city_nameAustin                  -1.871e+01  2.143e+00  -8.731  < 2e-16 ***
city_nameBoston                   1.404e+01  1.859e+00   7.555 4.18e-14 ***
city_nameChicago                 -4.829e+00  1.143e+00  -4.224 2.40e-05 ***
city_nameDenver                  -2.506e+01  2.828e+00  -8.859  < 2e-16 ***
city_nameLos Angeles             -4.038e+01  4.517e+00  -8.941  < 2e-16 ***
city_nameNashville               -5.883e+00  5.442e-01 -10.810  < 2e-16 ***
city_nameNew Orleans             -1.227e+01  1.345e+00  -9.126  < 2e-16 ***
city_nameNew York                 1.059e+01  1.412e+00   7.496 6.59e-14 ***
city_nameOakland                 -4.204e+01  4.965e+00  -8.468  < 2e-16 ***
city_namePortland                -4.292e+01  5.136e+00  -8.356  < 2e-16 ***
city_nameSan Diego               -3.886e+01  4.404e+00  -8.825  < 2e-16 ***
city_nameSan Francisco           -4.433e+01  4.976e+00  -8.909  < 2e-16 ***
city_nameSanta Cruz              -4.182e+01  4.937e+00  -8.471  < 2e-16 ***
city_nameSeattle                 -3.922e+01  5.193e+00  -7.551 4.31e-14 ***
city_nameWashington DC            6.477e+00  9.109e-01   7.110 1.16e-12 ***
cleaning_fee                     -7.559e-03  3.182e-04 -23.758  < 2e-16 ***
countryUnited States              1.730e+01  2.785e+03   0.006 0.995043
countryUruguay                    1.879e+00  4.838e+03   0.000 0.999690
```

```
extra_people                     -2.107e-03  4.827e-04  -4.365 1.27e-05 ***
first_review                      2.148e-04  2.930e-05   7.332 2.27e-13 ***
guests_included                   4.833e-02  6.014e-03   8.037 9.19e-16 ***
host_acceptance_rate              3.058e+00  3.627e-01   8.431  < 2e-16 ***
host_has_profile_pict            -2.483e-01  1.730e-01  -1.435 0.151237
host_identity_verifiedt          -5.797e-02  2.433e-02  -2.383 0.017171 *
host_is_superhostt                8.254e-01  2.197e-02  37.574  < 2e-16 ***
host_listings_count              -6.432e-03  6.334e-04 -10.155  < 2e-16 ***
host_response_rate                1.514e+00  1.876e-01   8.073 6.84e-16 ***
host_response_timewithin a day   -3.071e-01  2.073e-01  -1.481 0.138518
host_response_timewithin a few hours 1.757e-01 2.129e-01 0.825 0.409172
host_response_timewithin an hour  4.982e-01  2.136e-01   2.332 0.019691 *
host_since                        2.398e-04  1.843e-05  13.010  < 2e-16 ***
instant_bookablet                 7.866e-01  2.102e-02  37.428  < 2e-16 ***
is_business_travel_readyt         2.724e-01  4.010e-02   6.792 1.11e-11 ***
is_location_exactt               -7.343e-02  2.634e-02  -2.788 0.005309 **
latitude                         -3.328e-01  1.563e-01  -2.129 0.033274 *
licenseTRUE                      -8.061e-03  7.031e-02  -0.115 0.908724
longitude                        -1.099e+00  1.255e-01  -8.754  < 2e-16 ***
maximum_nights                    2.253e-10  4.140e-10   0.544 0.586343
minimum_nights                   -1.179e-01  5.153e-03 -22.878  < 2e-16 ***
monthly_price                     8.306e-04  1.657e-04   5.014 5.33e-07 ***
price                            -3.715e-02  4.998e-03  -7.434 1.05e-13 ***
require_guest_phone_verificationt -1.321e-02 8.186e-02  -0.161 0.871854
require_guest_profile_picturet   -4.237e-02  9.153e-02  -0.463 0.643485
requires_licenset                 3.645e-01  1.112e-01   3.277 0.001050 **
room_typePrivate room            -2.183e-01  2.875e-02  -7.595 3.07e-14 ***
room_typeShared room             -4.037e-01  7.073e-02  -5.708 1.15e-08 ***
security_deposit                 -7.552e-04  6.729e-05 -11.224  < 2e-16 ***
square_feet                      -4.331e-04  1.474e-04  -2.938 0.003304 **
weekly_price                      1.322e-03  1.537e-04   8.599  < 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: 90097  on 79984  degrees of freedom
```

From the summary, we can see that customers more care about the convenience and comfort of a room. Room seekers tend to book rooms which have temperature control, elevator and can check in quickly. As we expected, variable "country" is not significant because almost all observations are in the US. Customers also don't seem to care about hosts' license and profile picture. Though they don't quite care about the maximum nights, they do need to consider the minimum nights required to book a room. Logistic model also suggests that the most popular room_type is the entire home. If a host becomes a super host and does more verifications on the website, their effort will more likely to get payback. All these information are critically for us to help with those low booking rate listings to improve.

- Classification Tree

  We applied the classification tree as well. Our tree is shown as below:



Our prediction accuracy for decision tree is 76.97%. However, after we run the model, the tree only selects 3 variables, “availability_60”, “host_repsonse_time”, “minimum_nights” and eliminates all the other variables. According to the dendrogram, any Airbnb availability less than half day will predict the book rate as 0, any Airbnb availability more than half and has a certain amount of host_reponse _time will predict high booking rate as 0. For availability_60 less than 56.5 and minimum nights less than 1.5 will predict high booking rate as 1. And other will be 0. And any availability larger than 56.5 will predict high booking rate as 0. One potential reason for the model eliminates a lot of variables is because most of the variables are highly correlated with each other, and decision tree recognized the highly involved interaction of different variables so that it deletes the correlated one but leave the most significant one. However, since this model eliminates most of the variables, it ends up with very low accuracy so that we decide to not use decision tree anyway.

- Ridge (additional model output is in appendix)

We also applied a ridge regression with training data if there is high correlation variables in our dataset. Opposed to LASSO regression, Ridge wouldn’t take away any variables from the model. The significant coefficients are those separate far from zero, so we chose those variables whose absolute coefficients’ value are above 0.3. Accordingly, we draw a conclusion that the significant variables were elevator building, heating, host_acceptance_rate, host_is_superhost, host_response_time, instant_bookable, and requires_license.

According to the output, we found that an increase in host_acceptance_rate would increase the probability of high booking rate. In addition, compared with the listings without elevator and heating, the listings have these amenities tend to have high book rate. And compared with the listings don’t require licenses, the listings require license has less probability of having high book rate. What’s more, the probability of having high book rate would also increase owing to a super host, a shorter response time, or instant bookable.

However, in order to run out the regression, we translated all variables into a numerical format which resulted in a difficulty in interpreting the categorical variables. What’s more, with this model, our team discovered an accuracy of only 78.65% on our validation data. Although this was higher than the validation baseline of 75.11%, it was much lower than boosting model we tested, so ultimately we do not recommend a Ridge model for this problem.

- LASSO (additional model output is in appendix)

LASSO performs both regularization and variable selection, which can improve the prediction accuracy and enhance interpretability of model. It is supposed to be a better model for interpretation purpose. We used the training data to train the LASSO model. We do not need a separate validation data split here since LASSO uses cross-validation for their model choices.
Any variable not be taken away can be considered significant in LASSO. According to the output, only breakfast, dryer, oven, maximum_nights, minimum_nights, and monthly_price are excluded in this model. It indicates that these features are not useful when evaluating the importance and determining the correlation between high booking rate and each feature. Among the significant variables, country, host_acceptance_rate, host_is_superhost, host_response_rate,

host_response_time, instant_bookable, license, and reuqires_license seem to have more influences on predicting a high booking rate listing, since their absolute values of the coefficient are higher than others.

It should be noted with this model, our team discovered an accuracy of only 78.75% on our validation data. Although this was higher than the validation baseline of 75.11%, the prediction accuracy of this model was lower than boosting model we tested. To run this model, we converted all the categorical variables into numerical. Therefore, this conversion affected our interpretation results on the final model. Thus, we do not recommend the LASSO model for both interpretation and prediction purpose.

- Random Forests (additional model output is in appendix)

Since the random forests can avoid overfitting and reduce variance compared to a classification tree, we also tried the random forests model for prediction purpose. We used the training data to train the model.

According to the output, by sorting variable importance from high to low, city_name, minimun_nighs, availability_365, hosting_listings_count, and cleaning fee are the top 5 important variables for this model to predict whether an Airbnb listing will be a high booking rate. With this model, our team discovered an accuracy of only 83.66% on our validation data. Although this was higher than the validation baseline of 75.11%, it was slightly lower than the boosting model we tested. Additionally, since it is hard for this model to interpret the relationships between each variable and high booking rate, so ultimately we do not recommend a Random Forest model for this problem.

- Support Vector Machine (additional model output is in appendix)

We also tried Support Vector Machine since it works on both linear and non-linear classification that might improve the our prediction accuracy. We used the training data to train the Support Vector Machine model.

According to the output, based on the training data set, the SVM identified 26421 support vectors. By extracting variable weights from this SVM model, beds, host_is_superhost, cleaning_fee, instant_bookable, and maximum_nights are the top 5 relatively important variables this model used to separate the data set since they have larger absolute vector weights. With this model, our team discovered an accuracy of only 81.16% on our validation data. Although this was higher than the validation baseline of 75.11%, it was lower than the boosting model we tested and as it is hard to interpret, so ultimately we do not recommend an SVM model for this problem.

- Boosting (additional model output is in appendix)

In order to increase the accuracy, we employed boosting with our training dataset. According to the output, the variables are listed with the relative influence. The longitude, availibity_30, first_review, latitude and minimum nights are the top five significant in predicting the high book rate listing. Conversely, the hosting, kid-friendly, translation missing, wireless intercom and country are uselessly indicated by this model.

With this model, our team discovered an accuracy of 85.08% on our validation data, which is the highest accuracy among all the model we used. Thus, we recommended a boosting model only for the purpose of prediction.

Appendices

- Group member roles

All group members run the traditional logistic regression for double check purpose.

Junyuan Ma: take charge of cleaning the whole dataset with different method. And responsible for running logistic and neural network models.

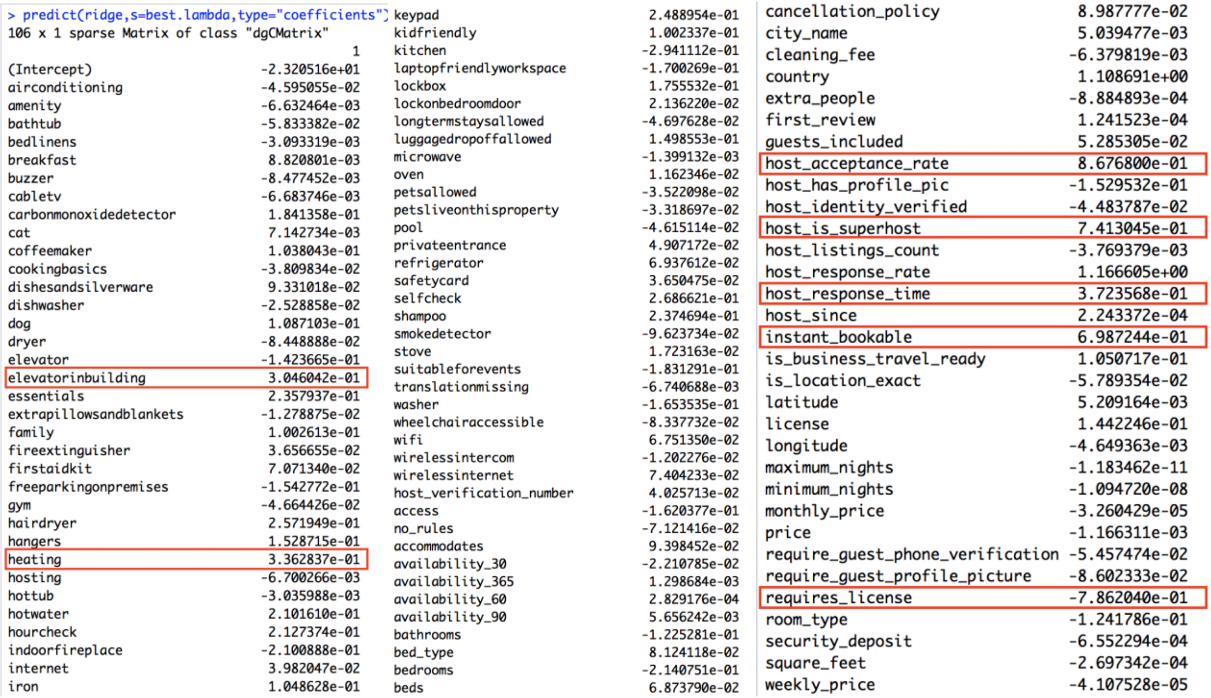
Jiaying Lu: take charge of cleaning the first 23 variables for the original version of dataset. And running the Random Forests, Support Vector Machine, and boosting models.

Yuran Wang: take charge of cleaning the middle 23 variables for the original version of dataset. And running the Ridge, LASSO, and boosting models.

Yifan Dang: take charge of cleaning the last 23 variables for the original version of dataset. And running the Classification Tree, neural network, and boosting models.

- Outputs

Ridge Details

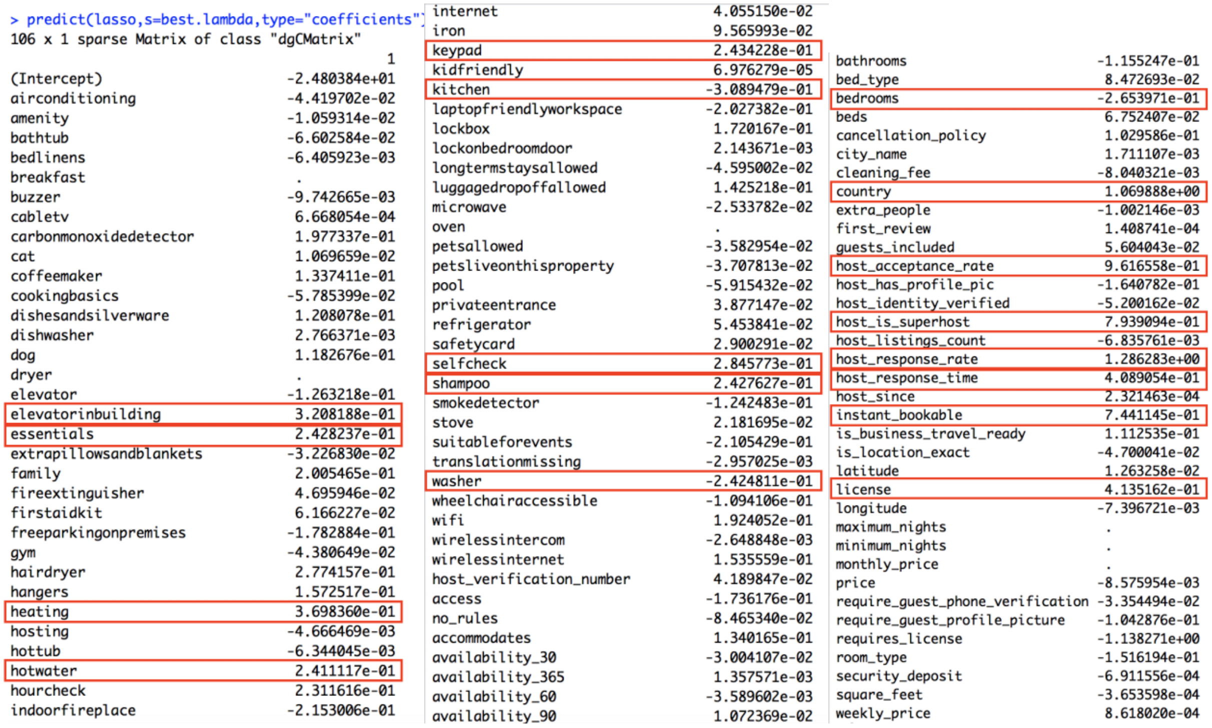


```
> predict(ridge,s=best.lambda,type="coefficients")
106 x 1 sparse Matrix of class "dgCMatrix"
                                        1
(Intercept)                 -2.320516e+01
airconditioning             -4.595055e-02
amenity                     -6.632464e-03
bathtub                     -5.833382e-02
bedlinens                   -3.093319e-03
breakfast                    8.820801e-03
buzzer                      -8.477452e-03
cabletv                     -6.683746e-03
carbonmonoxidedetector       1.841358e-01
cat                          7.142734e-03
coffeemaker                  1.038043e-01
cookingbasics               -3.809834e-02
dishesandsilverware          9.331018e-02
dishwasher                  -2.528858e-02
dog                          1.087103e-01
dryer                       -8.448888e-02
elevator                    -1.423665e-01
elevatorinbuilding           3.046042e-01
essentials                   2.357937e-01
extrapillowsandblankets     -1.278875e-02
family                       1.002613e-01
fireextinguisher             3.656655e-02
firstaidkit                  7.071340e-02
freeparkingonpremises       -1.542772e-01
gym                         -4.664426e-02
hairdryer                    2.571949e-01
hangers                      1.528715e-01
heating                      3.362837e-01
hosting                     -6.700266e-03
hottub                      -3.035988e-03
hotwater                     2.101610e-01
hourcheck                    2.127374e-01
indoorfireplace             -2.100888e-01
internet                     3.982047e-02
iron                         1.048628e-01
```

```
keypad                           2.488954e-01
kidfriendly                      1.002337e-01
kitchen                         -2.941112e-01
laptopfriendlyworkspace         -1.700269e-01
lockbox                          1.755532e-01
lockonbedroomdoor                2.136220e-02
longtermstaysallowed            -4.697628e-02
luggagedropoffallowed            1.498553e-01
microwave                       -1.399132e-03
oven                             1.162346e-02
petsallowed                     -3.522098e-02
petsliveonthisproperty          -3.318697e-02
pool                            -4.615114e-02
privateentrance                  4.907172e-02
refrigerator                     6.937612e-02
safetycard                       3.650475e-02
selfcheck                        2.686621e-01
shampoo                          2.374694e-01
smokedetector                   -9.623734e-02
stove                            1.723163e-02
suitableforevents               -1.831291e-01
translationmissing              -6.740688e-03
washer                          -1.653535e-01
wheelchairaccessible            -8.337732e-02
wifi                             6.751350e-02
wirelessintercom                -1.202276e-02
wirelessinternet                 7.404233e-02
host_verification_number         4.025713e-02
access                          -1.620377e-01
no_rules                        -7.121416e-02
accommodates                     9.398452e-02
availability_30                 -2.210785e-02
availability_365                 1.298684e-03
availability_60                  2.829176e-04
availability_90                  5.656242e-03
bathrooms                       -1.225281e-01
bed_type                         8.124118e-02
bedrooms                        -2.140751e-01
beds                             6.873790e-02
```

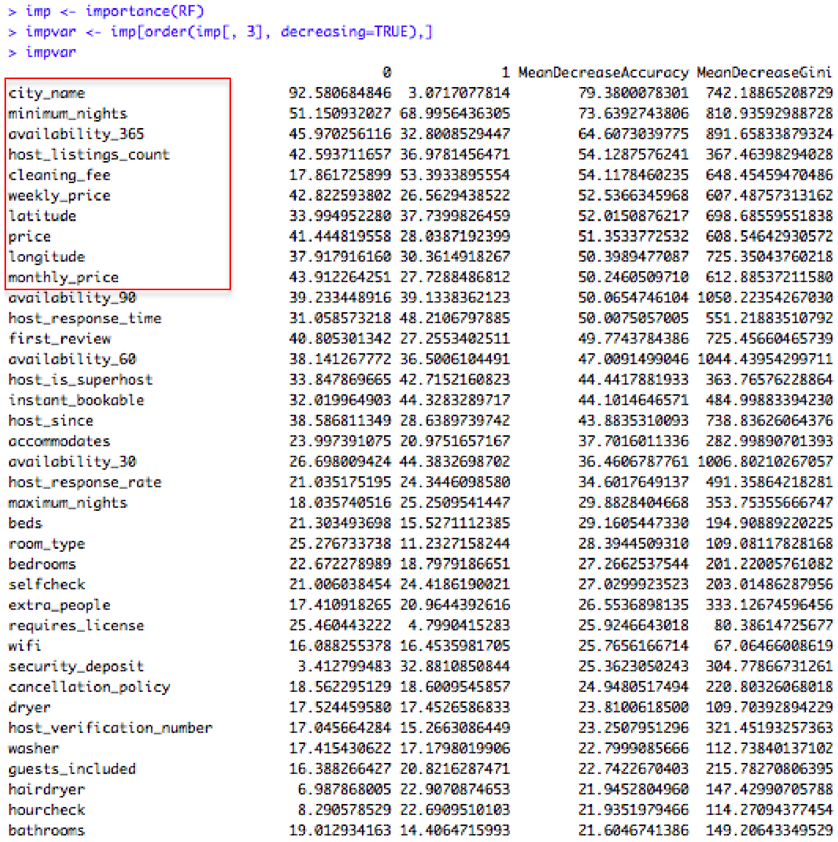
```
cancellation_policy               8.987777e-02
city_name                         5.039477e-03
cleaning_fee                     -6.379819e-03
country                           1.108691e+00
extra_people                     -8.884893e-04
first_review                      1.241523e-04
guests_included                   5.285305e-02
host_acceptance_rate              8.676800e-01
host_has_profile_pic             -1.529532e-01
host_identity_verified           -4.483787e-02
host_is_superhost                 7.413045e-01
host_listings_count              -3.769379e-03
host_response_rate                1.166605e+00
host_response_time                3.723568e-01
host_since                        2.243372e-04
instant_bookable                  6.987244e-01
is_business_travel_ready          1.050717e-01
is_location_exact                -5.789354e-02
latitude                          5.209164e-03
license                           1.442246e-01
longitude                        -4.649363e-03
maximum_nights                   -1.183462e-11
minimum_nights                   -1.094720e-08
monthly_price                    -3.260429e-05
price                            -1.166311e-03
require_guest_phone_verification -5.457474e-02
require_guest_profile_picture    -8.602333e-02
requires_license                 -7.862040e-01
room_type                        -1.241786e-01
security_deposit                 -6.552294e-04
square_feet                      -2.697342e-04
weekly_price                     -4.107528e-05
```

LASSO Details



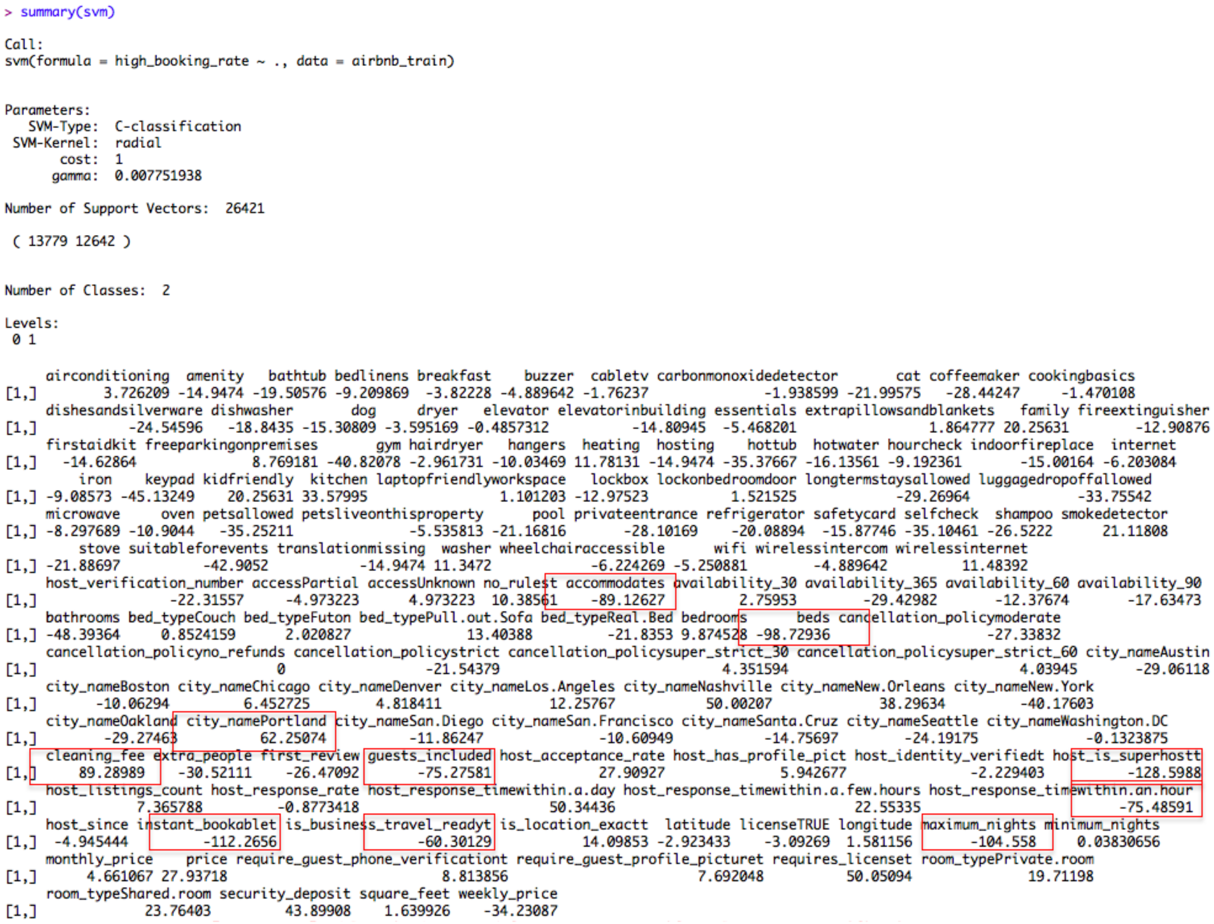
```
> predict(lasso,s=best.lambda,type="coefficients")
106 x 1 sparse Matrix of class "dgCMatrix"
                                              1
(Intercept)                       -2.480384e+01
airconditioning                   -4.419702e-02
amenity                           -1.059314e-02
bathtub                           -6.602584e-02
bedlinens                         -6.405923e-03
breakfast                          .
buzzer                            -9.742665e-03
cabletv                            6.668054e-04
carbonmonoxidedetector             1.977337e-01
cat                                1.069659e-02
coffeemaker                        1.337411e-01
cookingbasics                     -5.785399e-02
dishesandsilverware                1.208078e-01
dishwasher                         2.766371e-03
dog                                1.182676e-01
dryer                              .
elevator                          -1.263218e-01
elevatorinbuilding                 3.208188e-01
essentials                         2.428237e-01
extrapillowsandblankets           -3.226830e-02
family                             2.005465e-01
fireextinguisher                   4.695946e-02
firstaidkit                        6.166227e-02
freeparkingonpremises             -1.782884e-01
gym                               -4.380649e-02
hairdryer                          2.774157e-01
hangers                            1.572517e-01
heating                            3.698360e-01
hosting                           -4.666469e-03
hottub                            -6.344045e-03
hotwater                           2.411117e-01
hourcheck                          2.311616e-01
indoorfireplace                   -2.153006e-01
internet                           4.055150e-02
iron                               9.565993e-02
keypad                             2.434228e-01
kidfriendly                        6.976279e-05
kitchen                           -3.089479e-01
laptopfriendlyworkspace           -2.027382e-01
lockbox                            1.720167e-01
lockonbedroomdoor                  2.143671e-03
longtermstaysallowed              -4.595002e-02
luggagedropoffallowed              1.425218e-01
microwave                         -2.533782e-02
oven                               .
petsallowed                       -3.582954e-02
petsliveonthisproperty            -3.707813e-02
pool                              -5.915432e-02
privateentrance                    3.877147e-02
refrigerator                       5.453841e-02
safetycard                         2.900291e-02
selfcheck                          2.845773e-01
shampoo                            2.427627e-01
smokedetector                     -1.242483e-01
stove                              2.181695e-02
suitableforevents                 -2.105429e-01
translationmissing                -2.957025e-03
washer                            -2.424811e-01
wheelchairaccessible              -1.094106e-01
wifi                               1.924052e-01
wirelessintercom                  -2.648848e-03
wirelessinternet                   1.535559e-01
host_verification_number           4.189847e-02
access                            -1.736176e-01
no_rules                          -8.465340e-02
accommodates                       1.340165e-01
availability_30                   -3.004107e-02
availability_365                   1.357571e-03
availability_60                   -3.589602e-03
availability_90                    1.072369e-02
bathrooms                         -1.155247e-01
bed_type                           8.472693e-02
bedrooms                          -2.653971e-01
beds                               6.752407e-02
cancellation_policy                1.029586e-01
city_name                          1.711107e-03
cleaning_fee                      -8.040321e-03
country                            1.069888e+00
extra_people                      -1.002146e-03
first_review                       1.408741e-04
guests_included                    5.604043e-02
host_acceptance_rate               9.616558e-01
host_has_profile_pic              -1.640782e-01
host_identity_verified            -5.200162e-02
host_is_superhost                  7.939094e-01
host_listings_count               -6.835761e-03
host_response_rate                 1.286283e+00
host_response_time                 4.089054e-01
host_since                         2.321463e-04
instant_bookable                   7.441145e-01
is_business_travel_ready           1.112535e-01
is_location_exact                 -4.700041e-02
latitude                           1.263258e-02
license                            4.135162e-01
longitude                         -7.396721e-03
maximum_nights                     .
minimum_nights                     .
monthly_price                      .
price                             -8.575954e-03
require_guest_phone_verification  -3.354494e-02
require_guest_profile_picture     -1.042876e-01
requires_license                  -1.138271e+00
room_type                         -1.516194e-01
security_deposit                  -6.911556e-04
square_feet                       -3.653598e-04
weekly_price                       8.618020e-04
```

Random Forests Details (Partial Output)



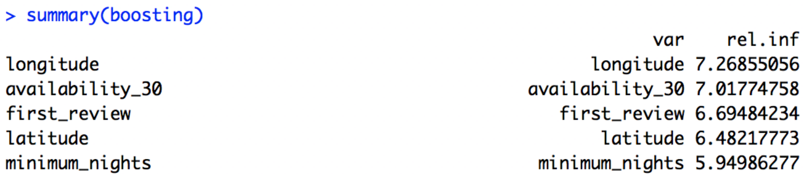
```
> imp <- importance(RF)
> impvar <- imp[order(imp[, 3], decreasing=TRUE),]
> impvar
                                    0          1 MeanDecreaseAccuracy MeanDecreaseGini
city_name                92.580684846  3.0717077814        79.3800078301   742.18865208729
minimum_nights           51.150932027 68.9956436305        73.6392743806   810.93592988728
availability_365         45.970256116 32.8008529447        64.6073039775   891.65833879324
host_listings_count      42.593711657 36.9781456471        54.1287576241   367.46398294028
cleaning_fee             17.861725899 53.3933895554        54.1178460235   648.45459470486
weekly_price             42.822593802 26.5629438522        52.5366345968   607.48757313162
latitude                 33.994952280 37.7399826459        52.0150876217   698.68559551838
price                    41.444819558 28.0387192399        51.3533772532   608.54642930572
longitude                37.917916160 30.3614918267        50.3989477087   725.35043760218
monthly_price            43.912264251 27.7288486812        50.2460509710   612.88537211580
availability_90          39.233448916 39.1338362123        50.0654746104  1050.22354267030
host_response_time       31.058573218 48.2106797885        50.0075057005   551.21883510792
first_review             40.805301342 27.2553402511        49.7743784386   725.45660465739
availability_60          38.141267772 36.5006104491        47.0091499046  1044.43954299711
host_is_superhost        33.847869665 42.7152160823        44.4417881933   363.76576228864
instant_bookable         32.019964903 44.3283289717        44.1014646571   484.99883394230
host_since               38.586811349 28.6389739742        43.8835310093   738.83626064376
accommodates             23.997391075 20.9751657167        37.7016011336   282.99890701393
availability_30          26.698009424 44.3832698702        36.4606787761  1006.80210267057
host_response_rate       21.035175195 24.3446098580        34.6017649137   491.35864218281
maximum_nights           18.035740516 25.2509541447        29.8828404668   353.75355666747
beds                     21.303493698 15.5271112385        29.1605447330   194.90889220225
room_type                25.276733738 11.2327158244        28.3944509310   109.08117828168
bedrooms                 22.672278989 18.7979186651        27.2662537544   201.22005761082
selfcheck                21.006038454 24.4186190021        27.0299923523   203.01486287956
extra_people             17.410918265 20.9644392616        26.5536898135   333.12674596456
requires_license         25.460443222  4.7990415283        25.9246643018    80.38614725677
wifi                     16.088255378 16.4535981705        25.7656166714    67.06466008619
security_deposit          3.412799483 32.8810850844        25.3623050243   304.77866731261
cancellation_policy      18.562295129 18.6009545857        24.9480517494   220.80326068018
dryer                    17.524459580 17.4526586833        23.8100618500   109.70392894229
host_verification_number 17.045664284 15.2663086449        23.2507951296   321.45193257363
washer                   17.415430622 17.1798019906        22.7999085666   112.73840137102
guests_included          16.388266427 20.8216287471        22.7422670403   215.78270806395
hairdryer                 6.987868005 22.9070874653        21.9452804960   147.42990705788
hourcheck                 8.290578529 22.6909510103        21.9351979466   114.27094377454
bathrooms                19.012934163 14.4064715993        21.6046741386   149.20643349529
```

SVM Details



```
> summary(svm)

Call:
svm(formula = high_booking_rate ~ ., data = airbnb_train)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.007751938

Number of Support Vectors:  26421

 ( 13779 12642 )


Number of Classes:  2

Levels:
 0 1

     airconditioning  amenity   bathtub bedlinens breakfast    buzzer  cabletv carbonmonoxidedetector       cat coffeemaker cookingbasics
[1,]        3.726209 -14.9474 -19.50576 -9.209869  -3.82228 -4.889642 -1.76237              -1.938599 -21.99575   -28.44247     -1.470108
     dishesandsilverware dishwasher       dog     dryer   elevator elevatorinbuilding essentials extrapillowsandblankets   family fireextinguisher
[1,]           -24.54596   -18.8435 -15.30809 -3.595169 -0.4857312          -14.80945  -5.468201                1.864777 20.25631        -12.90876
     firstaidkit freeparkingonpremises       gym hairdryer   hangers  heating  hosting    hottub  hotwater hourcheck indoorfireplace  internet
[1,]   -14.62864              8.769181 -40.82078 -2.961731 -10.03469 11.78131 -14.9474 -35.37667 -16.13561 -9.192361       -15.00164 -6.203084
         iron    keypad kidfriendly  kitchen laptopfriendlyworkspace   lockbox lockonbedroomdoor longtermstaysallowed luggagedropoffallowed
[1,] -9.08573 -45.13249    20.25631 33.57995                1.101203 -12.97523          1.521525            -29.26964             -33.75542
     microwave     oven petsallowed petsliveonthisproperty      pool privateentrance refrigerator safetycard selfcheck  shampoo smokedetector
[1,] -8.297689 -10.9044   -35.25211              -5.535813 -21.16816       -28.10169    -20.08894  -15.87746 -35.10461 -26.5222      21.11808
         stove suitableforevents translationmissing  washer wheelchairaccessible      wifi wirelessintercom wirelessinternet
[1,] -21.88697          -42.9052           -14.9474 11.3472            -6.224269 -5.250881        -4.889642         11.48392
     host_verification_number accessPartial accessUnknown no_rulest accommodates availability_30 availability_365 availability_60 availability_90
[1,]                -22.31557     -4.973223      4.973223  10.38561    -89.12627         2.75953        -29.42982       -12.37674       -17.63473
     bathrooms bed_typeCouch bed_typeFuton bed_typePull.out.Sofa bed_typeReal.Bed bedrooms      beds cancellation_policymoderate
[1,] -48.39364     0.8524159      2.020827              13.40388         -21.8353 9.874528 -98.72936                   -27.33832
     cancellation_policyno_refunds cancellation_policystrict cancellation_policysuper_strict_30 cancellation_policysuper_strict_60 city_nameAustin
[1,]                             0                 -21.54379                           4.351594                            4.03945       -29.06118
     city_nameBoston city_nameChicago city_nameDenver city_nameLos.Angeles city_nameNashville city_nameNew.Orleans city_nameNew.York
[1,]       -10.06294         6.452725        4.818411             12.25767           50.00207             38.29634         -40.17603
     city_nameOakland city_namePortland city_nameSan.Diego city_nameSan.Francisco city_nameSanta.Cruz city_nameSeattle city_nameWashington.DC
[1,]        -29.27463          62.25074          -11.86247              -10.60949           -14.75697        -24.19175             -0.1323875
     cleaning_fee extra_people first_review guests_included host_acceptance_rate host_has_profile_pict host_identity_verifiedt host_is_superhostt
[1,]     89.28989    -30.52111    -26.47092       -75.27581             27.90927              5.942677               -2.229403          -128.5988
     host_listings_count host_response_rate host_response_timewithin.a.day host_response_timewithin.a.few.hours host_response_timewithin.an.hour
[1,]            7.365788         -0.8773418                       50.34436                             22.55335                        -75.48591
     host_since instant_bookablet is_business_travel_readyt is_location_exactt  latitude licenseTRUE longitude maximum_nights minimum_nights
[1,]  -4.945444         -112.2656                 -60.30129           14.09853 -2.923433    -3.09269  1.581156       -104.558     0.03830656
     monthly_price    price require_guest_phone_verificationt require_guest_profile_picturet requires_licenset room_typePrivate.room
[1,]      4.661067 27.93718                          8.813856                       7.692048          50.05094              19.71198
     room_typeShared.room security_deposit square_feet weekly_price
[1,]             23.76403         43.89908    1.639926    -34.23087
```

Boosting Details (Partial Output)



```
> summary(boosting)
                                                   var    rel.inf
longitude                                    longitude 7.26855056
availability_30                        availability_30 7.01774758
first_review                              first_review 6.69484234
latitude                                      latitude 6.48217773
minimum_nights                          minimum_nights 5.94986277
```

- R Code (Partial)

```
######################## Logistic #######################
airbnb <-readRDS("~/Desktop/train.rds")
set.seed(12345)
validation_instn <- sample(nrow(airbnb), 0.3 * nrow(airbnb))
airbnb_validation <- airbnb[validation_instn, ]
airbnb_train <- airbnb[-validation_instn, ]
logistic <- glm(high_booking_rate ~ ., data = airbnb_train, family = "binomial")
summary(logistic)
log_pred_valid <- predict(logistic, newdata=airbnb_validation, type = "response")
log_pre <- ifelse(log_pred_valid>0.5,1,0)
table(airbnb_validation$high_booking_rate, log_pre)
```

```
accuracy_log_valid = sum(ifelse(airbnb_validation$high_booking_rate == log_pre, 1, 0)) /
nrow(airbnb_validation)
######################### Ridge #######################
library(glmnet)
ridge <- glmnet(data.matrix(airbnb_train[,c(1:105)]),airbnb_train$high_booking_rate,
family="binomial",alpha=0)
ridge.cv <- cv.glmnet(data.matrix(airbnb_train[,c(1:105)]),airbnb_train$high_booking_rate,
family="binomial",alpha=0)
best.lambda <- ridge.cv$lambda.min
predict(ridge,s=best.lambda,type="coefficients")
ridge.pred <- predict(ridge, s=best.lambda,newx = data.matrix(airbnb_validation[,c(1:105)]),
type="response")
class_ridge <- ifelse(ridge.pred>0.5,1,0)
accuracy_ridge <- sum(ifelse(airbnb_validation$high_booking_rate==class_ridge,1,0))/
nrow(airbnb_validation)
######################### LASSO #######################
lasso <- glmnet(data.matrix(airbnb_train[,c(1:105)]),airbnb_train$high_booking_rate,family =
"binomial",alpha = 1)
lasso.cv <- cv.glmnet(data.matrix(airbnb_train[,c(1:105)]),airbnb_train$high_booking_rate,
family="binomial",alpha=1)
best.lambda <- lasso.cv$lambda.min
predict(lasso,s=best.lambda,type="coefficients")
lasso.pred <- predict(lasso,s=best.lambda,newx=data.matrix(airbnb_validation[,c(1:105)]),
type="response")
class_lasso <- ifelse(lasso.pred>0.5,1,0)
accuracy_lasso <- sum(ifelse(airbnb_validation$high_booking_rate==class_lasso,1,0))/
nrow(airbnb_validation)
######################### Boosting #######################
airbnb <- data.matrix(airbnb)
airbnb <- data.frame(airbnb)
airbnb$high_booking_rate <- airbnb$high_booking_rate-1
airbnb$license <- as.factor(airbnb$license)
set.seed(12345)
validation_instn <- sample(nrow(airbnb),0.3*nrow(airbnb))
airbnb_validation <- airbnb[validation_instn,]
airbnb_train <- airbnb[-validation_instn,]
```

```
library(gbm)
boosting <- gbm(high_booking_rate~.,data=airbnb_train,distribution = "bernoulli",n.tree=10000,
          interaction.depth = 5,shrinkage = 0.05)
boosting_pred <- predict(boosting,newdata=airbnb_validation,type="response",n.tree=10000)
class_boost <- ifelse(boosting_pred>0.5,1,0)
accuracy_boost <-
sum(ifelse(airbnb_validation$high_booking_rate==class_boost,1,0))/nrow(airbnb_validation)
summary(boosting)
##################### SVM ##############################
airbnb <- readRDS("~/Desktop/train.rds")
test <-  readRDS("~/Desktop/test1.rds")
airbnb <- airbnb[,-77]
test <- test[,-77]
common <- intersect(names(airbnb), names(test))
for (p in common) { if (class(airbnb[[p]]) == "factor") { levels(test[[p]]) <- levels(airbnb[[p]]) }}
set.seed(12345)
validation_instn <- sample(nrow(airbnb), 0.3 * nrow(airbnb))
airbnb_validation <- airbnb[validation_instn, ]
airbnb_train <- airbnb[-validation_instn, ]
library(e1071)
svm <- svm(high_booking_rate~. , data=airbnb_train, scale = TRUE)
summary(svm)
w = t(svm$coefs) %*% svm$SV
svm_pred_valid <- predict(svm, newdata=airbnb_validation, type = "response")
table(airbnb_validation$high_booking_rate, svm_pred_valid)
accuracy_svm_valid = sum(ifelse(airbnb_validation$high_booking_rate == svm_pred_valid, 1,
0)) /nrow(airbnb_validation)
##################### Random Forests #########################
airbnb <-readRDS("~/Desktop/train.rds")
test <-readRDS("~/Desktop/test1.rds")
common <- intersect(names(airbnb), names(test))
for (p in common) { if (class(airbnb[[p]]) == "factor") { levels(test[[p]]) <- levels(airbnb[[p]]) }}
set.seed(12345)
validation_instn <- sample(nrow(airbnb), 0.3 * nrow(airbnb))
airbnb_validation <- airbnb[validation_instn, ]
airbnb_train <- airbnb[-validation_instn, ]
```

```
library(randomForest)
RF <- randomForest (high_booking_rate~.,airbnb_train,ntree=500,norm.votes=FALSE,
do.trace=10, importance=TRUE)
summary(RF)
imp <- importance(RF)
impvar <- imp[order(imp[, 3], decreasing=TRUE),]
rf_pred_valid <- predict(RF, newdata=airbnb_validation, type = "response")
table(airbnb_validation$high_booking_rate, rf_pred_valid)
accuracy_rf_valid = sum(ifelse(airbnb_validation$high_booking_rate == rf_pred_valid, 1, 0)) /
nrow(airbnb_validation)
```