Justin Jiayuan Tian

Brown University

Data 1030 Hands-on Data Science

https://github.com/quicklearnerjustin/1030project

Prediction of House Ratings on Airbnb

Introduction

I like the concept of 'sharing economy', which allows people to lend their spare properties to others in need so that it changes how people transport and travel. Sharing economy also helps travelers to save their money and space and provides opportunities for hosts to make profits. Thus, I am interested in investigating Airbnb's business and gaining an insight into its success and popularity.

Rating is a key to the success of Airbnb. It builds a sense of trust and reliability. However, according to an article on Forbes, what many guests don't understand is that anything less than a five-star review can cause serious issues for a host. The Airbnb platform actually delivers stress-inducing warnings if the hosts get four-star ratings. Anything less than five stars can have a serious and detrimental impact on a host's placement in the all-important search rankings. If this happens, to predict the ratings would be very hard and my goal is trying to demystify the secret of Airbnb. I downloaded the data of Airbnb in U.S major cities, which contain 72000 observations and 28 variables. The dataset was originally used to predict house prices in Delloite's machine learning competition, while my goal is to predict the ratings. Each of the observations has a unique ID number. Some of the data are text data while some of them are numerical data. Some of them are continuous while some of them are discrete. I am interested in the relationship between ratings and a bunch of features like the number of bedrooms and the bathrooms, prices, property type, room type and number of accommodates. For example, I can

test if there is a relationship between ratings and regions and prices. Is it possible that the ratings of houses in Boston are generally higher than those of New York? Or are relatively expensive houses rated higher or lower than cheaper houses? Do houses gain a higher score if customers are allowed to cancel their transactions for free? In my opinion, customers might tend to rate higher if they can rent the whole property than if they could only rent a room or share the kitchen and the living room with others. I can deal with these problems by using regression to test how other features affect ratings.

EDA

My target variable is *'Review_scores_rating'* which ranges from 0 to 100. The median is 95 and the mean is 92.59. From the figure, we could see that most of the score are densely distributed from 90 to 100.

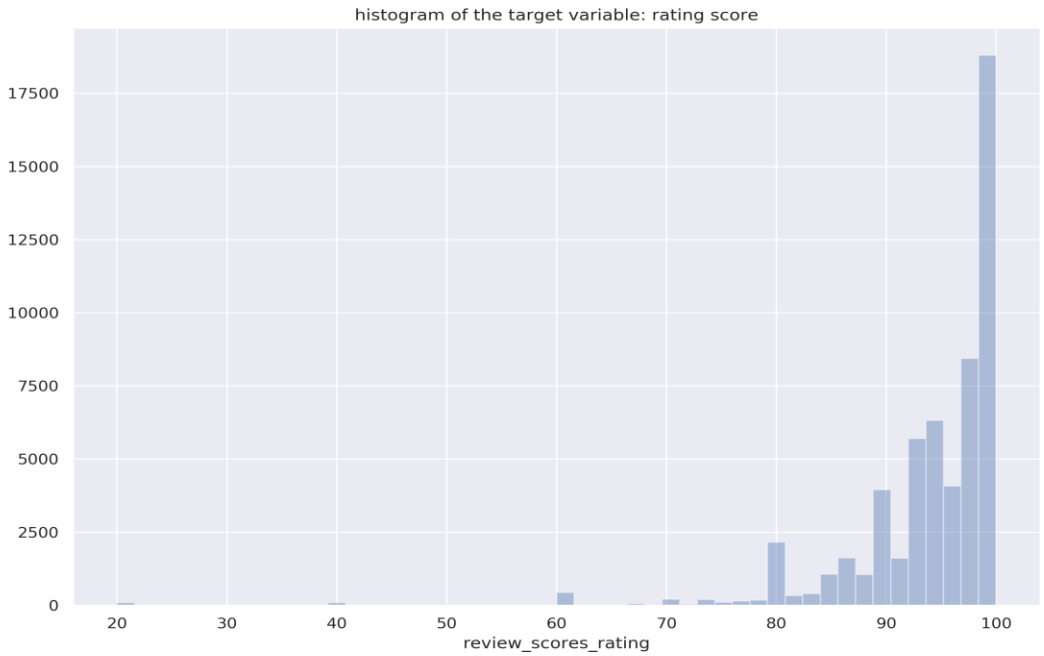

Figure 1: Histogram of review_scores_rating: the histogram shows the number of houses with different ranges of rating

A guess I made is that the more bathroom a house contains, the highly it would be rated. The reasoning is that more bathrooms tend to provide more privacy and convenience for tenants. However, the boxplot does not indicate any strong relationship.

Figure 2: Boxplot of No. of Bathrooms VS Rating: the boxplot shows the distributions of ratings for houses with different number of bathrooms

Then, a thought came up that the ratio between the number of bathrooms and the number of bedrooms might be a better indicator because a house with one bedroom and one bathroom provides more privacy than another house with three bedrooms and two bathrooms although the latter has more bathrooms. Unfortunately, I failed to find any significant relationship, either. Another thing I did is to figure out the relationship between cities and ratings. The finding is that the houses in San Francisco tend to be rated higher than those in LA.

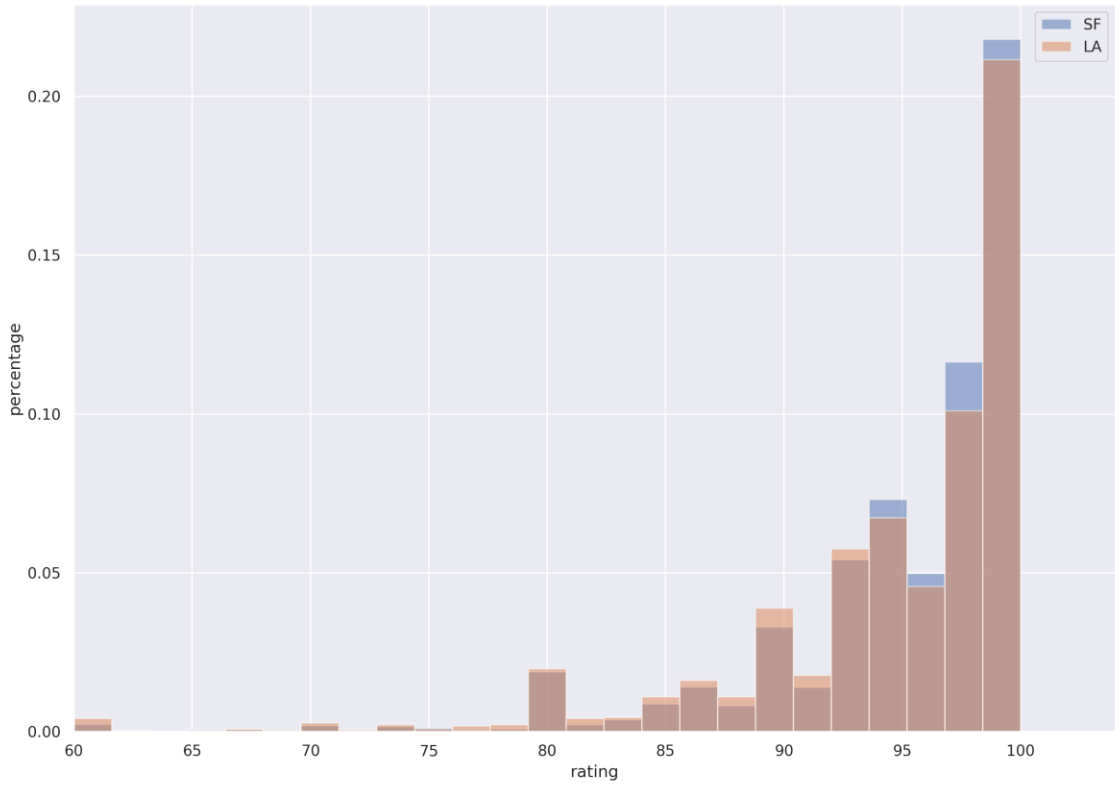

Figure 3: Histogram of Cities VS Rating: the histogram shows the percentage of houses with different ratings in San Francisco and Los Angeles

Room type is supposed to be related to rating as well because people should have better experiences if they can own the whole house than if they share it with other guests. However, the graph denies the assertion again.

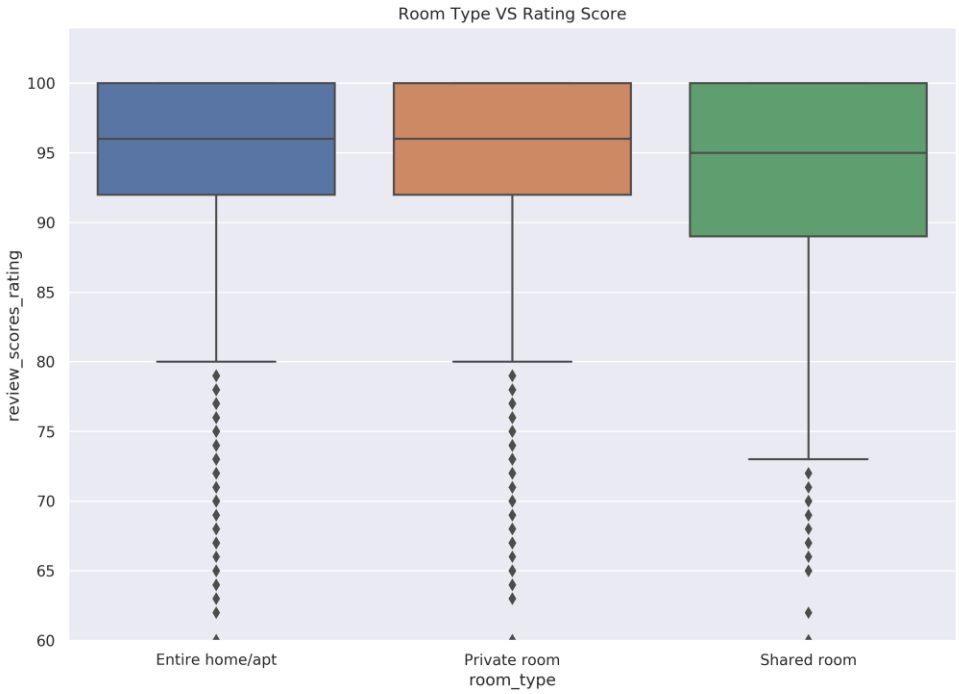

Figure 4: Boxplot of Room Type VS Rating: the boxplot shows distributions of ratings for houses of different types

Methods

Later, I work on with the data of wrong types. '*Last_review'*, *'first_review'*, and *'Host_since'* are converted to timedelta from strings. I also calculate the difference between the data and the max date plus one. I converted *'Host_response_rate'* to integers from strings. In order to preprocess my data, I dropped the missing categorical data because very few of the data are missing. Then, I used random forest imputation to deal with the missing continuous data. I set the value of the missing date as 0. I also dropped the missing data of my target variable, because it does not make sense to train the data with a missing target label. I also created a new

column called *'diff_last_first_review'*, which is the difference between the time of the last review and the time of the first review. Then, I used OneHotEncoder to scale categorical features and MinMaxSclaer and StandardScaler to rescale continuous features.

Then, I chose 4 models to predict ratings: Random Forest Regressor, Lasso linear models, Ridge linear models, SVR, and XGBoost. For each, I ran a k-fold cross validation and set n_folds equal to 5. I also ran 10 random seeds for each try of cross validation. For each random state, I look for the lowest mean squared error in cross validation to find the best parameters. MSE is a measure of the quality of an estimator as it measures the average squared difference between the estimated values and the actual value. Then, I figure out the R^2 score in the test set. In the random forest model, I tune the tree depth between 3 and 10 and the sample splits from 3 to 15. The best score I find is 0.08345 and the mean R^2 score is 0.036 with a standard deviation of 0.019. The mean MSE is 59.06. In the Lasso and Ridge linear model, I tune alpha with 20 values evenly from 10^ (-4) to 10^4. The mean MSE of Lasso is 59.808, and the mean MSE of Ridge is 59.57. Thus, we know my model is 2 standard deviations away from the baseline model. The best score I find is 0.06088 for both and the mean score is 0.05 with its standard deviation of 0.007 for Lasso and 0.048 and 0.007 for Ridge respectively, which are pretty close to each other and approximately 7 standard deviations away from the baseline model. My best score for XGBoost is 0.058, with a mean of 0.055 and its standard deviation of 0.005. The mean MSE is 54.471. The best score of SVR is 0.008, with its mean of 0.006 and its standard deviation of 0.0009. Its mean MSE is 58.87. One thing we notice is that SVR takes the longest time and it is not unexpected. As it is mentioned in an article, *Support Vector classification for large data sets by reducing training data with change of classes*, the SVM need to solve the problem of quadratic programming (QP). Since QP's routines have quadratic complexity the SVM need

huge quantities of computational time and memory for very large data sets because the training complexity of SVM is highly dependent on the size of a data set. The scores are shown in the table below.

Model	Best Score	Mean of R^2	Standard deviation
Random Forest Regressor	0.08345	0.036	0.019
Lasso Linear Model	0.06088	0.05	0.007
Ridge Linear Model	0.06088	0.048	0.007
SVR	0.008	0.006	0.0009
XGBoost	0.058	0.055	0.005

Results of Models

Results

As all of the R^2's are higher than 0, we can say that all of the models are slightly better than the baseline model (the cost function of the median of y). Surprisingly, the test scores are pretty low, implying that very few of the variance of y can be explained by x. I also figure out the global feature importance feature for each feature and make a plot of the top 20 most important features. We could see from the figure below that *host_reponse_rate*, *last_review*, and *first_review* are among the top 4 important features. This is in accordance with the fact that the hosts are afraid of receiving low ratings and that they want to provide best services to the guests. Thus, we could imply that the service provided by the host is crucial to the rating of the house and the interactions and communications between hosts and guests could help to raise the rating. Also, another thing we cannot ignore is that the price is an important factor as well. As we

cannot tell if the effect is positive or negative from the score, we could interpret this phenomenon in two different ways. An expensive house might provide great experiences and people tend to rate the house highly. However, it is also possible that a high price might make people think that the house is not worthy and that they waste their money, and thus, only give poor rates.

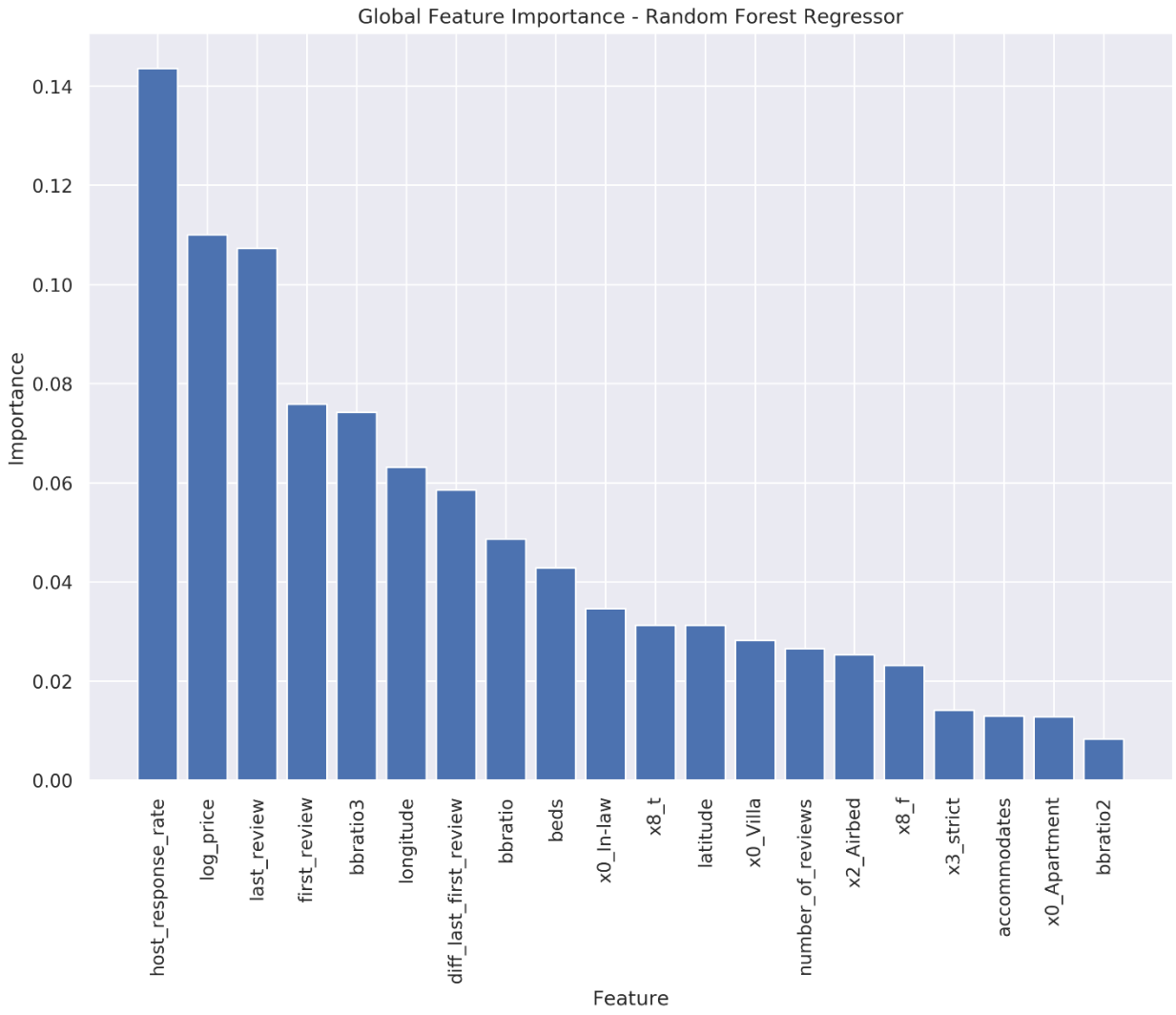

Figure 2: Plot of Global Feature Importance: the plot shows the importance scores of top 20 features in the model of random forest regression.

Outlook

The result that it is hard to accurately predict the rating of a house on Airbnb could be explained by the distribution of the ratings. As I mentioned before, the median of the ratings is 95 and the mean is 92.3. Both of them are pretty high. As the scores are pretty condense, no matter what the features look like, it is harder to predict accurately. This illustrates that nearly all of the hosts work hard to keep the ratings as high as possible and it is hard to predict the house prices under this situation as varying features might lead to similar target results. It is like a model which predicts if a person has cancer. If we do not rely on any technique and just assert that nobody has cancer, we might be 95% or even 99% accurate because most of the people in the world do not have cancer. However, if we use technology, the accuracy might worsen. Similarly, if we just predict any random house to have a 90ish score, we might be more accurate than a machine learning model. We could also draw a conclusion that not every quantifiable thing could be predicted well quantifiably, because we might not take deciding qualitative factors into account and data might not tell the most important secret.

Going back to what I mentioned earlier, the hosts might be under pressure of keeping ratings high. But the rating should be an objective, accurate and comprehensive tool to evaluate a house. Thus, on the one hand, the rating system motivates the hosts to provide the best services. On the other hand, it also leads to untrustworthiness to some degree. But it is not my point that rating is impossible to predict. Instead, in the future studies, there is still some more work we could do to improve our models. For example, we could collect more data other than the six major U.S cities, such as Providence, a smaller but beautiful capital of Rhode Island. We could also collect data of cleanliness of houses and the value for the price. We can assign cleanliness scores from 0 to 10. Another concern is that it might be hard to decide if a house is too expensive, and thus, we could value houses based on a ratio between the price of one night and the real price of the real estate.

References

https://www.kaggle.com/rudymizrahi/airbnb-listings-in-major-us-cities-deloitte-ml

https://ieeexplore.ieee.org/document/4811689

https://www.forbes.com/sites/sethporges/2016/06/29/the-one-issue-with-airbnb-reviews-that-causes-hosts-to-burnout/#5cc054a71eb3