How to succeed as an Airbnb Host

——Evidence from Text analysis and Machine Learning Approach

Yutian Lai
University of Chicago
5/26/2020

Abstract

Airbnb, one of the most outstanding marketplace in the sharing economy that facilitates peer-to-peer communication and trust, provides an alternative accommodation experience for its consumers as well as a challenge to hosts to strategically position themselves to attract travelers. To provide better guidance for hosts, this paper explores factors that affect the review score rating, which could serve as a metric to infer the satisfaction level of guests, through natural language processing and machine learning framework. The geospatial difference would also be detected by comparing the three localities of Tokyo, New York City, and London. The result implicated cleanliness, communication between hosts and guests, location and verification of super host are the main influencing factor in the final rating score across regions, and the three localities also display significant differences in terms of factors framing the results.

1. Data

1.1 Data Collection

The Airbnb dataset of the three cities was obtained at http://insideairbnb.com/get-the-data.html , recently updated on April, 2020. For each locality, the dataset contains information about 15000 host homes with 80 features describing almost all aspects of them, including host self-introduction, neighborhood description, number of bathrooms/bedrooms, etc.

1.2 Data Preprocessing

I divided the data preprocessing into four steps. First, I dropped features without sufficient data or unrelated to the response variable. Second, I created new variables based on current variables for the machine learning algorithm application. For instance, I extracted the text-form amenities information into several binary variables to indicate whether the host home has certain amenities. Third, I modified current variables to help them perform more effectively in machine learning algorithms. For example, I encoded certain numerical variables into binary variables(for instance, the variable “security_desposite” changes from “cost of the security deposit” to “whether the host requires security deposit”) as the availability of such information could be vital deal-breakers in guests' decision. Finally, I split and create gram features of the text of self_about(hosts' self-introduction) and house_description(description of the host home) so bi-gram analysis can be performed.

2. Methods

I plotted the distribution of the response variable: review_scores_rating in the three localities, confirming previous research regarding the skewness of rating towards high scores(Tussyadian & Zach, 2017). In all three localities, review scores center around 90 and peak near 100. Such dramatically high score is especially evident in New York City, with its low-score(<=85) covering only 5% of all homes, while the other two cities have more than 10% low-rating homes. Since there are still homes receiving scores relatively lower, the problem I aim to address in my paper is to understand what features can interpret the ratings of Airbnb homes.

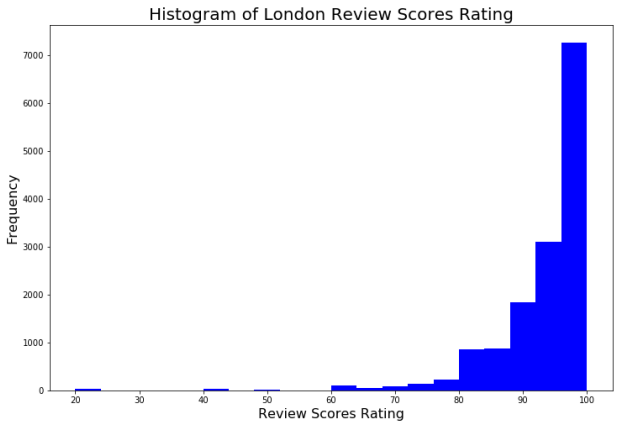

Figure1. Histogram of London Review Scores Rating

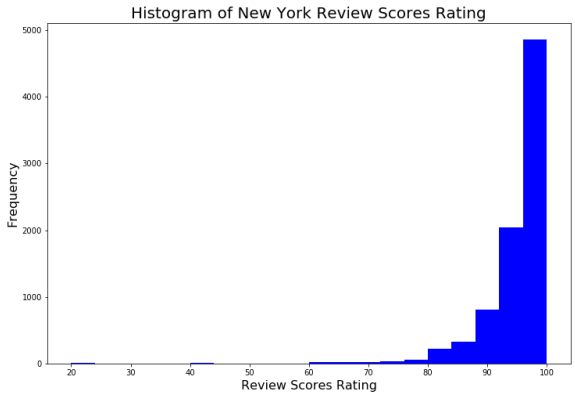

Figure 2. Histogram of New York Review Scores Rating

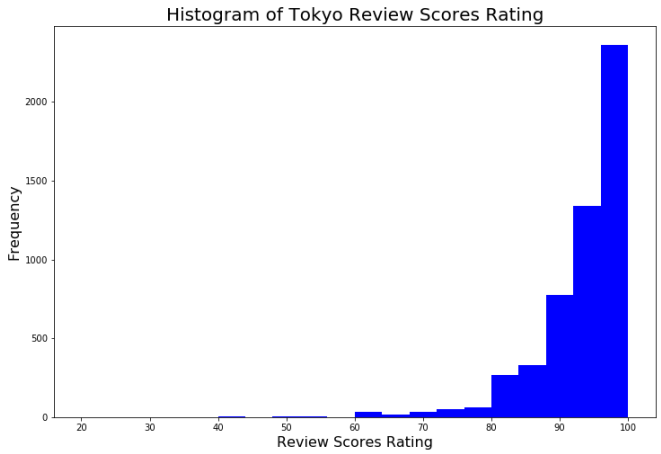

Figure 3. Histogram of Tokyo Review Scores Rating

2.1 Machine learning techniques

2.1.1 Preprocessing

To understand the variable correlation of my datasets, I drew heatmaps of the numerical features and the response variables of the three localities respectively (See Appendix 2). The interdependency of different features is not too strong. I have tested removing some features to reduce collinearity and keeping all of them, and it proves out that the performance of the latter is better. So no feature is moved after the collinearity analysis.

For the machine learning algorithms to perform more accurately, I conducted data normalization using standard scaler for numerical variables, deployed one hot encoding method for categorical features. And after tokenizing the house description and host self-introduction text data into both unigrams and bigrams, I built two pipelines based on the numerical, categorical, and text data. These two pipelines are the same except in the first pipeline, the text data is passed into bag of words function, while in the second pipeline, the text is processed with term frequency–inverse document frequency (TF-IDF) vectorizer. After these pre-processing steps, the three localities respectively have 10051(London), 10052(New York), 10049(Tokyo) features, and 14676(London), 8427(New York), 5288(Tokyo) samples. Finally, I split each datasets into two parts: 75% training set, and 25% test set.

I deployed 4 machine learning models: ridge regression, lasso regression, random forest, and gradient boosting(See Appendix 3), aiming to take advantage of both linear regression and ensemble methods. Among the 4 models, ridge and lasso regression were applied to both

pipelines, whereas random forest and gradient boosting were only applied for the first pipeline, since it proves by the two former models that the first pipeline generates better performance.

For each model fitting, I tuned the hyperparameters by a randomized search method, with "r2(proportion of variance in the response variable that is explained by the predictors)" as the scoring metric to judge results of different models, and optimized these models using 5-fold cross-validation with 10 iterations on the training set. For ridge and lasso, the parameter I optimized is alpha; in random forest model, I optimized the number of estimators, minimum sample split, and minimum sample leaf; in gradient boosting model, I optimized the number of estimators, learning rate, max depth, and alpha.

2.2. Bigram Analysis

I chose bigram analysis because rather than unigram/simple word counts, it allows researchers to investigate the context in which certain words are used. And to make comparative text analysis so as to understand how "successful" Airbnb hosts strategically market themselves, I divide the three datasets respectively into two groups of high review scores home(review_scores_rating=100) and low review scores home (review_scores_rating<=85), and perform bigram analysis on the host description and host self-introduction text of these sub-datasets.

3. Results

3.1 Machine Learning Algorithm Results

	Training r2	**Test r2**	**Mean Squared error**	**Max error**	**Mean Absolute error**
Ridge1	0.84	0.61	29.1	59.34	3.66
Ridge2	0.81	0.66	25.55	59.22	3.3
Lasso1	0.68	0.66	25.33	58.67	3.2
Lasso2	0.67	0.65	25.86	57.72	3.24
Random Forest	0.9	0.68	24.02	59.88	2.94
Gradient Boosting	0.84	0.69	23.1	59.74	2.93

Table 1. Model Performance of London Airbnb Homes

	Training r2	Test r2	Mean Squared Error	Max Error	Mean Absolute Error
Ridge1	0.74	0.68	12.56	32.54	2.39
Ridge2	0.69	0.67	12.73	31.39	2.37
Lasso1	0.66	0.65	13.4	31.82	2.43
Lasso2	0.65	0.64	14.02	32.2	2.47
Random Forest	0.87	0.62	14.78	29.81	2.38
Gradient Boosting	0.86	0.65	13.49	31.44	2.31

Table 2. Model Performance of New York Airbnb Homes

	Training r2	Test r2	Mean Squared error	Max error	Mean Absolute error
Ridge1	0.74	0.71	14.05	30.93	2.63
Ridge2	0.7	0.71	14.24	27.45	2.68
Lasso1	0.67	0.7	14.72	28.09	2.72
Lasso2	0.67	0.7	14.85	28.4	2.73
Random Forest	0.87	0.68	15.52	38.64	2.54
Gradient Boosting	0.86	0.7	14.51	30.91	2.59

Table 3. Model Performance of Tokyo Airbnb Homes

These six models have very similar performance in terms of r2 values in all three cities. Owing to the high dimensionality of the datasets, random forest and gradient boosting models are usually substantially overfitted as they have very high r2 around 0.9 in the training set, and the metric falls back around 0.7 in the test set. In linear regression models, the overfitting

problem is slighter. In the London dataset, though encountering the problem of overfitting, gradient boosting still proves out to be the model performing the best, with the highest test r2 of 0.69, and lowest mean squared error of 23.1 among all models. New York and Tokyo both have ridge regression (bad of words) as the best model, with r2 respectively being 0.68 and 0.71, and mean squared error being 12.56 and 14.05.

3.2 Interpreting the Models

	feature	gradient boosting
0	review_scores_cleanliness	0.477381
1	review_scores_communication	0.312448
2	review_scores_location	0.034144
3	number_of_reviews	0.008847
4	x11_f	0.007697
5	x11_t	0.007607
6	street	0.003886
7	much offer	0.003810
8	host_acceptance_rate	0.003129
9	x9_t	0.002587
10	home hand	0.002438
11	home please	0.002102
12	boast large	0.001972
13	original	0.001917
14	area bedroom	0.001807

Table 4. London Airbnb Homes Feature Importance Table (gradient boosting)

	feature	ridge(bad of words)
0	review_scores_cleanliness	2.719609
1	review_scores_communication	1.881774
2	review_scores_location	0.874736
3	x11_f	-0.481701
4	x11_t	0.481701
5	flat	-0.373484
6	also give	-0.254097
7	number_of_reviews	-0.223108
8	exquisite	-0.188712
9	floor first	0.188407
10	anywhere manhattan	-0.180656
11	designer	-0.179872
12	kitchen equipped	-0.175243
13	host_acceptance_rate	-0.171852
14	yankee	-0.168897

Table 5. New York Airbnb Homes Coefficient Weights Table(ridge with bag with words)

	feature	ridge(bad of words)
0	review_scores_cleanliness	3.061630
1	review_scores_communication	1.961779
2	review_scores_location	1.157020
3	x11_f	-0.589857
4	x11_t	0.589857
5	meter	-0.188597
6	x3_Asakusa/Ueno	-0.184333
7	good time	0.183026
8	shared room	0.182400
9	tokyo min	-0.176283
10	place people	-0.168128
11	他可是一个不错的顾问	-0.165896
12	tasting	0.165703
13	narita	-0.162606
14	specialized	-0.161084

Table 6. Tokyo Airbnb Homes Coefficient Weights Table (ridge with bag of words)

For each locality, I performed feature importance/coefficient weights analysis of the model with the best performance in terms of the r2 value. The results of the three cities have a lot in common, as their top features all suggest review score rating is mainly determined by review scores of cleanliness, communication, and location of the Airbnb home, and among the three common factors, cleanliness is valued most by guests. More interestingly, all three localities imply that being a super-host, as indicated by x11_t, significantly helps promote review scores. In terms of other crucial contributing factors, London and New York have much in common, where host acceptance rate and the number of reviews have high rankings, shedding light on what aspects guests pay the most attention to when they choose houses. These two cities also display differences, as New York guests show dislike for certain types of homes, with words indicating special features, such as “exquisite”, “designer” and “yankee” all negatively correlated with review score rating. Distinctive from the two localities, Tokyo guests emphasize more on “location” of Airbnb homes. Certain locations, including “Narita” and “Asakusa/Ueno”, are negatively affecting review scores, indicating guests might prefer to live in other districts.

Since all the three models, especially the gradient boosting model for London Airbnb homes, suffer from overfitting problem that might reduce their reliability, I supplement the current conclusion by analyzing other models(See Appendix 4,5,6). The results again reveal

that review scores of cleanliness, communication, and location are the most influencing features. This time, more conspicuous differences among the three localities emerge. London guests highlight the importance of house amenities and neighborhood around by mentioning “beautiful kitchen”, “modern room”, “furnished room”, and “lovely garden”, etc., and do not value host attributes that much. In New York, the host attributes grab the most attention. Time as host and host acceptance rate rank high in almost every model. In Tokyo, “location” still proves to be the most vital concern, with one group of districts, such as “Shibuya” negatively correlated with the review score, while the other group of districts, such as “Ginza” positively correlated with the review score, exhibiting guests' preference regarding where to stay when they are in Tokyo.

3.3 Text Analysis Results

The top 10 bi-gram comparison between high-score houses and low-score houses across the three cities (See Appendix 7,8) indicates the way “successful” and “failed” hosts position themselves are not substantially different. In terms of house/home description, all hosts addressed the equipment of the houses and convenience of the location. In New York, the high-score homes seem to be more “newly renovated”, so that might be a reason guests are in favor of them. In Tokyo, there appears to be an obvious difference that high-score houses are near “Shinjuku station” while the low-score homes are near “ Ikebukuro station”, corresponding with previous machine learning algorithm results that “location” is an important discriminating factor in Tokyo.

The host self-introduction proves to be quite similar between high-score hosts and low-score hosts, as the top bigrams in both groups suggest a sense of wish/welcome message sending by adopting phrases like “feel free” and “look forward.” This is most applicable to Tokyo hosts as there is no noticeable difference between high and low- score Tokyo hosts’ self-introduction. The other two cities, on the other hand, show that high-score hosts would address their own hobbies (especially the love for “traveling”) in the self-introduction to arouse resonance with the guests, whereas low-score hosts focus more on neutrally introducing the amenities and neighborhood of their houses, using phrases including “walking distance” and “subway ride” without the sense of enthusiasm radiating from effective self-introduction.

Reference

Tussyadiah, I. P., & Zach, F. (2017). Identifying salient attributes of peer-to-peer accommodation experience. *Journal of Travel & Tourism Marketing*, *34*(5), 636-652.

Appendix1: Attributes used in analysis

Numerical: 'accommodates', 'bathrooms', 'beds', 'bedrooms', 'price', 'guests_included', 'extra_people','minimum_nights', 'maximum_nights', 'number_of_reviews', 'review_scores_location', 'review_scores_cleanliness', 'review_scores_communication', 'time_as_host', 'host_response_rate','host_acceptance_rate'

Categorical: 'transit', 'host_has_profile_pic', 'host_identity_verified', 'neighbourhood', 'property_type', 'room_type', 'bed_type', 'security_deposit', 'cleaning_fee', 'instant_bookable', 'weekly_price', 'host_is_superhost', 'host_response_time' and the list of amenities in the corresponding locality

Text: 'self_about', 'house_description'

Response Variable: 'review_scores_rating'

Appendix 2.Heatmaps

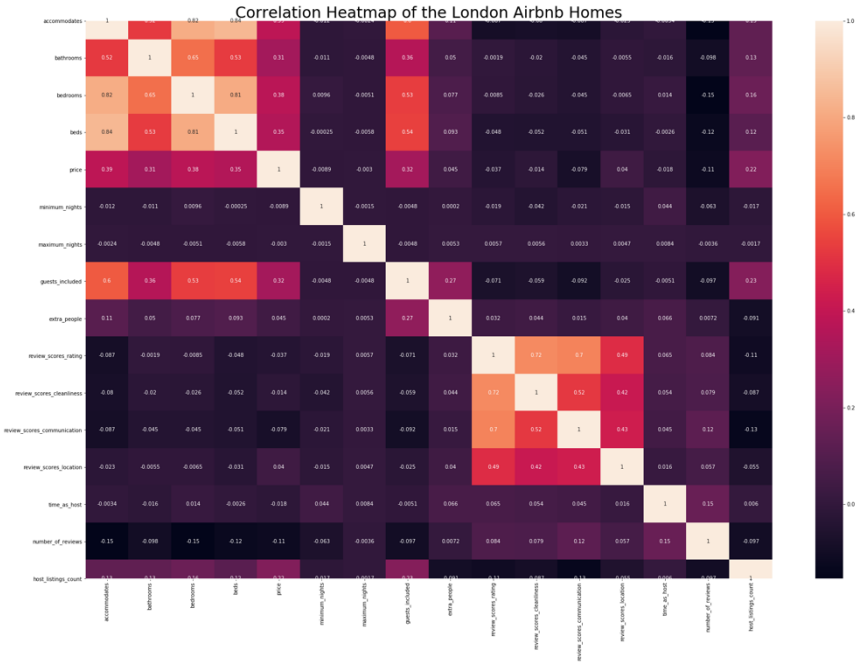

Figure1. Heatmap of London Airbnb Homes

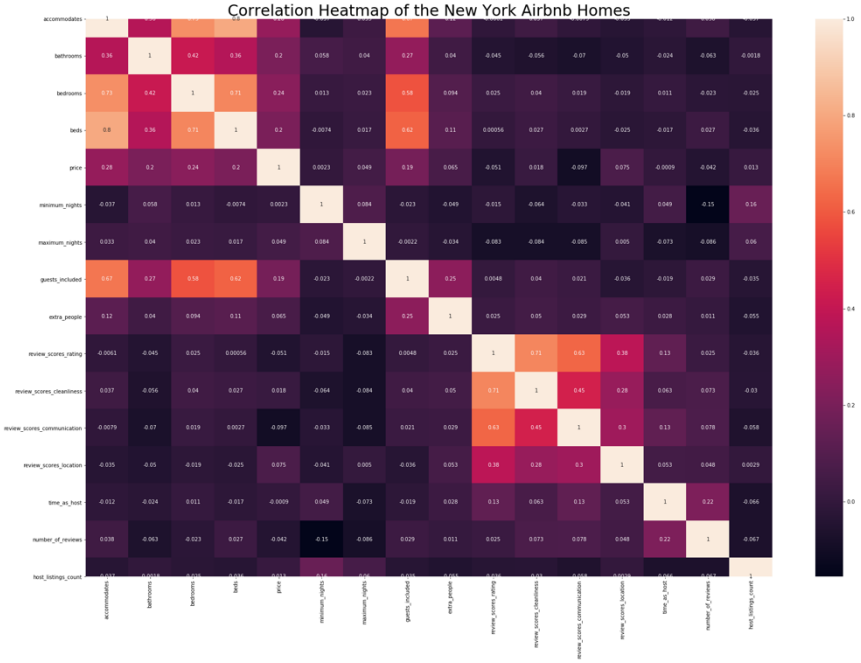

Figure 2. Heatmap of New York Airbnb homes

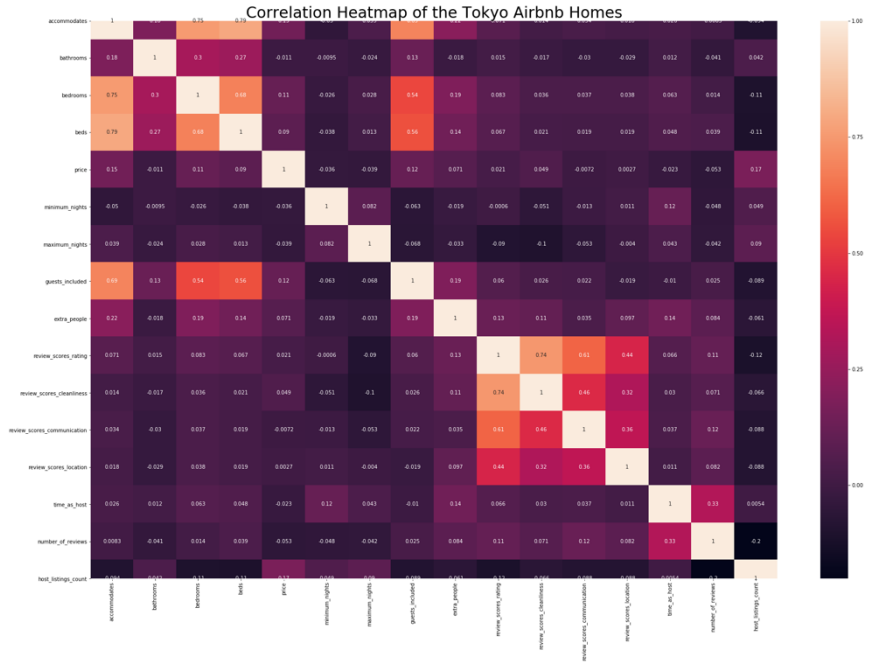

Figure 3. Heatmap of Tokyo Airbnb homes

Appendix 3. Model Introduction

Linear Regression: Ridge (TFIDF) /Ridge(Bag of Words)/Lasso(TFIDF)/Lasso(Bag of Words):

I adopted these two methods, aiming to approximate how linear combination of all the attributes could predict the rating of Airbnb houses. Since I did not remove features based on heatmaps, ridge and lasso could respectively help me conduct feature selection using l2 and l1 regulation to decrease overfitting.

Random Forest Regressor and Gradient Boosting

Random forest and gradient boosting are used owing to the effectiveness of ensemble methods. Random forest could find the best predictors among a random subset of predictors. Gradient boosting train predictors sequentially and finally form the best learner out of all the weak learners.

Appendix 4. Coefficient weights/Feature importance tables of London

	feature	ridge(bad of words)
0	review_scores_cleanliness	3.484939
1	review_scores_communication	3.130619
2	kitchen beautiful	1.626500
3	flat perfectly	-1.536597
4	hob oven	-1.499507
5	walk camden	-1.479074
6	room modern	-1.450935
7	government	-1.410651
8	cost	-1.405379
9	share flat	-1.303574
10	back soon	-1.280412
11	review_scores_location	1.256319
12	furnished bedroom	1.252755
13	much offer	-1.239625
14	ha always	-1.229664

Table 1. London Airbnb Homes Coefficient Weights Table (ridge with bag of words)

	feature	ridge(TFIDF)
0	walk camden	-10.560887
1	government	-9.953712
2	hob oven	-9.893350
3	kitchen beautiful	8.854857
4	professional couple	-8.787529
5	back soon	-8.504729
6	cost	-8.412489
7	bathroom shared	-8.149296
8	hospital	-8.007084
9	queen size	-7.825816
10	ha always	-7.720148
11	able share	-7.655592
12	large comfortable	-7.636694
13	garden lovely	-7.563018
14	warmest regard	-7.400118

Table 2. London Airbnb Homes Coefficient Weights Table (ridge with TF-IDF)

	feature	lasso(bag of words)
0	review_scores_cleanliness	3.725473
1	review_scores_communication	3.242501
2	review_scores_location	0.860570
3	street	-0.279452
4	accommodates	-0.000000
5	bathrooms	0.000000
6	beds	0.000000
7	bedrooms	0.000000
8	price	0.000000
9	guests_included	0.000000
10	extra_people	0.000000
11	minimum_nights	0.000000
12	maximum_nights	0.000000
13	number_of_reviews	-0.000000
14	time_as_host	0.000000

Table 3. London Airbnb Homes Coefficient Weights Table (lasso with bag of words)

	feature	lasso(TFIDF)
0	review_scores_cleanliness	3.735104
1	review_scores_communication	3.382070
2	review_scores_location	0.820763
3	accommodates	-0.000000
4	bathrooms	0.000000
5	beds	-0.000000
6	bedrooms	0.000000
7	price	-0.000000
8	guests_included	-0.000000
9	extra_people	0.000000
10	minimum_nights	0.000000
11	maximum_nights	0.000000
12	number_of_reviews	-0.000000
13	time_as_host	0.000000
14	host_response_rate	0.000000

Table 4. London Airbnb Homes Coefficient Weights Table (lasso with TF-IDF)

	feature	random forest
0	review_scores_cleanliness	0.458438
1	review_scores_communication	0.284273
2	number_of_reviews	0.016785
3	review_scores_location	0.012451
4	x11_t	0.006504
5	time_as_host	0.004991
6	price	0.004446
7	x11_f	0.004435
8	host_acceptance_rate	0.004181
9	street	0.003013
10	know guest	0.002386
11	guests_included	0.002279
12	boast large	0.002104
13	property london	0.001768
14	host_response_rate	0.001606

Table 5. London Airbnb Homes Feature Importance Table (random forest)

	feature	gradient boosting
0	review_scores_cleanliness	0.477381
1	review_scores_communication	0.312448
2	review_scores_location	0.034144
3	number_of_reviews	0.008847
4	x11_f	0.007697
5	x11_t	0.007607
6	street	0.003886
7	much offer	0.003810
8	host_acceptance_rate	0.003129
9	x9_t	0.002587
10	home hand	0.002438
11	home please	0.002102
12	boast large	0.001972
13	original	0.001917
14	area bedroom	0.001807

Table 6. London Airbnb Homes Feature Importance Table (gradient boosting)

Appendix 5. Coefficient weights /Feature importance tables of New York

	feature	ridge(bad of words)
0	review_scores_cleanliness	2.719609
1	review_scores_communication	1.881774
2	review_scores_location	0.874736
3	x11_f	-0.481701
4	x11_t	0.481701
5	flat	-0.373484
6	also give	-0.254097
7	number_of_reviews	-0.223108
8	exquisite	-0.188712
9	floor first	0.188407
10	anywhere manhattan	-0.180656
11	designer	-0.179872
12	kitchen equipped	-0.175243
13	host_acceptance_rate	-0.171852
14	yankee	-0.168897

Table 1. New York Airbnb Homes Coefficient Weights Table (ridge with bag of words)

	feature	ridge(TFIDF)
0	review_scores_cleanliness	3.023066
1	review_scores_communication	2.012767
2	flat	-1.306692
3	williamsburg jfk	-0.953300
4	exquisite	-0.939776
5	x5_Private room	0.933641
6	globe	-0.912901
7	review_scores_location	0.844676
8	also give	-0.830433
9	quiet comfortable	-0.826645
10	radio	-0.819729
11	fun fulfilling	0.801051
12	designer	-0.798140
13	stunning	-0.779048
14	listen	0.722349

Table 2. New York Airbnb Homes Coefficient Weights Table (ridge with TF-IDF)

	feature	lasso(bag of words)
0	review_scores_cleanliness	3.026230
1	review_scores_communication	2.028487
2	review_scores_location	0.763526
3	x11_f	-0.603290
4	flat	-0.329374
5	host_acceptance_rate	-0.199723
6	number_of_reviews	-0.180043
7	exquisite	-0.159862
8	globe	-0.143259
9	rare find	-0.134709
10	stunning	-0.127870
11	williamsburg jfk	-0.123344
12	time_as_host	0.112196
13	fun fulfilling	0.085956
14	importantly	0.069753

Table 3. New York Airbnb Homes Coefficient Weights Table (lasso with bag of words)

	feature	lasso(TFIDF)
0	review_scores_cleanliness	3.073856e+00
1	review_scores_communication	2.064565e+00
2	review_scores_location	7.147882e-01
3	x11_f	-6.423078e-01
4	host_acceptance_rate	-2.268486e-01
5	number_of_reviews	-1.797785e-01
6	time_as_host	1.441641e-01
7	price	-2.451104e-02
8	minimum_nights	2.121193e-02
9	x11_t	4.745758e-16
10	accommodates	-0.000000e+00
11	bathrooms	0.000000e+00
12	beds	-0.000000e+00
13	bedrooms	-0.000000e+00
14	guests_included	-0.000000e+00

Table 4. New York Airbnb Homes Coefficient Weights Table (lasso with TF-IDF)

	feature	random forest
0	review_scores_cleanliness	0.558040
1	review_scores_communication	0.134954
2	number_of_reviews	0.033551
3	review_scores_location	0.020946
4	x11_t	0.010725
5	time_as_host	0.008517
6	host_acceptance_rate	0.006884
7	price	0.005634
8	also give	0.005186
9	africa	0.004817
10	x11_f	0.004374
11	exquisite	0.003065
12	extra_people	0.002910
13	minimum_nights	0.002790
14	maximum_nights	0.002602

Table 5. New York Airbnb Homes Feature Importance Table (random forest)

	feature	gradient boosting
0	review_scores_cleanliness	0.596050
1	review_scores_communication	0.151098
2	review_scores_location	0.021239
3	number_of_reviews	0.020770
4	x11_f	0.015491
5	ac heat	0.012821
6	host_acceptance_rate	0.005333
7	x11_t	0.005172
8	price	0.004950
9	time_as_host	0.004365
10	minimum_nights	0.003824
11	tourism	0.002751
12	microwave coffee	0.002645
13	alumnus one	0.002255
14	travel around	0.002103

Table 6. New York Airbnb Homes Feature Importance Table (gradient boosting)

Appendix 6. Coefficient weights/Feature importance tables of Tokyo

	feature	ridge(bad of words)
0	review_scores_cleanliness	3.061630
1	review_scores_communication	1.961779
2	review_scores_location	1.157020
3	x11_f	-0.589857
4	x11_t	0.589857
5	meter	-0.188597
6	x3_Asakusa/Ueno	-0.184333
7	good time	0.183026
8	shared room	0.182400
9	tokyo min	-0.176283
10	place people	-0.168128
11	他可是一个不错的顾问	-0.165896
12	tasting	0.165703
13	narita	-0.162606
14	specialized	-0.161084

Table 1. Tokyo Airbnb Homes Coefficient Weights Table (ridge with bag of words)

	feature	ridge(TFIDF)
0	review_scores_cleanliness	3.665994
1	review_scores_communication	2.133918
2	review_scores_location	1.077961
3	x3_Asakusa/Ueno	-0.872699
4	takeko	-0.756255
5	x11_t	0.715077
6	x11_f	-0.715077
7	great length	-0.691162
8	여행하게된	0.638326
9	super	0.625475
10	x3_Shibuya District	-0.617687
11	thing give	-0.602274
12	situation hesitate	0.595667
13	place people	-0.590912
14	ginza tokyo	0.582963

Table 2. Tokyo Airbnb Homes Coefficient Weights Table (ridge with bag of TF-IDF)

	feature	lasso(bag of words)
0	review_scores_cleanliness	3.746943
1	review_scores_communication	2.146220
2	x11_f	-1.065800
3	review_scores_location	0.966697
4	accommodates	0.208355
5	bathrooms	0.105436
6	time_as_host	0.100183
7	time one	-0.096751
8	golden gai	0.079355
9	exclusive	0.077824
10	extra_people	0.076730
11	여행하게된	0.063049
12	host_acceptance_rate	-0.062873
13	land liberty	0.056137
14	super	0.053305

Table 3. Tokyo Airbnb Homes Coefficient Weights Table (lasso with bag of words)

	feature	lasso(TFIDF)
0	review_scores_cleanliness	3.762771
1	review_scores_communication	2.153907
2	x11_f	-1.102462
3	review_scores_location	0.961990
4	accommodates	0.239803
5	time_as_host	0.133888
6	bathrooms	0.109817
7	extra_people	0.089086
8	host_acceptance_rate	-0.080032
9	minimum_nights	0.057841
10	maximum_nights	-0.043366
11	guests_included	0.031561
12	x13_0	-0.031250
13	x15_0	0.021800
14	beds	0.002221

Table 4. London Airbnb Homes Coefficient Weights Table (lasso with TF-IDF)

	feature	random forest
0	review_scores_cleanliness	0.621555
1	review_scores_communication	0.094876
2	number_of_reviews	0.029014
3	review_scores_location	0.020108
4	x11_f	0.010315
5	x11_t	0.008375
6	time_as_host	0.007953
7	price	0.004564
8	accommodates	0.004131
9	host_acceptance_rate	0.003854
10	extra_people	0.003491
11	shop around	0.002676
12	beds	0.002511
13	directly connected	0.002470
14	我目前在东京拥有一家广告代理商	0.002390

Table 5. Tokyo Airbnb Homes Feature Importance Table (random forest)

	feature	gradient boosting
0	review_scores_cleanliness	0.608914
1	review_scores_communication	0.121688
2	review_scores_location	0.039136
3	number_of_reviews	0.025415
4	x11_f	0.013009
5	x11_t	0.012806
6	yen	0.005123
7	time_as_host	0.004271
8	price	0.003901
9	minimum_nights	0.002747
10	theme	0.002662
11	저희부부가	0.002587
12	accommodation	0.002499
13	accommodates	0.002440
14	directly connected	0.002274

Table 6. Tokyo Airbnb Homes Feature Importance Table (gradient boosting)

Appendix 7. Bigram of house description

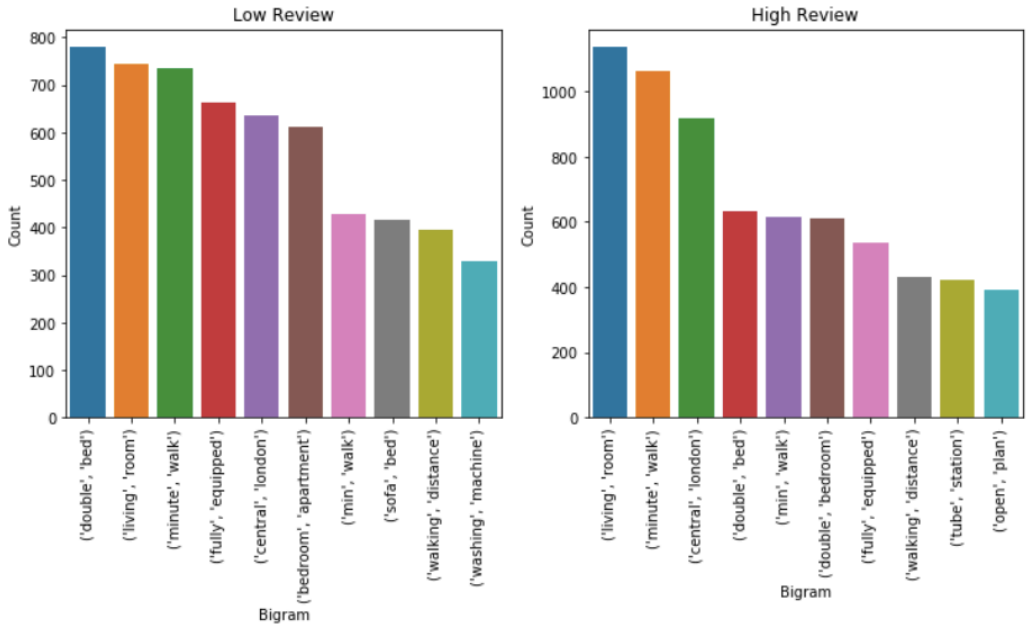

Figure 1. London House Description Bi-gram of Low-score and High-score Homes

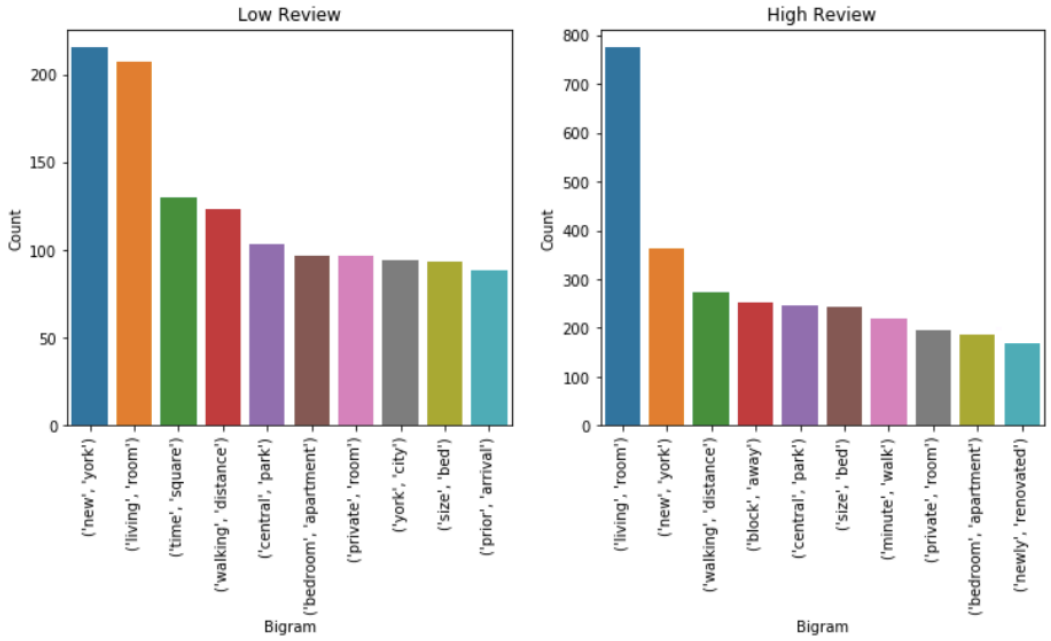

Figure 2. New York House Description Bi-gram of Low-score and High-score Homes

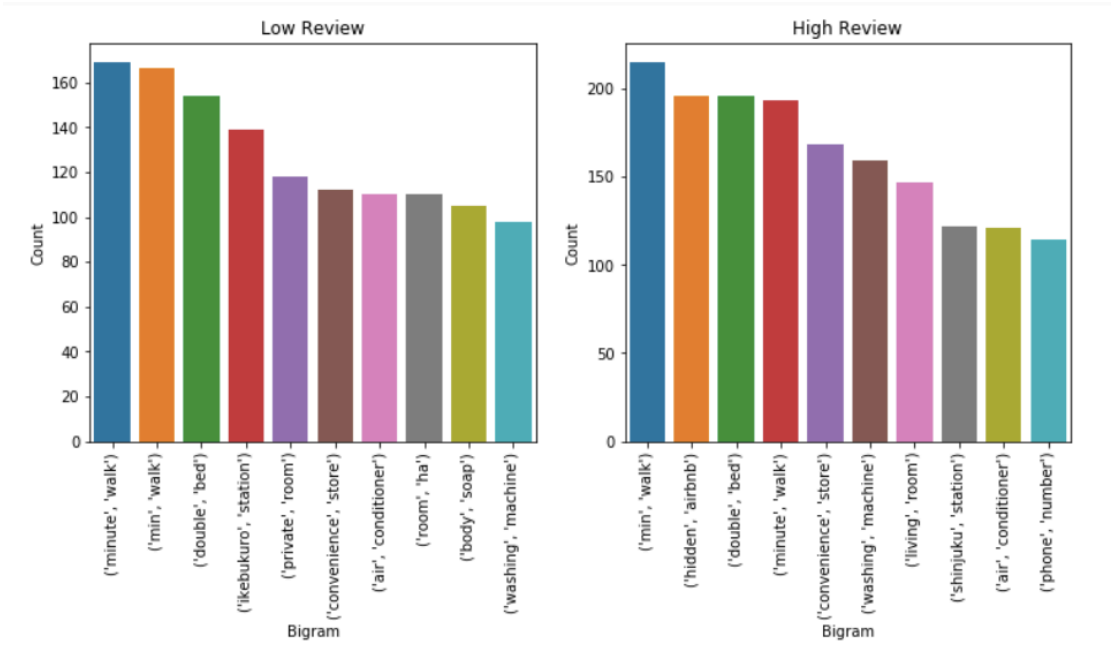

Figure 3. Tokyo House Description Bi-gram of Low-score and High-score Homes

Appendix 8. Bigram of host self-introduction

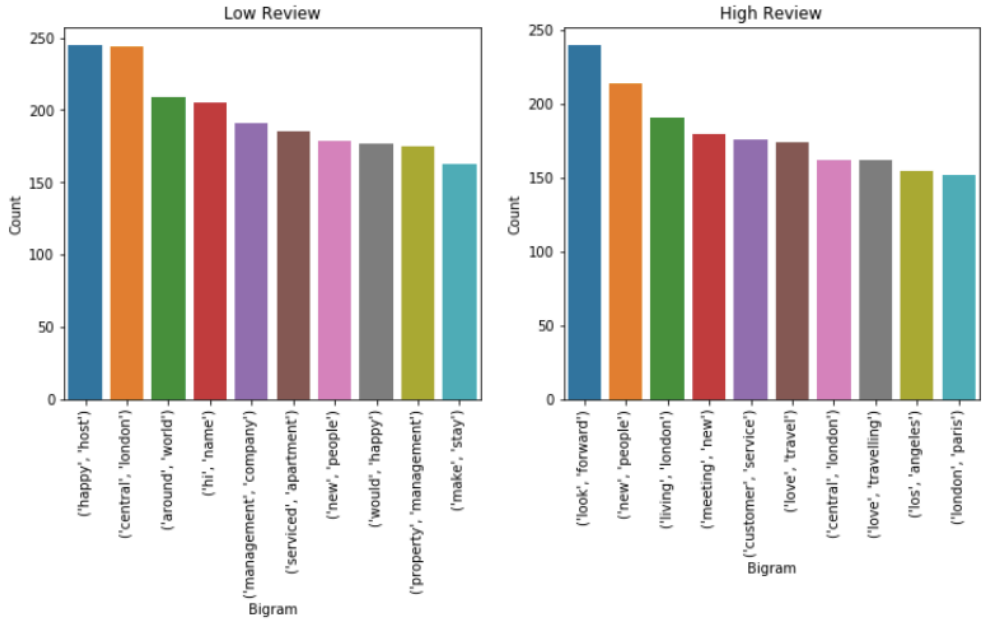

Figure 1. London Host-introduction Bi-gram of Low-score and High-score Homes

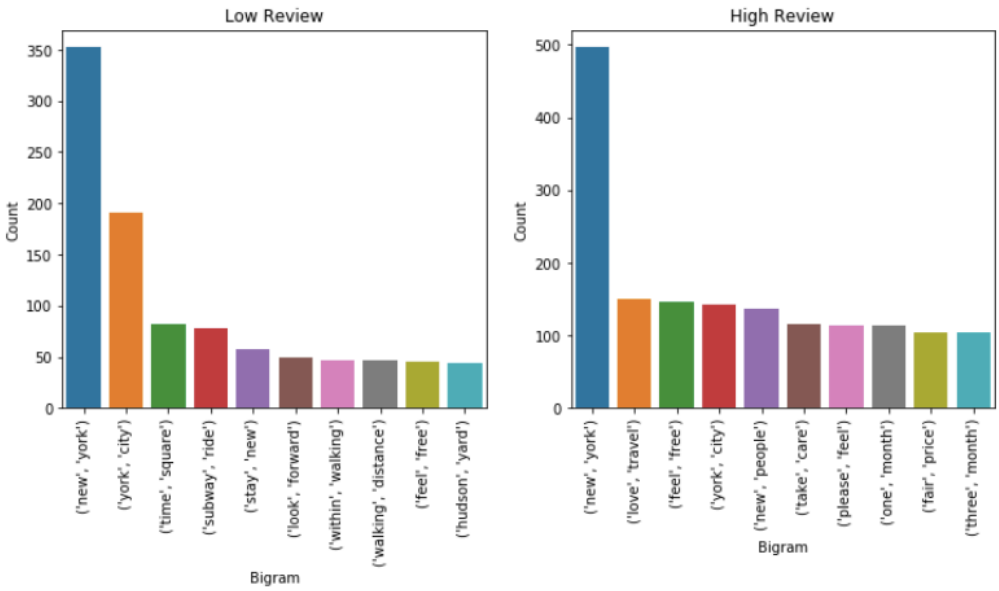

Figure 2. New York Host-introduction Bi-gram of Low-score and High-score Homes

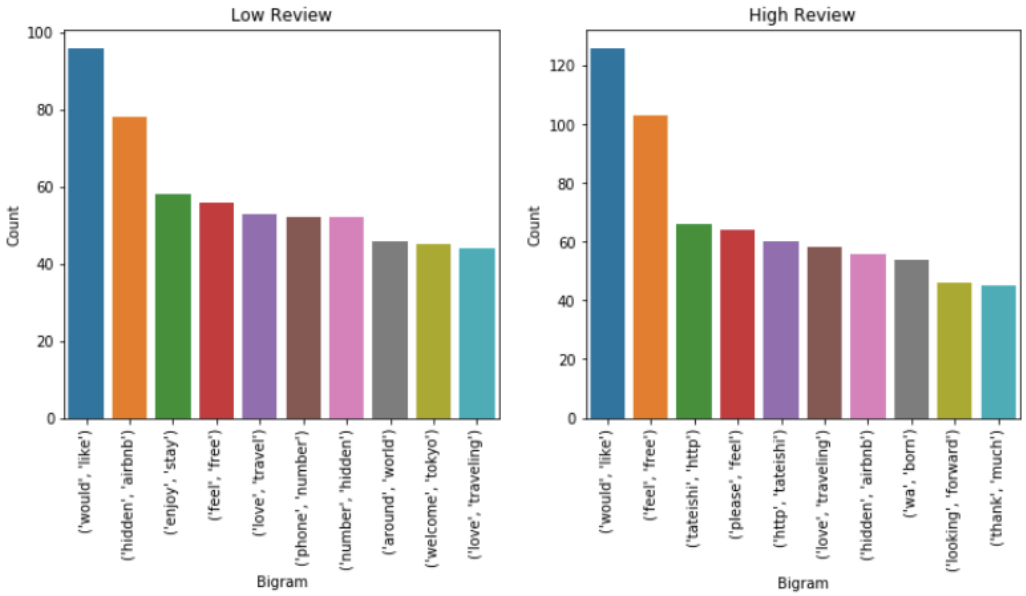

Figure 3. Tokyo Host-introduction Bi-gram of Low-score and High-score Homes