Increasing and Projecting Airbnb Review Counts

Kit Hui

**Table of Contents**

1. **Abstract. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**
2. **Introduction. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .3**
3. **Data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**
4. **Data Wrangling. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4**

4a. Dataset Cleaning. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .4

4b. Text Preprocessing. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .5

1. **Exploratory Analysis and Statistical Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .5**

5a. Cost. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

5b. Transportation and Host Identity. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

5c. Bedroom. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .6

5d. Time as Host. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .7

5e. Rating. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

5f. Bigram Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10

1. **Machine Learning. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .11**

6a. Preprocessing/Pipeline. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .11

6b. Models/Evaluation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .12

1. **Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13**
2. **Supporting Figures/Tables. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .14**
3. **Abstract**

Airbnb is a great platform for homeowners to rent out their residence and earn passive income; however, it could be difficult for homeowners to market an effective narrative to attract travelers. To better provide guidance for homeowner, this project explores factors that affect the number of review (a metric used to determine indirectly the lodging counts) through visual analysis and machine learning and predict the performance of homes. This project would create a universal pipeline for all homeowners thinking of marketing their homes on Airbnb and help to improve passive income for both Airbnb and homeowners. The project was limited to examining Airbnb homes in Los Angeles, CA and the best r2 value is 0.41.

1. **Introduction**

Airbnb has become a major player in the lodging industry taking 20% of US lodging market share in just 11 years since the company began in 2008. This phenomenon is owed not only to their range of economic lodging options, especially budgeted options, but also the ease of accessing lodging information from all over the world in one single platform. For many tourists, Airbnb also provides an immersive experience into the culture and people of the vacation destination by staying at a local residence instead of the similar hotel/motel experience. For these reasons, Airbnb have experienced a meteoric rise in popularity among youth travelers. On the other hand, hosts also enjoy the passive income that comes with renting out their home/space and potentially befriending world travelers that decides to stay at their venue. However, putting up their homes on Airbnb can be a complicated task and can be difficult to predict how their home will compete with neighboring Airbnb homes in attracting customers. Therefore, in this project, I will examine the number of reviews and how it correlates with description, the geographical location, price and other home characteristics that may lead to improvement in review counts. Number of reviews was selected because the Airbnb dataset does not provide numbers of previous occupants as an obtainable feature and numbers of reviews, in the general sense, typically correlates well with number of consumers that previously used the service.

Two main audiences would have an interest in the project: the hosts and Airbnb. Hosts can directly benefit from the study to condition their homes to high number of reviews. Based on the results coming out of the algorithm, hosts can improve their homes or language used when posting their home information or even how their home would predictably do in the market given some fixed conditions like bedroom numbers and geographical location. With the information the host can adjust operational cost as well as invest in amenities for a long-term profit like buying a television for the room. On the other hand, Airbnb also would also benefit from the study because it effectively tells a story on how the market is behaving for consumer interested in Airbnb services. Airbnb can use this and customer’s data to hypothesize a wholesome picture of the demand and supply of certain types of homes and use this information to motivate hosts to have certain characteristics to improve their services and drive business growth.

1. **Data**

The dataset was obtained at <http://insideairbnb.com/get-the-data.html> and the detailed Los Angeles listing were used. The listing dataset contains the general information about the host home like number of bedrooms, description of the home, price, etc as far back as 2008. This dataset has over 100 features and 45000 rows. Although the dataset do have many features, this dataset does have important features that are missing features which could negatively impact the prediction, including but not limited to time series of amenities offered, active days for which home is available for rent, etc.

1. **Data Wrangling**

The data wrangling steps could fundamentally be split into two portions. First, structure data cleaning where features unrelated to target feature (number of reviews) and features with insufficient data were removed. Second, unstructured data where the NLP of the text data processing were conducted.

1. DataFrame Cleaning

a. dropped unarguably unimportant features

b. feature engineered dataset (create new features and modify existing features)

2. Text Preprocessing

**4a. Dataset Cleaning**

Features were discarded based on one of three reasons. First, the feature is unrelated to target feature, second, the feature is a duplicate of another feature, and third, the feature has insufficient data to lead to an effective analysis/machine learning algorithm.

Features removed are in the following list: 'id', 'listing\_url', 'scrape\_id', 'summary', 'space', 'neighborhood\_overview', 'notes', 'access', 'interaction', 'house\_rules', 'thumbnail\_url', 'medium\_url', 'picture\_url', 'xl\_picture\_url', 'host\_url', 'host\_name', 'host\_location', 'host\_about', 'host\_response\_time', 'host\_response\_rate', 'host\_acceptance\_rate', 'host\_thumbnail\_url', 'host\_picture\_url', 'host\_neighbourhood', 'host\_listings\_count', 'host\_verifications', 'street', 'neighbourhood\_cleansed', 'neighbourhood\_group\_cleansed', 'city', 'state', 'zipcode', 'market', 'smart\_location', 'country\_code', 'country', 'latitude', 'longitude', 'is\_location\_exact', 'square\_feet', 'monthly\_price', 'minimum\_nights', 'maximum\_nights', 'minimum\_minimum\_nights', 'maximum\_minimum\_nights', 'minimum\_maximum\_nights', 'maximum\_maximum\_nights', 'minimum\_nights\_avg\_ntm', 'maximum\_nights\_avg\_ntm', 'calendar\_updated', 'has\_availability', 'availability\_30', 'availability\_60', 'availability\_90', 'availability\_365', 'calendar\_last\_scraped', 'number\_of\_reviews\_ltm', 'first\_review', 'last\_review', 'review\_scores\_accuracy', 'review\_scores\_checkin', 'review\_scores\_checkin', 'review\_scores\_location', 'review\_scores\_value', 'requires\_license', 'license', 'jurisdiction\_names', 'is\_business\_travel\_ready', 'cancellation\_policy', 'require\_guest\_phone\_verification', 'require\_guest\_profile\_picture', 'calculated\_host\_listings\_count', 'calculated\_host\_listings\_count\_entire\_homes', 'calculated\_host\_listings\_count\_private\_rooms', 'calculated\_host\_listings\_count\_shared\_rooms', 'experiences\_offered', 'reviews\_per\_month', 'host\_total\_listings\_count', 'host\_id'.

In the working features, ‘transit’, ‘security\_deposit’, ‘cleaning\_fee’, and ‘weekly\_price’ were modified to just a categorical feature by determining whether any information was provided. This is because for travelers some of these features could be a dealbreaker. For example, if travelers do not plan to drive around, information on easy public transportation is essential to moving around Los Angeles. Fees on top of renting the place is could also be a dealbreaker for people.

‘Time\_as\_host’ was engineered from ‘host\_since’ and ‘last\_scraped’. Although this can potentially be an important feature, host since may be misleading because the host can have multiple residence on Airbnb and the home posted is put up recently on Airbnb. Also. homes could be put up recently after a long hiatus could potentially cause major variance when comparing time as host and a response variable.

The data structure of “price” and “extra\_people” were modified from string to integer.

The review scores of rating, communication and cleanliness were filled with 0 in place of null values to represent locations that did not receive ratings.

All other homes with null values were dropped.

**4b. Text Preprocessing**

Only the ‘name’ and ‘description’ of the home were processed. The NLP are in the following steps:

1. Include only English texts
2. Lowercase all text
3. Tokenized sentences to words
4. Lemmatize text
5. Remove all stops words and numbers

2-gram features for ‘name’ and ‘description’ were also created.

1. **Exploratory Data and Statistical Analysis**

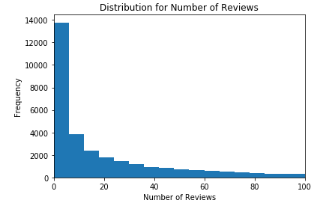
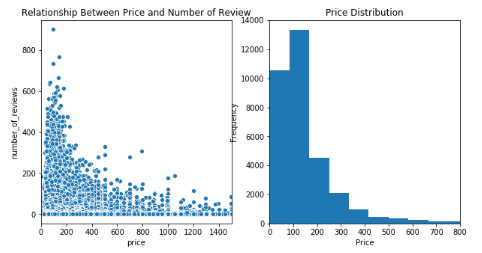
Fig 1. reveals that 50% of Airbnb homes receives less than 10 review counts with a mean of 35 number of reviews. The histogram quickly tappers off in an exponential decay fashion, showing an extremely left skewed distribution. The distribution shows most review counts concentrates in the top 25% at 42 counts, quadrupling the count at 10%, suggesting that renting out Airbnb may follow the 80/20 rule where 80% of the income goes to the top 20% Airbnb homes. Also, home with 0 reviews accounts for 17% of the data and home most review have 900 counts.

Figure : Distribution for Review Count

**5a. Cost**

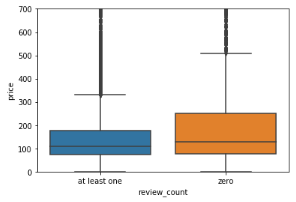
79% of all Airbnb homes are priced between $50 and $300 and this price region is most in supply as suggested by the distribution plot in Fig. 2. This price range also has the densest number of reviews at above 200 counts and most in number of reviews by count having 88% of all review counts. According to calculations this region contains 81% of all Airbnb homes with at least one review whereas it contains 71% of all Airbnb homes with 0 reviews. This not only suggests that this region is most in supply but also is the most in demand with smaller proportion of 0 reviews. Fig. 3 shows in the general scale, homes with at least one review tend to be slightly more economical than that of zero review looking at the median cost and substantially cheaper at the 95% percentile by a difference of about $150.

Figure : (left) Scatter plot of price and review counts. (right) Distribution of Price

Figure :Box plot of review count and price

**5b. Transportation and Host Identity**

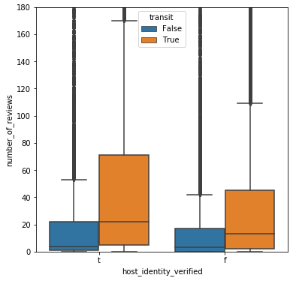
 In general, providing transportation information plays a significant role in improving number of reviews shown in Fig 4. The mean number of reviews for verified hosts with with transportation information is more than 2 times higher than those without at 53 and 23 respectively. The difference is particularly significant at 75% percentile for host that have their identity verified and provide transportation information, where there is 70 number of reviews compared to merely 20 number of reviews.

Figure : Host verification and number of reviews. t: True, f: False

The median for number of review is 4-8 times higher for those that provide transportation information at 21 and 18 number of reviews for verified host with transportation information and unverified host with transportation information. Although the bar plots appears to show weak correlation between identified host and number of reviews the mean proves otherwise where the mean number of reviews for verified host is 42 compared that of unverified at 28.

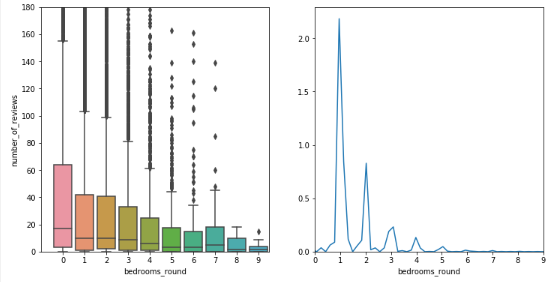
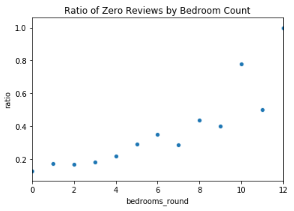
**5c. Bedroom**

Figure : (left) Bedroom Counts and number of Reviews. (middle) Distribution of Bedroom Counts. (right) ratio of zero review by bedroom counts

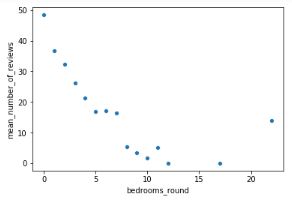
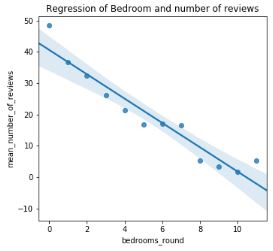
Focusing on the median of number of reviews for bedroom counts, a clear pattern of depreciation in reviews as bedrooms increase, which tapper off towards large number of bedrooms. (fig 5) This suggests that there is high demand of 1 bedroom Airbnb homes with highest 75% percentile review counts. The bedroom distribution shows that the 1 bedroom have supply 2 times higher than 2 bedrooms and more than quadruple of 3 bedrooms. Further in scatter plot for ratio between 0 reviews and at least one reviews suggests that lower bedroom counts have the highest demand to supply ratio with 1 bedroom at less than 0.2.

Figure : (left) Preprocessed scatter plot of mean number of reviews and bedroom counts. (right) Linear regression between mean number of reviews and bedroom counts

Fig 6 reveals a mostly linear relationship between mean number of review and bedroom counts. The two mean data points at 17 and 23 are only represented by one point each and removed. The resulting plot shows a high level of correlation between bedroom counts and mean number of reviews with a p-value of essentially 0. The slope suggests that for every room increases, a mean loss of 4 review would result.

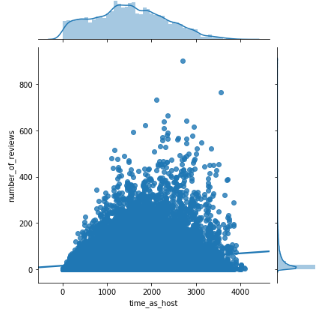
**5d. Time as Host**

Figure : (left) Joint (scatter and histogram) plot of time as host and number of reviews. (right) Violin plot of time as host and 0 review counts/at least one review

The distribution for time as host do not appear to be uniform as expected but peaks between 1500 and 2000 days. (fig. 7) This behavior could be understood in three stages. From right to middle, Airbnb experienced a raise in popularity and homeowners began joining the platform. Then the number of people joined peaked and the platform experienced a deceleration of homeowners joining the platform. Then a constant number of people began joining every year without much growth and saturates out. When examining the violin plot for zero review counts, we observed two peaks where one is around 1500 days and another at around 0, which at first is surprising because it would be logical that as time passes, zero review frequency tends to decrease. These two peaks could be explained logically and the general distribution. The peak close to zero days results from homeowners whom just joined the platform and have experience no or little booking from travelers. On the other hand, the other peak at 1500 days is the result of the saturation in demand while supply of Airbnb homes increase as shown in the general distribution.

The scatter plot shows a general positive correlation considering the regression line and the maximum number of reviews at set time intervals. However, the scatterplot does not show whether the linear regression line is the best regression fit. To measure the underlying behavior of time and number of reviews, mean number of reviews as a function of interval time was conducted.

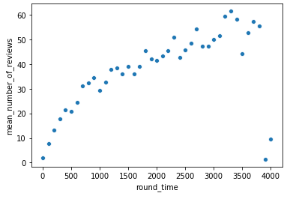
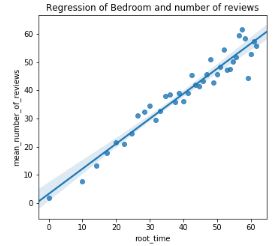




Figure : (left) Scatter plot of time interval as host and mean number of reviews. (right) Linear regression of mean number of reviews and time interval as host after transformation

The scatter plot after processing the data resulted in a seemingly quadratic behavior between time as host and mean number of reviews with two obvious outliers. After removing the outliers, a square root transformation was applied to root time and the result fits well on a linear regression model. The slope suggests for every root square of time, an 0.88 increase in mean number of reviews would result. This transformation was applied to the time as host data

Table 1: Linear Regression of the post-transformation of time as host and number of reviews

After the transformation, the correlation is severely weakened to 0.035, however with a p-value of a negligible number and thus statistically proving the importance of the regression model.

**5e. Rating**

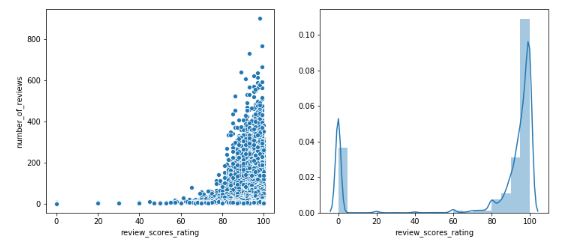


Figure : (left) Scatter plot of overall rating and number of reviews. (right) Distribution of overall rating

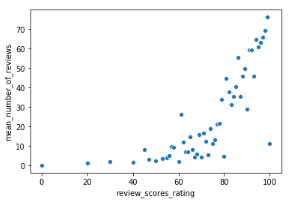
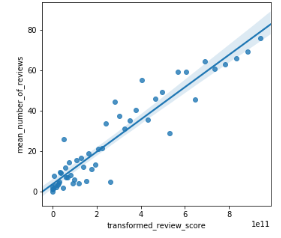
Fig 9 distribution plot reveals that it’s rare for homes to have a review rating of under 80 and over 50% of the Airbnb home have rating above 95. This could be interpreted as the cut off rating for homes to be attractive to market. This is further suggested by the scatterplot showing an exponential increase in maximum review counts as a function of overall rating and a cut off at around 78 using the maximum number of reviews as the indicator. The high number of 0 overall rating is the result of our data processing step where null values were filled with 0, accounting for 17% of the data. To better understand the underlying behavior of number of reviews response to overall score, mean numbers of reviews were measured against the score.



Figure : (left) Scatter plot of overall rating and mean number of reviews. (right) Linear regression of mean number of reviews and overall rating after transformation

The resulting scatterplot suggested an exponential relationship between overall rating and mean numbers of review. The best exponential fit using correlation as the metric was a power raise to the 6th. This transformation resulted in a high correlational fit between mean number of reviews and transformed review score. This transformation was done on the overall rating.

**5f. Bigram Analysis**

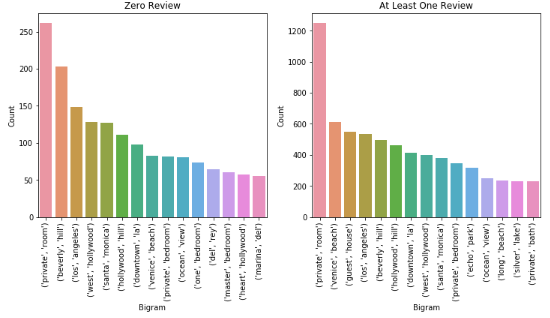


Figure : Bar graph of the most used bigram in the title for Airbnb homes differentiated by zero reviews and at least on review

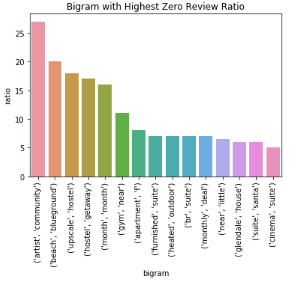
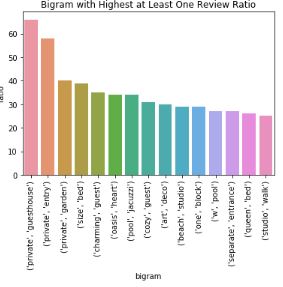
The bar graph for both classifications have common themes between their bigram suggesting that they are not discriminating factors. Both have many location dependent phrases such as west Hollywood and Beverly hills. Both also actively describe the rented room as private. However, homes with at least one review also uniquely mentioned private bathroom as one of the top 15 most used bigram in the title.

Figure : (left) Bar plot showing best bigram used in an effective title (right) Bar plot showing worst bigram used in a title

Since the word count shows little as to suggests what phrases to include in and exclude from titles, graphs with the largest zero review to at least one review bigram ratio and vice versa were plotted (fig. 12). These graphs show the most effective and least effective 2-word phrases to use in the title for attracting consumers. On the left graph, words used in the bigrams such as private and separate suggests a sense of luxury and privacy. They also use words that suggest warmth and hominess with phrases like private garden and cozy guest. On the other hand, most unfavorable words/theme in titles contains words like hostel and suite. Words and phrases seem generic and neutral like “furnished suite” or “gym near” without a sense of passion radiating from effective titles.

1. **Machine Learning**

**6a. Preprocessing/Pipeline**

5 machine learning models were trained which include two ridge regression for bag of words and Tfid, random forest, XG boosting, and deep learning. Prior to training the models, the list of features in the amenities column was extracted and resulted in 292 features. 

Figure 13: Heatmap of Kendall Correlation between Numerical Features

The heatmap (fig. 14) shows the Kendall correlation of numerical features and the target feature (number of reviews). Kendall was chosen because of it can determine nonlinear correlation between features. It shows multiple features with strong collinearly and we should apply some mitigating transformation to reduce interdependency. Two methods could be used to achieve this. PCA and removing features. Because PCA is computationally expensive, removing features was chosen. The resulting models resulted in weaker performance and no feature were dropped.

The pipeline for the machine learning portion starts with the numerical features. Since most data wrangling was already performed on the dataset, the numerical features were normalized using standard scaler for more effective training. The categorical features were encoded through one hot encoding and all features with less than 0.99(1-0.99) variances were dropped. The text data were subjected to two pipelines for the two ridge regressions. In the first pipeline, the text was passed into a bag of words function where the maximum number of features was limited to 10000 and a minimum threshold count of 20 words in the document. The text was tokenized into a unigram and bigram. In the second pipeline, the conditions were the same passing into a Tdif vectorizer.

**6b. Models/Evaluation**

The ridge regression was selected because of the l2 regularization effect to reduce overfitting. 3-fold cross validation with 10 iterations to optimize the alpha term evaluated based on r2 was conducted on the ridge regression.

Training on the random forest algorithm was also performed because of the versatility and effectiveness of ensemble models. A 3-fold cross validation with 10 iterations to optimize n estimators, minimum sample split and minimum samples lefts evaluated based on r2 was conducted on the random forest model.

XG boosting algorithm has been gaining widespread attention and won many Kaggle competition for it’s effectiveness and performance. A 3-fold cross validation with 10 iterations to optimize the n estimators, learning rate, maximum tree depth and alpha evaluated based on r2 was conducted on the XG boosting.

Deep learning models are effective learning algorithms on large dataset. The network was build from hybrid between LSTM and DNN network with 2 fully connected layers. The RNN network was built first by creating a embedding layer with a vocabulary size of 10000 feeding into the LSTM layer. The LSTM has 64 memory cells and dropout rate of 0.3 to prevent variance. The DNN layer starts with 256 units with a relu activation function. This is fed into a second layer with 128 units with a dropout rate of 0.3 and relu activation function. Finally, this is fed into a third layer with 64 units and l2 regularizer. This is followed by a batch normalization to speed up training and prevent vanishing/exploding gradient. The DNN and RNN networks are concatenated and passed into two fully connected network of 64 units. This ends with a final 1-unit layer with a linear activation function.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Training Set r2 | Testing Set r2 | Mean Square Error | Mean Absolute Error |
| Ridge Regression (Bag of words) | 0.4 | 0.3 | 2513 | 31.6 |
| Ridge Regression (Tfid) | 0.38 | 0.29 | 2519 | 31.8 |
| Random Forest | 0.76 | 0.39 | 2192 | 24.3 |
| XG Boosting | 0.62 | 0.38 | 2194 | 26.9 |
| Deep learning | 0.55 | 0.34 | 2336 | 26.1 |

*Table 2: Evaluation of the trained models.*

The performance of the models is shown in table 2. The ridge models have similar performance with the same r2 value and essentially the same mean square error and mean absolute error. Comparing the trained set and testing set, the model is slightly overfitted and may be a result of the curse of dimensionality. Random forest has the best mean absolute error and r2 but with most overfitted model of them all. The deep learning model has the worst r2 value of the advance models, this might be a result from insufficient data for better predictions.

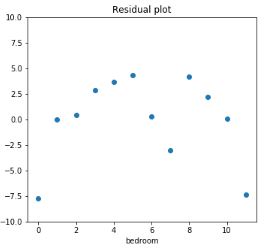
For the purpose of this project, I would choose either random forest since it has the lowest r2 value and mean absolute error. We ignore mean square error because the mathematics behind the calculation biases large deviations by squaring the number. Since high deviations of review counts do not occur frequency, we like to select mean absolute error that is smaller instead with reasonable r2 value. Even though XG boosting has the highest correlation value and relatively small variance, the mean absolute error is high compared to the other two models.

**7 . Future work**

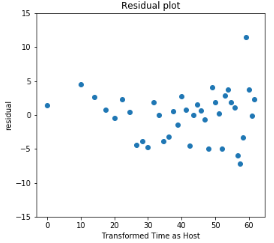
Because of the low r2 value, the trained models remain weak predictors for the number of reviews. The following are possible routes into building a better predictor

1. PCA. Many features exhibit collinearity. We can reduce overfitting and improve model performance with PCA
2. Use images of the residence to determine the contribution from sentiment of image
3. Remove extreme outliers like Airbnb home with over 5 rooms and how it impacts learning metrics
4. Combine Los Angeles datasets with other Airbnb datasets from the US to reduce variance
5. Explore home types and how it impacts other features
6. Remove commercial lodges like suites, hostel and hotel
7. Explore the calendar dataset and look for time series dependencies
8. Stack the existing models to build a more sophisticated ensemble model
9. Train the model with a bidirectional LSTM
10. Fine tune the deep learning model

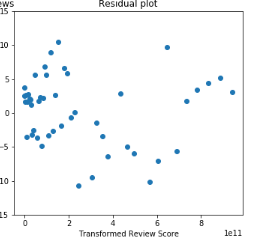
**8 . Supporting Figures**



Supporting Figure 1: residual plot for the linear regression of mean number of reviews and bedroom



Supporting Figure 2: residual plot for the linear regression of mean number of reviews and time as host interval after transformation



Supporting Figure : residual plot for the linear regression of mean number of reviews and overall rating after transformation