Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

Predicting Airbnb Ratings based on Owner Listings

Our task is to predict overall listing ratings on Airbnb based on the different attributes of the rooms and listing descriptions. The goal of our project is to predict the different attributes of the rooms and the ways hosts list their homes that entices guests on Airbnb to give higher ratings. This project could help future hosts cater to the wants of their guests and also help hosts prepare and market their listings most effectively to maximize their ratings and room usage. Airbnb is usually cheaper than staying in hotels in many cities in the US, and more users are beginning to rely on using Airbnb rather than traditional hotel rooms. As more customers begin to use the site, hosts need to understand how the market themselves most effectively and convince users to book their rooms.

We used listings data on Chicago from http://insideairbnb.com/get-the-data.html. Our dataset, after cleaning and dealing with missing values, consisted of 4295 examples and 52 attributes. We split out 500 examples each for testing and validation, leaving 3295 examples for training. The 52 features include qualities about the host (response time, superhost or no), property details (what type of property, number of bedrooms/bathrooms, amenities), as well as counts of words that were from lists we constructed. We used domain knowledge to construct three lists, corresponding to keywords related to Budget, Luxury, or Convenience. Since there were many textual data that consisted of hosts' descriptions of the property, neighborhood, etc., we wanted a succinct way to represent that in the model and give us information on how hosts should market their listings. We used a bag-of-words algorithm from NLTK and Counter to extract the counts of words that were found in our lists. This ended up being a lot more tedious than we thought, where we ran into issues with encoding and understanding how NLTK processed and matched words to our hard-coded lists. This task took much longer than we had anticipated, even after consulting with the Professor and TA's. It also resulted in losing a point on the status progress report.

Before running any machine learning algorithms on the data, we wanted to see if there were any general trends that were present in the data. In particular, we wanted to look at how the word count of different categories of words (i.e. budget, luxury, and convenience words) correlated with the overall review score of the particular listing. We did so using a bag-of-words algorithm variant that we wrote ourselves to count the number of times a word from a particular category appears in a given listing. As seen in Figures 2, 3, and 4, there is not a very strong correlation, but there does appear to be a slight upward trend such that listings with higher word counts that fit into each category tend to be rated higher on average. For budget and luxury words in particular, almost all of the listings with high counts of these words received high ratings from their users, while listings with lower counts have a much more varied distribution of ratings. This observation may be evidence of a correlation between listings that use certain buzzwords (especially for budget and luxury-related words) and user ratings. However, the weakness of the correlation makes it difficult to determine whether these words specifically have an impact on whether a user is more likely to rate a room highly or not.

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

In addition to finding correlations in the data, we also wanted to determine exactly how many missing features our data contained so that we could be aware of the impact that these missing features may have on our model. Most of the features only had a few missing attributes, which we didn't think would greatly impact the results of our experiment, but some of the features (like neighborhood overview and host information) had 1,000-2,000 missing attributes, which is almost 25-50% of the data for that feature (Figure 9). These features are usually optional for hosts to include in their listing, and many of them choose not to do so. Before running the machine learning experiments on our data, we wanted to get an idea of what we were missing in our data set so that we could make hypotheses about how the missing attributes may affect the outcome of our model.

We also wanted to look at a few other potential correlations that could confound the results of our experiment-- namely price vs. rating and number of reviews vs. rating. As seen in Figure 5, listings with lower prices had much more varied review scores while pricier listings were more likely to be rated highly. The price of the listing may be a confounding variable that affects the review score more than the language used in the listing description. Figure 6 shows the correlation between the number of reviews a given listing has and the review score of that listing. Similarly to the other charts, listings with more reviews also tend to be rated higher on average while listings with fewer reviews are much more varied in their review scores. This correlation could also indicate another confounding variable that may affect our results.

Finally, we also wanted to look at whether there was a correlation between the number of luxury/budget words in a listing and the listings price. As seen in Figure 7, there is a surprising downward trend as the number of luxury words increases. This result may be due to hosts using luxury words to persuade users into booking their cheaper rooms and thinking that they are receiving a deal on the listing. Hosts that have listings with higher prices may not feel the need to advertise how luxurious their listing is because the price itself provides an indication of luxury. Figure 8 shows the results of the correlation between budget words and listing price, which is much closer to what we expected to see. As the number of budget words increases, the average price of the listing decreases. The results from these two charts may indicate how hosts use different kinds of words in their listing description to try and entice users to book their rooms. (This part was done by Julia)

With a large number of features, we decided to try two methods -- K Nearest Neighbors and Linear Regression. For KNN, all the features (including property attributes like amenities and number of beds to the number of times certain phrases associated with luxury, budget, or convenience appeared in the host's descriptions) were used. We varied the number of neighbors used with a range of [1,50] as well as the distance measure (Lp norms) with a range of [1,7] and whether the distances were weighted or not. In Figure 11, we see that miscalculation error steadily decreases as number of neighbors increases but also plateaus around high 20's. We then obtained four models that performed best and they were all weighted with L1 norm, with the number of neighbors varying 29, 31, 33, 37. The validation and test set curves are seen in Figure 11. Using weighted L1 norms and number of neighbors set to 31

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

obtained the highest test set accuracy of 0.362. This accuracy is not very high, and is likely due to the sheer number of attributes we trained on and dimensionality issues. As discussed in class, KNN will factor in all the attributes given, and even with penalizing for further data points (weighting), there is likely a lot of noise that deters accurate predictions. We see that weighting is effective in that, across the selected pairs of parameters, that specifying weighting improves accuracy. From this, we hypothesized that having a model that would successfully eliminate training noise and irrelevant attributes could increase testing accuracy, so we decided to experiment with regression. (This portion was done by Judy)

We began with a simple linear regression using all 52 of our original attributes, training the model on our 3925 training examples, with validation and test sets of 500 examples each. Our main goal with linear regression was to understand which features of a listing had the greatest impact on ratings. The mean squared error of the first regression we ran was 36.023 and R-squared was 0.125. In terms of MSE, we found this to be low relative to the mean and standard deviation of our outcome variable. The variables with the largest absolute value coefficients were host response time, a dummy variable for whether or not the host is a "superhost" and a dummy variable for whether or not the listing has internet. It was somewhat surprising that more variables for whether or not the listing had amenities such as TV or a gym didn't have higher coefficients. These results speak to the fact that obtaining a good rating on an airbnb listing are more dependent on the host's behavior than anything else and this is valuable information for hosts! Unsurprisingly, the dummy variables for whether or not the room is shared and smoking is allowed had a pretty negative effect on rating, with values of -1.004 and -1.640 respectively. Of course, we expect to see more listings with lower prices to have shared rooms, but because price was included in our list of features, the model was controlling for price, so at a given price we would expect shared rooms to result in lower ratings and this makes sense. Again, this is valuable information for hosts to improve their ratings, because even at lower price points it may be more beneficial to them to provide accommodations that are not shared to improve ratings and generate more business. The feature with the lowest coefficient was maximum nights a guest can stay at a listing with a value of -.0009. Examining our data, we noticed that 70% of our data had maximum nights values of at least 100. This tells us that the opportunity to stay at a listing for an extended period of time is, in general, not important to how renters rate a listing. Another interesting piece of information we gained from looking at the coefficients in the model is that the coefficients on our bag of words variables pertaining to budget and convenience were negative, while the one for luxury was positive. Hosts can leverage this information to include more language in the descriptions of their listings that suggests luxury, rather than budget or convenience to improve their ratings.

With hopes to improve the accuracy of our results we ran another regression after dropping the features with low variance, using feature selection. This second regression actually did worse with a mean squared error of 39.903 and R-squared value of 0.031. The variables with low variance that were removed were mostly dummy variables equal to 1 if the listing had amenities such as wireless internet, TV, etc. The fact that this model achieved lower accuracy

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

tells us that these features play an important role in determining ratings. When running this model, the coefficient for the host being a superhost rose slightly (by about .1), perhaps pointing to the fact that if we don't control for various amenities, these things are included within the superhost category as more experienced hosts are probably more aware of which amenities are important to include in a listing. (This portion was done by Emma)

In terms of future work, we have a few next steps. The first is to implement feature selection with KNN, and hopefully rule out irrelevant attributes and noise in the training data to get more accurate representations of groupings. With 52 features, it is inevitable that we ran into issues with noise, weighting, and dimensionality. Reducing the number of features could also indicate which ones are truly important in predicting ratings. A second idea is to generalize this to other cities, and see which features are universally important in terms of ratings, and which are more specific to certain cities. The ultimate goal would to be provide a guide to Airbnb hosts of characteristics that affect how their listings are perceived and reviewed. Lastly, it would vastly improve our predictions if we could account for reviewer bias. It seems likely that certain users systematically rate at higher or lower levels. Having some way to account for these differences within users would likely improve estimates of coefficients, especially in linear regression.

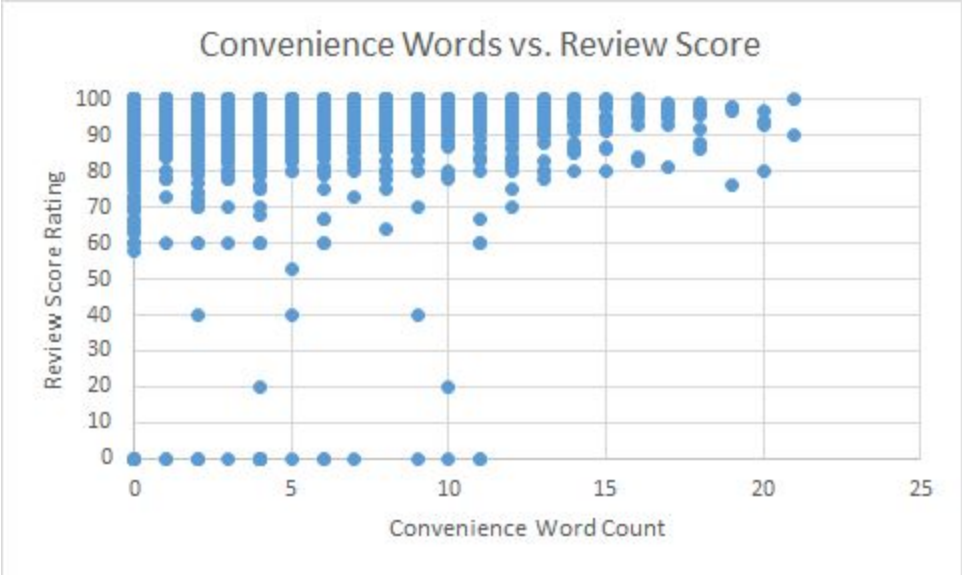

Figure 2: General correlation trends of the convenience word count and the corresponding review score for a particular listing. Listings with a review score of 0 do not have any reviews yet.

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

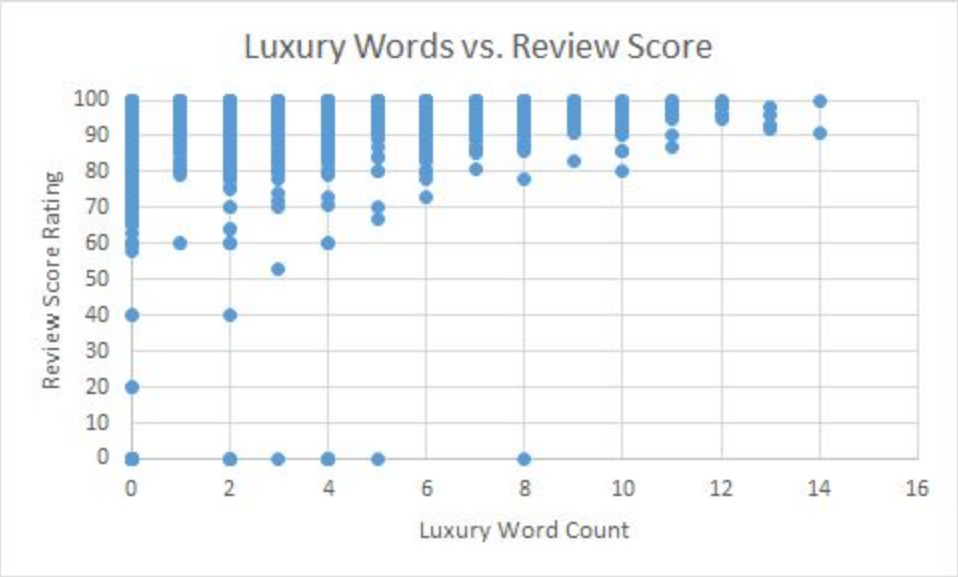

Figure 3: General correlation trends of the luxury word count and the corresponding review score for a particular listing. Listings with a review score of 0 do not have any reviews yet.

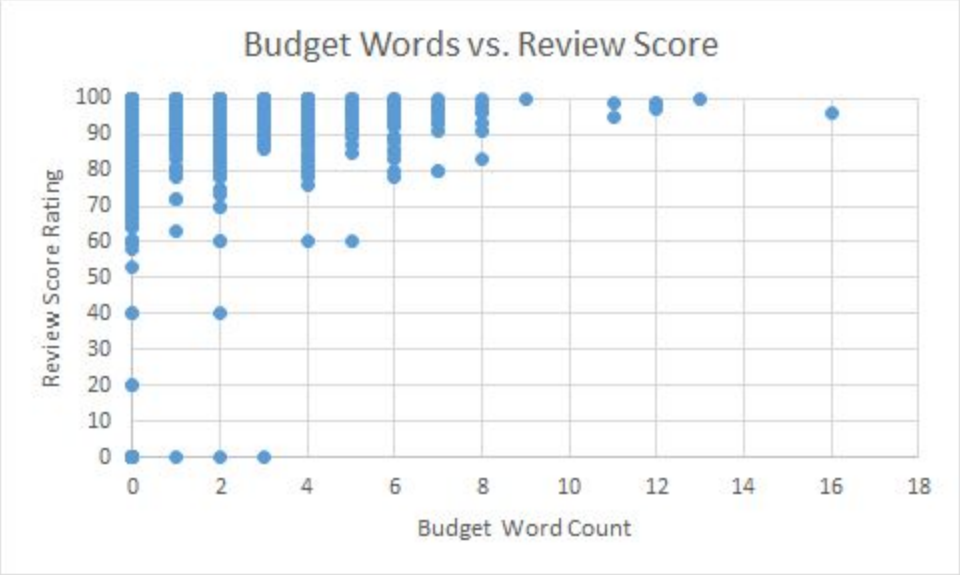

Figure 4: General correlation trends of the budget word count and the corresponding review score for a particular listing. Listings with a review score of 0 do not have any reviews yet.

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

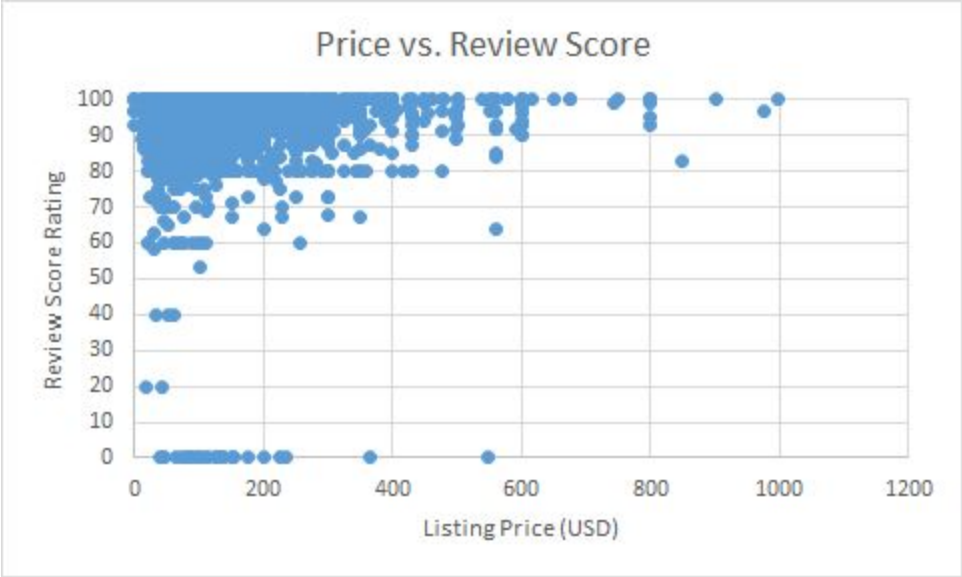

Figure 5: Correlation data of the listing price vs. the review score of a given listing. A review score rating of 0 indicates that the listing does not have any reviews yet.

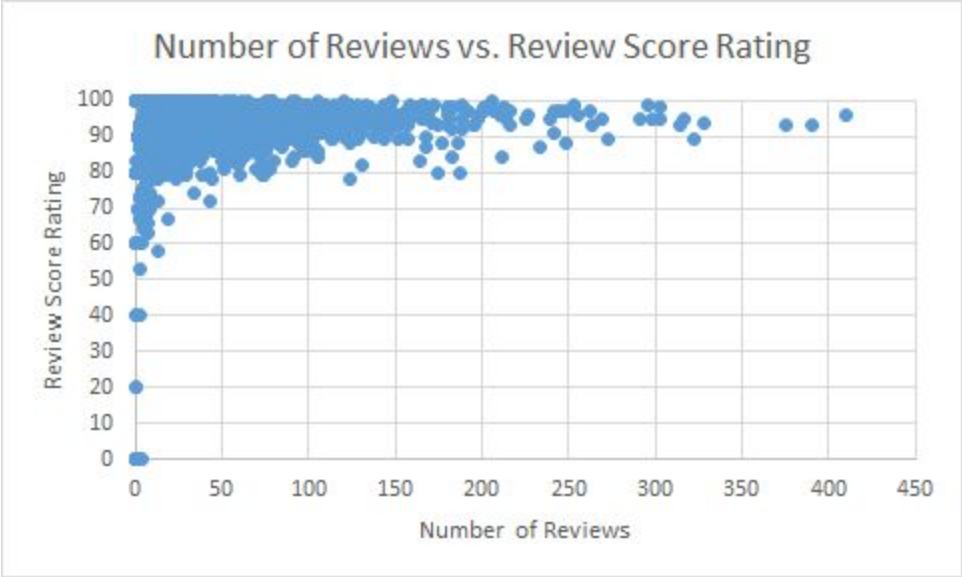

Figure 6: Correlation between number of reviews for a particular listing and the review score of that listing.

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

Figure 7: Correlation between the luxury word count and the price of the listing

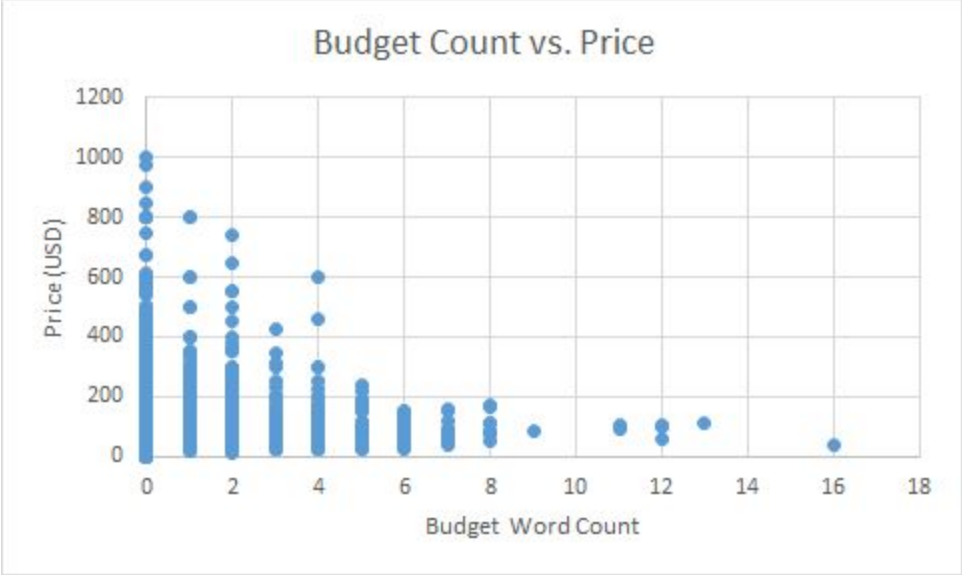

Figure 8: Correlation between budget word count and listing price

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

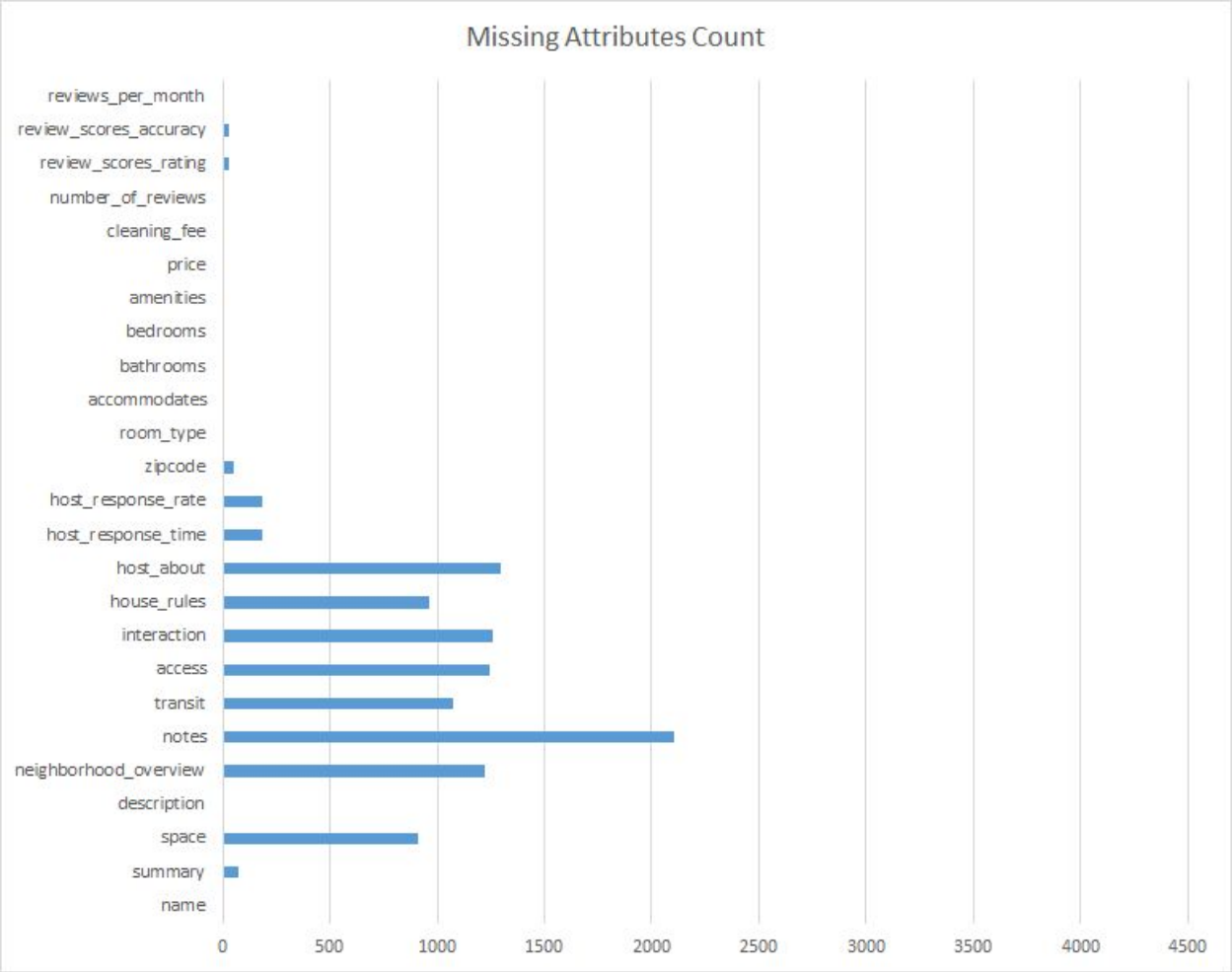

Figure 9: An overview of some of the features and their missing attribute counts. Note: this chart does not represent all of the features used, and any feature that was not represented on the chart had a missing attribute count of 0.

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

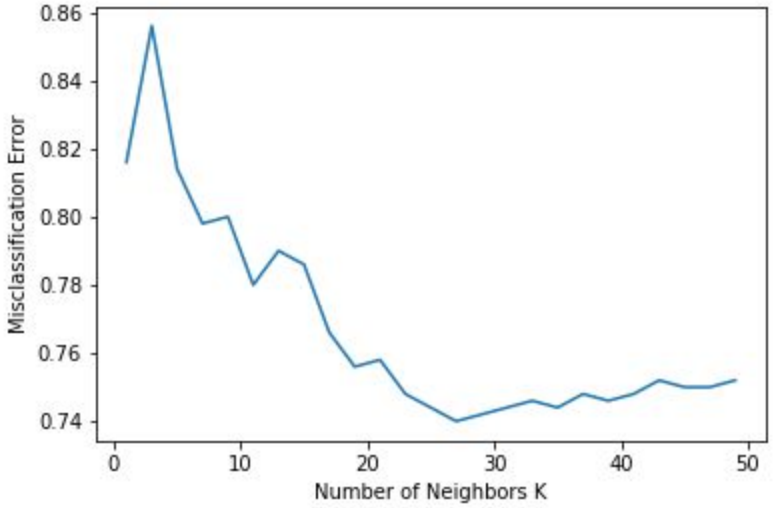

Figure 10: Miscalculation error by number of neighbors K

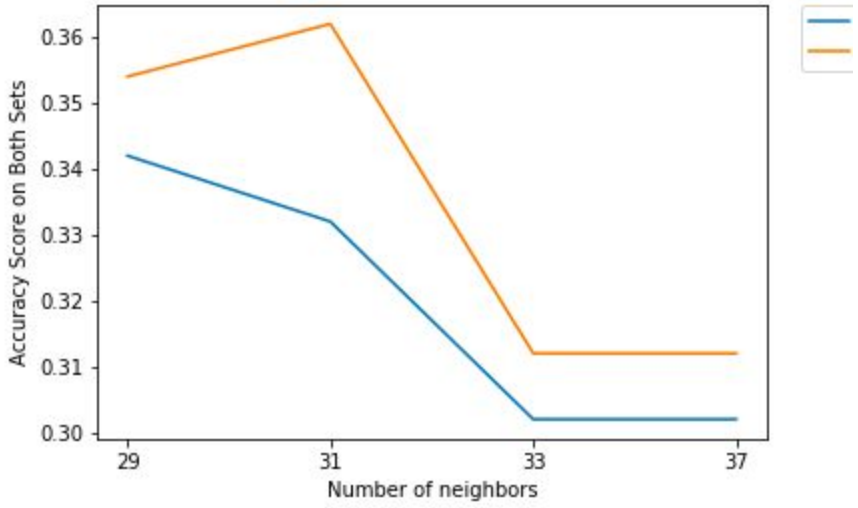

Figure 11: Validation (blue) set accuracy and test (red) set accuracy

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

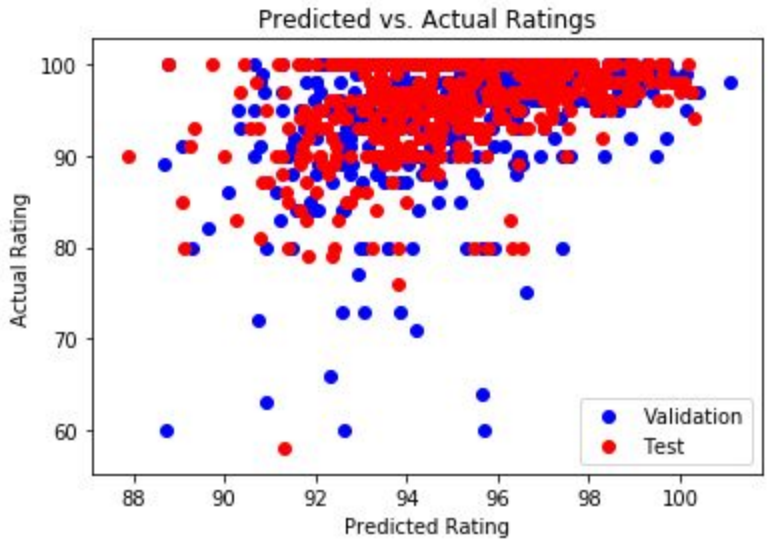

Figure 12: Predicted vs. Actual Ratings from LR without feature selection

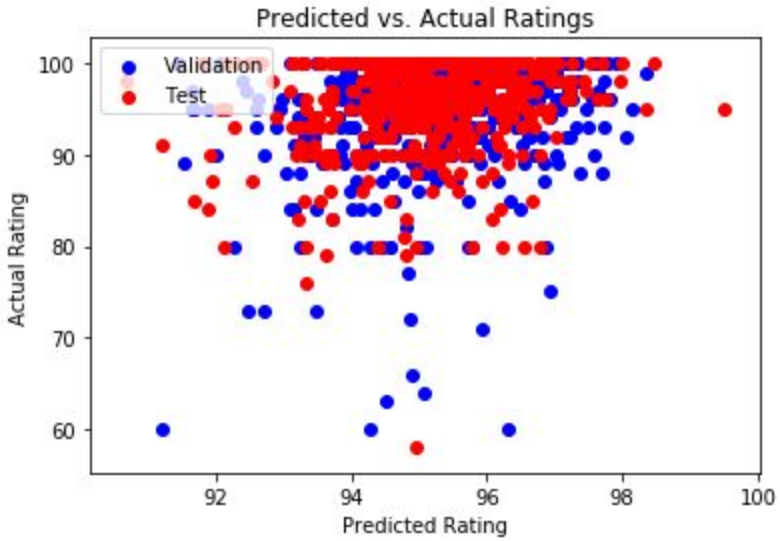

Figure 13: Predicted vs. Actual Ratings from LR with feature selection

Judy Tsai - paoyitsai2019@u.northwestern.edu
Julia Davis - juliadavis2018@u.northwestern.edu
Emma Crowley - emmacrowley2019@u.northwestern.edu
Northwestern EECS 349, Spring 2018

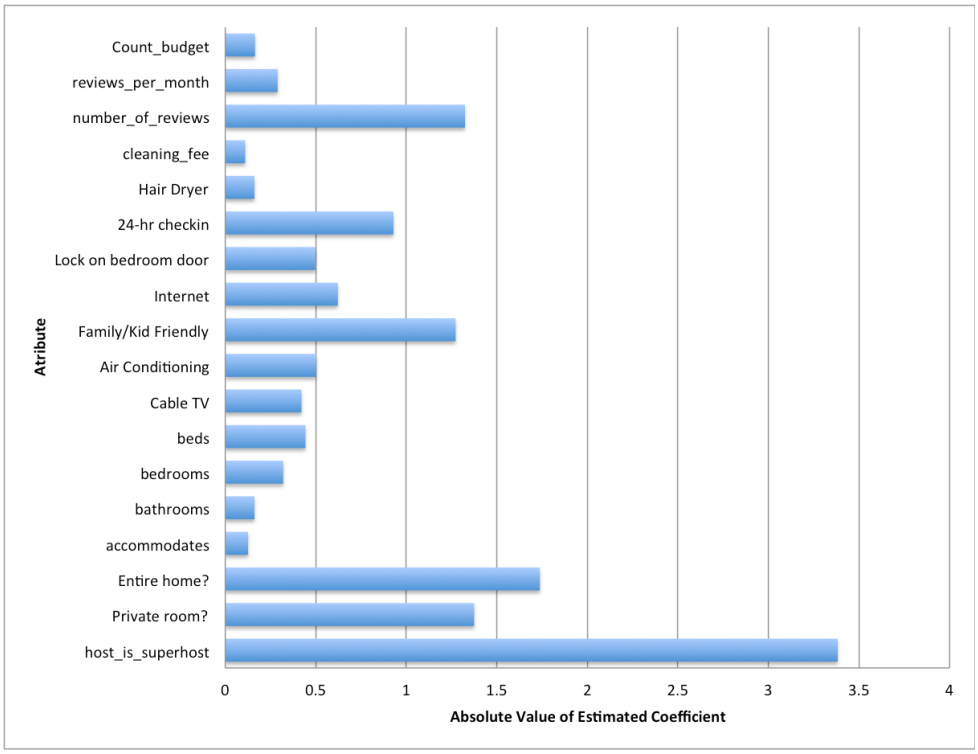

Figure 14: Absolute value of coefficients from LR with feature selection

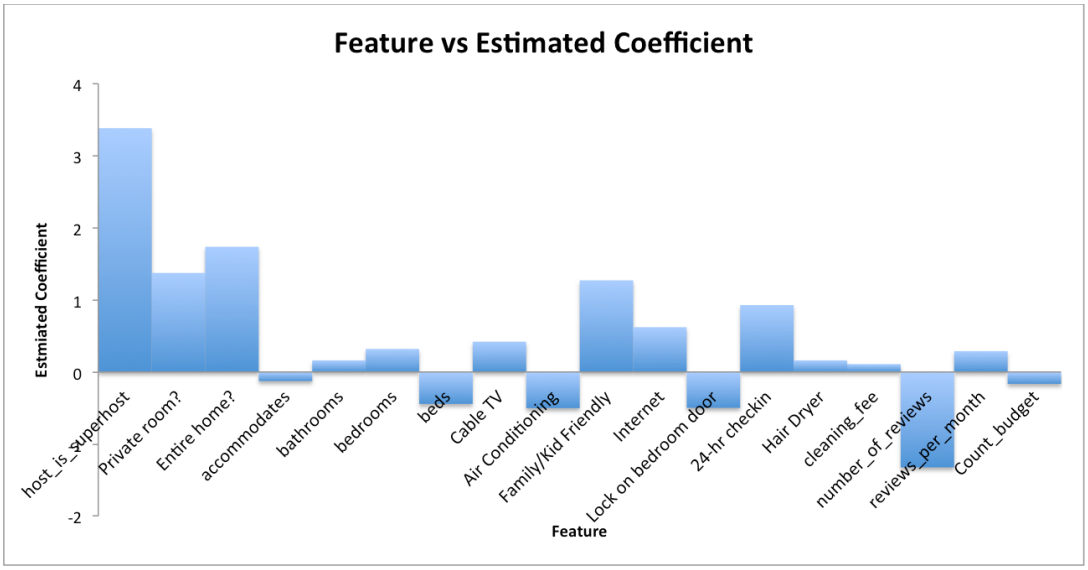

Figure 15: Estimated coefficients from LR with feature selection