PREDICTIVE MODELING ON AIRBNB LISTING PRICES

Josh Keating	Ethan Katnic	Christian Hahn	Robin Yang
University of Washington	University of Washington	University of Washington	University of Washington
jfk8@uw.edu	ethank5@uw.edu	hahnc3@uw.edu	ryang97@uw.edu

ABSTRACT

We empirically investigate host listings on Airbnb, a popular online hospitality service marketplace, to build predictive models on listing price based on key factors and variables. The data we used included ~4000 host listings in the Seattle area with 92 features (such as number of bedrooms) and ~85,000 reviews of those listings. In addition to this data, we ran a text sentiment analysis on each listing description, scoring positive or negative sentiment from 1 – 100. Unfortunately, we found this score insignificant towards predicting pricing relationships. Using the other features, we built several predictive models using different machine learning methods such as k-nearest neighbors, random forests, and neural networks. Our discoveries found that a neural network model had the best performance in predicting price with only an error margin of \$32 - \$35 dollars. The results serve as a step toward understanding the factors influencing the pricing trends of the relatively new online hospitality service market.

Author Keywords
Online Hospitality Service, Airbnb; Predictive Modeling; Neural Network; Feature Selection

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION
The online sharing economy has been growing extremely rapidly over the past couple of years. Airbnb is only one of many fee-based, peer-to-peer sharing platforms; although it is also one of the most popular and most successful. If you are unfamiliar with its service, Airbnb operates (both online or through a mobile application) as the middleman in booking homestays and lodgings. It allows homeowners to list their open rooms or houses to host short or long term stays and connects them with guests, taking a small commission in the form of service fees from each booking.

Founded in 2008, Airbnb now has over 5 million unique listings in over 81,000 cities and 191 countries [1]. Due to such recent and rapid growth of these markets, it is understandable that interest and curiosity levels are high. Peer-to-peer sharing doesn't operate on the same techniques as traditional businesses, and we are looking to identify what puts listings ahead in such environments.

While different from traditional business, online share services still have one important aspect in common: competition. With so many different listings worldwide, hosts are often competing against one another to get guests to book their rooms in order to make their money. Obviously, listing at higher prices can make hosts more money; but what allows a host to list at a higher price? This leads to the question of what features of the listing have the biggest impact on price, and can price be predicted based upon these features.

What features of Airbnb listings have the biggest impact on price, and can price be accurately predicted based upon these features?

The question above is the main focal point of the research we have done. In this modern age of the internet, data is both accessible and abundant. As mentioned earlier, Airbnb currently has over 5 million unique listings, with all relevant information easily obtained through web scraping techniques (gathering and recording data off the listing website URL). With this data we have the resources to attempt to answer our question. When making a host listing on Airbnb, a variety of information is needed. This includes standard quantitative information such as the number of bedrooms, bathrooms, etc. as well as data that includes more personalization such as host information and written descriptions of the property. Everything that the guests can see when making their decision on which listing to book is available in our dataset.

The research we are performing makes new contributions to the community through an enhanced understanding of factors leading to higher profits in online marketplaces and providing a model to predict where prices lie.

RELATED WORK
Before we began our research, we studied several works related to our question. As previously mentioned, the online share market has garnered massive attention during recent times. With Airbnb positioned as a powerhouse in this service, there has been much relevant research performed that we draw from.

The first paper we examined by the University of Paderborn examined the price adjustment for Airbnb hosts once they achieve a visible star rating using a dataset of listings from New York City. For those unaware, Airbnb keeps listing reviews hidden until hosts achieve at least 3 ratings. This is done in order to avoid having one bad review contaminate all future bookings. The researchers in this case wanted to examine how achieving a visible star rating would influence the prices hosts listed in a before and after investigation. The motive here is that in an online share, the product is not the only thing of value. There is the argument that trust is in fact sometimes even more important than the item in the listing itself. Services and products on the internet can be unreliable, and the aversion of risk is an important variable to guests when making decisions; particularly when deciding a place to sleep. The researchers' findings indicated that the upper quartile priced listings had a significant increase in price (+ 2.69 ∈) once visible ratings were achieved, while the lower 75% of values only saw a slight increase. In fact, the main driver of price for the top 25% was rating. The results indicated to us that the review scores can be an important, valid indicator of price for our model [2].

A second paper we examined was conducted by the Harvard Business School on listings in New York City as well. In this case, the topic was digital discrimination. The researchers questioned whether trust varied in online marketplaces based upon personal appearance factors such as profile picture, age, gender, race, etc. Based off host indicators such as rental prices and listing information, the researchers concluded that there was a significant prevalence of racial discrimination when renters had to choose between a white or black host. Non-black hosts charge approximately 12% more than black hosts for equivalent rentals. While this research was only performed on data from New York City, the researchers believe that the adversity of the dataset means their findings can be reflective across the entire Airbnb platform. In terms of answering our own question, we found this paper relevant through our own decision to add sentiment scores to our analysis. With variables such as rating and number of bedrooms as obvious indicators, we wanted to include the different methods of personalizing listings to our predictions in order to make a more well-rounded model [3]. These relevant works validated our assumption that there are indeed variables that have a large effect on listing prices. We also discovered that these variables are not limited to the basic properties of the host listing. With all the information made available, guests have much more to analyze while they judge listings.

DATA

The dataset we used is the Seattle Airbnb Open Dataset from Kaggle.com [4]. The set includes three separate files: calendar.csv, listings.csv, and reviews.csv. The information included is 3818 different Seattle based listings with 1.39 million calendar dates of their bookings from January 2016 through January 2017 and 84.8 thousand reviews. The focus of this research was on the 92 different features included in the listings. Notable features include the number of bedrooms, number of bathrooms, the neighborhood, and the room type.

Important to note is the presence of Airbnb reviews and sub-reviews within our feature set. Each listing includes guest-submitted scores of specific aspects of the Airbnb such as cleanliness, host communication, and location. Along with this is an overall rating for the Airbnb, which is not a direct summation or derivation of the sub-scores, but a separate score entirely. This distinction will be explored further in the review prediction section of our results.

In addition to these 92 features, we added one of our own. Using Microsoft Azure Text Sentiment Analysis API, we were able to run a sentiment analysis on the description of all the listings. After sending the description to the API, it returns to us a value from 1 – 100 with 100 being extremely positive and 1 being extremely negative. The intended use of this additional feature was to use it as a predictive feature in price. We assumed that more positive descriptions meant more colorful and descriptive language, leading to nicer and therefore more expensive listings.

We began our analysis with some data preparation, filling in NA review values with the review averages and converting the data types of percentages to nicer formats such as integers. We also decided to only include the frequently used property types. We removed property types like boats and cabins that only appeared rarely in the data. After this initial step, our data was cleaned and prepared for further analysis.

METHODS

All data analysis was performed using a Jupyter Notebook written in the Python coding language. In order to achieve a base-line estimate of our price predictor, we ran a basic ordinary least squares regression to look at which features correlate with the listing price. In this regression, we included all variables aside from those that were almost entirely null (empty values), or only contained a single value for greater than 90% of listings.

Dep. Variable:	price	**R-squared:**	0.581
Model:	OLS	**Adj. R-squared:**	0.577
Method:	Least Squares	**F-statistic:**	130.4
Date:	Tue, 04 Dec 2018	**Prob (F-statistic):**	0.00
Time:	20:47:45	**Log-Likelihood:**	-20310.
No. Observations:	3705	**AIC:**	4.070e+04
Df Residuals:	3665	**BIC:**	4.095e+04
Df Model:	39		
Covariance Type:	nonrobust		

Figure 1. Table of Initial OLS results (all variables)

The results of this test gave us an R-squared value of .581, which gives an indication that we can at least explain the majority variation of price using the selected variables. Unsurprisingly, we found features such as bathrooms, bedrooms, and number of guests had the highest correlation with price. Unfortunately, we also discovered that our sentiment score did not have a high significance with a p-value of 0.537. We had originally anticipated a significant correlation here as we assumed more positive sentiment in the description meant more descriptive and colorful language, leading to generally nicer listings. Due to the low scoring, we ultimately cut the score from our final selection of variables.

Our next step included some introductory data visualizations. The first we performed was a heatmap of several of our top factors.

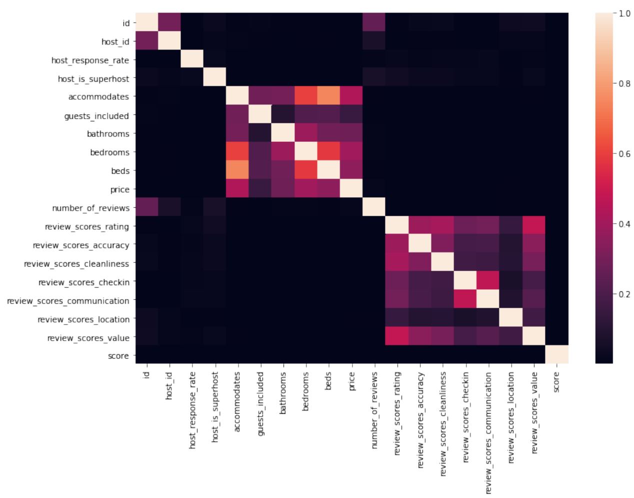

Figure 2. Heatmap of Correlations Between Several Important Features

We are most interested with the features correlated with price, which most notably includes the number of people the listing accommodates and the number of bedrooms. Another interesting observation was to see how each of the review sub-categories correlated with each other. It seems that review_scores_cleanliness and review_scores_value is the most correlated of the different sub-ratings.

Our next visualization was a comparison between the sentiment score and price in order to better understand the lack of significance we observed.

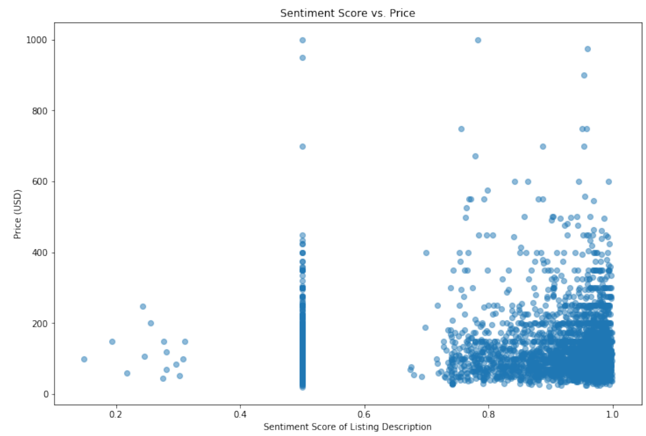

Figure 3. Sentiment Score vs. Price

While no clear correlation is apparent, some notable observations can still be made. The few listings that included the lowest sentiment all had lower prices, with none having above a $300 listing price. Those that had a neutral sentiment score of 0.5 displayed similar attributes to the high sentiment scores but included much less at the $500+ range. We assume that the results of the sentiment score are attributed to the fact that the online marketplace is intended to sell the product, and most hosts make positive descriptions in order to do so regardless of how nice the listings is.

Another important factor in listing price is the neighborhood in which the listing are located. Due to houses generally having a higher price in nicer neighborhoods, we assume that neighborhood is a likely predictor for listing price. In order to understand this better about our dataset within the Seattle area, we graph the average listing price per neighborhood.

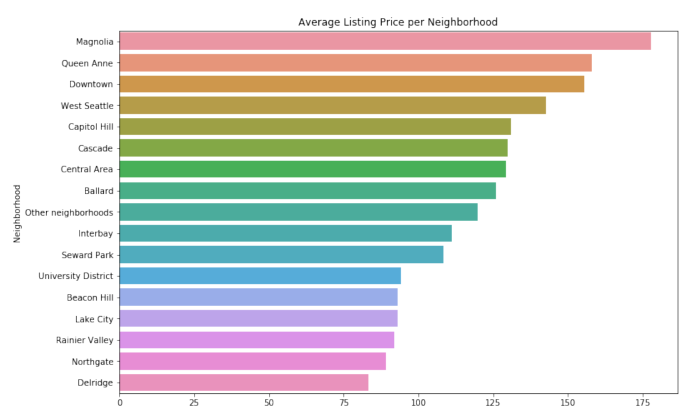

Figure 4. Average Listing Price Per Neighborhood

We observed that Magnolia stands out from the pack with an average price of nearly $175, nearly $100 more than the lowest average of Delridge at around $75 per listing. This strengthened our assumption that neighborhood has a large impact on listing price.

Our fourth and final visualization was the comparison between listing price and the number of reviews.

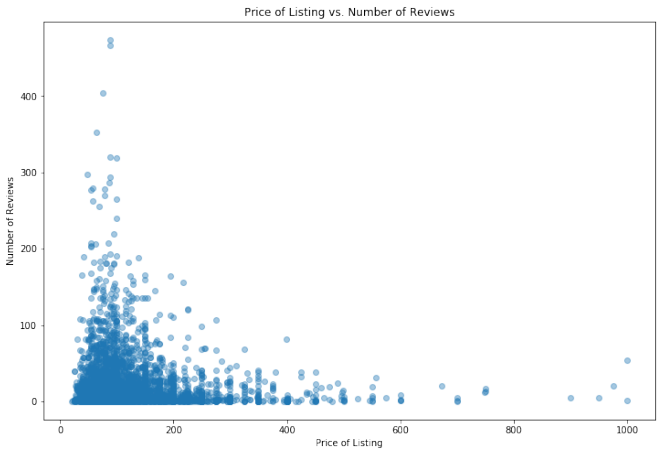

Figure 5. Price of Listing vs. Number of Reviews

The graph cascades down nicely and shows that higher priced listings have significantly lower number of reviews. This is most likely due to them being out of affordable range for most Airbnb users. However, with so many lower cost listings having a low number of reviews as well, it does not seem a good indicator towards predicting the overall listing price.

Having performed an exploratory regression and visualization analysis, we begin our process of building an accurate predictive model by splitting our data into training and testing sets. Our model was trained on 80% of the data available to us, and the remaining 20% was kept on hand to test the accuracy of our result. In this paper, we will first outline all the models we created before discussing their individual performances in the results section.

The first model we created was a k-nearest neighbors (KNN) model in which we varied the polynomial degree of regression, the number of neighbors, and the weight function. We ran our parameter grid search and performed a best parameters function to find observe that our best parameters included using the 11 closest neighbors, using the distance weight function when evaluating neighbor influence (defaulted to Euclidean), and a polynomial degree of 1.

Our second model was done using a random forests Regressor. In this model, we varied the polynomial degrees of regression, the number of trees in the forest, as well as the maximum depth of each tree. From the parameter grid search and best parameters function, we found the best combination of parameters included using 66 trees, a maximum depth of 4, and a polynomial degree of 1.

Our third and final model was also the highest performing. Using a neural network, we followed the same process with our previous two models to find the best parameters. Relu was the best activation function for the hidden layer, used a constant learning rate (over invscaling), using a stochastic gradient descent solver for weight optimization, and a polynomial degree of 1.

RESULTS (PRICE)

Having run and developed each of our three models, we concluded that the multi-layer perception neural net was the highest performing of them all. KNN reported mean errors in the range of 38 – 35, random forest with a 37 – 34 range, and the neural network scored in the 35–32 range. Comparing the spread of each model's errors against each the test data confirms the neural network as the most accurate of the three. We completed a number of visualizations to complement our results. First, we plotted the predicted vs. actual prices across each of the models to get a clearer view of performance. A somewhat accurate best fit line signified accuracy while listings at very high price points were indeed quite off.

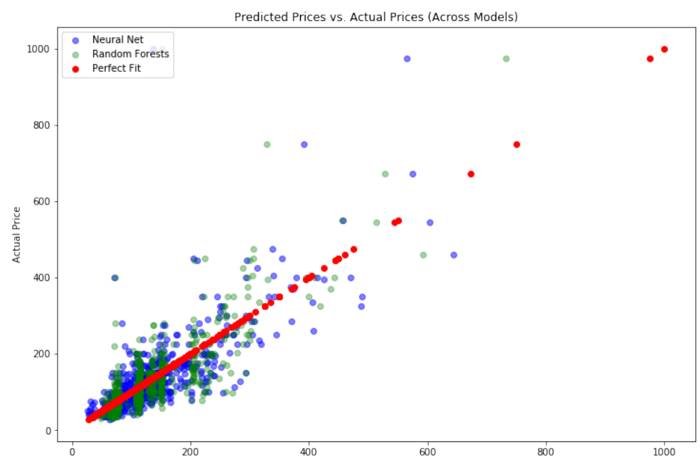

Figure 6. Predicted Prices vs. Actual Prices (Across Models)

In Figure 5, we visualized the predicted prices of our best two models, the random forests and the neural net methods. The neural net results can be distinguished in blue while the random forests model is in green. We also placed a red line to visualize what a perfect fit would have looked like. The general result follows a pattern similar to a slope of 1, indicating a somewhat accurate model was created. A noticeable difference between the two models is the highly concentrated clumping from the random forest at different intervals, likely due to the model predicting prices at common points such as $85, $100, and $150.

To get a better look at the errors across different models, we visualized the predictive error across each of the models.

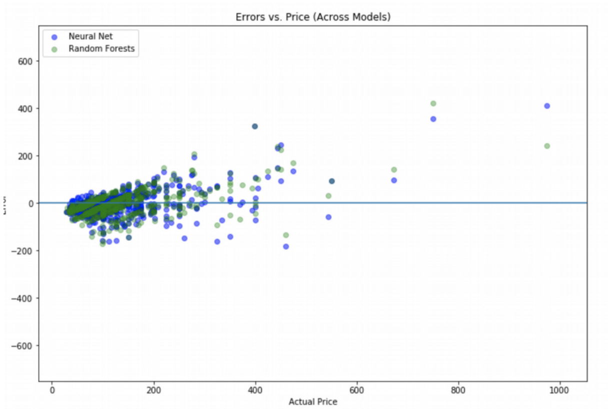

Figure 7. Errors vs. Price (Across All Models)

The clumping around a predictive error at 0 was a positive sign but understandably, there is less accuracy as the price goes up. This is likely do to high-end features of higher priced listings not being included in our study. While the performance of each of the models is comparable and the differences are slight, the neural net does seem to have greater consistency around the zero line.

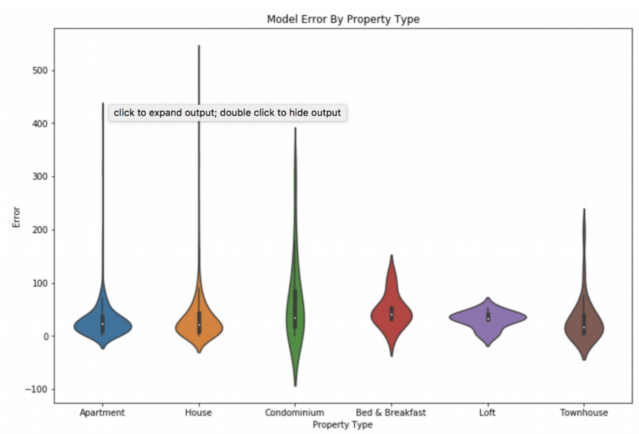

Figure 8. Model Error by Property Type

The violin plot above compares the errors across each of the property types for the Neural Net, our best performing model. Apartments, houses, and townhouses seem to have the lowest errors, but there are huge spikes due outliers on the extremely expensive end.

RESULTS (REVIEWS)

After performing the exploration into Airbnb pricing, we chose to briefly investigate Airbnb reviews in a similar manner. While reviews were not the main focus of our research question, we felt that we could perhaps gain insights into what makes an enjoyable Airbnb experience using similar factors as those included in our price analysis.

Our methods in this analysis were the same, first performing a regression, before building a predictive model and measuring accuracy in predicting the overall review score of an Airbnb.

OLS Regression Results

Dep. Variable:	review_scores_rating	**R-squared:**	0.719
Model:	OLS	**Adj. R-squared:**	0.715
Method:	Least Squares	**F-statistic:**	156.6
Date:	Tue, 04 Dec 2018	**Prob (F-statistic):**	0.00
Time:	20:48:03	**Log-Likelihood:**	-5878.0
No. Observations:	2424	**AIC:**	1.184e+04
Df Residuals:	2384	**BIC:**	1.207e+04
Df Model:	39		
Covariance Type:	nonrobust		

Figure 9. OLS Regression Results for ratings (with sub-reviews)

Although our least squared regression for ratings returned a high r-squared value of **0.719**, this was primarily because this equation also included the sub-review scores. Unsurprisingly, including these features lead to a much stronger ability to predict variance in the data. Intuitively, a feature such as review_scores_cleanliness can be a strong predictor of the overall review score. Additionally, other important features like price, number of reviews, and whether the host was a "superhost" did have a significant relationship to the overall rating as well.

We felt that including sub-reviews as features in our review score models would an invalid representation of what our model is attempting to predict. This is because the sub-reviews would not be accessible until the overall score has been submitted. For this reason, we chose to train separate models with and without the sub-reviews and compare.

To view the accuracy of the model without the sub-reviews, we conducted another multivariate linear regression, this time without the sub-reviews.

OLS Regression Results

Dep. Variable:	review_scores_rating	**R-squared:**	0.190
Model:	OLS	**Adj. R-squared:**	0.178
Method:	Least Squares	**F-statistic:**	16.94
Date:	Tue, 04 Dec 2018	**Prob (F-statistic):**	5.77e-86
Time:	20:48:08	**Log-Likelihood:**	-7162.7
No. Observations:	2424	**AIC:**	1.439e+04
Df Residuals:	2390	**BIC:**	1.459e+04
Df Model:	33		
Covariance Type:	nonrobust		

Figure 10. OLS Regression Results for ratings (without sub-reviews)

Unsurprisingly, this significantly dropped our R-squared value, to **0.190**, indicating that our remaining variables only account for **~20%** of the variation in Airbnb rating. We hypothesized that this is because the variables we are left with only represent a small fraction of what goes into a positive or negative Airbnb experience. Factors such as property type or price simply explain the listing and do not give much insight into a negative or positive experience. However, the regression did show significant factors. Higher price generally meant better reviews and the host being a "superhost" also aligned with higher ratings.

After this initial study, we followed the same steps to predict price, constructing a Neural Net model to predict rating. We get a negative mean absolute error of about -3.12 without the sub-reviews and -1.98 for the model with sub-reviews. This means that on average, our predictions are approximately three points (out of 100) away from true value. As expected, the model with sub-reviews performs with greater accuracy. It is important to note that the variance in overall review score was 26.64 meaning that the bulk of the review scores were in the mid-nineties. Given this low variance and the skew in the data, achieving this level of error is not as meaningful as it might appear on the surface. To help visualize this, we made a scatter plot to view if there was any correlation between the actual rating and predicted rating excluding the sub-reviews.

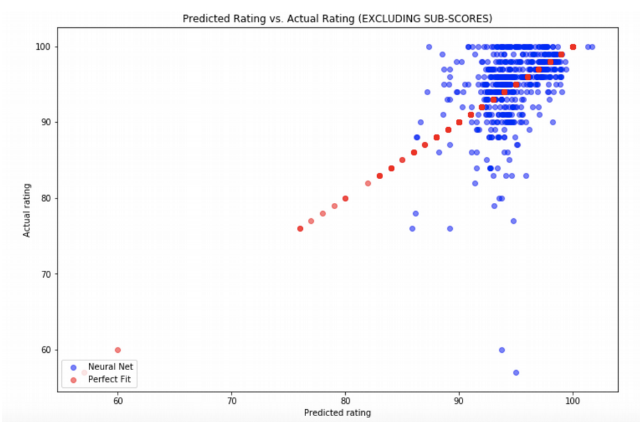

Figure 11. Predicted vs. Actual Rating (excluding sub-reviews)

This visualization shows a lack of correlation between the two as the Neural Net model did not perform as well without the sub-reviews. We can contrast this with following graph (Fig 12.) where review sub-scores were included.

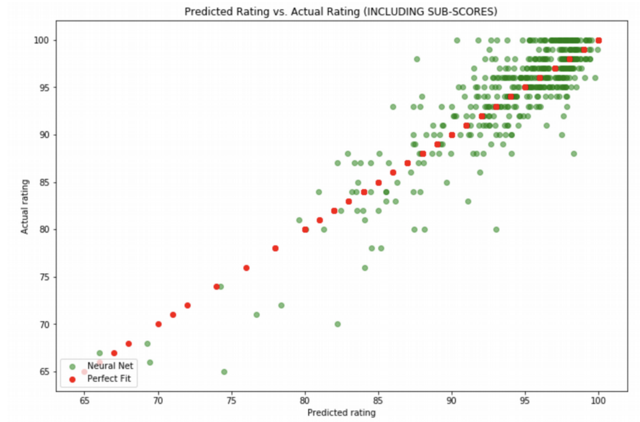

Figure 12. Predicted vs. Actual Rating (with sub-reviews)

These differences demonstrate that while we can reasonably predict Airbnb reviews if we have access to those sub-scores, without them, our predictions plummet in accuracy.

DISCUSSION

The results give us several insights into Airbnb reviews and listing prices. While our exploration into sentiment analysis of listing description did not yield a significant relationship with price, features of Airbnbs such as the super-host status of the property owner, neighborhood of the listing, and property-type of the home allowed us to predict listing prices with reasonable accuracy.

These insights can be used by Airbnb hosts to properly value their properties. Plugging in their own housing details into our pricing model could suggest a reasonable price given Seattle Airbnb trends.

Listers can also identify which features make for an expensive Airbnb and can invest in acquiring or developing those features to increase the value of their property. For example, with the knowledge that the “superhost” status correlates with both better reviews and higher listing prices, hosts can work with Airbnb to acquire this status. While we cannot say definitively whether the “superhost” status causes a higher price and higher rating or if it’s simply a correlative relationship, achieving “superhost” status could be worth exploring as a host looking to improve the general value of their listing.

Our brief look into review trends could provide similar benefits to listers who are looking to improve the overall rating of their listing. Observing that “cleanliness” has a strong positive relationship to the overall score, hosts can focus on improving cleanliness over other factors like the check-in process and communication quality, which do not share as strong of a positive trend.

From the guests’ perspective, relationships between features and price can be used to identify good deals that are priced reasonably to the Airbnb market and avoid unreasonably expensive listings. Using our exploration into neighborhoods and their pricing trends, a user unfamiliar to the Seattle area could recognize that an apartment in Delridge is not worth a $100 price tag, and the listing of the same type in Magnolia is a superior value.

The accuracy of our pricing model proves that we can predict listing price with reasonable reliability like our research question proposed, but the same cannot be said for our review-predicting model, which failed to provide significant predictive ability when sub-reviews were excluded.

FUTURE WORK

This study serves as an introductory analysis into Airbnb listings for the Seattle area. While the relationships between listing features, price, and review scores that we explored in this research were successful at some level for prediction, there are certainly a number of areas to be improved upon in future research.

Unfortunately, due to resource limitations we were not able to explore select features that our original literature review suggested would be valuable. In particular, as highlighted by the Harvard Business Review paper on host race, there may be latent variables related to a host’s bio that may influence the pricing and review scores of a listing. Going forward there is definite opportunity for further exploration of host features like gender and race.

Our research is also limited to listings in the Seattle metropolitan area and is not representative of Airbnb listings at the national or global level. As such, further work could incorporate data from additional cities and countries in an effort to decrease our sample bias. Expanding the scope of this research would allow for more complex comparisons between regions in categories such as income levels or socioeconomic status.

Additionally, Airbnb as a service shares many qualities with other online marketplaces. Competing hospitality marketplaces such as VRBO are likely to exhibit similar behavior. This intersection presents an opportunity for comparison studies to be conducted across similar hospitality marketplaces to show which shared factors are important in predicting pricing or reviews.

Our research marks a step toward a better understanding of predictive factors in the online marketplace and hospitality services. Our study serves as a brief introduction to a topic where additional work will provide more comprehensive insights into an emergent industry.

REFERENCES

1. AirBnb. 2018. Fast Facts. Retrieved from https://press.airbnb.com/fast-facts/.
2. Gutt, Dominik and Herrmann, Philipp, "Sharing Means Caring? Hosts' Price Reaction to Rating Visibility" (2015). *ECIS 2015 Research-in-Progress Papers.* Paper 54. ISBN 978-3-00-050284-2 https://aisel.aisnet.org/ecis2015_rip/54
3. Edelman, Benjamin G. and Luca, Michael, Digital Discrimination: The Case of Airbnb.com (January 10, 2014). Harvard Business School NOM Unit Working Paper No. 14-054. Available at SSRN: https://ssrn.com/abstract=2377353 or http://dx.doi.org/10.2139/ssrn.2377353
4. Kaggle. 2018. Seattle Airbnb Open Data Retrieved from https://www.kaggle.com/airbnb/seattle#listings.csv