Title
Airbnb Price Prediction in the Age of Social Distancing

Permalink
https://escholarship.org/uc/item/4x28127g

Author
Tran, Richard

Publication Date
2021

Peer reviewed|Thesis/dissertation
Airbnb Price Prediction
in the Age of Social Distancing

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Applied Statistics

by

Richard Tran

2021
ABSTRACT OF THE THESIS

Airbnb Price Prediction
in the Age of Social Distancing

by

Richard Tran
Master of Science in Applied Statistics
University of California, Los Angeles, 2021
Professor Yingnian Wu, Chair

The Coronavirus (COVID-19) pandemic has created a crisis in the tourism and hospitality industry. As the people began following social distance protocols and stay-at-home orders, how would this affect consumer behavior in the short-term rental market? Airbnb data from New York City was used to determine factors that influence Airbnb prices after the onset of COVID-19. Exploratory data analysis was used to select useful features to model the rental prices. Random Forest, Linear Regression, and XGBoost models would be evaluated on their overall performance using RMSE and R2 as metrics. Topic modeling, perceived cleanliness, and polarity analysis were used to determine how reviewers could influence the listings. The number of listings and the prices significantly dropped during the onset of COVID-19. Hosts were more interested in long-term guests in their stays. XGBoost, the highest performing model, weighted renting the entire home/apartment and number of accommodations as the most important factors in Airbnb price prediction.
The thesis of Richard Tran is approved.

Vivian Lew

Frederic Paik Schoenberg

Yingnian Wu, Committee Chair

University of California, Los Angeles

2021
To my sister . . .
who has always supported me
in everything that I do
# TABLE OF CONTENTS

1 Introduction ........................................................... 1

2 Exploratory Data Analysis .............................................. 3
   2.1 Obtaining the Data .............................................. 3
   2.2 Listings ......................................................... 4
   2.3 Reviews ......................................................... 8

3 Feature Engineering ...................................................... 11
   3.1 Variable Selection .............................................. 15

4 Methodology and Modeling ............................................. 18
   4.1 Random Forest ................................................. 17
   4.2 Linear Regression ............................................. 18
   4.3 XGBoost ........................................................ 19
   4.4 Latent Dirichlet Allocation ................................... 19

5 Results ................................................................. 21
   5.1 Topic Modeling ............................................... 23

6 Conclusion ............................................................... 27

7 Limitations and Future Direction .................................... 28

8 APPENDIX ............................................................... 29
LIST OF FIGURES

2.1 Map of Airbnbs in New York City ........................................... 3
2.2 Airbnb Prices of Entire home/apartments and private rooms by borough. .... 3
2.3 Number of reviews each year for Airbnbs in New York City. ................... 3
2.4 NRC Emotion lexicon for reviews. ............................................. 3

3.1 Histogram of log-transformed price ............................................ 15
3.2 Histogram of Average Sentiment for each review using Sentimentr ............ 15
3.3 Property vs Room type of NYC ............................................... 15
3.4 Variable Importance from Random Forest ................................... 17

4.1 Graphical Model Representation of LDA ..................................... 19

5.1 Importance of variables in XGBoost model Pre-Covid .......................... 26
5.2 Importance of variables in XGBoost model during Covid ........................ 26
5.3 Topic Modeling with LDA ...................................................... 26
## LIST OF TABLES

2.1 Summary statistics for numerical variables. ........................................ 4

5.1 Summary of results from modeling .................................................. 23

5.2 Linear Regression exponentiated coefficients ..................................... 23

5.3 Matrix of mean prices for average sentiment ..................................... 23

8.1 Linear Regression Output ................................................................. 29
CHAPTER 1

Introduction

In December of 2019, a new unknown virus was found to now be recognized as the novel coronavirus (COVID-19). The first case was confirmed in New York State on March 1, 2020. Shortly after, the World Health Organization declared the novel coronavirus a pandemic on March 11, 2020.[2] Since then, New York City was named an epicenter for COVID-19 during the spring of 2020. With the United States government enacting new lockdown measures in order to stop the spread of the virus, the world was brought to a standstill and devastated the economy. The travel and hospitality industry was severely hurt due to airlines cutting flights and consumers canceling travel arrangements. As the people began following social distance protocols, it was only a matter of time before they became restless. With the ability to work from anywhere because of stay-at-home orders, some people began to take full advantage of this once-in-a-lifetime opportunity.

When the pandemic began, people just stopped traveling; however, things started to change in the summer of 2020. According to Airbnb co-founder & CEO Brian Chesky, people started getting restless. “People want to travel, they just don’t want to get on airplanes,” Mr. Chesky said. “They don’t want to go for business. They don’t want to stay in the really big cities as prevalently as they used to. They don’t want to be in crowded hotel districts.” But, he said, “they do want to get out of the house. And so we think demand is going to be strong in the
future. I’m very optimistic, actually, about the industry.” These short-term rental stays were being rented out to those who wanted to work from home in another home. There was a yearning for travel so domestic vacations increased and so did the demand for Airbnbs.[2]

Airbnb is an online housing platform that provides users to list, discover, and book accommodations for their users all across the world. Hosts can offer their property spaces to guests for short or long periods of time. These accommodations differ from the typical hotel room because of the ability to live and witness the culture of living in the area of choosing. Airbnb’s listings offer a huge range of options ranging from community bedrooms to luxury housing, all in one platform. There is a peer-review system where guests can leave reviews after reservations are made. These reviews can often dictate how much owners can book/earn.

The objective of this research paper was to determine what factors affect Airbnb prices after the onset of COVID-19. In order to refine this study, the biggest tourist city in the United States and the epicenter of the coronavirus virus, New York City, was used as an example to see how consumer behavior has changed towards short-term rentals during the pandemic. With this increased understanding of the factors affecting the attitudes and pricing towards Airbnbs, consumers will be able to make informed decisions on their stays. The Airbnb hosts could develop more advanced pricing methods to increase profitability on their rentals.
CHAPTER 2

Exploratory Data Analysis

2.1 Obtaining the Data

InsideAirbnb is an independent website that allows users to explore how Airbnb is really being used in cities all around the world.[7] The data is sourced directly from the Airbnb site after being cleansed and aggregated. Listings can be deleted on the Airbnb platform since the data presented is just a snapshot at the particular time. The datasets include detailed information on the listings, reviews, calendar availability, and geo-filters for applicable airbnbs in each major city. For the purpose of this research, snapshots of the listings and reviews from January 2020 and January 2021 were used. The January 2020 data set had a total of 51361 listings and 1285935 reviews available. The January 2021 snapshot dropped off with a total of 37143 listings and 851941 reviews. The listings and reviews were filtered down by the following:

1. Only listings and reviews from after 3/1/2019 and 3/1/2020 until the following January snapshot were included

2. Listings must have an active review during this time period and be a part of the “reviews” dataset.

The outbreak of COVID occurred in the United States roughly around March. It was during this time that stay-at-home orders were enforced. The nine-month time period was there to keep the comparison between the two years consistent to avoid any possible seasonal changes. Some
hosts may not have their calendar updated so only recently reviewed Airbnbs were considered. After the data was imported, feature engineering was done on the ‘reviews’ dataset. These variables were then aggregated and combined with the ‘listings’ data.

2.2 Listings

In order to have a better understanding of the dataset, some exploratory data analysis must be done. The listings dataset spans Airbnbs across the five boroughs of New York City: Manhattan, Queens, Brooklyn, The Bronx, and Staten Island. In Figure 2.1 below, each dot represents a different Airbnb.[7] In Figure 2.1, the red represents Airbnbs that are entire homes or apartments. The green, light blue, and black points represent private rooms, shared rooms, and hotel rooms respectively.

Figure 2.1: Map of Airbnbs in New York City.
The 2020 and 2021 January listings dataset spanned across 106 and 74 variables, respectively. Only variables included in both datasets were considered. Airbnb offers a plethora of options and filters to refine the perfect stay for its consumers. Users will mainly use the Type of place, Calendar, Guests, and Price filters. There are a variety of extra filters to choose from including: Instant Booking, Cancellation flexibility, Range, Verified, Beds, Bathrooms, Bedrooms, Business Travel, Superhost, Amenities, Facilities, Property type, Unique stays, House rules, Review Scores, and many more.

At a quick glance, the average values in Table 1 stay consistent between the two datasets apart from Minimum Nights, Availability, Number of Reviews Past 12 Months, and Host Listings Count, and Price. According to Airbnb, “Due to changing travel trends, it’s no surprise that long-term stays are on the rise. In August, the number of people who searched for longer stays was up more than 50% compared to last year.” Hosts have transitioned into wanting their guests to stay for a longer minimum number of nights by almost three times the previous year. This is understandable, since hosting longer saves time on cleaning and administration tasks. Hosts often allow their guests to stay for longer at a discount if the stay is longer than 28 days.
<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Pre Covid</th>
<th></th>
<th>During Covid</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Price</td>
<td>148.63</td>
<td>294.02</td>
<td>10</td>
<td>10000</td>
</tr>
<tr>
<td>Minimum Nights</td>
<td>5.57</td>
<td>14.8</td>
<td>1</td>
<td>1125</td>
</tr>
<tr>
<td>Maximum Nights</td>
<td>560.95</td>
<td>534.57</td>
<td>1</td>
<td>10000</td>
</tr>
<tr>
<td>Accommodates</td>
<td>3.05</td>
<td>2.03</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>1.22</td>
<td>0.78</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Beds</td>
<td>1.65</td>
<td>1.26</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>30 Day Availability</td>
<td>14.54</td>
<td>12.15</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>60 Day Availability</td>
<td>31.57</td>
<td>23.86</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>90 Day Availability</td>
<td>49.06</td>
<td>35.59</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>365 Day Availability</td>
<td>154.54</td>
<td>134.86</td>
<td>0</td>
<td>365</td>
</tr>
<tr>
<td>Number of Reviews</td>
<td>40.21</td>
<td>57.69</td>
<td>1</td>
<td>685</td>
</tr>
<tr>
<td>Reviews Past 12 Months</td>
<td>16.17</td>
<td>18.76</td>
<td>1</td>
<td>460</td>
</tr>
<tr>
<td>Review Score Rating</td>
<td>94.03</td>
<td>7.98</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>Review Score Accuracy</td>
<td>9.63</td>
<td>0.79</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Review Score Cleanliness</td>
<td>9.34</td>
<td>0.97</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Review Score Accuracy</td>
<td>9.76</td>
<td>0.68</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Review Score Communication</td>
<td>9.75</td>
<td>0.71</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Review Score Location</td>
<td>9.62</td>
<td>0.68</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Review Score Value</td>
<td>9.39</td>
<td>0.88</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Host Listings Count</td>
<td>7.91</td>
<td>57.48</td>
<td>0</td>
<td>2044</td>
</tr>
</tbody>
</table>

Table 2.1: Summary statistics for numerical Variables.
Review rates during COVID were cut in half compared to the previous year. It is no surprise during this pandemic that people are less likely to travel short term to New York City during this time. Although the average Airbnb guest has opted towards longer stays, availability across the board has also increased. There are more vacancies in the listings on average during COVID. This supports the idea that Airbnb guests are staying longer but booking less frequently.

The most important takeaway from Table 2.1 is the average price before and after the onset of COVID. The mean dropped from $148.63 to $124.56 in just a matter of a year. It seems like property owners are scrambling to rent out their stays by significantly dropping the price in order to entice prospective guests.

Airbnb categorizes the stays into either an Entire home/apt, Private room, Share room, or Hotel room. Entire home/apartments and private rooms represented 97% of the Airbnb stays. Though there was significantly less amount of available Airbnb listings in 2020, the proportion stayed roughly similar. Entire homes/apartments encompass 60.5%, private rooms with 37.2%, shared rooms with 1.5%, and hotel rooms with .8% make up the Airbnbs in New York City with an average cost of $122 per night[3].
Figure 2.2: Airbnb Prices of Entire home/apartments and private rooms by borough.

The figure above shows the distribution of Airbnb prices across the five boroughs of New York City. The number of stays in each borough stayed consistent, but the distribution of prices seemed to be skewed right. Most stays tend to be below 200$ a night but some offer experiences that consumers can pay a premium for. Some examples are stays titled “Beautiful Brooklyn Brownstone Production Space” or “Extra-Lux Executive Penthouse With Private Rooftop,” allowing the guests to essentially rent out a larger, more luxurious space to accommodate more people. These Airbnbs are pulling the mean price higher than what is expected.
2.3 Reviews

According to Airbnb, the review rate for Airbnb’s is approximately 70%. Airbnb consumers are encouraged to leave reviews after completing their stay. There is a 14 day period after leaving the Airbnb for the host and guest to comment and share any interactions they had during the booking. These reviews are crucial to providing insights and details not directly apparent through the advertised listing.

In the figure below, there is a clear trend showing how much Airbnb has grown up until 2020. There is some seasonality to the number of reviews each year but the average number of reviews has only increased up until March 2020 when COVID-19 was announced as a pandemic.

![Number of Reviews each year](image)

Figure 2.3: Number of reviews each year for Airbnbs in New York City.
Despite the impact of COVID-19 in New York City, less than 1% of reviews had any mention of COVID compared to over 25% mentioning the word “clean”.

Figure 2.4: NRC Emotion lexicon for reviews.

Without reading through all of the reviews, it’s hard to get a good understanding of what is happening. The words from 10,000 reviews were extracted to estimate the emotions associated with them. From Figure 2.4, reviews tend to be similar and positive in both time periods.

To get a sense of what is being said, topic modeling was done on the listing reviews. In order to get the data ready, the text data must be turned into a document term matrix. Words from the comments were taken and sorted based on their neighbourhood. After these tokens were unnested, common stop words and adjectives were filtered out of the wordlist. These words were transformed into a document term matrix to perform Latent Dirichlet allocation (LDA) topic modeling.
Chapter 3

Feature Engineering

With the understanding of the data gained through exploration, it can be prepared in a way that’s useful for the model. Feature engineering is the process of using domain knowledge to extract features from the raw data. It offers flexibility, speed, and models that are more easily understood.

![Histogram of log-transformed price](image)

**Figure 3.1:** Histogram of log-transformed price.

From Figure 3.1, it’s clear that the distribution of price is not normally distributed throughout the boroughs. There are most likely several extravagant Airbnbs in New York skewing the price. Since the goal of this is to determine factors that affect Airbnb price, a logarithmic transformation was done on price to improve the linearity between the dependent
variables. This transformation would minimize the effects of these extreme values, and provide a more accurate model that represents the majority of the dataset. A Shipiro-Wilk’s test on normality was run on a 500 sample of the variable ‘price’. The p-value was 2.2e-16. At-risk 5%, we reject the null hypothesis, the sample price did not follow a normal distribution. After a log transformation, the p-value was .07216. We fail to reject the null hypothesis at 5% risk, the sample is normally distributed.

![Histogram of Average Sentiment per Review](image)

**Figure 3.2:** Histogram of Average Sentiment for each review using Sentimentr.

To analyze how consumers felt using Airbnb during the pandemic, the reviews were run through the Sentimentr package. Sentimentr was designed to quickly calculate the sentiment polarity of the text at the sentence level. If given a group of sentences, the sentiment is aggregated. Reviews for each listing available were run through Sentimentr in order to get the average sentiment. Figure 3.2 is the output for average sentiment for the years 2019 and 2020.
Some reviews were not long enough to give out a sentiment rating so their given value was 0. Similar to Figure 4, most reviews tend to be moderately positive (greater than 0 in polarity).

Since NYC was labeled as an early epicenter of COVID-19 pandemic in the United States [2], Airbnb started enforcing a 5-step cleaning process that all hosts were required to follow. A perceived cleanliness indicator was calculated through the reviews available, labelled as ‘clean’. A score of 1 was awarded to reviews that included the word “clean.” This occurred in over 25% of the reviews. A score of -1 was given to reviews that included words that are often associated with “dirty”, such as “filthy”, “unsanitary”, “bugs”, “mold”, “cockroaches, “infested”, and other negators by the word “clean”. If the review did not include the following, a score of 0 was given. The sum of the scores were calculated for each listing. If the score was positive, the listing was labeled as “clean”. 98.3% of pre-COVID and 98.1% of during COVID era listings were perceived as “clean”. These values are reinforced according to a large-scale sentiment analysis on Airbnb reviews across 15 cities, “Results indicate that 98.1% of reviews and 76.4% of the sentences are positive while only 1.06% of the reviews and 4.7% of the sentences are negative.”.[8]

The number of unique neighbourhoods in New York City added up to 220 over the past 2 years. While neighbourhood is obviously an important factor when booking a vacation stay, these variables needed to be trimmed. Only neighbourhoods that had at least 100 listings were kept in the dataset. Since the same issue occurred in the type of property, only neighbourhoods with at least 500 property types were kept in the dataset. The remaining listings were labeled as “other”. From Figure 3.3, it’s to be expected that most of New York City are apartments.
After extensive exploratory data analysis, the following transformations were done to the rest of the explanatory variables in order to prepare for inputting into the XGBoost model:

- Extracted word count of ‘comments’, ‘description’, ‘neighbourhood_overview’, and ‘host_about’
- ‘host_is_superhost’, ‘host_has_profile_pic’, ‘instant_bookable’ to (1=TRUE, 0=FALSE)
- ‘first_review’ date converted to number of active years on Airbnb
- Amenities extracted and converted into dummy variables
- ‘room_type’, ‘host_response_rate’, ‘borough’ converted from factor to separate dummy variables

Figure 3.3: Property vs Room type of NYC.
Several columns of the dataset included many NA values. For categorical variables with missing values, the mode was imputed and converted to the binary components. For numerical variables, the mean was imputed.

### 3.1 Variable Selection

Variable selection is an important process to ensure the model fits with accurate predictions on the response variable. R-squared and RMSE are great at assessing the quality of fit but it’s impossible to add a predictor to a model and make these values worse. Forward and backward selection was used to sacrifice goodness-of-fit to obtain a smaller model that will still fit well. Fitting a larger model ends up fitting noise that is not relevant to the response variable.

In backward selection, all possible predictors are fit into the model. This process was done using the step function in R. At each step, it considers how removing a predictor will affect the outcome variable. The AIC, a measure for quality of the model, is evaluated during this time for each variable. The variable that offers the lowest AIC at each step is removed one by one until no variable will improve AIC. The process is the exact opposite with forward selection. The initial model starts with 0 predictors. At each step, R attempts to add a variable to the model until it finds a good model or reaches the maximum number of variables. Any variables included in either of these two selection methods were kept.

A random forest was run on the collective variables from the forward and backward selection. The importance of each variable was calculated with R’s importance() method in the randomForest library. To cut down on the number of variables and keep the model simpler, the top 30 values according to the mean decrease in node impurity.
From this variable importance plot, it’s clear to see that what kind of room has the greatest impact on the price, followed closely by the number of accommodations needed for the stay.
Chapter 4

Methodology and Modeling

Machine learning models were used on the predictors of the variable selection results. These algorithms will use the data and various assumptions to predict the output variable, the daily price of an Airbnb. The dataset was sampled and split into 80/20 train and testing data. These samples were implemented into Random Forest, Linear Regression, and XGBoost models to evaluate their performance metrics in $R^2$ and root mean square error (RMSE).

4.1 Random Forest

Random forest is a supervised learning algorithm used for classification or regression. It has become very popular because of the satisfying results with little to no distribution assumptions on the data. This method was used to start because of its good predictive performance, high interpretability, and the ability to weigh the importance of variables.

Random forest uses a collection of randomly created decision trees to reach a final recommendation[5]. When building the decision trees, each split is considered at a random sample of $m$ predictors[5]. The value is $m$ was chosen to be the square root of $p$, in this case, $m=8$. At each split, only a subset $m$ of the predictors is considered. The algorithm doesn’t consider the majority of available predictors to avoid all of the bagged trees being highly correlated. Each decision tree will predict a value for the response variable by taking the average
of each tree, in this case ‘price’. This iteration was repeated for a total of 500 trees on a 10% sample of the dataset.

4.2 Linear Regression and Perceived Cleanliness x COVID interaction

Multiple linear regression is used to find a linear relationship between the output variable and their predictors. Linear regression follows the equation:

\[ y = b_0 + b_1 x_1 + \ldots + b_n x_n + \varepsilon \]

Where \( y \) is the dependent variable, \( b_0 \) is the slope-intercept, \( b_1 \ldots b_n \) are the regression coefficients, \( x_1 \ldots x_n \) are the independent variables, and \( \varepsilon \) is the error term.

Unlike random forest, there are several assumptions for linear regression:

- Linearity: There exists a linear relationship between the predictor and response variables
- Normality of residuals: Residuals follow a normal distribution
- Independence of variables: No multicollinearity
- Homoscedasticity: Residuals have a constant variance across every predictor

Through the feature engineering and variable selection, primarily through taking the log of ‘price’. Any variables that exhibited any kind of multicollinearity were removed from the model.

An interaction term was between ‘clean’, the perceived cleanliness of the Airbnb listing through the reviews, and ‘covid’, whether or not the listings were active during the COVID-19 pandemic or in 2019. This interaction, ‘covid’ x ‘clean’, was used to examine the impact of COVID on the price if the listing was perceived as clean and during COVID (‘covid’=1 & ‘clean’=1). If either variable is 0, there would be no impact as this slope would be 0. The rest of
the predictors from the previous variable selection were added to the model. The results of this linear regression are in appendix A.1.

4.3 XGBoost

XGBoost, also known as eXtreme, Gradient Boosting, is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. Like random forest, trees are being built but sequentially such that each subsequent tree aims to reduce the errors of the previous tree. Each of the trees learns from the previous ones and the residuals are updated. The XGBoost trees makes use of trees with fewer splits that are highly interpretable. After an initial model is defined that tries to minimize mean square error. Each new model would be associated with the residual of the previous model. This can be done for however many iterations until residuals have been minimized. The XGBoost model in R for the purpose of this study had the following parameters: step size shrinkage ‘eta’ of .05, maximum depth of tree ‘max_depth’ of 6, subsample ratio of training instances ‘subsample’ of .2, and minimum sum of instance weight ‘min_child_weight’ = 1.5, ‘gamma’ = 0.[6]

4.4 Latent Dirichlet Allocation

The reviews went through Latent Dirichlet allocation (LDA) topic modeling in order to have a deeper understanding of what is actually being said in the reviews. This is a generalative probabilistic model for collections of discrete data. It treats each document as a mixture of topics and each topic as a mixture of words. This allows the documents to overlap over each other in terms of content. [9]
Figure 4.1: Graphical Model Representation of LDA

From Figure 4.1, $\alpha$ represents a topic distribution for each document. This gets fed into the outer rectangle where the topic distribution, $\theta$, is calculated for each document, $M$. This information is fed into the second layer. This layer contains $N$-words in one document generated from the previous layer. These $N$-words generate $z$, a topic from the distribution of words. Now the distribution, $\beta$, is the per-topic-per-word probabilities with $k$ number of topics. This generates words for the individual words for each topic in $z$. At the end of the model, $w$, a set of $N$ words for the topic. The reviews could have a plethora of topics ranging from the stay, neighbourhood, cleanliness, or host. For the purpose of this study, the number of topics $k$ was set to 2. [10]
Chapter 5

Results

Two main metrics the models were evaluated on were $R^2$ value and root mean square error (RMSE). $R^2$ value is a statistical measure of the proportion of variability in the predicted variable explained by the regression model.

$$R^2 = 1 - \frac{\Sigma(y_i-\hat{y})^2}{\Sigma(y_i-y)^2}$$

RMSE measures the standard deviation of the residuals about the fitted regression line.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \Sigma y_i - \hat{y}^2}$$

These values were calculated for each model for the 80 / 20 training and testing data splits.

<table>
<thead>
<tr>
<th>Model Results</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>Rsquared</td>
</tr>
<tr>
<td>Random Forest</td>
<td>.446</td>
<td>.602</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>.444</td>
<td>.590</td>
</tr>
<tr>
<td>LinReg CovidXClean</td>
<td>.408</td>
<td>.637</td>
</tr>
<tr>
<td>XGBoost Covid</td>
<td>.334</td>
<td>.776</td>
</tr>
<tr>
<td>XGBoost Pre-Covid</td>
<td>.309</td>
<td>.790</td>
</tr>
</tbody>
</table>

Table 7.1: Summary of results from modeling
It’s clear here that the XGBoost model proved to be the most effective overall. Both versions produce output R\(^2\) values higher than the random forest and linear regression while minimizing RMSE. For XGBoost Covid, 80.1% of the variability in price is explained by the model. There is a slight gap between the train and testing results for these models most likely due to some overfitting.

For the XGBoost models, variable importance was calculated and graphed below in Figures 7.1 and 7.2. Both value ‘accommodates’ the highest; however, having a private room is important pre-COVID. This could be interpreted as most guests are not willing to share an entire place due to social distancing measures. Having a stay that’s an entire home/apt is more important during COVID. Borough Manhattan drops off the importance during COVID, possibly because this area had population density and concentration of COVID cases. Those that sought out Airbnb’s in New York may have been searching for more stays out of Manhattan and into the suburbs.
Figure 7.1: Importance of variables in XGBoost model Pre-Covid.

Figure 7.2: Importance of variables in XGBoost model during Covid.
5.1 Topic Modeling

To get a sense of what the reviews were saying, Latent Dirichlet allocation was applied to the document term matrix of reviews.

The number of topics $k$ was set to 2. Figure 7.3 depicts the visualization of the two topics extracted from the comments after filtering certain modifiers and stop words. The most common words in topic one include “location”, “host”, “apartment”, “subway”, and “close”. This suggests a topic of where the Airbnb is and what it is close in proximity to. Manhattan is also on this list of words for topic one. The common words in topic 2 include “clean”, “apartment”, “recommend”, “host”, and “room”, suggesting that this topic represents the condition of the stay.

Figure 7.3: Topic Modeling with LDA.
When checking the validity of the interaction term between CovidXClean, a linear regression with just these variables was run. Table 7.2 displays the coefficients from the model output with a y-intercept of 4.54. If a listing was perceived as clean Pre-COVID, the listing would be valued at 22% higher than one not clean. Since the value of the exponentiated coefficients when ‘covid’==1 is below 1, prices for Airbnb would have gone down despite the listing being perceived as clean in the reviews; however, there is still a difference in almost 10% when looking at the difference between a stay that is not perceived as clean versus perceived as clean during COVID.

In order to check how valid the user reviews were, this result was also compared to the review polarity calculated by ‘average_sent’ (polarity sentiment value calculated by Sentimentr package). This is just an alternative method of measurement using Airbnb guest reviews. The mean of average sentiment, was found to be .61. Table 7.2 represents a confusion matrix for the polarity above and below the mean versus pre and during COVID.

<table>
<thead>
<tr>
<th>Interaction Effect - Predictor Coefficients</th>
<th>Pre-COVID</th>
<th>During COVID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Clean</td>
<td>-</td>
<td>.875**</td>
</tr>
<tr>
<td>Clean</td>
<td>1.22***</td>
<td>.970</td>
</tr>
</tbody>
</table>

Table 7.2: Linear Regression exponentiated coefficients for Log(price) ~ covid + clean + covid * clean
<table>
<thead>
<tr>
<th>Mean</th>
<th>Pre-COVID</th>
<th>During COVID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Sentiment &lt; .61</td>
<td>$134.24</td>
<td>$121.66</td>
</tr>
<tr>
<td>Average Sentiment &gt; .61</td>
<td>$149.20</td>
<td>$127.96</td>
</tr>
</tbody>
</table>

Table 7.3: Matrix of mean prices for average sentiment

Though the predictor for perceived clean review during COVID was not statistically significant, these averages behave similarly to the interaction effect that was calculated. It’s very apparent the impact on COVID hit Airbnb hard in terms of pricing.
Chapter 6

Conclusion

The objective of this research paper was to explore what effects Airbnb prices during the onset of COVID and to do a comparison between the previous year. Most of the factors were related to the type of place and how many guests Airbnb can accommodate. These results can often help the consumer get a good deal on an Airbnb, or to help the hosts strategically price their stay.

To reach this objective, I began with gathering the data and conducted exploratory data analysis. This was done to have a better understanding of the dataset to feature engineer extra variables to run the model on. The number of variables was filtered down through forward and backward selection and these features were modeled using random forest, linear regression, XGBoost, and Latent Dirichlet topic modeling. These models would be evaluated on their overall performance by minimizing RMSE while maximizing the goodness of fit.

The results find that during COVID-19, Airbnb activity dropped dramatically. The number of active listings dropped by 67.5% even after 9 months into the pandemic. Hosts wanted more long-term guests since the average minimum number of nights doubled. XGBoost weighted having an entire home or apartment rental and number of accommodations as the highest of importance after COVID-19.

Through topic modeling, polarity, and sentiment analysis, it’s clear that most Airbnb guests leave positive reviews on the website. The topics in these reviews tend to be about the
surroundings or condition of the apartment. From the analysis, COVID did not have an apparent
difference on how the users left reviews.
Chapter 7

Limitations and Future Direction

While this study showed some significance on the impact of COVID-19 on Airbnb prices, there are some limitations and improvements that could be made. At the time of this study, the scraped data offered from InsideAirbnb were missing several impactful variables that most likely affect the price. These values, such as cleaning fees and security deposits, were present in the data years before. Oftentimes when trying to book a room, this is the deterrent as there is no cap to how high these values can be. There is usually a very high correlation between the daily price and the costs of these fees.

Some possible improvements to this research could be processing the data from every month of the year or broadening the scope of the study to other cities. New listings are often added so with this time series data, each listing can be followed and monthly progress and statistics could be calculated. Listings are often also taken off Airbnb. The months before each listing that leaves could be analyzed to find any trends. Seasonality effects can also be modeled to investigate prices around the holiday season.
## APPENDIX

Table 1: Linear Regression

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>log(price)</th>
</tr>
</thead>
<tbody>
<tr>
<td>room_type_EntireHomeOrApt</td>
<td>-0.30***</td>
</tr>
<tr>
<td>room_type_PrivateRoom</td>
<td>-0.72***</td>
</tr>
<tr>
<td>accommodates</td>
<td>0.11***</td>
</tr>
<tr>
<td>bedrooms</td>
<td>0.16***</td>
</tr>
<tr>
<td>beds</td>
<td>-0.04***</td>
</tr>
<tr>
<td>Lockonbedroomdoor</td>
<td>-0.03**</td>
</tr>
<tr>
<td>longitude</td>
<td>-0.71***</td>
</tr>
<tr>
<td>latitude</td>
<td>-0.50***</td>
</tr>
<tr>
<td>average_sent</td>
<td>0.05***</td>
</tr>
<tr>
<td>amenities_wc</td>
<td>0.001***</td>
</tr>
<tr>
<td>availability_90</td>
<td>0.001***</td>
</tr>
<tr>
<td>number_of_reviews</td>
<td>-0.0003***</td>
</tr>
<tr>
<td>review_scores_rating</td>
<td>0.01***</td>
</tr>
<tr>
<td>host_since_years</td>
<td>0.02***</td>
</tr>
<tr>
<td>number_of_reviews_ltm</td>
<td>-0.003***</td>
</tr>
<tr>
<td>calculated_host_listings_count</td>
<td>-0.004***</td>
</tr>
<tr>
<td>maximum_nights</td>
<td>0.0000***</td>
</tr>
<tr>
<td>TV</td>
<td>0.11***</td>
</tr>
<tr>
<td>maximum_minimum_nights</td>
<td>0.002***</td>
</tr>
<tr>
<td>minimum_nights</td>
<td>-0.003***</td>
</tr>
<tr>
<td>room_type_SharedRoom</td>
<td>-1.12***</td>
</tr>
<tr>
<td>review_scores_cleanliness</td>
<td>0.05***</td>
</tr>
<tr>
<td>Shampoo</td>
<td>0.09***</td>
</tr>
<tr>
<td>review_scores_location</td>
<td>0.05***</td>
</tr>
<tr>
<td>review_scores_value</td>
<td>-0.07***</td>
</tr>
<tr>
<td>review_scores_accuracy</td>
<td>-0.04***</td>
</tr>
<tr>
<td>host_is_superhost</td>
<td>0.06***</td>
</tr>
<tr>
<td>Borough_Manhattan</td>
<td>0.24***</td>
</tr>
<tr>
<td>Constant</td>
<td>-28.64***</td>
</tr>
</tbody>
</table>

Observations 9,784
R² 0.59
Adjusted R² 0.59
Residual Std. Error 0.44
F Statistic 501.10***

*Note: *p<0.1; **p<0.05; ***p<0.01

Table A.1: Linear Regression Output
REFERENCES


