**4.1 Project 1: Airbnb Pricing Predictions**

Ajuni Sohota

**Business Problem**

For those that want to use Airbnb, both the listers and renters will want to feel that they are being paid a fair amount and are paying a fair amount, respectively. Listers may feel overwhelmed when deciding what price to list their property as. They won’t want to list prices so high as to lose potential income from renters that feel that the price is outrageous, and they also won’t want to list so low as to be losing potential profits. With this analysis, answers to this predicament may become clearer. By understanding what kind of amenities, locations, or any other listing attribute may be indicative of price trends, listers will better be able to pick a price that is best for both them and potential clients. Airbnb would also benefit by providing a guide for clients based on geographic and attribute conditions for listings because it would give their clients peace of mind that they are listing at an appropriate price.

**Hypothesis**

For this project, I hypothesized that certain features would correlate strongly with list prices. Certain information such as whether a host is a “super host”, the types of amenities offered, and location would be important factors in how prices can be predicted accurately. I believe that there are relationships between the listing data and the pricing, and that from performing predictive analytics and various machine learning models, I will be able to predict prices for listings well.

**Methodology**

The first step in this project is to perform exploratory analysis where the dataset is analyzed and parsed through to find out what the dataset looks like and what kind of information I am working with. Through the EDA, I explored whether all of the columns were necessary for my analysis, and included explanations as to whether or not I wanted to keep them. I looked for null values and the distribution of the numerical values to determine if I wanted to exclude any of the information for the predictions. I made a correlation matrix for the numerical values to see what seemed to correlate most with pricing of listings.

I also mapped the price points of Airbnb listings to a map of San Francisco to see where listings may tend to be more expensive, or if there doesn’t seem to be a trend in that way. Then I officially dropped values and columns I didn’t need. I transformed data that needed to be altered to be numerical as well as used one hot encoding for the categorical data. After feature engineering, I did a train-test-split before creating the data pipeline. I performed base evaluations for various models and then performed these models afterwards with cross-validation. The train and test scores were then plotted to see which model was performing the best alongside with the previous cross-validation RMSE score mean. The next step was fine tunning the model followed by presenting the feature importance. XGBoost uses “second-order gradients” to derive it’s approximations whereas regular gradient boosting uses the base model of decision trees (). Another difference is that is uses advances regularization that can also improve your predictive outcome ().

**What Model Did I Use?**

The models that I used were between gradient boosting and XGBoost. Boosting works in the way that it builds models from “individual so called ‘weak learnings’ in an iterative way” (Elsinghorst, 2018). The name boosting can be understood by the way in which it works, such that each sequential individual model built is random but then puts more weight on predictions that are wrong or have high errors, so that it focuses on these during the learning process of the model (Elsinghorst, 2018). This can be thought of how we as humans learn from mistakes, but this model learns by errors as each individual model builds on itself. XGBoost is “Extreme Gradient Boosting” which you can assume by the name: XGBoost using gradient boosting but uses “more accurate approximation to find the best model” (Elsinghorst, 2018).

**Results**

After cross-validating my models, I picked XGradient Boosting Regression as the best model because it had the lowest RMSE score that was still between what is considered a good RMSE score. The learning curves also showed me that the model was scalable and had a good performance. After fine tunning the model, I got the best score using this model with an RMSE of 0.36.

**Table 1. Base Evaluations Models and RMSE Scores**

|  |  |
| --- | --- |
| **Model** | **Base Evaluation RMSE Score (Train Data)** |
| Linear Regression | 0.44 |
| Ridge Regression | 0.44 |
| Decision Tree Regression | 0.0 |
| Random Forest Regression | 0.15 |
| Gradient Boosting Regression | 0.35 |
| XGBRegressor | 0.35 |

**Table 2. RMSE Scores for Best Models**

|  |  |
| --- | --- |
| **Model** | **RMSE Mean with Cross-Validation** |
| Ridge Regression | 0.4822 |
| Random Forest Regression | 0.4064 |
| XGBRegressor | 0.4058 |
| **Model** | **RMSE Score after Fine Tunning** |
| XGBRegressor | 0.3647 |

**Table 3. Top Ten Feature Importance and Weight Score**

|  |  |
| --- | --- |
| **Feature** | **Importance Weight** |
| Bedroom Count | 0.169 |
| Accommodates | 0.028 |
| TV | 0.026 |
| Bath Towel | 0.020 |
| Bathrooms Count | 0.019 |
| Bathroom Essentials | 0.019 |
| Beds | 0.016 |
| Gym | 0.015 |
| Elevator | 0.014 |
| Shampoo | 0.013 |

**A picture containing window, shoji, building, station

Description automatically generated**

Figure . Distribution of Numerical Columns

**Chart, bar chart

Description automatically generated**

Figure . Most Advertised Property Types

Map

Description automatically generated

Figure . Map of Listings in SF by Property Type

A picture containing text, scoreboard

Description automatically generated

Figure . Correlation Matrix of Numerical Columns

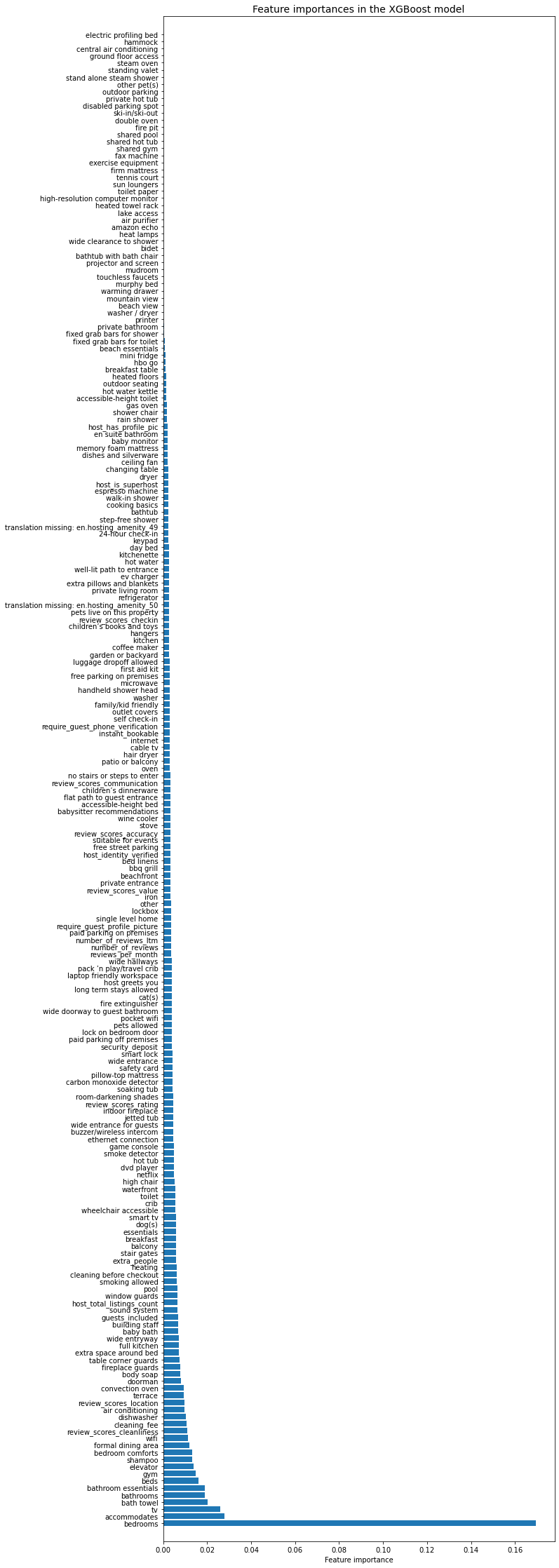


Figure . Feature Importance Graph

Chart, line chart

Description automatically generated

**Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated**

Figure . Learning Curves for Three Cross-Validated Models

**10 Questions Audience May Ask**

1. How do you think Covid-19 has affected the feature importance?

I think that many of the things that ranked on the top 10 would not change, however, the cleaning fee may increase which would impact the overall price. The other thing is that Covid-19 drastically disrupted the ability for people to travel as well as the safety in which people could even rent via Airbnb, leading to an initiative for Airbnb to use listings as safehouses for refugees.

1. Why get rid of the null values instead of replacing nulls with the median value?

The two options here are possible. I chose to get rid of null values because I wasn’t sure if they would impact the analysis if they were wrongly assumed as the median. I also didn’t see too many listings with null values that were in columns that I was keeping for analysis.

1. What is a good RMSE score?

A good RMSE score is between 0.2 and 0.5. It is also important to note that a smaller RMSE score indicates a better fit. (Malekineshad et al., 2020)

1. What are downstream applications of this research?

Downstream applications could be that the app would allow listers to be recommend pricing for listings. The other potential application is for renters to see how a price for a listing compares to others that have similar features, so renters can pick listings that feel like a fair list price.

1. What kind of problems arose while you were doing this research?

One of the problems that arose for me was that after my train-test-split, I realized I had done something wrong during the transformation of the categorical features because the train vs. test didn’t match up the way they should. From there I had to deal with the problem which laid in the way I was dealing with the amenities column, which was formatted with each amenity listed in a sentence separated by commas.

1. Why should we care what affects pricing of listings?

The reason we should care is so that as renters or rentees, you can either see what types of things will make your listing more lucrative and respectively then see what listings are a rip off. It will allow both to make informed decisions.

1. What are Airbnb’s competitors?

Airbnb competitors include websites like Vrbo, Booking.com, Tripadvisor, Agoda, Expedia, TUI Villas, TravelStaytion, HomeToGo, Plum Guide, and Google.

1. Do Airbnb’s competitors have price suggestions?

Yes, some do in a sense that they can recommend better or worse listings with similar prices, which helps the renter decide what they want to go with.

1. How did you pick which model to go with?

XGB Regressor won only by a tiny bit, but it had the best RMSE score.

1. How did you deal with the amenities column?

By first separating each amenity by commas and then using CountVectorizer to transform the amenities. After that I concatenated it back to the dataframe.

**References:**

Elsinghorst, D. S. (2018, November 29). *Machine learning basics - gradient boosting & xgboost*. Shirin's playgRound. Retrieved November 16, 2021, from https://shirinsplayground.netlify.app/2018/11/ml\_basics\_gbm/.

Malekinezhad, H., Nachtnebel, H. P., Klik, A., Hiriyannaiah, H., Walker, D. D., & Ágreda, E. A. P. D. (2020, December 2). *What's the acceptable value of root mean square error (RMSE), sum of squares due to error (SSE) and adjusted R-square?* ResearchGate. Retrieved November 16, 2021, from https://www.researchgate.net/post/Whats-the-acceptable-value-of-Root-Mean-Square-Error-RMSE-Sum-of-Squares-due-to-error-SSE-and-Adjusted-R-square#:~:text=Based%20on%20a%20rule%20of,more%20is%20acceptable%20as%20well.