Building a Generalizable Machine Learning Model to Predict Airbnb Price Across U.S. Cities

Bjorn Commers Jinlu Ma Alicia Teo Alexandra Rumyantseva

Abstract

This project attempts to build a generalizable price prediction tool for Airbnb rental units. Specifically, the objective was to see if it is possible to train a model on the data from one city (New York City) and then use that model to predict listing prices in a different city (San Francisco). Although we were able to generate good results using same-city data, the models had a difficult time generalizing across the two cities.

1. Introduction

Airbnb as a home-sharing platform has become one of the largest accommodation platforms in the world, with more than 2,000,000 listings globally. As Airbnb continues expanding to new cities, the uniqueness of each city and each property makes setting the right price a challenging task for property owners. Given the importance of location to property pricing, previous work was focused on learning on one city, and then predicting rental price for the same city. This paper aims to solve the pricing problem for property owners in cities new to Airbnb, and to explore the possibility of generalizing price prediction across cities.

The specific problem we explore is, for a given date, train a model on New York City data, and then accurately predict the Airbnb listing prices of properties in San Francisco. To develop a generalizable model, we constructed features that could have explanatory power across domains (different cities). We also explored whether using different machine learning models (Linear Regression, LASSO Regression, and Support Vector Regression) would enable better generalizability.

2. Related Work

With Airbnb being a relatively recent phenomenon, there is limited research in this area. A scan of past work on Airbnb pricing found that most research was focused on analyzing and predicting Airbnb prices within a specific city. Gibbs et al (2018) **[1]** analyzed the determinants of Airbnb pricing for 5 individual Canadian cities. Li et al (2019) **[2]** hypothesized that landmarks and room facilities affect Airbnb prices. They used Airbnb data from three cities (London, Tokyo, and Los Angeles), clustered properties according to distance to city landmarks, and applied a Linear Regression model to predict Airbnb prices within each city.

Looking into a related class of studies, house price prediction, we find a range of approaches towards predicting house prices. Wu (2017) **[3]** explored different feature selection methods together with Support Vector Regression to predict house prices in King County, USA. Chopra (2007) **[4]** modeled house prices as a product of intrinsic price (i.e. dependent on the house's features) and desirability of location, and then used a locally weighted linear regression to predict house pricing in Los

Angeles. In a similar study, Caplin (2008) **[5]** found that systematic patterns in house prices were greatly reduced when taking geography explicitly into account, and that using a nearest neighbour approach improved the results of his regression model. Finally, Malinowski (2018) **[6]** utilized clustering algorithms to divide a city into different areas, with the clusters then establishing the basis for property valuation in that area.

Based on past research, it is evident that location is extremely important in determining the price of a property. Given the importance of location, most research was focused on learning and then predicting prices for the same city. In this paper, we study the novel problem of generalizing price prediction - namely, being able to use a model trained on one city, to predict Airbnb prices in other cities.

3. Dataset

The public Airbnb dataset **[7]** was the main data source for this study. The dataset included 48, 378 properties in New York (NY) and 8,112 properties in San Francisco (SF).

We preprocessed the data in the following way:

1. Removed features with frequent missing fields
2. Set any 'NaN' missing values to 0
3. Converted categorical features (e.g. 'property type' and 'room type') to one-hot encoding
4. Removed irrelevant or uninformative features (e.g. picture url, host_picture_url)
5. Bucketed 'user review score' (the average review score for a property) from a continuous variable (e.g. score = '98") into score buckets of 10 (e.g. score = "80 - 90") and then hot-encoded
6. Log transformed 'price' (the price to stay a night at the property) to reduce skewness of the data

4. Task Definition

Our model takes in 61 inputs of a specific listing, such as host_is_superhost, bathrooms, bedrooms, number_of_reviews and outputs a predicted price for a given date in a year (September 2nd). The model is trained on New York City data, and tested on San Francisco data. The goal is to have a generalizable model that produces a small Mean Squared Error when used to predict San Francisco (SF) rental price when the model is trained on New York (NY).

5. Feature analysis

To enable generalization, we searched for features that would have explanatory power across cities and across properties. Past research from Gibbs et al (2018) and Wang et al (2017) who studied the price determinants of Airbnb properties, suggest that location (distance to the city center, transportation hubs, and major attractions) as well as review ratings significantly impact price.

We created 6 additional features as inputs to our model, appending the features to each property in the Airbnb dataset.

Section 5.1 describes each additional feature as well as our rationale for including the feature.

5.1. Description of Additional Features

5.1.1 Review Sentiment

Airbnb customer reviews is one of the most important factors in pricing a property: reviews generate traffic and increase the rental property's rank on the site; positive reviews attract more renters. Both of these contribute to a higher demand.

Therefore, each customer review is analyzed using sentiment analysis and added as an additional feature to the model. We used Python's TextBlob library to perform sentiment analysis, where it uses PatternAnalyzer to produce a polarity score from -1 (most negative) to 1 (most positive) for each piece of review. In the end, all of the review sentiment scores for each property is averaged for that property, which is added as a new feature to the model.

5.1.1 Poverty level

One of the most important area requirements by the tourists is their safety. The higher the area wellness is, the higher the listing price in it. Such an evaluation can be approached from different angles: crime or poverty rate, average income, etc. While the crime rate would probably be ideal, it was not easy to obtain and aggregate since it is separated into multiple categories and not organized by the zip code. Therefore, we chose to analyze the poverty rate instead. From census.gov, we extracted the total number of submitted tax returns per household and those that have earnings below the poverty level. From there, we calculated the poverty rate per zip code.

5.1.2 Commute time

Another criterion that tourists look at is how close they are to civilization and how useful is the transportation. Business travelers would also be concerned about commuting times to business districts.

We obtained data from AutoAccessoriesGarage on average commute times for each zip code in New York and San Francisco.

5.1.3 Distance to downtown

Location convenience can be also defined by the distance of the property to the city center. Properties who can claim to be right in the downtown tend to get a valuation boost. To get this data, we used the Foursquare API. For each property in NY and SF, we called the Foursquare API, and obtained the distance of that property from the city's downtown area.

5.1.3 Number of hotels within the area

A higher number of hotels in the immediate area indicates the competitiveness and demand of the area. For our analysis, we assumed that a half-mile radius around each property would be considered "immediate area". We again used Foursquare API, and searched for hotels within a 0.5mi radius of each property in the Airbnb dataset. The count of hotels within 0.5 mi of that property is added as a feature to the model.

5.1.4 Restaurant density

The number of restaurants nearby a rental property is an indicator of the popularity of the area, and an extremely important factor for tourists when choosing a short-term stay. Therefore, restaurant density near a property could potentially result in an increase of rental price. We used the Yelp API to get the number of restaurants in each zipcode in our dataset.

6. Experiments

6.1 Methods

We compared three different regression models. The first and baseline model is linear regression.

The second model is LASSO regression, which adds a penalty term to linear regression to encourage sparse models. Our hypothesis is that a more parsimonious model selected by lasso regression would allow for better generalization.

The third model is Support Vector Regression (SVR). The SVR uses the same principles as Support Vector Machine, aiming to maximize the margin between the closest points and the separating hyperplane. The regularization in SVR theoretically should lead to better generalization of SVR.

6.2 Results

6.2.1 Evaluation metric

Mean Square Error (MSE) is used as the evaluation metric for the model. MSE measures how close the predictions are to the actual value by averaging the sum of squares of the difference between the predicted values and the true values.

We will compare our model's results against the benchmark case where a model is trained on 80% of the Airbnb New York data and tested on 20% of New York data. The data includes information on the zipcode of each property, but none of our additional features such as hotel density.

The ultimate goal is to be able to train our generalizable model on NY, predict on SF data, and achieve similar levels of MSE as the benchmark. However, given that SF zipcodes and NY zipcodes are mutually exclusive, such a model would not be generalizable. A more realistic benchmark might be to aim to achieve the same test MSE as a model trained on NY without zipcode information, and then tested on NY.

Table 1: Benchmark MSE

Model	**Dataset**	**Test (NY) MSE**
Linear Regression	Airbnb dataset including Zipcode information for each property	0.10
	Airbnb dataset without Zipcode information	0.18

6.2.2. Initial results

We trained each of our models on 80% of Airbnb NY data (without zipcode information), holding out 20% of NY data as validation. We then tested the trained model on Airbnb SF data.

For all three models, MSE for SF was much higher than MSE for NY, indicating that all the models are overfitting on New York data.

Table 2: Overfitting on New York

Model	**Validation (NY) MSE**	**Test (SF) MSE**
Linear Regression	0.20	0.32
Lasso	0.25	0.35
SVR	0.16	0.30

6.2.3. Results after adding features

We added our additional features (Poverty level, commute time, distance to downtown, number of hotels within the area, restaurant density) to the Airbnb dataset, and trained each of our 3 models on New York data, and tested on San Francisco.

For all 3 models, MSE for SF was lower compared to the initial results. However, MSE for none of the models reached the benchmark of 0.10 or even 0.18.

Table 3: Reduced Test MSE after adding features

Model	**Validation (NY) MSE**	**Test (SF) MSE**
Linear Regression	0.14	0.25
Lasso	0.17	0.29
SVR	0.14	0.22

Result Discussion

Overall, SVR enabled the best generalization, and LASSO performed the worst in terms of generalization. It was interesting that LASSO performed worse on NY, and also generalized on SF worse than linear regression. This was contrary to our expectations that LASSO would generalize better than a simple linear regression. One reason is that several features are highly correlated (e.g. 'beds' and 'bedrooms' have a correlation of 0.67), which leads to LASSO dropping variables that could have had predictive power.

accommodates	beds	0.757100
beds	bedrooms	0.665712
accommodates	bedrooms	0.662277
Hotel room	Hotel	0.612392
accommodates	guests_included	0.597710
beds	guests_included	0.519018
accommodates	cleaning_fee	0.482448
	Entire home/apt	0.477641
host_since	host_identity_verified	0.475762
guests_included	bedrooms	0.473097

Fig 3: Top 10 most correlated features

Looking at the non-zero coefficients in the lasso model, 4 of our additional features were included. This indicates that the additional features were indeed useful.

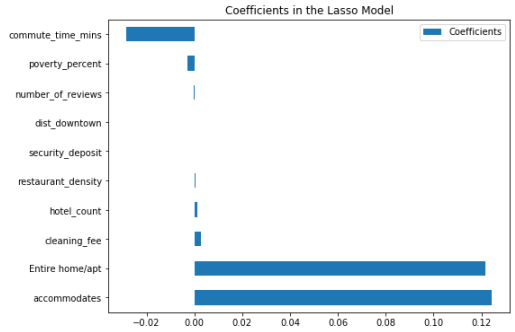

Fig 4: Non-Zero Coefficients in Lasso Model

6.2.4 Graphs & tables

Fig. 1 shows a plot of predicted prices versus actual prices, where blue represents actual price, and red represents predicted. We can see that there are quite a few outliers where our prediction missed by a significant amount.

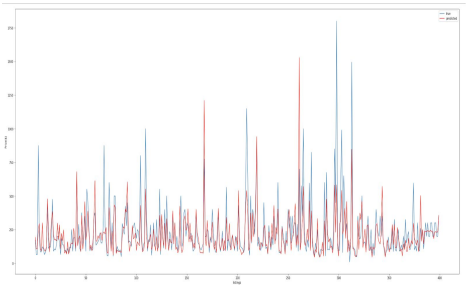

Fig. 1: SVR results plot

Fig 2 shows the distribution of Airbnb listing prices in NY. We decided to transform price by taking the log, because an initial look at the data showed that prices were extremely skewed to the left.

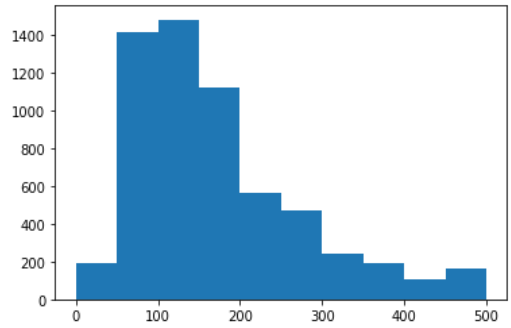

Fig. 2: Distribution of prices

7. Discussion

Compared to other previous work that we researched, a particular strength of our approach is that we approached location differently. Instead of analyzing location by hot-encoding zip codes, we used indicative feature of area and individual property locations. That allowed for some generalization among the areas, so the regression could adapt according to convenience or wellness of the area rather than its name.

On the other hand, results indicate that the weakness of model is the sole concentration on location analysis. While it is certainly an important factor in demand and thus pricing decision, it is still not the only one. Other factors that contribute to price discrepancy among different cities could include cost of living, for example. In addition, tourists look at photos of the property and decide whether they like it or not. Some landlords invest a lot in decoration, while others just make sure that guests have a place to sleep. Our model ignores this issue unless a specific concern was raised in the reviews - our sentiment analysis does catch negative feedback.

8. Conclusion

Although our models did demonstrate quite good results when being trained on same-city data, the results did not generalize across cities as well as we would have liked. This leads us to the conclusion that quantifying the difference between cities is quite difficult. For example, how does one quantify the “brand name” of New York City as compared to San Francisco? What about proximity to other wealthy nations from which tourists arrive? There are many number of reasons that two cities will differ in pricing of rental units and we have found it to be quite difficult to capture these nuances in our models.

9. Future Work

Other than continue exploring the underlying pricing differences across cities, our model can also benefit from property photos analysis. While the features that we considered such as location and number of tenants are important, tourists also carefully look at the pictures and rely on their feelings. While not impossible, human emotions are not easy to calculate and predict by a mathematical model. The colors used in the property design can be analysed from cold and warm perspective since it was shown in other papers that warm hues are more enjoyable and give a sense of comfort. However, other factors must be taken into consideration; and they are not easy to calculate from pixel manipulation.

Another interesting feature would be seasonality that includes both changes during the week and between the seasons and holidays. Current research shows that while some homeowners do raise their prices for the weekend, majority of them fail to practice dynamic pricing. Therefore, the current model can not benefit from such adjustment. As for different seasons and holidays, at this moment there is not enough calendar data available to make a general trend in price fluctuations. It will, however, be very useful in the future once enough data is collected.

References

[1] C. Gibbs, “Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings”, 2018.

[2] Y. Li, S. Wang, T. Yang, Q.Pan, J. Tang, “Price Recommendation on Vacation Rental Websites”, 2019.

[3] J. Wu, “Housing Price Prediction Using Support Vector Regression”, 2017

[4] S. Chopra, “Discovering the hidden structure of house prices with a Relational Manifold Model”, 2007.

[5] A Caplin, “Machine Learning and the Spatial Structure of House Prices and Housing Returns”, 2008.

[6] A. Malinowski, “An Approach to Property Valuation Based on Market Segmentation with Crisp and Fuzzy Clustering”, 2018.

[7] Airbnb public dataset. http://insideairbnb.com. Accessed: 02-Sep-2019