Understanding Airbnb Rental Prices in Nashville[*]

David Taylor

December 1, 2019

Introduction

In the past decade, cities around the globe have seen a sharp increase in the number of properties available for short-term rental. This growth has been supported by platforms such as Airbnb, which allows ”guests” to rent properties from ”hosts” for short stays. Nashville, being a popular destination for tourists around the world, has seen a large number of its properties converted to Airbnb listings to accommodate for tourist demand. Hosts can set the nightly prices of their properties on Airbnb, and this project aims to leverage statistical learning techniques to reveal the factors that drive the hosts' pricing decisions in Nashville. The goal of this analysis is not necessarily to build a model to accurately predict the price of an arbitrary Airbnb listing; rather, the goal is to build a model whose learned parameters reveal the factors that are most influential in the determination of Airbnb listing prices in Nashvile. This goal is relevant to members of the Vanderbilt community as we seek to understand the economic forces shaping the fast-growing neighborhoods around our institution.

Data

The data was downloaded from InsideAirbnb.com, an independent platform which scrapes listing data from Airbnb.com on regular intervals to provide information to the public that Airbnb not directly provide. The data used for fitting a model in this analysis was scraped in September of 2019, and it contains $n = 4958$ observations (listings on Airbnb.com) in $p = 59$ variables. The response variable for this analysis is `price`. The original data set was reduced from 7282 observations to exclude observations with a large number of missing data values as well observations with outlier price values (outside the middle 95^{th} percentile of price values). For each listing, there are features relating to the nature of the property (e.g., bedrooms, bathrooms, geographical coordinates, square feet) as well as the host of the listing (e.g., number of reviews, cancellation policy, average review scores). Additionally, I engineer several new

[*] The Python script used for this analysis can be viewed here

features for analysis: years since host joined the platform, distance from a popular tourist spot in Nashville, and dummy variables to indicate the type of the property and types of beds within the property. These engineered features are used to reduce dimensionality of the data set and provide the models more meaningful data to train on. See Appendix A for several informative figures regarding the dataset.

Methodologies

The data is first split into a training set (70%) and a test set (30%) at random. The metrics chosen to evaluate different models are sum of squared errors (SSE), mean absolute average (MAE) and R^2, allowing us to measure the variance in price explained by each model, as well as the absolute prediction error on the test set. We proceed first by fitting a simple OLS model to obtain baseline evaluation metrics. We then use k-fold cross validation with $k = 20$ to fit LASSO and Ridge regression models to the training data with the optimal regularization parameters α. Because these three models perform poorly on the test set, we move to tree-based methods. We fit a Random Forest regression model to the training set and achieve better results on the training set, suggesting that the high dimensionality of the dataset is best approached with a ensemble tree-based learning model. In light of this, we then fit an XGBoost model to the training set and achieve a better fit on the test set. XGBoost is a tree-based method that uses ensemble learning and gradient boosting to minimize errors in sequential models of boosting. It also uses L_1 and L_2 regularization to decrease the influence of irrelevant features. We tune the hyperparameters of this model and achieve the best results on the training set.

Results

The test results of all tuned models are summarized in the table below:

Table 1: Model Results

	OLS	Ridge	LASSO	Random Forest	XGBoost
R^2	0.536	0.537	0.473	0.743	0.755
SSE	2.439e7	2.436e7	2.774e7	1.353e7	1.287e7
MAE	$84.73	$84.37	$89.57	$59.73	$58.20
MAPE	48.05%	47.72%	50.75%	32.63%	30.85%

The XGBoost model is the best performer, achieving an R^2 value of 0.755 and MAE of $57.71. This suggest that the model explains close to 76% of the variance in price. It should also be noted that the most successful XGBoost model uses a high degree of L_1 regularization and a very small degree of L_2 regularization, suggesting that some coefficients have been set to 0 by the model.

However, the model is fairly poor at predicting the prices of new listings, as the mean absolute percent error is 30.85%. This is to be expected for two main reasons: 1) predicting the prices of properties is dependent on a high number of factors, some of which are likely not included in our dataset, and 2) there is much random noise stemming from the fact that hosts are able to set their properties' prices and are likely to act irrationally in the market on occasion. Much of this noise is present in the higher price ranges, as high-price listings are much more rare and suggest there may be non-linearity in the data. See Appendix B for visualizations of the model's performance. However, because the R^2 value is relatively high for this task, we can safely interpret the gain that each feature achieves in the model. In the XGBoost model, a feature's gain value refers to the sum of the decreases in squared errors that the feature induces in each tree that it is found on. It is essentially the feature's relative contribution to the model's accuracy. In the table below, we summarize the gain value and correlation coefficient with `price` of the 10 features in the model with the highest gain.

Table 2: XGBoost Regression Feature Importances

Feature	Gain	Correlation with Price
Host Num. Listings	7.59e5	0.499
Num. Bathrooms	4.96e5	0.454
Num. People Accommodated	2.91e5	0.412
I(Condominium)	2.88e5	-0.034
I(Super Strict Cancellation Policy)	2.58e5	0.269
Cleaning Fee	1.65e5	0.397
Num. Bedrooms	1.41e5	0.383
Host Cleanliness Avg. Review Score	1.38e5	0.006
Host Days Since First Review	1.30e5	-0.088
I(Host is Superhost)	1.25e5	-0.099

Analysis

We see that `Host Num. Listings` yields the largest gain in the model and positively correlates with `price`, suggesting that hosts with a larger number of listings on the platform are able to charge more for their properties due to their greater experience with hosting. `Num. Bathrooms`, `Num. People Accommodated`, and `Num. Bedrooms` unsurprisingly yield high gain with positive correlation with `price`, indicating that larger properties with higher capacity are more expensive. Similarly, condominium listings are marginally less expensive than house listings, as they are likely to be smaller. There are some other interesting inferences that we can make. First, we see that the presence of a `Super Strict Cancellation Policy` is an extremely important feature in predicting price, and it corresponds to a higher price. This is because the strict cancellation policies are rare and likely to be used only when the host's property is in high demand. Also, the bathroom count of a property is a much better predictor of price than bedroom count, indicating that consumers are more likely to pay a lower price for a multi-bed bedroom than for a larger space that requires more bathrooms. Also, we see that Superhosts are generally unable to charge more for their properties, suggesting users are willing to pay more for less-reviewed but higher-capacity properties. Perhaps the most interesting inference we can make is that the most important feature in predicting listing price is `Host Num . Listings`. There are real estate companies in Nashville operating as hosts on the platform with a large number of listings, as these companies are able to pool investor resources to buy and rent out many properties. Many Nashville residents and government officials take issue with this, and it is illegal in some cases. These firms are able to profit by renting out properties that might be either designated by law for long-term lease only or under rent controls by local governments. It also allows the firms to control large portions of the short-term lease market, squeezing out the smaller-scale hosts that Airbnb claims to support and raising prices as they please, knowing that tourists are likely to pay a higher price for a short stay in Nashville than a long-term resident would be able to pay. It is likely that, in the future, regulatory action is taken by local governments or Airbnb in order to stifle the growth of these firms, as this analysis has shown that their market share has become a driving force in the upward push on housing rental prices in Nashville, as well as hundreds of other cities around the world.

Appendix

A. Exploratory Data Analysis

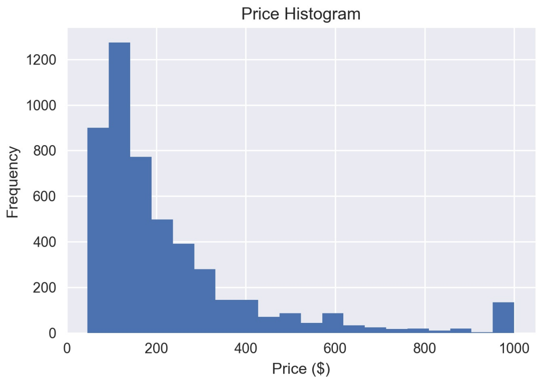

Figure 1: Price Histogram

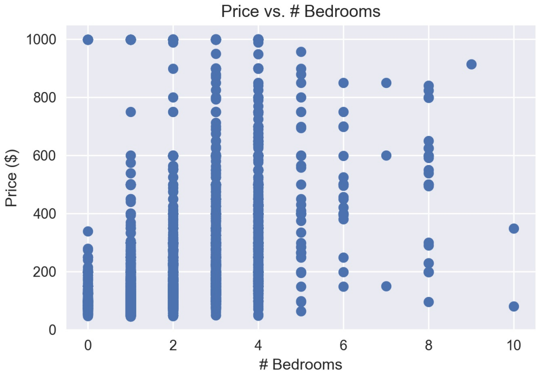

Figure 2: Price-Bedroom Scatterplot

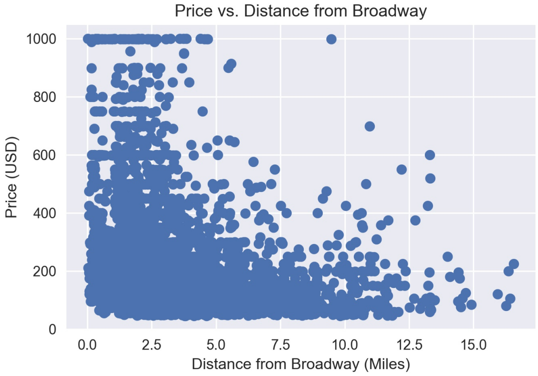

Figure 3: Price by Distance From Broadway

B. XGBoost Model Evaluation

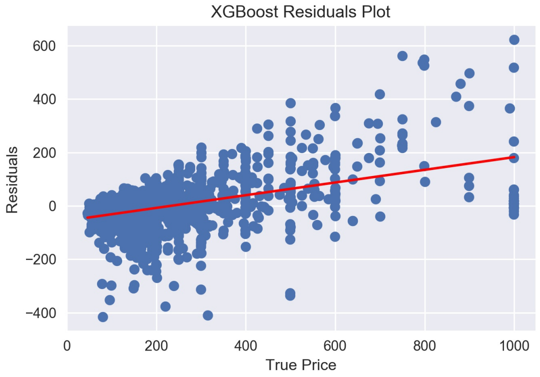

Figure 4: Residuals Plot

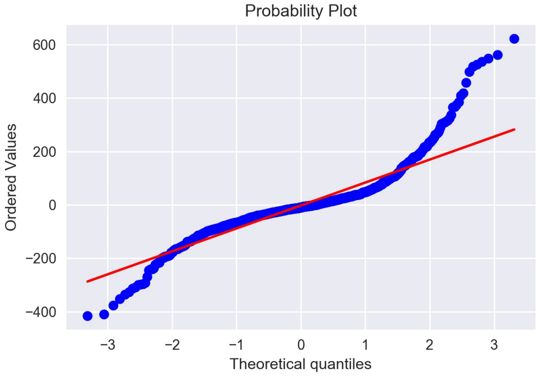

Figure 5: Q-Q Plot