Pricing Airbnb Rentals

James Gearheart, Bob Saludo,
Ryan Wallace, Daniel Zhuang

CS109a: Data Science
Fall 2016

Airbnb and Pricing

What is Airbnb?
Airbnb is a peer-to-peer online marketplace enabling people to list or rent short-term lodging in residential properties, with the cost of such accommodation set by the property owner.

What issues arise for hosts?
The decision of pricing a property on specific dates is a critical decision that every Airbnb client must address.

How can we help?
The goal of our research is to develop a predictive model that will provide the property owner with a competitive target price on specific dates for their property. This model will empower Airbnb property owners to competitively price their property and provide renters a fair market price.

Data Exploration and Cleaning

Description of Data
The data describes all properties listed on Airbnb in New York City during calendar year 2015. The *dependent variable* is daily price. Of the 54 *independent variables* available, notable examples are:

• zip code • square feet • average review score • number of beds
• number of bathrooms • latitude • longitude • property type.

The data are messy, and missing values are imputed with methods such as regression and conditional means. Categorical variables are encoded with either one-hot or quartile-based schemes.

Visualizations of Price and Predictor Variables

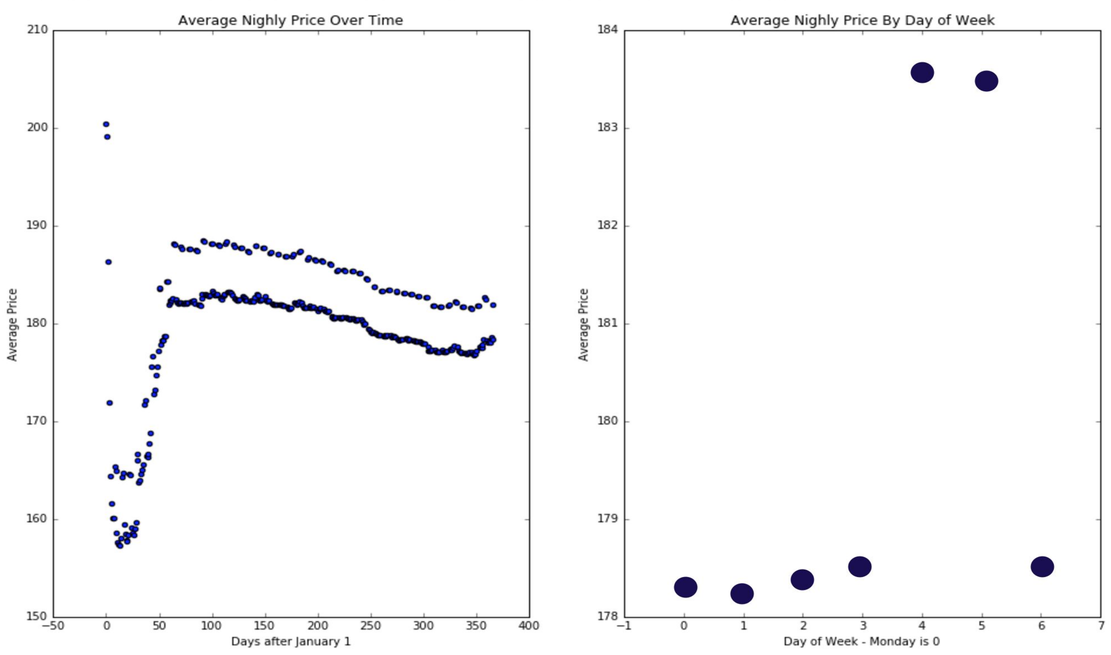

Figure 1. Average daily price against day of the year and day of the week.

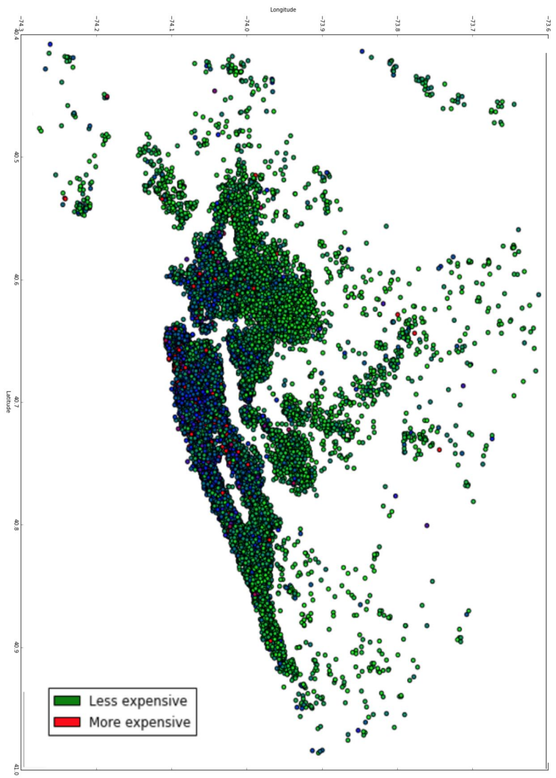

Figure 2. Heat map of daily price by geography.

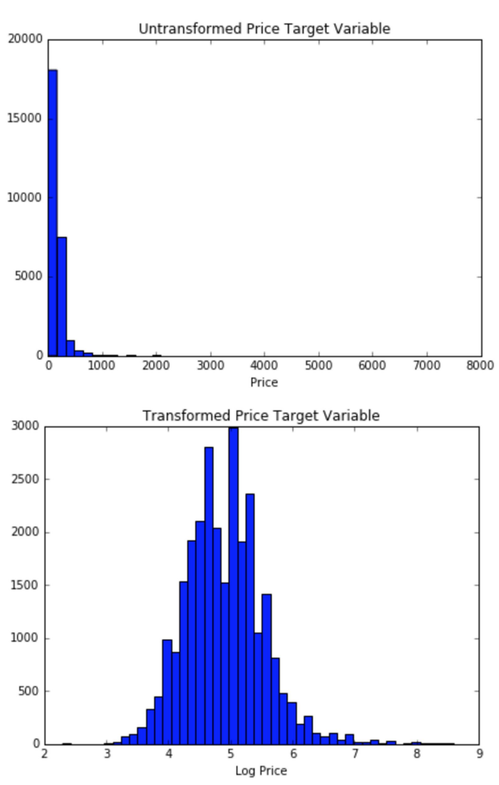

Figure 3. Histograms of daily price before and after log transformation.

Comparison of Modeling Approaches

Baseline Model
The baseline model is a linear and quadratic regression of log daily price against both continuous and categorical feature variables with any degree intuitive relationship with price. In test, the baseline achieves an R^2 of 0.63.

Ridge and Lasso Regression
Due the high number of predictors (73 after encoding of categoricals), the possibility of overfitting is a concern. In response, we fit and tune ridge and lasso regressions. We see only marginal improvement over the baseline, potentially symptomatic of cross-contamination between training and testing.

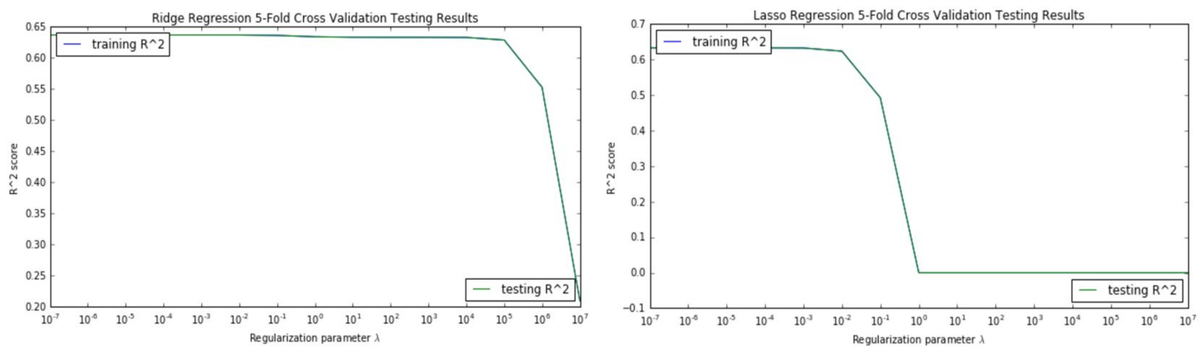

Figure 4. L1 and L2 regularization do not improve R^2 over the baseline model.

Random Forest Regression
By bootstrapping and fitting multiple regression models, random forest regression is a useful technique to minimize the bias of the predictions.

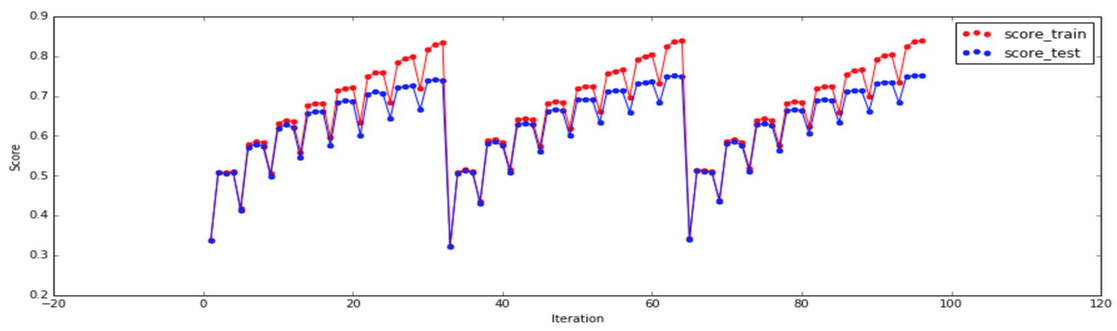

Figure 5. Tuning random forest regression over number of trees, max depth, and max features.

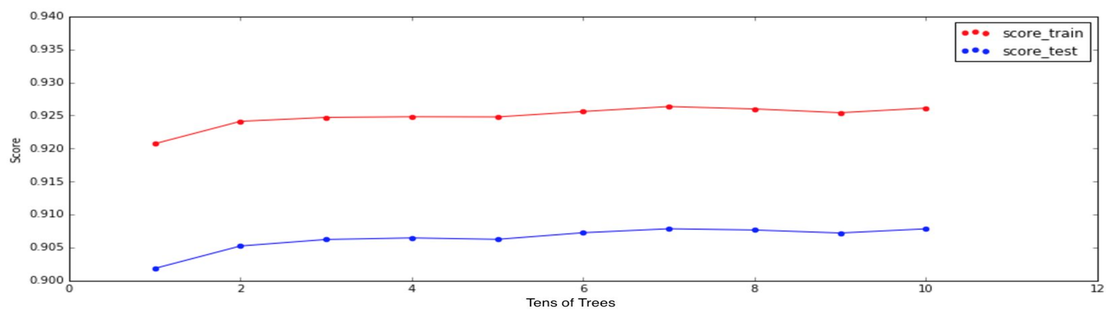

Figure 6. Additional trees confer marginal improvements in test R^2.

Tuning results in a random forest regression with 100 trees, a maximum depth of 14, and a maximum of 25 features with an R^2 of 0.92 in test data.

Prediction Intervals
In order to better understand the accuracy of our results and to provide a more useful recommendation to hosts, we build 90% prediction intervals using quantile regression forests. In test data, 80% to 90% of prediction intervals contain the true price, and have an average range of $40.

Conclusions and Future Work

It is possible to estimate the price of Airbnb rentals in New York City on a given night with a high degree of accuracy using random forest regression. Prices are highly dependent on location, features of the listing, and time of the year and week.

A potential issue with the above approach is violation of the OLS independence assumption, which may lead to inflated R^2 values. Future work may focus on addressing the independence assumption using a multilayer model, and generalizing the results to other locales by using additional data sets.

Citations and Links

Murray Cox: Inside Airbnb. www.data.beta.nyc/dataset/inside-airbnb-data/resource (2015)

Nick Amato: Predicting Airbnb Listing Prices with Scikit-Learn and Apache Spark. www.mapr.com/blog/predicting-airbnb-listing-prices-scikit-learn-and-apache-spark (2016)

Emily Tang, Kunal Sangani: Neighborhood and Price Prediction for San Francisco Airbnb Listings. (2015)