Write Up

Ezinne Nwankwo

February 2020

1 Abstract

In this paper, we seek to understand factors about a typical Airbnb listing that influence its price and popularity. We analyze a dataset with a total of 48,895 Airbnb listings in New York and 16 variables. We focus on feature engineering and fit two types of models: linear model using Bayesian Model Averaging and random forest with popularity and price as the outcome variables. Ultimately, our main finding in which all of our models agreed with was that room type and proximity to metro stations played a huge role in determining price and popularity.

2 Introduction

Airbnb has quickly grown into one of the largest online marketplace for arranging long or short term stays at homes and apartments owned by people, typically a cheaper alternative to booking hotels. Airbnb does not own any of the property on its site; instead it acts as a broker between customers and hosts and earns a commission from each booking. Hosts have full control over aspects such as pricing, how to advertise their listings, and how often their listing is available throughout the year. For this project, we are interested in modeling price and popularity of listings in order to best advise hosts on tactics that will maximize their profit and popularity on the platform. This includes (i.) identifying influential factors on the price and popularity of the listing (ii.) identifying heterogeneity across boroughs and neighborhoods, in particular which ones have the heaviest traffic and highest prices (iii.) identifying heterogeneity across listing type and (iv.) providing recommendations on listing location and names.

2.1 Data

3 Materials and Methods

We decided to implement two types simple models, Bayesian Model Averaging (BMA) in linear models and random forests (RF). For BMA, we built two

models separately for log price and log popularity using all the variables as a linear combination of predictors. For the BMA models, we exclude latitude and longitude because we don't believe that there is a linear relationship with the outcome variables. We use pretty standard parameters for these models like a Cauchy prior for the predictors (see Sensitivity Analysis section where discuss the use of different priors) and a uniform prior over the model space, which assigns equal probability to all the models. We use an MCMC algorithm with 10^{16} iterations to sample from the model space of 2^{15} models. Lastly, we approximate the marginal inclusion probabilities of the predictors by taking the p-values from p simple linear regressions. (In R code, this parameter is initprobs = "marg-eplogp"). The posterior inclusion probabilities (PIP) of each predictor was primarily what we used to determine influential factors in price and popularity of a listing.

For the RF models, we used all the predictors including longitude and latitude since this class of models does not make any linearity assumptions between predictors and outcome variable. We build these models using 1900 trees each built on subsamples of 19000 data points. Due to computational challenges, we could not go beyond this number of trees and subsamples. For the split criteria, we randomly sampled $m = \frac{p}{3} = 4$ variables as candidates for the split (Note that this is the default value for regression) and $n_{leaf} \geq 5$. We used a variable importance measure that is based on an increase in node purity, or Gini-based importance. It is is calculated based on the reduction in sum of squared errors whenever a variable is chosen to split.

We chose to build these two models because from our initial exploratory analysis, there was no clear linear trend for some of the predictors. Thus to provide further justification to our conclusions that we discuss in the following section, we also fit a non-linear model for comparison.

4 Results

4.1 Exploratory Data Analysis

4.2 Main Results

From our models, we were able to identify the top influential factors for price and popularity. For price, the BMA and RF models agree that room type and boroughs are very influential factors. BMA flagged the variable $name_host_special as being significant with the highest$

Our model was very sensitive to outliers. When we fit our models, we noticed some outliers in the residuals plot, and thus decided to remove points that were three times the standard deviation away from the mean of the outcome. After refitting, the model residual plots looked much better and had better R^2 values. Unfortunately, for the popularity model, the residual plots indicate that our models are not accounting for some relationship in the data very accurately (See residual plots below). For the BMA models, we also varied the choice in prior by testing out the g-prior for g = 1, 5, 8, 50, 100, 500, 1000. The results did

not change for the different priors. Lastly, for the RF models, we varied m and the minimum n_{leaf}. The final values that we used for those parameters did not result in a change in MSE.