page cover

Table of contents

Overview

part 1: problem identification

part 2: Data-driven analysis

1. data visualization
2. bivariate analysis between dependent variables and predictor
3. models
4. linear regression and semi-log regression
5. logistic regression
6. classification

part 3: recommended decisions

Appendices

Overview

Airbnb, founded in 2008, provides an online platform for local hosts to rent accommodations for outside travelers. In general, there are three categories of roles associated with the company: guests, hosts, and the company itself. In this project, we are going to build four different models from the perspective of guests, using two continuous dependent variables and two nominal variables. Followed by each model with various predictors, we will conduct a comprehensive analysis and determine if we should rent the host’s accommodation.

Part 1: problem identification

Given various datasets in different locations, we will focus on the area of San Francisco. As a role of guests, we can predict a price range with given preference, such as known days of stay, numbers of companions etc., all of which will be included in our dataset. Therefore, as continuous dependent variables price and log price, we will predict whether the hosts’ listing prices are above or under our budgets using linear and semi-log regression models. In addition, as nominal dependent variables, we will classify whether the hosts’ listing prices are overpriced or underpriced using logistic regression model and classification model.

Part 2: data-driven analysis

1. data visualization

(a bunch of intuition and pictures)

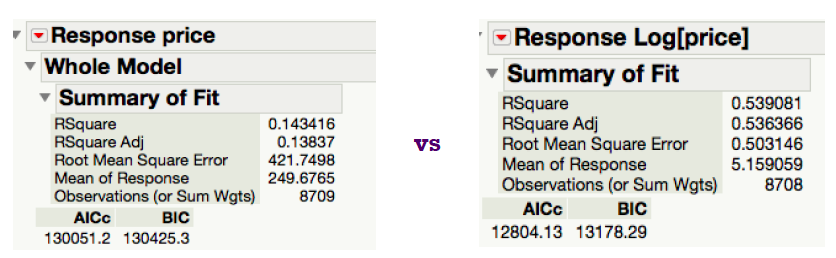
1. bivariate analysis

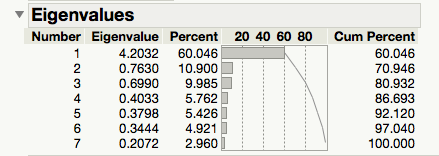
(data preprocessing, data type, a few examples will be enough)

1. models

a. linear and semi-log regressions

According to above bivariate analyses, we can prove that 22 predictors are strongly correlated with price and log price. Comparatively, we found that as continuous dependent variable, semi-log price model performs much better than the linear model.

Among the 22 predictors, 7 have similar meanings, which complicate the model. Therefore, we can apply factor analysis to see if we can simplify the model with similar predictive power. Principal Component Analysis result follows:



we can see that only one factor has an eigenvalue greater than 1, and grouping all 7 review scores and run a regression, we have following regression report:



R-square decreases and AIC/BIC increase, so in this semi-log model, factor analysis is not efficient. Even though column cannot be reduced by factor analysis, we can reduce rows by clustering analysis. There are some outliers in “price” and “security deposit”, excluding them help us reduce 174 rows, and increase R-square by 1.6%.

