Predicting Airbnb Housing Prices in London

Xiao Wan & Xiaozhou Yu & Xiaoyu Zhang
Instructor: Dr. Yifan Hu

Dec.15 2016

CONTENTS

LIST OF FIGURES

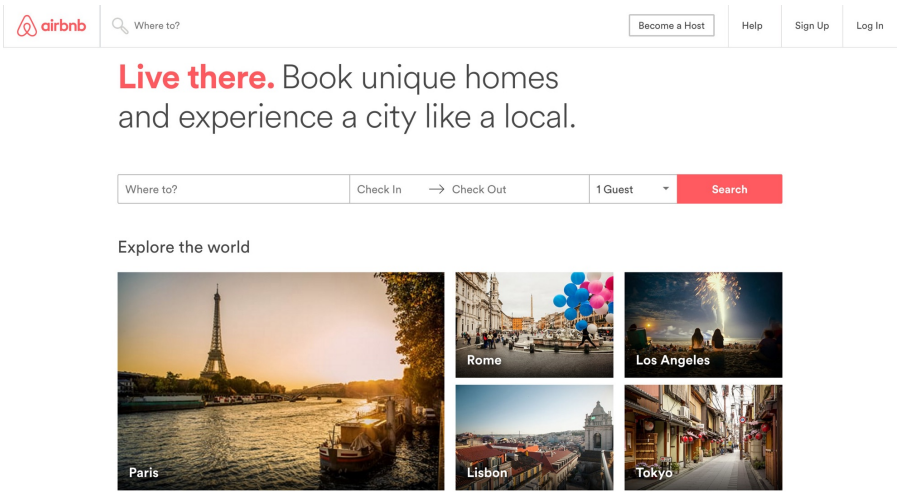


Figure 1: An example of Airbnb (from http://www.airbnb.com)

1 CHALLENGE

Our challenge is focused on making predictions that revolve around the housing price of Airbnb. Based on the Airbnb existing data, we will create a model to explore and predict the housing price of new listings in London. On one hand, unlike traditional hotels, the housing price is set by house owner in Airbnb. In fact, there are many factors can impact the price of a new list such as location, room type, and published reviews. In this case, we need to take more consideration in feature selection for a better prediction. On the other hand, London is the political, economic and cultural capital of Britain and one of the biggest markets in Airbnb. In June 2016, there are more than 42,000 records of listing in London. We need to deal with a large number of features of different types and handle missing value and outliers.

2 BACKGROUND

2.1 Introduction to Airbnb

Airbnb is an online marketplace and homestay enabling people to rent a short-time home. The cost of such accommodation is set by the house owner. Airbnb receives percentage fee from both guests and hosts in conjunction with every booking (Wikipedia-Airbnb, 2016). Because of the better located, unique place and cheaper cost(generally 30%-80% lower than available hotels), Airbnb has been a boon to travelers in all over the world. Over the past several years, Airbnb has become the market leader in the temporary accommodations industry.

2.2 Motivation and Business Implication

The housing price is one of the most significant factors considered when planning a trip, especially so for price sensitive people. People would expect a reasonable cost of the house according to their requirements via Airbnb. Thus, there exists a legitimate motivation for a prediction model that encompassed the relationships between various features and housing prices.

FOR POTENTIAL TENANTS Our model can assist travelers in choosing the best-fitting, within budget houses.

FOR HOUSE OWNERS Our model will help house owners make better pricing decisions.

FOR AIRBNB Our model will help the company to detect fraudulent listings. Specifically, it can identify transactions with the unusual prices (a big deviation from our predictions). Then the company can have a further investigation on that.

2.3 Methodology and Evaluation

METHODOLOGY The problem of predicting housing prices can be considered a regression problem since the predictive values fall within a continuous range. Through regression, we will explore the relationship between the features we have selected and the housing prices.

EVALUATION To avoid the influence of different price level, we use relative error instead of an absolute error. The results are evaluated on RMSPE–Root Mean Square Percentage Error. The RMSPE is calculated as:

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \widehat{y_i}}{y_i})^2} \tag{1}$$

y_i denotes the actual price of a new listing and $\widehat{y_i}$ denotes the corresponding prediction.

3 DATA EXPLORATION

3.1 Dataset Description

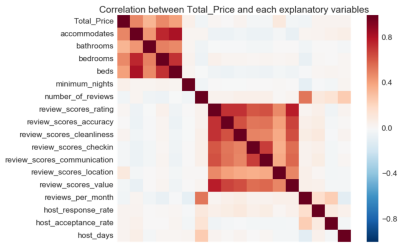
The dataset is obtained from Inside Airbnb (http://insideairbnb.com) which offers independent, non-commercial data scraped from open Airbnb listings. The original dataset contains 42,646 records and 95 variables.

3.2 Data Cleaning

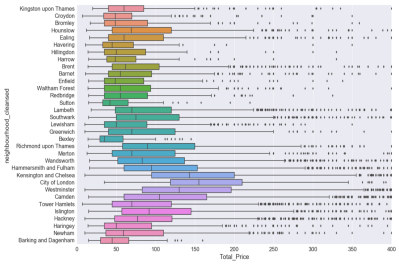
1. Remove columns that are either irrelevant or null.
2. Convert columns into desirable formats. For example, convert price to double, bedrooms to integer and host date to datatime.
3. Combine column price and cleaning fee into one column: total price. Use it as the target variable.
4. Add dummy variables for categorical variables such as room type and experience offered.
5. Remove records with unreasonable prices. Fill in null records with proper values.

After data cleaning, the dataset contains 38,541 records and 72 variables.

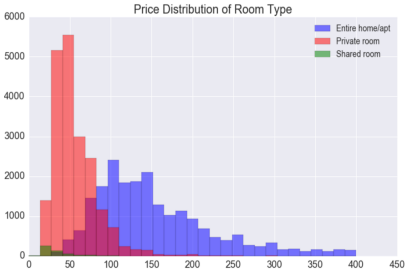
(a) Distribution of Actual Sample Prices



(b) Correlation between Price and explanatory variables



(c) Distribution of Price per Neighbors



(d) Distribution of Price per Room Type.

Figure 2: Exploratory Data Analysis

3.3 Exploratory Data Analysis

FIGURE 2-A The price ranges from 10 to 300, its distribution is positive skewed.

FIGURE 2-B Correlation between price and explanatory variables.

FIGURE 2-C The distribution of price is significantly different by neighborhood. It's natural that costs rise as the houses locate nearer to the center of the city.

FIGURE 2-D The distribution of price is significantly different by room type.

4 BASELINE MODEL

This is a simple model that predicts the housing prices based on two features - Room Type and Neighborhood.

4.1 Methodology

First, we divide all training data into different groups based on the two features mentioned above - Room Type and Neighborhood. Then we calculate the average housing prices for each group. Last, for each new house, we predict its housing price to be the average housing price of the group it belongs. Note this is a static model, houses in the same group will always get the same price prediction.

4.2 Results

Using the evaluation formula mentioned in Section 4, the baseline model gives an RMSPE of 46.8%.

5 ADVANCED MODELS

In this section, we start using machine learning algorithms such as regression and ensemble in order to improve prediction accuracy. Five models are built in total and Gradient Boosting gives the best performance in terms of RMSPE.

5.1 Model Introduction

MULTIPLE LINEAR REGRESSION Housing price as target variable, all the other features introduced in section 3 as predictors.

LASSO Add L1 norm regularization to Multiple Linear Regression model to reduce over-fitting.

BAGGING Fits base learners on boostrapped training data, averaging each individual predictions to reach a final prediction.

RANDOM FOREST Build fully-developed decision trees on boostraped training data, averaging each individual prediction to improve the predictive accuracy and control over-fitting.

GRADIENT BOOSTING Again, use fully-developed decision trees as base learners. This time build trees sequentially: each tree is fitted on the residual given by the previous prediction.

5.2 Model Evaluation and Selection

Figure 3 shows each model's RMSPE on the training dataset through 10-fold cross-validation. In general, low RMSPE indicate better model performance. Gradient boosting (Green Line) gives the lowest RMSPE on training data. As a result, we use this as our final model and try parameter tuning on it to improve its performance.

5.3 Parameter Tuning on XGBoost Model

In this subsection, we start parameter tuinng on XGBoost Model. Parameters of interest are Learning Rate & Number of Estimators, Maximum Depth, Gamma, Subsample & Colsample, Reg alpha & Reg lambda. Each parameter is tuned and validated on a 10-fold dataset.

MAXIMUM DEPTH Maximum tree depth for base learners. A larger value indicates a more complicated tree. The ideal value is 10.

GAMMA Minimum loss reduction required to make a further partition on a leaf node of the tree. The ideal value is 0.3.

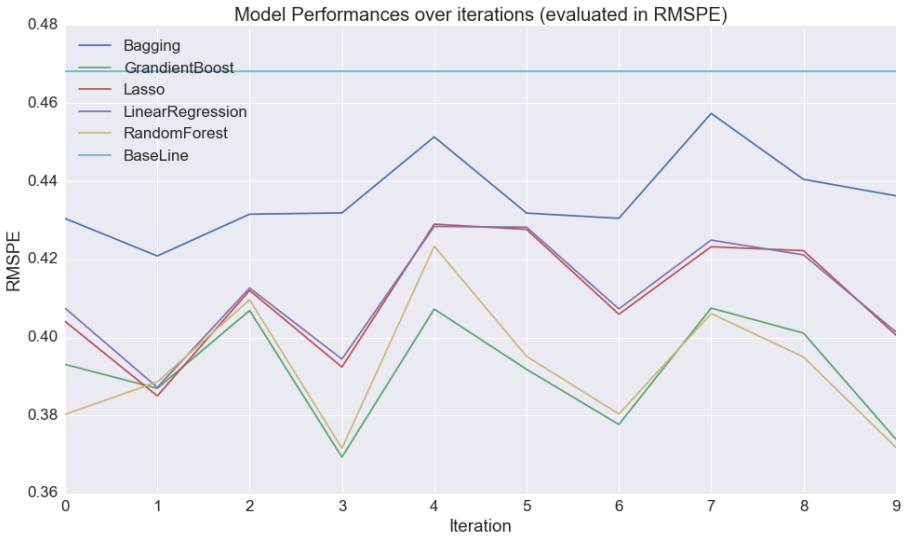


Figure 3: Model performance on training data

SUBSAMPLE & COLSAMPLE Subsample ratio of the training instance and subsample ratio of columns when constructing each tree. Ideal values are 0.9 and 0.8 respectively.

REG ALPHA & REG LAMBDA L1 and L2 regularization term on weights. Ideal values are 0.5 and 6 respectively.

LEARNING RATE & NUMBER OF ESTIMATORS Boosting learning rate and Number of boosted trees to fit. Use a small learning rate and a large number of estimators for a robust model.

6 RESULTS AND DISCUSSION

6.1 Error Analysis

Since XGBoost algorithm gives the best performance according to cross-validation, we build a final model with it using the parameters given above. The RMSPE on testing data is 36.29%.

Next, we further investigate the performance of XGBoost model through error analysis. The error rate is given as

$$\text{ErrorRate} = \frac{\text{PredictedPrice} - \text{ActualPrice}}{\text{ActualPrice}} \tag{2}$$

If the error rate of a certain sample is positive, it means predicted price of this house is higher than actual one. To see the distribution of deviation, we draft a picture of error rate vs actual price (Figure 4: Error distribution). The actual price data has been binned into 10 price groups to generate a histogram.

According to the histogram, predicted price values scattered in a reasonable range around actual price between 59 and 150, but deviated from lower and higher actual price in positive and negative direction respectively.

Figure 4: Error Analysis of XGBoost Model

Therefore, we can conclude that our model is a bit conservative. It can make good prediction performance for houses at middle prices, but over-predicts houses at prices lower than 59, and under-predicts houses at prices higher than 150.

Through ANOVA and importance assessment of features based on information gain, current model is under-fitting but not over-fitting. Therefore, one possible error source is the lack of potential key features. Some important information is hidden in variables dropped from original data, such as description, amenities and reviews. These variables can influence rental price, but can be utilized directly only after manual interpretation and marking or text mining process. Other important information is hidden in missing features, such as house age, neighborhood environment, transportation accessibility. Last but not least, irregularity of pricing exists in homestay marketplace. House owners' decision of actual prices can be arbitrary and subjective, and it increases the difficulty of prediction.

6.2 Possible Improvements

1. Improve the complexity of the model by utilizing more explanatory variables such as amenities (including kitchen, parking plot, internet, TV, heating, etc.), neighborhood environment (including crime rate, distance to downtown, grocery or restaurant, etc).
2. Alter objective and evaluation functions used by XGBoost algorithms.
3. Use techniques such as feature engineering.

REFERENCES

[1] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[J]. *arXiv preprint arXiv:1603.02754*, 2016.

[2] Ng A, Deisenroth M. Machine Learning for a London Housing Price Prediction Mobile Application[J]. 2015.

[3] Brennan, Morgan (16 Sep 2011). "The Most Amazing And Absurd Places For Rent". Forbes. Retrieved 13 December 2012.

[4] Aarshay Jain (1 Mar 2016). "Complete Guide to Parameter Tuning in XGBoost".