Using Machine Learning to Predict Airbnb Rental Price in New York City

Author: Tuong D. Vu
Supervisor: Prof. Marco Bee

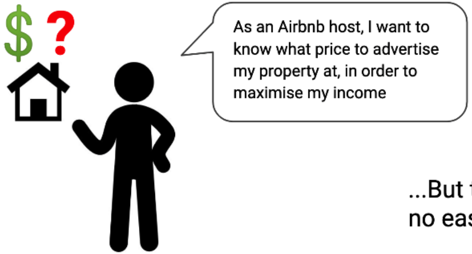
Università degli Studi di Trento
Department of Economics and Management

Master of Science in Economics
Master Thesis

29 January 2021

Agenda

Motivation



...But there is currently no easy way to do this

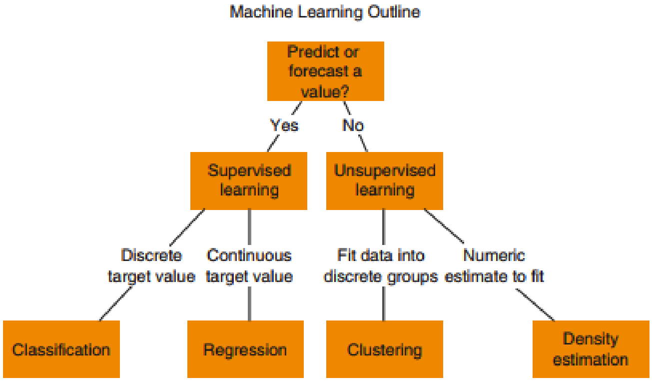
Research Questions

- Which models perform best to predict Airbnb listing price in New York City?
- Which features of an Airbnb listing are most important in predicting the price?

Methodologies

- Data Collection
- Data Preprocessing
- Exploratory Data Analysis
- Model Fitting

Machine Learning



Problem Formalization

The goal : approximate a target function `f` for the output variable rental price (Y) based on a set of predictors such as `bathrooms`, `accomodates`... The relationship between `price` (`Y`) and its predictors $X = (X_1, X_2, ..., X_p)$:

$$Y = f(X) + \epsilon \quad (1)$$

Then, the rental price of a listing can be predicted by:

$$\hat{Y} = \hat{f}(X) \quad (2)$$

Quantitative Measures of Performance

- We use **mean squared error** to characterize a model's predictive capabilities:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \quad (3)$$

- Best models gives the lowest **test** MSEs instead of the lowest training MSEs.

Quantitative Measures of Performance

- We use the **coefficient of determination** (R^2) to measure the proportion of the information in the data explained by the model:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- An R^2 value of 0.8 means that the model can explain 80 percent of the outcome's variation. An R^2 of 1 indicates that the regression predictions perfectly fit the data.

Variance-Bias Tradeoff

The expected test MSE, for a given value x_0, can be broken down into three parts as followed (James et al., 2013):

$$E(y_0 - \hat{f}(x_0))^2 = \mathrm{Var}(\epsilon) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\hat{f}(x_0)) \quad (4)$$

- $\mathrm{Var}(\epsilon)$: the variance irreducible error term.
- $[\mathrm{Bias}(\hat{f}(x_0))]^2$: model's squared bias,i.e how close the target function f to the real relationship between the predictors and and outcome.
- $\mathrm{Var}(\hat{f}(x_0))$: how much the value of the target function f will vary if we use different training data.

Variance-Bias Tradeoff

- Equation 4 means that, minimizing the test MSE $=$ reducing the combination of bias and variance.
- However, it's impossible to reducing *both*:
  - Overly simple model $\Rightarrow$ low variance, but high bias.
  - Complicated model $\Rightarrow$ low bias, but high variance.
- The good strategy: try various models with different variance-bias tradeoff levels to decide which is the best model, i.e the one with lowest test MSE.

Models

1. Linear Regression
2. Ridge Regresion
3. Lasso Regresion
4. XGBoost

Linear Regression

- We can specify the hedonic price function of Airbnb listings as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{5}$$

- Advantages: simple, intuitive, has a theoretical justifies (Hedonic pricing theory (Rosen, 1974))
- Disadvantages: tend to overfit data (Harrell Jr, 2015)

Ridge Regression

- Ridge coefficient estimates minimize:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \quad (6)$$

- The tuning parameter λ can by found by using a cross-validation technique.
- Disadvantages: Ridge regression does not perform *feature selection*,i.e it does not set any of the parameter estimates equal to 0

Lasso Regresion

- Least Absolute Shrinkage and Selection Operator (LASSO) coefficients, $\hat{\beta}^L$, minimize the quantity:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p} \mid \beta_j \mid = RSS + \lambda\sum_{j=1}^{p} \mid \beta_j \mid \quad (7)$$

- The tuning parameter λ in can by found by using a cross-validation technique.
- Advantages: Simulataneously reduce the model's variance and conduct feature selection.

XGBoost

- short for e**X**treme **G**radient **Boost**ing package.
- An efficient and scalable implementation of gradient boosting framework by Friedman, 2001
- Advantages: provide state-of-the-art results for diverse problems, including regression, classification and ranking.
- Disadvantages: Interpretability is very hard to achieve.

Data

- Data source: a dataset with 50,599 Airbnb listings in NYC is available from "Inside Airbnb", 2019 website.

Table 1: Summary Statistics

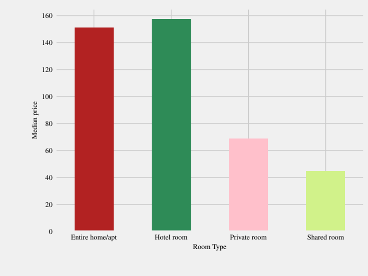
	mean	std
price	138.085	118.185
host_is_superhost	0.234	0.424
host_listings_count	7.775	54.391
host_identity_verified	0.486	0.500
accommodates	2.906	1.911
bathrooms	1.140	0.421
security_deposit	172.822	406.817
cleaning_fee	54.161	54.671
...	...	...
pets_allowed	0.165	0.371
private_entrance	0.210	0.407
self_check_in	0.260	0.438

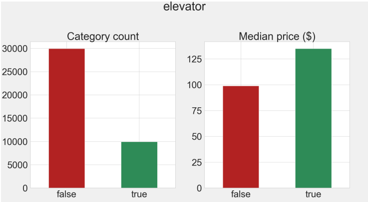
Data Preprocessing

- Data Filtering: Eliminate listings consider "inactive", which has not been reviewed.
- Data Cleaning:
  - Dealing with Missing data by either by dropping features with majority of null values or by data imputation.
- Data Transformation:
  - Z-score Normalization
  - Log Transformation to remove skewness
  - One Hot Encoding Categorical Features
  - Data Binning

Exploratory Data Analysis

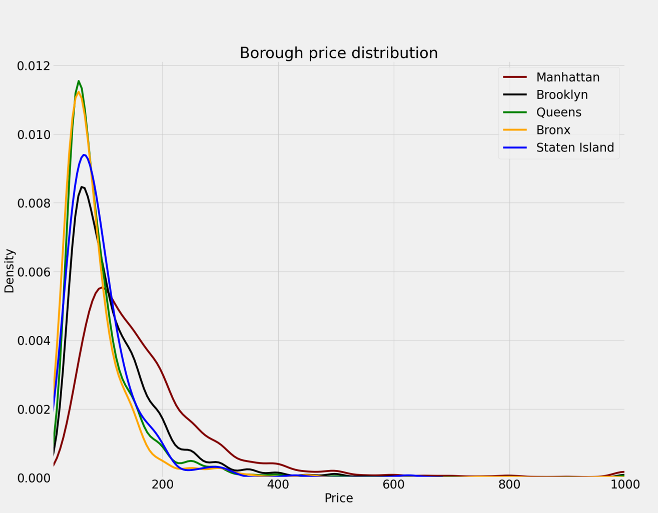
We use graphical to get a sense/glimpse of potential effect of each feature on the price





Exploratory Data Analysis

Or, we can observe a price difference between the NYC boroughs:



Best Model?

Table 2: Results

ML Algorithm	Training MSE	Test MSE	Training R^2	Test R^2
Linear Regresion	0.1291	8.5E21	0.7019	≈ 0
Ridge Regression	0.1291	0.138	0.7019	0.6857
Lasso Regression	0.1351	0.1441	0.688	0.6718
XGboost	0.0798	0.1173	0.8157	0.7328

- Gradient boosting with all features (XGBoost) performs the best among all models followed by Ridge regression.
- Linear Regression model suffers from overfitting.
- While Lasso's performance is not as good as Ridge Regression and XGBoost, Lasso performs feature selection. The final model contains only 153 variables while eliminating 125 variables.

Most Important Features?

Table 3: XGBoost Top 20 Important Features

Feature	Weight
room_type_Entire home/apt	0.336396
bathrooms	0.032001
neighbourhood_Midtown	0.025008
neighbourhood_Hell's Kitchen	0.018545
neighbourhood_East Village	0.015763
property_type_Other	0.015168
neighbourhood_Bedford-Stuyvesant	0.014314
neighbourhood_West Village	0.014031
neighbourhood_Chelsea	0.013612
neighbourhood_Lower East Side	0.011874
neighbourhood_Bushwick	0.011854
neighbourhood_Upper West Side	0.011682
neighbourhood_Washington Heights	0.011659
neighbourhood_SoHo	0.011582
room_type_Shared room	0.011304
neighbourhood_Greenwich Village	0.010347
room_type_Hotel room	0.009697
neighbourhood_Theater District	0.008575
neighbourhood_Williamsburg	0.008490
neighbourhood_Crown Heights	0.007979

Most Important Features?

- The most critical feature is whether the type of listing is an entire home or not. The second most important feature is the number of bathrooms.
- Location features play an essential role in predicting price.

Future Works

- Experiment with the data with Neural Network.
- Find a way to include listing's photo quality as a predictor.
- Incorporate customer reviews feature through sentiment analysis.

References I

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Harrell Jr, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Inside Airbnb [[Online; accessed 04-Dec-2019]]. (2019).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of political economy*, *82*(1), 34–55.

Thank you for your careful attention!

Questions and Answers