Modeling Price and Popularity of AirBnB listings in New-York

Melody Jiang Raphael Morsomme Ezinne Nwankwo

Case Study 2 - Stat 723

February 1, 2020

Overview

Philosophy

Dirty data limit us to simple models.
Very dirty data:
- "last day of the month" type of data: price, frequency of booking, constant over time?
- improbable values: minimum length $> 1,000$.
- even the dependent variables are shaky: $popularity = numberbooking/availab$?.
- ideally, *tidy* data (Wickham, 2009) with one row per booking
Focus on data cleaning and feature engineering over modeling.
EDA will motivate the creation of new variables and the cleaning of the data.

EDA - A City of Two Tales

¡Figure (histogram/density) showing short vs. long stay¿

EDA - Are you available?

¡Figure (histogram/density) showing distribution of most recent stay¿

EDA - Attractions

¡Map showing effect of an attraction on price¿

Data Cleaning

Drawing on the EDA, focus on active listings for short stay:
Keep listings with
(i) last review ¡ 1 year old [lose $15,000$]
(ii) minimum number days < 30 (short type of stay) [lose XXX]

Feature Engineering - Proximity

EDA shows impact of attraction on price. This suggests the creation of a variable measuring the proximity of a listing to attractions. The proximity variable is defined as the average proximity of the listing to the attractions

$$proximity(X) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{dist(X, attraction_i)}$$

where

$$dist(x, y) = \mid latitude_x - latitutde_y \mid + \mid longitude_x - longitude_y \mid .$$

is the Manhattan distance.
Similarly, we compute the proximity to the closest metro station.

Feature Engineering - Textual Data

Sentiment analysis of listing name
- "documents" too short for topic modeling - Afinn dictionary (gradual rating)

$$Sentiment(X) = \frac{1}{n}\sum_{i=1}^{n} dictionary(x_i)$$

where $Afinn(x) \in \{-5, -4, \ldots, 5\}$.
Origin of host name
- use name frequency as a proxy

Models

Linear regression model $Y = X\beta$ where X consists of:
proximity metro, proximity attraction, host name frequency, listing name sentiment, [newly created variables]
X1, X2, X3 [regular variable]
Random forest (n $= 1,500$, m $= 2/3$)
BMA (setting)

Influential Factors

Variable Importance metric from the random forest ($n = 1,500$, $m = 2/3$)
Posterior Inclusion Probability from the BMA

Sensitivity Analysis

Vary the setting of the RF: different levels of pruning, different values for m.
Vary the priors in the BMA: prior1, prior2, prior3

Results - Influential Factors

¡Table of variable importance¿
¡Table of PIP¿

Results - Q3

¡Figure for Q3¿

References

Whickam, H.
Tidy Data
Journal, month year