Kevil Khadka
Dr. Darrin Weber
STAT-493
25-Feb-2020

Mining Airbnb dataset using Machine Learning algorithms and R concepts

Introduction

Airbnb is a vastly growing and a trusted community online marketplace that connects people that owned real estate properties and tourists interested to rent short-term or long-term lodging. With the growth of internet-facilities, online social networking, mobile technology, location-based services, Airbnb is gaining a huge advantage to the peer-to-peer economy growth. Airbnb site offers customers variety of accommodation regarding locations, experiences, price to choose from. Since the company started in 2008, it has now over 150 million total users, more than 2 million of people staying in an Airbnb per night, 6 million of global Airbnb listings worldwide, and 35-billion-dollar valuation based on recent stock sale (Airbnb-statistics). The global compound growth rate of Airbnb is increased to “153%” since 2009. Looking at this significant development, Airbnb’s competitors would easily replicate the same business model which makes long-term growth further challenging. As we see very similar structures (low fees, easy and fast accessible service) within the industries, there is limited differentiations found among the companies’ core business models which are basically the number and diversity of listings, performance on the platform, customer and worker relationship. This project focuses on those differentiations and how machine learning algorithms could be helpful to predict better model.

Problem

1. Dealing with missing value?
2. What can we learn about different hosts and areas?
3. What can we learn from predictions? (for example: locations, prices, reviews, etc.)
4. Is there any noticeable difference of traffic among different areas and what could be the reason for it?
5. Predicting price of Airbnb rental room

6. How Airbnb is differed to hotel industry?
7. What aspects of the rental experience do people like and what aspects do they dislike?
8. Based on reviews, this project focuses on a text analysis to compare between Airbnb and hotel.

Dataset Source

This project will use two main datasets: hotels dataset and Airbnb dataset. The hotels' datasets are obtained by data scraping from the website like TripAdvisor, Expedia. And it is easy to access the Airbnb dataset from their website, insideairbnb.com.

Competition

There are many exploratory data analysis projects done in Airbnb dataset. Various projects are based on machine learning algorithms and deep learning to predict Airbnb price for properties in main locations like New York, London, Berlin etc. Despite that multiple projects were carried out on predicating the listing prices, none of them has been performed the comparison between hotel and Airbnb. This project would also explore to experiment and compare different machine learning algorithms in price prediction.

Methods

1. Data Exploration

- Structure and features of dataset, exploring missing data

2. Data Visualization

- Price vs Location, Seasons, Neighborhood area, availability rate, room type
- Demand for Airbnb rentals over the year
- Growth of Airbnb in higher traffic cities vs low traffic cities
- Home sharing vs commercial use; and its earning revenue
- Airbnb VS Hotel industry

3. Machine Learning

- Split data into train/test sets
- Predicting price
- Linear Regression model
- K-means clustering
- Random Forest (Maybe Decision tree)

Results

The main aim of this project is to create an analysis which could help people to make more informative decision when they travel across the world and stay with Airbnb. We would also compare which model provides the best results by comparing their MSE and R-squared.

Resources

1. Airbnb-dataset. http://insideairbnb.com/get-the-data.html
2. Airbnb-statistics. https://ipropertymanagement.com/research/airbnb-statistics
3. Hotel Data and Reviews.
   https://data.world/datafiniti/hotel-reviews/workspace/file?filename=7282_1.csv