Project 2 Proposal
Regression
AirBnB Arts & Crafts Class Experiences

Kristen Tokunaga

Background

Airbnb Experiences offers a hosting platform for locals to advertise and schedule activities that introduce travelers to their city, craft, cause, or culture. Visitors can use this system to find, schedule, and rate their experiences.
https://www.airbnb.com/help/article/1581/what-are-airbnb-experiences
https://fortune.com/2017/10/23/airbnb-ceo-experiences-new-york/

Problem

The floral industry is facing disruption—through Costco, grocery stories, online deliveries, DIY culture, etc. A local flower shop owner wishes to diversify her business to reinvigorate customers' appreciation for the art of flower arranging and share her gift in a new way with those who love to learn crafts. In consideration of the market demand for flower arranging classes and features to include to make her class more popular, I want to analyze what are the most important characteristics of an Airbnb Experience listing, and how do they influence average rating?

Data

Webscraped Airbnb Experience listings that are returned in a search of "Crafts":

Variable	Type	Description	Use for Model?
Location	String/Object	→ turn into zip? Region? Country?	If use, may need to limit to avoid having too many variables
Language(s)	String/Object	→ turn into # lang?	
Duration of experience	Float		
Cost	Int		
Number of people per experience	Int		
Number of reviews	Int		Can't use if pre-listing (not useful)
Includes -drinks	N/Y (0/1)		
Includes -food	N/Y (0/1)		
Avg past ratings?	Float		Business can't influence Existing listings vs. pre-listing Can't use if pre-listing (not useful)
Host - # guests hosted	Int		

Number of photos on listing	Int		
Number of guest photos	Int		Can’t use if pre-listing (not useful)
Number of tags	Int		
~~How long been listed~~			

Known Unknowns/Barriers

- Timing of listings
    - Note: duration of listing may not be meaningful, but time of day/week/month/year that the host offers the experience may influence guest interest
- Frequency of classes
    - Is this experience available all the time? Or limited based on host availability?
- Host’s descriptions of the experience may have a significant effect on whether guests sign up for the experience
- Average review ratings are the only available indicator of success/popularity of an experience—number of bookings, revenue, and profit are not publicly available
- Limitations of variables that can be used in this analysis, based on features that can be influenced pre- vs. post-listing (ex. guest photos, number of reviews, average of past ratings)
    - Also important to note that the host may not have direct influence over these variables and so even if these variables had high impact, they would not lead to actionable recommendations for the hosts to increase their reviews
- The Airbnb Experiences website may not display all information consistently (e.g., may omit sections that are N/A) and I will need a way to gracefully deal with those observations/variables

MVP

Linear regression model using variables that are “clean” (i.e., consistently represented across all observations in my sample, such as duration and price).

Ex. add utilities into house (e.g., adding microwave)

Who's market – posting vs. renting

scrapy

Backstory:

Using information we scrape from the web, build linear regression models from which we can learn about movies, sports, or categories.

Data:

- **acquisition**: web scraping
- **storage**: flat files
- **sources**: (as listed below or any other publicly available information)

- movie: boxofficemojo.com, imdb.com
- sports: sports-reference.com

Skills:

- basics of the web (requests, HTML, CSS, JavaScript)
- web scraping
- `numpy` and `pandas`
- `statsmodels`, `scikit-learn`

Analysis:

- linear regression is required, other regression methods are optional
- We recommend at least 1000 rows and 10 features. Make sure not to have too many categorical features.

Deliverable/communication:

- organized project repository
- slide presentation
- visual and oral communication in presentations

- write-up of process and results
- 4 minute presentations
- Project Logistics

Design:

- iterative design process
- "MVP"s and building outward
- stand-ups/scrums (1 minute progress updates to the class)

More information:

We'll learn about web scraping using two popular tools - BeautifulSoup and Selenium. You must know the very basics of HTML. We can also evolve the way we use Jupyter notebooks; during this project, we begin to use the notebook as a development scratchpad, where we test things out through interactive scripting, but then solidify our work in python modules with reusable functions and classes.

We'll practice using linear regression. We'll have a first taste of feature selection, this time based on our intuition and some trial and error, and we'll build and refine our models.

This project will give you the freedom to challenge yourself, no matter your skill level. Find your boundaries and push them a little further. We are very excited to see what you will learn and do for Project Luther!