STAT222 Individual Write-Up

Predicting Income for AirBnB Listings

Ziwei Liu

Abstract

Heuristic project aiming to help airbnb communities grow by digging up factors with significant impact on listing income and building a well-performing, generalizable predictive model. The group consists of four individuals including Yiheng He, Yizhang Lin, Ziwei Liu (the author), and Zihan Ye. The idea of choosing this project was out of consideration for a lack of expertise in specific scientific areas among the group as well as a shared interest in market profits. After the expectation of each member was articulated, the choice for the topic was unanimous. This is meant to be a mid-semester report synthesizing all of our efforts so far, with my individual contribution more discussed in details.

Contents

1 Problem Description

Likewise to what we stated in our first presentation, in this project, we are trying to answer the question that how much money one can expect to make on a yearly basis, given his or her listings? Due to the high volume of properties one can observe for each listings on airbnb nowadays, plus constraints such as location and budget, this question is really non-trivial. The bedrock of our analysis is the following hypothesis:

$$H: \textit{There exists a subset of features that we can use to make a good model.}$$

Here by 'good' we meant excellent performance on the data set for which we would use for training, and satisfying performance on other data sets that we may use for testing. We also found state-of-the-art or widely circulated articles such as Phillips et al. (2017) and Martinez et al. (2017). to cross validate our work, so that we don't risk going blindly or moving towards directions of whimsy. Of course, not all of the papers provided a comprehensive review of the problem - some of them were mainly focusing on a subset of features such as the reviews - at the end we would hope to justify our work by either: 1. finding more literatures to compare; 2. taking tests on more data sets to see how our results can be generalized.

With that said, our end goal is to predict income. We would certainly hope to have income information of every listings on airbnb, but oftentimes such information is unavailable due to privacy concerns. We thus agreed on one naive yet very intuitive way of defining income: $occupancy \times price$. That is, if we can manage to establish relationship between variables and average occupancy/price in a year, we would be in good shape for estimating the income.

In summary, we aim to explore the following dependencies:

$$\begin{aligned} price &\sim f(x_1 + ... + x_l + z_1 + ... + z_k), l << p \\ occupancy &\sim g(x_1 + ... + x_i + z_1 + ... + z_j), i << p \\ income &= price \times occupancy \end{aligned}$$

where x's are given features; z's are features created on top of the givens such as sentiment scores, etc.

2 Data Description

Available at insideairbnb.com, in total of seven files are provided for each city: calendar.csv, listings.csv, listings_full.csv, neighbourhoods.csv, neighbourhoods.geojson, reviews.csv, reviews_full.csv. Cities worldwide can typically be found on this website. For our purpose, we chose that of New York city as our training data, as it contained enough data points to make our analysis generalizable. The namings of each file pretty much reflect their contents, where listings and reviews are two data sets that would support the main body of our analysis. Calendar is also an important data set that records time-based information about listings prices, from which time series analysis could be drawn. Finally the neighbourhoods data track the geographic information of listings, i.e. which county the surrounding neighbourhood of a listing belongs to. It may be useful as a categorical variable or later as we attempt to make some geographical visualizations.

The focus of this report will be on listings and review. To begin with, listings.csv is a short, abbreviated version of listings_full.csv. Its columns are:

name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365.

This is a clean, nice formatted data with some basic descriptive features and the price information that requires little to none data cleaning. Given it was at the EDA stage, we did not do anything further than loading and cleaning, but later some simple yet heuristic model fitting such as linear regression or logistic regression would be performed on this short data before they are applied to the full data set.

Listings_full.csv is a more comprehensive version of listings that stores over 50000 observations with more than 100 variables, of which 42 are quantitative and 64 are categorical. It contains super illustrative features such as list of amenities, booking/cancellation policy, and text information provided by the hosts. There are of course missing data given its size, here I only summarize the actions we take for the various missing values, as these steps were not primarily carried out by me. More logics and technicality are to be found in reports by my teammates.

Type of missing value	Treatment
Name (String)	Ignore for now: chances are they serve no predictive role
Numerical	Replace with zero or mean/median
Categorical	Create a new level: 'na'
Text description	As a new variable, record if there is one available on a 0-1 basis

Last but definitely not least, reviews.csv and reviews_full.csv contain comments left by customers who stayed at the various properties.

	listing_id	id	date	reviewer_id	reviewer_name	comments
0	2595	17857	2009-11-21	50679	Jean	Notre séjour de trois nuits.\r\nNous avons app...
1	2595	19176	2009-12-05	53267	Cate	Great experience.
2	2595	19760	2009-12-10	38960	Anita	I've stayed with my friend at the Midtown Cast...

Using a series of joins between tables enabled by the Python pandas package, the above table was obtained. The most informative column is apparently `comments`, which contains the actual review corpus. The language composition, however, was not pure English thanks to our large sample size. For simplicity of analysis, we henceforth decided to focus only on English reviews. Comments that utilized emoji were also not considered. As per the strategy for handling missing data, new indicative binary variable was created to show wether there is a comment or not. The main cleaning processes in this part were conducted by the author using pandas and regular expressions. They are also key for subsequent analyses, as will be reflected in the following section.

The above sums up the team's efforts on data. As hinted previously, we did not believe all of the variables

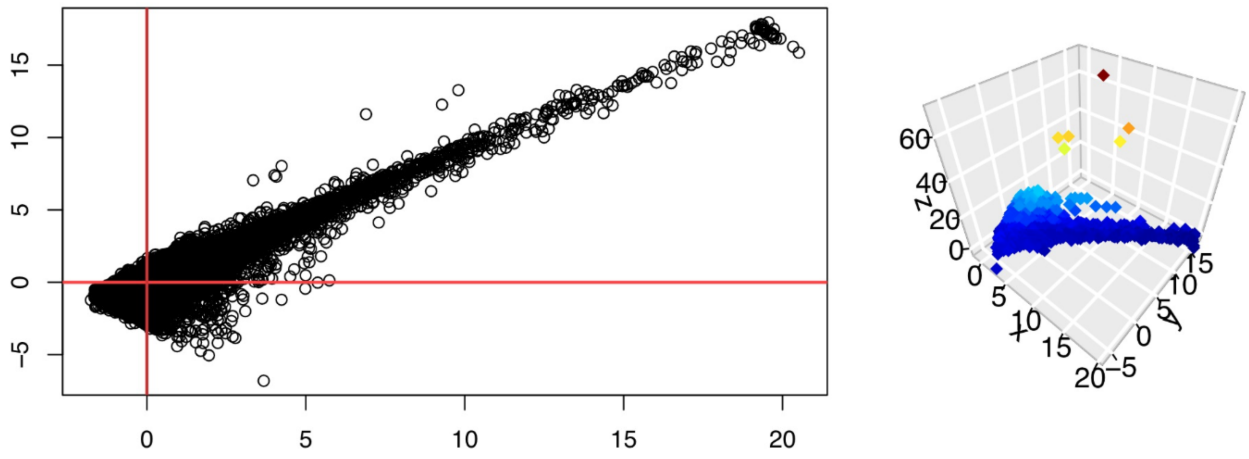
would end up contributing to our goal. The next section will discuss ways to derive insights from the data as well as some simple approches to address its high dimensionality.

3 Methods

As part of EDA, we would like to pull information that could pertain to our goal from the data. In awe of the curse of dimensionality, we wanted to see how we could conduct some simple variable selections. This section introduces three approaches that were directed accomplished by the author or correlated to his effort. The first two were based on the listings data set, whereas the third part was a detour from the EDA that would be proven vital for subsequent work.

3.1 Principal Component Analysis

Developed by Hotelling (1933), principal component analysis (or PCA) is a powerful, unsupervised procedure to project some data matrix $X_{n\times p}$ onto lower dimensions, while preserving its internal variability. If we specify an unit vector $v \in R^p$, then Xv is exactly the projected version of X. It turned out if we could find a collection of orthogonal v's that can maximize the variance of the projected matrix $\frac{1}{n}v^T X^T Xv$, then we obtained the principal components or PCs, each acting like a "hologram" of the original X matrix. We might plot in 2D or even 3D to see the interactions between the PCs to partially understand any patterns in the data. Also since we had the PCs, it would be very straightforward to pass them to models such as linear regression, but meanwhile interpretability would be sacrificed, i.e. we won't be able to easily know how each variable was leveraging the significance of the covariates.



3.2 LASSO regression analysis

Alternatively, LASSO regression can also shrink model size. According to Hastie, Tibshirani, and Wainwright (2015), it will generate sparsity in the linear space of coefficients, and variables picked through a series of bootstrap-style cross validations are deemed as important. On remembering such a powerful ad hoc trick, the author discussed with Zihan Ye, the latter did the actual analysis.

3.3 Word Cloud

One of the underlying threads in this project is to incorporate Natural Language Processing to extract sentiments of the reviews and to potentially use that as a predictor for income. The intuition was in fact both from the Martinez et al. paper and the author's personal experiences. To have that accomplished, one would need to first gain an understanding of the data. Zihan Ye kindly assisted me in this direction as he pulled a series of summary statistics about the comments text such as length, number of capitalized words, number of exclamation marks, etc. A word cloud was then made to show more frequently appeared clauses as bigger blocks. While it might not be the most rigorous practice, it followed the widely reknown philosophy "a picture is worth a thousand words" in modern industry. A traditional histogram by conparison, was not as expressive when it comes to the visual effect of the frequency for specific words, and it could subconsciously forge an impression of the "distribution" of words, which we by no means attempted to show or prove.

4 Results

4.1 PCA

Early we argued about the variance preserving capability of PCA. However, as we pulled more information from standard libraries api such as proportions of variance explained by each PC, we realized that we were over-optimistic about the performance of PCA on our data set, as we discovered even the first few PCs could account for only 10% - 20% of the total variance, whereas we would normally expect figures like 40% or 50%. The cause could be attributed to two main reasons: First, we actually only considered all numerical features to pass to the PCA, because handling multiple categorical variable was like the Achilles's heel for this procedure. On top of that, the neglected quantitiative variables could happen to be driving the variance of the numerical variables, hence the below average PC performance we have seen. It was also due to this problem that the author thought of running LASSO regression analysis.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.1940950 2.0363446 1.8406727 1.66258287 1.40803775
## Proportion of Variance 0.1660018 0.1429896 0.1168302 0.09531661 0.06836449
## Cumulative Proportion  0.1660018 0.3089915 0.4258217 0.52113827 0.58950277
```

4.2 LASSO

The LASSO approache made some fairly sensible suggestions of features to be included, such as number of bedrooms, length of reviews, etc. It also kept seemingly irrelevant feature like host_id, a result that ought to be justified by the team.

4.3 Word Cloud

The word cloud revealed that most people would tend to leave positive comments such as "great location", "great host". It also showed neutral, objective descriptions including "New York", "Subway Stations". Negative comments are hard to be found in the graph, but as we were certain of their presence by previous data cleaning, we would find ways to account for them later when it comes to sentiment analysis.

5 Conclusions

In summary, as we moved along the EDA procedures and took a few glimpses into model fitting, we built up confidence towards meeting our initial expectations. We gain intuitions from the data, and LASSO regression gave us a baseline for model fitting. References that could guide our next steps were also found on websites. We have been blessed with a good start, now the rest of this project is expected to be spent on implementing the underlying ideas for most of the time.

As we look ahead, more data cleaning and study for their relationship need to be done before benchmarking additional models. We also have to conduct in-depth analysis on suitability of machine learning algorithms that are out there for our mission, including their interpretability, estimated cost for training, linearity versus nonlinearity, etc. It is also imperative for the group to work more cohesively and form more productive meetings. Nevertheless, with the group's experiences and proven capabilities, we still hold high expectations for the final outcome as we put together more individual efforts.

6 Appendix

	id	listing_url	scrape_id	last_scraped	name	summary	space	description	experiences_offered	neighborhoo
0	2595	https://www.airbnb.com/rooms/2595	20191204162729	2019-12-07	Skylit Midtown Castle	Beautiful, spacious skylit studio in the heart...	- Spacious (500+ft²), immaculate and nicely fu...	Beautiful, spacious skylit studio in the heart...	none	Centrally lo heart of Ma
1	3831	https://www.airbnb.com/rooms/3831	20191204162729	2019-12-07	Cozy Entire Floor of Brownstone	Urban retreat: enjoy 500 s.f. floor in 1899 br...	Greetings! We own a double-duplex brownst...	Urban retreat: enjoy 500 s.f. floor in 1899 br...	none	Just the urban cen
2	5099	https://www.airbnb.com/rooms/5099	20191204162729	2019-12-06	Large Cozy 1 BR Apartment In Midtown East	My large 1 bedroom apartment has a true New Yo...	I have a large 1 bedroom apartment centrally l...	My large 1 bedroom apartment has a true New Yo...	none	My neig Midtown E
3	5121	https://www.airbnb.com/rooms/5121	20191204162729	2019-12-06	BlissArtsSpace!	NaN	HELLO EVERYONE AND THANKS FOR VISITING BLISS A...	HELLO EVERYONE AND THANKS FOR VISITING BLISS A...	none	
4	5178	https://www.airbnb.com/rooms/5178	20191204162729	2019-12-05	Large Furnished Room Near B'way	Please don't expect the luxury here just a bas...	You will use one large, furnished, private roo...	Please don't expect the luxury here just a bas...	none	Theater d restaurants

Figure 1: Snapshot of listings_full data set

	Features	Coefficients
0	host_response_rate	-0.005086
1	longitude	-0.058669
2	bathrooms	0.074845
3	bedrooms	0.046156
4	extra_people	0.002414
5	maximum_nights_avg_ntm	0.001703
6	review_scores_cleanliness	0.017512
7	review_scores_checkin	-0.016002
8	review_scores_communication	-0.011754
9	review_scores_location	0.035581
10	review_scores_value	-0.017127
11	available	0.011659
12	maximum_nights_y	0.004681
13	nwords	0.001649
14	prop_cap	0.023720
15	host_id	0.003090
16	accommodates	0.131548
17	maximum_nights_x	-0.014804

	Features	Coefficients
18	minimum_maximum_nights	0.094756
19	maximum_maximum_nights	0.000454
20	availability_90	0.024713
21	number_of_reviews	-0.006975
22	number_of_reviews_ltm	-0.024868
23	calculated_host_listings_count	-0.009975
24	calculated_host_listings_count_shared_rooms	-0.014084
25	has_space	0.017592
26	has_notes	-0.003449
27	has_transit	-0.012162
28	has_host_about	0.002310
29	host_verifications_count	-0.013207
30	neighbourhood_cleansed_Midtown	0.018095
31	neighbourhood_group_cleansed_Manhattan	0.160025
32	property_type_Apartment	-0.060865
33	property_type_Boutique hotel	1.473231
34	room_type_Entire home/apt	0.111206
35	host_since_year_2019	0.003403

Figure 2: LASSO results

References

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press.

Hotelling, Harold. 1933. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology* 24 (6). Warwick & York: 417.

Martinez, Richard Diehl, Anthony Carrington, Tiffany Kuo, Lena Tarhuni, and Nour Adel Zaki Abdel-Motaal. 2017. "The Impact of an Airbnb Host's Listing Description 'Sentiment' and Length on Occupancy Rates."

Phillips, Paul, Stuart Barnes, Krystin Zigan, and Roland Schegg. 2017. "Understanding the Impact of Online Reviews on Hotel Performance: An Empirical Analysis." *Journal of Travel Research* 56 (2): 235–49. https://doi.org/10.1177/0047287516636481.