DSCI 553 Final Project

USC Viterbi
School of Engineering

Airbnb Listing Price Prediction

Jiwoong Kim |
Kai-Ping (Rory) Wang |
Phuong Ngo |

Table of Contents

I. Project Introduction & Hypothesis

Airbnb has become a familiar name to travelers, tourists and property owners. Airbnb's popularity is accounted for by the fact that it lets its users, commonly known as hosts, to list part or all of their residences as travel lodgings, allowing them to make profit from their extra bedroom or the holiday bungalow that they rarely use. This opens up more options for tourists and travelers, as they have access to more local and homely places yet still get some of the experiences of staying at a hotel at a fraction of the price. Hence, there is a need from both the hosts and the renters to know the market price, usually determined by the location of the property, its characteristics and amenities, and the timing of the booking (holiday season or slow season).

This report examines data gathered from Airbnb's listings, along with their descriptions and reviews, in conjunction with the occupancy model suggested by Inside Airbnb. Popular listings will generally have similar descriptions or contain certain frequent words. In addition, those listings also get relatively positive reviews from guests. This means that a listing's texts are likely to influence its price. The report goes through the process of building a price prediction model and adding text as additional features by transforming text data into categorical or numerical variables in order to test this hypothesis.

II. Literature Review

Listing price prediction is not new and has been done before. Doing a simple search on Towards Data Science, we can find some similar projects done on different Airbnb data. For example, an analysis was done against the London data. Ten features were identified, such as accommodation capacity, cleaning fee, listings per host and so on. The creator built a neural network around these features. There is also another project on the Edinburgh data, project also identifies ten important features, some of them are days available to be booked out of the 90 days, the extra fee per person, the number of reviews, and so on. This project attempted to look for how the listing's accessibility to points of interest (i.e. walking distance to tourist attractions) affects pricing through some geospatial analysis, but it is not so distinct in this research model.

III. Data Description

Seeing that not much has been done on examining correlation between listing's text and its price, the project sets out to explore this, specifically on New York Airbnb data. Two datasets are used in this project, listing data with all listing information and reviews data. They are both provided by Inside Airbnb, an open-source Airbnb online database.

- **Airbnb Listing data for New York City, NY, USA**: The dataset contains 44,666 records with 48 fields, including host information like response time and days active; listing information and features like amenities and location; reviews information like the number of reviews and rating scores.
- **Airbnb Review data for listings in New York City**: The dataset contains 1,003,064 review comments and the listing ID numbers of the properties reviewed by guests to help the analysis.

IV. Analysis & Results

Data Cleaning

The process of cleaning data for regression modeling is primarily adopted from Laura Lewis' previous work, but largely customized to account for the differences in the NYC dataset. The process involves (1) dropping features that are considered to be non-predictive, (2) dropping features that contain mostly null values, (3) identifying features (by column names) that might be highly correlated to each other, verifying the correlation and dropping redundant ones, (4) dropping Boolean/categorical features that do not contain sufficient number of instances in each category, and (5) reviewing the remaining features one by one to perform further data manipulations as necessary, such as converting the format of values, replacing null values to its own categorical value, dropping the rows with null values, retaining top *n* categories for features with too many varieties, and binning numbers into categories.

Then the dataset is assessed for multicollinearity and then is then standardized, and categorical features are one-hot encoded. A train-test split is performed with a test size of 0.2.

Building Regression Models and Evaluation

There are six models built for price prediction using six different machine learning methods widely used for modeling regression problems:

- Ridge regression: Linear regression with L2 regularization
- Lasso regression: Linear regression with L1 regularization
- ElasticNet regression: Linear regression with combined L1 and L2 regularization
- XGBoost: Gradient-boosted decision trees
- Four-layer neural network with L1 regularization
- Four-layer neural network with some collinear features removed

Because this is a regression problem, both MSE and R-squared are looked at as metrics for evaluation of the models, and R-squared value is the primary metric for selection of the baseline and final models. Among these models, the XGBoost model performs the best with the R-squared value of 0.6710 (Fig.1). This implies approximately 67% of the variance of price could be explained by the baseline model developed. The performance of each model can be seen in Fig.1

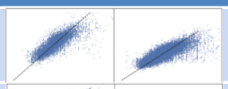
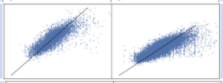
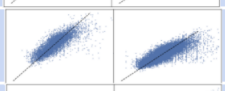
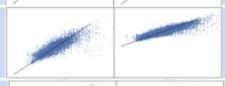
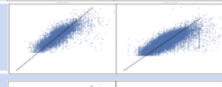
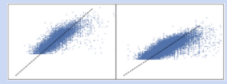
	Training MSE	Testing MSE	Training R^2 score	Testing R^2 score	Test set | Train set
Ridge Regression	0.1931	0.2059	0.6344	0.6167	
Lasso Regression	0.1937	0.2057	0.6334	0.6172	
ElasticNet Regression	0.1933	0.2057	0.6340	0.6172	
XGBoost	0.1128	**0.1768**	0.7864	**0.6710**	
Four-Layer NN	0.1988	0.2115	0.6236	0.6065	
Collinear-Feature Removed NN	0.1999	0.2107	0.6216	0.6080	

Fig.1 – Six prediction models and their performances

Model Improvement - Topic Modeling

Since correlations between texts and listing price is the main hypothesis, a topic modeling method is carried out on listing description in order to turn a group of words frequently appear together into categorical topics. All descriptions are pre-processed by removing stop-words, any punctuations and special characters. Words are then turned into vectors by Countvectorizer to get the most frequent words. These words tend to be about location and amenities as some of the top frequent words are apartments, bedroom and located (Fig.2). Next, descriptions are turned into topics using Latent Dierechlet Allocation (LDA) in Natural Language Processing (NLP). It's an unsupervised classification method to classify texts in a document to one or various topics. Descriptions are then divided into three, five and ten topics with words that make up that topic. Looking only at 5 topics, one can see that more than 13,000 descriptions are classified to each topic 2 and topic 3 respectively (Fig.3). These groups of topics will be used to test and improve the prediction model.

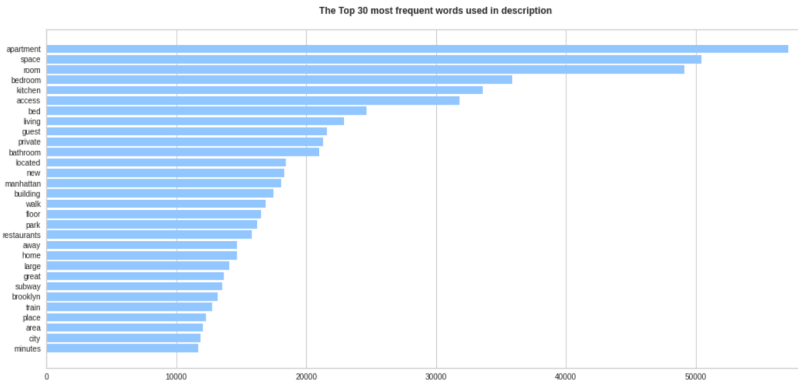


Fig.2 - Top 30 most frequent words used in Listing

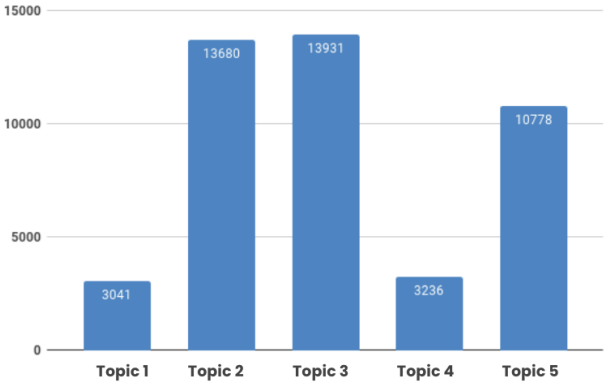


Fig.3 - Number of descriptions allocated to each topic

Model Improvement - Sentiment Analysis

Different from descriptions, which are mostly in English, reviews are written in different languages. For this project, only reviews written in English are used. Using the same pre-processing and vectorizing method, we are able to get 30 frequent words presented in reviews. Again, location and amenities seem to be the top topics based on the words (Fig.4). Sentiment analysis is then implemented to get a sentiment score for each review, which consists of positive, negative, neutral and compound scores. Compound scores range from -1 to 1 and is the sum of all the other scores. It shows how a review leans towards positive or negative. All the scores seem to be rightly skewed, meaning that there are mostly positive reviews (Fig.5).

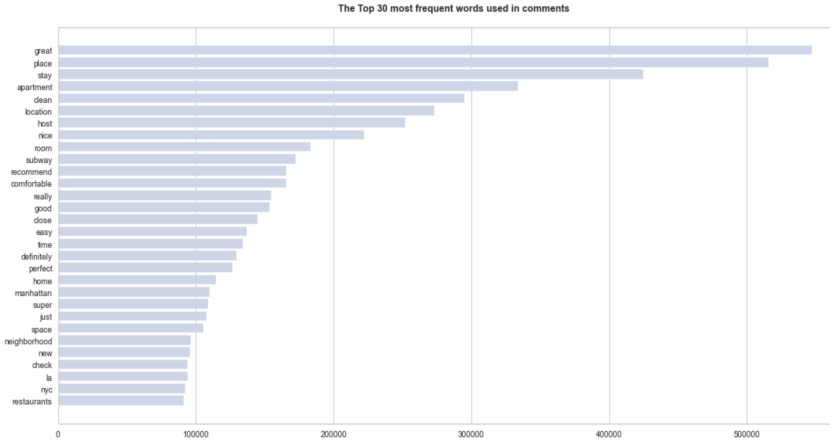


Fig. 4 - Top 30 most frequent words used in Reviews

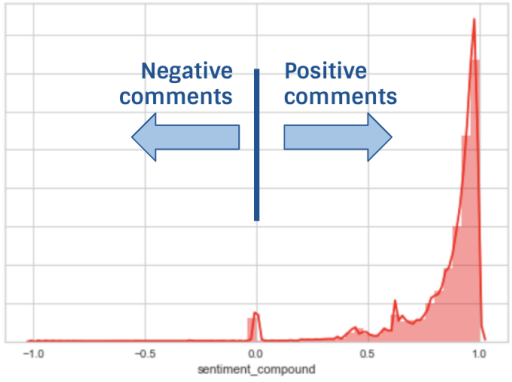


Fig. 5 - Sentiment score distribution

Model Evaluation

The categorized topics obtained from topic modelling for listing descriptions are then incorporated into the model with the best performance, XGBoost. The value of R-squared of this new model sees a slight improvement from 0.6710, the baseline, to 0.6735. Then an additional feature, which is the sentiment scores obtained from sentiment analysis for user reviews, is added to the model. The R^2 for the test set sees a significant improvement of 0.7103

from the baseline. Adjusted r-square is also calculated to make sure that the changes in the number of features do not give some unwanted favor to the models with more features and produce similar results. This confirms our hypothesis that there is a correlation between the sentiment of the reviews for properties and their listing prices.

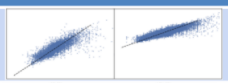
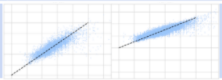
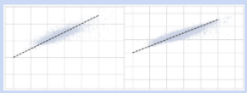
	Training MSE	Testing MSE	Training R^2 score	Testing R^2 score	Test set | Train set
XGBoost	0.1128	0.1768	0.7864	0.6710	
XGBoost + topic modeling	0.1111	0.1755	0.7896	0.6735	
XGBoost + topic modeling + sentiment analysis	0.0877	**0.1357**	0.8151	**0.7103**	

Fig. 6 – Prediction model and its evaluations after adding text features

Final Model

Given the results, the final model selected from the analysis is the XGBoost regression model add on text features. The model explains 71% of the variation in price and proves that there is a correlation between listing's texts and its price. Topics remain in the top 20 important features of the model, further proving the hypothesis (Fig.7).

Fig. 7 – Features importance in the final model

V. Conclusions & Future Work

From the project, we learn not only about the important features that are affecting the price of an NYC accommodation listed on Airbnb, but also several text classification techniques, such as language detection, regular expression, LDA, sentiment analysis, and so on. We also get the chance to build models using multiple algorithms such as linear regression, decision tree, and so on.

Since only English reviews are used when conducting text classification, reviews written in other languages are not factored in. In addition, topic categories were not distinctive enough due to the nature of real estate listing descriptions, which is why the model was not much improved by this feature.

One way to improve the model is by pre-processing and translating the comments prior to text classification in order to analyze those written in other languages. LDA model can be further improved by more sophisticated pre-processing of the text or adjusting LDA parameters to additionally generate more distinctive categories.

VI. Appendix

References:

1. Inside Airbnb. http://insideairbnb.com/get-the-data.html
2. Lewis, Laura. “Predicting Airbnb prices with machine learning and deep learning”. Towards Data Science. https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6
3. Carrillo, Graciela. “Predicting Airbnb prices with machine learning and location data”. Towards Data Science. https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a