Paper 1326-2017

Price Recommendation Engine for Airbnb

Praneeth Guggilla, Snigdha Gutha, Dr.Goutam Chakraborty, Oklahoma State University

ABSTRACT

Airbnb is the world's largest home sharing company and has over 800,000 listings in more than 34,000 cities and 190 countries. Therefore, the pricing of their property, done by the Airbnb hosts, is crucial to the business. Setting low prices may hinder profits, while setting high prices may result in no bookings at all.

In this paper, we have suggested a pricing methodology for Airbnb hosts that helps in overcoming the problems of overpricing and underpricing by observing the factor that determine occupancy rate and price. Through this methodology, we are trying to identify key factors related to Airbnb pricing such as:

1. Those influential in determining a price for a property
2. The relation between the price of a property and frequency of its booking
3. Similarities among successful and profitable properties

As a part of this methodology, we built a scrapping tool to get details of New York City host user data along with their metrics. Using this data, we will build a pricing model to predict the optimal price of an Airbnb home.

INTRODUCTION

The purpose of this project is to build a price recommendation engine for Airbnb hosts. We chose this topic in particular due to the raising concern among some Airbnb hosts that their properties get very less bookings or no bookings at all. We wanted to analyze what factors are driving the occupancy rate and we hypothesized that price will play a significant role in getting bookings along with other factors. Through our analysis we found that price did impact occupancy rate. Keeping that in mind we analyzed further to come up with factors that play a significant role in determining price of a property.

The dataset compiled for this project serves as a foundation for additional research. We did not have any information on the number of bookings a property got, but we have reviews for the properties. So we used reviews count as an indicator for bookings. We assumed that most guests leave a review after staying at a property.

PROJECT CONSIDERATIONS

IDENTIFICATION OF POTENTIAL BENEFACTORS

This study will particularly benefit Airbnb hosts and their customers. In general any property listing service can benefit from this study. Pricing a property is very different especially when there is only little variation across. In our project we tried to capture every minute detail from amenities to surrounding neighborhood. This will also keep hosts informed about what guests are looking for while booking a property.

CONSTRAINTS AND LIMITATIONS

One of the major concerns is that since we do not have exact occupancy rate, we used review count as an indicator of occupancy, the results can be stated more confidently provided we had occupancy rate. We scraped web data and we noticed some listings which were updated months before, so presence of dormant data might bring bias in the results.

DATA COLLECTION, CLEANING AND CONSOLIDATION

We scraped data from Airbnb website on December 3rd 2016 for New York City and our dataset has all the listings reported as of that day on the Airbnb site. We captured around 40,228 listings and following information related to them.

- Information related to listing description given by host in text format such as location, house rules and transit and also latitude and longitude.
- Information related to location, response time and response rate of the host.
- Information about property type, room type, bathrooms, bedrooms, amenities, minimum nights, maximum nights, allowed number of persons, cancellation policy and calendar availability.
- Information about daily price, weekly price, monthly price, cleaning fee, extra person fee, security deposit
- Information about review ratings on cleanliness, location, communication, check-in and all reviews of a listing.

For better understanding of host pricing pattern and booking information we created new variables from reviews dataset and listings dataset. We created new variables zip code, street name, and neighborhood using latitude and longitude information. Again grouping neighborhood locations to five major neighborhoods Bronx, Brooklyn, Manhattan, Queens, and Staten Island and created new variable neighborhood_grouped. In order to determine occupancy rate, we created a new variable review_count to denote the number of reviews per listing id. We looked at the distribution of review count and availability to decide the occupancy rate. We have coded all the listings with review count less than 20 and availability more than 30 days as low occupancy and everything else as high occupancy. After final coding the proportion of high occupancy to low occupancy was 55:45. Overall we used 65 variables used for our analysis.

There is missing information for weekly price, monthly price, cleaning fee, security deposit and extra person fee. We imputed them with zero as those fees are not applicable to certain listings. From our initial analysis, we suspected some extreme observations. So we looked at the distribution of all continuous variables and removed all the extreme observations after three standard deviations. Some such observations were, listings that were charging $10,000 for one day and listings with maximum nights as a four digit number. We have also excluded all those listings with no reviews as we cannot determine occupancy rate. To remove bias in the data, we only considered those listings which have at least one review and with a minimum availability of at least 5 days in the 90 day calendar period. Post cleaning the raw dataset we were left with 23,860 listings for our analysis.

Most of the variables had skewed distributions, in order to bring them back to normal distribution, we applied logarithmic transformation on all right skewed variables and power transformation on all left skewed variables. The following table shows the list of transformations applied on the variables.

Variables	**Transformations**
Price	Log transformation
Security deposit	Log transformation
Extra_person fee	Log transformation
Cleaning_fee	Log transformation
Review_scores_rating	Power transformation
Review_scores_cleanliness	Power transformation
Review_scores_location	Power transformation
Review_scores_value	Power transformation
Review_scores_accuracy	Power transformation
Review_scores_communication	Power transformation
Review_scores_checkin	Power transformation

Table1. Transformations of various variables

TEXT TOPICS

We have amenities information in a text field. We did text parsing on amenities information and created text topics and text clusters out of it. We later used the topics extracted as predictors in the regression models built further.

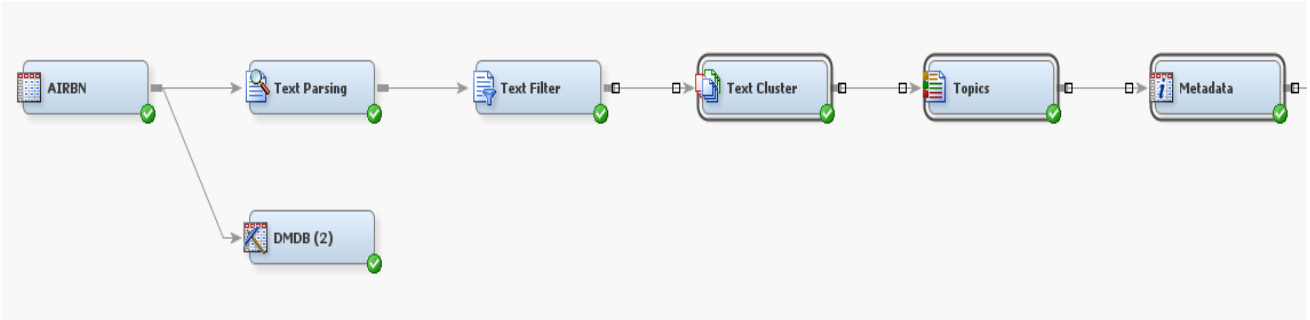

Fig1. Process flow of text topic creation

The following diagram shows the list of text clusters and topics created from the amenities information.

Cluster ID	Descriptive Terms	
1	'hot tub' hot tub building elevator 'indoor fireplace' fireplace indoor washer +premise free +park 'cable tv' family kid	...
2	+pet live property 'first aid kit' aid first kit extinguisher +miss en translation 'wireless intercom' buzzer intercom 'wireless internet'	...
3	wireless 'wireless internet' +heat kitchen +condition air +essential friendly detector tv dryer shampoo +hanger family kid	...
4	+miss en translation 'bedroom door' bedroom door lock +hanger +essential wireless 'wireless internet' detector kitchen +heat +condition	...
5	building elevator washer +park free +premise family kid 'cable tv' dryer hour tv 'first aid kit' aid first	...
6	'first aid kit' aid first kit 'carbon monoxide detector' monoxide extinguisher iron detector hour family kid shampoo 'laptop friendly workspace' laptop	...
7	'indoor fireplace' fireplace indoor extinguisher washer 'cable tv' +premise free +park family kid 'first aid kit' aid first kit	...

Fig2. Text clusters derived from amenities

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.355	0.171	intercom,wireless intercom,buzzer,tv,wireless	6	4958
Multiple	2	0.442	0.161	translation,+miss,en,+essential,detector	5	6099
Multiple	3	0.366	0.164	first,kit,aid,first aid kit,extinguisher	6	6916
Multiple	4	0.295	0.162	+pet,live,property,+allow,cat	6	3339
Multiple	5	0.314	0.157	lock,door,bedroom door,bedroom,intercom	4	4303
Multiple	6	0.304	0.165	elevator,building,wheelchair,accessible,doorman	7	4912
Multiple	7	0.241	0.162	+park,free,+premise,suitable,+event	3	2826
Multiple	8	0.423	0.174	friendly,laptop friendly workspace,laptop,workspace,kid	10	4357

Fig3. Text topics derived from amenities

From the table we can see that the text topics formed are unique in their composition of terms which depicts the varieties of aspects these text topics are covering and appears to be independent from each other. Further down the process during variable reduction process we have also checked for the existence of correlation among these topics.

MODELLING

MODEL FOR OCCUPANCY

Initially we built a multiple regression model to determine the factors effecting occupancy rate. Prior to proceeding with model building, we used variable clustering node in SAS Enterprise Miner for variable reduction. It uses oblique principal component methodology to create clusters by grouping variables that are correlated. Variable cluster node outputted sixteen clusters and we selected sixteen representative variables to represent those clusters based on the minimum (1-Rsquared) ratio. We also have text reviews data, which we used after cleaning and filtering for our modeling purposes. We built text topics from the amenities information given in the data and used them further for modeling. We built two models here, one with text topics and one without text topics. The performance of the model improved after including text topics. For honest assessment we used 70% of data for calibration and 30% of data as validation data. The validation misclassification rate for this model turned out to be around 39.84%.

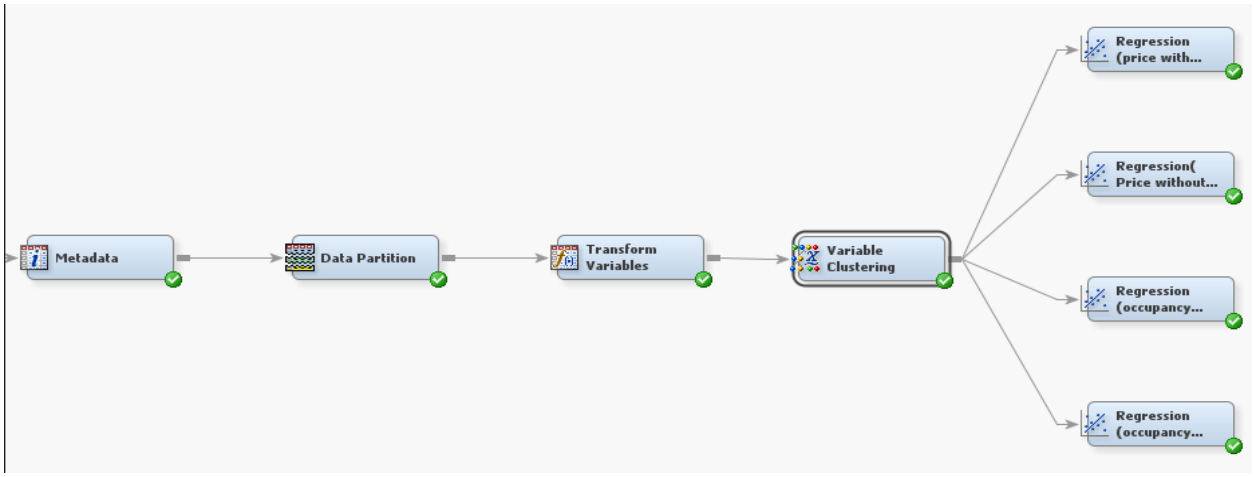

Fig4. Process flow of regression model

The following table shows the list of variables selected by variable Clustering node for modeling occupancy and price.

Variables	
Neighbourhood Group Clenliness	Bathrooms
Security Deposit	Bedrooms
Property Type	Beds
Review Scores Cleanliness	Room Type
Review Scores Communication	TextTopic Raw1
Review Scores Location	TextTopic Raw2
Accommodates	TextTopic Raw6
Extra People	TextTopic Raw7

Table2. Variables Selected for Further Processing

Price turned out to be an important factor as expected in predicting occupancy followed by security deposit, cleaning fee and number of extra people allowed. Interestingly review ratings given on cleanliness by other customers also played an important role in determining occupancy.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-4.8369	0.1925	631.24	<.0001		0.008
LOG_cleaning_fee	1	0.0259	0.0107	5.81	0.0159	0.0247	1.026
LOG_extra_people	1	0.0262	0.0109	5.84	0.0157	0.0241	1.027
LOG_price	1	0.3005	0.0453	44.07	<.0001	0.1059	1.351
LOG_security_deposit	1	2.5282	0.3407	55.07	<.0001	0.0759	12.530
PWR_review_scores_cleanliness	1	0.1375	0.0704	3.82	0.0507	0.0208	1.147
PWR_review_scores_communication	1	-0.6956	0.0934	55.48	<.0001	-0.0835	0.499
PWR_review_scores_location	1	-0.0416	0.0753	0.31	0.5805	-0.00596	0.959
accommodates	1	-0.0318	0.0172	3.41	0.0649	-0.0342	0.969

Fig5. Regression output showing top continuous predictors

Talking about categorical predictors guests are preferring flexible and moderate cancellation policies over strict ones. Guests are preferring Airbnb in Bronx, Queens and Staten Island when compared with that of Brooklyn or Manhattan and Condominium, Camper and Boutique hotels are in demand when compared with Villas or Timeshare properties. Most guests are preferring shared room over entire home or private room. The following diagram shows the odds ratio estimates.

Odds Ratio Estimates

Effect		Point Estimate
cancellation_policy	flexible vs super_strict_30	999.000
cancellation_policy	moderate vs super_strict_30	999.000
cancellation_policy	no_refunds vs super_strict_30	0.982
cancellation_policy	strict vs super_strict_30	999.000
neighbourhood_group_cleansed	Bronx vs Staten Island	1.549
neighbourhood_group_cleansed	Brooklyn vs Staten Island	0.788
neighbourhood_group_cleansed	Manhattan vs Staten Island	0.680
neighbourhood_group_cleansed	Queens vs Staten Island	1.016

property_type	Boutique hotel vs Villa	999.000
property_type	Camper/RV vs Villa	999.000
property_type	Castle vs Villa	999.000
property_type	Condominium vs Villa	1.214
property_type	Timeshare vs Villa	<0.001
room_type	Entire home/apt vs Shared room	0.312
room_type	Private room vs Shared room	0.500

Fig6. Noticeable odds ratio estimates of categorical predictors

MODEL FOR OCCUPANCY WITH TEXT TOPIC INCLUDED

We built a model similar to the above model, but this time we included text topics derived by mining amenities information. Once again Price turned out to be the most important predictor followed by Security deposit and cleaning fee. The standardized estimate of price increased a little when compared with that of previous model. And the impact of reviews decreased a little. Text topics also turned out to be significant predictors. Text topic 2 is discussing about essential amenities and Text Topic 1 is talking about internet connection. The overall model improved a little in terms of validation misclassification rate by 2%. The new validation misclassification rate turned out to be 38.6%. Categorical predictors gave almost similar results with little changes in the odds ratio estimates.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-4.7199	0.1963	578.38	<.0001		0.009
LOG_cleaning_fee	1	0.0411	0.0108	14.52	0.0001	0.0392	1.042
LOG_price	1	0.3216	0.0462	48.39	<.0001	0.1133	1.379
LOG_security_deposit	1	2.6069	0.3455	56.92	<.0001	0.0782	13.557
PWR_review_scores_cleanliness	1	0.1058	0.0710	2.22	0.1363	0.0160	1.112
PWR_review_scores_communication	1	-0.6565	0.0943	48.43	<.0001	-0.0788	0.519
PWR_review_scores_location	1	-0.0281	0.0761	0.14	0.7124	-0.00402	0.972
TextTopic_raw1	1	-1.7488	0.1116	245.55	<.0001	-0.1500	0.174
TextTopic_raw2	1	0.2303	0.0657	12.27	0.0005	0.0333	1.259
TextTopic_raw6	1	0.4613	0.0939	24.11	<.0001	0.0461	1.586
TextTopic_raw7	1	-0.2011	0.0945	4.53	0.0334	-0.0196	0.818

Fig7. Regression output showing top continuous predictors

MODEL FOR PRICE

We used variables that came out of variable clustering node as predictors here, without text data this model is able to explain about 68.74% of variation in the data.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	3.6628	0.0222	164.99	<.0001
G_neighbourhood_group_cleansed	0	1	-0.1929	0.00584	-33.04	<.0001
G_neighbourhood_group_cleansed	1	1	-0.0575	0.00417	-13.79	<.0001
G_property_type	0	1	-0.1484	0.0464	-3.20	0.0014
G_property_type	1	1	-0.0391	0.0172	-2.27	0.0231
G_property_type	2	1	0.00521	0.0162	0.32	0.7480
LOG_security_deposit		1	1.0831	0.0539	20.08	<.0001
PWR_review_scores_cleanliness		1	0.1028	0.0122	8.43	<.0001
PWR_review_scores_communication		1	-0.1181	0.0160	-7.36	<.0001
PWR_review_scores_location		1	0.3875	0.0127	30.47	<.0001
accommodates		1	0.0706	0.00291	24.31	<.0001
bathrooms		1	0.1238	0.00793	15.61	<.0001
bedrooms		1	0.1392	0.00571	24.39	<.0001
beds		1	-0.0424	0.00439	-9.67	<.0001
guests_included		1	0.0223	0.00301	7.43	<.0001
room_type	Entire home/apt	1	0.5238	0.00624	83.97	<.0001
room_type	Private room	1	-0.0360	0.00581	-6.21	<.0001

Fig8. Regression output showing parameter estimates

Security deposit turned out to be a significant predictor in determining price followed by Room type, Reviews about location, Neighborhood cleanliness, Property type, beds and reviews given by other guests.

MODEL FOR PRICE WITH TEXT TOPICS INCLUDED

We built a model similar to the one above, but this time we included text topics from amenities information. This improved the performance of model by a little extent. The proportion of variance explained by this model turned out to be 69.78%.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	3.6561	0.0222	164.95	<.0001
G_neighbourhood_group_cleansed	0	1	-0.1817	0.00580	-31.30	<.0001
G_neighbourhood_group_cleansed	1	1	-0.0496	0.00412	-12.05	<.0001
G_property_type	0	1	-0.1391	0.0456	-3.05	0.0023
G_property_type	1	1	-0.0128	0.0170	-0.75	0.4516
G_property_type	2	1	-0.00074	0.0159	-0.05	0.9631
LOG_security_deposit		1	1.0000	0.0533	18.78	<.0001
PWR_review_scores_cleanliness		1	0.0991	0.0120	8.25	<.0001
PWR_review_scores_communication		1	-0.1159	0.0158	-7.34	<.0001
PWR_review_scores_location		1	0.3763	0.0125	30.00	<.0001
TextTopic_raw1		1	0.1356	0.0186	7.29	<.0001
TextTopic_raw2		1	-0.1051	0.0111	-9.44	<.0001
TextTopic_raw6		1	0.2900	0.0157	18.44	<.0001
TextTopic_raw7		1	-0.0934	0.0160	-5.85	<.0001
accommodates		1	0.0669	0.00288	23.25	<.0001
bathrooms		1	0.1224	0.00783	15.63	<.0001
bedrooms		1	0.1437	0.00563	25.54	<.0001
beds		1	-0.0411	0.00431	-9.53	<.0001
guests_included		1	0.0223	0.00296	7.52	<.0001
room_type	Entire home/apt	1	0.5084	0.00621	81.84	<.0001
room_type	Private room	1	-0.0347	0.00571	-6.08	<.0001

Fig9. Regression output showing parameter estimates

The results look similar to the above model except for small changes in the parameter estimates. Text topic 1 discusses about Wi-Fi and TV, Text topic 6 discusses about having elevator, doorman. They have a positive impact on price, which implies that guests are ready to spend money over those. Text topic 2 talks about essential amenities and Text topic 7 talks about parking. They have a negative impact on price which means guests are not willing to pay extra for these facilities, they don't want these to make any difference in the price, while we can see that essential amenities played a major role in determining occupancy rate, we need to keep in mind while pricing a property.

CONCLUSION

We would like to conclude by summarizing the results and insights obtained from our analysis. We found out that price turns to be very significant in determining occupancy rate followed by security deposit. And people are giving utmost importance to cleanliness and reviews by other customers before renting a property. Bronx and Queens are having good business over Manhattan or Brooklyn, it might be because of presence of friends or cousins in these cities that there is no necessity for renting a property. Essential amenities, Internet connectivity and Television has pretty good impact on deciding occupancy, also we need to keep in mind that people are not willing to pay extra for essential amenities. Again Security deposit turns out to effect price. People are willing to spend extra on internet connectivity and television. Location and neighborhood has to be kept on mind while pricing a property.

FUTURE WORK

The scope of the project can be extended by performing sentiment analysis on the text reviews and we can always compare and contrast the numerical rating results with the text results to get more insights. We can always build an optimization model on top of this to come up with optimal prices.

Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Praneeth Guggilla
Oklahoma State University
Phone: 405-780-5330
Email: praneeth.guggilla@okstate.edu

Snigdha Gutha
Oklahoma State University
Phone:
Email: snigdha.gutha@okstate.edu

Dr.Goutam Chakraborty
Oklahoma State University
Email: goutam.chkaraborty@okstate.edu