Airbnb Host Pricing Strategies

Detecting the Influence of the Superhost Status

Ryerson University

Houssem Rekik
Prepared for Dr. Tamer Abdou
CIND820: Big Data Analytics Project
July 2021

Table of Contents

Github

Listings: https://github.com/Houssemrekik/CIND820-Final-Project/blob/main/listings-v2.ipynb
Reviews: https://github.com/Houssemrekik/CIND820-Final-Project/blob/main/reviews.ipynb

1. Abstract

Airbnb has created the superhost program which provides benefits to distinguished hosts. Superhosts are experienced hosts who provide a shining example for other hosts[1]. For non-superhosts, pricing is yet a critical factor in the long-term success or failure of the accommodation.

The purpose of this study is to predict the price that a superhost would apply to an accommodation in the city of Toronto belonging to a non-superhost, and to test if a significant difference exists between superhost predicted prices and non-superhost actual prices, given a set of relevant features from the dataset[2]. Multiple regression algorithms will be applied to select the best model, followed by a hypothesis testing will help answer that question

We will exploit the "insideAirBnb" dataset, that gathers 15542 listings in Toronto as of April 2021. Along with the numerical features, the dataset includes a key textual attribute which is the guests 'reviews (410918). Mining these reviews would be of a substantial added value to the regression model.

We will use a dictionary-based technique-the AFINN lexicon- to assign a score to each review and aggregate the scores into a new attribute that will be added to the predictors

The selected model will be trained on the superhost listings and tested on the non-superhost listings to predict the prices that superhosts would apply on non-superhosts listings. Hypothesis testing is the technique that will be used to judge if a significant difference experienced (superhosts) and less experienced (non-superhosts) accommodation suppliers

Themes: Regression-Text Mining-Hypothesis Testing

2. Introduction

Online peer-to-peer marketplaces, which are part of the "sharing economy" business model, allow people to offer rooms or entire houses to tourists, with Airbnb being the biggest and most famous example[3], founded in San Francisco-California in 2008. The pricing is yet a challenge for hosts willing to share(rent out) their accommodations: "For many hosts, finding the right price for their space can be both time-consuming and challenging... Even many experienced hosts told us that they find pricing difficult, especially as seasons change, special events come to town, and more listings emerge in their neighborhood." [4]

[1] https://www.airbnb.ca/help/article/828/what-is-a-superhost
[2] http://insideairbnb.com/get-the-data.html
[3] Borg, J., Camatti N., Bertocchi, D., Albarea, A. The Rise of the Sharing Economy in Tourism: Exploring Airbnb Attributes for the Veneto Region-Italy (2017)
[4] https://airbnb.design/smart-pricing-how-we-used-host-feedback-to-build-personalized-tools/

Some of the common mispricing strategies that hosts can undertake are pricing too high too early, pricing too low too long, and pricing the same throughout the year. That is why Airbnb offered an automated source of pricing and released its first tool in 2012[5]. Some of these popular tools are "Price Tips" and "Smart Pricing", which also aim to reduce the gap in market knowledge between professional and casual hosts[6]. Another proxy for the knowledge gap is the superhost status, which is a reward given by Airbnb to its most experienced hosts who provide the guests with the best stays.

The "Superhost" qualification is automatically evaluated every three months, and owners must satisfy four conditions to obtain and keep the badge: **a)** receive at least 10 bookings in a year **b)** respond to their guests quickly and maintain at least 90% response rate **c)** rarely conceal confirmed reservations, 1% or lower **d)** satisfy most of their guests and maintain a 4.8 overall rating[7]. Therefore, owners must devote more energy to their listings

Reviews in peer-to-peer marketplaces have a great importance and may explain the price variation, that is why analyze them and attribute a score to each listing

The main research question of this project is to determine how significantly different is the superhost pricing from the non superhost one. In other terms, if the regular host were a superhost, what price would have he applied, given hi experience as a superhosts and his better knowledge of the expectations of Airbnb guests ?

To address this question, I will apply four machine learning regression algorithms to the subset of superhost listings: Random Forest, KNN, XGBoost, and Linear Regression (after assessing its five assumptions). For the evaluation techniques, I will combine both train/test split and K-folds. Another combination of inserting/removing 2 categorical variables (amenities & neighborhood) will raise the total number of models to 16. This work also encompasses a text mining portion, as we will use a dictionary-based approach to analyze these the guests'reviews

When building the models, feature selection filter method will be used to only retain the potent attributes

The models will be trained and tested on the superhost subset, and then the best-performance model will be applied to predict how superhosts would have set the daily rates for listings belonging to non superhosts. Conclusions will be drawn upon testing the difference between these 2 groups

---

[5] Hill, D., How much is your spare room worth? SPECTRUM.IEEE.ORG, 2015
[6] Casamatta G., Giannoni S., Brunstein D., Host type and pricing on Airbnb: Seasonality and perceived market power
[7] https://www.airbnb.ca/help/article/829/how-do-i-become-a-superhost

3. Literature Review

Airbnb, as a tech-heavy business model, has been the subject of a plethora of academic research. However, there is no similar approach in trying to predict prices based on dividing the dataset into two subdatasets, i.e., superhosts and non superhosts, such approach represents the originality or the main contribution of this work.

We will expose a summary of the most related articles to our topics : first is pricing and the superhost status, second is the sentiment analysis on the guests 'reviews

3.1. Pricing and the "superhost" status

3.1.1. Liang, S., Schuckert, M., Law, R., Chen, C.C., 2017. Be a "Superhost": the importance of badge systems for peer-to-peer rental accommodations.

In this paper, the authors state that both the volume and the valence of online reviews are critical for product and service suppliers and can boost their online sales, and that consumers tend to pay more attention to "popular" products that are shown with more reviews, while they make their final purchase decisions based on the rating distribution of those target products

The author hypothesized that prospective guests are willing to pay a premium price for the offer with the "Superhost" badge due to its higher added value. The dataset: a web crawler that scraped accommodations in Hong Kong in 2015. Initially retrieved data showed 3830 listings, after removing duplicates and cleaning 1872 hosts were included in the final analysis, with only 2.9% of them having earned the superhost badge

The dependent variables are the Review volume and the Rating (review valence). The independent variables are Superhost Badge (1 or 0) and the price. More control variables were included.

Model: negative binomial for the model targeting the review volume as the number of reviews distribution is not normal and the dependent variable is discrete with many zeros

The values of some variables like price and the number of photos are higher than others, so they were transformed to avoid extremely high coefficients in the results. Price was divided by 1000 and number of photos by 10. Also, the collinearity was measured by the VIF (variance inflation factors). All VIF values were less than 10, which means that collinearity cannot influence the accuracy of the results

It was found out that a significant interaction effect exists between the superhost badge and the price, guests are happy to spend more for accommodations with the superhost badge. Also, an accommodation with the superhost badge is more likely to receive reviews

3.1.2. Berensten A., Rojas M., Waller C., (2019). What is the Value of Being a Superhost?

The authors state that sellers have to provide incentives to attract buyers and get ratings and reviews. One way to do this is offer a low price to start – this will attract buyers. Pricing can be challenging as it is highly plausible that hosts do not know their true quality compared to the average market quality when they launch into the rental activity

It was shown that hosts with high ratings, and among them those with the superhost designation, charge higher average prices (and yet receive more bookings), have a higher occupancy rate and earn more revenue.

The key aspect is to explore the role of the rating and the superhost status through Airbnb by testing whether the coefficient associated with superhost status is statistically significant

The dataset was provided on 4 major cities: Amsterdam, Rome, Miami and San Francisco: highly touristic, sufficiently diverse locations Superhost fractions go from 10.2% in Miami to 17.8% in San Francisco in restricted samples of respectively 4259 listings 2057 listings. The restriction comes from considering only entire apartments, whose daily rates fall within 3 standard deviations, with at least 10 bookings to exclude high-season-rentals-only

The standard OLS regression model was used for each of the 4 cities. The dependent variable is the daily rate (in log). The independent variables are the superhost status, the size of the apartment(maximum number of guests), number of photos in the listing, number of reviews, instant book, type of cancellation, and neighborhood as a dummy variable. The number of bedrooms, number of bathrooms are highly correlated with the size of the apartment, so they were eliminated

Three variant models were constructed for each regression in every city by adding the number of features in each model from an initial model of 4 variables. Adjusted R2 was higher each time a new variable was added. The daily rate mean difference between superhosts and non superhosts was statistically significant in Amsterdam ($p<0.05$) and in Miami ($p<0.01$)

3.1.3. Chattopadhyay M., Mitra S., (2019). Do Airbnb host listing attributes influence room pricing homogenously?

The objective of this work was to investigate and identify the key determinants from various offered amenities that are influential in room pricing and to compare the performance of 3 machine learning algorithms to explore whether the room pricing determinants are similar or locally generalizable across 11 cities in the US or not.

The study employed a comparative approach by using three different methods—OLS, random forest (RF), and conditional inference tree(CTree)[8]—applied on a vast quantity of data from the Airbnb listing dataset for 11 cities in the US (151,955 observations and 143 explanatory variables) after compilation from insideairbnb.com .

Complexity characterizes the relations between determinants and room price because in a dynamic rental market, listing variables may exert a nonlinear impact on room pricing, which leads to infer those non-parametric approaches are a better alternative to deal with the nonlinear influence of room pricing determinants

RF involves a large number of trees generated from random samples, such as 1000 decision trees, for better predictive performance. Due to this randomness in trees, each of the individual trees produces a different prediction. Thus, the mean prediction for the predictions of the individual trees is regarded as the overall prediction of the RF

The explanatory variables in the RF and the Ctree are ordered based on their variable importance, which is determined based on the value changes in the dependent variable in response to the change in explanatory variables. Checking multicollinearity[9] : VIFs for all the variables are less than 10. The correlation matrix also shows there is no collinearity problem

OLS regression was applied to full samples of the whole dataset, and separately, the 11-individual city-specific individual samples were used to implement the RF, CTree, and OLS regression models

Performance measures : The evaluation of the predictability of a model applies the root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). The lower these measures are, the better. Whereas the explanatory potential of a model can be evaluated through R^2 (in-sample fit measures)

In all the 11 cities, Random Forest regressor provided the best results in the 4 performance measures, R^2 ranges from 73% to 88%. It was followed by Ctree and in third place OLS.

[8] The only difference between conditional inference trees and decision trees is that conditional inference trees use a significance test to select input variables rather than selecting the variables that minimizes the information impurity https://www.geeksforgeeks.org/conditional-inference-trees-in-r-programming/#:~:text=Conditional%20Inference%20Trees%20is%20a%20tree%2Dbased%20classification%20algorithm.&text=The%20only%20procedure%20that%20makes,that%20maximizes%20the%20information%20measure.

[9] In multiple regression, multicollinearity may exist between two or more variables even in absence of high correlation between variables that may lead to redundancy. The VIF for a given predictor represents the amount of variance of a coefficient as a result of multicollinearity, and values above 10 indicate problematic amount of multicollinearity

3.1.4. Wang D., Nicolau J.L. (2016) Price determinants of sharing-economy-based accommodation rental: A study of listings from 33 cities on Airbnb.com

The price determinants were categorized into 5 categories: host attributes, site and property attributes, amenities and services, rental rules, and online review ratings score. In the current study, I will focus on all of these categories except the rental rules for lack of data.

A sample of 180,533 accommodation rental offers listed by Insideairbnb.com in 33 cities is analyzed. Only listings with at least 1 review are considered to ensure actual transactions

Two linear regression models were applied : OLS and QR(quantile regression). The main difference between the model types is that OLS regression models are based on the conditional mean of the dependent variable, whereas QR models are based on the conditional t^{th} quantile of the dependent variable, where $t \in (0, 1)$[10], which allows to uncover hidden price-response patterns that exist depending on the level of prices.

The main finding is that superhost status consistently leads to higher prices: $p < 1\%$ in the OLS model and all the 5 quantiles in the QR model. This increment is more noticeable among lower-prices listings than among higher-priced listings

3.2. Sentiment analysis on the guests' reviews

Customers rely more on the reviews than the rating scores (Chevalier and Mayzlin, 2006). According to Archak, et al. (2011), numerical or bimodal ratings do not accurately capture the information embedded in the reviews and may not express precise information to prospective shoppers.

3.2.1. Lawani A., Michael M., Mark T., Zheng Y. (2018) Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston

The use of scores or ratings oversimplifies quality measures by assuming that quality is a unidimensional measure. This study uses the contents of the written reviews to extract the sentiment hidden in the review and uses it as a quality measure. The authors apply sentiment analysis to extract value, opinions or attitudes from the reviews

The data used in this study are from the Airbnb platform for Boston and were retrieved from Inside Airbnb during the month of September 2016. It has 2051 individual hosts

[10] QR goes beyond the analysis of the conditional mean of a dependent variable, providing a more comprehensive description of the conditional distribution. Rather than estimating the average response of the dependent variable to changes in the explanatory variables, QR measures the effects of individual explanatory variables on the whole distribution of the dependent variable

Using sentiment analysis, the opinions in the reviews are mined, and a score is derived. The mean score of the reviews for each room is used as a proxy for the quality of the room

AFINN lexicon was used to extract the sentiment from the words, it is a program based on unigrams (single words). Each unigram is assigned a score from -5 to +5. The total score of a review is given by the sum of the scores of the words in that review.

Only the reviews written in 2016 were used as customers on online platforms focus on more recent comments (Pavlou and Dimoka, 2006). N-gram-based approach for text categorization and texcat package were used to retrieve the reviews written in English. Also “Host cancelled this reservation … this is an automated posting" reviews were dropped from the dataset. In total, 22,651 reviews were mined and the average of the review score per room is used as a proxy for the room quality

The detailed procedure of scoring is as follows: cleaning the reviews (removing punctuation, numbers, extra spaces and non-textual contents, removing the stopwords, stemming, matching each stem with a unigram in the list of sentiment words in the AFINN lexicon, calculating the final score of a review which is the sum of positive and negative matches

The price (logPrice) estimation was performed with OLS regression. One model was built without the sentiment analysis score, and another one integrated it. The second model provided a better price estimation. Also, the sensitivity analysis showed that hosts with rooms of high-quality set higher prices compared to hosts of low-quality rooms.

3.2.2. Kiatkawsin K., Sutherland I., Kim J.Y. (2020) A Comparative Automated Text Analysis of Airbnb Reviews in Hong Kong and Singapore Using Latent Dirichlet Allocation

A useful summary of the three different approaches used for automating text analysis tasks can be derived from article.

First is the **dictionary-based** approach. A set of words (unigrams) with topics and emotions is predefined. Then, words in the documents are detected and compared to the unigrams to help produce the outcome, which is mostly a score. However, this is non-contextual and considers the words in isolation of the global meaning of the expression/sentence, which makes the dictionary-based approach unable to capture sarcasm, metaphors, or idioms.

Second approach is the **feature extraction**, which is like the first, but the difference is the use of ML algorithms to define the features. The process suggests that the algorithms, after being trained, should be able to detect features from a text corpus. These models require large datasets with known features to learn from.

The third and most recent approach is based on word co-occurrences, to understand how words are gathered to convey meaning, assuming that any word appearance in a given context is not random.

The algorithm utilizes a **document-term matrix** to map the frequency of word co-occurrences, LDA (Latent Dirichlet Allocation) being the most commonly used modeling technique. However, LDA produces results that do not focus on the meaning, and requires some degree of classification and subsetting of the data before extracting topics

3.3. Implications from the literature review

From what has been discussed above, many important concepts and approaches will be implemented in this work:

-non-parametric and parametric regression algorithms: Random Forest and Linear Regression respectively, after assessing the assumptions of the latter

-VIF as a measure of collinearity between explanatory variables

-transforming the dependent variable which is the price into log (or Box Cox transformation which is more general)

-RMSE and MAE as performance indicators. We will highlight the difference between them and define which one will be a better measure

-to ensure that the listings (or the datapoints) used in the model are active and their prices are not outdated, we will remove the listings whose last review is older than 2019. In the same way, text analysis will only treat recent reviews (in 2019 or later)

-the dictionary-based approach and the AFINN in mining the text reviews

4. Dataset

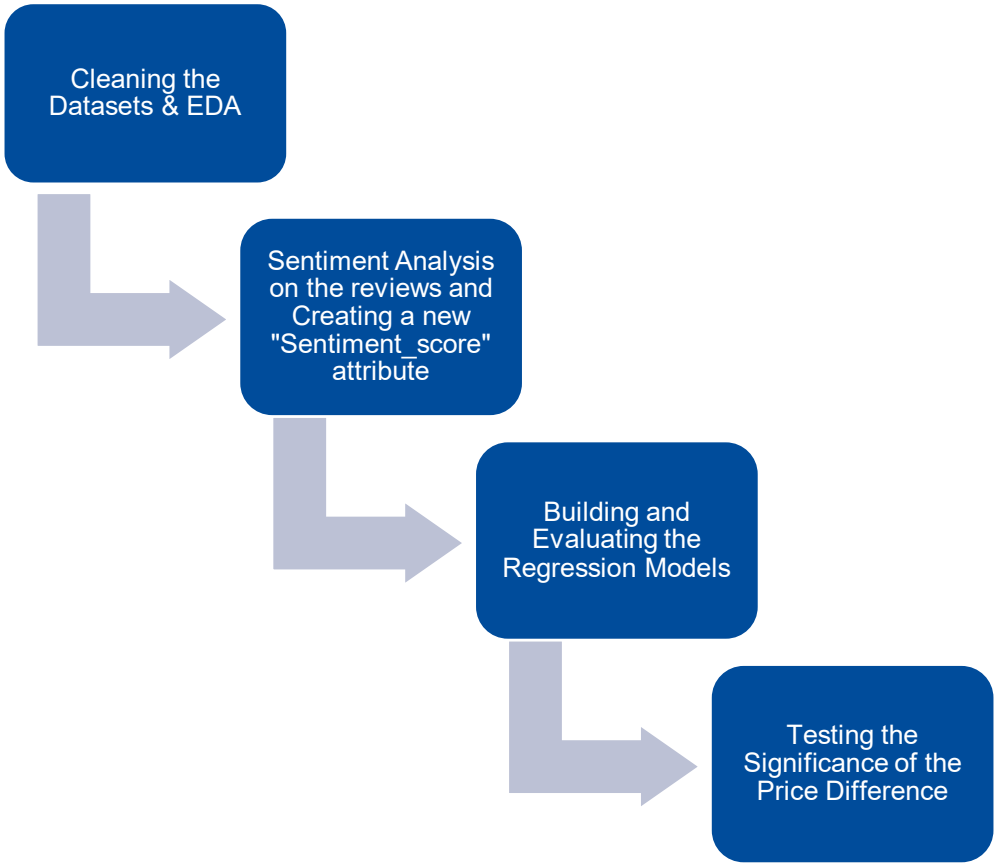
Our data was extracted from Insideairbnb.com which is an independent, non-commercial set of tools and data that allows to explore how Airbnb is really being used in cities around the world.

Our data comprises two datasets: listings and reviews. The listings were extracted from the city of Toronto in April 2021. It initially has 15542 listings and 73 attributes.

The reviews dataset bulks 411,220 reviews from 2009 to 2021 for 11952 listings.

The superhost status distribution is 3955 superhosts, 11170 non superhosts and 417 missing values.

5. Approach



5.1. Cleaning the data and EDA

In this phase, we will deal with the missing values and the outliers with values imputation if needed, analyze the distribution of the numeric attributes, and the correlations. We will also remove non relevant attributes from the “listings” dataset.

For the categorical variables: label encoding for ordinal features (exp: first_review, last_review, room type), and oneHot encoding for nominal features like amenities, neighborhoods

To improve understanding the data, univariate and bivariate relationships between the daily rate and the other variables will be examined.

5.2. Sentiment Analysis on the reviews and Creating a new "Sentiment_score" attribute

First, the “reviews” dataset will be prepared for the sentiment analysis. Comments before 2019 will be removed as well as automatic postings “Host cancelled this reservation. This is an

automated posting". Non-English comments will also be dropped using “langdetect” language detection package

Then, I will use a dictionary-based approach to attribute a score to each review and assign it to the corresponding listing as a new “sentiment_score” attribute. If a listing has more than one review, the median will be taken as the rating.

A drawback of using the raw AFINN score is the that longer texts may yield higher values simply because they contain more words. To adjust for that, we will divide the score by the number of words in the text.

Upon completing that step, we will merge the ‘listings’ dataset with the ‘reviews’ dataset, based on the listing_id column, to end-up with one dataset.

5.3. Building and Evaluating the Regression Models

This is the main step of this study. We will train and test four regressors and combine multiple models on the “superhost” subset. The regressors are: Random Forest, Linear Regression, KNN, XGBoost.

Feature selection: the filter method will be applied. The independent attributes that have a high correlation between each other will be removed and one will be kept to avoid mutli-collinearity. Threshold is 0.6. At the same time, a relatively high collinearity exists between the dependent variable (price) and 2 other dependent variables (accommodates and bathrooms_text). The filter method is preferred to the wrapper and the hybrid methods as these ones are computation costly. In calculating correlations, ‘Spearman’ method is the one to be applied as most of the attributes are discrete

We will also apply Spearman test of hypothesis to deem whether two variables are uncorrelated (the null hypothesis) and keep the variables with p-value > 0.05

Models and evaluation techniques: the regressors will be applied on 3 combinations of attributes: excluding the oneHot variables (neighborhood_cleansed & amenities), including the neighborhoods but excluding the amenities, and including both oneHot attributes. For each regressor, we will apply 2 evaluation techniques: trainset/testset split & k-folds, the purpose is to compare and select the best performer

Linear Regression Assumptions: we will evaluate its five following assumptions: linear relationship between the predictors and the response variable, normality of the error terms, no Multicollinearity among Predictors, independence of the error terms, and homoscedasticity

Performance measures: RMSE and MAE will be calculated for each model and each combination of attributes. RMSE is robust to skewness but sensitive to outliers (as when optimizing it, it looks to optimize the mean), it assures to get unbiased forecasts. On the other

hand, MAE protects outliers but is sensitive to skewed distributions (as when optimizing it, it looks to optimize the median). Since the price distribution is skewed and outliers were removed (keeping 99% of the observations), RMSE will be our primary performance measure.

Best model evaluation: the **effectiveness** against three competing algorithms, **efficiency** regarding the time execution, and **stability**: we will vary k- the number of folds- when running the cross validation and draw a curve with k in the X axis and RMSE in the Y axis. If the line goes up exponentially that means that the model is unstable. If the line goes up and down, this is a sign of overfitting (the model is doing a very good job on the training set but is highly biased on the testing set). The optimal case: the line goes up and at a certain k = a goes flat, in which case we can say the model is stable at k = a

5.4. Testing the Significance of the Price Difference

Price prediction: the models will be trained and tested on the "superhost" subdataset. Then the best model will be trained on the "superhost" subdataset but tested on the "non_superhost" subdataset so the output will be the prices that "superhosts" would apply on "non_superhost" listings.

Hypothesis testing: based on whether the actual and predicted "non_superhost" prices are normal or not, we will apply either parametric or non-parametric tests

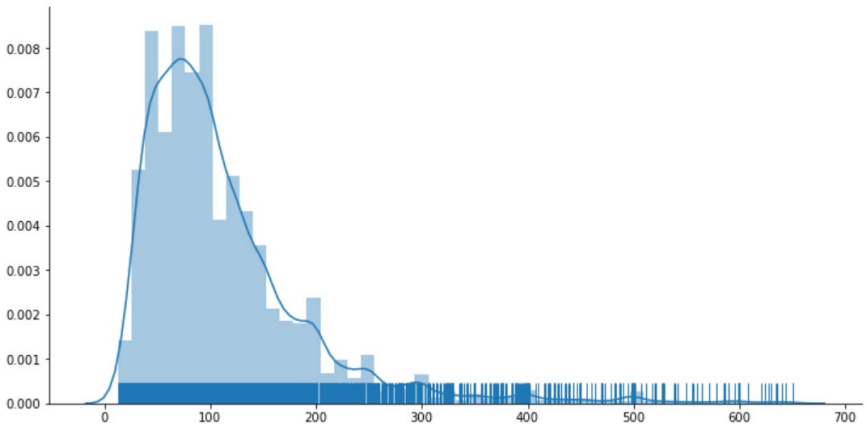
6. Results

6.1. Exploratory Data Analysis

Fig1. Statistical description of numerical attributes

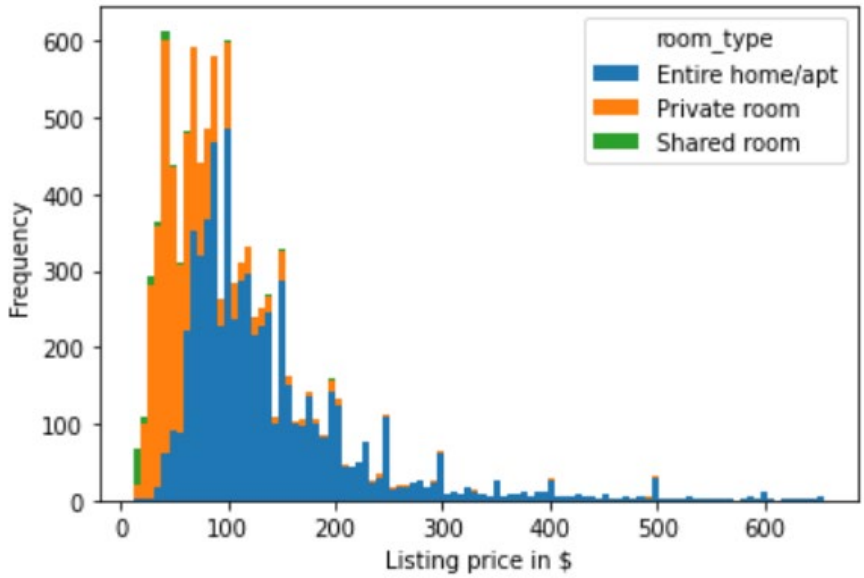
	accommodates	price	minimum_nights	number_of_reviews	calculated_host_listings_count	sentiment_scores_adj
count	9181.000000	9181.000000	9181.000000	9181.000000	9181.000000	9181.000000
mean	3.123625	113.844570	23.557020	40.379806	5.010239	36.515614
std	1.903233	82.817354	30.355241	63.654609	9.324220	21.351499
min	1.000000	13.000000	1.000000	1.000000	1.000000	-66.666667
25%	2.000000	60.000000	5.000000	4.000000	1.000000	25.948592
50%	2.000000	94.000000	28.000000	16.000000	2.000000	33.035714
75%	4.000000	140.000000	28.000000	49.000000	4.000000	41.428571
max	16.000000	650.000000	1000.000000	828.000000	72.000000	300.000000

Fig2. Price distribution



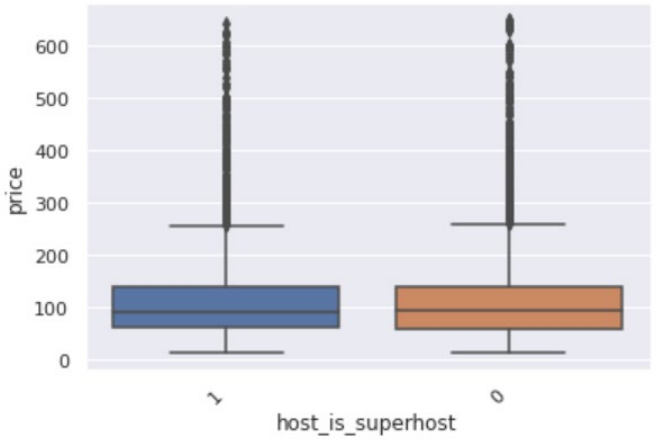
The price distribution is highly skewed to the right (more weight on the left tail of the distribution), as 75% of the prices are less than $140

Fig3. Bivariate distributions: Price vs room_type



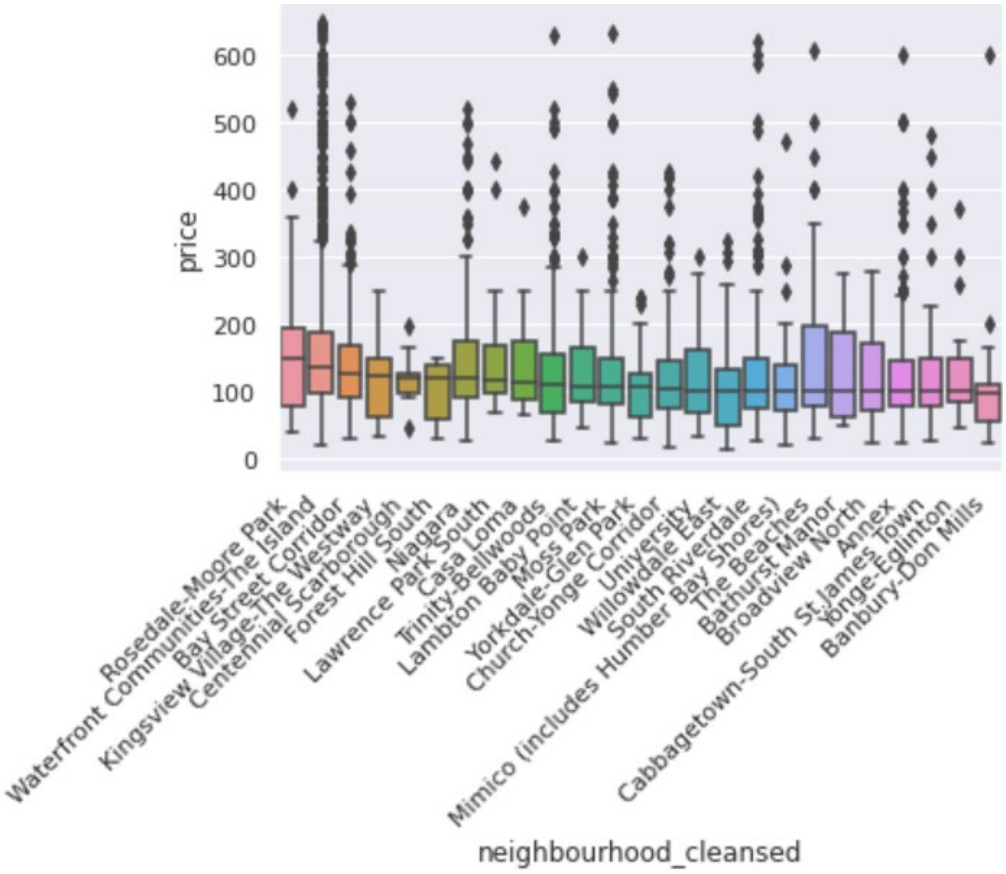
Shared rooms are not very common in Airbnb Toronto listings. As price approaches $100, the great majority of listings become entire homes/apartments

Fig4. Bivariate distributions: Price vs host_is_superhost



The mean of superhosts' prices is slightly higher than non-superhosts

Fig5. Bivariate distributions: Price vs neighborhood_cleansed (sorted)



The closer the listing is to the lake, the main transit stations, the beach or to colleges, the higher the price

6.2. Feature selection

After removing the correlated attributes and confirming with Pearson correlation tests, we end up with 12 independent attributes on the superhost subdataset: 'accommodates', 'minimum_nights', 'number_of_reviews', 'last_review', 'instant_bookable', 'calculated_host_listings_count', 'room_type_Entire home/apt', 'room_type_Private room', 'room_type_Shared room', 'sentiment_scores_adj' , 'amenities', 'neighborhood_cleansed'. The last two will be subject of oneHot encoding

6.3. Results of the regression models on the superhost subdataset

We started with a baseline model that encompasses 10 independent attributes, excluding the amenities and the neighborhood_cleansed

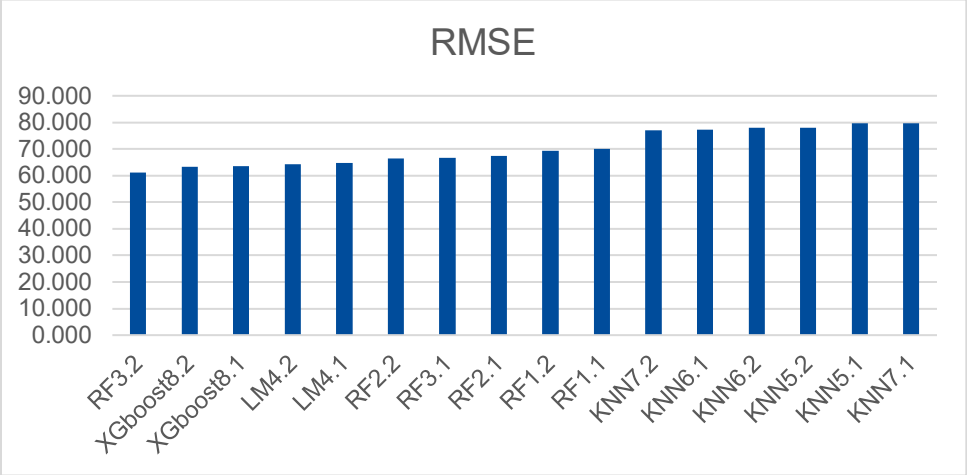
Fig6. Results summary sorted by RMSE

MODEL	Attributes included	Total attributes	Evaluation technique	RMSE	MAE	Execution_time
RF3.2	Baseline + neighborhood + amenities	778	K-fold cross validation	61.031	38.513	229.163
XGboost8.2	Baseline + neighborhood + amenities	778	K-fold cross validation	63.350	39.067	77.440
XGboost8.1	Baseline + neighborhood + amenities	778	K-fold cross validation	63.504	37.350	4.962
LM4.2	Baseline	10	K-fold cross validation	64.175	41.787	0.089
LM4.1	Baseline	10	Train/test split	64.644	41.516	0.013
RF2.2	Baseline + neighborhood	149	K-fold cross validation	66.459	42.197	72.297
RF3.1	Baseline + neighborhood + amenities	778	Train/test split	66.709	39.453	14.263
RF2.1	Baseline + neighborhood	149	Train/test split	67.390	41.400	4.575
RF1.2	Baseline	10	K-fold cross validation	69.351	45.078	32.019
RF1.1	Baseline	10	Train/test split	70.073	44.393	2.253
KNN7.2	Baseline + neighborhood + amenities	778	K-fold cross validation	76.940	50.897	14.315
KNN6.1	Baseline + neighborhood	149	Train/test split	77.321	51.278	0.631
KNN6.2	Baseline + neighborhood	149	K-fold cross validation	78.030	52.191	2.410
KNN5.2	Baseline	10	K-fold cross validation	78.095	52.166	0.295
KNN5.1	Baseline	10	Train/test split	79.615	52.910	0.066
KNN7.1	Baseline + neighborhood + amenities	778	K-fold cross validation	79.749	50.889	3.319

Our most effective regressor which includes the whole set of features & evaluated using the cross-validation technique is the Random Forest regressor (350 trees). However, in terms of efficiency, it ranks last with almost 3min50s execution time. Nevertheless, this model will be used in predicting prices of the non_superhosts listings based on the knowledge of superhosts

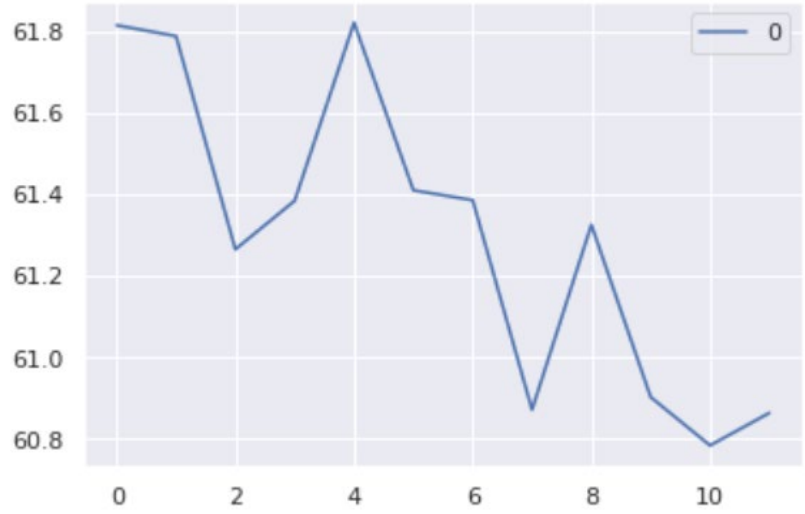
When evaluating the five linear regression assumptions, only one was satisfied which was the independence of the error terms (absence of autocorrelation). That is why results of the linear regression were added only for information.

Fig7. RMSE visualization



When varying 'k' the number of folds, our best model looks rather stable as the accuracy varies only by 1.71% between the max RMSE and the min RMSE. It reaches its maximum accuracy at rmse_val_stab[10] which corresponds to k=13

Fig8. Evaluating the stability of the model RF3.2



According to our best model, the top 5 features (using the feature importance function) are: how many guests the listing accommodates, whether the listing is an Entire home\apt, review scores, number of reviews, minimum nights required

The top 5 amenities: dishwasher, indoor fireplace, bbq grill, patio or balcony, paid parking off premises

The top 5 neighborhoods: Etobicoke West Mall, Waterfront Communities-the Island, South Riverdale, Moss Park, Roncesvalles

6.4. Price prediction and Hypothesis testing

We predicted the non-superhosts listings 'prices by the best model RF3.2, which had been trained on the superhost subdataset. The RMSE is 63.80, slightly higher than the reference model RF3.2 applied on the superhost subdataset

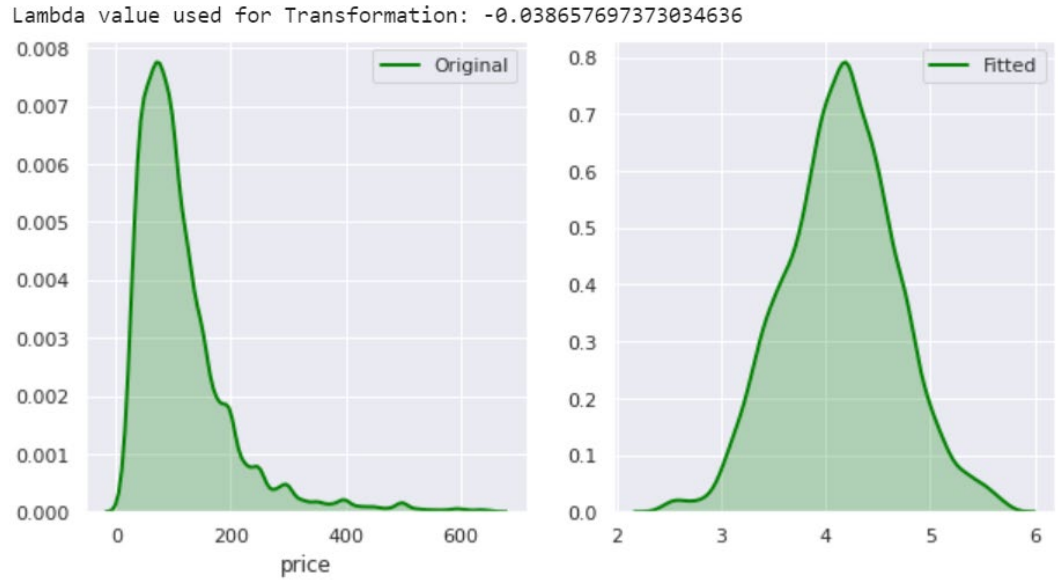
To answer our main research question about whether a significant difference between superhost and non superhost pricing exists, we will apply a non-parametric test, as the price_actual and price_predicted are not normally distributed (with Shapiro-Wilk normality test)

Mann-Whitney U Test (two-sided) suggests a null hypothesis such that the two population distributions are identical. P-value was almost 0, so we can conclude that the distributions are different. The one-sided test confirmed this result, and better leads to conclude that the actual non_superhosts price distribution is less than the predicted by superhosts

6.5. Improving the accuracy

For the sake of better results, we will use the Box-Cox transformation on the dependent variable, the new price is the fitted price

Fig9. Actual and fitted price after transformation



The new RMSE = 0.367 after fitting the price and applying the model RF3.2 on the non_superhost listings is impressively improved, however when comparing the new predicted

fitted prices and actual fitted prices, the hypothesis tests (Mann-Whitney and Kruskal-Wallis) stated that the distributions were identical

7. Conclusions

It has been confirmed that at 95% of confidence, superhosts would apply higher prices than superhosts if they were to offer the same listings. One interpretation is that they utilize efficiently their market knowledge and experience to apply higher prices and generate more income. All else being equal, non_superhosts may either devote more time and energy to get more reviews and higher ratings to increase their prices, or they can add some amenities that guests value the most, like a dishwasher or a paid-parking option

Nevertheless, this work suffer some limitations. First, the available features can be limited and fail to explain an important proportion of the price variance. When applying PCA on the whole dataset (778 attributes), the explained_variance_ratio reaches 90% when the number of attributes attains 430.

Second limitation, the price is static. The results would be more realistic if the price in the dataset was dynamic and evolves in time to deal the seasonality aspect of renting on Airbnb.

Finally, the models were trained on less data (superhost listings) than the test set (non_superhost). This may bias the results.