Determinants of number of Monthly Reviews for Airbnb Listings: Evidence From New York City

Abstract

The main goal of this Airbnb project is to find out the association between the average monthly reviews and characteristics that are available in a public dataset. These characteristics include aspects such as location, borough, and price. Based on the attributes of the dataset, three popular machine learning algorithms - logistic regression, decision tree, and random decision forest were chosen for this project. This paper contains a systematic approach at each step starting from the selection of decision variables, to the fitting process to the resultant model formations. Out of three chosen models, the decision tree method achieved the highest predictive accuracy with decent interpretability. From cross referencing across all the methods and after conducting in-depth analysis we can conclude that title of the listing, availability, price, and minimum nights required are the most important factors in predicting a better monthly review performance.

Key words: Logistic regression, Decision Tree, Random Decision Forest

Contents

1. Introduction 3
3. Data Exploration 3
4. Data Mining Methodologies 5
5. Conclusions and Recommendations 12
6. Limitations 12

1. Introduction

Since 2008, Airbnb has become one of the most popular accommodation choices for travelers around the world. This report will focus on a public dataset from the Airbnb's in New York City and will aim to discover if there are any patterns in the data. This dataset contains 16 variables including both categorical and numeric variables with approximately 40,000 data points. However, not all variables are usable in analysis. After detailed exploration, we determined the key variables which would be relevant to an analysis and decided to drop the unique variables. For data points, we excluded cases with missing information or illegible data; lastly, we also dropped a handful of outliers that were outside the three standard deviation range.

The final list of variables that were used for analysis are listed below with a short description:

Table 1 Variable definition

Variable	Definition
Neighborhood group	Borough in New York City. ex. Bronx, Brooklyn, Manhattan.
Latitude & longitude	Location of the building
Room type	Entire home/apt, private room ad shared room
Price	Price in US dollars
Minimum nights	The required minimum nights to book the stay
Calculated host listings	Total number of apartments/rooms belonging to the same host
Availability_365	number of available days in a year

2. Data Exploration

The first step in exploring the data was plotting the correlation matrix so that we could see the relationship between our variables. As can be seen from Figure 1: there is a relatively weak positive relationship between the *number of reviews* and *availability* **(0.3)**, *reviews per month* and *availability* **(0.39)**, *calculated_host_listings_count* and *availability* **(0.37)**. After plotting numerous combinations, we found that the two plots in Figure 2 are most relevant for our research question.

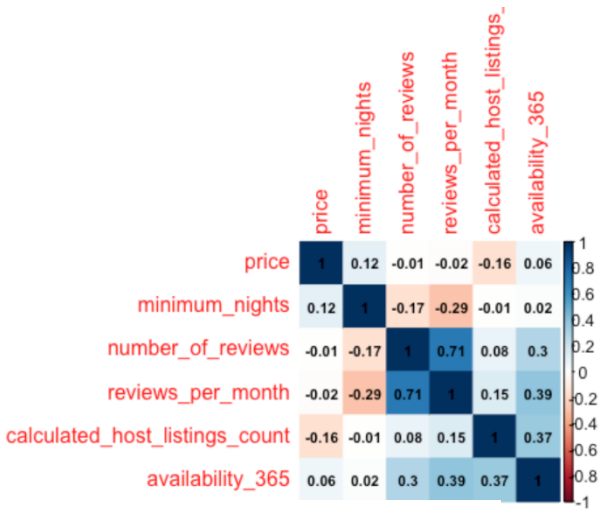

Figure1: Correlation Matrix

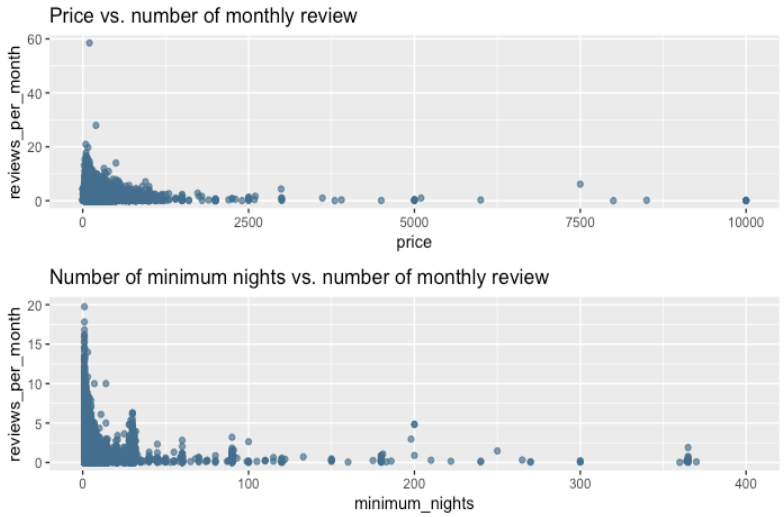

Figure 2 Scatter plot of main variables

From the plots above, we can clearly observe an inverse relationship between the number of monthly reviews and price variables; that is, the hosts tend to receive a higher number of reviews when the price is lower. The same pattern is also observed in Figure 2 such that the number of reviews increase as the number of minimum nights requested decreases.

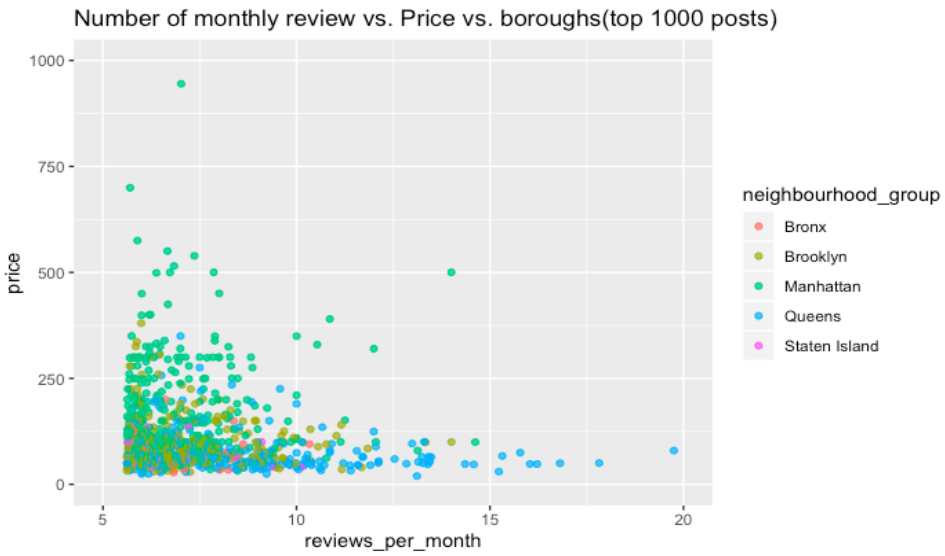

Figure 3 Monthly review vs. price, breakdown by neighborhoods

The next step in our analysis was to determine if this phenomenon was replicated when we broke the data down by neighborhood (borough). We found that lower price listings which received a higher number of reviews were mostly in Queens; whereas, listings in Manhattan were more expensive and therefore, had fewer reviews. This is in accordance with the previous observations for the entire city and in the geographical landscape of New York itself.

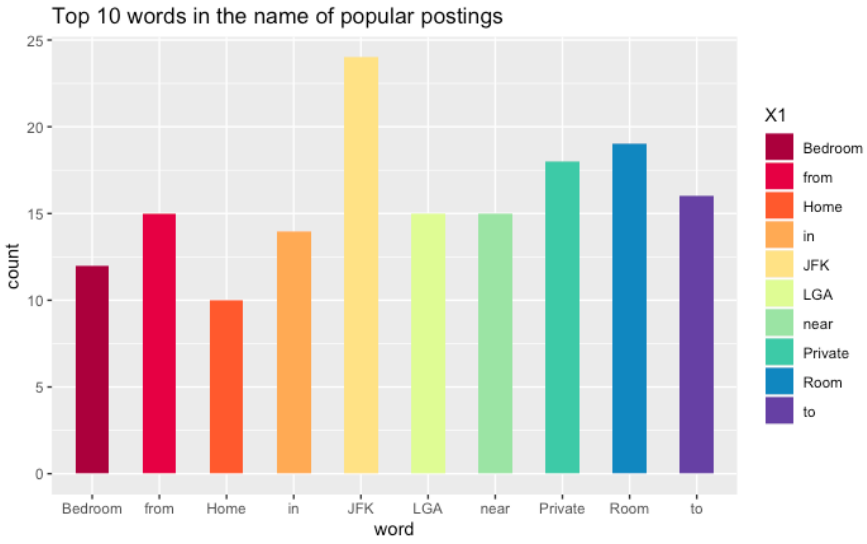

Figure 4 Top 10 words in the name of popular postings

Our next step was to analyze the words that were used in the title of the most popular posts and discover commonalities. In Figure 4 we found that the words - near, private, JFK, LGA were the most frequently occurring in our dataset. It's worth noting that "JFK" and "LGA" are the airport codes for two of the big New York airports. The takeaway from this exercise was that it seemed users preferred being near an airport or a landmark for convenience and also preferred privacy when it came to accommodation.

After careful consideration, we finalized the research question for this project as –

"**What factors contribute to a listing having higher monthly reviews?"**

In the course of this paper, we aim to give insights to the Airbnb market in NYC and provide recommendations for hosts. To accurately determine these factors and provide robust recommendations, we decided to use three different data mining approaches - logistic regression, decision trees, and random forest. This project would add value to the literature by focusing on the determinants of Airbnb listing's monthly review.

4. Data Mining Methodologies

There are myriad data mining techniques and methodologies that exist. However, selecting the correct technique to apply is almost as important as the model itself. After much deliberation, we

chose three techniques with different strengths that could provide insights on our dataset; namely, Logistic Regression, Decision Trees and Random Forest methods.

4.1 Logistic Regression

Logistic regression is a form of the regression-based analysis in which the dependent variable is dichotomous. It is often used to find the relationship between the dependent (binary) variable and other independent variables. To execute this method, we created a new dummy variable based on the original response variable: reviews per month. In our setup, the dummy variable would compare the number of reviews for an instance with the mean of the number of reviews and would be assigned a value of '0' if it was lower than the mean or '1' if the value of the instance was equal to or greater than the mean of the number of reviews. The data was split via the 80-20 method with 80% being used for training. The summary for the regression can be found below:

```
glm(formula = top ~ neighbourhood_group + room_type + price +
    minimum_nights + calculated_host_listings_count + availability_365,
    family = binomial, data = Def_T)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4965  -1.1313  -0.3178   1.1318   7.3073

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     0.9549718  0.0970565   9.839  < 2e-16 ***
neighbourhood_groupBrooklyn    -0.1510067  0.0903781  -1.671   0.0948 .
neighbourhood_groupManhattan   -0.0505379  0.0918769  -0.550   0.5823
neighbourhood_groupQueens       0.0427661  0.0956720   0.447   0.6549
neighbourhood_groupStaten Island -0.1210391  0.1616390  -0.749   0.4540
room_typePrivate room          -0.3951577  0.0370501 -10.665  < 2e-16 ***
room_typeShared room           -0.6455145  0.1056251  -6.111 9.88e-10 ***
price                          -0.0030113  0.0002081 -14.471  < 2e-16 ***
minimum_nights                 -0.1076780  0.0034378 -31.322  < 2e-16 ***
calculated_host_listings_count -0.0226128  0.0036098  -6.264 3.74e-10 ***
availability_365                0.0002595  0.0001250   2.076   0.0379 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5 Summary for the logistic regression

We can observe that the *price, minimum_nights* and *calculated_host_listings* variables are significant at 99% confidence level, and the negative coefficients show a negative relationship. That is as *price* and *# of minimum nights* increase there is a high likelihood that the number of reviews will

decrease.

Table 2 Logit Confusion Matrix

	Predicted 0	1
True 0	1419	1309
1	476	1968

The Logit Classification accuracy on the test set was **0.655** with the AUC score being **0.663**.

4.2 Decision Tree & Random Decision Forest

Reason for selection and Model Assumption

This method is a particularly good choice for this dataset since most of our explanatory variables have very skewed and non-normal distributions. Decision trees are extremely adept at classifying this type of data since it doesn't require normal distribution on the input variables. It also is extremely intuitive and is much easier to decipher than black-box methods such as SVM (James et al., 2013). It easily showcases the classification of the data and highlights the importance of each feature. There is a tendency for it to over fit the data but that can easily be overcome by applying k-fold validation, tree pruning, and random forest.

Model Formulation

The independent variables are price, minimum nights, host listings count, availability in a year, neighborhood group (borough) and room type. The dependent variable is the dummy variable that determines whether monthly rating is above the average (0: below,1: above).

1. Decide the training and test proportion

Table 3 The training and testing proportion

Training: Test	0.7: 0 .3	0.75: 0.25	0.8: 0.2
Accuracy (training set)	0.682	0.682	0.682
Accuracy (test set)	0.678	0.678	0.68

Before we fit our data into the decision tree classifier, we tested three different proportions to see which gave us the highest accuracy. Since it had the highest accuracy, we decided to use the 80-20 proportion. Therefore, our number of instances will be 20687 for training & 5172 for test. Figure 6 depicts the distribution of our dataset and we can see that the numbers between the two classes are quite balanced, this proportion will constitute the baseline to check for predictive power later.

Figure 6 Target distribution

2. Tree Pruning

Table 4 Accuracy on training and test set

Accuracy on training set (full depth tree)	Accuracy on test set (full depth tree)
0.994	0.648

When we first grow the full tree without pruning, the result is over fitted on the training set as expected. To circumvent this, we experimented with different maximum depths for our dataset and found that our tree achieves the highest accuracy when we grow our tree to level ten (shown on Figure 7). Therefore, we will be using ten as the maximum depth for the following analysis.

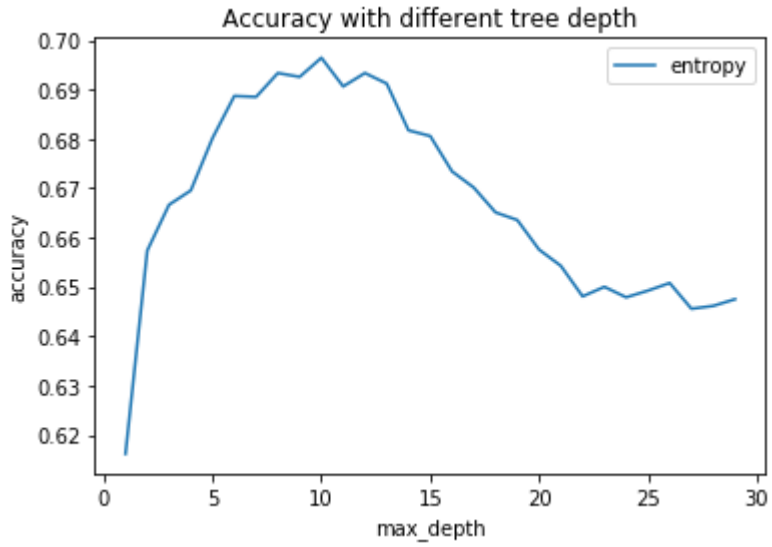

Figure 7 Accuracy with different tree depth

3. Calculating the Accuracy (confusion matrix, accuracy, ROC)

Table 5 Confusion Matrix

	Predicted 0	1
True 0	1697	1031
1	541	1903

Table 6 Accuracy on training and test set

accuracy on training set(pruned)	accuracy on test set(pruned)
0.72	0.696

Table 7 AUC score

AUC score	0.7

The confusion matrix indicates a 30.4% error rate in our prediction. That is, it has a 69.6% accuracy. Next, to determine discriminant power, we compared the accuracy to the proportion of 0 in the test set is 52.7%. This indicates that the model still has discriminant power and is reflected in our ROC score of 0.7, which is acceptable in the industry.

4. 10-fold Validation

Table 8 Tree depth and accuracy

Tree depth	Accuracy
3	0.67 (+/- 0.065)
5	0.68 (+/- 0.060)
10	0.68 (+/- 0.069)
Full tree depth	0.62 (+/- 0.062)

To further check accuracy of the test set, we applied the 10-fold validation on our dataset such that each part of the data could be assigned as the test set which would give us a more robust accuracy rate for our model. For depth 10, the result was 0.68, which is lower than our pruned 10 level decision tree's accuracy (0.696). This indicates there are some other test sets with lower accuracy in the data.

5. Random Forest (Confusion matrix, accuracy, ROC)

Table 9 Confusion Matrix

	Predicted 0	1
True 0	1908	820
1	785	1659

Table 10 Accuracy and AUC acore

Accuracy on test set	0.69
AUC score	0.689

Yet another way to prevent overfitting of data is the random forest method. This method randomly takes a given number of subtrees and grows them into full ID3 trees without pruning. By

using majority vote assumption, it can address the overfitting problem of the full depth trees. We find that our result is a little bit lower than the accuracy of the pruned decision tree (0.696).

6. Random Forest with 10-fold validation

Table 11 Average Accuracy on test set

Average Accuracy on test set	0.66 (+/- 0.06)

When we conduct the random forest approach with 10-fold variation we find that it has a slightly lower accuracy than before the k-fold validation. This is expected since this number is the average number from all the sub folds, and there may be other folds with lower accuracy in the dataset.

7. ROC curve

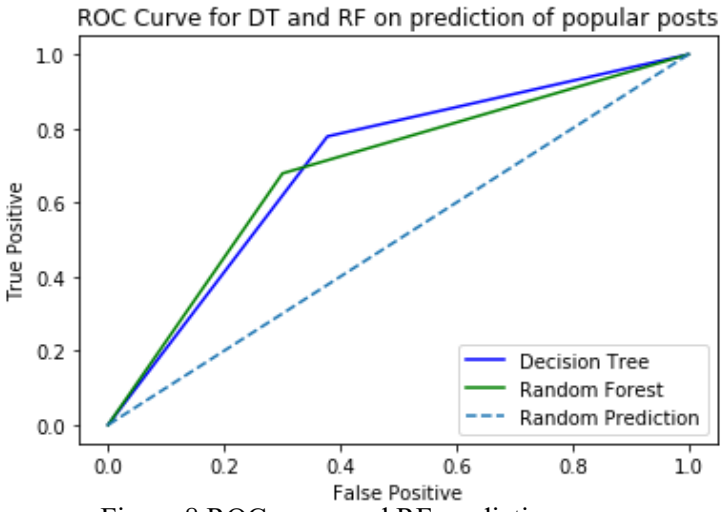

Figure 8 ROC curve and RF prediction

Table 12 AUC scores

Decision Tree AUC Score	0.7
Random Forest AUC Score	0.689

Lastly, we plotted the ROC curve for both the Pruned decision tree and random forest methods. The decision tree method scored a slightly higher AUC score. This might be the result of random selection of features in the random forest. Let us consider the importance of each feature in the decision process.

8. Interpretation for Decision Tree

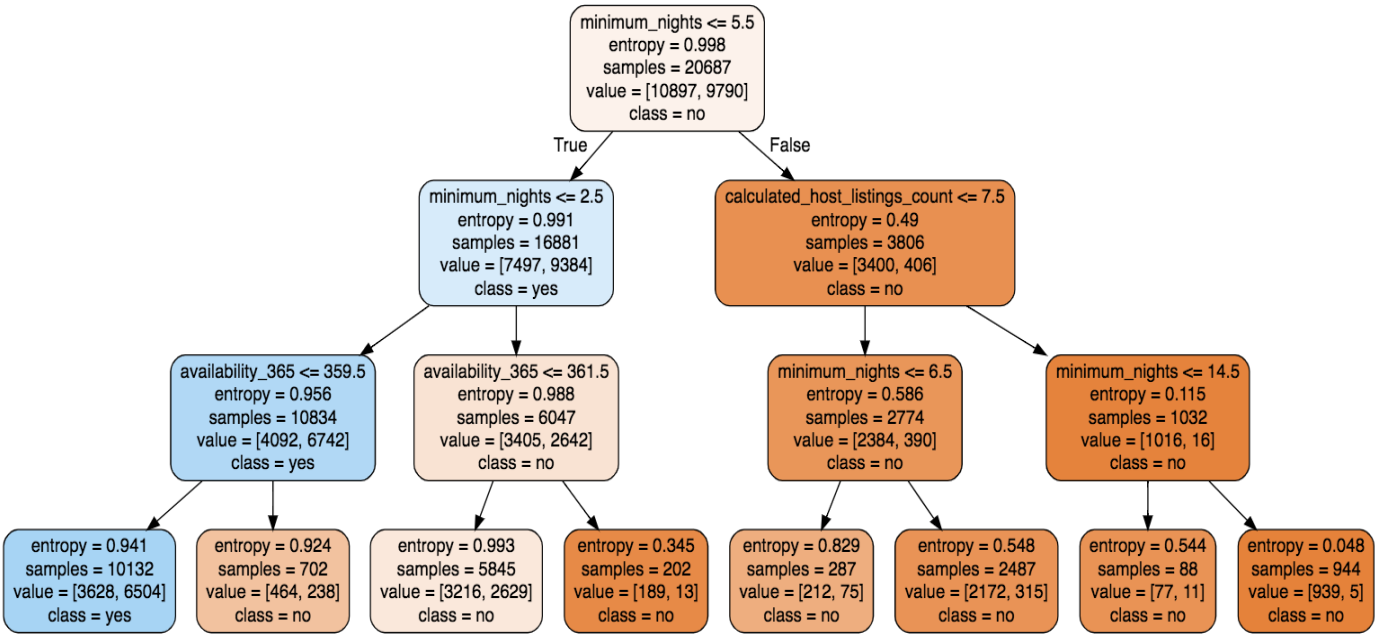

Figure 9 3-level pruned tree

The above tree plot is a 3-level pruned tree and really highlights the ease of interpretation for the decision tree approach. We can easily determine that rooms that have required minimum nights less or equal to 2.5 and have availability lower than 359.5, have a higher possibility of acquiring more monthly reviews. Due to our limited space, we are not able to showcase the full 10-level tree that was our result. However, it can be viewed here if needed.

Finally, Table 13 displays the level of importance of each variable for both the methods. We can see that for both approaches the minimum nights, price, and availability seem to be the most important factors; whereas, room type and neighborhood are less likely to affect the number of monthly reviews. We can also see the difference of feature importance between two methodologies in this section. For instance, while minimum night has the highest importance in the decision tree, it is only treated as the third-highest feature in the random forest method. This might also be the reason our decision tree outperforms the random forest.

Table 13 The level of importance of variables for decision tree and random forest methods

Importance of Features	Decision Tree	Random Forest
price	0.124	0.309
minimum_nights	0.516	0.157
calculated_host_listings_count	0.079	0.083
availability_365	0.223	0.418
neighbourhood_group_Bronx	0.001	0.003
neighbourhood_group_Brooklyn	0.01	0.006
neighbourhood_group_Manhattan	0.013	0.006
neighbourhood_group_Queens	0.006	0.004
neighbourhood_group_Staten Island	0.001	0.002
room_type_Entire home/apt	0.021	0.005
room_type_Private room	0.004	0.005
room_type_Shared room	0.002	0.002

5. Conclusions and Recommendations

From the results of all three models - it emerges that price, minimum nights and availability are the three key factors that contribute towards having higher reviews. As price and the number of minimum nights goes down, the number of reviews goes up; whereas, greater availability leads to higher reviews. Therefore, if hosts aim to attract more guests and have higher reviews then we would suggest setting slightly lower room prices, have lower minimum nights required as well as having greater availability throughout the year.

In terms of model performance, the decision tree method outperformed the other two (having the highest AUC) and was followed by random forest and then logistic regression. As we highlighted above, a reason for Decision tree outperforming could be due to the difference in weightage given to important features, such as minimum nights.

6. Limitations

No dataset or model is perfect since it is a collation of information and by its very nature it's bound to have some limitations. For this dataset, the limitations are:

1) There can be unobserved factors, which also affects monthly reviews, but we cannot put them into the model. There can also be reverse causality, for instance, the monthly review may drive the price to go higher. We do not claim there is a causal relationship, we just say that price, minimum nights and availability are associated with reviews.

2) There is a lack of information about the quality of the reviews. Further data points on 'star rating', host location, etc. would have added significantly to the model.
3) For logistic regression the interpretation is not straightforward as the independent variables affect the dependent variable in the form of log odds. That is, while we may be able to predict the "probability" of an event occurring, due to the nature of log odds, it becomes harder to directly predict effect.
4) The limited nature of the dataset also affects the separation power of the decision tree, making it less likely to produce a perfect model.
5) The difference of feature importance influences the result of random forest, as the subtrees were randomly chosen. There is the possibility that the features with less importance were picked more times than those with higher impact.

References:

Becerra, Manuel, Juan Santaló, and Rosario Silva. 'Being Better vs. Being Different: Differentiation, Competition, and Pricing Strategies in the Spanish Hotel Industry'. *Tourism Management* 34 (February 2013): 71–79. https://doi.org/10.1016/j.tourman.2012.03.014.

Chen, C.-F., & Rothschild, R. (2010). An Application of Hedonic Pricing Analysis to the Case of Hotel Rooms in Taipei. Tourism Economics, 16(3), 685–694. https://doi.org/10.5367/000000010792278310

Hua, N., Khaldoon "Khal" Nusair, & Upneja, A. (2012). Financial characteristics and outperformance. International Journal of Contemporary Hospitality Management, 24(4), 574-593. http://dx.doi.org.ezproxy.is.ed.ac.uk/10.1108/09596111211226833

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An Introduction To Statistical Learning.

Kalehbasti, Pouya Rezazadeh, Liubov Nikolenko, and Hoormazd Rezaei. 'Airbnb Price Prediction Using Machine Learning and Sentiment Analysis'. *ArXiv:1907.12665 [Cs, Stat]*, 29 July 2019. http://arxiv.org/abs/1907.12665.

Ki, S., & Jang, S. (Shawn). (2011). Room Rates of U.S. Airport Hotels: Examining the Dual Effects

of Proximities. Journal of Travel Research, 50(2), 186–197. https://doi.org/10.1177/0047287510362778

Ma, Yixuan, Zhenji Zhang, Alexander Ihler, and Baoxiang Pan. 'Estimating Warehouse Rental Price Using Machine Learning Techniques'. *International Journal of Computers Communications & Control* 13, no. 2 (13 April 2018): 235–50. https://doi.org/10.15837/ijccc.2018.2.3034.

Marchenko, Anya. 'The Impact of Host Race and Gender on Prices on Airbnb'. *Journal of Housing Economics* 46 (December 2019): 101635. https://doi.org/10.1016/j.jhe.2019.101635.

Monty, B., & Skidmore, M. (2003). Hedonic Pricing and Willingness to Pay for Bed and Breakfast Amenities in Southeast Wisconsin. Journal of Travel Research, 42(2), 195–199. https://doi.org/10.1177/0047287503257500Oskam, Jeroen, and Albert Boswijk. 'Airbnb: The Future of Networked Hospitality Businesses'. *Journal of Tourism Futures* 2, no. 1 (14 March 2016): 22–42. https://doi.org/10.1108/JTF-11-2015-0048.

Wang, Dan, and Juan L. Nicolau. 'Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.Com'. *International Journal of Hospitality Management* 62 (April 2017): 120–31. https://doi.org/10.1016/j.ijhm.2016.12.007.

Yang, J. (2012). Identifying the attributes of blue ocean strategies in hospitality. International Journal of Contemporary Hospitality Management, 24(5), 701-720. http://dx.doi.org.ezproxy.is.ed.ac.uk/10.1108/09596111211237255

Y. Li, Q. Pan, T. Yang and L. Guo, "Reasonable price recommendation on Airbnb using Multi-Scale clustering," 2016 35th Chinese Control Conference (CCC), Chengdu, 2016, pp. 7038-7041.

Zervas, Georgios, Davide Proserpio, and John W. Byers. 'The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry'. *Journal of Marketing Research* 54, no. 5 (October 2017): 687–705. https://doi.org/10.1509/jmr.15.0204.