4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15091

Understanding the effects of Covid-19 on P2P hospitality: Comparative classification analysis for Airbnb-Barcelona

Juan Pablo Argente del Castillo Martínez, Isabel P. Albaladejo
Universidad de Murcia, Spain

Abstract

The Covid-19 pandemic has produced significant changes in tourism markets around the world. The large amount of data available on the Airbnb platform, one of the world's largest hosting services, makes it an ideal prospecting place to try to find out what the aftermath of this event has been.

This paper explores the entire Airbnb housing stock in the city of Barcelona with the aim of identifying the key characteristics of the homes that have remained operational during the 2019-2021 period. We carried out this analysis by using two classification methods, the random forest and logistic regression with elastic net. The objective is to classify the houses that have remained on the platform against those that have not. Finally, we analyze the results obtained and compare both the general performance of the models and the individual information of each variable through partial dependence plots (PDP).

We found a better performance of Random Forest over logistic regression, but not significant differences in the relevant variables chosen by each method. It is worth noting the importance of the geographical location, the number of amenities in the home or the price in the survival of the homes.

Keywords: *Airbnb; Covid-19; Survivability; Random-Forest; Logistic-Regression .*

1. Introduction

The Covid-19 pandemic has caused all kinds of effects in the different affected markets and economies, especially in tourism. Spain, and more specifically the city of Barcelona, has notably suffered the consequences of the different policies that have been carried out to control the devastating effects of the virus. According to the Statistics of tourist movements at the border of the Spanish Statistics Institute (INE), the number of international tourists arriving in Cataluña has dropped from 19.4 millions in 2019 to 5.7 millions in 2021. This fall in arrival of turists has had non-zero effects in hosting services.

Currently, one of the most popular accommodation services in the city of Barcelona is the Airbnb platform. It brings together both professionals from the world of hospitality and non-professionals, offering a range of different solutions far superior to conventional media. This fact, together with the greater competitiveness in prices, has made the platform become the reference for the vast majority of tourists who visit the city (Gutierrez J. et al, 2017).

In this paper, we analyze the characteristics that best define the group of dwellings that has remained available during 2019-2021. This is a matter of importance for the owner in terms of achieving regular income in the long term (Lladòs-Masllorens et al., 2020). It is also a key factor for local managers in tourism-dependent economies (Wachsmuth, D., & Weisler, A. 2018), one of the most external-shock dependent, like the city of Barcelona. Proof of this are the scenarios that both covid-19 and 2008 crises left behind. We use datasets from Airbnb which contain the different characteristics of the dwellings of Barcelona and correspond to the month of November for both 2019 and 2021.We decided to use these timestamps because in November 2019 the pandemic had not yet started and in November 2021 the process of suspending anti-covid measures in Spain had already begun, along with the end of travel bans to other countries and the end of the second round of vaccinations for the majority of the population. In addition the reason for choosing the month of November is to avoid the seasonality effects of tourism data, since Barcelona, along with most tourist destinations in the world, suffers an increase in tourists in the summer months.

To find the characteristics that best define the "dwelling-survivavility" we have carried out two classification methods with the same set of variables: random-forest and logistic regression with elastic-net regularization. We have found evidence that the most determining variables have been the geographical location, the price, the experience of the host and the level of equipment of the dwelling.

2. Data Handling

Both datasets from November 2019 and November 2021 were collected from Inside Airbnb, an online service that provides these datasets for different cities around the world. It has been used in similar studies (Gibbs et al. 2018) due to the large amount of data it provides and the ease with which it can be obtained.

After a minimal cleaning of the datasets, with the aim of maintaining as much of the sample as possible, the datasets decreased to 12,337 dwellings in 2019 and 9,540 in 2021. Furthermore different operations were carried out in order to operationalize the dataset. We did 'one hot encoding' of categorical variables, character count of variables that are strings, or count of items in lists, wich is the case of the amenities.

Once the datasets were cleaned separately, we created the dummy variable that will be used as the dependent variable, old_homes. We obtained this variable by comparing the unique identifiers of the different ads, so if we happened to find the same identifier in 2019 and 2021, we will add a 1 to the variable and 0 in the event that this assignment does not occur. This variable allowed us to find that only 4,366 ads were shared between both moments of time, so 8,001 homes were lost since 2019. Only 5,174 have been recovered as of November 2021. We can observe the spatial distribution of the 2021 dataset in Figure 1. The red dots indicate the three main tourist points of the city, from top to bottom: Sagrada Familia, Ramblas and Puerto. In yellow we see the dwellings that were maintained throughout the period and in blue the new ones, slightly more grouped in these three neighbourhoods.

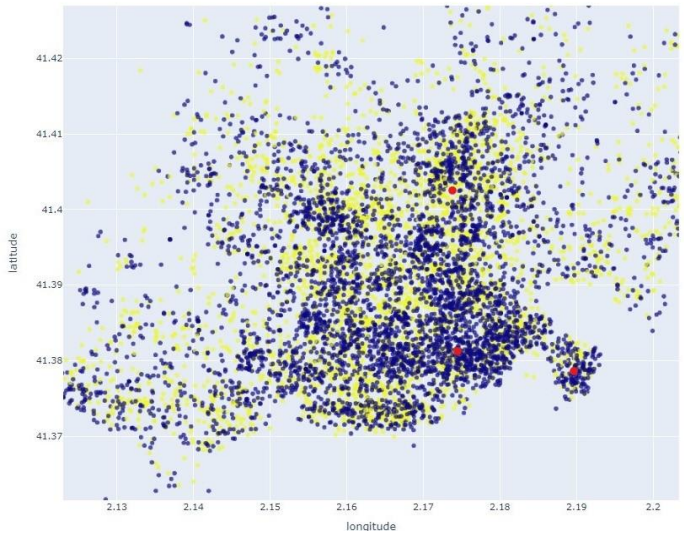

Figure 1. Spatial Distribution of 2021 Airbnb Dataset

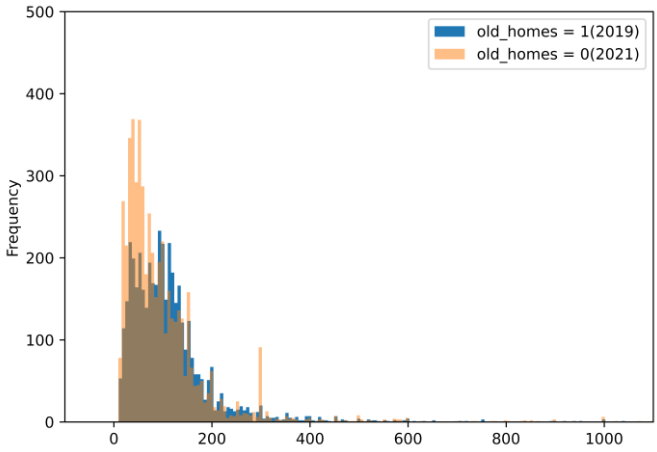

Figure 2. Price frecuency distribution of 2021 Airbnb Dataset.

Figure 2 shows the frequency distributuion of home's prices. We can also find certain differences, the most important being the predominance of the cheapest prices in the newer houses, and a more uniform distribution in the houses that remained during the entire period. One last aspect to highlight is the professionalism of the hosts. This quality is usually measured in the literature (Lladós-Masllorens et al, 2020) by the total number of ads that the host has on the platform. This variable is known as calculated_host_listings_count and has had significant variations from 2019 to 2021, increasing the average level in the period, as we can see in figure 3.

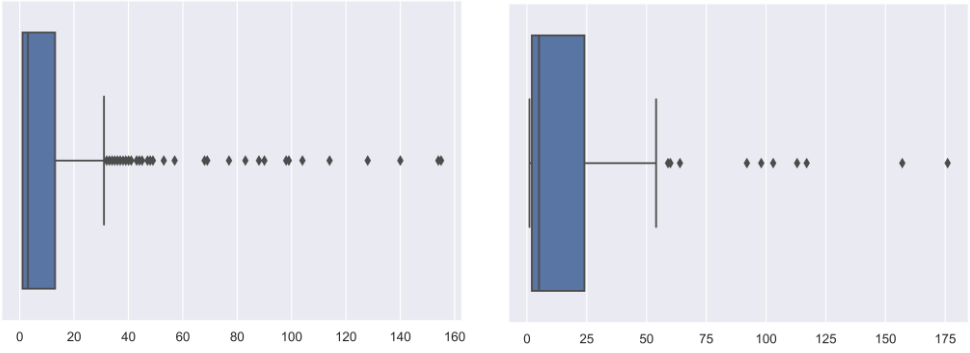

Figure 3. Box plots of calculated_host_listings_count. (left 2019 and right 2021).

All the variables that were provided with the original datasets were used, with the exception of those who implied somehow temporality (f.e. first_review, last review, …) or were descriptions that had little to no relationship with the analysis (f.e. host_neighbourhood, …). In total, 72 variables were used.

3. Modeling the survivability

We have used two different aproaches to understand the differences between those dwellings that were kept on the platform for the whole period, wich we refer as old houses, and those who did not, wich we refer as new houses: Random Forest Classifier and Logistic Regression with Elastic Net. To assure we do not overfit the samples, two different datasets were established within the sample, a training set with 80% of the values and a test set with the remaining 20%.

3.1 Random Forest Classifier

The hyperparameters for the Random Forest (Brieman, 2001), wich were optimized by Grid Search, are the following: 760 trees, the square root of the total available variables as the maximum number of variables to consider in each tree, and an execution without Bootstrap. It should be noted that entropy has been used as a criterion to measure the quality of the sample divisions and not the Gini impurity.

3.2 Logistic Regression with elastic net classifier

Also for the Logistic Regresssion a Grid Search was carried out to obtain the hyperparameters, these being: elastic net penalty function with an l1/l2 ratio of 50%, 1000 maximum iterations and the SAGA algorithm (Defacio, Bach, Lacoste- Julien, 2014) for the optimization of the problem.

4. Results

Table 1 shows the classification output of both methods and Table 4 the confussion matrix. As can be seen in both, random forest presents a precision 15% higher than the logistic regression with elastic net. According to random forest the five most important variables to define survivability are latitude, longitude, price, the amount of amenities and the professionalism of the host. It is important to notice that these variables are relevant in both methods but only the random forest allows us to see the non-linearities between them.

To analize the effects of these variables we use the partial dependence plots (Friedman, 2001). Each point of these plots indicate how many of the trees (in average) that random forest builds have been classified for the class "old_homes" across all observations, given a fixed level of the variable we're looking at. Figure 6 indicates these partial dependence for the obtained key variables.

Table 1. Random Forest Classifier and Logistic Regression Output

	Variable	Precision	Recall	F1-Score	Support
RF	0	0.78	0.69	0.73	658
	1	0.78	0.85	0.82	872
Log_e-n	0	0.65	0.31	0.42	658
	1	0.63	0.87	0.73	872

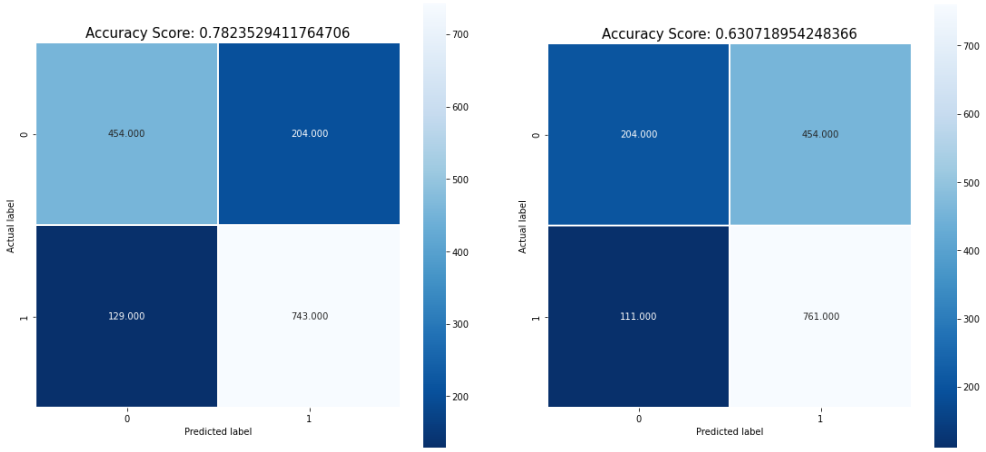

Figure 4. Confusion matrix of Random Forest Classifier(left) and Logistic Regression-elastic net(right).

In the first place, the partial dependencies of latitude and longitude indicate a greater tendency to predict a "new" house in the geographical points (41.38,2.175), wich coincides with Las Ramblas point. This shows that the houses in the pool with the best geographical location, in terms of restrictions for the covid-19 pandemic, have managed to remain active during this time. Secondly, we can relate higher prices to houses that have remained during the pandemic. The bulk of the new houses are located in the cheapest levels, which normally correspond to private rooms in shared houses or collaborative housing solutions, which were more likely to close in the evaluated period. In relation to the amenities of the properties, we can infer that fewer amenities are associated with lower probabilities, that is, with "new" homes, while the best-equipped homes are those that have achieved endure. Finally, the professionalism of the owner of the home, measured by the total number of homes owned by the host (calculated_host_listings_count), provides information that is only valid in the lower ranges of the variable, since that is where the majority of the sample is found. In this case, a decrease in probability is observed as professionalism increases,

indicating that owners with a greater number of homes have withdrawn their homes from the market, unlike the majority of small owners, who have been more open to staying.

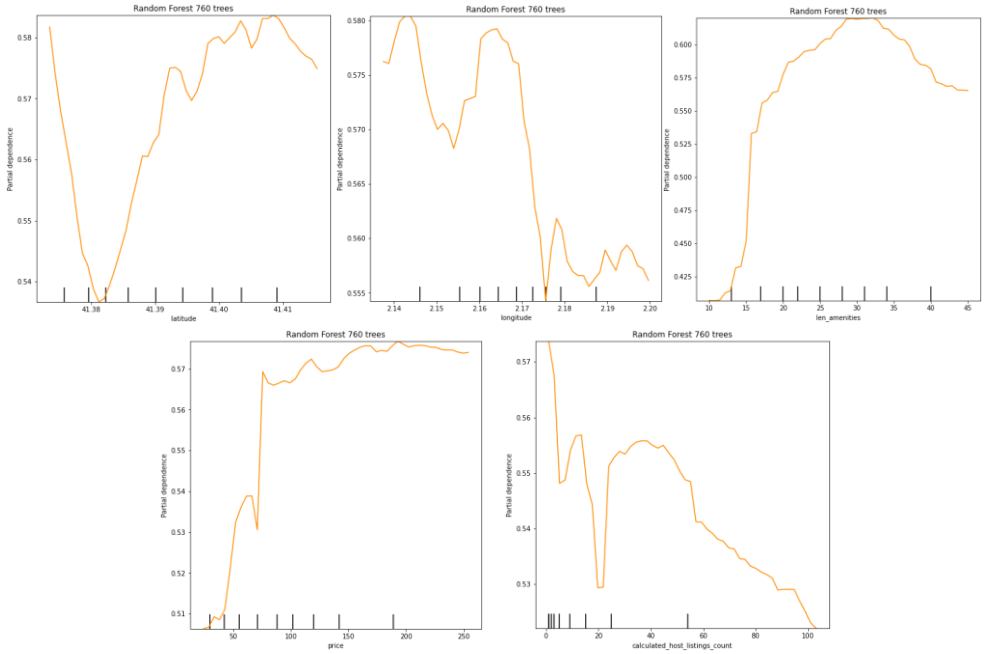

Figure5. Partial Dependence Plots of most important features according to Ranfom-Forest.

4. Conclusion and Prospects

The classification proposed in this study of the Airbnb accommodations of Barcelona allows us to conclude that the main variables affecting their survaivability are those that also define quite well the prices on hedonic models that are generally used to determine the factors defining the prices of this kind of accommodations. (Casamatta et al. 2022). In particular, the most touristic neighborhoods and the most densely populated with housing are going to be what suffer the most during non-normal stages. We can also infer that professionalism plays a key role in survival, and that although it normally implies better decision-making, in the face of this type of event it also implies a certain level of vulnerability.

The results of the study allow us to intuit that the robustness of the market, understanding it as the survivability rate of the dwellings on such difficult periods, could be improved by regulating property in order to avoid professional proprietaries. This regulation would also help to avoid gentrification that most tourist cities suffer, as we have seen in Barcelona in

Las Ramblas neighborhood. These densely populated areas are more prone to develop big clusters of accommodations that are not based in shared economy but in hospitality firms.

In future works we will explore more specific effects of these attributes, mediated by macroeconomic variables, as well as also analyze the effect on robustness of a pure shared economy model. New techniques that allow deeper analysis, such as neural networks and the use of natural language processing, are also interesting to complete the analyses.

References

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Casamatta, G., Giannoni, S., Brunstein, D., & Jouve, J. (2022). Host type and pricing on Airbnb: Seasonality and perceived market power. *Tourism Management*, 88, 104433.

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing*, *35*(1), 46-56.

Gutiérrez, J., García-Palomares, J. C., Romanillos, G., & Salas-Olmedo, M. H. (2017). The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona. *Tourism management*, 62, 278-291.

Lladós-Masllorens, J., Meseguer-Artola, A., & Rodríguez-Ardura, I. (2020). Understanding peer-to-peer, two-sided digital marketplaces: pricing lessons from Airbnb in Barcelona. *Sustainability*, 12(13), 5229.

Wachsmuth, D., & Weisler, A. (2018). Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and Planning A: Economy and Space*, 50(6), 1147-1170.