G VICEPRESIDÈNCIA
O I CONSELLERIA
I INNOVACIÓ,
B RECERCA I TURISME
/ AGÈNCIA TURISME
ILLES BALEARS

Universitat de les Illes Balears

una manera de hacer europa

Unión Europea
Fondo Europeo de Desarrollo Regional

Report on the Airbnb rental market for the Balearic Islands

Technical report

William Nilsson
Department of Applied Economics
University of the Balearic Islands

Veronica Leoni
Department of Applied Economics
University of the Balearic Islands

Paolo Figini
Department of Economics
University of Bologna

This is a technical report for the project:
Econometric Analysis of the rental market of apartments for tourists in the Balearic Islands.
(Estudi Econòmic de la Vivienda vacacional)

Contact information: **william.nilsson@uib.es, 971 17 13 77**

Purpose and models

From the descriptive analysis we found that about 71% of the active listings on the Balearic Islands had at least one reservation day in August 2016. Accordingly, even in high season there are no guarantees that offering an apartment/room into the market in Airbnb means that tourists will be accommodated there. In this section we evaluate the factors that are important for having at least one reservation day in August. In the same way we also study the probability to dropout from Airbnb during the either July or August, i.e. during high season. Of course, we cannot know if this actually means that the host withdraws from trying to rent the apartment/room, or if he/she simply abandoned Airbnb as a channel to find possible guests. A third focus is to see which factors that are important for having a high revenue in August, given that at least one reservation day was recorded. In each of these cases we are both interested to analyze which variables that matter, but also to see if these variables can be used to predict the outcome on a test sample that not was used to estimate the models. In a sense this reflects the uncertainty that the hosts are facing. When the outcome is binary we use Logit, Classification Tree, Random Forests and Boosting. Random Forests and Boosting are often better for prediction, but harder to interpret in terms on how important the variables are. When we analyze revenues we use Linear Regression, Regression Tree, Random Forests and Boosting. Finally we study tourist saturation in terms of tourists per capita in both the hotel sector and the Airbnb sector. We group the municipalities into clusters.

Preparation of variables

Linear regression and Logit models require preparation of variables that is not necessary in tree based models. The following local variables can also be useful in the latter models, but another alternative to capture local effects is to include longitude and latitude into the model. We created three local variables by calculating a weighted average from 0.25% of the closes observations. 0.25% corresponds to just above 50 apartments/rooms and more weight was given the closer the apartment was. The tricube function was used to weight the observations. These variables are "Local Revenue – July", "Local reservation days – July" and "Local no day in July". The original variable of "no days" refers to 1 if the apartment did not have any reservation day in July. Accordingly the "Local no day in July" is a weighted average of 0 and 1, and the results can be interpreted as a proportion with no rented day.

Some variables, such as revenues, have some very extreme high values, but also a lot of 0. For these variables we apply an inverse hyperbolic sine transformation. This transformation allows maintaining zeros, while for positive values the transformation resembles a logarithmic transformation. Since "Local Revenue –July" does not include 0, we simply used a logarithmic transformation. In addition to these variables we also include a dummy variable to indicate that no day was rented, for example in July. For the variable security deposit an inverse hyperbolic sine transformation was done, and we added a dummy equal to one, given that the security deposit was used, i.e. above 0. This separation of the variables into a quantitative and qualitative part is not necessary for the tree based models.

All variables that measure an amount of money was original specified in US$. The inflation rate has been very low and the exchange rate has been fairly stable over the period and we simply used the exchange rate of 0.9EUR/$.

In the data a variable is included for the average rating that the apartment has received from previous guests. In case of not having any rating a missing value was found for the variable. Based on this information we created a qualitative variable with categories; "Never", "Low" (<4), "Moderate" [4-4.5),

"High" [4.5-5) and "Perfect" [5]. A set of dummy variables were created to be used in the linear regression model and the Logit model. The reference case was "Never". The reason that we refer <4 to "Low" is because the average rating is very high.

Topic 1: Analyzing the probability to have at least one rented day in August 2016.

In this analysis the dependent variable is binary, i.e. we only have to categories; "No" or "Yes", where the latter is having at least one day rented in August 2016. The initial analysis is made by a Logit model, and much more variables are needed to code qualitative variables into dummy variables. In models based on classification and regression trees this is not necessary. In the Logit model we also include some dummy variables to capture a particular nonlinear effect of the time since the listing was created. These possible effects can be captured automatically in models based on trees. The purpose with the Logit model is to obtain marginal effects of the included variables, while the Random Forest and Boosting are consider more competitive for making predictions.

Logit

The marginal effects from the Logit model are found in table 1. All marginal effects are evaluated at the average of the explanatory variables. The marginal effects are interpreted keeping the other variables constant. Notice that having no day rented in July also implies that both reservation days and revenues are zero. It is not possible to do a ceteris paribus interpretation for no day rented. Performance in June does not seem to matter much.

Table 1. Marginal effects on the probability that the apartment is rented at least one day.

	Marginal effects	Standard errors	z	P>|z|
No day rented in July	-0.4069***	0.0471	-8.63	0.000
Reservation days in July	0.0123***	0.0008	15.28	0.000
Ihs(Revenues in July)	-0.0209***	0.0056	-3.72	0.000
No day rented in June	0.0025	0.0461	0.05	0.956
Reservation days in June	-0.0017*	0.0009	-1.79	0.073
ihs(Revenues in June)	0.0045	0.0066	0.69	0.489
ln(published nightly rate, EUR)	-0.0110**	0.0052	-2.12	0.034
Entire home (i.e. not shared)	0.0148	0.0098	1.51	0.131
Number of bathrooms	-0.0034	0.0036	-0.94	0.347
Number of bedrooms	-0.0080*	0.0045	-1.78	0.076
Maximum guests	0.0042*	0.0023	1.85	0.065
Minimum days for reservation	-0.0016	0.0012	-1.31	0.192
Instant bookable (dummy)	0.0401***	0.0064	6.27	0.000
Security deposit (dummy)	0.0570	0.0366	1.56	0.119
ihs(Security deposit, EUR)	-0.0070	0.0056	-1.24	0.215
Created date (numeric/365)	0.0146***	0.0031	4.73	0.000
Created in July 2016 (dummy)	0.0879***	0.0082	10.77	0.000

table continues for previous page

	Marginal effects	Standard errors	z	P>|z|
Created in June 2016 (dummy)	0.0514***	0.0104	4.92	0.000
Created in May 2016 (dummy)	0.0437***	0.0116	3.75	0.000
Created in April 2016 (dummy)	0.0222*	0.0126	1.76	0.078
Rated class, Low (dummy)	0.0977***	0.0115	8.50	0.000
Rated class, Moderate	0.1117***	0.0078	14.31	0.000
Rated class, High	0.1221***	0.0073	16.74	0.000
Rated class, Perfect	0.0825***	0.0068	12.12	0.000
Ibiza (dummy)	-0.0037	0.0096	-0.38	0.704
Menorca	-0.1432***	0.0197	-7.27	0.000
Formentera	-0.0612***	0.0207	-2.96	0.003
ln(Local revenues – July)	-0.0030	0.0079	-0.38	0.707
Local no days – July	-0.0464	0.0395	-1.18	0.240
Local reservation days, July	0.0032**	0.0015	2.07	0.038
Apartments: 21654	Pseudo R2 : 0.2610			

Notes: Marginal effects are estimated based on a Logit model. ***,**,* indicate significantly different from zero at 1, 5 and 10% significance level.

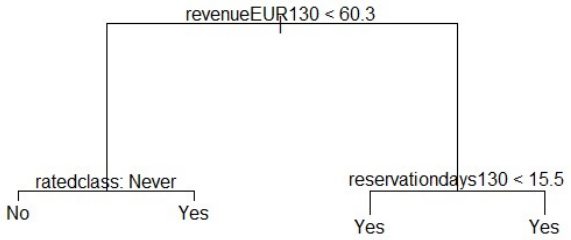
If the listing is instantly bookable the probability to have at least on rented day in August increases with about 0.04. Apartments that were introduced in the recent months before also seem to have higher probability to have at least one rented day in August. The marginal effects are also fairly large for all the classification with ratings, compared to the group without a review from previous guests. The probability is also lower for apartments in Menorca (-0.14) and Formentera (-0.06) compared to Mallorca.

Classification Tree

A single classification tree was grown based on a training data set. The most important split is based on the revenue in July, where the separation is done at 60.3 Euros. Other variables that we find in the tree, is "Rated Class", which distinguish those never being rated from different classifications, and also the number of reservation days in July.

The value on the terminal node refers to the class that is represented to the largest degree. This tree was built on a training data consisting of 15000 observations and the test data is 5654 observations. 78.0% of the test observations were correctly classified. It is important to keep in mind that 71.4% of the test sample had the apartment/room rented at least one day in August, and classifying all observations to the largest group would imply that 71.4% of the observations would be correctly classified.

Figure 1. Classification tree

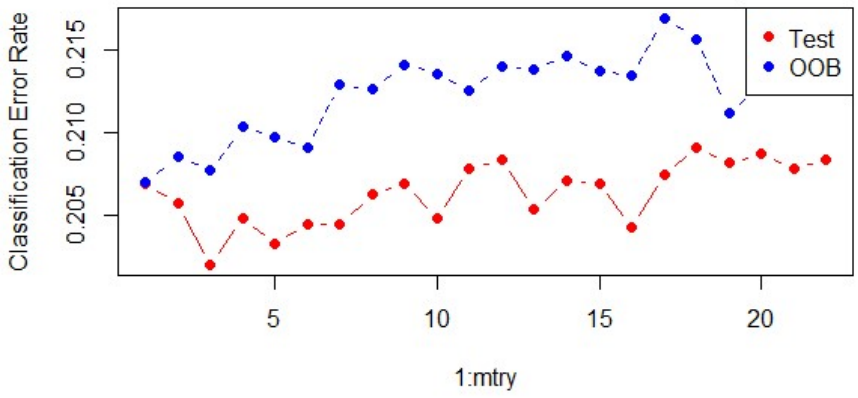


A very simple classification rule could be created based on knowing if the apartment had more than one reservation day in July. If the predicted classification for August would be that exactly the same apartments/rooms were rented as in July, this would provide that 77.0% of the observations would be correctly classified. In this perspective, the classification tree only provides a slight improvement, despite the availability of a large amount of variables.

Random Forests

Random Forests is a technique that combines many classification or regression trees into a common prediction. An important tuning parameter in Random Forests is how many of the variables that are randomly chosen to compete in each split in the tree growing process. To find the optimal number of variables to use, we repeated the model from 1 to 22, where "22" refers to a model where the best variable from all of the 22 variables is chosen in each split of the tree. This case refers to Bagging, but as we can see in graph Random Forest with fewer variables is better. The graph shows the classification error rate on both Out-Of-Bag (OOB) observations and the test data. We choose to use m=3, i.e. in each split we allow one out of three randomly chosen variables to conduct the split.

Figure 2. Classification error rate



For each split that is done in the tree growing process the decreas in the Gini index is stored for each used variable. Calculating the average decrease is accordingly a measure for how important a variable has been in the trees.

Table 2. Variable importance

Variable	Mean Decrease Gini
Revenues in July	650.1
Reservation days in July	545.4
No day rented in July	441.9
Created date (numeric)	433.2
Local reservation days – July	388.6
Longitude	378.4
Local no days – July	378.0
Published nightly rate	377.5
Latitude	377.3
Local revenues – July	359.3
Rated class	263.6
Security deposit (amount)	226.6
Revenues in June	190.6
Maximum guests	175.6
Minimum days for reservation	171.5
Reservation days in June	160.2
Number of bathrooms	149.8
Number of bedrooms	137.1
No day rented in June	67.6
Instant bookable (dummy)	58.5
Island	46.0
Entire home (i.e. not shared)	25.1

The most important variable in this model is the revenues in July and the second most important variable is the amount of days that the apartment/room was occupied in July. The third variable, “No days rented in July” is actually equal to one when zero days was rented, and these two variables contain overlapping information. It is important to understand that the table does not measure the importance of a variable for prediction, since other variables would possibly work as surrogates in case of leaving out a particular variable. Hence, dropping “No day in July” would not reduce prediction error, because reservation days in July would replace its contribution. Note that longitude and latitude can be included directly in a tree, while this is less useful in a Logit Model. These variables capture general geographic information that helps to predict the outcome. In addition to these variables we also have weighted averages for a small area close to the position of the apartments. 79.9% of the observations in the test data were correctly classified. Estimating the model with all observation and using OOB-observations to evaluate the performance to predict, means that 79.35% were correctly classified. Random Forests outperforms a single tree when it comes to predicting, but the performance is not particularly impressive. Below we show a confusion matrix based on Random Forests, using the complete data, but with OOB-observations for prediction.

Table 3. Confusion matrix

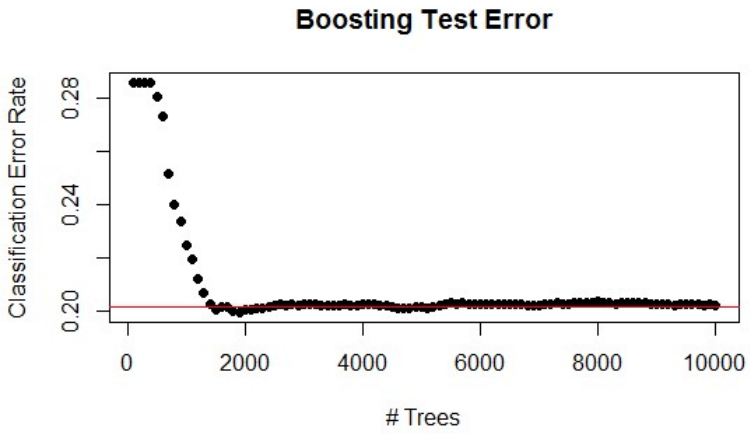
		Model predicted class:		
		No	Yes	Classification Error
Actual class:	No	3761	2461	0.3955
At least one rented day	Yes	2011	13421	0.1303

Among all apartments that did not have any reservation day, i.e. "No", the model incorrectly predicts "Yes" for almost 40% of the cases. Among the apartment that actually had at least one day rented in August, the model incorrectly classifies to "No" in 13% of the cases. In addition to the model estimated above, we eliminated all variables referring to the performance of the apartment the previous months (i.e. June and July). Using OOB-observations, 73.9% were correctly classified, which is fairly far away from the complete model. This indicates that the remaining variables are poor substitutes for performance related variables registered in the previous months, i.e. in this case, in particularly July.

Boosting

Boosting is technique that requires fine-tuning of more parameters. Initially, we specified 10000 trees and a shrinkage parameter to 0.01, but the model showed tendencies of over-fitting, so we changed the shrinkage to 0.001.

Figure 3. Classification error rate



We tried some different values of interaction depth. We finally used 8, but there are very small differences using other values. In addition, the results are only slightly better than Random Forests, given that the best range of trees was used. The best result was a model that correctly classified 80.0% of the apartments in the test data.

Table 4. Variable importance

Variable	Relative influence
Revenues in July	46.3
Reservation days in July	21.2
No day rented in July	9.6
Rated class	6.5
Created date (numeric)	6.3

table continues from previous page

Variable	Relative influence
Published nightly rate	2.2
Revenues in June	1.2
Local no days – July	1.2
Longitude	0.9
Latitude	0.8
Instant bookable (dummy)	0.8
Local reservation days – July	0.7
Island	0.4
Minimum days for reservation	0.4
Reservation days in June	0.3
Security deposit (amount)	0.3
Number of bedrooms	0.2
Maximum guests	0.2
No day rented in June	0.2
Number of bathrooms	0.2
Local revenues – July	0.1
Entire home (i.e. not shared)	0.1

Boosting identified the same top four variables as was found in Random Forests. The fifth most important variables was "Rated class", that was found less important in Random Forests.

Topic 2: Analyzing the revenues in August 2016

The dependent variable in this analysis is the logarithm of the revenues in August 2016. Only apartments/rooms with at least some revenue during the month are accordingly included. We prefer to separate the models in this way, instead of having to deal with a censoring of the dependent variable. It is, however, important to remember that it is not a random sample that is analyzed. In the same way as for the Logit model, we use the linear regression primarily to evaluate the effects of the variables, while we do not evaluate its performance to predict on a test sample. This is done for the models that we expect to perform better for that purpose.

Linear Regression

Table 5 includes coefficients for the linear regression when log of revenues in August is the dependent variable. This a selected sample where only apartments with positive revenues are included. The performance in July is clearly important. Notice that it is not possible to make a ceteris paribus interpretation of these coefficients. The reason is that a change in having no days rented in July will also mean a change in the number of reservation days and revenues.

Table 5. Linear regression, Dependent variable: ln(Revenues in August)

	Coefficients	Standard Errors	t	P>|t|
No day rented in July	1.7185***	0.0904	19.02	0.000
Reservation days in July	0.0029**	0.0014	2.07	0.038
Ihs(Revenues in July)	0.2680***	0.0126	21.20	0.000
No day rented in June	0.6015***	0.0942	6.38	0.000
Reservation days in June	0.0023	0.0017	1.40	0.161
ihs(Revenues in June)	0.0818***	0.0136	6.02	0.000
ln(published nightly rate, EUR)	0.3884***	0.0136	28.45	0.000
Entire home (i.e. not shared)	0.2186***	0.0212	10.30	0.000
Number of bathrooms	0.0407***	0.0091	4.45	0.000
Number of bedrooms	0.0232**	0.0108	2.15	0.032
Maximum guests	0.0106*	0.0054	1.97	0.049
Minimum days for reservation	-0.0096***	0.0030	-3.22	0.001
Instant bookable (dummy)	0.0162	0.0141	1.15	0.250
Security deposit (dummy)	-0.3380***	0.0863	-3.92	0.000
ihs(Security deposit, EUR)	0.0563***	0.0138	4.08	0.000
Created date (numeric/365)	0.0469***	0.0068	6.87	0.000
Created in July 2016 (dummy)	0.0098	0.0266	0.37	0.711
Created in June 2016 (dummy)	0.1068***	0.0247	4.32	0.000
Created in May 2016 (dummy)	0.0490*	0.0252	1.94	0.052
Created in April 2016 (dummy)	0.0348	0.0264	1.32	0.188
Rated class, Low (dummy)	0.1172***	0.0304	3.86	0.000
Rated class, Moderate	0.0847***	0.0204	4.15	0.000
Rated class, Hight	0.1072***	0.0180	5.96	0.000
Rated class, Perfect	0.0639***	0.0190	3.37	0.001
Ibiza (dummy)	0.2716***	0.0218	12.46	0.000
Menorca	-0.0196	0.0522	-0.38	0.707
Formentera	0.2973***	0.0466	6.37	0.000
ln(Local revenues – July)	0.0826***	0.0191	4.33	0.000
Local no days – July	-0.0484	0.0939	-0.52	0.607
Local reservation days, July	-0.0108***	0.0036	-3.02	0.003
Constant	-0.6000	0.4068	-1.47	0.140
Apartments: 16373	R2-adj:	0.4685		

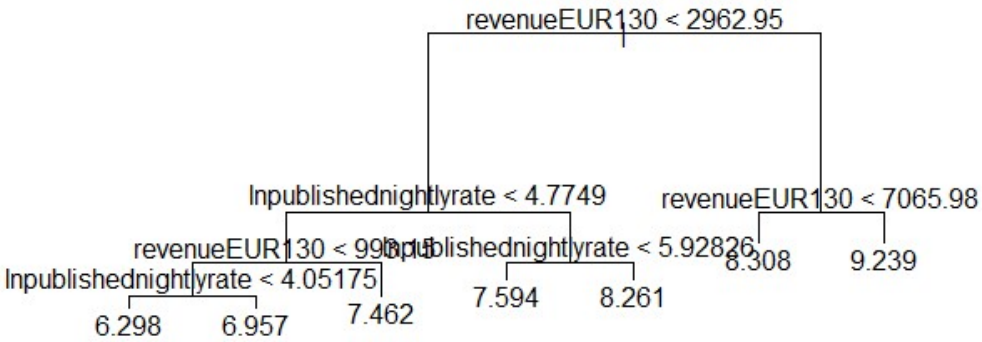
Notes: Regression only includes apartment with at least one rented day in August. ***,**,* indicate significantly different from zero at 1, 5 and 10% significance level.

It is interesting to see that hosts that require longer minimum stay pay a price in terms of lower revenues. Another interesting effect is that having received reviews from previous guests are important for the revenues, but the effect is very similar for different ratings.

The revenues are also about 31.2%(=100*exp(0.2716)) and 34.6% higher in Ibiza respective Formentera compared to Mallorca. An increase in 1% of the published nightly rate would increase the revenues with about 0.39%.

Regression Tree

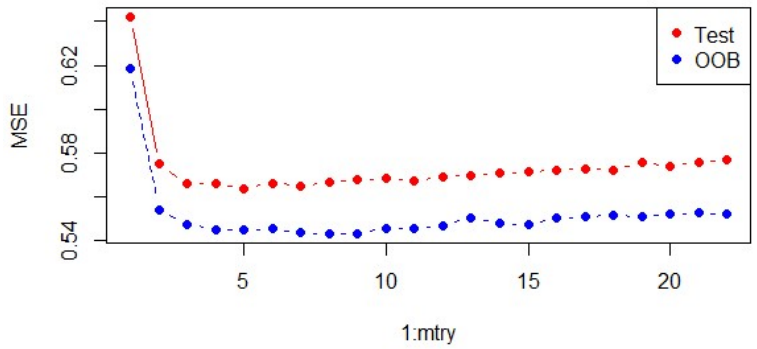
Figure 4. Regression tree



The most important variable is revenues in July, and the split is done at about 2960€. In the tree, log of published nightly rate is also influencing the log revenues in August. Notice that a logarithmic transformation of explanatory variables does not influence the result in tree based models. The value on the terminal nodes is the average log revenues in August for the group that are found in that particular terminal node.

Random Forests

Figure 5. Mean square error



The model was estimated with Random Forests with 1 to 22 randomly chosen variables in each split and an overview on how this decision is related to the performance of the model is obtained. Both results from the test data and OOB-observations are included in the graph. For the analysis below we use m=5.

Table 6. Variable importance

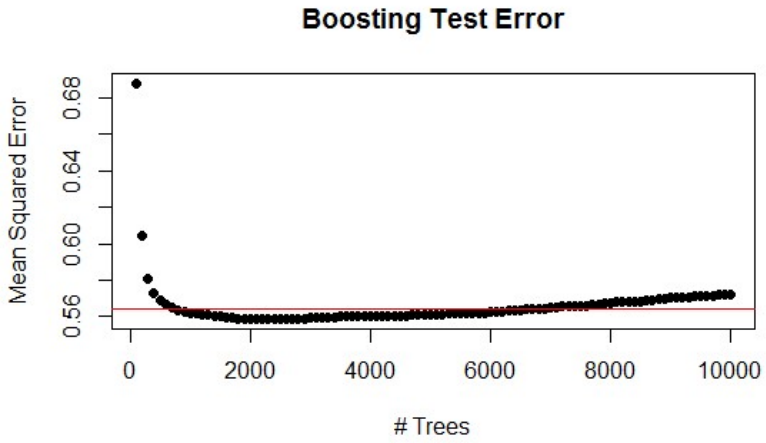
Variable	Increased Node Purity
Revenues in July	2287.0
Published nightly rate	1752.7
Security deposit (amount)	680.0
Revenues in June	643.3
Longitude	621.5
Latitude	586.5
Local revenues – July	551.9
Created date (numeric)	541.6
Reservation days in July	488.1
Local reservation days – July	458.2
Local no days – July	455.1
Maximum guests	423.7
Number of bedrooms	355.2
Number of bathrooms	313.4
Rated class	250.5
Entire home (i.e. not shared)	235.1
Minimum days for reservation	220.2
Reservation days in June	193.1
Island	119.1
No day rented in July	98.0
Instant bookable (dummy)	55.7
No day rented in June	42.0

The explained variance is 49.6% in the test data. Using OOB-observations for the full data the variance explained is 50.0%. Variable importance can be measured with the increase in node purity averaged over all trees. The increase in node purity is the average reduction in residual sum of squares (RSS). The variable "revenue in July" is the most important variable in this respect, but the published nightly rate is also very important. The reason that security deposit occupies a fairly high position is, probably that requiring a security deposit is related to the value of the property. Note that "no day in July" seems to be fairly important in the linear regression, but is not important at all in this model. The reason is that the variable could be omitted because having zero revenues in July is of course that the apartment/room did not have any reservation during July.

Boosting

Boosting was done with the shrinkage parameter of 0.01 and 10000 trees were grown. Over fitting was found to start at about 3000 trees. A systematic search over different interaction depth provided the highest result on the variance explained with an interaction depth of 9. On the test data, explained variance was 50.2%. This is an improvement compared to Random Forests. The difference is fairly small.

Figure 6. Mean squared error



The horizontal line refers to Random Forests. At about 3000 trees Boosting starts to over fit the data. 1088 trees should be used when the amount is chosen based on 5-fold cross validation. In that case, the explained variance is 49.7%, which is very close to what was found for Random Forests. For the test data there is, however, still advantage to fit more trees.

Table 7. Variable importance

Variable	Relative influence
Revenues in July	50.9
Published nightly rate	22.1
Longitude	3.7
Created date (numeric)	3.5
Revenues in June	2.8
Security deposit (amount)	2.3
Latitude	2.0
Reservation days in July	1.9
Local revenues – July	1.7
Local reservation days – July	1.6
Local no days – July	1.5
Maximum guests	1.3
Number of bathrooms	1.1
Rated class	1.0
Minimum days for reservation	0.9
Entire home (i.e. not shared)	0.6
Number of bedrooms	0.4
Reservation days in June	0.3
Instant bookable (dummy)	0.1
No day rented in July	0.0
Island	0.0
No day rented in June	0.0

The importance of the variables are evaluated at the cross validated amount of trees. The results are similar to what is found for 3000 trees, except that the magnitude of the relative influence of revenues in July is lower in that case.

Topic 3: Analyzing the probability to dropout from Airbnb in July or August 2016

For the complete data we find that 11.3% of the active listings in June left Airbnb during either July or August 2016, that is, during the months of high season. In this sample we only include listings that were last scraped in June or later. If the date of last scraped was before, we do not include the listings in the model, because we consider that they abandon the system before the high season. Of course apartments that were introduced into the system in July and August (or later) are also excluded from the sample.

Logit

Table 8 includes the marginal effects on the probability to dropout from Airbnb in either July or August 2016. It is very interesting to see that we cannot reject the hypothesis that the marginal effects for the performance related variables in June, nor May is zero.

Table 8. Marginal effects on the probability to dropout in July or August 2016.

	Marginal effects	Standard errors	z	P>|z|
No day rented in June	0.0035	0.0366	0.09	0.924
Reservation days in June	-0.0011	0.0008	-1.44	0.150
Ihs(Revenues in June)	-0.0007	0.0054	-0.13	0.893
No day rented in May	-0.0279	0.0474	-0.59	0.557
Reservation days in May	-0.0015	0.0010	-1.42	0.157
ihs(Revenues in May)	-0.0056	0.0064	-0.87	0.386
ln(published nightly rate, EUR)	-0.0451***	0.0044	-10.24	0.000
Entire home (i.e. not shared)	-0.0041	0.0078	-0.52	0.602
Number of bathrooms	0.0133***	0.0031	4.34	0.000
Number of bedrooms	0.0095**	0.0038	2.49	0.013
Maximum guests	-0.0059***	0.0019	-3.05	0.002
Minimum days for reservation	-0.0044***	0.0011	-4.14	0.000
Instant bookable (dummy)	0.0054	0.0056	0.97	0.334
Security deposit (dummy)	0.0204	0.0295	0.69	0.490
ihs(Security deposit, EUR)	-0.0014	0.0047	-0.30	0.761
Created date (numeric/365)	0.0023	0.0027	0.86	0.387
Created in June 2016 (dummy)	0.0071	0.0080	0.88	0.379
Created in May 2016 (dummy)	0.0309***	0.0096	3.23	0.001
Created in April 2016 (dummy)	0.0119	0.0099	1.21	0.227
Rated class, Low (dummy)	-0.0750***	0.0053	-14.09	0.000
Rated class, Moderate	-0.0912***	0.0045	-20.48	0.000
Rated class, Hight	-0.1394***	0.0051	-27.50	0.000
Rated class, Perfect	-0.0849***	0.0043	-19.92	0.000
Ibiza (dummy)	0.0670***	0.0077	8.67	0.000

table continues from previous page

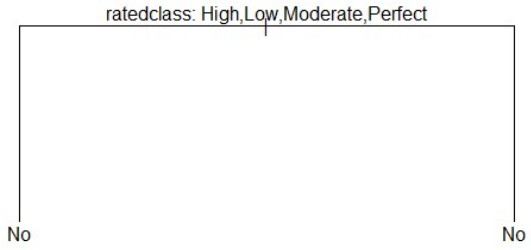
	Marginal effects	Standard errors	z	P>|z|
Menorca	-0.0105	0.0168	-0.63	0.530
Formentera	0.0495**	0.0212	2.33	0.020
ln(Local revenues – June)	-0.0045	0.0052	-0.86	0.392
Local no days – June	0.0352	0.0393	0.90	0.370
Local reservation days, June	0.0075***	0.0023	3.23	0.001
Apartments: 19677	Pseudo R2 : 0.0964			

Notes: Marginal effects are estimated based on a Logit model. ***,**,* indicate significantly different from zero at 1, 5 and 10% significance level.

Very strong marginal effects are found for all of the categories of being rated. For example, the probability to dropout is 0.075 lower for the class with lowest rating, compared to the group without any review at all. Apartments in Formentera have a 0.0495 higher probability to dropout compared to apartments in Mallorca. The marginal effects are evaluated at the average of the explanatory variables, and all marginal effects are interpreted considering all other variables constant.

Classification Tree

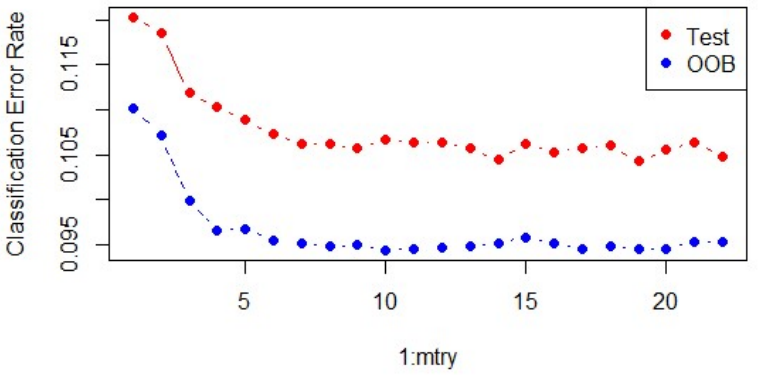
Figure 7. Classification tree



Both terminal nodes predict "No", that is, the majority is predicted to remain active. The proportion of dropout is 6.9% compared to 21.0% among those that never has been rated. This model classifies all observations to "No", and the misclassification rate is equal to the proportion that actually drops out from the system.

Random Forests

Figure 8. Classification error rate



In this case we choose m=14. 89.6% of the test observations are correctly classified. Classifying all of the observation in the test data to "No" would mean a rate of 88.5%. Measuring the correct classification rate for the complete data with OOB-observations implies a rate of 90.3%.

Table 9. Confusion matrix

		Model predicted class:		
		No	Yes	Classification Error
Actual class:	No	17323	129	0.0074
Dropout	Yes	1774	451	0.7930

The model fails to detect almost 80% of the dropout that occurred; only 451 were correctly classified to "Yes". On the other hand, predicting dropout, when this did not occur was only done for 0.7% of the apartments that actually remained in the system.

Table 10. Variable importance

Variables	Mean Decrease Gini
Security deposit (amount)	478.3
Created date (numeric)	401.0
Published nightly rate	386.4
Longitude	375.9
Latitude	352.5
Local revenues – June	325.5
Local no day – June	308.5
Local reservation days – June	295.3
Rated Class	195.9
Minimum days for reservation	141.1
Revenues in June	128.7
Maximum guests	111.2

Table continues from previous page

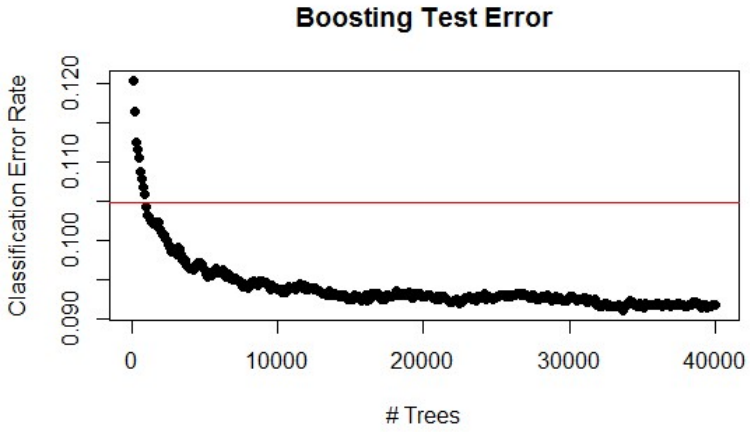
Variables	Mean Decrease Gini
Number of bathrooms	96.2
Reservation days in June	79.0
Number of bedrooms	75.7
Revenues in May	61.8
Reservation days in May	48.4
Instant bookable (dummy)	36.7
Entire home (i.e. not shared)	20.1
No day rented in June	12.3
No day rented in May	10.6
Island	5.6

Many variables are found to obtain a moderate mean decrease in the Gini index. Security deposit turn out to be the most important variable, which is, somewhat surprising given that the coefficient for the variable was not even significantly different from zero in the Logit model. The low importance of the information on islands is simply due to that longitude and latitude effectively separates the islands, and the variable "Island" does not contribute with new information. It is important to keep in mind that classification is very difficult in this example.

Boosting

Initially 10000 trees were grown, but even with shrinkage parameter of 0.01 it was not clear if more trees could improve the model further. Estimating the model with 40000 trees no apparent signs of over fitting was found, but the model also had stopped to improve. 91.0% were correctly classified with an interaction depth of 8.

Figure 9. Classification error rate



We also used 5-fold cross validation to choose the number of trees without using the test data and the best choice of number of trees were 13280. 91.1% of the test observations were correctly classified in that case. Boosting is the preferred model to predict dropout from Airbnb in this study of listings in the Balearic Islands.

Table 11. Variable importante

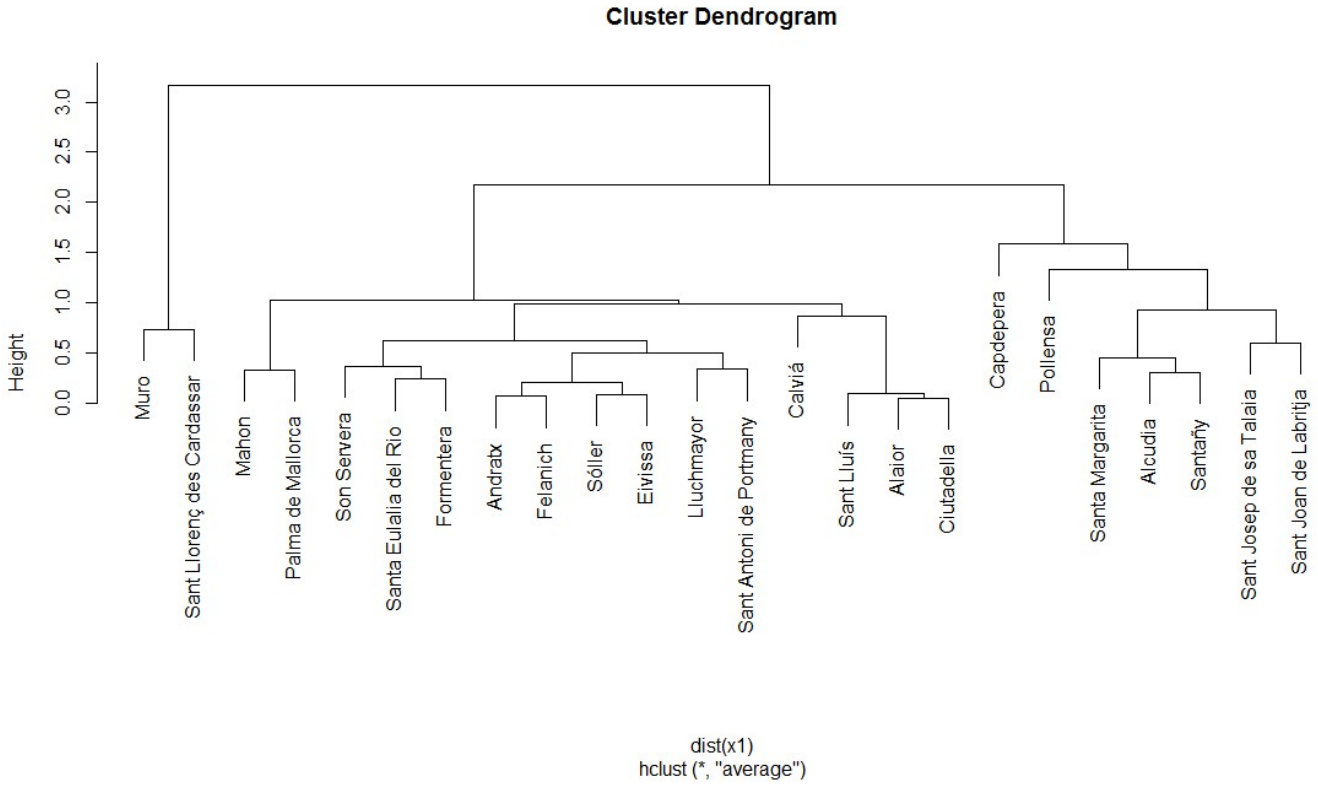
Variable	Relative influence
Security deposit (amount)	24.0
Published nightly rate	11.0
Rated Class	9.3
Created date (numeric)	8.7
Longitude	8.7
Latitude	8.1
Local revenues – June	6.4
Local reservation days – June	6.0
Local no day – June	6.0
Minimum days for reservation	2.6
Revenues in June	2.1
Number of bathrooms	1.4
Reservation days in June	1.3
Maximum guests	1.2
Number of bedrooms	0.8
Revenues in May	0.8
Reservation days in May	0.6
Entire home (i.e. not shared)	0.4
Instant bookable (dummy)	0.3
Island	0.2
No day rented in June	0.1
No day rented in May	0.1

The relative importance of the variables was evaluated at the optimal number of trees according to the cross validation. It is very interesting to see that the only performance related variable that is found reasonably high is “Rated Class”, while measures of performance in June or May are not important. The local variables are, in fact, more important compared to revenues or days rented in June. This conclusion was also found for Random Forests. We expected that performance related variables would be more important. This makes the task to predict the dropout very difficult.

Topic 4: Clustering analysis on tourist saturation in August 2016.

Based on data on bed places and occupation rate in August 2016 we are interested to cluster the municipalities into homogeneous groups to detect different kind of tourist saturation. For the analysis we use tourists per capita on an average day in August 2016 in both sectors. We refer you to the main report for the definition of the variables. We only include municipalities where data on occupancy rate for the hotels in August 2016 was available. The cluster analysis was made after standardizing the variables to give the variables the same weight in the algorithm. Hierarchical clustering was used and the dentrogram is found in Figure 10.

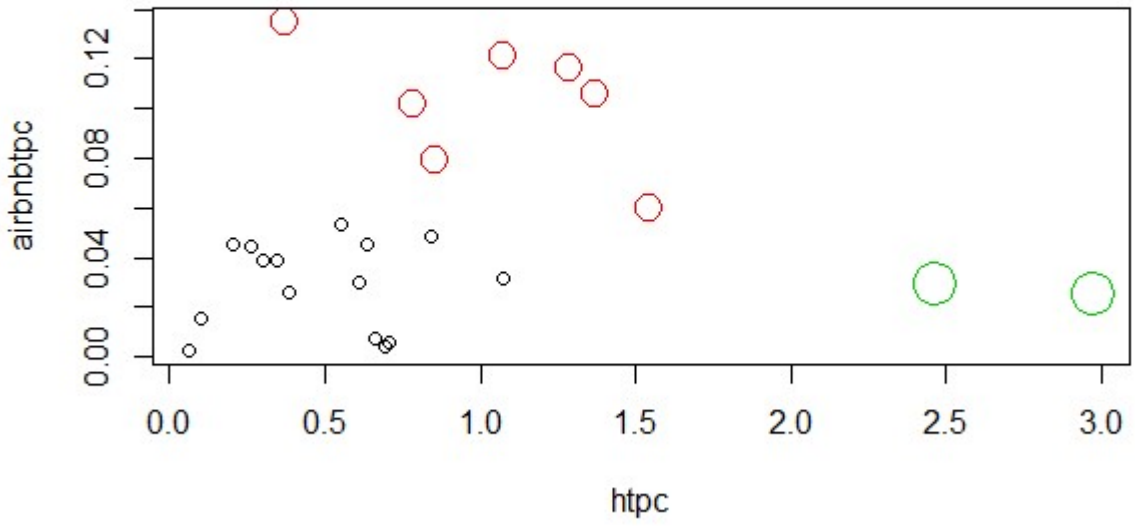
Figure 10. Cluster dentogram



A dentrogram based on hierarchical clustering show all objects as their individual cluster and the process of grouping these, starting with the closest pair. We use average linkage for this process. The vertical axis is proportional to the distance between the groups that are joined together. Based on the dentrogram it is reasonable to settle for three different clusters in this case. Muro and Sant Llorenc de Cardassar is the first cluster. The municipalities from Mahon to Ciutadella are another cluster and finally the third cluster is Capdepera to Sant Joan de Labritja. In figure 11 we represent these clusters with different sizes of circles. This graph is based on the original variables, while the clustering was done with standardized variables. “htpc” refers to hotel tourists per capita, and “airbnbtpc” refers to Airbnb tourists per capita. Muro and Sant Llorenc de Cardassar are municipalities with very high saturation in terms of tourist per capita in the hotel sector, but very low saturation in terms of tourist per capita in the Airbnb sector. Hence, these are very touristic areas, but the Airbnb sector contributes very little to additional saturation. Another cluster consists of Capdepera, Pollensa, Santa Margarita, Alcudia, Santañy, Sant Josep de sa Talaia and Sant Joan de Labritja. This group has fairly high saturation in the hotel sector, and, in comparison to other municipalities, also for Airbnb tourists per capita. Notice that a high value for Airbnb tourists is around 0.08-0.14 tourists per capita, while the scale for tourists in the hotel sector is much higher. The largest group constists of Mahon, Palma de Mallorca, Son Servera, Santa Eulalia del Rio, Formentera, Andratx, Felanich, Sóller, Eivissa, Lluchmayor, Sant Antoni de Portmany, Calvía, Sant Lluis and Ciutadella. These municaplities have low saturation in terms of Airbnb tourists per capita, and low, moderate or even fairly high values of hotel tourists per capita.

Using this way to measure tourist saturation, it is clear that the saturation is much more driven by the hotel sector compared to the Airbnb sector in August 2016 among these tourist intense municipalities. The values on Airbnb are likely overestimated because of using the hosts' specification of maximum number of guests as the number of tourists. On the other hand it is important to keep in mind that we only analyze tourist rentals done in Airbnb, and other renting channels are not analyzed.

Figure 11.



Conclusions

In this technical report we have studied four different aspects of the Airbnb activities. We study the probability that the apartment was rented at least one day in August 2016. For the group that did have at least a day rented in August we evaluate which factors are important for the revenues. We also evaluate the probability to dropout from the Airbnb portal. Finally, we analyze tourist saturation in terms of the hotel sector and the Airbnb sector for tourist intense municipalities. The main conclusion from this analysis is that prediction is very difficult based on the available variables. Information on performance in June and July is important for probability to have at least on day rented in August. Despite this, it is difficult to predict the performance, and renting apartments in Airbnb involves a lot of uncertainty for the hosts on the Balearic Islands. Surprisingly, the performance related variables are not important for the decision to leave Airbnb in July or August, i.e. in high season. Of course, we cannot know if these apartments only dropped out from the system, or actually stopped renting the apartment for tourist. The performance related variables are related to days rented and revenues, but the data does not contain information on costs. These costs could, of course, be important for the decision to continue offering the apartment on this market. Another important conclusion is that tourist saturation is much more related to tourists in the hotel sector compared to the Airbnb sector in August 2016.