Apartment Pricing – Crete, Greece

Introduction

The purpose of this project was to help predict prices for our newly built small and mid-sized apartments hosting 2-6 guests in Crete, Greece. The project has utilized multiple prediction models on Airbnb data for Crete, Greece. The data was filtered to property types of apartments/lofts that could accommodate up to 6 individuals. After running different prediction models, the project advises the company to focus on the apartments in the Rethymnon municipality as those might garner higher rental prices, however, the predictive power of the model is weak for this municipality. On the other hand, the model is comparatively better at predicting prices in the remaining 3 municipalities. The model predicts these prices for municipalities given certain property types, amenities, and a few other requirements. These are based on the tuned Random Forest model with the loss function being the root mean squared error (RMSE).

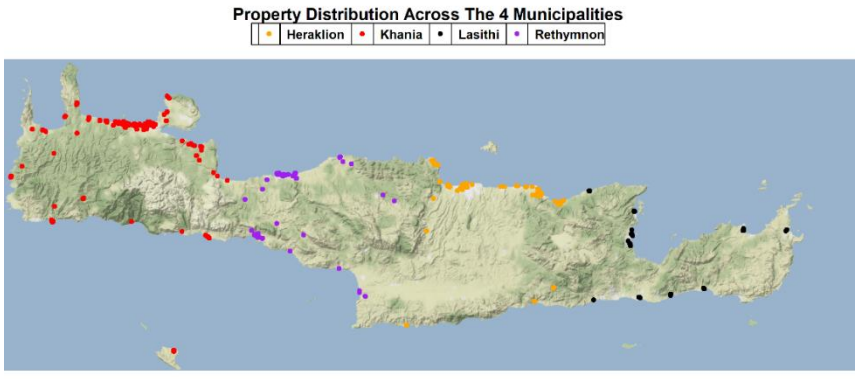
Data Selection

The models built for the purpose of this project utilized the Airbnb dataset scraped during the last week of December 2021, for Crete, Greece. The dataset contained information on numerous types of properties, reviews, municipalities, host related variables, and amenities. However, to be in line with the company vision of marketing the small and medium-sized apartments, the project focused on property types of 'Serviced Apartments, Entire Loft, Home/Apt' from the dataset. It was further filtered on the number of people the properties could accommodate, which in this case was between 2-6 people. A major downside of this filtering was the extreme reduction in the size of the data set from around 20,000 observations to only around 500 observations. Out of these 500 observations, 80% were used to train the models and the remaining 20% were used as the hold out which served as live data to test the models. That said, for the purpose of this project, prediction exercises were conducted on these limited observations to come up with an initial pricing mechanism for feasibility purposes, which can be improved upon later on by doing this exercise on a larger dataset.

Data Engineering

Since the source of the dataset was web-scrapping, it was dirty. Our analysts had to conduct intensive data cleaning and preparation prior to running the predictive models. A major part of the preparation process was to extract amenities provided at each apartment and pool them into relevant categories. Some of the amenities had to be dropped considering the relevancy with our built properties and the size of the dataset. Amenities such as lake access, private pool, hot tub and a few others which were not relevant in our case; our apartments, for example, currently have a shared pool and there are no lakes near our properties, only the ocean that is covered by another amenity. Moreover, variables were divided into numeric, categoric, and dummy types for easier analysis.

Given that the data was scrapped from Airbnb Greece website, some of the neighborhoods were written in Greek which R-Studio could not pick up, making it difficult to make any modifications. However, after careful deliberation, it was decided to divide the observations into 4 major municipalities of Crete, based on the latitude and longitude information of the properties. The map below shows the properties from the dataset mapped as per



their municipalities on the island of Crete.

The dataset also had missing values in various variables, both factor and numeric variables. For factor variables, a flag variable was created that took the value 1 where the original variable was NA and 0 otherwise and the NAs in the original variable were replaced with the word "missing". For numeric variables, flags were created in the same fashion, however, the NAs were replaced with the median of the original variable. The reason for creating the flags is that we believe these were missing at random as we did not find a particular pattern in the missing values.

Furthermore, we looked at the distributions of all the variables to see if there was a need to convert to log scale. Only 2 variables had to be converted in log scale; the number of reviews and minimum number of nights, because of these having along right tail. Apart from these variables, others had a near normal distribution. We also checked for the association of all the individual variable with our y-variable (price), where all had either a linear or near-linear relationship with it, hence no higher-order polynomials of the variables were considered.

Additionally, we placed a cut off at apartments with a price tag of Euros 300 per night as our small and medium-sized apartments not luxury and are targeted towards the middle and upper-middle class of the economic strata.

Moreover, boxplots of property types against price were plotted, conditioned on amenities and municipalities to check for viability of interaction effect on the price. Based on these, we created around 26 interactions between property type and several amenities plus the interaction between property type and their municipalities, to uncover any amenity or municipality specific price effect.

Finally, since the number of observations were extremely small due to specification of our project and the number of available variables were comparatively large, we decided to use the LASSO model as a tool drill down into the variables that LASSO assigned the value of zero to the coefficients. It helped us narrow down the number of predictors to only 54. This was also beneficial because our knowledge of the apartment pricing was limited, and we were not very sure about the relative importance of all the variables.

Out of these 54 predictors, unique variables were identified and then used in the following prediction exercise.

Prediction

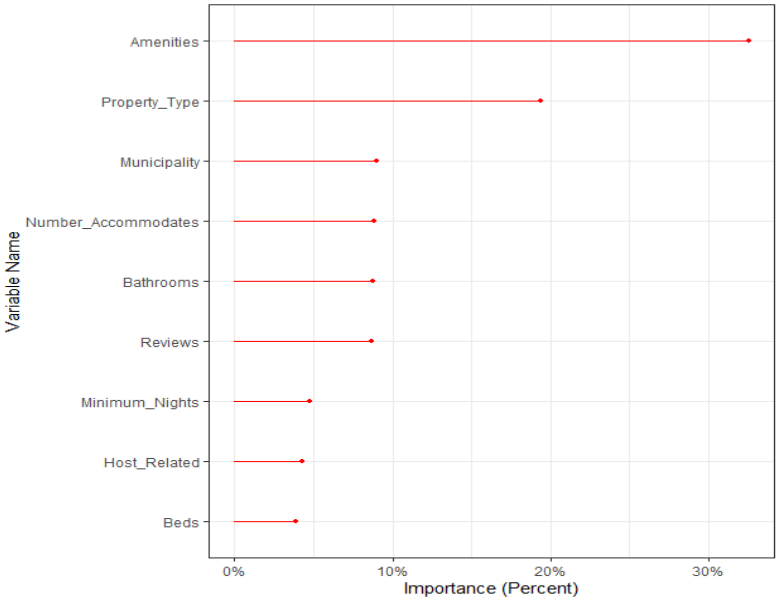
For the purpose of prediction, we ran a total of 4 different prediction models: a simple Ordinary Least Squared (OLS) model containing basic variables, Classification and Regression Tree (CART) with pruning, two Random Forest (RF) models where 1 was provided with the tuning parameters and the other was run on automatic tunning and lastly a Gradient Boosting Machine (GBM) model. The results of the cross-validated Root Mean Squared Error (RMSE) on the training sample are provided in this adjacent table. The difference between the two RF models is negligible, however, keeping in mind that the same models would be run again on a larger dataset version of our dataset, we advise to use the RF model with provided tunning parameters due to time efficiency.

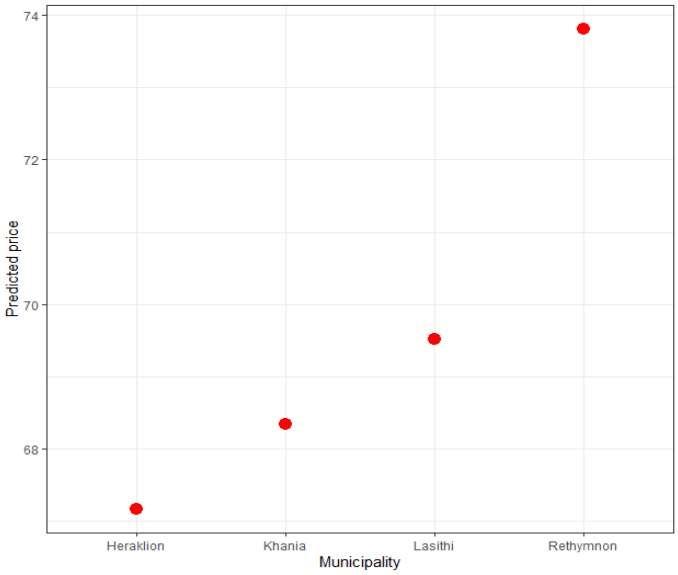
Model Type	**CV RMSE**
OLS	31.5
CART with pruning	31.6
Random Forest 1: With Tuning Provided	27.9
Random Forest 2: Auto Tuning	27.8
GBM	30.2

We further ran diagnostics on the RF1 model using Variable Importance (VI) plots, Partial Dependency (PD) plots, and checking subsample performances. For variable importance plots, we grouped together similar variables and re-calculated their importance to gauge the relative importance of these variable groups in predicting the prices. Amenities stood at the top with a relative importance of around 34% followed by the

type of property with 20% relative importance, as shown in the graph below. Based on this, we recommend the management to focus on the types of amenities provided in the apartment towers.

We additionally drilled down into some of these important variables to create PD plots to see how changing the specific predictors value while keeping other predictors constant predicted the prices. Looking at the PD plot for property type suggests focusing more on entire





serviced apartments, which may garner higher prices relatively, while keeping other things constant. Similarly, drilling down into municipality suggests focusing more on our properties in the Rythmnon municipality as the model predicts relatively higher prices compared to other municipalities, while keeping other things constant.

	RMSE	Mean Price	RMSE/Price
Municipality			
Heraklion	29.9	69.0	0.4
Khania	33.2	74.4	0.4
Lasithi	35.4	98.0	0.4
Rethymnon	44.4	69.8	0.6
Property Type			
Entire Home/Apt	26.4	35.0	0.8
Entire Loft	21.0	54.9	0.4
Entire Serviced Apartment	36.4	78.0	0.5
All	35.1	75.2	0.5

We also looked at subsample performances for these variables to see how individually these would impact on the price prediction. As shown in the table here, this contradicts with the outcome of the PD plots, where here Lasithi municipality seems to garner higher price predictions with an error of 0.4 Euros per predict Euro in the price. However, when it comes to entire serviced apartments, this is in line with the PD plot results suggesting focusing more on these property types where the prediction error of 0.5 Euros per predict Euro in the price of 78 Euros.

Conclusion

Based on these model predictions, it is very difficult to drill down into certain predictions. Perhaps it is possible the contradiction between the PD plot results and subsample analysis results could have been because of the very small size of the dataset. If the management is willing to invest in another project where the focus is given on data collection so that the resulting dataset is large enough, the project recommends using the Random Forest model with provided tuning parameters to predict the prices for our apartments in Crete, Greece.

However, if the management is to go ahead with the current predictions, the project advises to weigh the importance of a simple OLS model as well. A very important idea to keep in mind here is that the OLS model could have been chosen as well, due to its simplicity and the smaller number (13) of predictors in comparison to the size of the dataset. The Random Forest model was chosen here because of the small RMSE in the training dataset as well the holdout dataset in comparison to other models. Nevertheless, OLS model's RMSE was

roughly 4 Euros higher than the best RF model and if the management is ready to ignore this marginal amount, we advise to utilize the OLS model for prediction on the live data.