Predictive Pricing Model for Apartments in Toronto

Rauhan Nazir

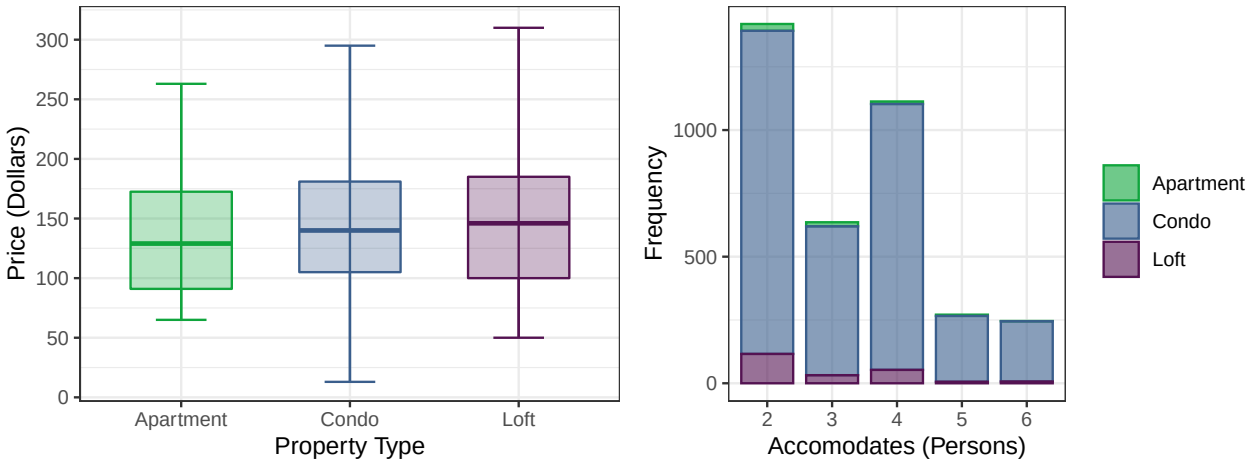
2/10/2022

Executive Summary

The purpose of this project was to help a company in setting their price for their newly launched apartments in Toronto, Canada. This company operates and deals with small and mid-sized apartments hosting 2-6 guests. There were 4 models that were created, namely OLS Linear Regression, Random Forest (parameters provided & Auto-Tuning), Cart & GBM, and their performance was evaluated based on their RMSE. The best model that turned out to be was Random Forest with Auto-Tuning.

Data Selection

The data set that I initially downloaded from the Airbnb website contained 15,435 listings, however it had to be cleaned and filtered down to only the listings that were relevant and similar to the apartments that my company was about to launch, making sure that the data we used in our models was a result of conscious decisions made, ensuring that we were not feeding any irrelevant data into the models that made us compromise on the quality and accuracy of the models. Hence, a lot of time and detailed thought process was spent on the data cleaning and data preparation. The listings in the data set contained all type of listings including the ones that were not apartments or were not similar to apartments, like private suits, so I filtered the data down to the ones that were relevant, namely, "Entire loft", "Entire serviced apartment", "Entire home/apt" and "Entire condominium (condo)". Even from those listings I filtered down to the ones that accommodated 2 to 6 people, as these were the only ones that matched our criteria. I also removed the percentage signs from the percentage columns such as the host acceptance rate and the host response rate. The dollar sign was also removed from the price variable, which is our target variable. After that I changed the values in the property type to make them cleaner and remove the redundant words such as "Entire condominium (condo)" was changed to "Condo".



Data Engineering

There was only one listing that was scraped on the 9th of January, so I just dropped it. After that I started performing basic data checks and the first one was to check if data had rows with only NAs and duplicate values, as it made no sense to keep them. However, I did not find any such rows. Another check that I performed was to see how many columns had more than 50% NA values, to figure out which variables needed more attention, and drop them especially if they are not important predictors. There were 3 variables that had all the values as NAs, including the "bathrooms", so I just dropped them right away. However, Bathroom is an important feature for predicting the apartment price, so I extracted that information as a numeric variable from a column named bathroom_text. I also dropped all the other irrelevant variables at this stage as well. Another important factor while valuing the property is the kind of amenities an apartment comes with. In the initial data set, the amenities were included in one variable as a list, and for us to be able to include all those features in the models, it was necessary to extract all of them into separate dummy variables, and for that I had to go through extensive data cleaning and preparation. There were a lot of amenities, even those that did not make much of a difference. So, after extracting all the amenities, I dropped the ones that had minimal and no impact and clubbed those which were quite similar to each other. For instance, TV and Netflix, and children, crib, and baby were clubbed into one separate dummy variable. For this process, the domain knowledge was extremely important, and I had to go through extensive research.

There was no point including variables that had no variance in them, as they would not help predicting the price at all, one example was the room type variable. The filtered data had only one type of room, so I dropped that variable. As mentioned earlier, the number of bathrooms can make a difference in the property value and there were a couple of conscious decisions that I made while cleaning this variable. The data set included bathroom values in decimals as well, that do not make much sense and for easier and clearer interpretation, I rounded them to whole number. For instance, 1.5, 2.5 and 3.5 bathrooms were just considered to be 2,3 and 4 respectively, the rationale behind this was that I considered the half bathroom as a whole bathroom as well.

For all the variables that I converted to factors, percentages, numeric and dummy, I put a prefix of f, p, n and d respectively so that I can easily recognize and then only kept those, apart form the target variable, as these were the final variables that were going to be used to make the predictive models. One important decision that I had to make was that how the missing values were going to be dealt with. There was a different approach taken for different variable, based on my domain knowledge. For instance, for the variable missing bed, I used the variable, number of accommodates, to apply a sense check. Where there were 3 people accommodated, I included number of beds as 2, and this is the ratio that was applied throughout.

I then checked the distribution of our target variable (Price), to see if there was a need to transform it my taking a log of it to make it more normally distributed. However, upon inquiry I found out that there was no need for taking the log and absolute values resembled more to that of a normal distribution. I also filtered down to the apartments that were priced at 350 dollars or less, since it was very rare for the apartments to be priced above that and since our apartments are going to small or mid-sized, there was no need to include the ones above 350 dollars. Finally, for the cleaning and preparation phase I pooled values within certain variables, which did not have a significant difference between their mean values, before converting them as factors.

After all the data cleaning and data preparation, following are the variables that were used in the models that were created to predict the price of an apartment:

- *Dummies*: Binary variables consisting of all the amenities that are being offered by host.
- *Size variables*: This includes numeric variable like number of beds, number of baths, number of people it accommodates, and minimum nights.
- *Factor variables*: For each Neighbourhood, type of property, including flag and factorized variable of size variables.

- *Reviews variables*: Review score rating and the number of reviews the apartment gets each month.
- *Host variables*: Dummies for host verification and if they a super host or not.

Table 1: Ranking of Models CV RSME

	CV RMSE
OLS	53.47537
CART	54.98241
Random forest 1: Tuning provided	51.04889
Random forest 2: Auto Tuning	50.18117
GBM	51.08318

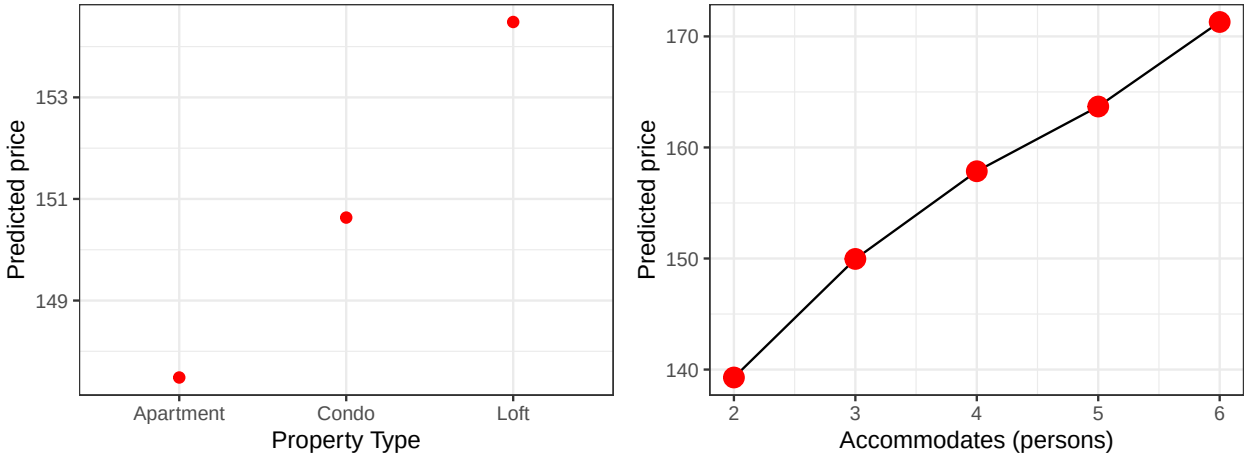
- *Interaction Terms*

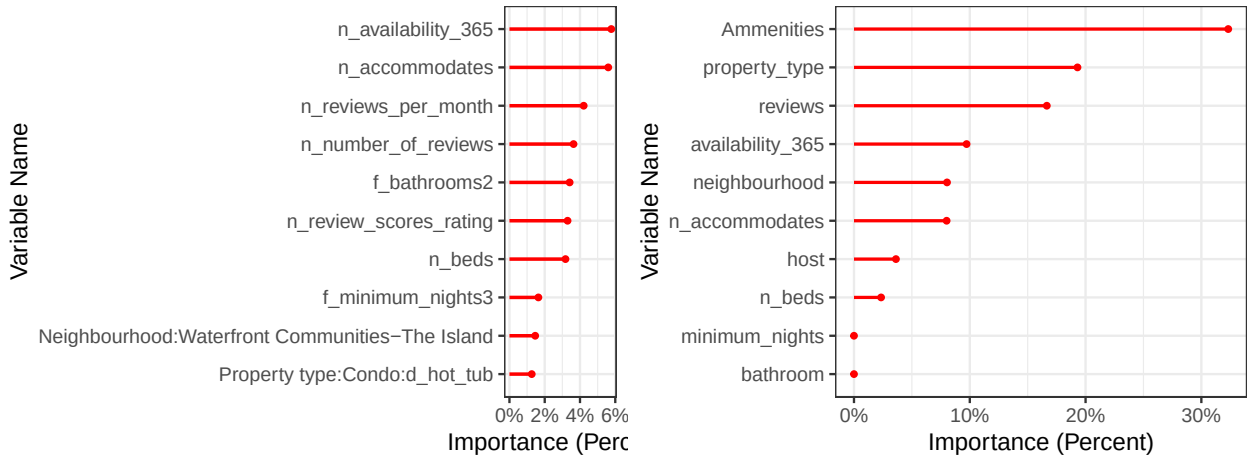
Prediction

After all the data cleaning there were 3685 listings left. Of these listings, 20% were randomly selected as a holdout set (test set), while the rest were used as a training set. The training set underwent 5 test fold cross validation, allowing our models to perform more efficiently and provide the coefficients with least amount of overfitting. The result of these cross-validated RMSEs were used to then evaluate the model performance and was used as primary criteria to select the best model.

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) are provided in this adjacent table. It can be seen that the best model is Random Forest with Auto-tuning.

I also made Partial Dependence Plots where it keeps everything else other than the predictive variable, we want to see association of with the target variable. We created two such graphs, one with the property type and the second one with the number of people it accommodates. In the property type, Lofts were the ones which were priced the highest while for the number of people it accommodates, price went up when the persons accommodated increased. 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. For our best model it showed that the amenities were the most important (almost 40%) and the property type and reviews were the next most important variables (almost 20%).





Conclusion

The best predictive model was Random Forest with Auto Tuning, while in the case study it was Random Forest with tuning parameters. We could have shortlisted the same if we wanted our model to be more time efficient as the difference between their RMSEs was minimal, and Auto-tuning takes a lot of time. But we still went with Auto-tuning as we only wanted to choose the best model in terms of the least error ($48.7 – The possible error in predicting the price). The recommendation to the company according to the final results is that they should invest in lofts in the Water-Front community, which is loaded with at least the basic amenities (the more the better), with the option of accommodating the maximum possible number of persons, as it will allow them to price their listings the highest and potentially earn highest profit. The RMSE of our model, compared to the one is higher, however there are lot of factors that have changed over time. For instance, the data in the case study was from 2017 and our data is from 2022, and one potential change is the variety and range of apartments being offered.

Predicted vs actual prices