Zsófia Katona
https://github.com/zsofiarebeka/DA3_Assignment-2

Toronto Airbnb Price Prediction Analysis

1. Introduction

This report aims to present the findings for my analysis conducted for our partner company, focusing on small and mid-sized apartments for shorter rentals. The objective of this report is to provide insights into pricing strategies for Airbnb listings in Toronto. The data set contains 20,386 Airbnb listings, spanning across various attributes including host details, property attributes, pricing, availability, and guest reviews.

2. Data cleaning and feature engineering

Data preprocessing included dropping unnecessary columns and cleaning the price column. I handled missing values by imputation and the exclusion of irrelevant columns. I replaced the missing bedroom and bathroom values with the number of beds and converted the bathroom descriptions to numeric values. The feature engineering incorporated formatting datetime and binary values. Moreover, I filtered for property types that corresponded to small and mid-sized apartments. Besides pooling values for fragmented variables, I created bins for apartments based on the number of bathrooms and reviews. It's important to mention that the amenities column didn't contain any valuable information, possibly because the data was scraped recently, on the 8th of January, 2024.

3. EDA

I transformed and renamed the columns with numerical variables, and renamed the categorical variables to oversee the data types easily. In the exploratory data analysis, I have analyzed the price distributions and found that the 90th percentile is around $300 a day (*Figure 1*). Therefore, I opted to drop the higher values to examine the pricing patterns and characteristics of the more typical listings within the dataset, which are likely to be more representative of the general market trends. I have found that prices of houses and rental units exhibit similar ranges (from $50 to $175), with the mean around $133 (*Figure 2*). I have found that entire rooms and apartments are more expensive, where the price converges to $150, meanwhile the mean price for private and shared rooms are around $75. Surprisingly, there is no significant price difference between private and shared rooms despite the private rooms' added value in terms of privacy (*Figure 3*). Even considering the number of accommodates, there are no large differences in different property types. The largest price gap is accommodations suited for 2 and 6 people. The increasing tendency in price is evident, and it increases around $50 at each additional guest (*Figure 4)*.

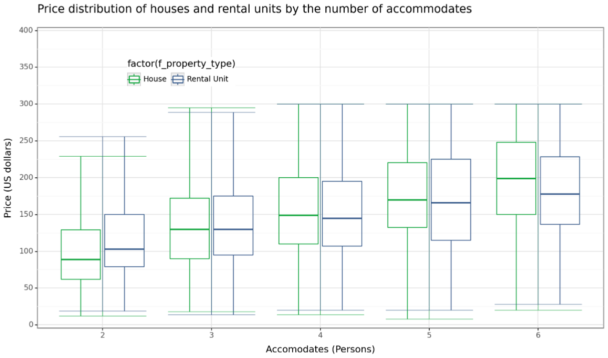


Figure 4: Price distribution of houses and rental units by the number of accommodates

4. Model building and evaluation

OLS

As the first model, I constructed 4 OLS regression models to predict the daily price of airbnb listings. Building the models incrementally, I started with regressing the price on the number of guests as based on the EDA, it was evident that it significantly influences the price (*Figure 5*). In the other models, I included property characteristics and added interactions. The R squared of the models increased gradually (from 0.183 to 0.391). Comparing the first regression to the other models, it became visible that room types have a greater impact on the target variable. I also conducted a comparison between the OLS and a Lowess model which revealed similarities between the two approaches and suggested that the linear model could easily find a good fit. Despite the high BIC values moving around 113,000, on *Figure 6*, we can see the RMSE of the models range between 50 and 59, with Model 4 fitting the data best among all the four models (50.559).

LASSO

As the second model, I defined different sets of features, including binary variables about host and availability and used different interactions. The variables were matched to see how different room types interact with property types and how different property types interact with instant bookability, availability, and superhost status. I created 8 different models and after performing cross-validated linear regressions, the results showed similar output to the OLS models. Standardly, I set the n fold to 5 and split the data into 4 categories. The RMSE values of the LASSO model ranged from 53 to 58. The RMSE values of the train and the test set seemed to align and go in the same direction, suggesting that the model is not overfitting and the model is generalizing well to unseen data. We can see a significant improvement when the number of variables reaches 6 and a slight improvement after when it reaches 13. However, there was a slight split between test and training set RMSE values when the number of coefficients reached 18. The best performing model appears to be Model 3, obtaining 12 variables for both RMSE Training and RMSE Test, with respective values of 53.32 and 53.41. (*Figure 7-8.*)

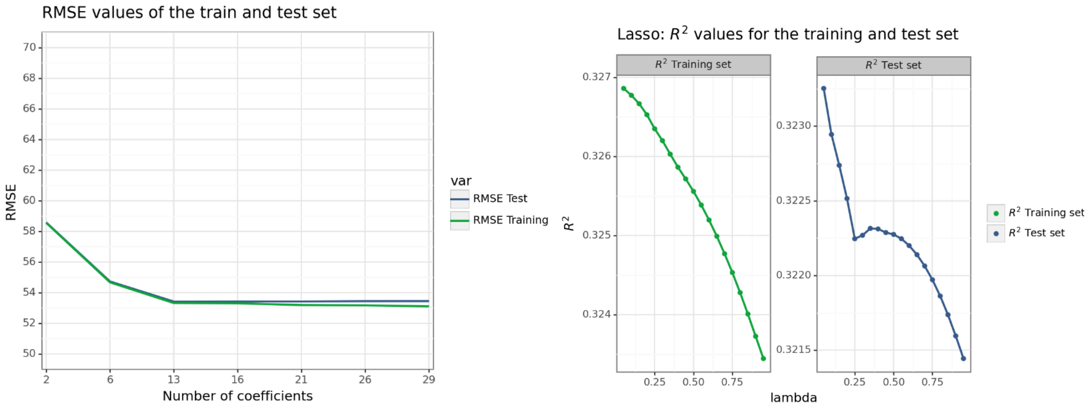


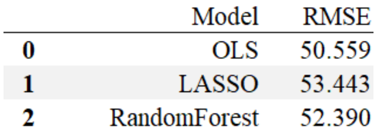
Figure 7-8: RMSE values and R2 values of the training and test set

To analyze the LASSO model better, I also inspected the R-squared values. As expected, the stronger regularization causes the R-squared values to decrease for both the training and test sets. The values for the training set are consistently higher, suggesting that the model might be overfitting to some extent, as it performs better on the data it was trained on compared to the unseen data. Using the best lambda value (0.2), none of the coefficients were close to 0, suggesting that even the best lambda value might be too weak, allowing the coefficients to retain relatively higher values. The RMSE value of the model (54.79) further proves the weakness of the regularization parameter. According to *Figure 9*, comparing the actual prices to the predicted prices of Model 3 indicated that the model is underfitting up to $150, but above that price the model turns into overfitting. Around the same price range, we can observe a weak cluster of values.

RandomForest

For the RandomForest model, I grouped basic variables, reviews, amenities and interactions and calculated the optimal number of variables for the trees. Due to the fresh data, my amenities column only consisted of dummies regarding the host. This also can be beneficial considering, the report is made for a company dealing with accommodations, therefore hosts. This way, these variables get more emphasis and we can provide better feedback for the company. RandomForest gave a similar range of RMSE values, from 52 to 58. According to the best estimator, the best performing tree obtains 12 predictors with the minimum of 5 observations in each node and has an RMSE value of 52.38. Exploring the feature importance values showed me that the most important predictors are the private room type, the number of guests, the beds, the reviews per month and if there is one bathroom in the apartment (*Figure 10*).

5. Key findings and interpretation



	Model	RMSE
0	OLS	50.559
1	LASSO	53.443
2	RandomForest	52.390

Comparing the lowest RMSE values of the models tells us that the most complex OLS model with interactions has the best predictive performance.

To compare if the feature importance values from RandomForest would actually have a better predictive power, I added another OLS model. Interestingly, regressing the price on the variables with the highest feature importance yielded worse results than expected. Comparing the RMSE value of the most complex linear regression with variables selected by hand (50.559) and the RF adjusted regression (54.705) shows us a significant difference. This indicates that while certain features may have high importance in RandomForest, their predictive power may not translate well to other types of models.

6. Limitations and recommendations for the company

The substantial amount of missing data may limit the analysis. Moreover, the absence of the amenity column restricts model insights. LASSO and RandomForest assumptions may not fully apply. Truncating prices at 90th percentile overlooks values above $300, which affects accuracy. (*Figure 11.*)

The company should adopt specific pricing strategies, emphasizing private rooms and adjusting guest and bed numbers. Price increases per additional guest should be limited to $50. For 2 guests, they should aim for $100, and around $200 for 6 guests. I would suggest higher prices for rental units accommodating 2 guests, and slight overpricing for houses hosting 6 guests should be acceptable.

7. Conclusions

Overall, the most complex OLS model with interactions performed the best predictive performance, obtaining an RMSE value of 50.559. However, it's important to mention that while feature importance values from RandomForest appeared promising, their translation intro predictive power in the OLS model yielded worse results, with an RMSE of 54.705. The LASSO model demonstrated moderate performance, with the models' RMSE values from 53 to 58.

Appendices

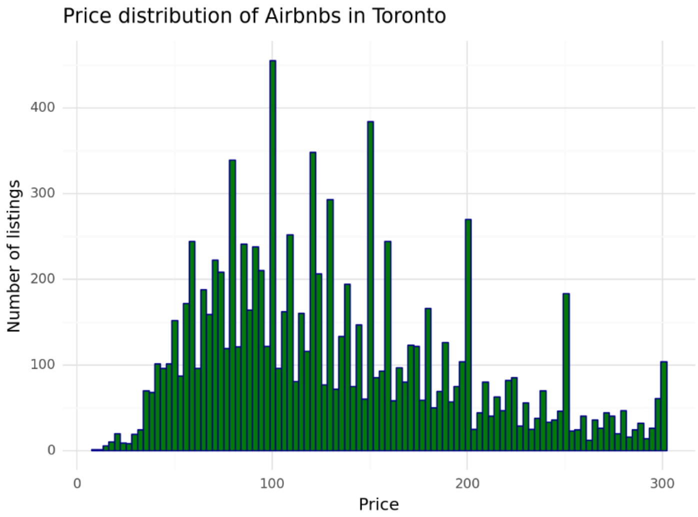


Figure 1: Price distribution of Airbnbs in Toronto

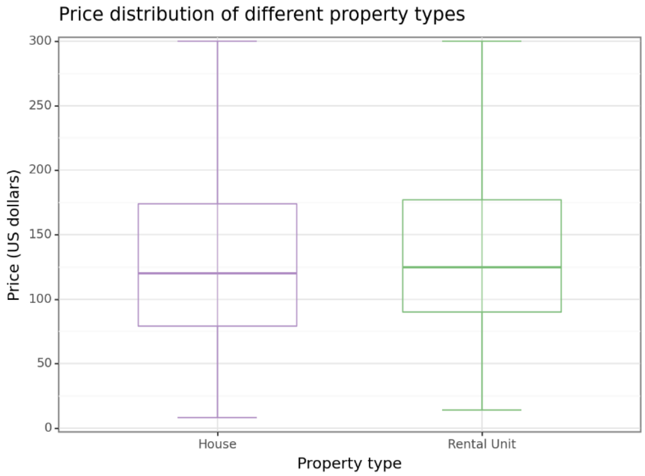


Figure 2: Price distribution of different property types

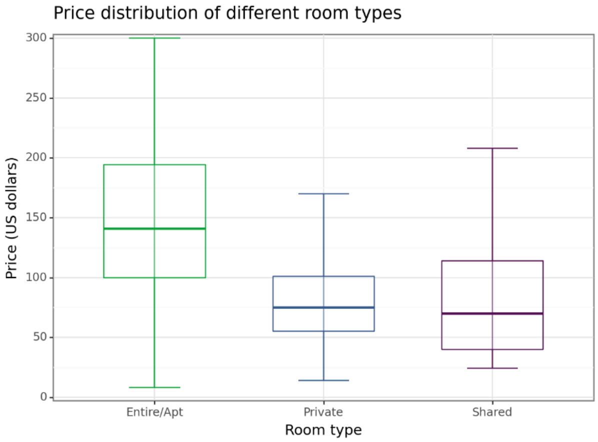


Figure 3: Price distribution of different room types

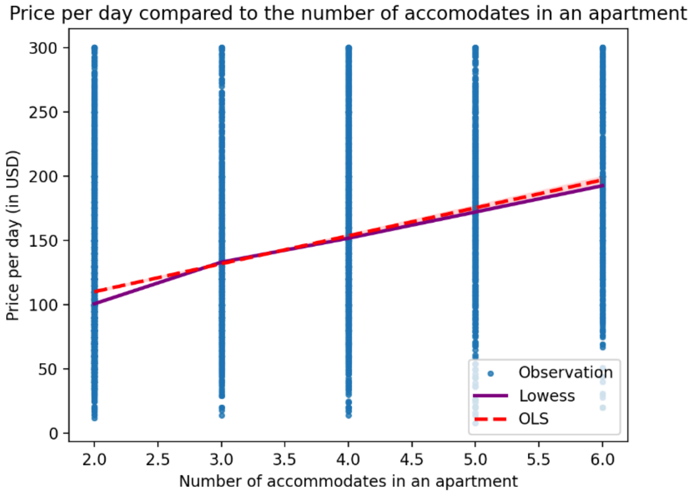


Figure 5: Price per day compared to the number of accommodates in an apartment

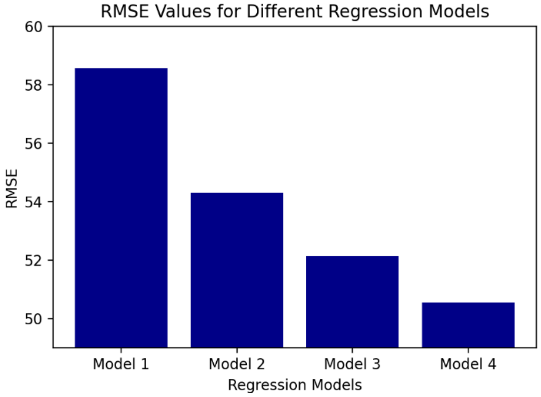


Figure 6: RMSE values for different OLS regression models

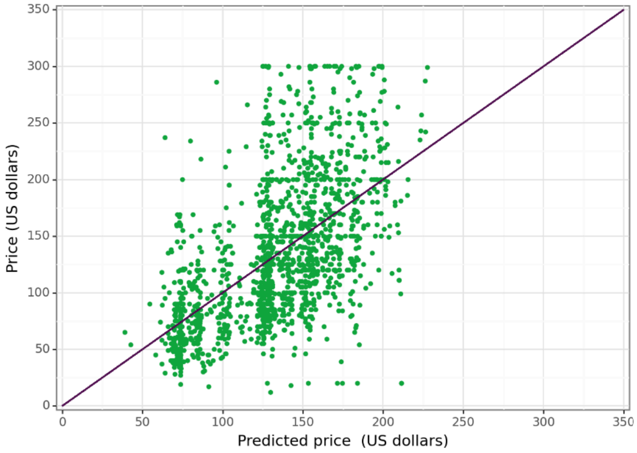


Figure 9: Comparing the actual values to the predicted values of Model 3, created by LASSO

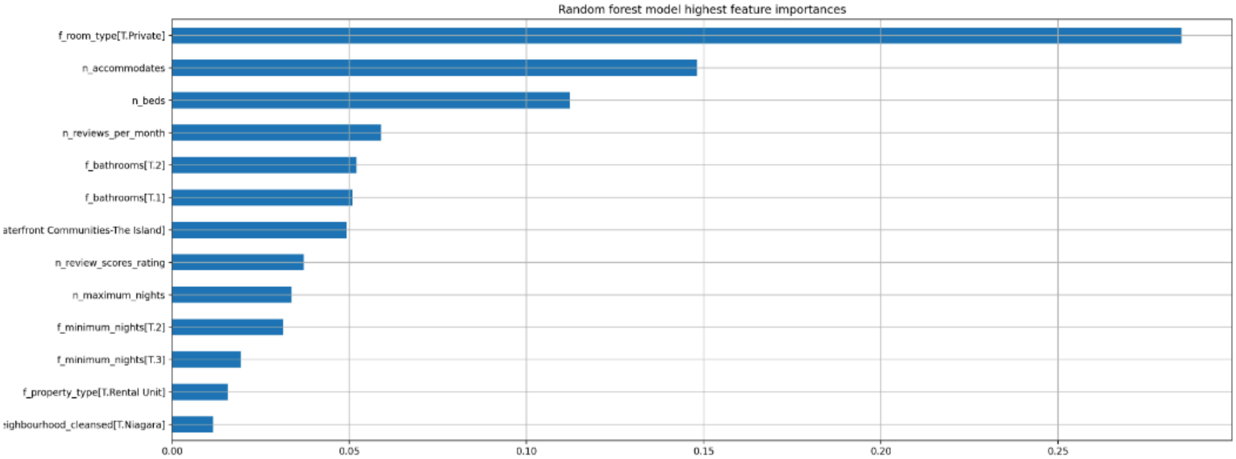


Figure 10: Random Forest model highest feature importance values

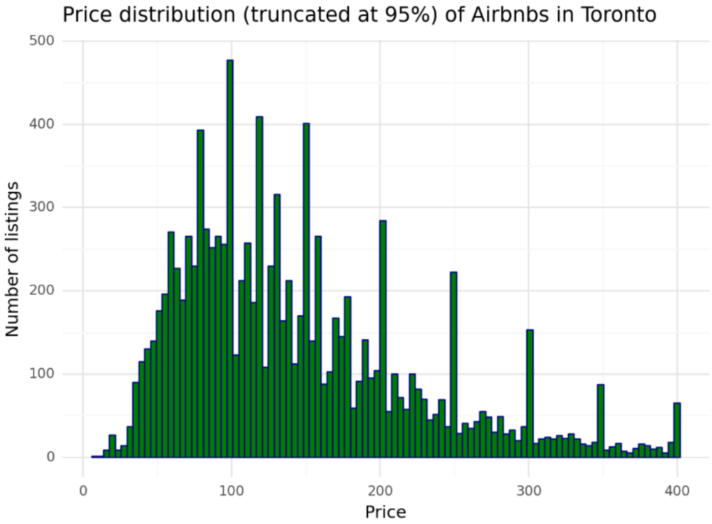


Figure 11: Price distribution (truncated at 95%) of Airbnbs in Toronto