Data Analysis 3: Assignment 2

Muhammad Arbash Malik – 2202071
GitHub

Executive Summary

The purpose of this project was to help the company in setting the prices for their newly launched apartments in Lisbon, Portugal. From the data brief, the company wants to deal with small and mid-sized apartments hosting 2-6 guests. There were 5 models that were created, namely Random Forest (parameters provided), Lasso Model, OLS Linear Regression, Cart, and GBM, and their performance was evaluated based on their RMSE.

Feature Engineering

The data set that I initially downloaded from the Airbnb website contained 22,605 listings, however it had to be cleaned and filtered down to only the listings that were relevant and similar to the apartments that the 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 data set had a lot of irrelevant property types for our business objective for e.g. room in serviced apartment, private room in home, room in hostel etc. so I filtered the data down to the ones that were relevant. 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. Feature engineering was done to ensure clean and accurate predictor variables were used in modelling.

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 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. For all the variables that I converted to factors, numeric, and dummy, I put a prefix; f_, *n_*, and *d_* respectively so that I can easily recognize and then only kept those, apart from the target variable, as these were the final variables that were going to be used to make the predictive models.

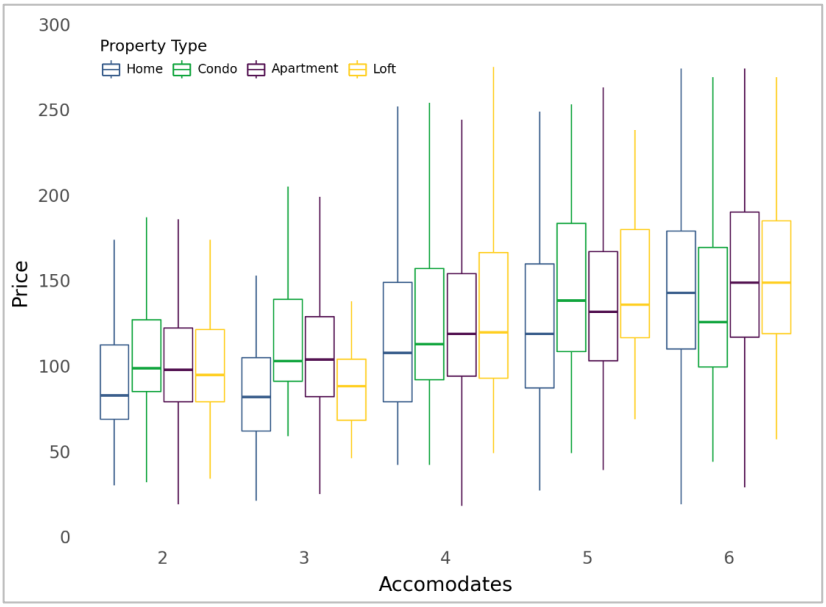
One important decision that I had to make was how the missing values were going to be dealt with. From our dataset, the missing values were in 3 variables only from which I took the decision on removing them instead of creating flag variables or imputing them. Lastly, I then checked the distribution of our target variable, 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 transforming it, and absolute values were sufficient to be used. I also filtered down to the apartments that were priced at 276 dollars or less (95% percentile), since it was very rare for the apartments to be priced above that and since our apartments are going to small or mid-sized , and it helped making the distribution more normal as well.

Modelling Decisions

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 as well as if the host is a super host and the property is instant bookable.
- Numeric variables: This includes numeric variables like number of beds, number of baths, number of bedrooms, number of people it accommodates, maximum and minimum nights.
- Factor variables: For each Neighborhood, 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.
- Interactions: This included interactions between property type & dummies as well as number of accommodates & dummies.

An interaction check was done between property type and number of accommodations to see if the interaction was necessary to add in the model (see below).



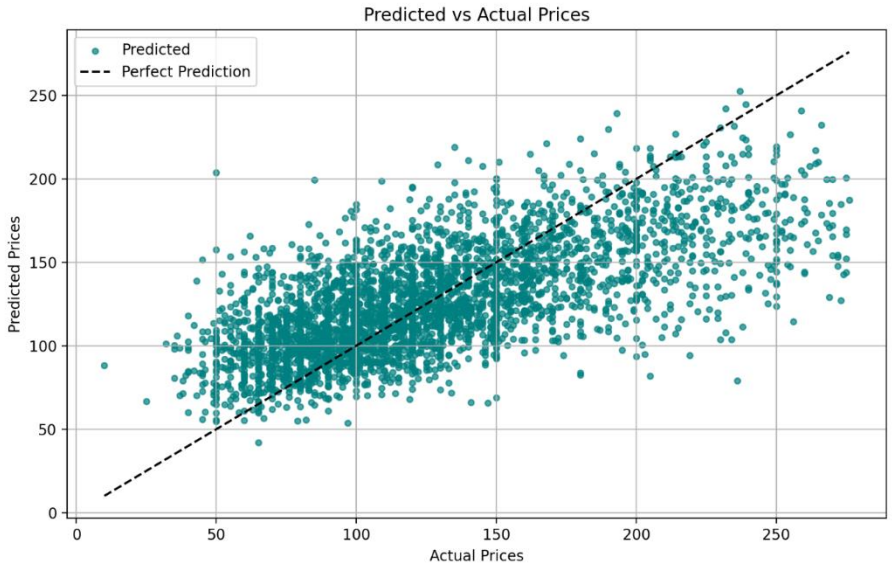
As per the visualization the decision was taken not to add this interaction in our predictor variables. For the rest, a lasso model was used on our most complex set of predictors (155 variables) to shortlist the ones with the best performance (112 variables).

Model Comparison

As mentioned earlier, there were 5 models that were created, namely Random Forest (parameters provided), Lasso Model, OLS Linear Regression, CART, and GBM, and their performance was evaluated based on their RMSE (Training Dataset)

Model	CV RMSE
GBM	36.80
OLS	37.45
LASSO	38.75
Random Forest	39.65
CART	42.46

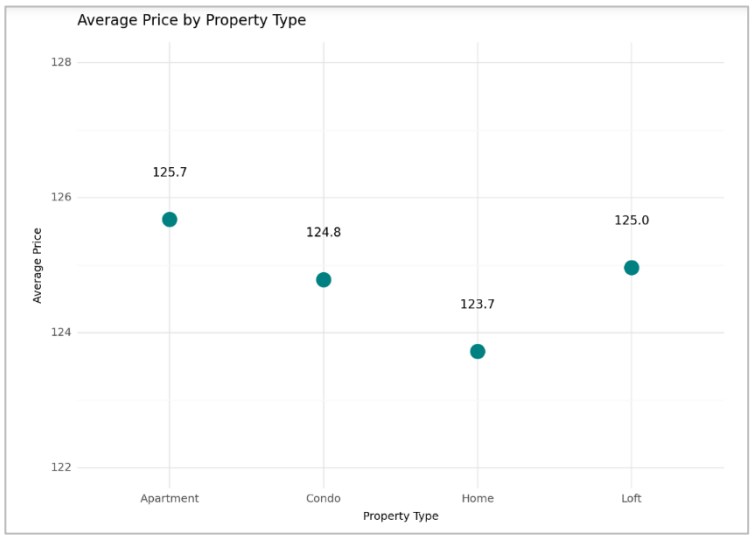
The best performing model was the Gradient Boosting Method with the lowest RMSE of 36.80 folloed by OLS with a RMSE 37.45, and LASSO with a RMSE of 38.75. Since it was the best performing model, gradient boosting method was also used in predicting on the Test Dataset (see below).



From the prediction plot, our model is good at predicting model, but if you notice that for actual price values above 200, our model under values the property.

Partial Dependance Plots

From the random forest model, I was able to capture some important information which may help the company in deciding their apartment size (see below).



Comparing the property types, the cheapest property type is Home while the most expensive property type is Apartment, although the difference is just 2 Euros. From the number of accommodates graphs, there is a very small increase in price when the number of accommodates increases from 2 to 3. But there is a drastic price increase when the number of accommodates go from 3 to 4.

Limitations

One of the biggest limitations in this project was that there was no variable available that represented the distance from the city center. Since Lisbon is a big tourist destination, most of the customers would be tourists. And the distance from city center would be one of the most important features. Another limitation is the number of observations that remain after filtering for our business objectives. Since only one source of dataset was used (Airbnb), I would recommend sourcing data from other websites that include different listings than we have.

Recommendations

From the plots, the company can easily see the average price of each property type as well as the average price for the number of accommodates. The recommendation to the company according to the this is that they should invest in apartments and lofts, 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.