Summary Report - Assignment 2

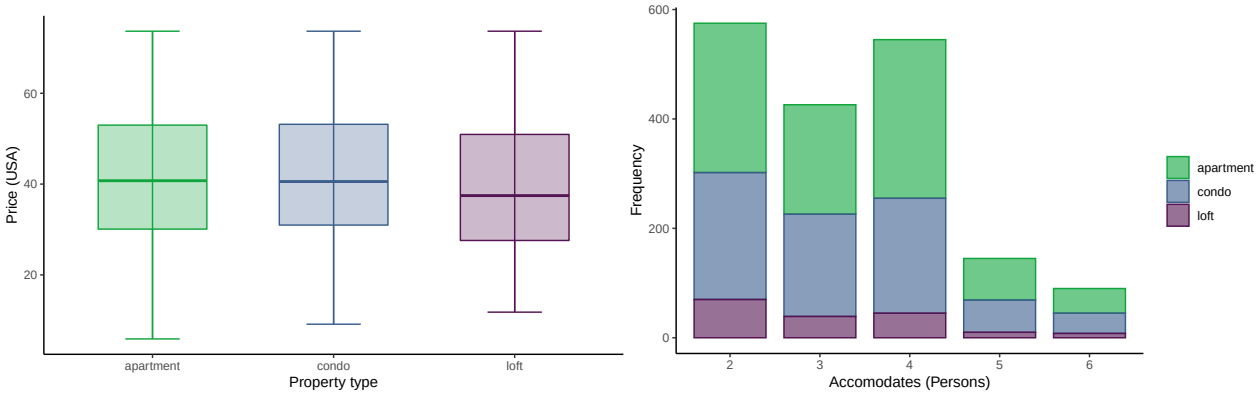
Maryam Khan

Introduction

The purpose of this project is to help a small to mid-sized firm price an apartment that can accommodate 2 – 6 people in **Istanbul**. To help the firm make this decision I built a price prediction model using various regression techniques and models. The data for this prediction analysis was scraped from the Inside Airbnb from 29 December 2021 to 31 December 2021. To make this prediction I will be looking at a few key predictors to make the prediction more accurate. These predictors some basic attributes about the property, reviews and ratings, a few variables regarding the host and various kinds of amenities that describe the apartment. For this project I will be using 5 different types of models which include the OLS, Cart, two different Random Forest and GBM. I will be using these models to pick my final prediction based on the prediction power looking at the RMSE and R squared values.

Data Cleaning and Preparation

To start working on the models, I had to start with some preliminary cleaning of the data. Before I began cleaning the data my raw data set contained 22,695 observations and 74 variables. I dropped some of the variables in the beginning that included urls and some descriptions that could not be used in the analysis. After doing some rudimentary cleaning I proceeded to sort out the amenities. For this I first split the amenities in different dummy variables and then they similar amenities were grouped together into one variable for example, the different WIFI and internet variables were all grouped together as one. This helped me in reducing the number of variables, initially, I had 2093 amenities which were reduced to 172 after being grouped together. For more details the data cleaning code has been uploaded here data_cleaning.R. The main variable of interest, the price, was also converted to USD as in the raw data the prices were given in Turkish lira. Furthermore, the all the observations with missing price values were dropped. The data was further prepared for the analysis by filtering for apartments, condominium and lofts that can host 2 to 6 people as that is the requirement of our company. I also imputed some of the variables that contained missing values. Variables like number of bedrooms, bathrooms and the number of beds had missing values. After imputation of these predictors I had a data set which had no missing values. For more details the data preparation code has been uploaded here data_prep.R.



Variables

Another enormous challenge for this project was the future engineering and deciding which variables to include in my modelling. I did this by dividing my variables into groups of 5:

- Numeric variables: These variables define size for example, number of beds, number of accommodates, number of bathrooms and bedrooms, the minimum number of nights required to rent the place.
- Factor variables: These are categorical variables that are either in string or numeric form. For example, the neighborhood the apartment is in or the type of property (apartment, condominium, or a loft).
- Dummy variables: These are binary variables mostly describing the amenities of a property like the WIFI, pool, beachfront facing property, kitchen supplies, elevator etc.
- Review variables: These describe the reviews and ratings characteristics for an individual property. Number of reviews, the rating for those reviews and the mean monthly reviews received.
- Host variables: These variables describe the characteristics of the host of the property. They are binary variables like host is a superhost or if the host is verified.

Exploratory Data Analysis

Once all the data was cleaned and the variables were grouped, we looked at the main target variable which is the price. The price distribution was close to normal, hence, there was no need of taking a log of the variable. The next step was to look for interactions, for this I checked the relationship between property type and some of the amenities and if there was a noticeable price difference then the dummy variable was included as an interaction term.

Modeling

Once all the data was prepped for doing our regression analysis, I divided the data into the test and train samples, where 80% of the observations were part of the train sample and 20% of the observations were part of the test sample. These test and train samples were used in all machine learning models. The detailed code can be viewed here data_prediction.R.

Linear Regression: I first started with building my OLS model and based on LASSO I selected all the non-zero coefficients and included them in my OLS models. The first model was only with n_accommodates variable, model 2 and 3 had 7 and 91 variables respectively and model 4 was the over-fitted model with all the variables. According to my OLS regression results, model 3 proved to be the best predictive model out of these 4. Even though the BIC value increases after the model 1, the increase in the value is not that significant from model 1 to model 3 and model 3 has the lowest RMSE in the training set. Furthermore, looking at the R squared value of the model we can say that this model explains 34.7% of the variation in price. When BIC and Cross validation produce conflicting results, we should prefer to side with the cross-validation results since it's not based on auxiliary assumptions (BEKES, 2021). Since the difference between them is small it will not be big mistake if I pick model 3 over model 1 and 2.

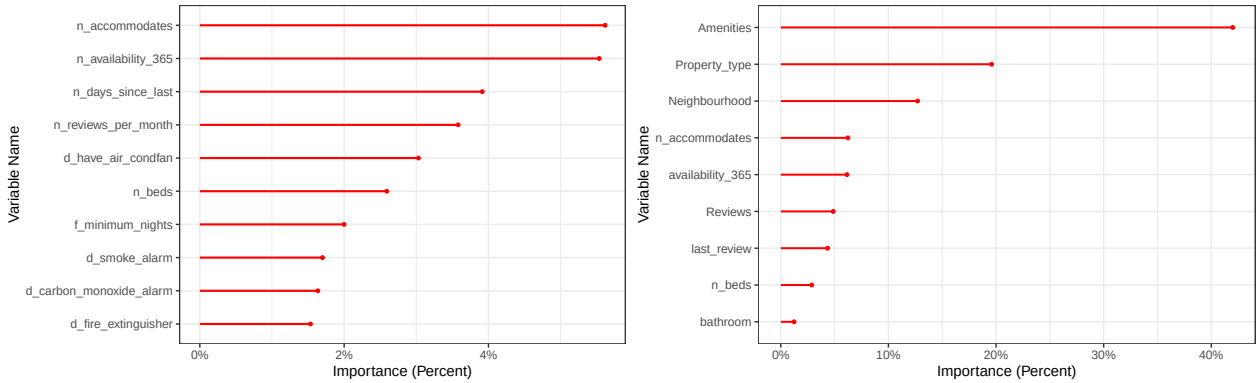
Random Forest: The Random Forest software creates hundreds of regression trees and combines them in one and gives an average. Since it uses bootstrap aggregation, the model is able to produce accurate results. For this regression model I will be using the same training and holdout samples. Based on my regression results the lowest RMSE value was at 5 for the terminal nodes and 12 variables in each node giving me an RMSE value of 12.4. I also used the auto tuned the same version of the model, in which the algorithm picks the variable and node values automatically. The cross-validated RMSE for autotune produced a cross-validated RMSE of 12.3 with 5 terminal nodes and 92 variables for each split.

Variable Importance: The variable importance helps us in identifying the important variables that impact price. In my analysis the room availability, number of accommodates, days since last review, reviews per month and if the property has an air conditioner or fan were the top 5. The top 10 variables according to

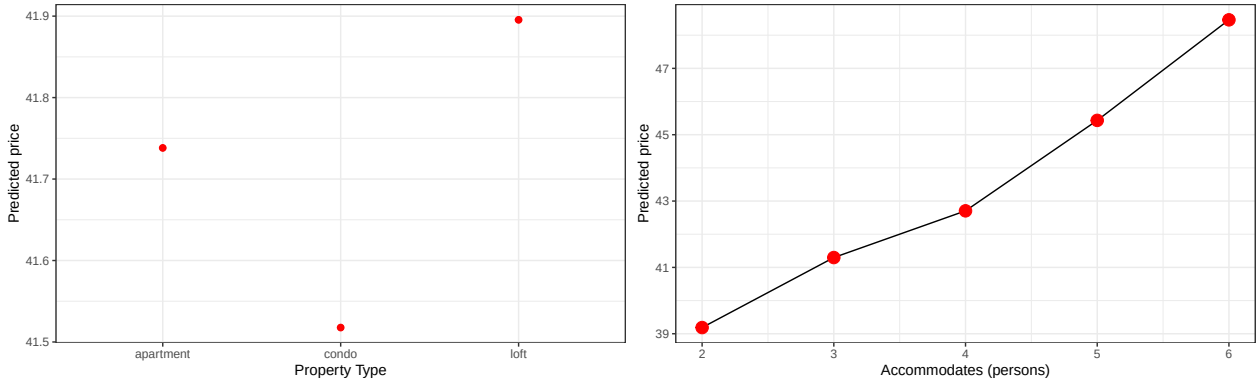
Table 1: Horse Race of Models CV RSME

	CV RMSE
OLS	12.6
CART	13.8
Random forest 1: Tuning provided	12.4
Random forest 2: Auto Tuning	12.2
GBM	12.4

importance can be seen below. Furthermore, I grouped the variables to get a bird eye view as seen in the graph below and amenities account for 50% of the variable importance.



Partial Dependencies: To further investigate I created partial dependency plot for the number of accommodates and the predicted price and as we can see in the graph below there is a fairly linear relationship between the number of people accommodated and price. We also created a plot with property type and price and as in the graph below it can be seen that renting a condo is more expensive than an apartment.



Conclusion

After conducting this analysis, according to the results my best model was the Random Forest model with autotuning as that gave the lowest RMSE value. However, since Random Forest is more like a black-box model it is difficult to explain the details of the regressions to clients, therefore, I will suggest that they pick the OLS model. The difference in RSME values of both models is 40 cents which is not that big of a difference to avoid complexity. Therefore, the client should invest in either condo or apartments as they yield higher prices and since we are able to predict the prices better for 5 beds they should invest in a property with 5 beds.