Business Report: Building Price Prediction Models For Copenhagen Apartments

Shahana Ayobi

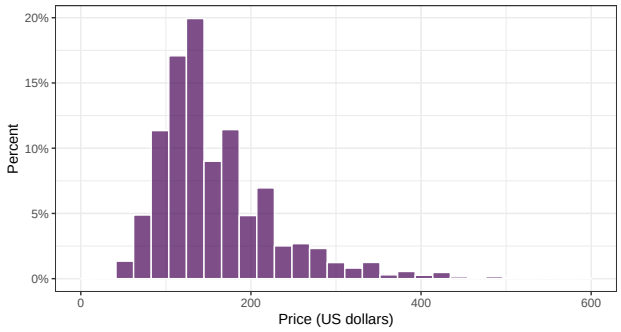
2023-02-12

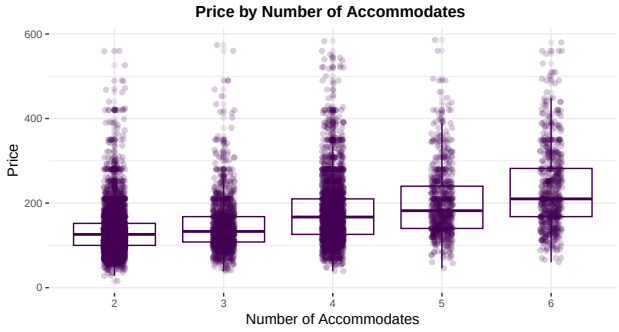
Introduction

This report's objective is to give a thorough explanation of the price prediction model. The primary objective of this project is to assist a business in setting a price for brand-new flats that have not yet hit the market. The data is gathered from Inside Airbnb, which can be obtained here, to develop a price prediction model for a business operating small and mid-size apartments hosting two to six guests in Copenhagen, Denmark. Five price prediction models—OLS, Lasso, Cart, Random Forest, and GBM, Gradient Boosting Machine—are produced as a result of data cleaning, munging, and analysis. The project's ultimate goal is to complete a better prediction model as measured by relative RMSE values.

Data Cleaning

The original data set consists of a single data table with 75 columns and 13,820 observations. The information relates to one-night rental rates for the period of December 29, 2022. Price per night, per person, expressed in Danish Krone, is the desired variable which is then converted to US dollars. For further cleaning, unnecessary columns were removed, the "amenities" column was transformed into binary variables. factor variables for predictors like neighborhood and property types were also created. The data was filtered for units that fall between 2 and 6 accommodates. Price per night included extreme values exceeding 1000 USD per night, which comprised fewer than 1% of the observations, therefore price was filtered to less than 600 USD and the observation where price is missing was dropped. As shown below the price distribution is skewed with a long right tail while log price distribution is close to normal. However, prediction is carried out on price per night for model simplicity, also the distribution of number of accommodates for price is also shown in the Figure where as the number of guests increases, prices increases as well.





Data Analysis and Feature Engineering

Feature engineering entails deciding on the type of predictor variables to include as well as the functional forms of predictors and potential interactions. Basic variables include the main predictors such as the number of accommodates, number of beds, property types, number of days since the first review, and its flag variable, and number of bathrooms. Basic addition includes key factorized variables like neighborhoods and minimum nights. Review Variables contain important guest review predictors such as review score rating and its flag indicating missing values, and factored total number of reviews. Polynomial level is made up of squared terms for number of accommodates as well as squared and cubic terms for days since the first review and dummies include binary values for all amenities. Three types of interactions were produced: X1 which includes property type multiplied by the number of accommodates. X2 contains property type, air conditioning, refrigerator, WIFI, coffee maker, microwave and hot water dummy variables and X3 includes property types times neighborhood, accommodates times neighborhood groups, and all amenities.

Modeling

The best model provides the best prediction in the live data. The original cleaned data is split into two random parts by 20% to 80% ratio in order to avoid over-fitting. The Holdout set includes 20% and the rest 80% is work data set. Then, the work dataset is split into train and test datasets and 5-fold cross validation is run on the train dataset. Then the best model is chosen based on the lowest average of 5 CV RMSE result. Eight basic OLS regression models from simplest to the most complex one were utilized to find the best model for our analysis. 5-fold cross-validation RMSE suggests that Model 7 regression has a better performance and it has the lowest RMSE value of 59.19 USD for the test set.

Models

It is critical to run and evaluate various models for a given data set. The following models and algorithms were used to
- **OLS** and **LASSO** CV-RMSE results using model 7

- **CART**, **Random Forest**, **GBM** using dummy variables for basic level variables, basic additions, review variables, and amenities.

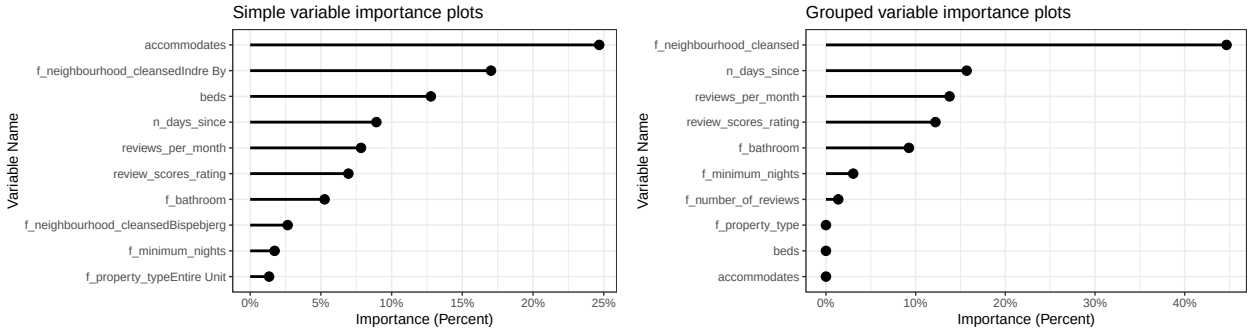
As a result Random Forest model has a relatively better performance for 5-fold cross validated work set, while LASSO works best for holdout set, as can be seen from the table of models below. The data's 5-fold cross validation RMSE is 58.79 dollars for Random Forest which is the lowest among all models, which is 0.2735 dollars less than the RMSE for LASSO. LASSO outperforms all models in Holdout set with RMSE value of 57.6185 dollars and 5 fold cross validated average RMSE of 59.0655 dollars which is lower than GBM model. It also gives the highest R-squared with value of 33.21% despite penalizing for interaction terms and shrinking some coefficients to zero. Considering external validity and model performance for live data, LASSO is chosen as the best model that has relative better performance in both sets.

Table 1: Evaluation of Models

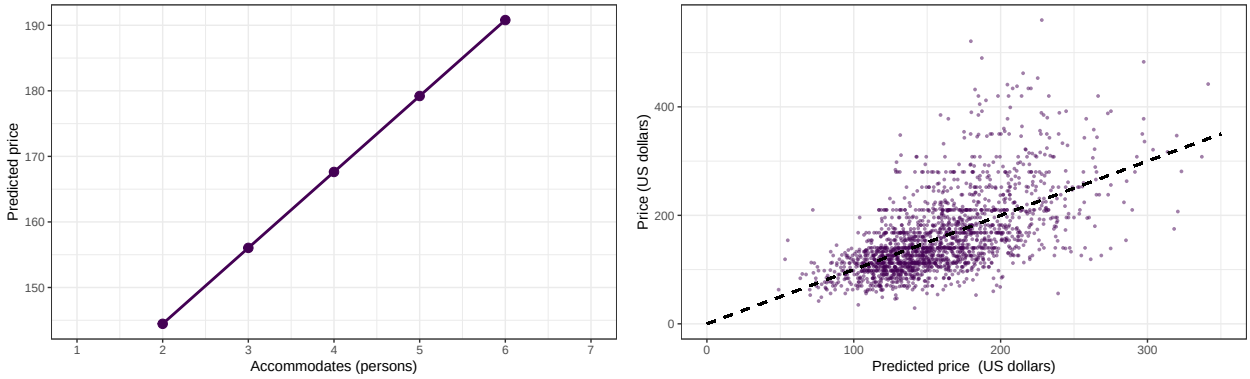
	CV RMSE	Holdout RMSE	CV Rsquared
OLS	59.2794	57.6941	0.3269
LASSO	59.0736	57.6750	0.3319
CART	62.3001	61.4069	0.2574
Random forest	60.0738	58.2892	0.3133
GBM	59.4176	57.5352	0.3247

Diagnostics

LASSO is an algorithm that fits a model by shrinking the coefficients and even shrinks some of them to zero is chosen as the best model. Diagnostic tools, on the other hand, can be used to uncover information about the patterns of association that drive prediction. Some examples are as follows: **Variable Importance Plot** depicts the average importance of fit when an x variable or set of x variables is used. Plot of variable importance for The top ten most important variables show that number of accommodates and beds, Indre By neighborhood, review variables are the most important. The importance of grouped variables reveals that neighborhoods, reviews, and days since first review are important.



Partial Dependence Plot depicts how average y varies for different x values in relation to all other predictor variables. The partial dependence plot is based on the holdout set's predictors. The partial dependence plot for the number of accommodates and the price shows that the price rises as the number of accommodates increases. **Actual vs Predicted Price** Another post-prediction diagnostic is a comparison of predicted and actual prices. The graph below shows that prediction performs better for lower prices than for higher prices. However, for prices above 200 dollars, the model seems to not fit well.



Conclusion

The purpose of this report was to develop a more accurate model for predicting Airbnb prices in Copenhagen for small to mid-size apartments. Five models were depicted in order to compare and contrast their performance. The basic LASSO model with lambda value of 0.01 , which highlights meaningful characteristics about the nature of Airbnb apartments in Copenhagen was the best model with a 59.065 dollars RMSE, while it performed best in the holdout set with a 57.618 dollars RMSE. While OLS came in second place with 57.64 dollars RMSE for holdout set. In this study, simple models such as LASSO and OLS outperformed complicated models such as GBM and CART. Therefore, LASSO is chosen as the best model; however, results can be different depending on the dataset. The neighbouhood, days since first review, property type, number of bathrooms, the number of accommodations, review scores rating and reviews per month are key price drivers based on post prediction diagnostics.