Data Analysis 3: Assignment 2 Report
Corinne Williams

Introduction

Working to predict the prices of small to midsize apartments hosting 2-6 guests in Austin, TX, USA, 4 different predictive models were used, each incorporating the same predictor variables for the outcome variable of price, in order to compare each model's effectiveness in prediction. As such, the models used were: (1) Linear Regression with OLS, (2) Linear Regression with LASSO, (3) CART (Regression Tree), and (4) Random Forest. After various data cleaning and feature engineering, in order to provide the best possible price suggestions for the client, the following features were selected as predictor variables: number of guests, number of bedrooms, number of bathrooms, review scores, property types, availability of the unit a year in advance, number of reviews, whether an Airbnb host was a Superhost or not, required minimum length of stay, and required maximum length of stay. The choice to leave out zip codes/neighborhoods as predictor variables is due to the fact that after analyzing the fits of the models, the RMSE values (measurements of fit) were better without such variables. Moreover, the decision was also made to leave out various interactions as their scatter plot distributions did not show much evidence of a relationship that would highlight variable dependence.

Reaching for higher external validity and accounting for variation, each of the models' RMSE values were cross-validated using a training set (70% of original data) and test set- in this case also the holdout set- (30% of original data).

Model 1: Linear Regression with OLS

The Linear Regression with OLS was a simple model using all the predictor variables listed above. The cross-validated RMSE value was found to be about $43.16, meaning on average, the predictive model is off by about $43.16 from the actual apartment prices.

Model 2: Linear Regression with LASSO

Using the same predictor variables, the Linear Regression with LASSO model had a cross-validated RMSE value of about $43.45, meaning on average, the predictive model is off by about $43.45 from the actual apartment prices.

Model 3: CART (Regression Tree)

Using the same predictor variables and creating 50 fits (5 folds evaluated 10 times, each time with different parameters), the Regression Tree model had a cross-validated RMSE value of about $46.79, meaning on average, the predictive model is off by about $46.79 from the actual apartment prices. This was the highest cross-validated RMSE among the models, accounting for the worst fit.

Model 4: Random Forest

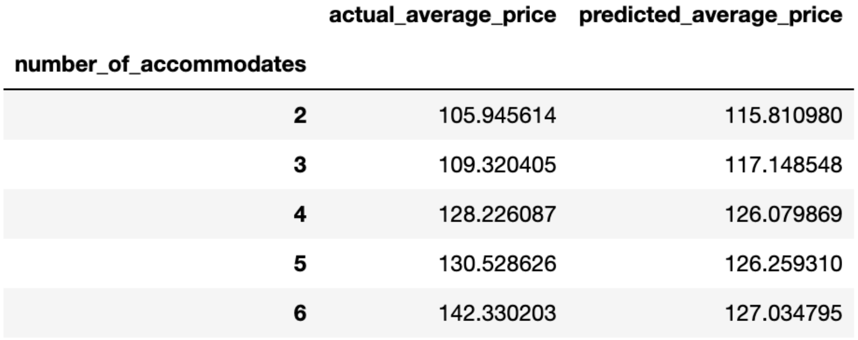
For this model, the same predictor variables were used as well. 60 fits were created (5 folds evaluated 12 times, each time with different parameters. 6,8,10, and 12 features were considered for splitting each tree node, and the minimum number of observations that would be required in a node were set to 5,10, and 15. The best parameters for the random forest were 8 predictor variables with terminal nodes having a minimum of 5 observations. The cross-validated RMSE was the lowest among all of the models (best fit), with a value of $38.64, meaning on average, the prediction model is off by about $38.64 from the actual apartment prices.

Predictor Variable Importance

Each predictor variable was then tested for the amount of importance in predicting the outcome variable. Availability of the unit a year in advance was the most important variable in predicting price, followed by the number of reviews, and property type (namely, the entire rental unit). This is something to consider when pricing the apartments on the market in real time.

Prediction and Partial Dependence Plot

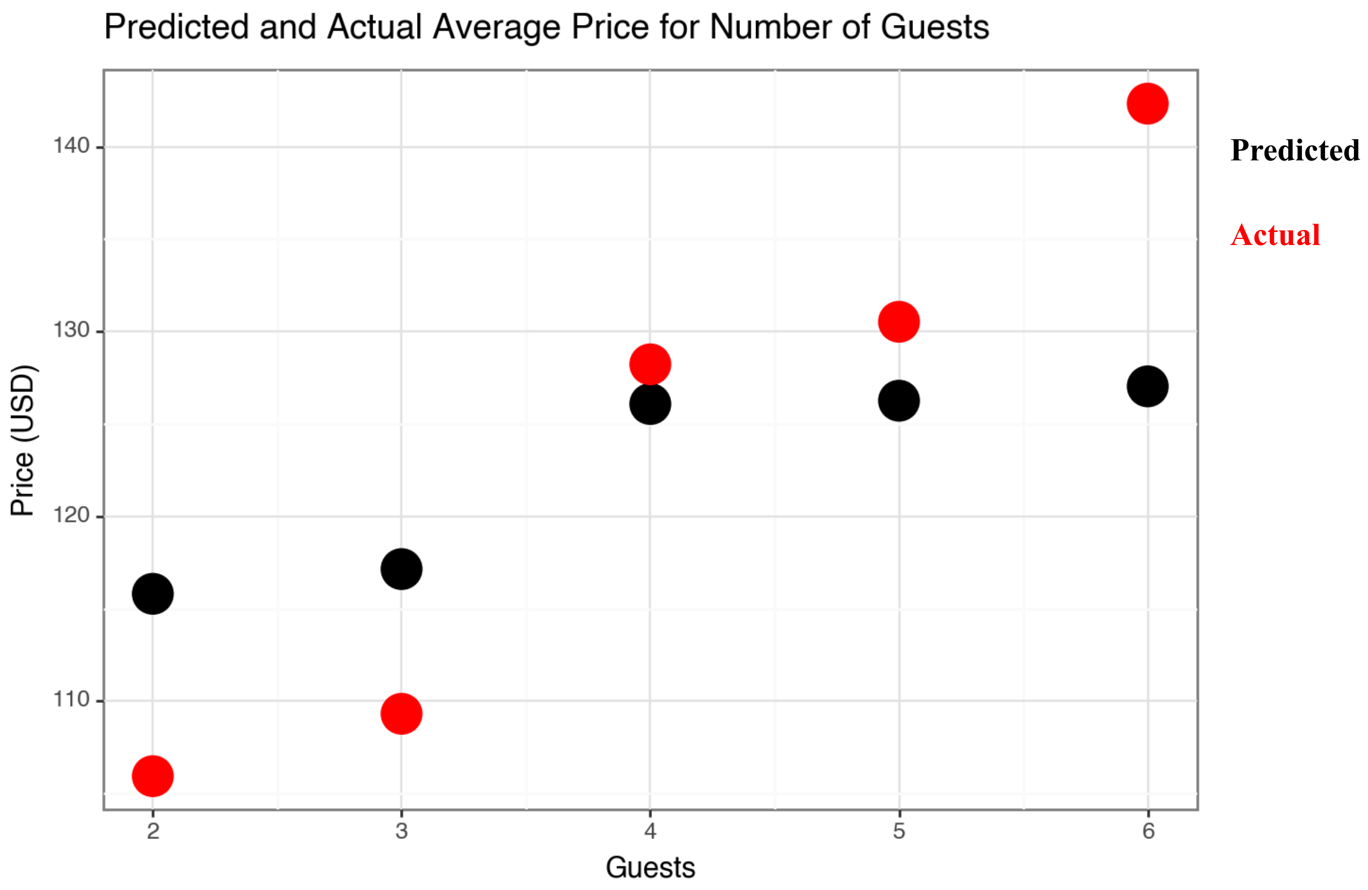
As the client is interested in the pricing of their apartments according to size of the unit (small to midsize), specifically those fitting 2-6 guests, using the best model (Random Forest), the price of apartments was predicted using the holdout set, keeping in mind the number of guests. A data frame was created to display the actual average price of apartments accommodating 2-6 guests, versus Random Forest's predicted average prices of apartments accommodating 2-6 guests.



	actual_average_price	predicted_average_price
number_of_accommodates		
2	105.945614	115.810980
3	109.320405	117.148548
4	128.226087	126.079869
5	130.528626	126.259310
6	142.330203	127.034795

From the dataframe, one can see that the predicted average prices (predicted by the Random Forest model) are not far off from the actual average prices provided by the data set. In fact, the predictions for apartments accommodating 3-5 guests seem to be very close. To account for the discrepancies with predicted average price for 2 and 6 guests respectively, there could be many factors coming into play, and further research would need to be done to properly handle the discrepancies.

Using the same information as above, a partial dependence plot was created to visualize the effects of the variable, guests, on the outcome variable, price, holding all other predictor variables constant.



Interestingly, the model seemed to overfit the data with 2-3 guests, and underfit the data with 4-6 guests. This could be due to a number of different things such as the model being too complex for 2-3 guests and too simple for 4-6, the possibility that certain features and/or interactions weren't accounted for that might explain the discrepancy, etc; however, further research would need to be conducted to figure out the cause.

Summary

The client should be advised to price the apartments in Austin, TX somewhere between $115-127 a day, depending on the number of guests it can accommodate. Further recommendations come from the importance of the predictor variables. It is advised that the client should pay attention to adjust the price of the apartment based on availability a year in advance. Number of reviews being the second highest variable of importance may not seem like it applies to this case; however, it is important as the client could decide to reprice the apartment once it gets to a certain number of reviews (if placed on the Airbnb market). Further research would need to be done in order to see what the minimum number of reviews would be in order for the company to reprice their listing. Lastly, as the property type consisting of the entire rental unit was of the third highest importance, it would be beneficial for the client to list the entire property rather than renting out a private room or section of the unit.