Individual Project

**Set-Up**

In my ‘Set-Up’ and ‘Data Loading’ sections, I loaded in the required libraries and removed the street address and description. I also set the seed to reproduce results. Ultimately, my results were not reproducible, and almost every time I ran my models, I received new predictions.

**Exploratory Data Analysis**

In this section I graph the x variables against the latestPrice to get an idea of how the variables are related, as well as which variables are suitable as categorical variables.

**Feature Engineering**

I by far spent the most time in this section. I tried making many different variables categorical, and then reverted them back. I tried summing together different variables and manipulating them. Since my results weren’t very reproducible, this was difficult. I settled on adding together the internal features to make a holistic numOf feature. I also turned zip codes into areas around Austin, hoping that different areas would represent a higher cost of living. I also turned the year built into age of the house, hoping that it would help define the age. Finally, I found a household income per zip code data set from the Census Bureau and assigned an income to each zip code. I then removed zip codes.

**Splitting and Control**

I used this section to split my testing and training set as well as define my control.

**Repeat of Problem 4**

I redid all the models from problem 4 of the homework, and found that in the homework problem my models were getting progressively better, where as in this case my models were sporadic. In some iterations they were better but on the next run they would be worse, again possibly due to the seed.

**Extra Models**

I ran a boosting, forward selection, backward selection, stepwise selection, selection with interactions, ridge regression, lasso regression, and a PCR model to test my models. I found none of these models were significantly better than the previous problems.

**Results**

I found that if I took the average of most of the models, my predictions produced a smaller OOS MSE. Therefore, I did the same when calculating my predictions. All in all I’m fairly disappointed with the way my model performed, I attempted in depth and thoughtful feature engineering and found that it didn’t help my models perform better than the homework problem simplified models. I did find that the median income variable that I added was a very strong indicator of house prices, for obvious reasons.

**Group Project Contribution**

In the group project, I produced the models for Decision Tree Regression, Boosting Regression, BART Regression, Boosting Classification, Random Forest Classification, Decision Tree Classifier, and Bagging Classification. The slides were evenly split between the members. I compiled and submitting the markdown of everyone’s code. I also helped interpret results for other members. All in all everyone contribute a fair amount to the project and I was happy with my team.