Ben Rohlfing- STAT 490 Project Writeup

**Abstract:**

Before performing analysis on this dataset, I set out to learn what general factors can lead to higher house selling prices in King County, WA. Initial predictors were eliminated before performing various regression techniques due to general uselessness and multicollinearity. A combination of feature selection and ridge regression and LASSO were then performed in order to try and eliminate relatively insignificant predictors due to the size of the dataset. Then a regression tree was developed along with a random forest for comparison as to determine some of the most significant predictors. After some more analysis, it was determined that the final model would include 13 predictors (Subject to change), including sqft\_living, grade, and surprisingly latitude being some of the most significant predictors.

**Introduction:**

The dataset is based on the selling prices of 21,613 homes in King County, Washington (Seattle and Seattle Metro Area) from May 2014 to May 2015. There are initially 21 predictors used to try and predict the housing prices `price' in this area. Some of the predictors, such as "id", "date", and "zipcode" seem to have no real bearing on this model. Other predictors, such as number of bedrooms and bathrooms per house, seem to have more bearing on predicting home prices, though. Other variables include how big the house is in terms of square feet, how big the area around the house is, how many bedrooms/bathrooms are in the house, what kind of view it has, the condition and construction quality of the house, the year it was built/renovated, its location within the county in terms of latitude and longitude, and the area in square feet of the surrounding 15 homes. In terms of which predictors I think will be the most significant, I would have to think the amount of space a house has would be very important. I also think the year it was built/renovated would also influence its price, along with the quality the house is currently in, especially in terms of its condition.

**Analysis**

Before performing any real analysis on the dataset, I first performed some data cleanup and adjusted some variables to make them cleaner and more palatable to a regression model. The first thing I did in this pre-analysis phase was to eliminate some unnecessary variables, including “id”, as this number meant nothing to the home price since its just random identification of each house. The “zipcode” predictor was also initially discarded, as the zipcode numbers throughout King County didn’t seem to follow a specific pattern and were all over the place. The date of each house also didn’t factor into the final dataset, but it was used to adjust a couple of other variables, and thus was not completely discarded. Before adjusting any other variables, I found a typo in the dataset where the number of bedrooms of the house was accidentally entered in as 33 while there were only two bathrooms in the house. Seeing as this is implausible, I immediately discarded this point from my dataset.

Next, some variables were adjusted. Since the predictors “yr\_built” and “yr\_renovated” were recorded via the actual year these events occurred, inference becomes difficult. So, I took the substring the year that was recorded in the date (since it was recorded as a string), made it numeric, and subtracted this number by the year it was built/renovated, depending on the variable in question. Now, these predictors were recorded in terms of the number of years since these things happened, with 0 saying that it happened in the year the house was bought. This made inference on these predictors simpler to perform and understand.

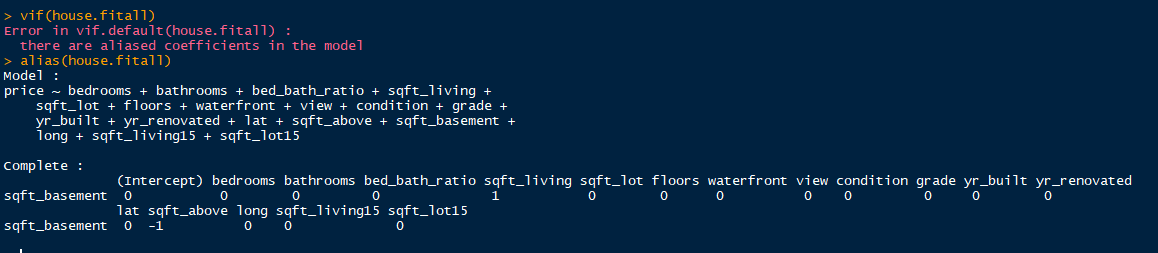
Another predictor, called “bed\_bath\_ratio”, was created by dividing the number of bedrooms in each house by the number of bathrooms in each house. This way, what seems to be an important factor in home prices was created and can be analyzed further.

I tried to avoid transformations whenever I could, but the “sqft\_lot” and “sqft\_lot15” variables were distributed in such a way that a 1/X transformation seemed to be necessary for both of them. This was done to try and reduce the range of X values these predictors had, as they were skewed heavily to the right with a thick tail on both of them.

Finally, I turned both the latitude and longitude predictors into dummy variables, as these values alone don’t seem to be very intuitive in interpreting or predicting. So, for latitude, I divided the county into two sections: north of 47.5 degrees (including city of Seattle, denoted as a 1) and south of 47.5 degrees (everything south of the city of Seattle, denoted as a 0). For the longitude, I once again divided the county into two sections: west of -122.1 degrees (urban section of King County, denoted as a 1), and east of -122.1 degrees (rural part of King County, denoted as a 0). Given these distinct sections, latitude and longitude seem to be easier and more intuitive to interpret.

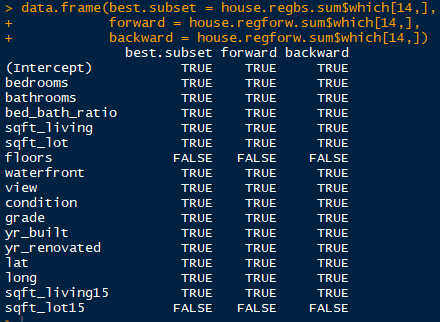
I created a second identical dataset with the same adjustments for the ridge regression/LASSO techniques so I wouldn’t have to eliminate the predictors in the original dataset immediately. I then separated both datasets into training and testing data using a 70% cutoff after applying all the changes made to some of the predictors.

Once this was done, I fit all 18 initial predictors into a model. When this happened, pretty much every single predictor was significant to a certain degree. With this many predictors, I tried to check for multicollinearity, but when I tried to run the ‘vif()’ function, it couldn’t because there was a linear dependency among the predictors. The ‘alias()’ function was run to determine which variables this applied to. The variables in question turned out to be ‘sqft\_living’ and ‘sqft\_above’. Considering ‘sqft\_above’ + ‘sqft\_basement’ = ‘sqft\_living’, this does not come as a surprise.

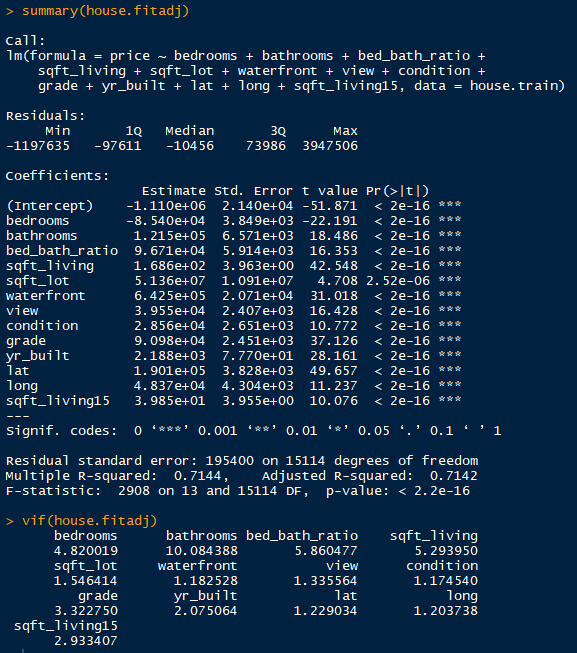


So, in order to figure out which of these variables to eliminate, I ran a ridge regression and LASSO on this initial dataset. When these were run, ‘sqft\_above’ was found to be significantly less important than ‘sqft\_living’ based on the fact that both techniques forced the ‘sqft\_above’ coefficient to be zero about 10 predictors faster than ‘sqft\_living’, while ‘sqft\_basement’ was the first predictor coefficient forced to zero. Based on these results, ‘sqft\_above’ and ‘sqft\_basement’ were eliminated from the model, as ‘sqft\_living’ already accounts for the information in these predictors, while being more significant to the prediction power of the model. The RMSE for both methods was just under 196,000.

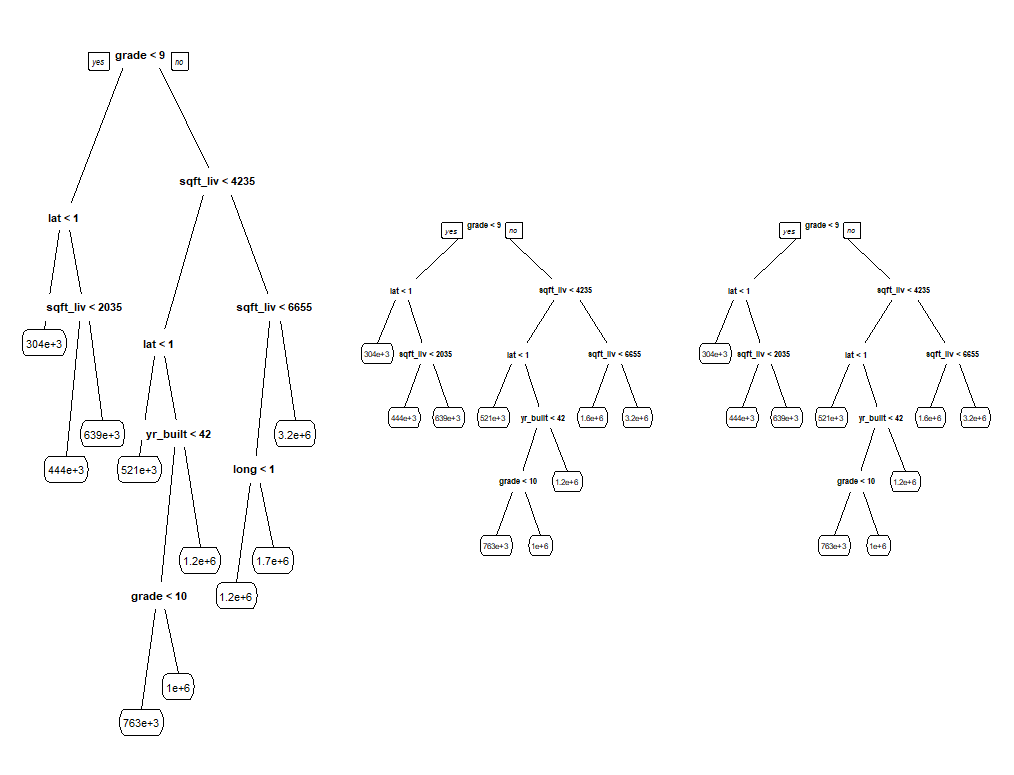
After getting rid of these predictors, I ran the remaining 16 predictors through various feature selections, such as best subset selection, forward selection, and backward selection, using the metrics of adjusted R2, Cp, and BIC for each method. When examining the results of each method, all of the metrics, especially the BIC metric, suggested that only 14 predictors be used in the final model, with all of the metrics within every method agreeing that the predictors ‘floors’ and ‘sqft\_lot15’ should be eliminated. The same 16 predictors were run again through ridge regression and LASSO, and while the predictors eliminated in the feature selection were eliminated fairly early with the ridge regression and LASSO, the predictor ‘yr\_renovated’ was eliminated first, suggesting that this predictor is relatively insignificant. This along with the fact that ‘yr\_renovated’ would have been the next predictor eliminated using the spectrum of feature selection features lends further evidence to this predictor’s relative unimportance, and was thus eliminated from the final model.



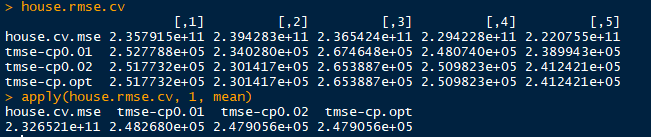
An adjusted model was then fit using the remaining 13 predictors. The summary of coefficients yielded mostly reasonable results. However ‘bedrooms’ yielded a negative coefficient in this adjusted model, even throughout the ridge/LASSO process, and ‘yr\_built’ yielded a heavy negative coefficient in the adjusted model, which is consistent with the ridge/LASSO results. I checked for any lingering multicollinearity issues to try and explain these coefficients. The ‘yr\_built’ predictor did not suffer from this, but the ‘bedrooms’, ‘bathrooms’, and ‘bed\_bath\_ratio’ have higher variance inflation factors, especially ‘bathrooms’, as it was around 10. This led me to perform further investigation into which of these three I could potentially eliminate from the model for the sake of simplicity.



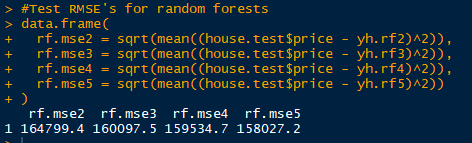
So, the next thing I did was to create a regression tree, prune it, and then create a regression forest, all for the sake of determining which predictors are significant/insignificant, including the three mentioned in the last paragraph. Whenever I developed the initial regression tree, it gave me 9 splits, with the predictors ‘sqft\_living’, ‘grade’, ‘lat’, and ‘long’ being the only predictors with tree splits. Seeing as how this tree is bushy, I tried to prune it a little bit using the techniques learned in the class. More specifically, I found the smallest tree with the highest cross-validated error less than the sum of the smallest cross-validated error and the smallest standard deviation of the cross-validated error. When I did this, I was able to cut the number of branches in the tree from 9 to 8. I further refined this by calculating the geometric mean of the cp of the trees with 7 and 8 branches.



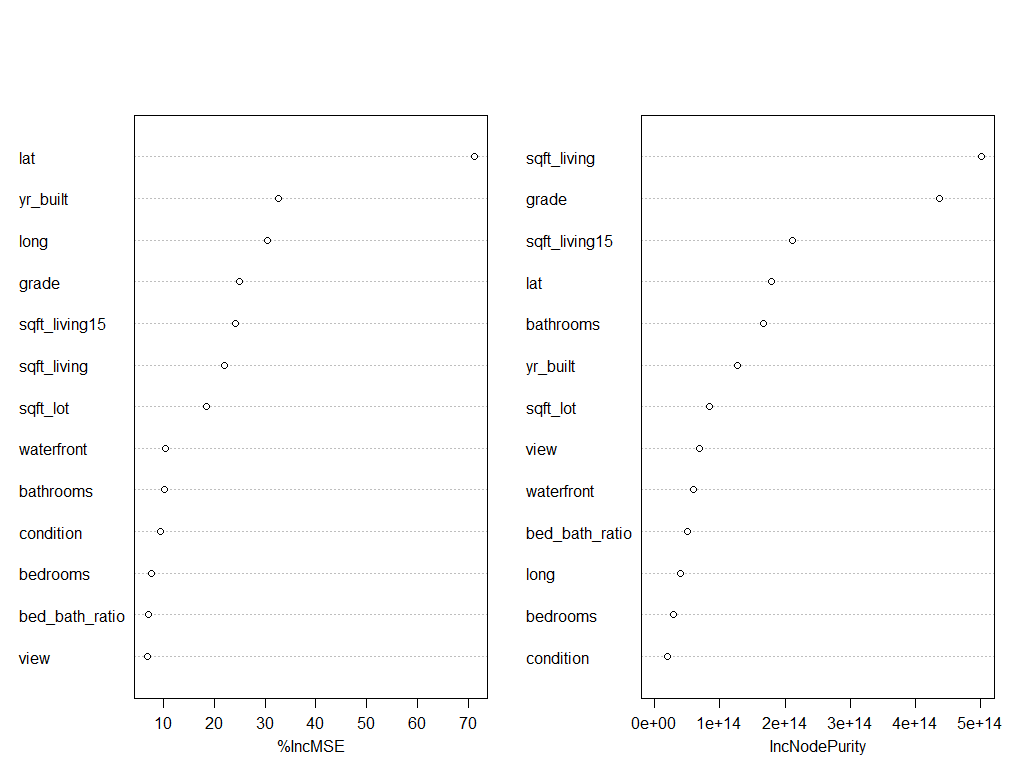
I then calculated the test RMSE for each of the three trees. This really didn’t yield promising results, as the test RMSE stayed roughly the same, no matter the cp value. Five-fold cross-validation was also performed on these trees, with the test RMSE being calculated for each of the five folds in each tree. An average was taken for each tree, and the tree CV RMSE’s were significantly lower than the regular CV RMSE, thus rendering the test data necessary.



In order to attempt to obtain better results, I developed several regression trees, with the number of predictors randomly selected at each node varying from two to five. Using the out of bag test RMSE’s for each individual tree, I determined that 150 trees would be a more than sufficient number for the number of trees in each forest. This way, there are enough different trees that calculating the overall test RMSE won’t be influenced by just a few predictors. I calculated these test RMSE’s for each forest, and each one showed a significant drop from just one regression tree (over 50%), with 4 predictors yielding the lowest test RMSE.

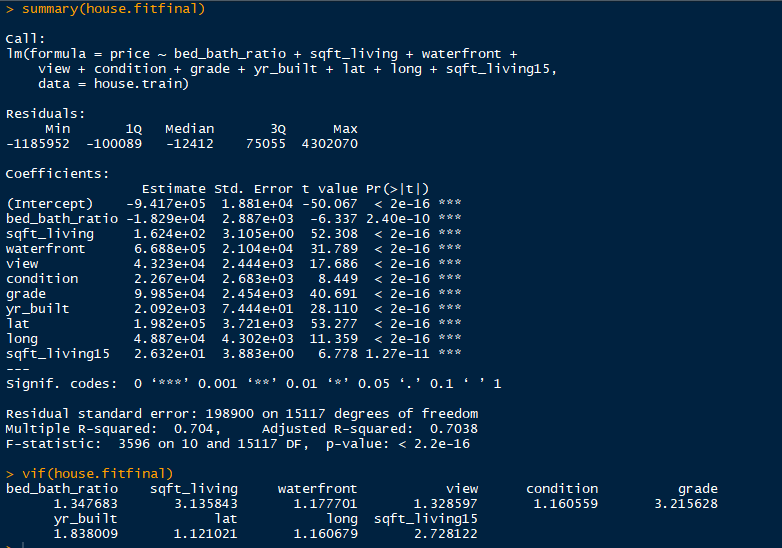


This adds validity to the model and forests, and looking at the percent increase of each predictor lends credibility to the significance of the latitude and longitude predictors. I was initially skeptical about their inclusion in the model, especially after changing them to dummy variables, but no matter what method I’m using, especially this one, it says these two are significant. So these two are staying. ‘Grade’ and ‘sqft\_living’ also proved to be important, lending further evidence to their inclusion in the final model. However, the bed/bath predictors vary in relatively insignificant importance for each forest, meaning that I can’t reasonably eliminate any of them due to the lack of evidence from the forests that one predictor is more important than the other.

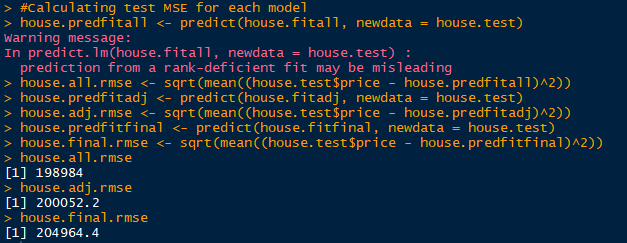


With the bed/bath predictor saga not settled, I decided to run ridge regression and LASSO one last time on the adjusted model with 13 predictors. Both of them suggested that the predictor ‘sqft\_lot’ seems to be fairly insignificant when running this, as this is one of the first predictors eliminated using both methods. This makes sense, as this predictor is the least significant in the adjusted model, and had been eliminated from the model fairly early in previous runs of ridge and LASSO. This predictor was thus removed from the final model.

Another model was then run without ‘sqft\_lot’. All of the predictors seemed significant, but the ‘bedrooms’ predictor coefficient is still very negative, even after fitting another model without ‘bathrooms’ and ‘bed\_bath\_ratio’. Only whenever I also removed the ‘sqft\_living’ predictor did the ‘bedrooms’ coefficient become positive, which tells me this predictor was causing this phenomenon. Seeing as though ‘sqft\_living’ has been a more significant predictor throughout this entire process, ‘bedrooms’ was eliminated from the final model along with ‘bathrooms’, as these two are closely related. Every one of the coefficients except for maybe ‘yr\_built’ at this point in the model made sense, and little multicollinearity seems to exist at this point as well.



Finally, for each of the three models (18 predictors, 13 predictors, 10 predictors), I calculated their test RMSE’s, and while they slightly increased as predictors were being eliminated, the increase is so small that it accounts for the ever so slight decrease in model accuracy, but overfitting is most likely not as much as a problem anymore, so this is not a problem.



**Results**

The predictors that I found to be most important in predicting the selling price of homes in King County, WA were ‘bed\_bath\_ratio’, ‘sqft\_living’, ‘waterfront’, ‘view’, ‘condition’, ‘grade’, ‘yr\_built’, ‘lat’, ‘long’, and ‘sqft\_living15’ to varying degrees. Here are interpretations of a few of the most significant variables:

* Grade: While holding all other predictors fixed, as the grade of construction of the house goes up by one point, the price of a home goes up by about $98,850.
* Sqft\_living: While holding all other predictors fixed, as the square footage of a house goes up by 1 foot, the price of a house goes up by about $162.40.
* Lat: While holding all other predictors constant, as the house goes from the southern part of the county to the northern part of the county, the selling price of the house will go up by about $198,200.
* Long: While holding all other predictors constant, as the house goes from the eastern part of the county (rural) to the western part of the county (urban), the selling price of the house will go up by about $48,870.

The latitude and longitude predictors are of much intrigue here considering that I transformed them into dummy variables instead of leaving them as relatively meaningless numbers. Their significance actually increased whenever this occurred. The ‘yr\_built’ predictor is also interesting in the fact that it is positive, even throughout the ridge/LASSO procedures, even though the higher units for this predictor indicate older houses. I feel like this is due to factors that were left out of the final model, including the year it was renovated, as I can see older houses being more likely to be renovated than older houses. However, this was eliminated early on from the model due to its relative insignificance. Other factors that may have influenced this include what neighborhood it was in, which would be difficult and time-consuming to figure out.

**Limitations of Study and Conclusion**

As described in the last paragraph, specific neighborhoods would prove to be very hard to track and determine which neighborhoods are more upscale than others, even with the zipcode data. Also, this model really only works for King County, as the latitude and longitude cutoffs were so specific for this county such that they actually seemed to matter when creating the dummy variables. Most other regions would not have this kind of cutoff, if any. Thus, if this model were used to predict home selling prices in any region of the country, it would be inaccurate, and latitude and longitude would probably have to be removed entirely to create a more accurate model for any general case. But for this specific region, latitude and longitude are fine. There are no other variables that seem to be region specific though, especially the construction quality of a home and its size.