Stat 502 Class Project Weekly Updates

Competition: King County real estate price prediction competition

Team Kaggle Name (if Relevant): _____Y_____

Team Members: _____Eryn Blagg,__Yiqun Jiang_____

Week 10 (March 23-27) Summary of Activity and Progress

Preprocessing: we transformed waterfront and zip code to be factors and delete date and property.

Model: This week we tried to use a simple linear regression model to predict house price but one problem we meet is with this method, we may get negative predictions. It's commonsense that house price should be positive so we modify our model a little bit to be absolute linear regression model. We forced the prediction to be positive to get more reasonable result.

Future plan: one thing we can do to improve our results I think is input engineering. Besides we would try more models.

Week 11 (March 30-April 3) Summary of Activity and Progress

Now we tried using glam to use as a smoothing to some of the variables. Our goal with this is to try and get rid of the negative values that we have in our first attempt at a model. This makes sense as it would be the next logical step in keeping our original model to be a linear model. For our results, we no longer have any negative prices on our houses, but we still have a relatively high error rate, but its nice to see that we no longer have negative prices in our models prediction.

Week 12 (April 6-10) Summary of Activity and Progress

We used log(price+1) to be our prediction object this week and for the preprocessing part, besides what we did before, we also deleted yr_built and yr_renovated columns and derived a

new column from 2020-max(yr_built, yr_renovated) representing age of the house. And we applied linear regression on the modified data set. After that, we applied random forest and kNN on it too but get larger errors. So our next step may be try some other methods and try to combine some results together to see whether there are some improvements.

Week 13 (April 13-17) Summary of Activity and Progress

At the beginning of this week we got kicked off of kaggle, so a lot of time has been spent on hold with them. But that has now been resolved.

This week we seemed to have hit a plateau on our error. After we combined the age of the house last week, we re-ran our models of knn, random forest, elastic net, etc. to compare to our linear model that we have as our current lowest estimate on our predictions on Kaggle. These models are coming up with estimates are in the range of [0.22, 0.3], which is much higher than our 0.18. So, now we are taking a new approach. At the end of this week we have been starting to separate some of the variables into similar groups: spects of the house, location, etc. and finding models based on each of these subgroups and trying to stack the models to see if that helps our process and our predictions.

Week 14 (April 20-24) Summary of Activity and Progress

we tried to boosting our best result. We fitted a random forest model for the residuals. With λ equals -0.1, -0.05, 0.05, 0.1, the prediction errors of lm Fit Values+ λ *Residual Fit Values are 0.1914708, 0.1961755, 0.1883854, 0.1907908 respectively, which is higher than original error rates. We expanded the range of λ, but still, no results get better than original ones.

Week 15 (April 27-May 1) Summary of Activity and Progress

This week we tried to add date column to the regression model to see whether it helps to improve the results and we found the result increased a little bit from 0.18681 to 0.18583. And we also tried to put longitude and latitude back but error rate increases. It seemed that current model is the beat we could do for linear regression. The stack model failed due to limited computing power. We kind of got stuck at this point.

Week 16 (May 4-May 8) (Finals Week) Summary of Activity and Progress

This week we summarized all results we have and wrote the report based on that. We made a PowerPoint and recorded a video presentation.