Stats 202| Homework: 8 | Kaggle Competition Discussion
Team: AIRealtors

Team Members:
Sagar Ganapaneni: sagar123, Matt Pecevich: mpecevic and Delbar Yousefi: delbary

Best Model Submitted: Ridge Regression with Lambda derived from CV after removing removing Outliers in the test data.

Approach:
Our team approached the Kaggle competition in three stages:

1. Data exploration and cleaning,
2. Model building/predictor understanding and
3. Model tuning and improvement.

Our team explored the data in R using histograms, box plots, and the summary command. Immediately we noticed a few potential outliers and high leverage points, we then moved to simple linear regression and used the residual plots as justification to remove these points, most of which had high sales values, very high SqFt values or extreme acreage values.

With a clean data set we created models of increasing complexity to understand the most important predictors and to get a feel for the most successful method. Our models started at simple linear regression (including best subsets) but advanced to polynomial regression, ridge regression, lasso regression, random forests, boosting, and generalized additive models with splines.

We submitted each of these models but achieved the best performance with one of the earlier models, ridge regression. In this step we learned the most important predictor is SqFt. The other predictors improve the estimate but every type of model included SqFt.

In the model tuning stage we tuned our parameters, better examined our training data and attempted to improve our existing models by double checking for more high leverage points or outliers, bootstrapping the parameter estimation in our best performing models, and incorporating some interactions.

Unfortunately, we were not able to improve upon the simple ridge model and since our sqrt(error) is nearly double that of the top teams we assume we are missing something very big in the model, like a different model type or a certain predictor.