Final Project Report
Ryan Rizzo
Statistical Learning

My approach for the Allstate Kaggle Competition was to compare two approaches: Ridge and LASSO regressions, and see which one would fare better. It ended up that LASSO achieved better scores through cross-validation, although not by much.

Some of the variables in the dataset had missing values which messed up the newX section of the main for loop, so I did not include those variables in my model. I included all continuous variables since there were only 14 and those data points seemed a little bit more valuable when looking at graphs of their correlation to loss. For categorical variables, I looked at boxplots mostly as well as the LASSO coefficients vs log(lambda) graph which we have gone over a lot in classes and on exams. I tried to make sure I included all of the variables that stood out on that graph, as well as some other categorical variables which seemed to show tendencies on the boxplots.

The cvfit aspect of the code was the bottleneck which disallowed me putting all the variables into the model. It just simply took too long for all the lambda values to run through all the variables. I also attempted a LOESS smoother but I moved away from that method because it didn’t seem to be doing any more good than the LASSO and Ridge regression models.