Team N
Ben Czekanski

The competition that I chose to enter was the Liberty Mutual Fire Peril competition. This competition was determining the severity of insurance claims. What made this competition unique was that most of the claims were zero, and so the score was a Gini Coefficient rather than the MSE that we are used to for continuous outcomes. Fortunately, I found a function on the Kaggle website that could calculate the Gini Coefficient easily.

The data for this competition consisted of a target outcome variable and more than 200 masked predictor variables for the insured, including data about crime, geographic and demographic status, and weather measurements. Due to the large number of variables, I initially attempted to use Principle Component Analysis. This turned out to be difficult because of a lot of missing data, and when I was able to implement PCA, it did not significantly improve my results. The main method that I used to create my predictions was CART, and I used the complexity parameter as my "knob". I performed cross-validation to find the optimal cp for the data I was predicting. I thought that CART was a good model for this data because the predictor variables were both categorical and continuous, which is something that CART can handle.

The large size and high dimensionality of this dataset made running my cross-validation very time intensive, and as a result I predicted only based on the basic 17 predictor variables, rather than including all 200+ variables. Including more variables should definitely have improved the fit of my model, but it took more than an hour to run as it is. Because of the long time that my code took to run, I have included a plot of my cross validation results. Another drawback to the data provided was that the meaning of the predictor variables were hidden, which made it difficult to perform EDA, as there wasn't really any meaning to the variables that I was showing.