Notes on process

Phil Hoxie and Otto Nagengast

Final project

Statistical Learning

23 May 2017

We optimized LASSO, KNN, and CART to generate predictions. The outcome variable of interest in the Prudential Life Insurance competition is a factor variable of insurance risk with eight levels. We used LASSO because it is a regression model and it has a continuous outcome. We then assigned these continuous outcomes to the closest integer. We used KNN and CART because they are classification models.

The training data consist of 59,381 observations and 128 variables. This made optimizing KNN and CART time-consuming. Phil had the insight to use Principal Component Analysis (PCA) to reduce dimensionality. Three principal components account for 99 percent of the variation in the data. We fit the KNN and CART models using these three principal components instead of the dozen or so most important variables as identified by shrinkage. This principal component approach not only reduced computation runtimes but also made our models more accurate. In the end, however, LASSO generated the most accurate predictions.