John Saul

## Method Selection

While examining the data selected for the retention project, there were noticeable spikes in the data, specifically around the number of terms a student had been at TC before they stopped attending classes. The fourth term shows a significant drop off in retention and another one occurs at the tenth term. These kinds of mid-range changes in the data are not well tolerated with linear and logistic regression models. Decision Rule and Decision Tree models are more capable of adjusting the model to accommodate these changes, but are prone to over fitting. While I could balance this risk with a significant amount of cross-validation (such as Leave Out One Validation), I also can mitigate over fitting issues by using Random Forest algorithm, which involves a random selection of features in each Decision Tree training model and aggregating the resulting models into a final product. To that end, I am selecting the Random Forest algorithm for building my retention model.

The data in my data set is intentionally data from several years ago, so that once the model has been built, I can test and validate it’s model “goodness” with new information from future terms. This will help me check the results of the model to ensure that I haven’t over fit even with cross-validation.