Nick Holyk
Chris Peake
CSC367 - Settimi
Third Deliverable
10 November 2014

For our project, we wish to gain insight into what goes into being an effective offensive baseball player using recent player data. For a given year, we would like to determine whether a baseball player will have a strong offensive outing based on past offensive performance and personal demographics. In addition to this, we want to find what set of values that contain a large amount of players with strong offense.

Essentially, we wish to accomplish both a classification, as well as a clustering task. By using past data and demographics to predict a player's offensive output for next year, we are classifying their predicted output. Our second task looks for clusters of strong offensive players to determine the associations between their coinciding factors.

The measurement of what makes a good offensive player is determined by a number of statistics. Originally, we felt that batting average alone would be sufficient to determine offensive capability, but batting average alone does not correlate well to runs scored. Now, we will use the on-base percentage plus the slugging percentage determine a binary classification of whether the offense will fall above or below a chosen percentile. This classification will use past year's statistics such as batting average, OPS, at-bats, and slugging percentage, as well as demographics such as age, batting side, height, and weight. Our clustering task will use the same variables, without the binary classification, to determine clusters of demographics and statistics that net strong offense.

Currently we are working on finding the best set of variables to use to predict past performance. We have been using both methods of binning, equal depth and equal width, to simplify classification of variable values as well as using different combinations of variables to find strong predictors. We have been using classification decision trees to discover not only which variables reliably predict future performance, but also which offensive statistics, things like BA, OPS, and Slugging Percentage, can be accurately predicted with the data on hand. Should the decision tree classification fail to produce strong predictors, we will move to testing the data with the K-Nearest Neighbors classification method. Once we find a strong set of predictors we will move on to our clustering task. Using the strong predictors found through classification, we will attempt to find clusters of values from those variables, combined with demographic information, to determine sets of values that contain a large amount of strong offensive players.

Our method to validate the classification model uses a simple testing and training set. We will take three quarters of the data to build our training set, and use the remaining to test whether our offensive classification was correct. We are capable of testing the accuracy of our classification since we can look ahead one year and see what the the player's actual offensive numbers were.

One issue with our data set is that there are multiple possible measures for what determines strong offense. It has taken longer than expected to determine the set of variables that consistently predict our end goal. Currently we can predict which players will do poorly, with an accuracy of ~ 72%, but we have yet to manage to find a group of variables that will differentiate between average offensive players, and those that excel.