Sachin Lal and Rahul Matta
5/27/15

Status Report

We are attempting to predict if a player in the NBA will be in the Hall of Fame based on their statistics of the first four years of their career. Most teams that win the NBA championship have at least one player on the roster that will end up in the Hall of Fame. For a general manager of a team, it is important to be able to predict if a young player will be in the Hall of Fame. With the knowledge from our model, a general manager can decide if he would like to continue to build a team around a young player. Also the classifier will be helpful in determining if the general manager should give the young player, who is on their rookie contract, a contract extension.

We are using http://www.basketball-reference.com/ for all our data. The following 30 attributes are in our original data set and a binary HOF attribute:

Season	Age	Tm	Lg	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS

All the players in our data set have made the all-star team at least once in their career and they started playing in the NBA after the 1980 season (when the three point line was introduced). In total we have the first four seasons of statistics for 172 players. Our data set includes current NBA players and players who recently retired, who technically are not in the Hall of Fame, but we decide to alter the HOF attribute for obvious candidates such as Kobe Bryant, LeBron James, and Shaq O'neal. We partitioned the data into training, testing, and validation. There are 120 players in our training set, 26 in our testing set, and 26 players in our validation set.

We are using the python package Sci-kit Learn to train, test, and validate our models. The first model we tried was K-Nearest Neighbor, where k=1, which yielded a 77% accuracy. We also tried KNN where K=5, which yielded a lower accuracy of 75%. In both models, the weights of each attribute are uniform and we are using the minkowski distance measure.

Our next steps are to use cross-validation to find the optimal parameters for our current nearest neighbor model. This would include discovering the correct weights for our features. It is extremely important to use cross validation in our project since our training set is very small. Also we would like to train several other models including SVM, Decision Trees, and Linear Regression. For each model we create we will need to perform cross validation to find the optimal parameters. After choosing the best model we will create our website to display our findings from the project.