Final Project Report

Application Description:

Our application is designed to predict stats of rookies in the NBA by using their statistical data from their college careers in the NCAA based on per 40-minute metrics (points, rebounds, assists, steals, blocks), position in the college team, and college team performance (win percentage, points per game, rebounds per game, assists per game, steals per game, block per game).

The database contains a large amount of data distributed across four relations. This data was scraped and/or found from various online repositories such as draftexpress.com and basketballreference.com in the form of csv files. A python script was created to load the csv data into a sqlite database for use in the web application. The database schema and E/R diagram are shown below. In essence, the 4 relations contain ncaa player data, ncaa team data, NBA draft data, and NBA rookie data.

In order to make predictions, two models were trained using the data: a random forest and nearest neighbor. Random forest was used for its ability to make accurate, generalizable predictions across very different domains. A nearest neighbor model was used since we wanted to show NBA players who are similar to the NCAA player chosen.

The way the application works is a user can input an NCAA player’s per 40-minutes stats into boxes for points, rebounds, assists, steals and blocks. Then the web application will output rookie season predictions for each of the 5 stats. The web application will also output names and stats of the nearest neighbors in order to provide real-life comparisons. Alternatively, the user can choose a year from a drop-down box, and then select an NCAA player from that year using another dropdown. This dynamically queries the database to find the players who played that year in order to display them as options. Finally, the user has the ability to insert a player into the NCAA table. This works by using a simple update statement.

E/R Diagram:

The diagram below only uses attributes used in the model. We obtained a massive amount of data, not all of which could be feasibly incorporated in the time we had. For future work, we definitely intend on expanding the scope of the model to incorporate more of the available data. For example, stats such as three-point percentage, free throw percentage, and eFG% could yield interesting discoveries. For now, the data in the database that was not used is located in the attributes “Other Rookie Data,” “Other NCAA Data,” and “Other Team Data.” The full extent of the database schema can be found below in the list of database table.

../../../../Downloads/Untitled%20Diagram%20(2).png

Assumptions:

* College does not change conference
* Players do not change College (does occur, but is rare)
* Only looking at past 17 years
* Every drafted NBA player went to at least one year of college in the NCAA
* Every NBA player was drafted
* All NCAA conferences are equally challenging (should change in the future)
* No two NBA players have the same name (this is actually upheld as of the late 1980s, and our data does not extend that far back)
* Every player plays for the team they were drafted for during their rookie year

Database Tables:

**NCAA**- Contains player stats and information from the NCAA season before they were drafted for the NBA.

* ncaa(Player,Team,G,MP,PTS,FG,FGA,FGpercent,TwoP,TwoPA,TwoPpercent,ThreeP,ThreePA,ThreePpercent,FT,FTA,FTpercent,ORB,DRB,TRB,AST,STL,BLK,TOV,PF,Year,PTSperGame,FGAperGame,PTSperPlay,TSpercent,eFGpercent,FTAperFPA,ThrePAperFGA,ASTperGame,ASTperFGA,ASTperTOV,PPR,BLKperGame,STLperGame,PFperGame)

**teamdata**- Contains data for every college season to gauge team performance. This is combined with NCAA data to yield interesting results.

* teamdata(Rk,School,Season,G,W,L,WLpercent,SRS,SOS,ConfW,ConfL,HomeW,HomeL,AwayW,AwayL,TmPoints,OppPoints,MP,FG,FGA,FGpercent,ThreeP,ThreePA,ThreePpercent,FT,FTA,FTpercent,ORB,TRB,AST,STL,BLK,TOV,PF)

**draft**- Contains draft information for each player such as what pick they were drafted at, what team they were drafted to, and perhaps most importantly, position played.

* draft(Rk,Year,Lg,Rd,Pk,Tm,Player,Age,Pos,Born,College)

rookiedata- Contains player data and statistics for rookie season.

* rookiedata(Rk,Player,Age,Tm,Lg,Season,G,GS,MP,FG,FGA,TwoP,TwoPA,ThreeP,ThreePA,FT,FTA,ORB,DRB,TRB,AST,STL,BLK,TOV,PF,PTS,FGpercent,TwoPpercent,ThreePpercent,eFGpercent,FTpercent,TSpercent)