P3 – Data Analysis Plan

Overview

The problem that this project will tackle relates to the value of a basketball player's contract in the NBA. We want to look at how a player's performance, shown by their recorded stats, correlate to the contract that they have and potentially see if we can predict the value of a player's contract given the players stats. In the NBA, especially with a salary cap, it is important to utilize the money that you give to players wisely. In order to win the championship, teams must not give out large contracts to players that do not actually deserve it. You must have players that perform at or above their contract value, in order to be successful in the league. Our project could help NBA GM's give better contracts to players by giving them a sense of how much value a player has. To get the data for this project, we looked at basketball-reference and celticshub for player stats and contract sizes. Our project's dataset of 1442 rows consists of a list of all the players in the NBA from 2013-2015 along with all of their recorded statistics (including advanced stats). Every player's contract value is represented by a salary cap percentage, which represents how much of the team's total salary goes to the player. The features of this dataset are a player's stats and the target value is salary cap percentage. We will use this data to predict the value of any given player's contract.

Questions/Hypotheses

We have developed quite a few questions regarding our dataset. We are all very interested in the topic related to our data. Some of our questions we had were, is there correlation between player performance and contract value? Which stats are most important in determining contract value? Which players are overperforming or underperforming their contract value, in relation to the model prediction? We have all made a few predictions or hypotheses related to our data. They are, if a player has a high PPG stat, then they will have a higher contract value, and vice versa. If +/- (plus minus) is included in the model, then it will have a low correlation in relation to cap percentage. If a player is between 23-30 years old, then they will on average have a higher average contract value than players of other ages. If a player has a high PER stat, then their contract value will be proportionately higher, and vice versa (meaning that PER will be highly correlated

with contract value). We look forward to testing these hypotheses in the next step of our project.

Data Analysis Plan

Our project is looking to have a regression problem here, as after some debate we decided that it would be better to have our problem predict the actual salary instead of determining what class/group it would belong in. For the ML algorithms we are planning to use we went back and forth on this and decided that we would try to use regression versions of LinearSVC, Gaussian, KNeighbors, and Lasso in the manner we did in homework 5 and in other labs and would use the one that produced the best results in terms of accuracy and how closely its fit was. For feature hyper tuning we are planning on using GridSearch with min max scaling as we liked the functionality that was available with this. For feature extraction, we plan on using iterative feature selection to get the mor important features in our data. For the various variables we are interested in using, we will mainly be using scatter plots in the same manner we saw in online classes where we will use Scatter plots on the advanced states with these states as the X axis and the salary as the Y axis and try to identify various trends with these to see what variables might have the best correlations. We are also interested in using Violin plots for displaying the distribution and spread of our statistics where we will once again have all our features in the x axis and our target variable (the percentage of the cap) as the y to see how the distribution is centered and whether or not certain trends are occurring.