Project Proposal

Alex Ojemann

2022-11-04

Topic

For my project I intend to explore the factors that best predict NBA player pay and how these factors change over time. This topic interests me because I have been a fan of the NBA since I was seven years old and I believe this project can help me gain insight into what makes players valuable.

Data

Two data sets I intend to use in this project are a player stats data set and a player salary data set. Both contain entries for each NBA player in each year of their career. The stats data set spans from 1950-2017 and the salaries data set spans from 1990-2017, so the range of years I will explore will be from 1990-2017. The links to each data set on Kaggle are below:

Player Stats: https://www.kaggle.com/datasets/drgilermo/nba-players-stats?select=Seasons_Stats.csv
Player Salaries: https://www.kaggle.com/datasets/whitefero/nba-player-salary-19902017

Model

To best predict NBA player salaries I intend to use a decision tree. In the book Predictive Analytics by Eric Siegel, Siegel writes "To use a decision tree, to predict an individual, you start at the top (the root) and answer yes/no questions to arrive at the leaf." To build the decision tree, you determine the factor that is most telling and split the data down the middle. The example Siegel uses is a decision tree designed to predict whether a mortgage gets prepaid, and the most telling factor was "Is the interest rate < 7.94%?" If the answer to that was yes, the risk of prepayment was 3.8%, if not the risk ballooned to 19.2%. The tree builds on itself by again finding the most telling factor, this time within each of the yes and no groups from the first question. We can repeat this process as many times as we want to finish building our decision tree. However, Siegel warns that decision trees can overlearn if given too many factors. To avoid this, Siegel used a measure called lift, which is ratio of the target response and the average response, and compares it between the training data and the test data, which is about 20% of the overall data set aside from training the model so we can determine its effectiveness on data that the model isn't already exposed to. The example used by Siegel is a classification problem while mine is a regression problem, so my measure of effectiveness will have to be slightly different.

Timeline

The first step will be to inner join the two data sets together on the player and the year. It must be an inner join because we only want data where there exists both a set of stats and a salary. This will be done this weekend. The following two weeks I will build the structure of the model. The week after that I will work on analyzing the effectiveness of the model and decide on where the tree should be cut off to avoid overlearning. The final week and/or any remaining time will be used to further analyze how the model has changed over time by making multiple decision trees for different periods within 1990-2017 and comparing them.