Jonathan Greenberg

Milestone #3 Report

4/26/2020

I have made progress in what my plan is, because thats what a project is, constant iteration of what the goal is and how to get there.

Originally I was going to predict on a game-by-game basis whether a player will get injured. However, I found that retrieving the data for each game was taking very long and was too granular of a level to predict at…So I pivoted my goal to predict on the season level. Meaning, using a player’s features: their age, position, team, weight, height, wingspan, and a bunch of game statistics from the previous season (points, games played, minutes played, personal fouls, etc.), can I predict the type of injury a player is most likely to sustain in the following season? Additionally, can we cluster certain types of injuries or certain types of players?

I have done a considerable amount of data cleaning, transforming, and manipulation. I have also filtered out players that did not play at least 15 minutes/game I still need to do a few more things for this project:

1. Need to make a table with a key explaining acronyms or terms that are not used frequently such as DNP - did not play, DTD - day to day

2. Add in a few more body parts to account for any missing ones that my function didn’t catch, such as “neck” or anything else that I can find.

3. Vectorize/embed/one hot encode my three columns that contain words. This will enable my prediction, analysis, clustering.

Here are the three Google Colabs showing my work:

1. Cleaning: <https://colab.research.google.com/drive/1InZ1b1YDH97RYYncTBs5TKvWcNoX9PPr>

2. Add NBA Stats from previous season: <https://colab.research.google.com/drive/1RzC52c1mc4AbtqrTNDVwd7RAzpFdy286>

3. Predictions and Analysis: <https://colab.research.google.com/drive/1rJcQoJOE7w9tjGaxqRrcWU0Q9WvVS6YX>