Jonathan Greenberg
Milestone #4, #5 Report
5/3/2020
The following is my progress thus far and not on track wit the current milestone requirements.

This past week I had a great meeting with professor Avi Rosenfeld on the NLP aspect of my project, and he recommended I use a simple encoding (one hot) for my body, severity, and injury variables. We dove into the data and quickly found that the “severity” column was of little use, as there are only 3 classes, and two of them (DNP and DTD) were in fact the same thing, the NBA just called them differently starting in 2014.

I wanted to simplify the “Body” and “Injury” columns so I did a word count and took only the top ten classes in each and left those in the dataframe, and deleted the rest. Basically, focus on the most common injuries.

I have a few research goals:

1. Using all the features except Injury type (because this data is only known after the injury) predict the body part of a player that will get injured
2. Using all the features except Body part (because this data is only known after the injury) predict the Injury type of a player.

These first two I will try using logistic regression, random forests, and any other machine learning algorithms I decide to deploy.

3. Find out which features are most important in the predictions.
4. Visualize the players multidimensionally, project them in space.

I have started with my machine learning predictions and you can find the results here in the following colab: I have solid accuracy, precision and recall in my random forests, which is promising (:

https://colab.research.google.com/drive/1rJcQoJOE7w9tjGaxqRrcWU0Q9WvVS6YX