Capstone Project 2 Proposal

Sports have been games centered around numbers and statistics, but none more so than baseball. The development of sabermetrics in recent years has brought incredible insight into the game that players and enthusiasts have been outraged by the fact that the game might be more about statistics than skill. But what sabermetrics aim to find is efficiency. General managers and statisticians use sabermetrics to get the most out of their players and pitchers are one of the first set players that are looked into when wanting to maximize efficiency.

The reason being is because pitchers aren’t typically seen as “everyday” players. Starting pitchers typically pitch once every 5 days and usually only last 6 innings where as relief pitchers see as many innings in a week. This becomes an issue when teams are paying pitchers the same, or more, than everyday players who play all 9 innings in a game 5 days a week.

One way to improve pitcher efficiency is to prevent injuries for if a pitcher is injured, they’re getting paid not to play. There have been numerous studies that look into the correlation between pitch count and arm injuries and found a positive correlation so now pitchers are healthier but pitching less. This paper also statistically quantifies that pitchers are most effective 20 to 70 pitches into a game. Despite all these studies, it still doesn’t solve the problem on how to maximize a pitcher’s efficiency.

As a general manager of a baseball team, your objective to to get as many wins without spending too much. In other words you want to maximize the following formula:

$$\mathit{Dollars\ per\ win} = \frac{\sum (\mathit{Player\ Salaries})}{\mathit{Total\ Wins}} = \frac{\sum (\mathit{Player\ Salaries})}{\mathit{Runs\ Scored} - \mathit{Runs\ Allowed}},$$

$$\mathit{Runs\ Allowed} \sim \mathit{ERA} \sim \frac{\mathit{Runs}}{\mathit{Innings\ Pitched}} = \frac{3 \times \mathit{Runs}}{\#\ \mathit{of\ Outs}}.$$

A win is when your team is able to score more runs than your opponent and assuming your team is already signed (players salaries are already determined), the pitcher's role in the equation is *ERA* or Earned Run Average. ERA is correlated by the number of runs a pitcher allows divided by the number of innings pitched which itself is correlated by the number of outs the pitcher is able to get.

The goal of this project is to develop a model that could reliably predict outs (or the outcome of a pitch) given factors about what the pitcher is about to throw. The hopes for this model is to be used by pitchers and coaches/managers to take factors like pitch type, pitch location, pitch speed, pitching hand or pitcher, batter side or batter, and count to effectively predict the outcome of the pitch. This will hopefully lead to pitchers throwing unnecessary pitches and lowering pitch count. Less pitches for the pitchers means less strain and fatigue on the arm while still getting outs at a high level.

Major League Baseball Gameday offers play-by-play, or pitch-by-pitch, data from every game in the season dating back years. This data comes in JSON format which can easily be extracted in Python to gather the necessary information from each pitch. I plan to take data from the past three complete season (2015-2017) where it will then be cleaned and engineered to be passed through a deep neural network. Depending on the success of the model, the code, presentation, and model itself could be presented at the conclusion of this project.