Capstone Project 2 Proposal

Baseball is a chess match between the batter and pitcher. What pitch will the pitcher throw next? Where in the zone will he throw it? Is the batter able to make an educated guess on what is coming next, or is he totally clueless and just reacting to what is coming? Sports is beginning to adopt data science and analytics to answer questions like these, and is what I want to do for my 2nd capstone project.

- **What is the problem you want to solve?**

I want to make predictions of pitches coming from certain pitchers based on factors like count, score of the game, how many runners on base, if a righty or lefty is batting, and many others that could go into making predictions like this. From these predictions, I can analyze what factors were the most important in determining what pitches were thrown. I'm also able to choose specific pitchers, so comparing Cy Young award winners with bottom of the rotation guys could be interesting to look at. I would guess the Cy Young winners will be less predictable than pitchers who are not as good.

- **Who is the client and why do they care?**

The client for this project would be the managers, pitching, and hitting coaches who are interested in this data, as well as the players themselves. Any leg up they can get on the opposition is a good thing and if there is any way to reliably predict pitches coming, that would be an advantage for the hitters. Hitting coaches would be interested in this data as well. They

can coach their guys to jump on a fastball on a certain count or sit back on an off-speed pitch in certain situations. This data could be even be good for baserunners, once the hitters get on base. If they and the coaches had an idea of what pitches were coming, baserunners could steal on off-speed stuff as it will give them a better chance to swipe the base.

- **What data are you using? How will you acquire the data?**

The data for this project comes from a competition on Kaggle. It is broken up over 4 years of data from 2015-2018, and has most, if not all pitches thrown in the MLB over those 4 years.

- **Briefly outline how to solve the problem**

The most natural way I can think of at this time to model this problem is as a supervised classification problem. With the classes being the type of pitch the pitcher will throw. Looking at the information on Kaggle's site, there are 16 different types of pitches, with all the different off-speed pitches present, as well as both types of fastball. I could potentially approach this problem by using two different approaches. One using all 16 different classes, and the other dealing with new classes created by bundling all of those pitches into "fastball", "off-speed", "breaking ball", and "other"classes. Fastball group containing four seam and two seam fastball and cutter. Off-speed containing change up, knuckleball and splitter. Breaking ball containing curveball, knuckle-curve, screwball, sinker, and slider. Finally the "other" group containing the rest of the pitches, pitchout, intentional ball, and the unknown category.

Doing the analysis once with the groups would give an idea of overall trends of pitches, then doing it again with the specific pitches would allow me to get more detail out of the models. I plan on using the base logistic regression model, as well as k-neighbors classifier, random forest classifiers and even build a deep learning neural network if that is possible.

- **What are the deliverables?**

The way I want to deliver this is through a presentation. Presentations and public speaking are not my strong suit and doing so would give me more practice in that area and will allow me to be more comfortable speaking in front of people.