Evan Benton

Capstone 2 Project Proposal: “Hey Batter Batter!”

Background: John Doe is a rich multi-billionaire businessman who likes to invest his money into different things all the time. Recently John has been really into baseball and is looking to buy a team. He wants to start small before moving into a bigger corporation so he is starting with the Korean baseball league. The thing is, John is also competitive and likes to win. He knows the key to a good baseball team is the pitching staff. John would like to know the top 5 teams with who will be predicted to have the highest win/loss percentage based on pitching statistical analysis from previous years for the season of 2022 so he knows what’s the best team to buy.

Problem statement formation: (SMART) How can teams increase their win/loss percentage above 0.55 by performing better in the top 5 most important pitching stats.

- **Context**
    - John Doe is looking to buy a team in baseball and is looking for the best teams predicted to have the highest win/loss ratio based on their pitching statistics from previous years.
- **Criteria for success**
    - Having a predicted higher win/loss percentage of 0.55 or above which is generally considered positive.
- **Scope of solution space**
    - Increase training of pitchers to perform better in certain stats and accuracy
    - Having a variety of pitchers that specialize in certain areas while others don’t
    - Having a good defense behind the pitchers to support them
- **Constraints**
    - Requires money and potentially lots of it, even for a multi-billionaire, to buy an entire organization

- **Stakeholders**
    - Pitching staff
    - Players
    - Player Analysts
- **Data sources**
    - Team/Player game records to find all statistics that happen throughout the season
    - Accounting for money management and concerns

Plan: Since this is a predictive problem, we'll be leaning towards linear regression analysis. Pitching encompasses a multitude of stats that can affect a game's outcome. From walks, strikeouts, giving up runs, giving up homeruns or hits can all affect whether or not a team wins a game. This leads me to believe that multivariable linear regression is most likely best in predicting a team's win percentage using SelectKBest features. Looking into some baseball background might be the best way to also determine which stats are most important to include into the regression model based on sport analyst data. However, baseball is also a sport that has used statistics for mathematical analysis for quite some time now and have also formed their own formulas for predicting win percentage. The Pythagorean Win-Loss formula is the formula that has been created over the years and might also be an option for our model. This leaves us with 3 different options for predicting win percentage. Trying all 3 might also be a good option for comparison with each other.