CS287 – Data Science I

Eamon Deffner, Frederick R. “Randy” Carlson  
Subject: Data Science Check-in #1

Working Title: “Use of Data to analyze Base Running Statistics.”  
*Abstract: How much does base running speed affect offensive statistics? This paper mines data from BaseballSavant on sprint speed and batting averages to answer this question. We use the data science capabilities of anaconda python to answer these questions. We found that there is little to no advantage to single base hitting due to speed, but we did find that there is an advantage to speed in slugging. This means that training for longer sprints (from home to 2b) may yield better offensive results than training for the shorter sprint to first.*   
  
This document serves as a progress report on our project. Our team wishes to provide a condensed overview of this project, its progress to date, and a timeline.

2) Our goal for this week was to analyze which players have an added contribution to on base percentage, slugging percentage, and on base plus slugging (OPS) based on speed. Although we found small correlation coefficients between these things in our model, we found that there is notable contribution to these statistics from the types of “hustle plays” that we are analyzing. In addition, we created some visualizations for our data, including one showing that players that are involved in these hustle plays more often have a higher contribution to their OPS than other players. This contribution is enough to separate an average offensive player from an above average one.

3) Challenges and Questions: One factor that we may be missing in our model is the role that stolen bases play. Faster players can steal more bases, giving them the extra base more often on plays where it is not calculated as a part of their slugging percentage. This would be an added stretch goal, as our original focus was to study the importance of speed for getting on base.  
   
4) Timeline   
Week 1: Extract data from baseball savant, create a data frame, and do some exploratory data analysis (EDA).  
Week 2: Identify types of data (numeric, categorical, ordinal), determine independent and dependent variables, and create linear regression models; confirm the linearity of relationships.   
Week 3: Find players with the highest offensive profile addition from speed. Visualize this data using tables and graphs to show if speed impacts OPS.  
Week 4: create a model to determine how much of a role *speed*, *batting average above expected*, *slugging above average*, and *percent of at-bats that result in ground balls and hits* determine a player’s OPS. Compare how results in “hustle plays” affect OPS for average, fast, and slow players. We will do this analysis by building a fake player with a given speed and an offensive profile particular to his speed.  
Week 5: Use clustering to run the same regression as in week 3, breaking players into groups, and then see if speed matters differently for each group of players and how this affects their overall offensive contributions.  
Week 6: Read in a file with salary values for players from 2017-2022. Cluster players based on *on base percentage* (OBP),*slugging percentage* (SLG), and speed. We will use a linear model to determine how speed plays a role in how players are paid.

5) Open Challenges and Questions: We will still need to incorporate techniques like Bootstrapping and Non-Parametric analysis into our overall paper. There are also opportunities to look at this effect of speed to new rules changes such as the shift rule.

6) Revised week-by-week timeline: None