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 <data source a> and <data source b> 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.  
   
1) Speed as a Factor of *on-base percentage plus slugging* (OPS) in the Game of Baseball. Using linear modeling and clustering of players, we hope to create a model for determining how important speed is for a player getting on base on ground balls and how speed impacts a player’s slugging percentage on hits that are balls in play. After creating a model for each statistic, we hope to cluster players based on these abilities and see if there are groups for which speed matters more or less.

2) Our goal after this week was to have a completed and cleaned data frame. This week, our other goal was to build the linear model for slugging and on-base percentage, with data from 2017-2021 being used as training data and 2022 used as our test set. We have accomplished this goal and will be moving into seeing if certain players have significantly higher additions to their statistics based on speed.

3) We have found that while speed has some impact on slugging for most players, on average, it has little to no effect on batting average on ground balls for the league as a whole. However, it appears that speed does improve their batting average for some people that are fast. Next week, we aim to test this by seeing if the players with the highest addition to their batting profile from these “hustle plays” are the fastest or if something else is at play.  
   
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 speed affects 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

6) Revised week-by-week timeline: None