DS2500 Project Proposal

Wesley Chapman

For my project, I would like to focus on MLB player statistics, specifically OPS (On-base percentage plus slugging percentage). This will be the value I am trying to predict. I will look at data from hundreds of MLB players, and using 10-15 other statistics I will attempt to predict their OPS. The more statistics I use as independent variables, the better my model will perform, and I believe this range of 10-15 statistics will be sufficient here. Since OPS only applies to position players and not pitchers, the other statistics I use will be offensive or batting metrics. These independent variables will include statistics such as batting average, home runs, walks, strikeouts, stolen bases, runs batted in, on-base percentage, slugging percentage, total bases, singles, doubles, triples, etc.

To gather data, I will use the website Baseball Savant (baseballsavant.mlb.com). This site has data for any baseball statistic I would need in the form of tables. It allows me to customize which columns I want to use and which I don't. The data can be downloaded straight from the site as a CSV file, eliminating the need to clean and scrape data from elsewhere, and making the Python work more straightforward.

Regarding ethical considerations, there are few, with baseball statistics being publicly available. But as with any use of data, I want to ensure that my results do not promote any form of bias or violate the privacy of MLB players. Since I am not using any personal data such as age, hometown, high school, etc, this shouldn't be an issue.

The goal of my analysis is to create a model that is successful at predicting an MLB player's OPS based on their other offensive statistics. In order to predict OPS, I would compile these 10-15 other statistics, and use multiple linear regression to predict a value for OPS based on the given player's statistical values. With multiple linear regression, I am hoping to find a Pearson Coefficient value of 0.9 or higher. Baseball as a sport is very predictable using statistics, which is why I am striving for a high Pearson Coefficient value. If

I can obtain a Pearson Coefficient value above 0.9, I would deem my model successful in predicting OPS.

This project will be done individually, so no division of work will have to be done. I believe with my knowledge of the nature of baseball statistics, the project will be manageable solo.