MLB Batting Stance, Performance Analysis

Project Abstract:

In this project, I am looking to analyze Major League Baseball statistics from Baseball Savant and Google Cloud's Statcast, and specifically the newly released dataset on batting stances, or where a batter stands during their at-bat. Despite Baseball's reputation as potentially the most analytically-studied sport, there is no consensus on an optimal batting stance, and it is currently considered to come down to player preference. Through this project, I hope to find connections between batting stance factors, such as distance from the pitcher, angle, stance width, etc., and hitting outcomes, including hits, runs, times hit-by-pitch, and more. There is a great depth of conclusions that could be found within this dataset, from simple questions such as "Does standing closer to the plate increase the odds of getting hit by a pitch?" to harder questions like "Do power hitters have marginally different stances compared to contact hitters?" or "Can contact percentage be predicted from batting stance alone?".

For the exploratory data analysis of this project, I'll look to spend a solid amount of time visualizing the data to find strong correlations, both to eliminate overlapping predictors and to quickly find basic, interesting conclusions. As the project evolves, I'll look towards the logistic regression performed in class to find relevant, strong predictors from batting stances for a number of different targets from batting outcomes. Where the previous steps are successful, I'll expand by building decision trees, random forests, or XGBoost models as appropriate, to see if batting stance, as a whole or just a factor of it, can accurately predict batting performance in any specific way. Within this project, there is also potential for using clustering algorithms to find groupings of batting stance ideas, which may then be able to be labeled based on batting characteristics.

The batting stance dataset on MLB hitters contains a ton of useful predictors and targets that have the potential to lead to interesting conclusions, both high and low stakes. I am confident that through this project, at the very least, I'll gain useful insights into a key component of baseball, that which has up to this point remained mostly nebulous throughout the history of the sport.

Milestone Report:

My project began with selecting relevant data. Baseball Savant and Google Cloud's Statcast contains an almost overwhelming amount of statistical baseball data. I first identified my independent variables, or predictors, for this project to be mainly the "Batting Stance Leaderboard" data. This includes top down X and Y positions of the center of mass of a batter in the batter's box, their average foot separation, average stance angle, and the average point they make bat contact, relative to the plate and themselves. In the interest of adding features, I included swing length and swing speed, as both measures are batting characteristics independent of pitch scenario, as a generalization. The side the batter bats from was also recorded, taking special note of switch hitters as well.

Given the problem statement of this project, the factors to select for prediction were wide open. With the most broad view of the problem, to see if batting stance can predict hitting performance, a general metric of offensive performance is best. For this, On Base Plus Slugging, or OPS, was chosen, as it is considered a strong comprehensive assessment of offensive production. OPS combines the percentage a batter gets on base per plate appearance with the total number of bases a player records per at-bat. For additional metrics, I included On Base Percentage, Slugging, Batting Average, and Home Runs (as a percentage and a tally), as they are also used when discussing and comparing hitters. I'm doubtful that stance plays a significant role in comprehensive batting statistics, so more granular statistics were also chosen. Hit by pitches (as a percentage and a tally) were taken first as a relevant side-question to the analysis. Also chosen were Exit Velocity, Launch Angle, Sweet Spot Percentage, Barrel Batted Rate, and Solid Contact Percent. These statistics take defense out of the equation, and just look at how correctly or efficiently a ball is hit, which can more reasonably be suggested to be affected by stance.

Altogether, the data comes out to 28 columns and 993 rows of data. Due to the nature of the stance records being new data collected from new technology, batting stances only go back to 2023. The data is additionally filtered to only consider batters with at least 50 plate appearances. In other baseball statistical analyses, a batter is only considered "qualified" for analysis and reference when they are above 500 plate appearances in a season. With that additional filtering, the data is cut to only 266 rows. The dataset as a whole, through EDA, is plotted to visualize their distributions, seen in Figure 1.

At this point in the project, I've completed the EDA phase, and have started to answer questions brought up in the project's abstract. I first addressed the question of whether batting stance alone can predict the rate at which a batter gets hit by a pitch, and unfortunately, the answer seems to be a strong no. Scatter plots, which can be seen in Figure 2, and correlation matrices show no indication of potential correlation. Experiments conducted with Linear Regression, Logistic Regression, and Random Forest Regression and Classification resulted in extremely poor performance, with R^2 values and error values outside the realm of consideration for further work. This specific

idea has proven to be a failure, and at this point my conclusion is that batting stance cannot be used alone to predict how often a batter is hit by pitch.

There are far more questions to be answered, however. Beyond those addressed in the abstract, I'd like to also look at whether stance affects performance against certain pitch types, as it is oftentimes ascertained that you should, say, move up in the box against pitchers with drop or rise to hit before the ball breaks. This question requires the coalescence of more data, which is unlikely to be an issue.

In addition, we can ask whether certain stances or stance factors are linked with decreased performance. This can be relevant information for avoiding stances that are, say, too open or too far back in the box.

Central to this project is to still discover if Batting Stance can be used as a relevant predictor towards batting performance, in the abstract or direct, so the project's scope will center around this idea. The goal remains to examine the potential through different levels of targets. There is risk to this assessment, and it is certainly possible that stance alone cannot function as an indicator towards batting performance. In this case, the project will still have an important, relevant conclusion for baseball understanding, but other avenues will be examined as well.

Visualizations:

Figure 1 - Variable Distributions in the Full Dataset as of this Milestone

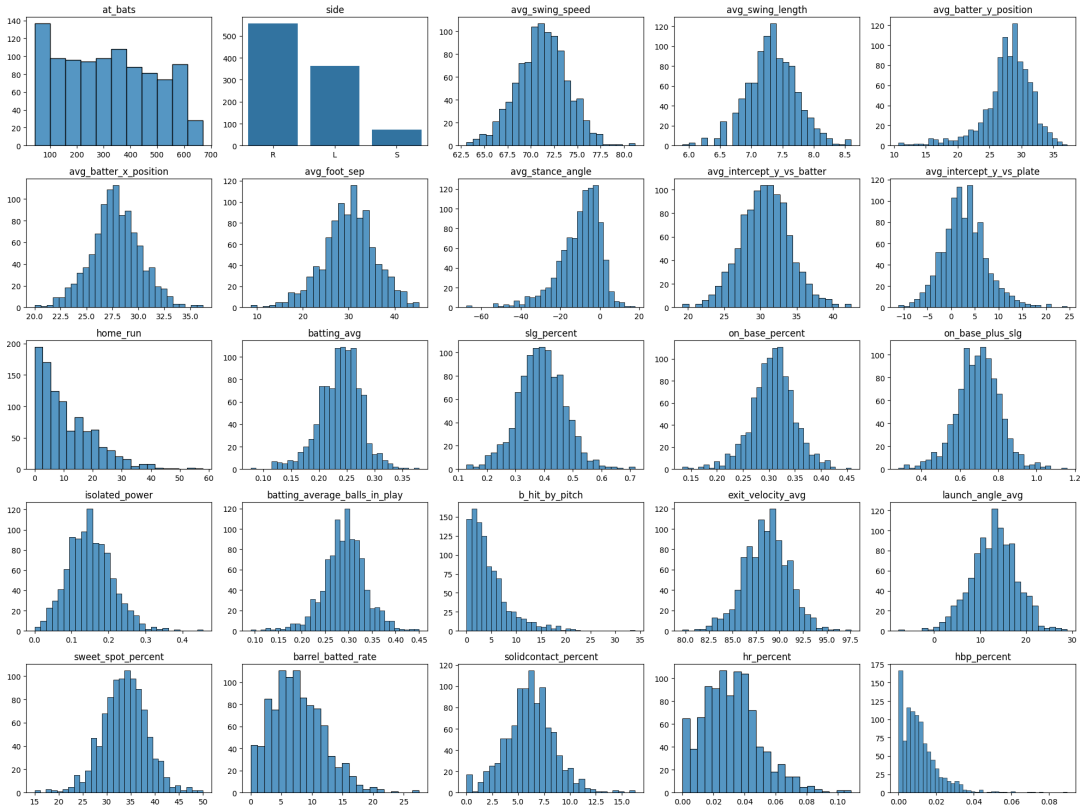

Figure 2 - Hit by Pitch Percent and Tallies plotted against Batting Stance Factors