DS2500 Project Report

Wesley Chapman

Problem Statement and Background

This project analyzes the relationship between twelve MLB batting metrics, and a batter's OPS (on-base plus slugging), which has become a common metric for measuring a batter's overall talent from year to year. The twelve batting metrics used are singles, doubles, triples, home runs, strikeouts, walks, batting average, runs batted in, total bases, stolen bases, barrel percentage, and in-zone percentage.

My lifelong love for baseball led me to choose it as a topic for my final project. Growing up, I loved to collect baseball cards and compare player's stats on the backside of the cards. Although I didn't understand what the majority of the numbers I was looking at meant, they intrigued me. I developed my own standard for each statistic that made a player good in my mind. As I grew older and followed other sports, I realized that none come close to baseball in its focus and emphasis on statistics. Everything about the sport has some sort of metric, whether it be the spin rate of a pitch, the heat map of an outfielder's positioning, or the barrel rate of a batter. When deciding on a project topic this semester, I couldn't help but consider baseball. It felt like a perfect way for me to merge my passions of data science and baseball.

The ability to predict a player's OPS offers many practical uses. It allows teams, scouts, fans, or whoever to predict how well a batter will perform in the future. MLB scouts may use this to find players who have the potential to perform very well based on their secondary stats. For example, if a player's secondary stats predict a high OPS, but their actual

OPS does not match this, it could be a sign that their ceiling in the future could be much higher. MLB front offices could use this to evaluate whether or not to sign a player. Team managers may use this as a way of crafting a line-up for a game. Fans who participate in fantasy baseball would be able to use this as a means of drafting a winning team. Players could even use this themselves to hone in on a certain aspect of their game that they need to improve; say a player has a very low barrel percentage which is leading to a low predicted OPS.

By using predicted OPS and not actual OPS - which may seem like the more logical option - we can see how an accumulation of several other statistics influences the player's OPS. Predicted OPS becomes a more inclusive, valuable metric to study when analyzing a player's offensive performance. Regular OPS can be effective in this same way, although there are instances when it may be misleading, for example with a small sample size of at-bats, or with a player who strives in certain offensive areas but not others.

While using predicted OPS is effective at measuring offensive performance, it doesn't encapsulate an individual's overall ability as a baseball player. But if we want to find a metric that would analyze that, there would be some implications. We would need to merge all aspects of the ballplayer, both offensive and defensive, into one statistic. Doing this would create less of a clear measure of the individual strengths of the player. Hypothetically, if we do this and get a seemingly high score for a player, we do now know that they are performing well, but we are unsure what about them has led us to this conclusion. Using predictive OPS would be a more effective way of measuring the player's specific skillset.

Introduction to the Data

The data used in this project is sourced from baseballsavant.mlb.com, a website dedicated to exclusively baseball statistics. The site allows you to customize data tables to include any MLB players, past or present, as well as hundreds of offensive and defensive statistics corresponding to the players.

My table used the following criteria:

- Years: 2019, 2021-2024 (skipping 2020 because of shortened season)
- Minimum plate appearances: 502 (qualified batters only),
- Columns: Singles, doubles, triples, home runs, strikeouts, walks, batting average, runs batted in, total bases, stolen bases, barrel percentage, and in-zone percentage.

Baseball Savant allows for the download of the tables in CSV format. This level of accessibility to the public leaves little room for privacy or ethical concerns with the data. Since the data is entirely MLB-related, and does not include personal player information, there are no implications regarding ethics or privacy.

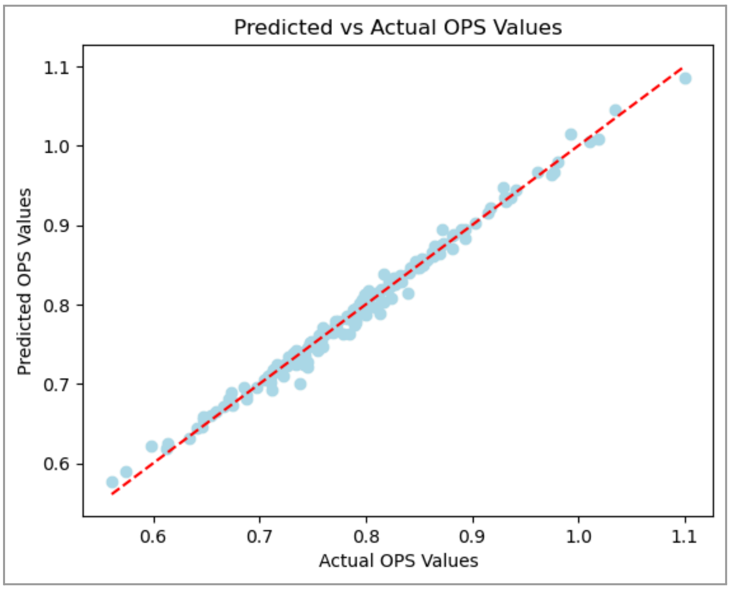
Data Science Approaches

To analyze the relationship between my independent variables and the dependent variable of OPS, I decided to use multiple linear regression. With this method, I am able to compare the predicted OPS values with the actual value for each player. I can also examine the predicted OPS alone to determine how well a player should theoretically be playing.

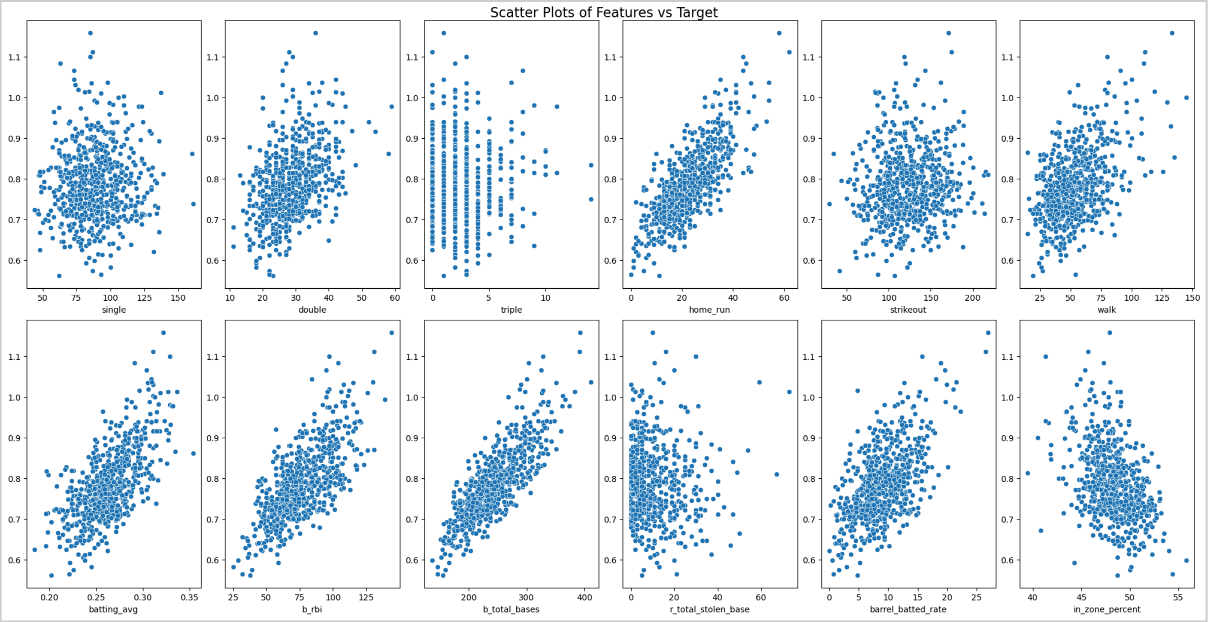
To evaluate the relationship between my predicted and actual values for OPS, I used two data science approaches, mean squared error and r-squared (coefficient of determination). Mean squared error finds the difference between each pair of predicted and actual values, squares it, and determines what the mean of these values is. By doing this, we are able to effectively evaluate the success of the multiple linear regression model. A smaller mean squared error would mean more of a relationship between my independent offensive variables and OPS. An r-squared model is another means of evaluating the success of my multiple linear regression model. R-squared values range from 0-1 and indicate how much of the variance in the dependent variable is due to the given independent variables. In the case of r-squared, a higher value would be better, showing that the independent variables have a significant influence on the dependent variable, which in our case is OPS.

Results and Conclusions

My data science approaches proved the multiple linear regression model to be very successful. The mean squared error evaluation returned an extremely low value of 0.00008457, meaning the predicted OPS values were very successful. Through the r-squared evaluation, we got a result of 0.99000652, which tells us that 99% of the variance in OPS can be explained by the twelve independent variables. Both of these values prove to us that the multiple linear regression model is highly effective at predicting OPS values for MLB players.



This scatter plot of predicted versus actual OPS values is an excellent visualization of the accuracy of the multiple linear regression model. The red dotted line represents an exact prediction, so you can see how close and even spot-on many of the data points were.



Another interesting visualization is this pair plot, showing the individual relationship between each independent variable and OPS. Some variables such as total bases and home runs have strong correlations to OPS, while others, triples, for example, have little to no correlation. This illustrates how multiple linear regression incorporates several individual linear relationships that may have little in common, into one more comprehensive relationship.

In conclusion, it is clear through the usage of several data science approaches, that our twelve offensive independent variables have a significantly strong relationship with the dependent variable - a player’s OPS value.

Future Work

With the success of this multiple linear regression model, we could use it, as stated before, as a means of evaluating a player's potential OPS ceiling, and whether or not we think they will be a successful offensive player. As OPS is an effective indicator of a batter's overall performance, this model would be very qualified to do evaluations and could be used by player scouts, MLB teams, fans, etc.

A next step could be to try to create a similar model for pitchers. Being able to predict performance for batters is great, but there is more to baseball than just hitting. Pitchers are equally as important for a team's success. A statistic that could be used as the dependent variable in this hypothetical model is ERA (earned runs average). ERA tells you how many runs a pitcher tends to give up, and it would be useful in multiple linear regression in the same way OPS is for batters.