# Predicting MLB Batting Average for the 2025 Season

By Madison Nafarrete
San Diego State University
CS 577
Professor Kang

**Abstract**

Analytics have become a big part of sports in order to optimize a player or team's performance. Predicting a player's batting average for next season is a complex and fun challenge in sports analytics. This project will use statistical methods and machine learning models to predict a player's performance based on data collected from previous seasons. Some of the variables that will be analyzed will be batting averages, plate appearances, walks, and even advanced metrics like OPS+SLG and wOBA. These variables will be included in order to make an accurate prediction for next season. Techniques such as linear regression and cross-validation will be used to help discover key variables that contribute to a player's batting average. Results from this project can offer insights to MLB teams, baseball fans, and even other sports teams to use analytics as a tool to improve the performance of athletes. This project can provide insight into the various factors affecting a player's batting average throughout the season, using both new and traditional baseball metrics to make accurate predictions. As someone who grew up playing softball and watching baseball, I decided to undertake a project that would attempt to predict the 2025 season's batting average. Predicting the MLB batting average for the next season, both for the league and individual players, is a challenging task, but I thought it would be interesting for sports analytics, as it can lead to players refining their routine, and baseball teams finding ways to maximize their performance, which could also be fun for fans in terms of how they place bets or draft players for fantasy leagues.

**Introduction**

The dataset used is a custom CSV file from baseballsavant.mlb.com, which allows users to select and display specific batting statistics. The dataset consists of variables that are basic hitting stats such as hits, batting average, strikeouts, walks, slugging percentage and on base

percentage. Additionally, there are advanced metrics, including wOBA (weighted on-base average), xwOBA (expected weighted on-base average), on-base plus slugging (opsslug), whiff percentage, and swing percentage. The main variables for predicting batting average will be hits, wOBA, on-base plus slugging, and slugging percentage. The primary objective of this project is to predict the 2025 MLB season batting average for the league and select players using various regression methods, and to determine which model is best suited for this task.

**Approach**

The analysis began by ensuring the data was clean, with no missing or negative values that could compromise the prediction's accuracy. Additionally, since this was a prediction for the 2025 season, we ensured that the final predictions were for players who played in the 2024 season and are slated to play next season. Pandas uploads the CSV file and turns it into the data frame, and manipulates the data to make the data frame easier to analyze. To identify suitable variables for the regression model predictions, Matplotlib was used to create visualizations during the exploratory data analysis. In the exploratory data analysis, the seaborn library was also utilized for data visualizations, producing a scatterplot and correlation heatmap to reveal the relationships between variables. The sklearn library proved crucial for model predictions, particularly in making the actual prediction for the 2025 season, which involved running regressions to find the best model for predicting the batting average for next season.

**Data Analysis**

In the initial analysis of the data set, I made a histogram of the distribution of batting average. The shape of the distribution is bell-shaped and symmetric, which is reminiscent of a normal distribution. Building regression models will become easier and more reliable, with more interpretable data and accurate predictions. Next, the correlation heatmap shows all the variables

and their correlation with each other. The darker the red, the more the two variables being compared have a positive relationship with each other; the darker the blue, the more negative correlation a variable has with the other. When looking at the relationships, batting average and its relationships with other variables were targeted in order to make the prediction. The ones that showed the best relationship with batting average were on base percentage, wOBA, on base plus slugging, slugging percentage, and hits. Following the correlation heatmap, five scatterplots visualize the relationships between batting average and on-base percentage, wOBA, on-base plus slugging, slugging percentage, and hits, all of which exhibit a positive and consistently linear relationship. A cross-validation was performed for a logistic regression to compare it to running a linear regression for the predictive model. Before running the linear regression, cross-validation on the linear regression model was used to make sure the linear regression was as accurate as possible. Cross-validation helped ensure that the linear regression would not overfit the data and also perform well with unseen data, such as future statistics like predictive stats. After the cross-validation, a linear regression was implemented to train the data set in order to make predictions. Finally, the linear regression model was chosen to make the actual prediction for batting average, because it had lower root mean squared error (rmse) than a logistic regression which means a more accurate prediction of batting average.

**Results and Discussion**

The cross-validation root mean-squared error in each fold for the linear regression is relatively small, which means the prediction could be really close to the true value of next season's batting average. After the cross-validation, I trained a linear regression model in order to make the batting average prediction. The mean squared error and the R-squared for the model indicated that the linear model would make a good prediction for the 2025 MLB batting average.

After running the linear regression model, prediction for the league average for next season turned out to be .266. There were also predictions for individual players such as Paul Goldschidt of the Cardinals, he hit .245 this season, and is predicted to hit .251 this upcoming season. This type of prediction could be beneficial towards sports betting if fans wanted to try and predict batting average for their favorite players or favorite teams. This could also be beneficial for sports organizations as well because it allows for an organization to make decisions financially as well based on these predictions. Players can also benefit because it allows players to identify trends within their routine that helped with their batting average and other metrics they can look at to improve batting average. The downside of predicting next season's batting average, there is really no way to tell how accurate the prediction can be because there are a lot of outside factors besides batting stats that affect batting average. For example, players getting hurt throughout the season can affect a player's batting average because it shrinks their sample size, which can make the prediction less accurate. Additionally, players who don't get signed or retire before the next season start are included in this data set, so that can affect the prediction as well. Overall, predicting batting average for next season was interesting because it has a possibility to open up predictive analytics that can benefit sports teams, players and fans.

## Sources

CSV file provided with files for submission

https://baseballsavant.mlb.com/leaderboard/custom?year=2024%2C2023%2C2022%2C2021%2C2020%2C2019%2C2018%2C2017%2C2016%2C2015&type=batter&filter=&min=q&selections=pa%2Ck_percent%2Cbb_percent%2Cwoba%2Cxwoba%2Csweet_spot_percent%2Cbarrel_batted_rate%2Chard_hit_percent%2Cavg_best_speed%2Cavg_hyper_speed%2Cwhiff_percent%2Cswing_percent&chart=false&x=pa&y=pa&r=no&chartType=beeswarm&sort=xwoba&sortDir=desc