Frank Barbera and Alex Ipeker

Roth

DS 325

5/2/2025

Old Hitting Metrics vs. Advanced Hitting Metrics in MLB

Introduction

In the MLB, the art of hitting is analyzed as a science, people seek answers for why teams offensive production is greater than others, they use different metrics to understand teams offensive production giving reason to why some teams are better than others. With all the new metrics that have been made over time, there are still some traditional metrics that have always been around that hold more weight to offensive production. Despite the hype around advanced hitting metrics such as exit velocity, launch angle, and barrel rate, traditional statistics such as batting average and OBP are stronger predictors of offensive productivity in the MLB. By using hitting statistics from the 2023 MLB season, we applied regression models to compare the predictive power of traditional and advanced hitting metrics to see runs scored. By viewing the R-squared, MSE, and featured plots they showed that traditional metrics outperformed advanced ones in both linear and random forest models.

Methods

We used a dataset of the 2023 MLB batting statistics of 363 different players. The features we used for our dataset were different for traditional metrics and the advanced metrics, for the traditional we used batting average, on base percentage, and slugging percentage. For the advanced metrics we used exit velocity average, launch angle average, barrel rate, and hard-hit percentage. The goal of this was to predict r_runs, which is runs scored by player, between traditional hitting metrics and advanced hitting metrics. The cleaning and preprocessing that took place removing double quotations from the dataset and replacing them with single quotations, removing any whitespace in the column names so there was no unwanted spaces, used feature engineering to to create new column identifiers for identify player names, and training and testing the date with a 80/20 split. Assumptions that were made before modeling was the player hitting metrics correlated enough to runs scored to make an accurate prediction and the data was trustworthy being it was from the MLB. The models that were used were linear regression (LinearRegression) and random forest regression (RandomForestRegressor) on the data. The purpose behind using linear regression (LinearRegression) was to show the direction and magnitude of influence the features had, it helped create a benchmark for the features to see how they perform under the simplest assumptions. It also revealed whether metrics like batting average and exit velocity average had a direct relationship with players scoring runs. Linear regression (LinearRegression) also tested our utility of the advanced metrics to the traditional metrics. The purpose for using random tree regression (RandomForestRegressor) was to help handle multicollinearity, irregular data distributions, and any synergies between features. It also helped evaluate which features contributed most significantly to predicting runs scored.

Results

When looking at our results, we used R-squared and Mean Squared Error to see how well each model predicted runs scored by players. For the traditional metrics, the linear regression (LinearRegression) expressed the R-squared to be 0.459 and the MSE as 308.924. For the advanced metrics for linear regression (LinearRegression) the R-squared was 0.0629 and the MSE was 531.0757. Looking at the results for the random forest regression (RandomForestRegressor) for the traditional metric showed the R-squared was 0.316 and the MSE was 387.74. The random forest regression (RandomForestRegressor) for the advanced metrics showed the R-squared as -0.0402 and the MSE as 589.52. Now looking at the R-squared for the train and test sets, for the training set it represented a R-squared of 0.4345 and for the test set the R-squared was 0.4549.

Model Type	Features Used	R2	MSE
Linear Regression	Traditional	0.4549	308.92
Linear Regression	Advanced	0.0629	531.08
Random Forest	Traditional	0.3158	387.74
Random Forest	Advanced	-0.0402	589.52

Discussion

The results express how the traditional metrics are better for predicting runs scored by a player than the advanced metrics. As it was expressed the R-squared for the linear regression (LinearRegression), the traditional metric had a 0.459 meaning that 45.9% of the variance in runs scored can be explained by the traditional metrics. This is a moderately strong relationship that the stats are good predictors for runs scored. For the linear regression (LinearRegression) of the advanced metric it was 0.0629 meaning that 6.29% of the variance in runs scored is explained by the advanced metrics, this shows a weak linear relationship meaning that the stats are not strong predictors (Figure 2). This gives evidence to how the traditional metrics are better predictors for runs scored by a player. The MSE for the traditional metric and advanced metric for the linear regression (LinearRegression) expressed traditional having a MSE of 308.924 and the advanced having a MSE of 531.0757. This means that the error for traditional metrics have a lower error percentage than the advanced metric, expressing what the R-squared showed that the traditional metrics show better predictions for runs scored. Now looking at the random forest regression (RandomForestRegressor) for both the advanced and traditional metrics show similar explanations as the previous model. The R-squared for the traditional metric was 0.316 and the MSE was 387.74. This expresses that the R-squared had a 31.6% variance in runs scored and average squared error of predictions is moderate meaning it still makes reasonably good predictions. Now looking at the advanced metrics, the R-squared showed a negative value meaning that the model could not find any meaningful patterns between runs scored and the advanced metrics. For the MSE had the highest prediction error out of all the models, this expresses that advanced metrics were inaccurate and unreliable when determining their impact on runs scored. Some hurdles we experienced while making the code, we wanted to use WAR

which is wins above replacement instead of runs scored because they expressed a better metric for wins. Another hurdle was cleaning the data, it showed difficulty at first but then when using copilot it helped figure out the problems. A successful outcome we saw was our thesis being correct for the data expressed that the old hitting metrics better determined runs score than the advanced hitting metrics. This project does inspire future work for it expresses limitations within the metrics. The project can move to examine the limitations the advanced metrics alone have upon offensive production, expressing that the advanced metrics are more glamor stats to actual offensive production stats. The project can also move to looking at the offensive production of specific players by not viewing runs scored, for runs scored is a team metric rather than individual metric. In conclusion, our project answered our thesis accurately by using reputable data and with the use of accurate models.

FIGURE 1

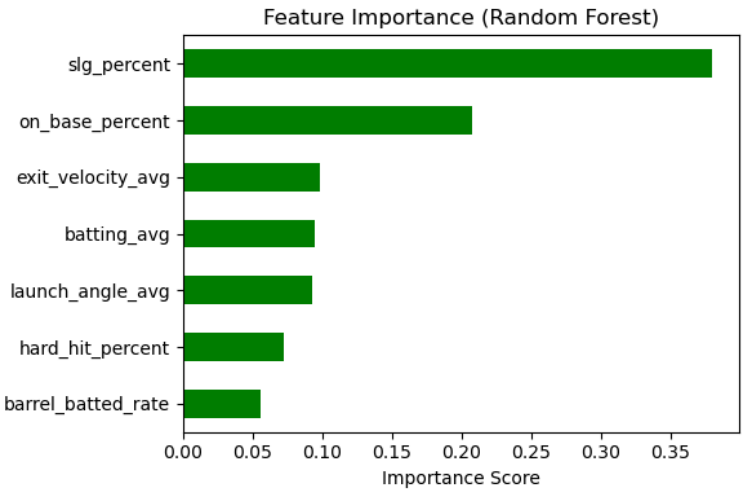

Figure 1, Shows that traditional metrics (SLG and OBP) are ranked as a more important feature when compared to advanced metrics.

FIGURE 2

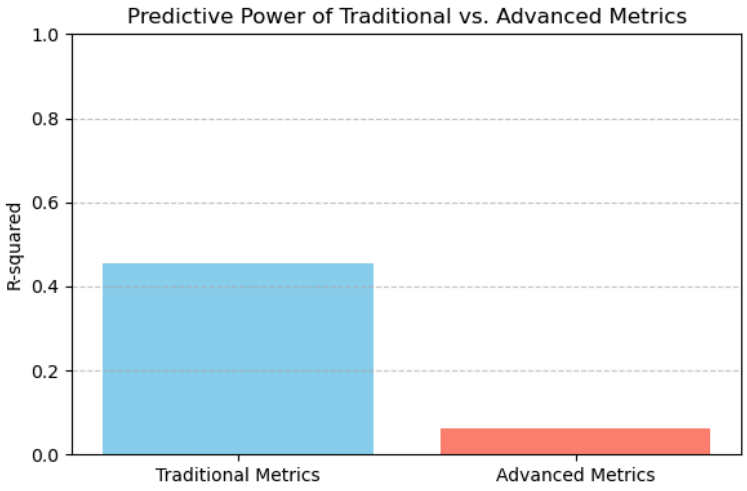

Figure 2, Shows that with R2, traditional metrics have a much higher predictive power than advanced metrics

FIGURE 3

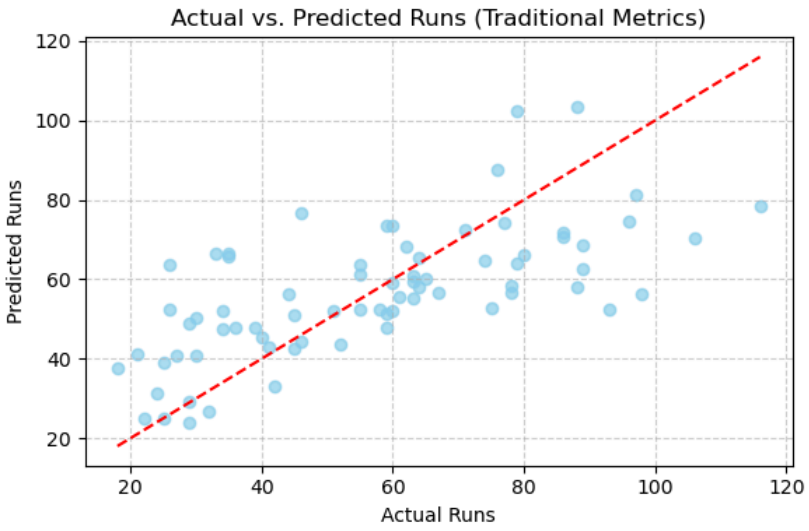

FIGURE 4

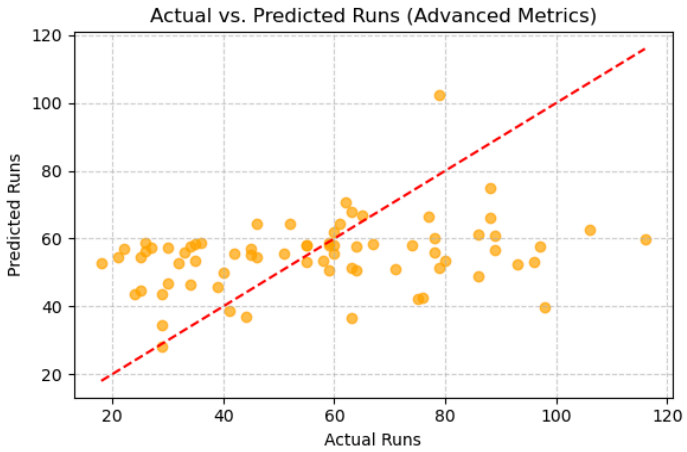

FIGURE 5

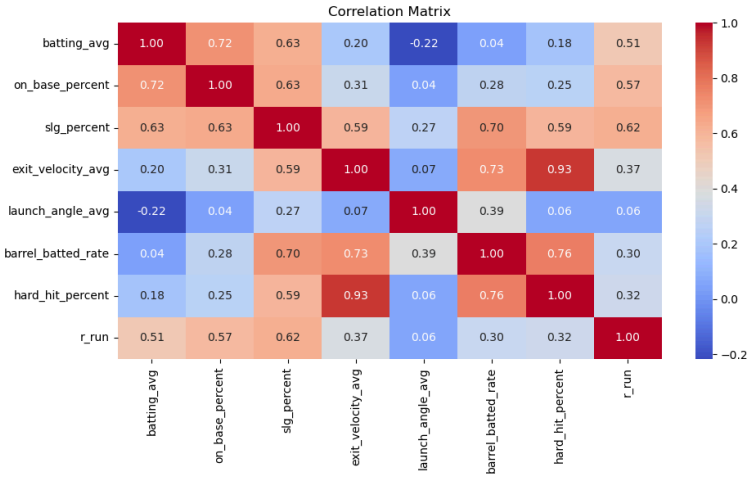

Figure 5, Correlation matrix shows that the traditional metrics have hotter spaces when correlated with r_run

References - Data Set

Baseball Savant. *Statcast Player Batting Data – 2023 Season*. MLB Advanced Media. Retrieved from https://baseballsavant.mlb.com/statcast_search

- Dataset came from Baseball Savant, which is maintained by MLB.com. It's an official source of Statcast data and includes both traditional and advanced metrics for every MLB player in the league.

"How do you make a confusion matrix?" prompt. *ChatGPT*, 25 Sep. version, OpenAI, 28 April. 2025, chat.openai.com/chat

“Fix Errors” prompt. *Copilot*, OpenAI, 28 April. 2025