Exploring Metrics for Predicting Home Runs in Baseball Using Machine Learning

Mickey Shamah

Introduction

In recent years, many baseball franchises have relied heavily on data analysis to predict and improve performance. The ability to quantify player performance has led to the documentation of various metrics that can be used to evaluate a player's batting abilities. Using a dataset of Major League Baseball (MLB) home run metrics for all players from the 2015- 2022 seasons, with a higher barrel plate appearance percentage (Brls/PA) denoting success, the specific metrics that correlate with a high success in baseball can be analyzed.

The dataset includes information on a variety of batting metrics.

- Batted ball events: any balls put into play by the batter, including fly balls, line drives, and ground balls.
- Launch angle: the vertical angle at which the ball leaves a player's bat after being struck.
- Sweet spot percentage: the percentage of batted balls hit within a certain optimal range (between 8 and 32 degrees) of launch angle and exit velocity.
- Max ev and average ev: the maximum and average exit velocity, respectively, of batted balls.
- Fly ball line drive ev and ground ball ev: the exit velocity of batted balls hit in the air and on the ground, respectively.
- Max distance and average distance: the maximum and average distance, respectively, that batted balls travel before landing.
- Average home run: the average distance that home runs hit by the player travel.
- Hard hit 95 mph+ and hard hit percentage: the number and percentage, respectively, of batted balls hit at a speed of 95 mph or greater.
- Hard hit swing percentage: percentage of swings taken by the player that resulted in batted balls hit at a speed of 95 mph or greater.

- Total barrels: the number of batted balls hit with an optimal combination of exit velocity and launch angle.
- Barrels batted balls percentage and barrels plate appearance percentage: the percentage of batted balls and plate appearances, respectively, resulting in a barrel, which is a batted ball with an optimal combination of exit velocity and launch angle.

The relationships between these metrics and Brls/PA can determine which factors contribute to a player's success in generating high-quality batted balls. By examining the correlations between the metrics and the success metric and the levels of success of different machine learning models, one can gain insight into which values are most consequential for generating high-quality batted balls.

The primary objective of this analysis is to identify the key batting metrics that significantly contribute to generating successful plate appearances in the MLB. By analyzing a dataset containing home run metrics for all players from the 2015-2022 seasons, the metrics that are most indicative of a player's batting performance and overall offensive contribution can be deduced. Additionally, I hope that these inferences will provide a plausible solution to a statistical anomaly that has perplexed baseball enthusiasts and analysts alike. The home run rates have followed a distinctive wave-like trajectory with a sharp and drastic increase over time (Figure 1). The data report will explore whether this trend can be attributed solely to players and teams utilizing data to improve their performance, or if there are other factors at play.

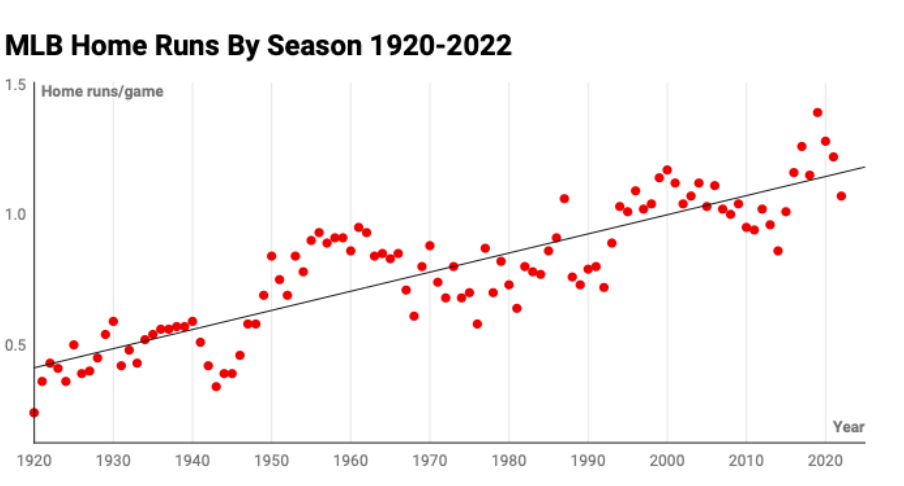

Figure 1: Home Run Rates per Year

By identifying the most decisive factors that contribute to success, I hope to find a strong correlation between certain metrics and performance so players and teams can optimize their training and improve their performance on the field while making strategic in-game decisions and ultimately improve a team's chances of winning respectively.

Machine Learning Models for Analyzing Baseball Metrics

In order to identify which factors are most indicative of a hitter's success, a correlation analysis was conducted which was used to represent a hitter's success compared to the success metric used for this analysis (Brls/PA) which can be seen in the heatmap below (Figure 2):

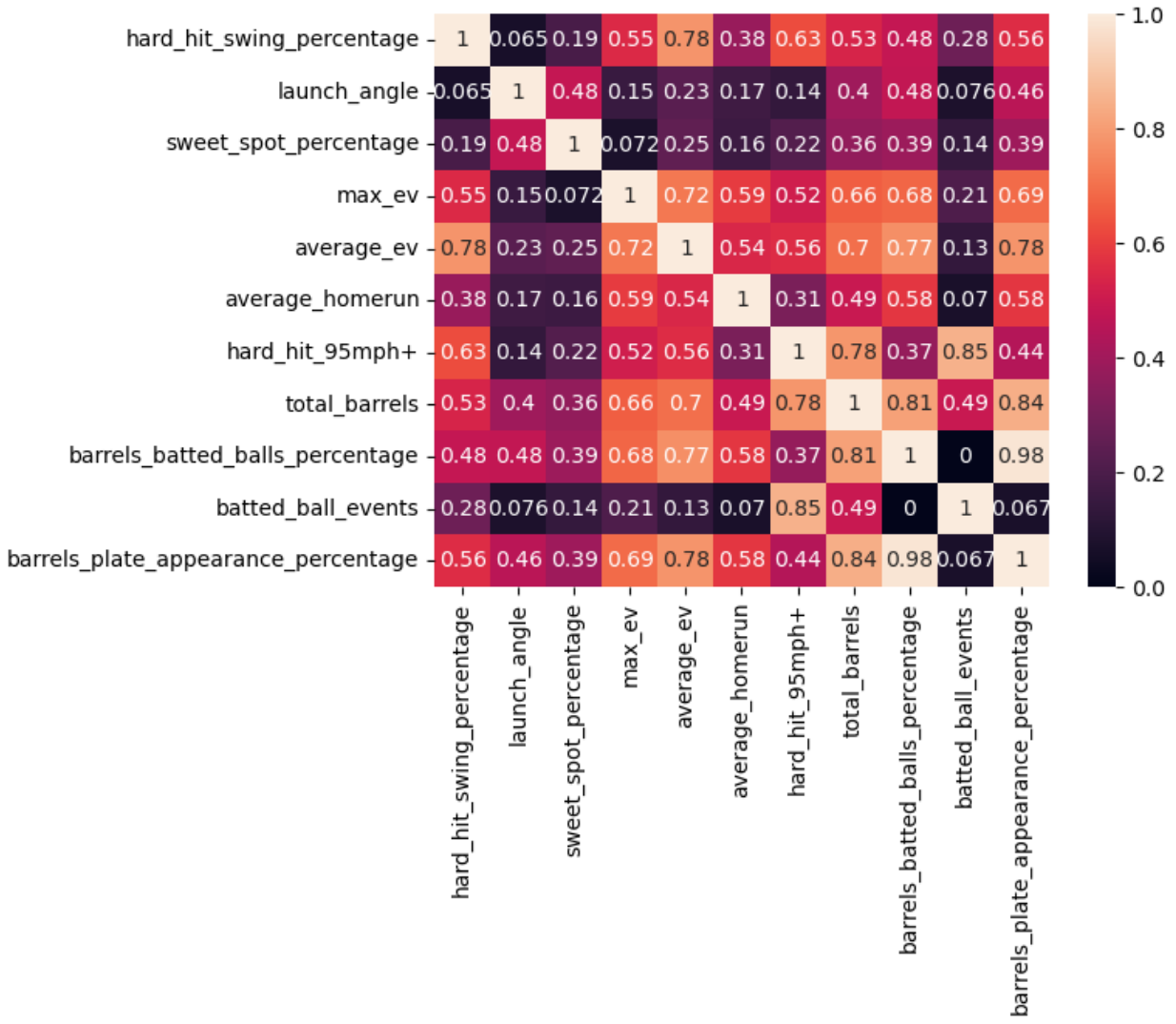

Figure 2: Correlation Matrix between each attribute on the dataset

In the above diagram, the larger value (on a scale from 0 to 1) represents an attribute that has a higher positive correlation, while the lower number indicates a lower positive correlation to the success metric. Several variables have a moderate to strong positive correlation with the success metric, including total barrels, barrels batted balls percentage, average exit velocity, and average home run. On the other hand, batted ball events, sweet spot percentage, and hard hit 95 mph+ have weaker correlations with the success metric.

This initial analysis provides insight into which variables may be most critical for generating successful plate appearances in the MLB. Furthermore, by using machine learning models such as linear regression and random forest, the relationship between these key variables and the success metric can be further analyzed. This will help to determine the strength of the correlations and how accurately they can be used to predict success.

Initially, a linear regression was formulated to determine the effectiveness of the prediction model. In such model, the R-squared value of 0.943 indicates that the model can explain around 94.3% of the variation in the barrels per plate appearance percentage using the selected input features. The higher the R-squared value, the better the model fits the data, suggesting that the selected input features are moderately good predictors of success.

The Mean Squared Error (MSE) of .145 stipulates that on average, the predicted values of the model deviate from the true values by .145, which is relatively low and suggests that the model is performing well.

Barrels batted balls percentage has the greatest weight on performance, followed by hard hit swing percentage, and total barrels which shows that these features have the strongest positive relationship with success. Alternatively, average ev and hard hit 95mph+ have weaker relationships with success, as indicated by their lower coefficients.

The random forest model had a mean squared error of 0.138 and an R-squared value of 0.93, suggesting that the model fits the data well and can explain around 93% of the variation in the barrels per plate appearance percentage. The feature importances were also calculated for this model, with barrels batted balls percentage having the highest and average homerun having the lowest significance.

The plot below shows the actual versus predicted values for both models, along with the ideal line representing perfect prediction. As seen in Figure 3, both models have similar patterns of

predictions, with some deviations from the ideal line. However, the random forest model appears to have slightly better predictions overall, as the data points are more tightly clustered around the ideal line.

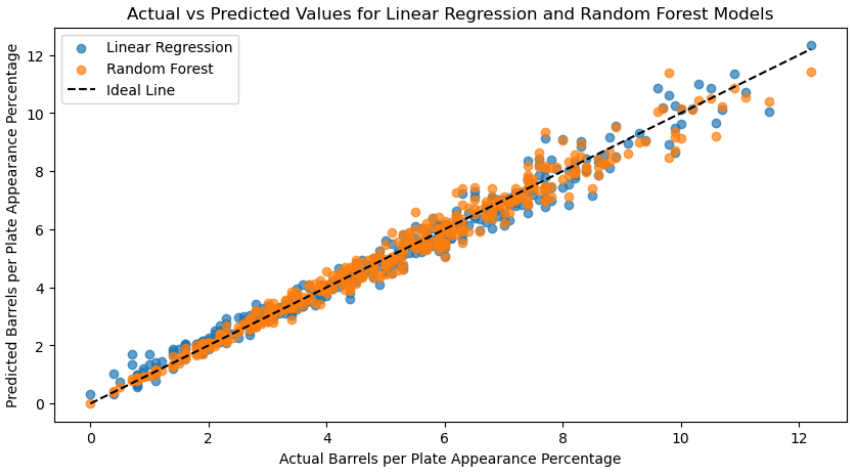

Figure 3: Scatter plot representing the Actual vs Predicted Values for Linear Regression and Random Forest

It is pivotal to note that while these models provide useful insight, they are not without limitations. For instance, additional variables such as weather, ballpark, and altitude that were not accounted for in the analysis may skew potential inferences. Another constraint is the reliability of the dataset, particularly in the year 2020 where the data may be altered due to the shortened season caused by the pandemic. As a result, the data for that year may not be entirely representative of a typical season. The schedule was much smaller, and teams had to deal with a wide range of challenges related to the pandemic, such as games being postponed or canceled due to outbreaks among teams.

The performance of the linear regression and random forest models was analyzed for each year. The plot (Figure 4) shows that for the year 2020, the mean squared error (MSE) experienced a 72% increase compared to the previous and following years, while the R-squared value is 92.7%

of what it typically is on average. This exemplifies that the data in 2020 is less reliable compared to other years.

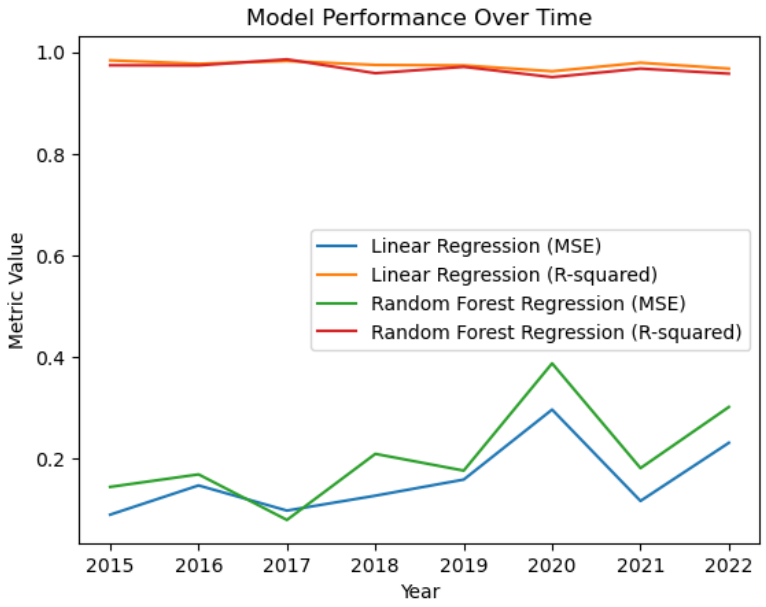

Figure 4: Line Graph representing the effectiveness of both Models

This analysis has provided valuable insights into the factors contributing to a hitter's success in the MLB, as measured by the Brls/PA metric. The correlation analysis, linear regression, and random forest models have identified key variables that demonstrate a positive correlation with the success metric, such as barrels batted balls percentage, hard hit swing percentage, and total barrels. These findings can be instrumental for coaches and trainers in designing targeted player development programs that focus on enhancing these attributes.

Common Metrics Among Top Performing Baseball Players

Along with model success, I was wondering if there were any specific statistical similarities that the top-performing players commonly possessed. To explore what top players have in common, a bar chart was created to show the percentage difference of the metrics for players who ranked in the top 10% for the success metric compared to the mean of all players in the dataset. The graphs below (Figure 5 and Figure 6) encapsulate such data.

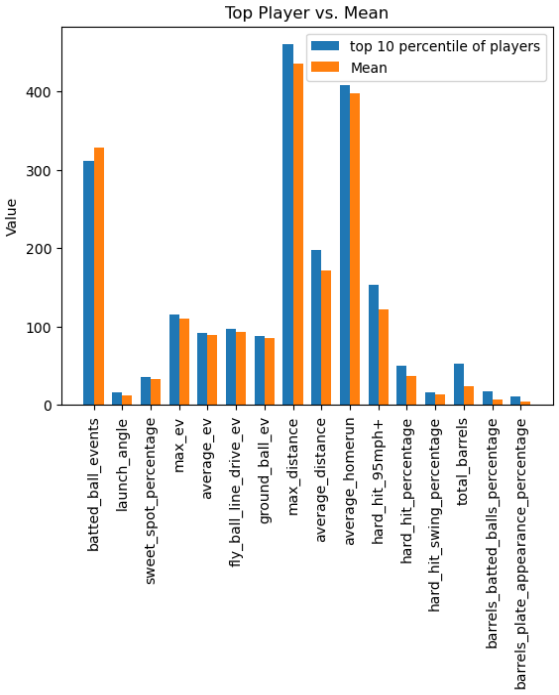

Figure 5: Performance metrics of the 90th percentile of players Relative to the mean

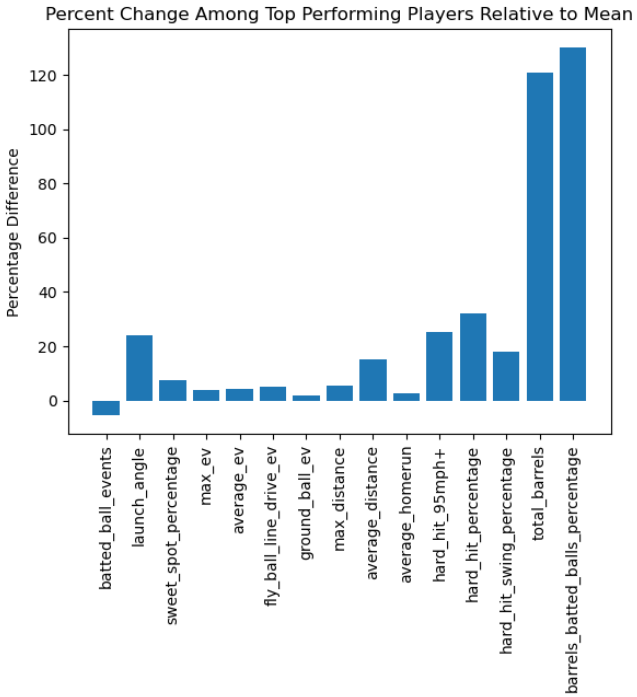

Figure 6: Metric Percent Change among the 90th percentile of players relative to the mean

Several common traits were identified among the top-performing players. These players tended to have a higher hard hit swing percentage (18% greater than the mean), better launch angles (23.7% greater than the mean), and greater hard-hit rates (29.5% greater than the mean). These metrics contribute to their ability to generate a higher number of barrels per plate appearance. Additionally, the top performers tended to have a higher percentage of their batted balls classified as barrels, further demonstrating their proficiency at producing high-quality contact consistently.

Interestingly, top performers in the league tend to have a lower batted ball events (BBE) metric on average compared to the mean of all players. As Eno Sarris discusses in *The Art of Not Swinging,* this could be attributed to a more selective approach at the plate, resulting in fewer overall batted balls but a higher quality of contact when they do make contact. A prime example of this selective approach can be observed in Juan Soto, a highly successful MLB player known for his plate discipline and ability to generate high-quality contact. Former Giants and current Diamondbacks slugger Evan Longoria mentioned in the article, “Over Time I have become more selective in a specific area.” By focusing on discipline and being more selective in their swings, these players achieve better outcomes and potentially maximize their performance

Furthermore, the article highlights the importance of balancing selectivity with the need for aggression. Padres outfielder Trent Grisham, known for his disciplined approach, acknowledges that being too passive can hinder performance. He states, “You’re going to get pitches to hit and you have to hit those, you’re a hitter first. So it’s a balance. Use it as an asset and not as a hindrance; it can hinder you if you’re too passive.” These insights from professional players offer valuable perspectives on the significance of striking a balance between discipline and aggression in order to optimize performance.

In addition, a new prediction model was created to estimate the Brls/PA metric for each top performer. The model was trained on a dataset of home run metrics for players from the

2015-2022 seasons, and it achieved a mean squared error of 0.140 and an R-squared value of 0.976. A comparison of the model's predicted Brls/PA values with the actual Brls/PA values for the top 10% of players is shown in Figure 7 below.

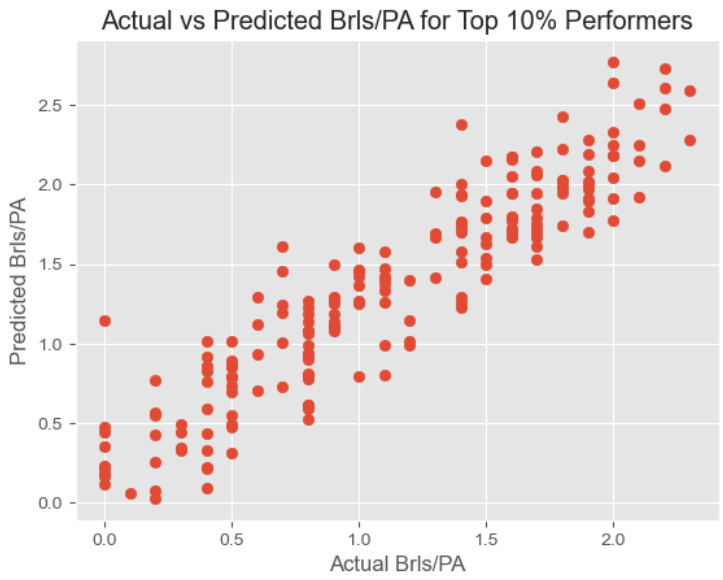

Figure 7: Scatter Plot of Actual vs Predicted Brls/PA for Top 10% Performers

The scatter plot in Figure 7 provides a visual representation of the relationship between the actual Brls/PA values and the predicted Brls/PA values generated by the model for the top 10% of players. It can be observed that the majority of the points lie close to a diagonal line, indicating a strong positive correlation between the actual and predicted values. This suggests that the model is generally accurate in predicting the Brls/PA for top performers.

The model's success in predicting Brls/PA for top performers also emphasizes the importance of the identified common attributes among these players. As top performers tend to have higher exit velocities, better launch angles, and greater hard-hit rates, it is crucial for coaches and trainers to focus on these areas during player development programs. By emphasizing these specific

attributes, teams can help players maximize their potential and contribute more effectively to the overall success of the team.

Determining the Correlation Between the Year and the Home Runs per Plate Appearance

As previously discussed, the home run rates have followed a distinctive wave-like trajectory with a sharp and drastic increase over time (Figure 1). This surge in home runs has raised questions and speculations about potential factors driving this phenomenon. To determine the correlation between the year and the home runs per plate appearance, two machine learning-based models, RandomForestRegressor and K-Nearest Neighbors (KNN) with k equal to 5, were developed and tested on available data. The results are portrayed in Figure 8.

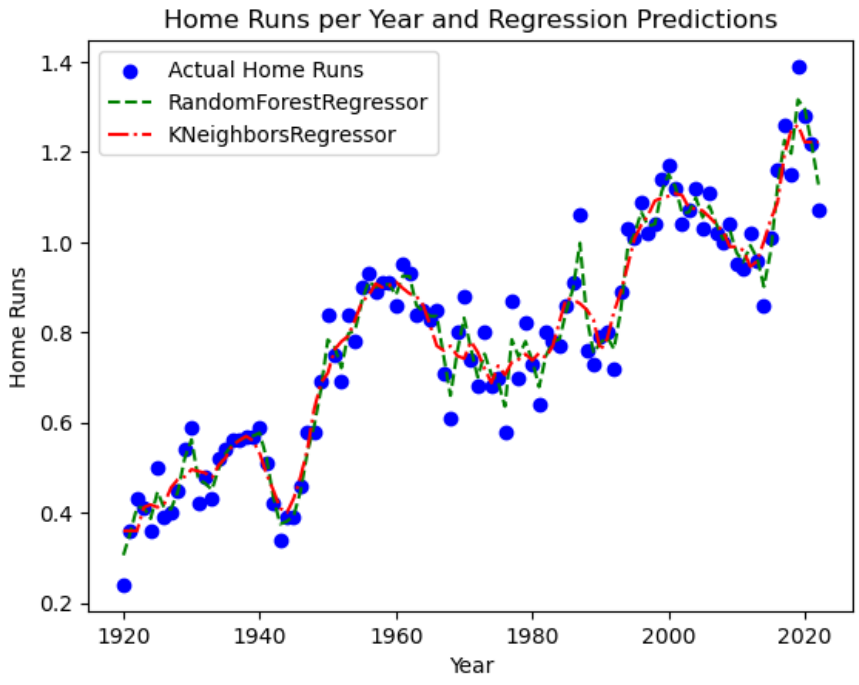

Figure 8: Actual vs. Predicted Home Runs per Year using RandomForestRegressor and KNN

The KNN model slightly outperformed the Random Forest model, with a marginally better MSE, and R-squared value. However, both models were unable to fully explain the underlying factors contributing to the increase in home runs. In 2018 MLB conducted a study with the intention of determining such factors. MLB's study, which analyzed Statcast (a tracking technology that allows for the collection and analysis of a massive amount of baseball data) batted ball data, found that “although there was not a substantial change in the percentage of batted balls falling within the right ranges of exit velocity and launch angle to create a home run, there was a remarkable change in the rate of home runs themselves.” This observation points to better carry as a primary factor, with the study discovering a decrease in the drag coefficient (a dimensionless quantity that is used to quantify the drag or resistance of an object in a fluid environment, such as air or water) of MLB balls since 2015.

While the decrease in drag coefficient could be considered a factor, there are aspects of the MLB report that can be construed as ambiguous. For example, the study did not provide a clear explanation for the decrease in drag and better carry, which warrants further investigation.

In contrast, my data report's findings challenge the MLB study's conclusions regarding the role of launch angles. This analysis indicates that the launch angle of top players in baseball is 22.4 percent greater than the mean, suggesting that launch angles may have a more serious impact on home run rates than initially believed.

Even as various theories and speculations have formed over the years, there is still no concrete explanation for the abnormal trends in home run rates. Further research is needed to uncover additional factors influencing these rates and to explore the implications of these findings on player training, coaching strategies, and equipment development. This deeper understanding of the complex relationships between home run rates and various factors will contribute to more informed decisions within the sport and potentially lead to innovations in player performance and game strategy.

Conclusion

The analysis of MLB data throughout the 2015-2022 seasons led to the identification of key batting metrics that significantly contribute to successful plate appearances. Some of these metrics include barrels batted balls percentage, exit velocity, and hard hit swing percentage. Machine learning models further validate these findings by providing data-driven evidence of the impact of specific metrics on a hitter's success. These models can not only prove existing knowledge but also uncover previously unknown relationships between various factors and a hitter's performance.

However, despite the wealth of knowledge provided by analytics and machine learning models, baseball remains a sport with many unexplainable anomalies that we may never truly understand. These occurrences remind us that the game still retains vast elements of unpredictability and human intuition, which are not easily captured by metrics and statistics. Researchers and analysts will always look to find a new correlation or data point that may help uncover the secrets behind these unexplained phenomena. This ongoing quest for knowledge highlights the vigorous nature of baseball and the sport's continuous evolution, fueled by the unification of data-driven analysis and the intangible elements that make the game so captivating yet timeless.

References

Castrovince, Anthony. “Key Takeaways from MLB Study of HR Rates.” *MLB.com*, MLB, 16 Feb. 2023, https://www.mlb.com/news/mlb-report-on-baseballs-home-run-rates-c278120310.

Nathan, Alan. “The Physics of Baseball Alan M. Nathan University of Illinois.” *The Carry of a Fly Ball*, http://baseball.physics.illinois.edu/carry.html.

Sarris, Eno. “Not Swinging Is an Art Form. but Is It Bad for the Sport?” *The Athletic*, 5 May 2022, https://theathletic.com/3292610/2022/05/05/mlb-plate-discipline/.