DSC 680
Benjamin Bartek
March 2, 2025

Analyzing MLB Player Performance Metrics and Game Outcomes

Business Problem

Major League Baseball (MLB) is a data-rich sport where player performance statistics play a crucial role in team management, player trades, scouting, and strategic game planning. Teams rely on comprehensive analytics to evaluate batting and pitching performance, determine key predictive factors of success, and optimize decision-making. However, identifying the most influential metrics and developing accurate prediction models remain challenges. This study seeks to address these gaps by analyzing batting and pitching data from the 2023 MLB season to derive actionable insights for teams, analysts, and stakeholders.

Background/History

Since the introduction of sabermetrics by Bill James in the 1970s, baseball analytics has transformed significantly. Advanced metrics such as On-Base Plus Slugging (OPS), Wins Above Replacement (WAR), and Fielding Independent Pitching (FIP) have become standard in evaluating players beyond traditional statistics like batting average and earned run average (ERA). With the rise of machine learning and data visualization, MLB teams increasingly leverage analytics to gain competitive advantages. This study builds upon this historical foundation by employing modern regression models, machine learning algorithms, and visual analytics to identify trends and predict player performance.

Data Explanation

The dataset for this analysis was obtained from Kaggle's 2023 MLB Player Stats, consisting of two CSV files: one containing batting data and another containing pitching data. The batting dataset includes statistics such as Batting Average (BA), On-Base Percentage (OBP), Slugging Percentage (SLG), Home Runs (HR), Runs Batted In (RBI), and Strikeouts (SO). The pitching dataset includes Earned Run Average (ERA), Strikeouts (SO), Walks & Hits per Inning Pitched (WHIP), Innings Pitched (IP), Home Runs Allowed per 9 Innings (HR9), and Walks per 9 Innings (BB9) (Vinco, 2023).

Data preparation involved cleaning missing values, encoding categorical variables, and engineering additional performance metrics. Exploratory Data Analysis (EDA) provided insights into the distribution of player statistics and relationships among variables, utilizing histograms, scatter plots, and correlation heatmaps. Figure 1 displays the correlation heatmap for batting statistics, highlighting key relationships between metrics, while Figure 2 presents the correlation heatmap for pitching statistics.

DSC 680
Benjamin Bartek
March 2, 2025

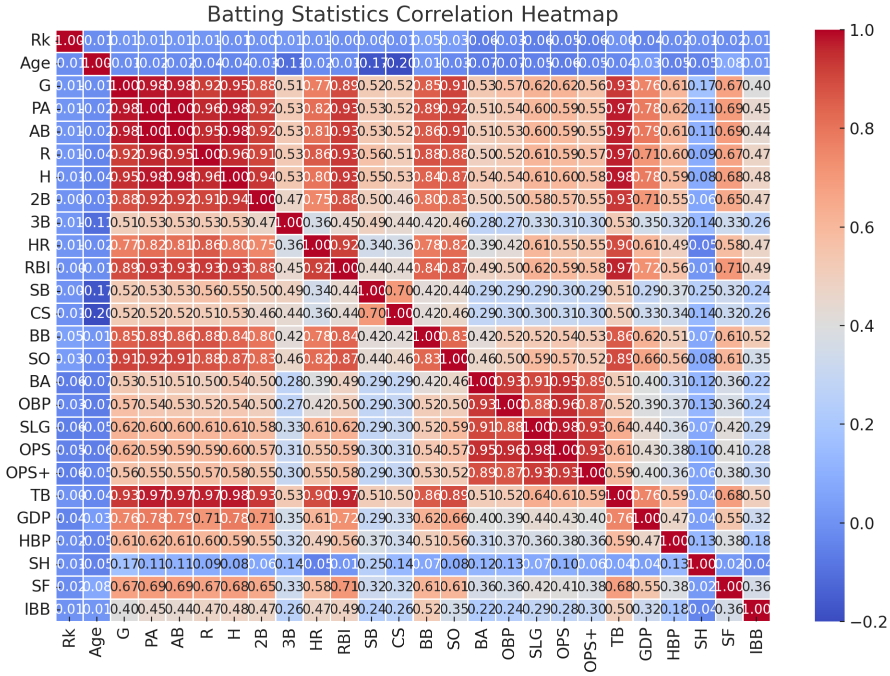

Figure 1: Batting Statistics Correlation Heatmap

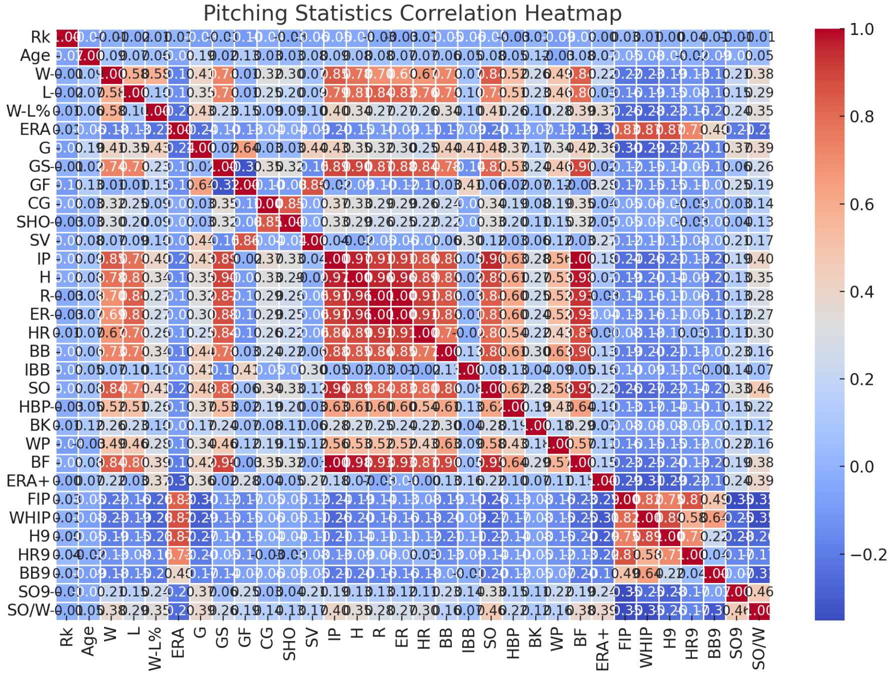

Figure 2: Pitching Statistics Correlation Heatmap

DSC 680
Benjamin Bartek
March 2, 2025

Methods

The study employs a combination of descriptive analytics, correlation analysis, and predictive modeling. Descriptive statistics provide an overview of key batting and pitching metrics. Team and position-based analyses aggregate statistics to compare performance trends across teams and player roles. Regression modeling using Linear Regression and Random Forest Regressors is applied to predict OPS for batters and ERA for pitchers. Feature importance analysis helps identify the most influential metrics in predicting player performance. Cross-validation techniques ensure that model accuracy and generalizability are tested effectively.

Analysis

Descriptive statistics revealed significant variability among MLB players in both batting and pitching metrics. The mean batting average across all players was 0.252, while the top quartile of hitters exceeded 0.290, demonstrating the importance of contact ability. Home run distribution was highly skewed, with elite power hitters reaching totals of 40 or more, whereas a significant portion of players recorded fewer than 10 home runs. Slugging percentage (SLG) and On-Base Percentage (OBP) both showed strong correlations with OPS, with SLG exhibiting the highest correlation at 0.85. These findings are visually represented in Figure 3, which shows team-level OPS comparisons.

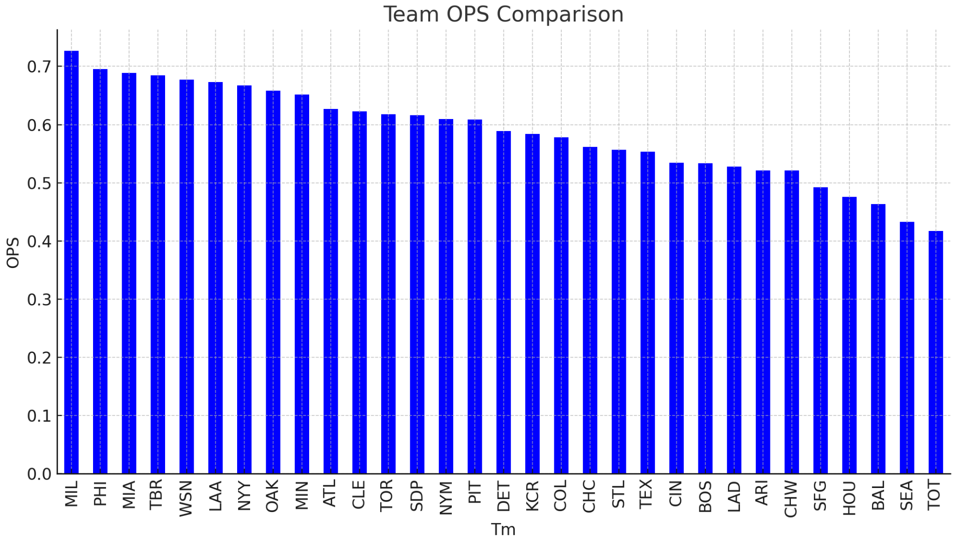

Figure 3: Team OPS Comparison

Team performance analysis indicated that teams with higher collective OPS scores also ranked among the league leaders in total runs scored. The correlation between OPS and runs scored was measured at 0.91, reinforcing its effectiveness as a predictor of offensive production. On the pitching side, teams with lower WHIP and HR9 values tended to maintain better ERA rankings, demonstrating the importance of controlling baserunners and limiting home runs. These results are illustrated in Figure 4, which compares team ERA rankings.

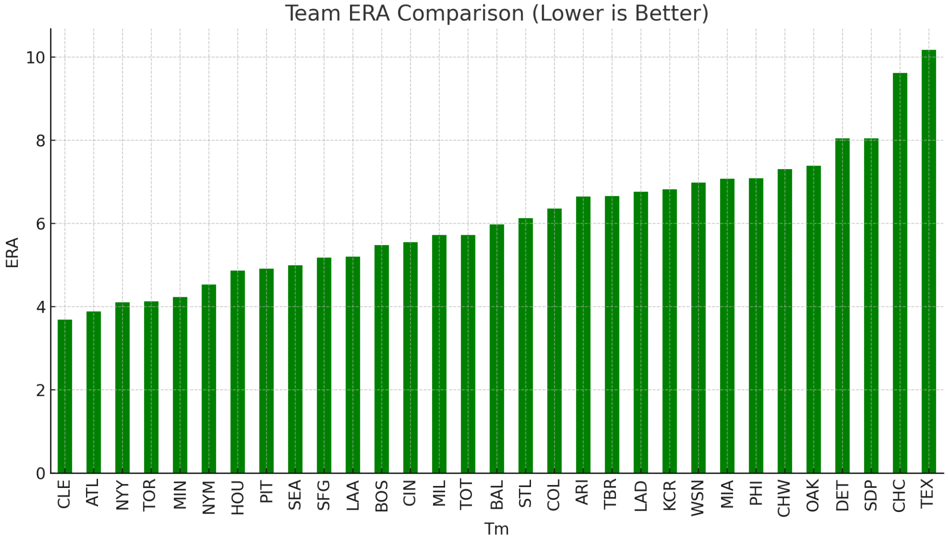

Figure 4: Team ERA Comparison

Regression modeling for batting performance confirmed that SLG and OBP were the strongest predictors of OPS. Linear Regression achieved an R-squared value of 0.999, nearly perfect, while Random Forest Regression produced a slightly lower R-squared of 0.997. These results indicate that OPS is highly predictable given a player's slugging and on-base capabilities. Pitching performance models were also effective, with Linear Regression producing an R-squared of 0.913 and Random Forest an R-squared of 0.871. Feature importance analysis identified WHIP, SO9, and BB9 as the most influential variables in predicting ERA, highlighting the importance of efficiency in limiting baserunners and generating strikeouts. Figure 5 and Figure 6 visually present the most important features identified in the predictive models for batting and pitching, respectively.

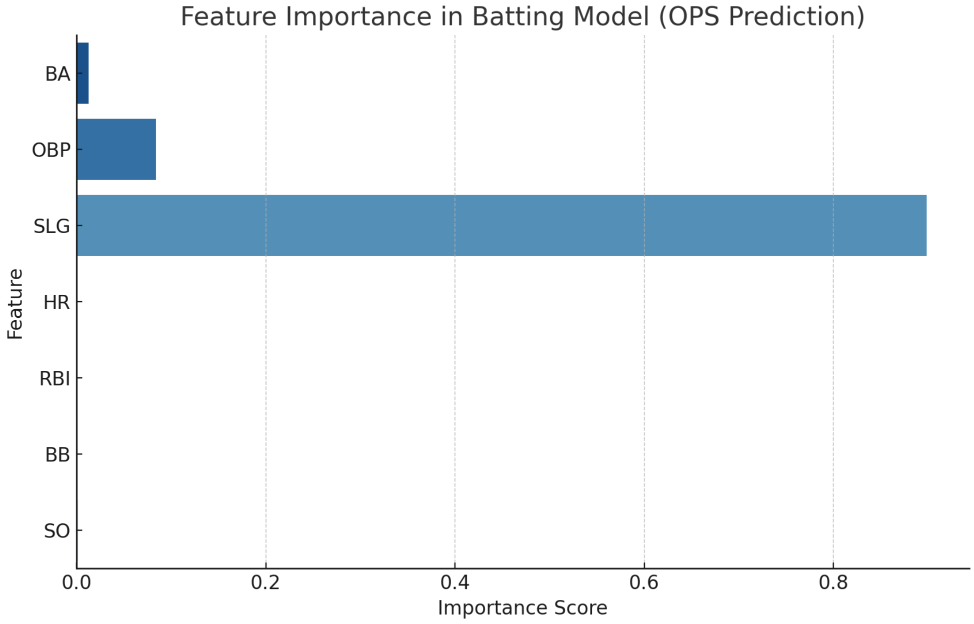

Figure 5: Feature Importance in Batting Model

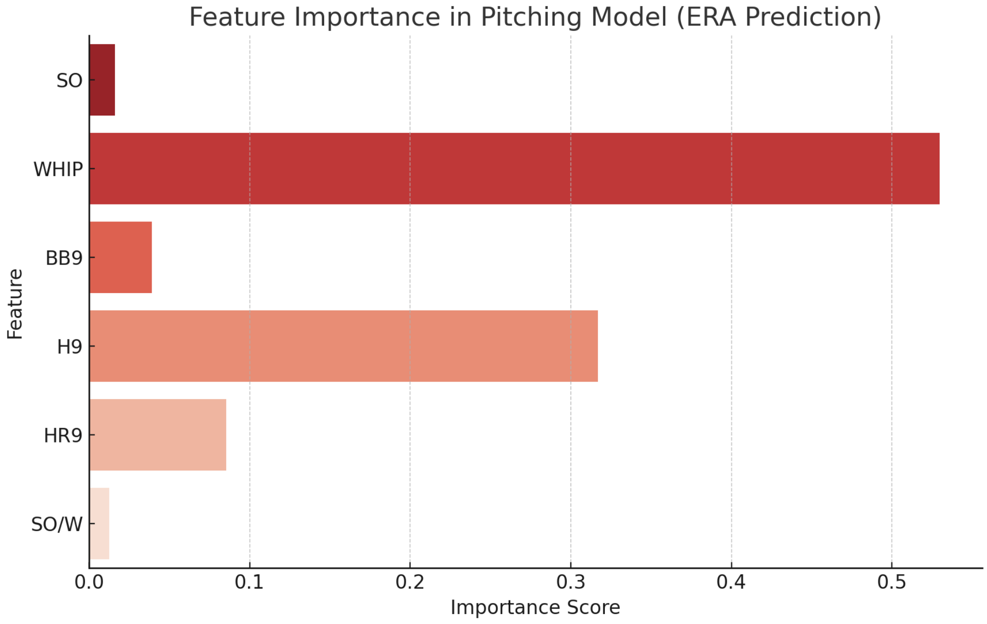

Figure 6: Feature Importance in Pitching Model

Additional insights were drawn from position-based performance analysis, which revealed that outfielders and first basemen typically led in OPS due to their offensive roles, whereas middle infielders and catchers showed lower OPS values but contributed more defensively. Among pitchers, starting pitchers tended to have lower ERA values compared to relievers, reflecting their ability to sustain performance over multiple innings.

Conclusion

This study demonstrates the power of data-driven decision-making in MLB. By analyzing player statistics and applying predictive modeling, we identified key performance drivers that influence success. Teams can use these insights to optimize scouting, lineup decisions, and trade evaluations. Further research should incorporate advanced sabermetrics and multi-year performance trends to enhance predictive accuracy.

Assumptions

This analysis is based on the assumption that the dataset is accurate and complete. It assumes that player statistics from a single season are representative of overall performance and that external factors such as injuries, weather conditions, and park effects do not significantly impact the trends observed. Additionally, it is assumed that standard MLB rules and game structures remain consistent across the dataset.

Limitations

One limitation of this study is that it is based solely on one season's worth of data, which may not account for long-term player trends or seasonal variances. The dataset does not include advanced sabermetrics such as Wins Above Replacement (WAR) or expected statistics like xBA (expected batting average), which could provide further insights. Furthermore, while machine learning models were used, they may not fully capture the complexity of human performance and strategic decision-making in baseball.

Challenges

A major challenge encountered in this analysis was handling missing or inconsistent data, requiring cleaning and preprocessing to ensure accuracy. Feature selection was another challenge, as determining the most predictive metrics required iterative modeling and validation. Additionally, computational resource constraints limited the use of more complex deep learning models that could have potentially improved prediction accuracy.

Future Uses/Additional Applications

This methodology can be expanded to analyze multi-season trends, providing deeper insights into player consistency and development. The models used can also be applied to minor league data for prospect evaluation. Additionally, integrating biomechanical and health data could enhance injury prediction and performance optimization for teams looking to improve player longevity and effectiveness.

Recommendations

MLB teams should prioritize OPS over traditional metrics like batting average when evaluating hitters, as it better correlates with run production. For pitching, reducing WHIP and walks per nine innings should be key objectives, as they strongly influence ERA. Future studies should incorporate additional sabermetrics and advanced machine learning techniques, including neural networks, to improve predictive accuracy. Additionally, leveraging real-time data from MLB Statcast can help refine models for in-game decision-making.

Implementation Plan

To implement these findings effectively, teams should integrate the predictive models into their existing analytics platforms. Scouting departments can use model outputs to inform player evaluations and trade decisions. Training programs should focus on improving the key statistical metrics identified, such as slugging percentage for hitters and WHIP for pitchers. Finally, teams should collaborate with data scientists to continuously refine models with updated data each season.

Ethical Assessment

Ethical considerations in baseball analytics include ensuring fair representation of players and avoiding over-reliance on data-driven decisions that may not account for intangible factors such as leadership and teamwork. Additionally, privacy concerns must be addressed if biometric and personal health data are integrated into future analyses. Transparency in model decision-making is also critical to ensure stakeholders understand and trust the results.

References

1. Vinco, V. (2023). 2023 MLB Player Stats [Data set]. Kaggle. https://www.kaggle.com/datasets/vivovinco/2023-mlb-player-stats

DSC 680
Benjamin Bartek
March 2, 2025

Appendix A

Data Dictionary

The following table provides definitions for key variables used in this analysis:

Column	Description
BA	Batting Average – total hits divided by at-bats
OBP	On-Base Percentage – frequency of reaching base via hits, walks, and hit-by-pitches
SLG	Slugging Percentage – weighted measure of power-hitting performance
OPS	On-Base Plus Slugging – combination of OBP and SLG to measure overall offensive value
HR	Home Runs – total home runs hit
RBI	Runs Batted In – total runs scored due to a player's hits
SO	Strikeouts – total strikeouts recorded
ERA	Earned Run Average – average runs allowed per nine innings by a pitcher
WHIP	Walks & Hits per Inning Pitched – measure of baserunners allowed per inning
SO9	Strikeouts per 9 Innings – average strikeouts per nine innings pitched
BB9	Walks per 9 Innings – average walks allowed per nine innings pitched

DSC 680
Benjamin Bartek
March 2, 2025

Audience Questions

1. What were the key findings from your analysis, and how do they impact MLB team decision-making?

2. How did you handle missing or incomplete data in your dataset?

3. Why did you choose OPS and ERA as the primary metrics for evaluating player performance?

4. What were the biggest challenges you faced while conducting this analysis?

5. How does your predictive model compare to existing baseball analytics models used by MLB teams?

6. Were there any surprising correlations or trends in the data that stood out to you?

7. What additional data sources or metrics would you include in future research to improve accuracy?

8. How can teams practically implement your findings in scouting, game strategy, or player development?

9. What ethical concerns should MLB teams consider when using data-driven decision-making?

10. How would you improve the machine learning models used in your study for future applications?