Moneyball 2.0 – Predicting Player Value from Performance Metrics

By Drake Rupple

WGU Data Analytics Capstone Project

Student ID: 012431628

Executive Summary

As Sabermetrics evolves, it becomes easier to quantify a Major League Baseball player's performance. Machine Learning algorithms can use advanced statistics to accurately predict how well a player will do over a period of time. This project hopes to similarly leverage linear regression models to answer a simple question: Can past player performance be used to indicate value compared to their salary?

This project will address this hypothesis: *A combination of performance metrics can reliably predict a player's Wins Above Replacement (WAR), and comparing WAR to salary will reveal significant inefficiencies in how players are valued.*

To predict WAR, this project uses the Python programming language along with many of its modeling packages that enable clustering and linear regression. With data obtained from the MLB and Baseball-Reference, a valuable model has been produced that is capable of providing scalable and predictive results.

By looking at the past, teams can accurately predict a player's WAR without the need for proprietary models. Comparing this derived statistic to a player's salary indicates who in the league is undervalued, or which ones are being paid too much.

Introduction

In Major League Baseball, player valuation has historically been influenced by subjective scouting reports and legacy statistics like batting average or RBIs. However, as the league has become more data-driven, front offices rely heavily on advanced metrics to evaluate performance and make multi-million-dollar decisions. Despite this shift, financial inefficiencies persist. Teams often overpay for name recognition or past success, while undervaluing players who contribute meaningfully but lack traditional appeal.

This project is motivated by the foundational work outlined in *Moneyball* by Michael Lewis, which showed how under-resourced teams could compete by exploiting market inefficiencies through statistical analysis. That philosophy still applies today, but the tools have evolved. With access to detailed public data and machine learning methods, it is now possible to build predictive models that estimate a player's value using objective performance indicators.

The research question driving this project is: *Can past player performance metrics reliably predict Wins Above Replacement (WAR), and does comparing predicted WAR to salary reveal inefficiencies in player valuation?* The hypothesis is that *A combination of performance metrics can reliably predict a player's Wins Above Replacement (WAR), and comparing WAR to salary will reveal significant inefficiencies in how players are valued.*

This project's scope includes MLB player data from 2024, focusing on batters and pitchers with sufficient playing time. It excludes minor league players, injury projections, and proprietary team scouting inputs. The goal is to use publicly available data and open-source tools to develop a regression model that predicts WAR, compute a WAR-per-dollar index, and visualize these insights through an interactive dashboard. By doing so, this project provides a scalable framework for identifying undervalued talent and supporting data-informed roster decisions.

Literature Review

This project draws on foundational and contemporary works in baseball analytics to inform its approach to modeling player value. The selected literature spans philosophical, practical, and technical perspectives, each supporting different components of the project.

Moneyball: The Art of Winning an Unfair Game

Lewis, M. (2003)

Michael Lewis's *Moneyball* introduced the baseball world to a radically different perspective on player value. The book details the Oakland Athletics' efforts under General Manager Billy Beane to build a competitive team on a limited budget using undervalued performance metrics like on-base percentage (OBP). The central thesis, that market inefficiencies in player valuation can be exploited through data, forms the philosophical backbone of this project.

This work justifies the project's core question: Can player value be measured more effectively using statistical models than conventional means? It also frames the use of WAR and salary comparison as a modern extension of Beane's original insight, seeking value not where it is obvious but overlooked.

The Only Rule Is It Has To Work

Lindbergh, B. & Miller, S. (2016)

In this book, two sabermetricians take control of an independent league baseball team and attempt to apply data-driven methods to real-time game and roster management. They encounter the limitations of analytics, especially in communicating data insights to coaches and players.

This work reinforces the need for models to be accurate, usable, and interpretable. It directly influences the decision to build visual dashboards and create a salary efficiency index that can translate model output into actionable insights for decision-makers. It also highlights the importance of contextualizing statistics within real-world constraints, which shapes how model outputs can be evaluated.

Predicting Baseball Player's Value through Machine Learning: "Traditional Metrics" vs "Advanced Metrics"

Franco, T. M. (2024)

This master's thesis applies multiple machine learning models to predict a player's annual average contract value (AAV) using traditional and advanced performance statistics. Franco finds

that models incorporating advanced metrics like WAR, wOBA, and FIP outperform those relying on legacy stats.

Franco's study supports the technical approach to this project. Specifically, the use of regression modeling and the emphasis on sabermetric inputs. It validates that machine learning techniques can effectively estimate player value and provides a precedent for evaluating model performance using accuracy metrics like R2 and RMSE. It also demonstrates the importance of feature selection in building predictive models.

Methodology (CRISP-DM Framework)

Business Understanding

Organization Need

Major League Baseball franchises are constantly pressured to build competitive teams within budgetary constraints. The cost of elite talent has increased, while the margin for error in player acquisition has shrunk. Front offices must identify good players and cost-effective players who deliver high performance for relatively low financial investment.

Stakeholders

- General Managers and Front Office Analysts who oversee roster construction and payroll allocation.

- Scouts and player development staff who can benefit from a clearer picture of a player's expected value.

- Ownership and finance executives who must evaluate return on investment for contracts. This model provides a scalable, explainable approach to support these decisions using publicly available data.

Data Understanding

Data Sources

- MLB.com: Official league and team statistics

- Baseball-Reference: WAR, salaries, advanced metrics

Initial Exploration

- Key statistics include OBP, SLG, K%, BB%, HR, SB, DRS, FIP, ERA, and innings played.

- Salary data ranged from league minimum to $40M+ per season, showing wide variance unrelated to performance.

Data Preparation

Cleaning Steps

- Missing values are handled by either imputation or row exclusion.

- Salary values normalized to millions for better scaling.

- WAR values are normalized for use in the linear regression model.

Feature Selection and Engineering

- Choose predictors with known correlation to WAR: OBP, SLG, BB%, K%, DRS.

- Create the derived metric WAR-per-million of salary.

Modeling

Model Used

- Linear Regression.

Tuning

- Features standardized before modeling.

Evaluation

Metrics Used

- R2: Target >= 0.65 for predictive strength.

- MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error): Target < 0.5 WAR error.

Insights

- A list of top undervalued players is generated based on WAR vs salary.

Validation Approach

- Train/test split (80/20).

Deployment

Visualizations and Dashboard

- Power BI Dashboard includes:
    - WAR vs Salary Scatterplots.
    - Filterable tables by team and player.
    - WAR/$ Rankings

GitHub Repository

- Contains modular Python scripts for data processing, modeling, and visualization

- Jupyter notebooks for exploratory analysis and iteration.

- README for reproducibility and setup.

- Final report and figures included as static outputs.

Reusability

- New seasons or player datasets can be plugged into the existing model pipeline.
- The tool can be extended to evaluate minor leaguers, free agents, or trade candidates.
- It can be the foundation for a live team valuation tool that updates the team.

Results and Findings

Preliminary Observations

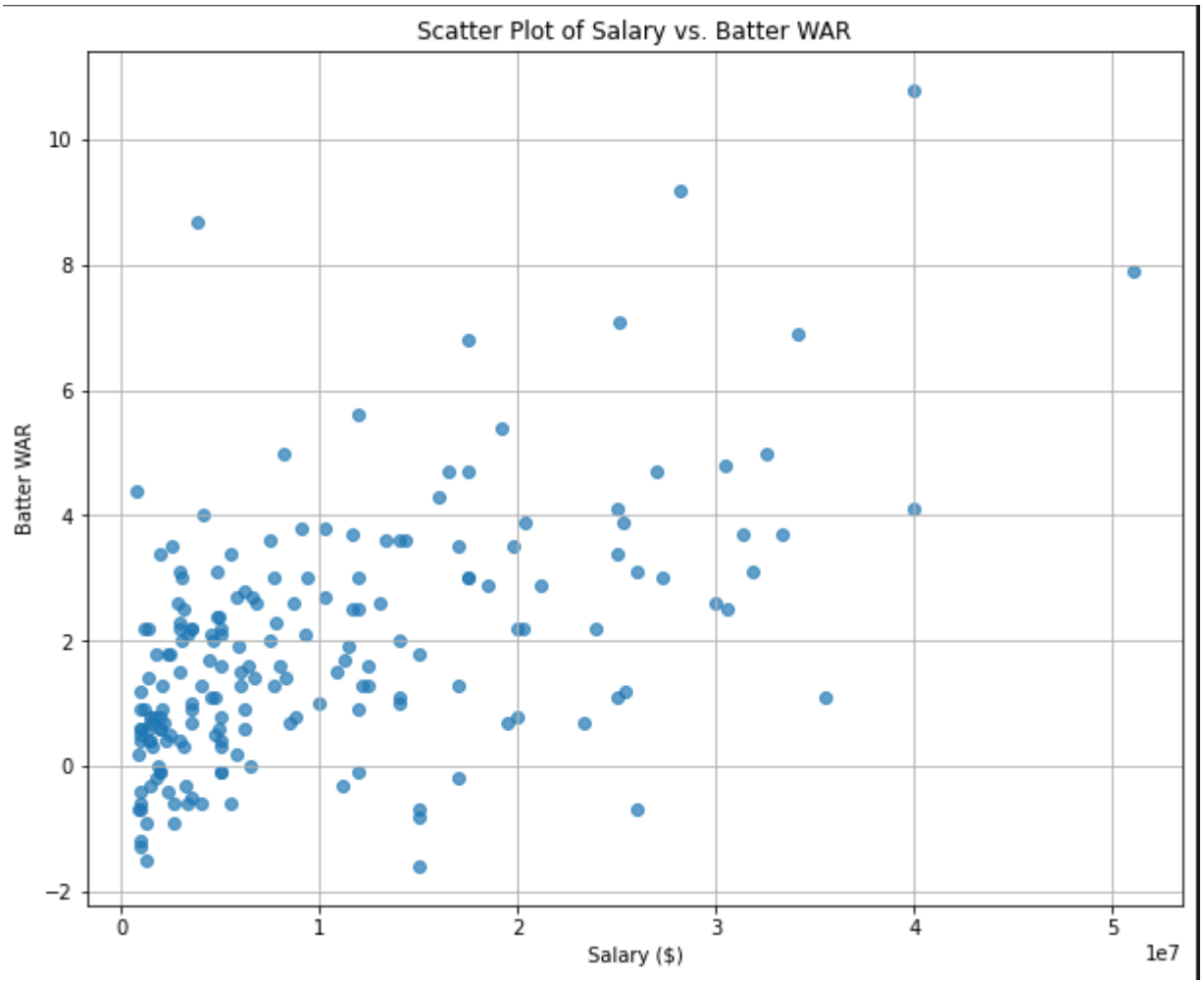

Figure 1: A scatterplot comparing player salary to WAR.

Comparing player salaries to WAR reveals no (or very weak) correlation between performance and pay. This justifies the assumption that there is little empirical reason why players' contracts are what they are. Despite these financial inefficiencies, player WAR is well distributed (Figures 2 and 3).

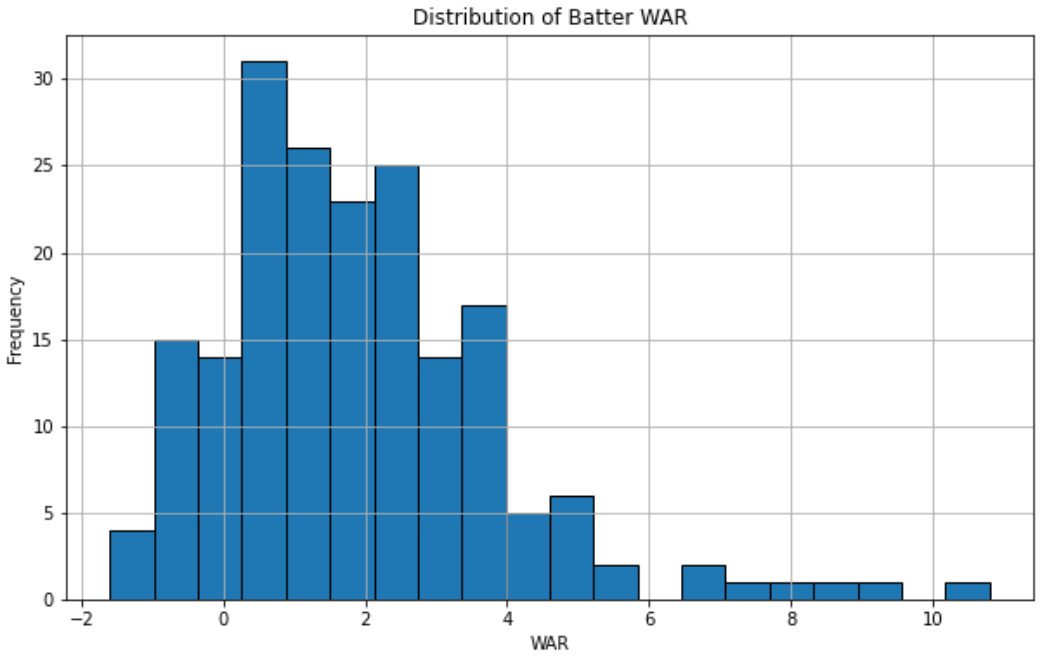

Figure 2: Distribution of WAR for batters.

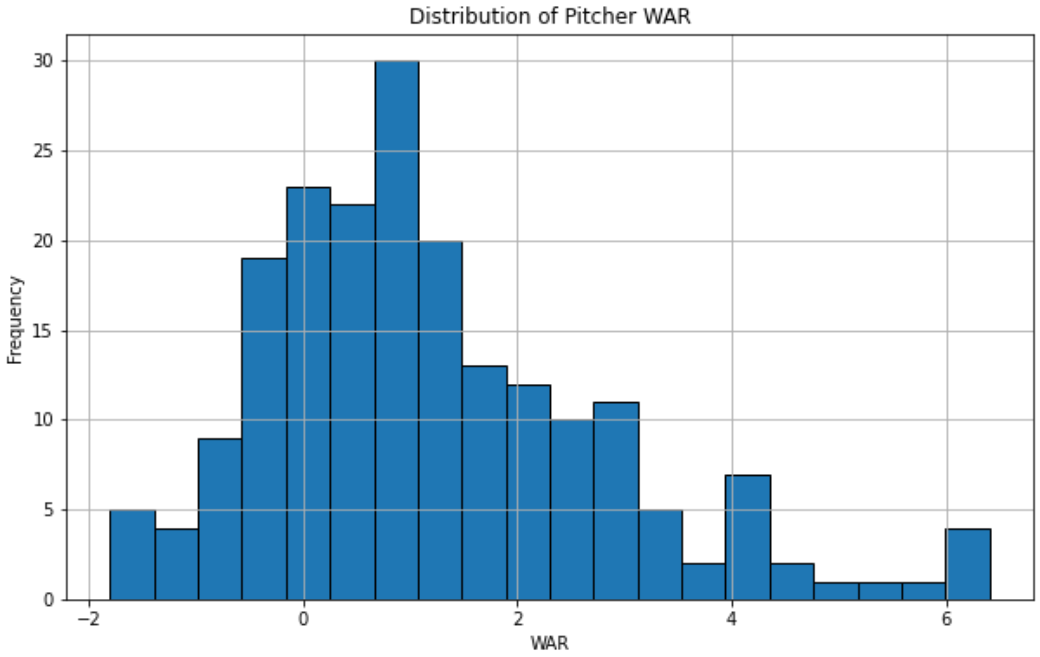

Figure 3: Distribution of WAR for pitchers

The bell curve distribution of WAR implies an even spread of player contributions, yet the absence of correlation with salary highlights a misalignment between pay and performance. This pattern suggests that many teams continue to overpay for reputation or positional scarcity rather than output, leaving significant value on the table. This project attempts to remedy this irrational approach by predicting WAR and comparing it to a player's salary.

What is Wins Above Replacement (WAR) and why predict it?

Wins Above Replacement is a comprehensive statistic that estimates the total value a player contributes to their team compared to a "replacement-level" player, who could easily be acquired from the minor leagues or bench. It incorporates offense, defense, baserunning, and positional value into a single number expressed as wins. A player with a WAR of 5.0 is estimated to have contributed five more wins than a replacement player would have in the same role.

A simplified, conceptual version can be expressed as:

$$\text{WAR} = \text{Batting Runs} + \text{Base Running Runs} + \text{Fielding Runs} + \text{Positional Adjustment} + \text{League Adjustment} - \text{Replacement Runs}$$

Each component is converted into runs, then adjusted and normalized to represent wins above replacement using a runs-to-wins conversion (Typically 10 runs ~ 1 win).

Although WAS is calculated retrospectively using full-season data, predicting WAR is strategically valuable because:

- Teams need to forecast future value when making contracts, trades, or roster construction decisions.

- Proprietary WAR models are often inaccessible; being able to predict WAR from public metrics democratizes talent evaluation.

- Predicting WAR allows for cost-efficiency analysis, enabling teams to identify players who will likely outperform their salary.

- It helps reduce financial risk by estimating how much production a player will likely offer before a contract is signed.

In short, predicting WAR enables proactive, data-driven decision-making and supports the core goal of building winning teams efficiently.

Building a Predictive Model

Using season data found on Baseball-Reference, this project provides a wealth of features contributing to a player's WAR.

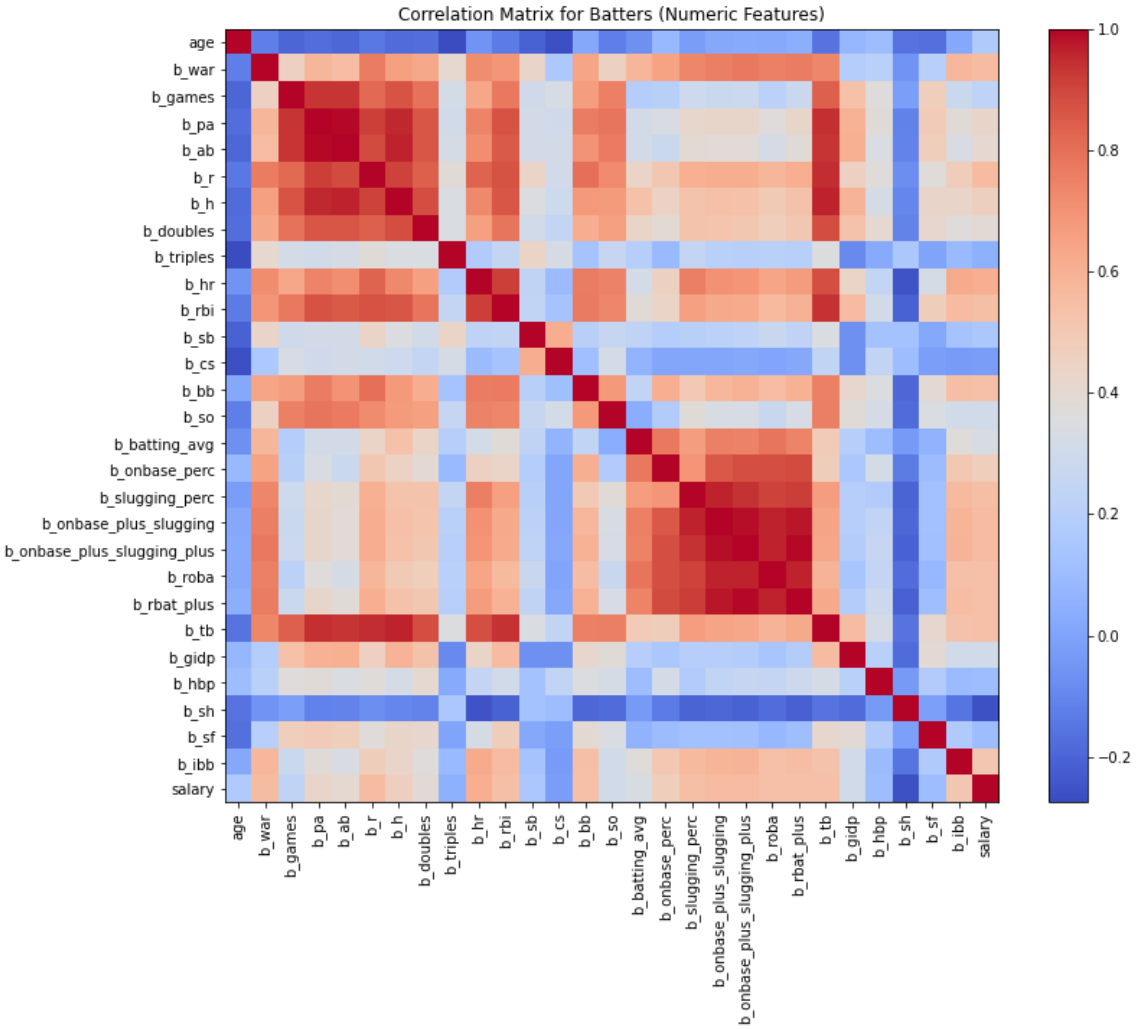

Figure 4: Correlation matrix showing relationship between various stats for batters

As we can see from the second row of Figure 4, many stats correlate positively with WAR. For this model, all numerical features were used. In other words, even those that did not positively correlate with WAR were included in the model. This decision allowed the regression algorithm to determine the actual predictive value of each variable rather than relying solely on correlation, which does not account for multivariate interactions.

With the complete feature set prepared, a linear regression model was trained on the normalized WAR values using an 80/20 train-test split. The model achieved a mean squared error (MSE) of 0.0042 on the test set, translating to an approximate error of 0.65 WAR when scaled back to the original value. Additionally, the model returned an R2 score of 0.7227, meaning it could explain over 72% of the variance in player WAR.

These results suggest that the model performs well predicting player value using accessible, non-proprietary performance metrics. While some features had limited standalone

correlation with WAR, their combined predictive power contributed to a robust model with relatively low error.

Identifying Undervalued Players

Using a derived efficiency metric comparing WAR to salary, the following figure plots the distribution of players based on their performative value. From left to right, players move from overvalued to undervalued.

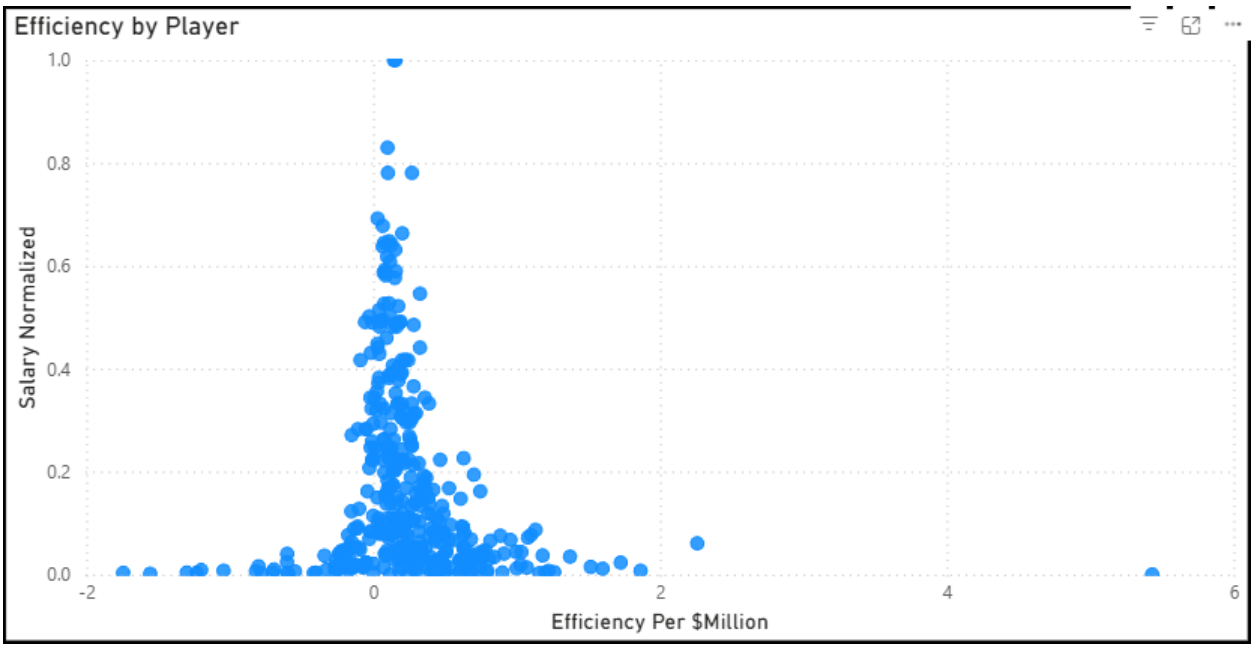

Figure 5: Players plotted by Efficiency/$million of salary

From this, it is easy to determine the league's top 10 most undervalued players (Figure 6).

Top 10 Undervalued Players

Player Name	Team	Salary	Eff/$Mill
Jackson Merrill	SDP	810,000.00	5.43
Jarren Duran	BOS	3,850,000.00	2.26
Joey Bart	PIT	1,180,000.00	1.86
Ernie Clement	TOR	1,970,000.00	1.73
Kevin Newman	ARI	1,375,000.00	1.60
Shane Baz	TBR	1,450,000.00	1.52
Geraldo Perdomo	ARI	2,550,000.00	1.37
Dylan Lee	ATL	1,030,000.00	1.26
Jorge Alcala	MIN	1,145,000.00	1.22
Jason Heyward	2TM	1,000,000.00	1.20

Figure 6: Most undervalued players

This table provides an immediate glimpse into which players are paid too little based on their performance. For team management, this could be an opportunity to offer a more enticing contract or for current owners to increase payout for retention. The following figure examines the efficiency at a team level, highlighting which organizations are the least efficient regarding payroll and performance.

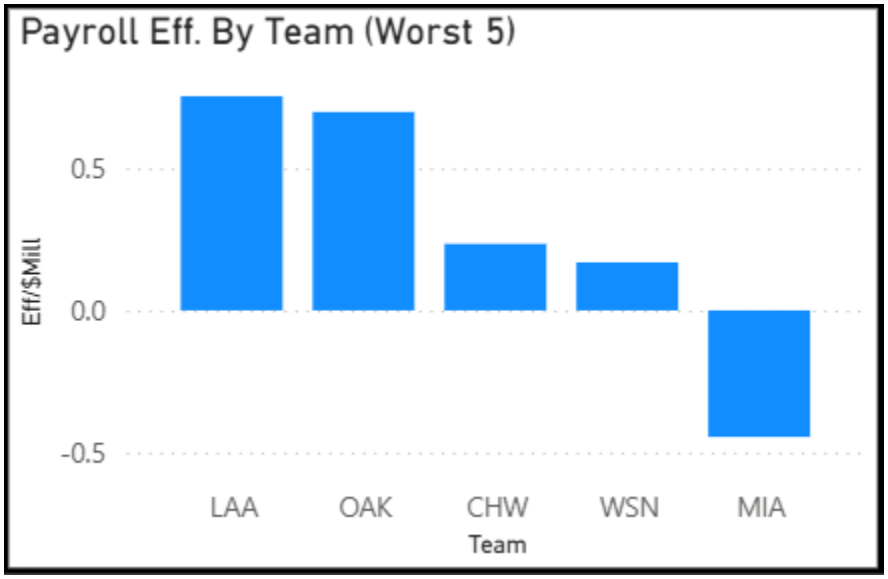

Figure 7: Least efficient teams

These findings would indicate that these teams must reevaluate how they recruit their players. Indeed, these teams finished at the bottom of their divisions during this season.

Discussion

The findings of this project highlight a critical disconnect in Major League Baseball: teams continue to spend millions on player contracts with little statistical alignment between pay and performance. As shown in Figures 1 through 3, player WAR follows a relatively normal distribution, indicating that contributions are spread across the league. Yet the scatterplot between salary and WAR demonstrates no significant correlation. This misalignment suggests that name recognition, historical performance, and positional scarcity may drive financial decisions more than objective metrics.

Interpretation in Real-World Decision-Making

The ability to predict WAR using a simple linear regression model and compare it directly to salary has clear implications for decision-makers. General managers, analysts, and finance executives could use tools like the efficiency index presented in Figures 5 and 6 to identify players likely to outperform their contracts. This could influence trade decisions, free agent signings, and contract extensions. Likewise, team-level efficiency comparisons (Figure 7) offer a macro view of how well (or poorly) an organization allocates payroll to actual on-field production.

This model, while simplified compared to proprietary systems used by MLB teams, demonstrates that significant insight can be drawn from public data alone. For smaller-market franchises, such tools could be affordable, transparent, and reproducible alternatives to expensive internal analytics systems.

Strengths and Weaknesses of the Model

Strengths:

- Strong predictive power with low average error.

- Based on accessible data, allowing complete transparency and reproducibility.

- Produces actionable outputs

Weaknesses:

- Linear regression does not capture non-linear relationships or interaction effects well.

- Assumes a static context (e.g., ignores injury history, team-level coaching quality).

- Relies on a simplified version of WAR, which may differ slightly from more advanced calculations.

Ethical Considerations

Valuing players strictly based on statistical efficiency raises critical ethical questions. While it can highlight undervalued players and support financially sustainable team-building, it also risks reducing human beings to economic units. For example, this approach may undervalue clubhouse leadership, veteran mentorship, or community presence, qualities that matter but are challenging to quantify.

Furthermore, models like this could be used to justify suppressing player salaries, especially for young players or those under team control, even when they are performing at elite levels. Therefore, while the model provides valuable insights, it must be used alongside human judgment and broader organizational values, not in place of them.

Conclusion

This project set out to answer a straightforward question: *Can a combination of performance metrics reliably predict a player's WAR, and does comparing this expected value to salary reveal inefficiencies in how MLB teams spend?* The results strongly support the hypothesis. Using a linear regression model trained on publicly available statistics, the project achieved a high R2 score and low error, indicating that WAR can be effectively estimated without proprietary tools. More importantly, comparing predicted WAR to player salary revealed significant market inefficiencies; many players are over- or underpaid relative to their performance.

These insights confirm what many have suspected since Moneyball's early days: despite an increasing reliance on analytics, financial decisions in baseball often lag behind performance data. The model developed in this project provides a transparent, data-driven method to highlight undervalued players, assess team-level payroll efficiency, and support more innovative roster construction.

Practical Recommendations for MLB Front Offices

- Use WAR/$ metrics to complement internal scouting and proprietary models to flag potential value acquisitions.

- Target undervalued players for trades or contract extensions before the market correction catches up with their performance.

- Conduct team-level efficiency reviews to assess whether payroll is allocated proportionally to performance.

- Incorporate transparent models like this into internal presentations to communicate performance value across departments.

Potential Improvements and Next Steps

- Model enhancements: Explore regularized regression (Ridge/Lasso), ensemble methods, or time-series forecasting for multi-year WAR projections.

- Additional features: Integrate park factors, injury history, or leadership metrics to refine predictions.

- Broader scope: Apply this model to minor league players, free agent markets, or international prospects using translated stats.

- Deployment: Package the model into a reusable app or dashboard that updates seasonally with new data.

In conclusion, this project demonstrates that with sound methodology and open data, it is possible to build practical, interpretable tools that offer meaningful insights into professional

baseball teams. By predicting WAR and evaluating it in the context of salary, teams can move one step closer to building competitive and financially efficient rosters.

References

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. W. W. Norton & Company.

Lindbergh, B., & Miller, S. (2016). *The only rule is that it has to work: Our wild experiment building a new kind of baseball team*. Henry Holt and Company.

Franco, T. M. (2024). *Predicting baseball players' value through machine learning: "Traditional metrics" vs "Advanced metrics"* (Master's thesis). Universidad Autónoma de Madrid. Retrieved from https://www.researchgate.net/publication/382249191_Predicting_Baseball_Players_Value_through_Machine_Learning