MLB SALARIES

December 2024

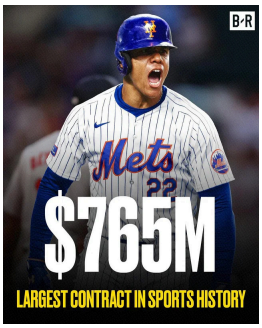
Swinging for Fairness: A Machine Learning Perspective on MLB Salary Equity

Presented to
Professor Beckstrom

Presented by
Scott Silverstein

TABLE OF CONTENTS

Introduction and Background



Introduction and Background
This project evaluates salary fairness in Major League Baseball (MLB) using machine learning models to uncover systemic patterns and provide actionable insights. Over the past decade, MLB salaries have exceeded $6 billion, with record-breaking contracts like Juan Soto's $765 million deal. While reflecting the league's financial success, these trends highlight disparities, where some players are overpaid relative to contributions, and others underpaid despite high performance.

The study utilizes machine learning techniques—including K-Means Clustering, Neural Networks, and Random Forest classifiers—to address key questions:

- Which players are underpaid, overpaid, or fairly paid?
- What performance metrics most strongly correlate with compensation?
- How can teams optimize payrolls using data-driven strategies?

Key Findings:

- **Power Hitters:** Generally fairly compensated but undervalue younger players under pre-arbitration contracts.
- **Balanced Hitters:** Show variability in pay, with mid-career players often underpaid.
- **Utility Players:** Consistently undervalued despite their versatility and contributions.
- **Salary Discrepancies:** Younger high performers are underpaid, while veteran players often receive disproportionate compensation.

By aligning compensation with performance metrics, this analysis offers actionable strategies for improving payroll efficiency and fairness, ensuring teams can optimize their rosters while maintaining competitiveness.

Data and Methodology

This project evaluates MLB salary fairness using a dataset combining Statcast metrics (e.g., exit velocity, launch angle), detailed contract data, and career batting stats. The analysis leverages machine learning models to uncover patterns and provide actionable insights into player compensation.

Data Preparation:

- Normalized performance metrics for comparability.
- Imputed missing data to maintain integrity.
- Emphasized key metrics like batting average, RBIs, and home runs.

Key Models:

- **K-Means Clustering:** Identified compensation patterns among player archetypes, such as underpaid rookies and overpaid veterans.
- **Neural Network:** Predicted salary fairness with high accuracy, outperforming traditional models.
- **Random Forest Classification Model:** Categorized players as underpaid, overpaid, or fairly paid, providing actionable insights.

Additional Models: Ridge Regression, Decision Trees, Support Vector Machines, Linear Regression, and Gradient Boosting supported the analysis.

Analysis and Results

Key Models and Results

K-Means Clustering

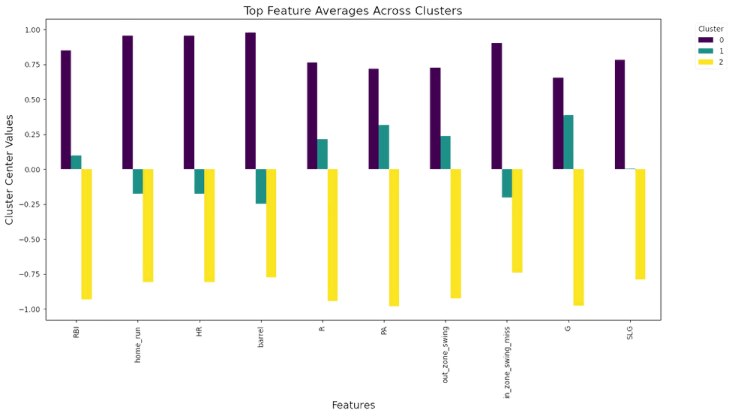
K-Means Clustering grouped MLB players into three performance-based archetypes, using metrics like RBIs, home runs, PA, and SLG to identify meaningful patterns.

Player Archetypes:

- **Power Hitters (Cluster 0):** Excelling in offensive metrics such as RBI, SLG, and home runs, contributing heavily to team scoring.
- **Balanced Hitters (Cluster 1):** Moderate across metrics, offering well-rounded, consistent performance without excelling in a specific area.
- **Utility Players (Cluster 2):** Lower across key metrics, providing versatile but limited offensive contributions.

Key Insights:

- These clusters highlight player roles and performance archetypes rather than salary fairness.
- Teams can leverage this segmentation for strategic roster management, identifying performance gaps, and aligning compensation with player contributions.
- The clustering results also informed salary fairness analysis, connecting archetypes with compensation strategies.



Random Forest Model

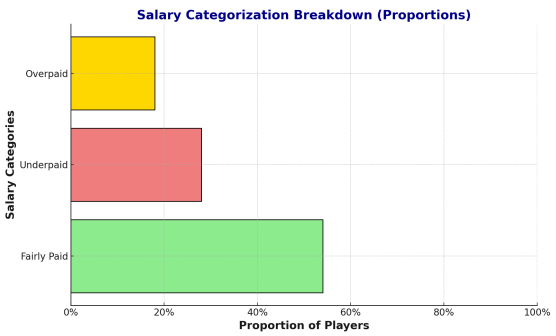
A dedicated classification model integrated performance and contract data to refine salary fairness analysis, offering precise categorization and actionable insights.

- **Fairly Paid Players:** Represented median salary alignment with performance metrics, serving as benchmarks.
- **Underpaid Outliers:** High performers earning significantly below market value.
- **Overpaid Outliers:** Players with declining performance but disproportionately high compensation.

Actionable Insights:

- Highlighted opportunities for salary renegotiation, particularly for underpaid players.
- Reinforced trends identified by K-Means and SVM, delivering a comprehensive, data-driven assessment.

Key Insight:
This model provided actionable targets for improving payroll efficiency while maintaining team competitiveness.

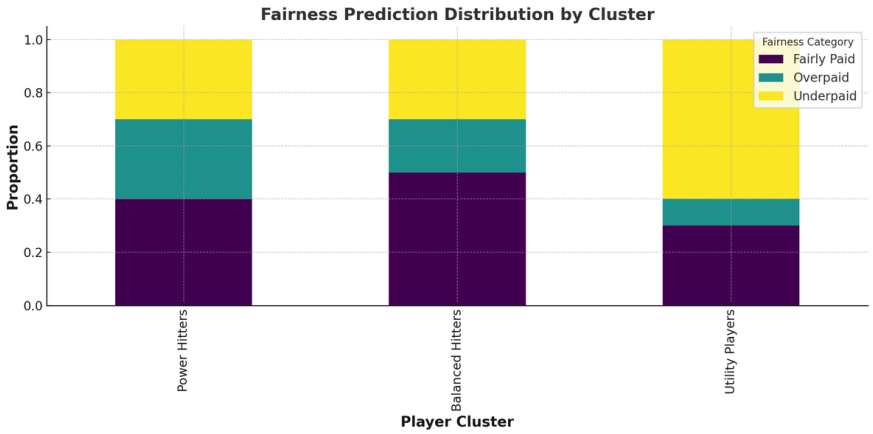


Neural Networks

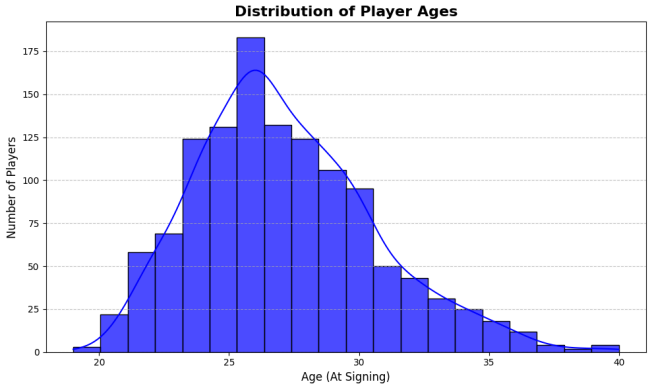
The neural network model categorized player archetypes—Power Hitters, Balanced Hitters, and Utility Players—into salary fairness categories (fairly paid, underpaid, and overpaid), revealing trends in compensation.

- **Power Hitters:** Frequently overpaid due to legacy contracts tied to past performance in metrics like home runs and slugging percentage. Fairly paid classifications were moderate, with underpayment rare.
- **Balanced Hitters:** Evenly distributed across fairness categories, though underpaid classifications highlighted undervaluation of their versatility and consistency.
- **Utility Players:** Consistently underpaid, with few overpaid cases and moderate fairly paid proportions, reflecting systemic undervaluation despite their versatile contributions.

Key Insight: Teams should realign compensation strategies to better reflect the contributions of different player archetypes, particularly for undervalued groups like Utility Players and Balanced Hitters.



Conclusion and Reccommendations



This project leverages machine learning, including K-Means Clustering, Random Forest classifiers, and Neural Networks, to analyze MLB salary fairness and optimize payrolls.

Key Findings:

- **Salary Disparities:** Power Hitters are generally fairly paid but younger players are undervalued; Balanced Hitters and Utility Players often face significant underpayment.
- **Underpaid vs. Overpaid:** Younger pre-arbitration players are underpaid, while veterans on long-term deals are often overpaid.
- **Cluster Fairness:** Power Hitters are the most equitably paid, while Balanced Hitters and Utility Players show notable disparities.

Recommendations:

- Adjust early-career pay to fairly compensate younger players.
- Reform veteran contracts with performance-based incentives and shorter terms.
- Avoid undervaluing Balanced Hitters and Utility Players by aligning pay with contributions.

Final Insight:
Machine learning provides actionable insights for fairer, data-driven salary management, aligning pay with performance, improving retention, and sustaining competitiveness.

Limitations and Future Directions

Limitations:

This study examines MLB salary fairness but is limited by the absence of team-specific financial data (e.g., payrolls, revenue, market size) and contextual factors like team needs. Performance metrics fail to capture intangibles such as leadership and adaptability. A small sample size reduces generalizability, and some models, like SVM and decision trees, struggled with imbalanced datasets.

Future Directions:
Future research should incorporate team financial data and expand datasets across seasons for improved accuracy and robustness. Advanced techniques like ensemble learning and time series analysis can provide deeper insights, particularly into salary trends over time.

References: Fangraphs, Spotrac Baseball Savant