Final Capstone Report

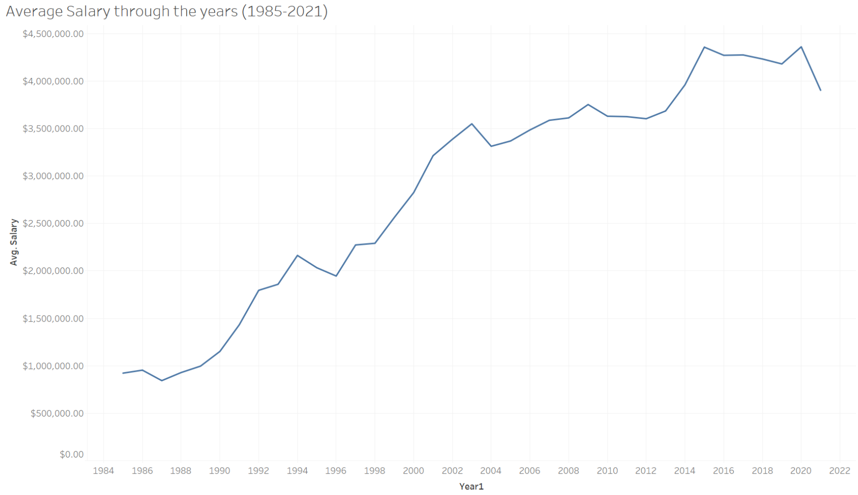
Predicting MLB Player Salaries: A Batting Performance Analysis

July 31, 2023

Hector Guillermo Guerrrero Jauregui
BrainStation
Vancouver, BC

Table of Contents

1. Introduction



Major League Baseball (MLB), a multi-billion dollar industry, has seen impressive economic growth, setting a new revenue record in 2022 of over $10.8 billion. With the average player salary reaching $4.22 million, up 14.8% from previous years, the intriguing dynamics of salary distribution among players have come into focus.

Teams in the MLB, regardless of size, face the challenge of building winning lineups on a budget while ensuring fair player pay. What really sets a player's salary? This project aims to answer this by predicting MLB salaries through a machine learning model, offering insights that could reshape roster planning in an era of growing salaries.

This study looks into batting stats, player performance, and salaries to see if more money means more wins, and if pay is really fair. By exploring these questions, we hope to uncover information that could change how teams handle player contracts, budgeting, keeping players, and overall planning in the MLB.

Drawing on historical baseball data from 1985 to 2021, the goal of this project is to develop a machine learning model capable of predicting MLB hitters' salaries. Such a model holds the potential to enhance our understanding of a baseball player's value and assist in identifying overvalued or undervalued players during team roster construction.

2. Value Add

The value add is a potential reshaping of how teams approach contracts, budgets, and roster planning, offering a data-driven guide in an industry where financials are continually escalating.

- Player Contract Negotiation
- Budget Planning
- Fairness and Transparency

3. The Data

Our study uses five datasets:

- WAR Dataset (Baseball Reference): 121,375 rows, 48 columns. Covers all parts of player performance, including batting and pitching.
- Batting Dataset (baseballdatabank - GitHub): 112,184 rows, 22 columns. Focuses on batting statistics for each season.
- People Dataset (baseballdatabank - GitHub): 20,811 rows, 24 columns. Contains players' biographical details.
- Salary Dataset (baseballdatabank - GitHub): 46,450 rows, 14 columns. Shows player salaries from 1985 to 2022, crucial for our analysis.
- Teams Dataset (baseballdatabank - GitHub): 3,015 rows, 47 columns. Offers team-based statistics for each season.

After merging, our dataset had **103,554 rows and 103 columns**. We trimmed it to **22,010 rows** by focusing on the salary data and removing pitchers, narrowing our study to the most relevant information.

4. Data Wrangling, Cleaning and Preprocessing

The process of building and preprocessing the dataset was a highly significant part of the project, laying the groundwork for all subsequent analyses.

1. Merging Datasets: As a starting point, five different datasets were combined to create one unified dataset. This was a key step, allowing for a complete look at baseball statistics and setting the stage for everything that followed in the project.

2. Adding Key Features: Essential baseball metrics like batting average (BA), on-base percentage (OBP), and slugging percentage (SLG) were calculated and added.

3. Cumulative Features: Cumulative features reflecting players' career statistics were calculated and added. These features became the most important predictors in the later modeling phase.

4. Time Span Setting: The analysis was restricted to post-1985 to maintain statistical consistency and account for changes in the game.

5. Dropping Irrelevant Columns: Unnecessary columns like personal, pitching, and team stats were removed to focus on batting performance.

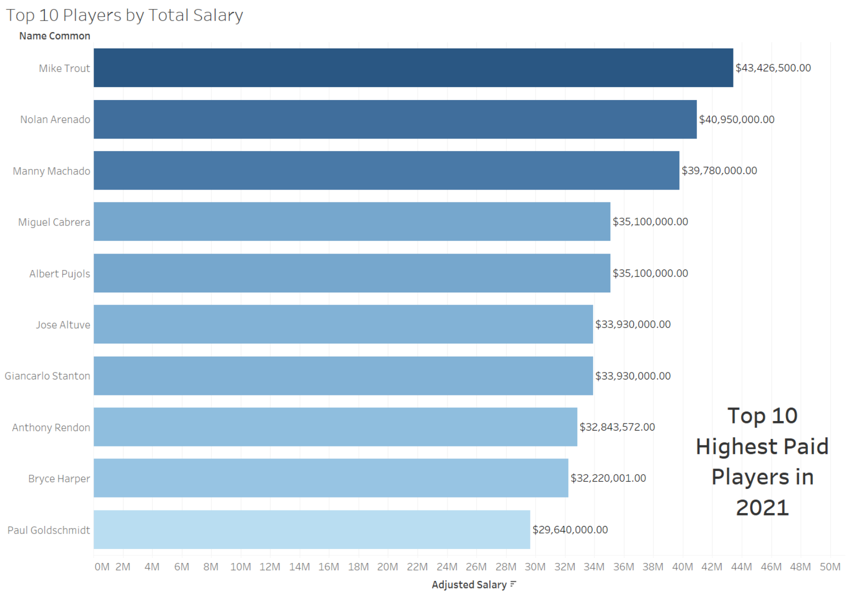
6. Handling Null Values: Special handling of null values, especially those from the 1994 season due to a labor strike.

7. Salary Adjustments: Salaries were meticulously adjusted to account for inflation. This critical adjustment ensures fair comparisons of player salaries across different years.

The result is a clean, tailored dataset containing 22,010 rows and 68 columns, ready for exploratory data analysis and modeling.

5. Exploratory Data Analysis and Feature Selection

The Exploratory Data Analysis (EDA) focuses on the Major League Baseball (MLB) players' data, exploring relationships between player performance, demographics, and salary.

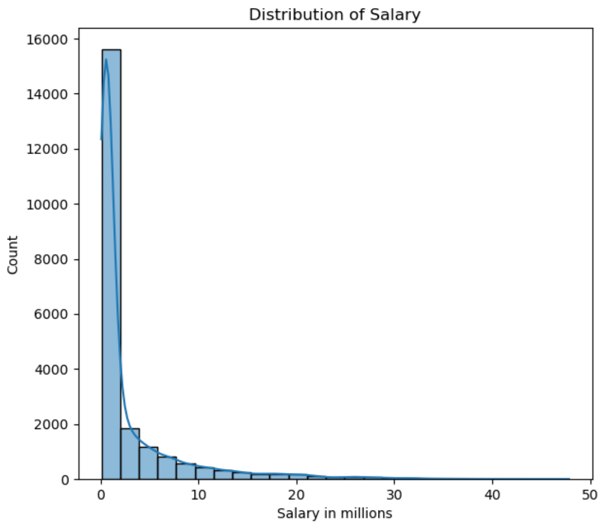


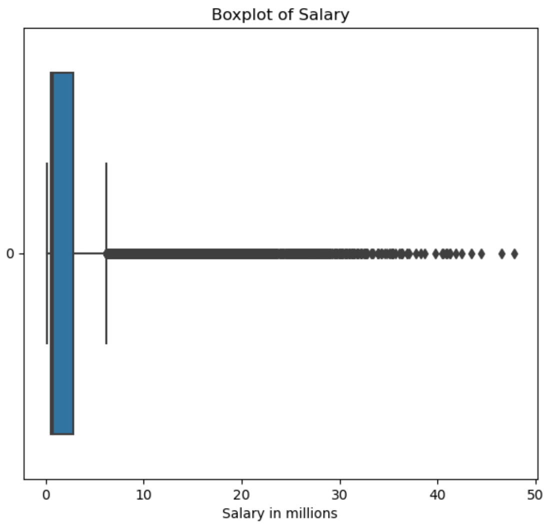
Top 10 Highest Paid Players in 2021

In just 2021, several of the top-earning batters are commanding contracts surpassing $30M, with Mike Trout leading at roughly $43.5M. Additionally, there are veteran players like Albert Pujols and Miguel Cabrera, who earned $35M, stemming from an extensive long-term agreement penned in 2012. These present and past giants of the game can equally be regarded as statistical anomalies.

Key Findings

- **Distribution of Salary**



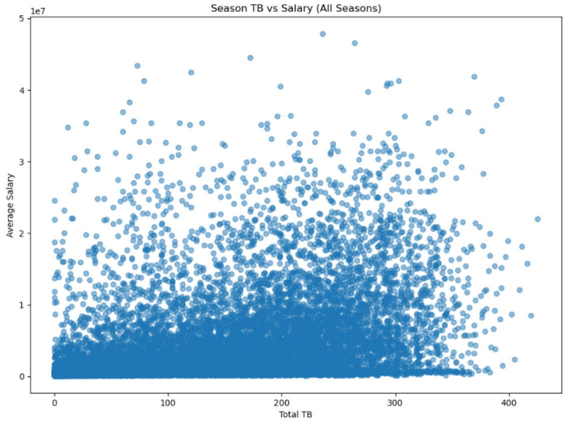


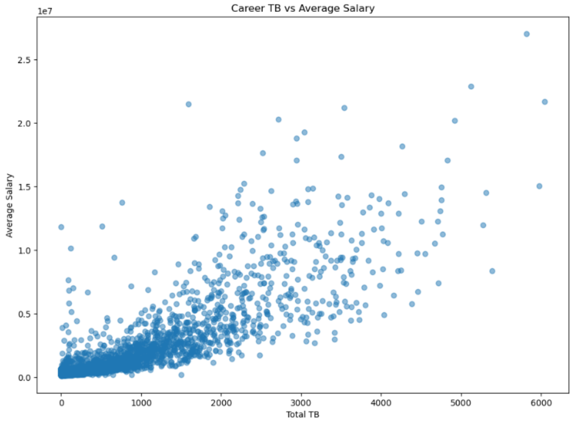
- Average Salary: $2,928,191
- Standard deviation: $5,158,169
- Minimum Salary: $72,500
- Median: $670,350
- Max Salary: $47,850,000

The long tail to the right in the histogram and a similar pattern in the boxplot indicate a right-skewed distribution, where most players earn below the mean salary of $2.93 million. The skewness emphasizes the presence of a few superstar players with exceptionally high salaries, which pull the mean upwards and widen the distribution. The boxplot would likely show many outliers on the higher end, reflecting these high earners, while the majority of players earn closer to the median of $670,350.

- **Relationship Between Salary and Performance Metrics**

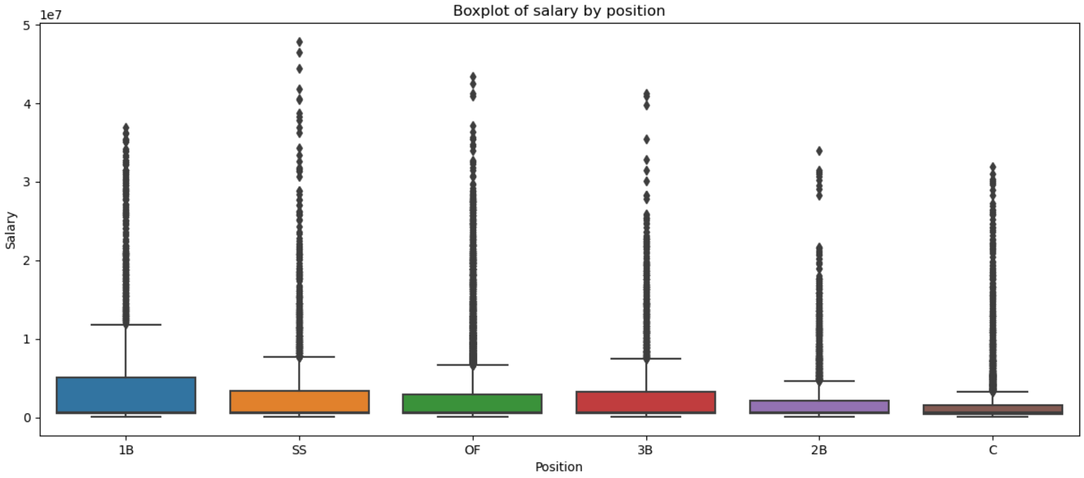
The Exploratory Data Analysis (EDA) helped us uncover important connections between a player's salary and how they perform on the field. We found that a player's overall career achievements have a bigger effect on their salary compared to just one season's performance.





The plots reveal a significant positive relationship between Total Bases (TB) and player salaries, this is just one example of a key performance metric of the many we analyzed. Most of them show the same pattern.

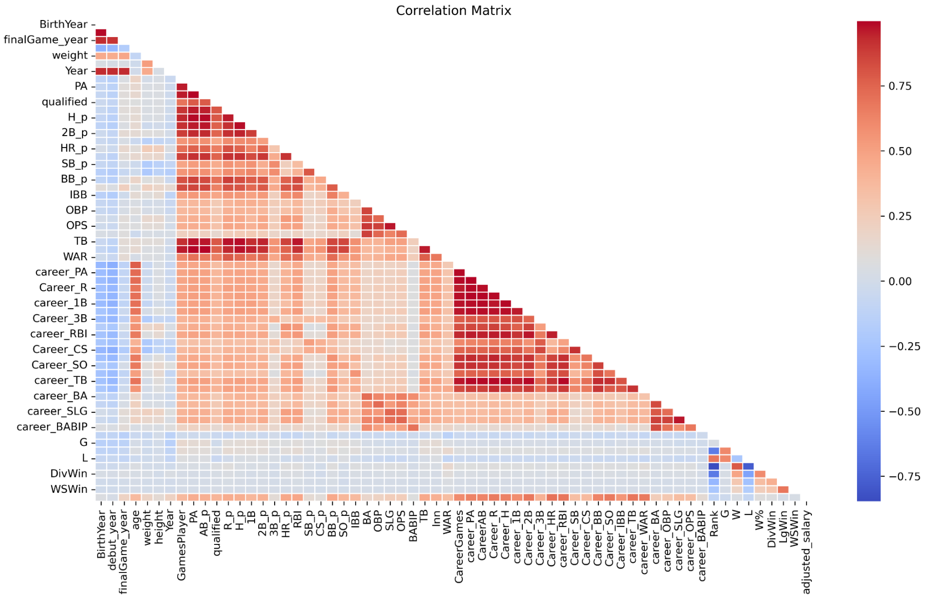
- **Relationship Between Salary and Positions**



This boxplot suggests that the position a player plays on the field could be an influential factor in determining their salary. This could be due to different skill requirements, injury risks, or supply and demand dynamics for players in each position.

Multicollinearity and Feature Selection

We used SelectKBest to pick the top 10 features that have the most influence on a player's salary. We also applied our understanding of baseball to choose some extra metrics and remove others.



6. Modeling and Results

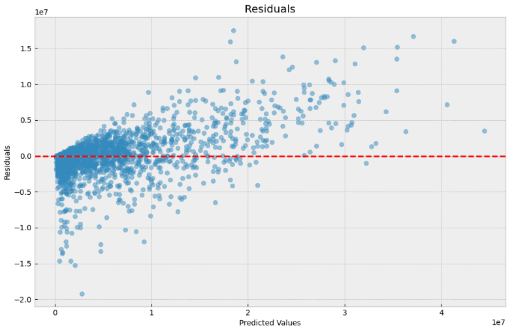
The aim was not only to build models that accurately predict salaries, but also to derive insights on what aspects of the game most significantly influence a player's earnings. For this purpose, we started with a simple linear regression model as our baseline, and then explored more complex models such as Ridge and Lasso Regression, Decision Trees, Random Forests, and Gradient Boosting. Each model was evaluated based on its predictive accuracy.

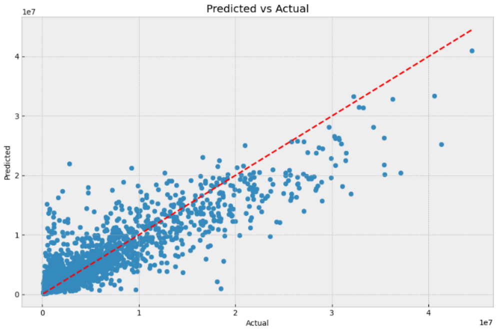
Hyperparameter Optimization: GridSearchCV was utilized across all models to find the best hyperparameters by performing an exhaustive search over a specified parameter grid.

Models

- Linear Regression (Baseline):
  - $R^2 = 0.620$.
  - RMSE = ±$3,222,420.69
- Ridge Regression
  - $R^2 = 0.62$
  - RMSE = ±$3,223,089.73
- Lasso Regression
  - $R^2 = 0.62$
  - RMSE = ±$3,229,146.47
- Decision Tree
  - $R^2 = 0.785$
  - RMSE = ±$2,422,400.47
- Random Forests
  - $R^2 = 0.826$
  - RMSE = ±$2,180,027.92
- **Gradient Boosting Regressor**
  - **$R^2 = 0.841$**
  - **RMSE = ±$2,080,887.37**

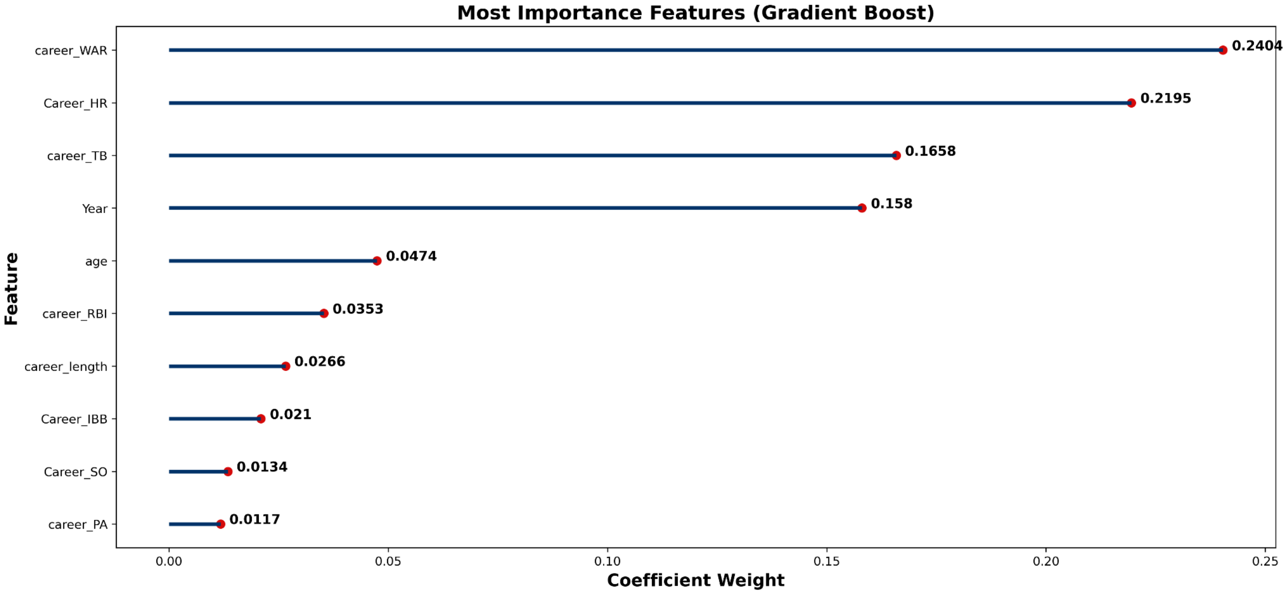
The Gradient Boosting model had the best performance, with an R2 score of 0.841 and an RMSE of about $2.1 million on the test set. This means that about 84.1% of the variance in the salaries can be explained by the model, and the predictions are typically off by about $2.1 million.





- **Homoscedasticity:** The residuals seem to be randomly scattered around the zero line. This suggests that the assumption of homoscedasticity (constant variance) of the residuals is reasonably met.
- **Independence:** The residuals do not display any clear patterns (they look quite random). This suggests that the assumption of independence is not violated.

Feature Weights



The feature importance of the Gradient Boosting model shows which features are the most influential in predicting a player's salary. The most important features are Career War, Career HR, and Career TB.

- **Career WAR**: This is the most important feature according to the model. It suggests that a player's Wins Above Replacement throughout their career is the most influential factor in determining their salary.
- **Career HR and career TB**: These are the second and third most important features respectively. They represent a player's Home Runs and Total Bases in their career. This implies that players who have hit more home runs and achieved more total bases throughout their career tend to have higher salaries, which makes sense given that these are indicators of a player's offensive performance.
- **Career Length and Age:** These features are also quite important. This suggests that more experienced players, and older players, tend to earn higher salaries.
- **Career SO, Career IBB, Career R, career PA, Career BB, career RBI, Career H**: These statistics from the player's career also play an important role in their salary.

The less important features appear towards the bottom of the plot. These include the player's position (POS), their batting hand (bats), their throwing hand (throws), and their team (teamName). These features have some influence on the salary, but much less than the player's performance statistics.

Gradient Boosting Regressor Behavior (Best Model)

In conclusion, the model explains approximately 84.1% of the variation in players' salaries (Test R2: 0.941) with an average error (RMSE) of $2,080,887.37. It likely performs better for lower to middle salary players, where salaries are tied to measurable performance metrics. Higher salary players, influenced by factors like marketability, may pose a challenge, as indicated by the wider spread of residuals for those salaries. The model performs reasonably but may benefit from refinement to predict the salaries of superstar players more accurately.

7. Conclusion, Limitations and Future Steps

This study successfully explored the complexities of Major League Baseball players' salaries using a robust predictive model. The standout was the Gradient Boosting model, with an R^2 of 0.841 and an RMSE of ±$2.1 million, highlighting key factors such as Career WAR, Career HR, and Career TB in determining salaries.

Limitations

- Exclusion of Non-Statistical Factors: Other influences on salary, such as marketability, leadership, and fan appeal, were not considered in the analysis.
- Salary Structure Limitations: The analysis does not consider the complex nature of player contracts, including long-term agreements, salaries prior to arbitration, bonuses, and other financial arrangements that could influence a player's official salary. This may lead to discrepancies between the model's predictions and real-world salaries
- Outliers: The presence of outliers, such as superstars with unusually high salaries or injured players with altered performance, could skew the analysis and impact the general applicability of the model.
- Multicollinearity Among Features: Even though efforts were made to minimize multicollinearity, some variables might still be intercorrelated, impacting the interpretation of individual feature importances

Future Steps

- Incorporate Advanced Metrics: Including advanced metrics like Statcast could enrich the analysis and provide more detailed insights into player performance. Utilizing such data could help in identifying subtle patterns and relationships that were not captured using traditional statistics.
- Analyze Pitchers: Extending the analysis to pitchers could offer a more complete picture of MLB salary dynamics.
- Investigate Market Influences: A study of market-related influences on player salaries could provide a deeper understanding of non-performance-related factors.
- Time-Series Analysis: Developing a time-series analysis could enable a more nuanced view of salary trends and player performance over time.
- Robust Handling of Outliers: Implementing robust techniques to handle outliers, such as superstars and injured players, could make the model more applicable to a broader range of players.