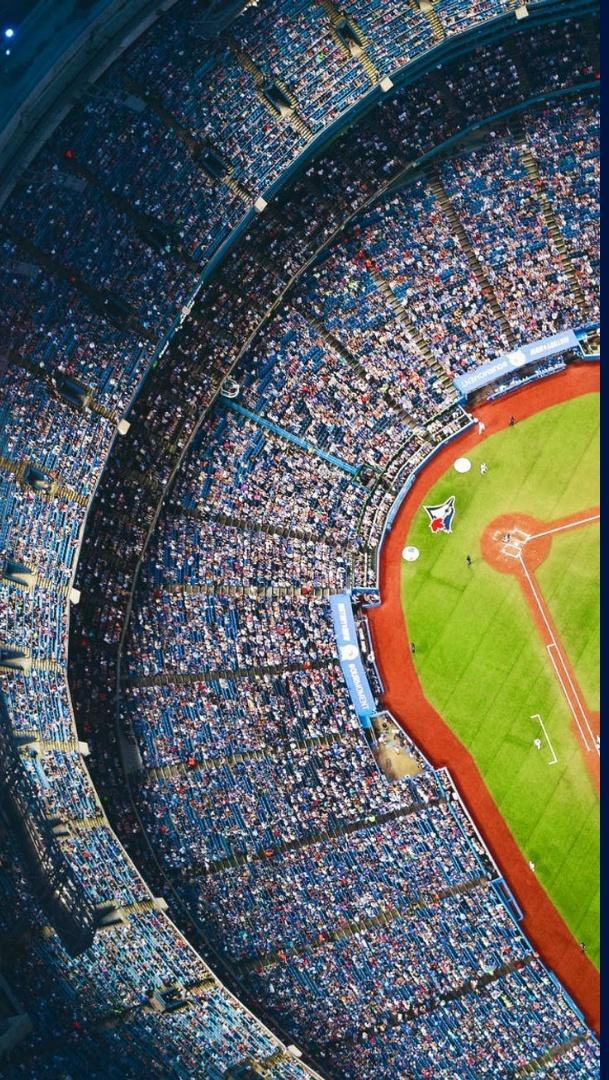


Predicting MLB Player Salaries: A Batting Performance Analysis

An In-Depth Exploratory Analysis of Factors Influencing Player Compensation

By: Hector Guerrero, Data Scientist

Introduction

The objective is to understand the monetary value of MLB players based on batting metrics and develop a predictive model to estimate salaries and uncover key factors driving player compensation in professional baseball

Value Add

- PLAYER CONTRACT NEGOTIATIONS
- BUDGET PLANNING
- FAIRNESS AND TRANSPARENCY
- BUSINESS STRATEGY

The Data

DATA SOURCES

- **GITHUB - BASEBALLDATABANK**
- **BASEBALL REFERENCE**

DATA SETS

- **WAR DATASET**

  121,375 rows and 48 columns.

  Main features: WAR, PA (Plate Appearances), Pitcher (Y/N), Age.

  All players from 1871 to 2022 season.

- **BATTERS DATASET**

  112,184 rows and 22 columns.

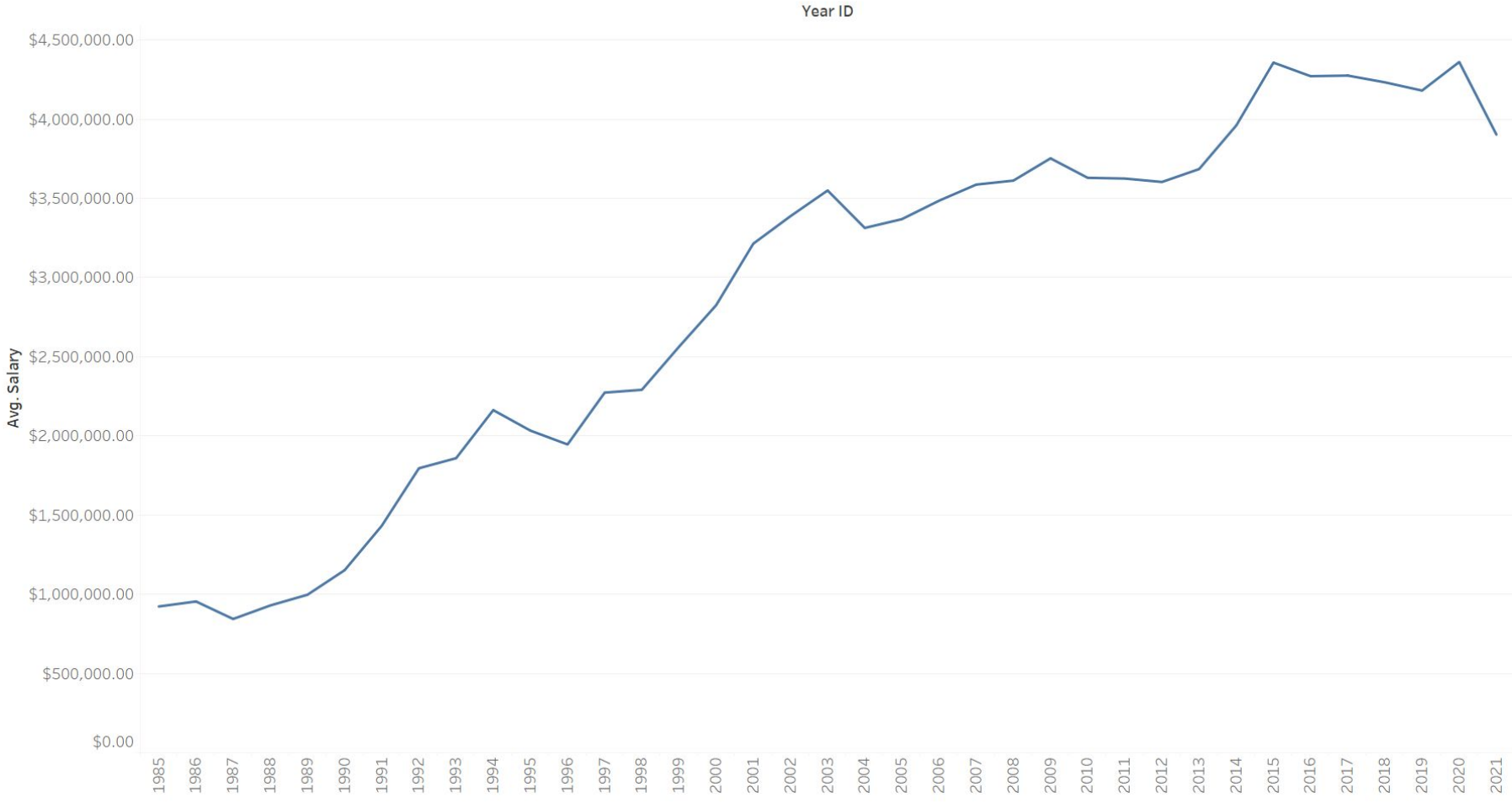
  Main Features: R (Runs), H (Hits), HR (Home Runs), RBI (Runs Batted In)

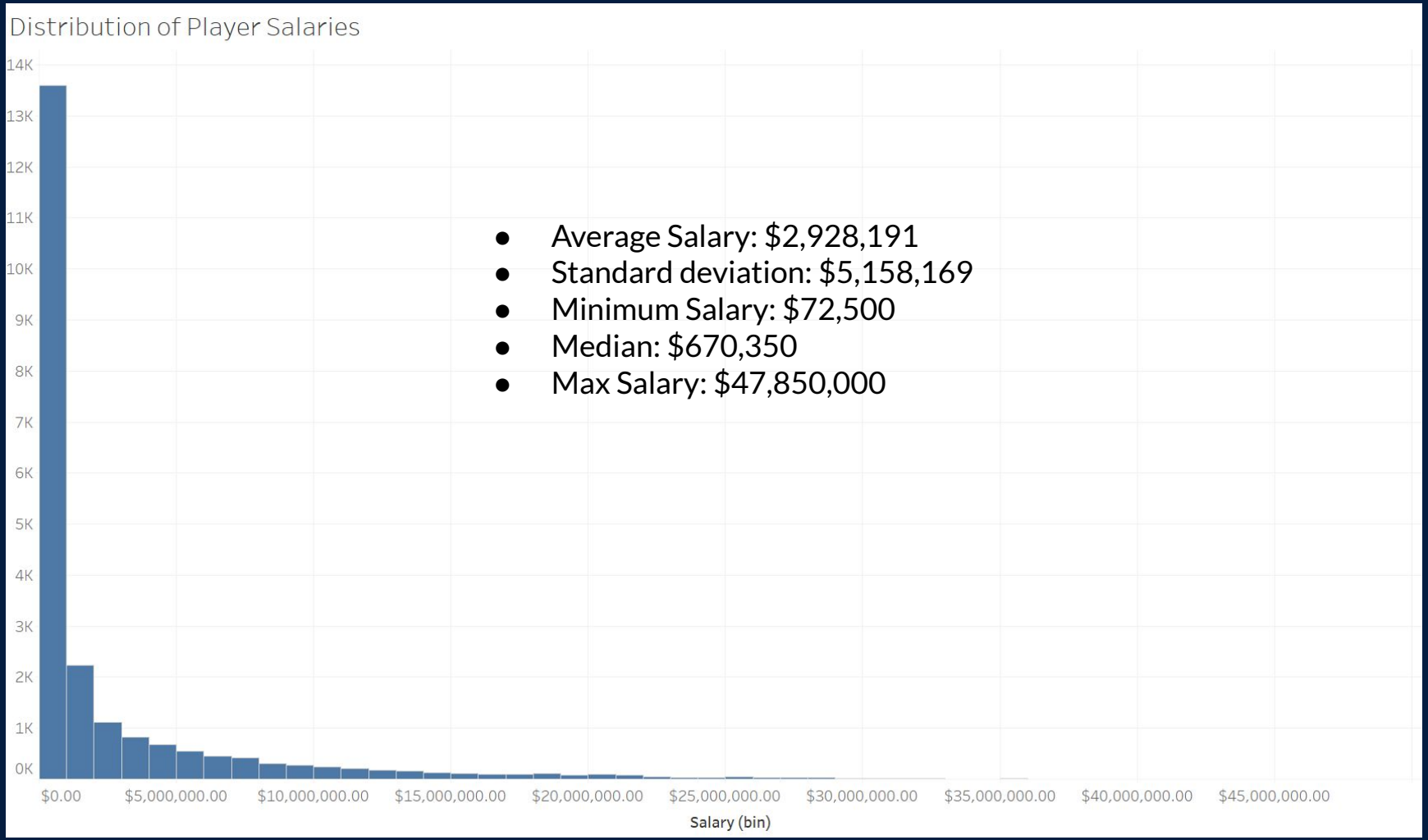
- **SALARY DATASET**

  46,450 rows and 19 columns.

  Main Features: Salary

Average Salary through the years (1985-2021)





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

The Process

DATA COLLECTION

- Baseball Reference, GitHub Baseballdatabank
- WAR, Batters, Salary

DATA CLEANING AND PREPROCESSING

- Merging datasets
- Creating Key Features
- Adjusting Salary for Inflation

EDA AND FEATURE SELECTION

- Key Insights
- Relationships between Features and Target
- Correlations and Multicollinearity
- Selecting Most Important Features for Modeling

MODELING

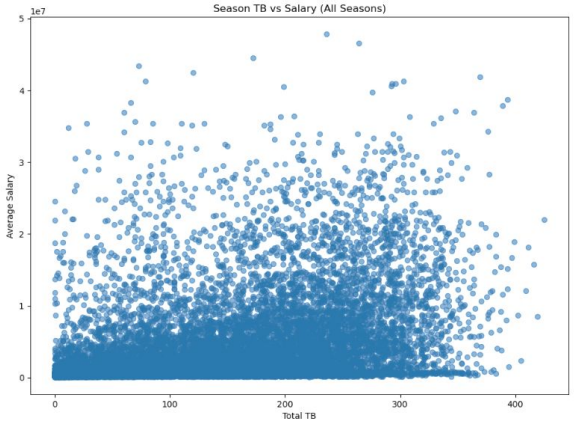
- Train-Test Split (80%-20%)
- Scaling and OneHotEncode
- Tune Hyperparameters with GridSearchCV

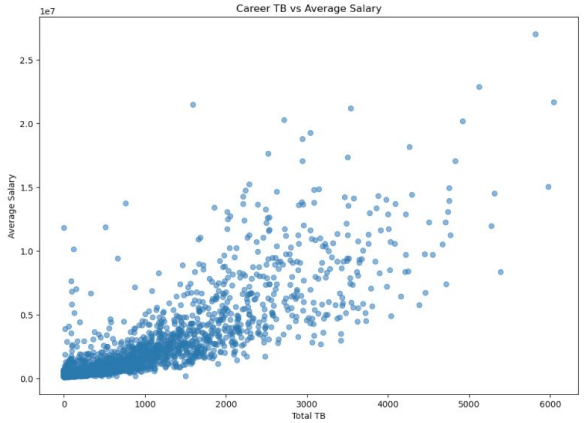
FINAL MODEL AND TESTING

- Choose Best Performing Model
- Most Important Features (Feature Weight)
- Test Model and Evaluate Performance
- Limitations and Future Steps

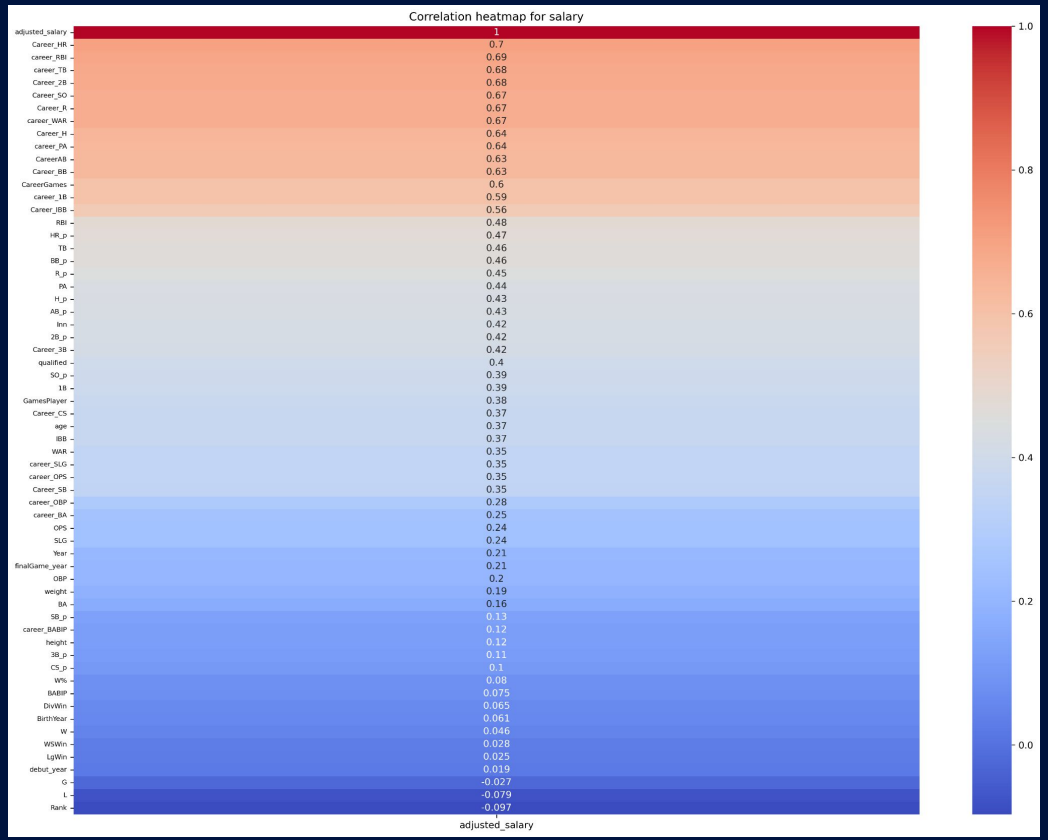
Key Insights

Season Stats vs Career Stats





Key Insights



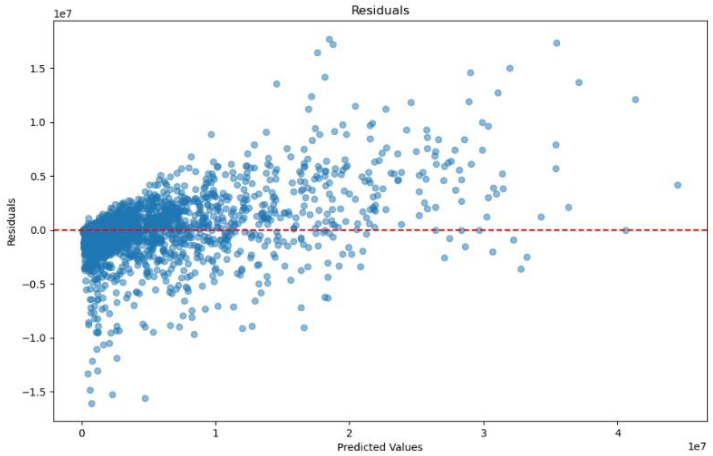
Correlation and Feature Selection (SelectKBest and Domain Knowledge)

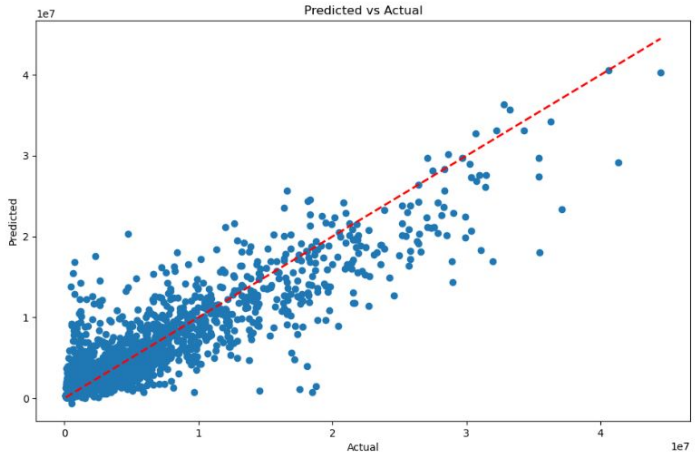
- 'Year', 'age', 'career_length', 'bats', 'POS', 'teamName', 'career_PA', 'Career_R', 'Career_H', 'Career_HR', 'career_RBI', 'Career_BB', 'Career_SO', 'Career_IBB', 'career_TB', 'career_WAR'

Models and Results

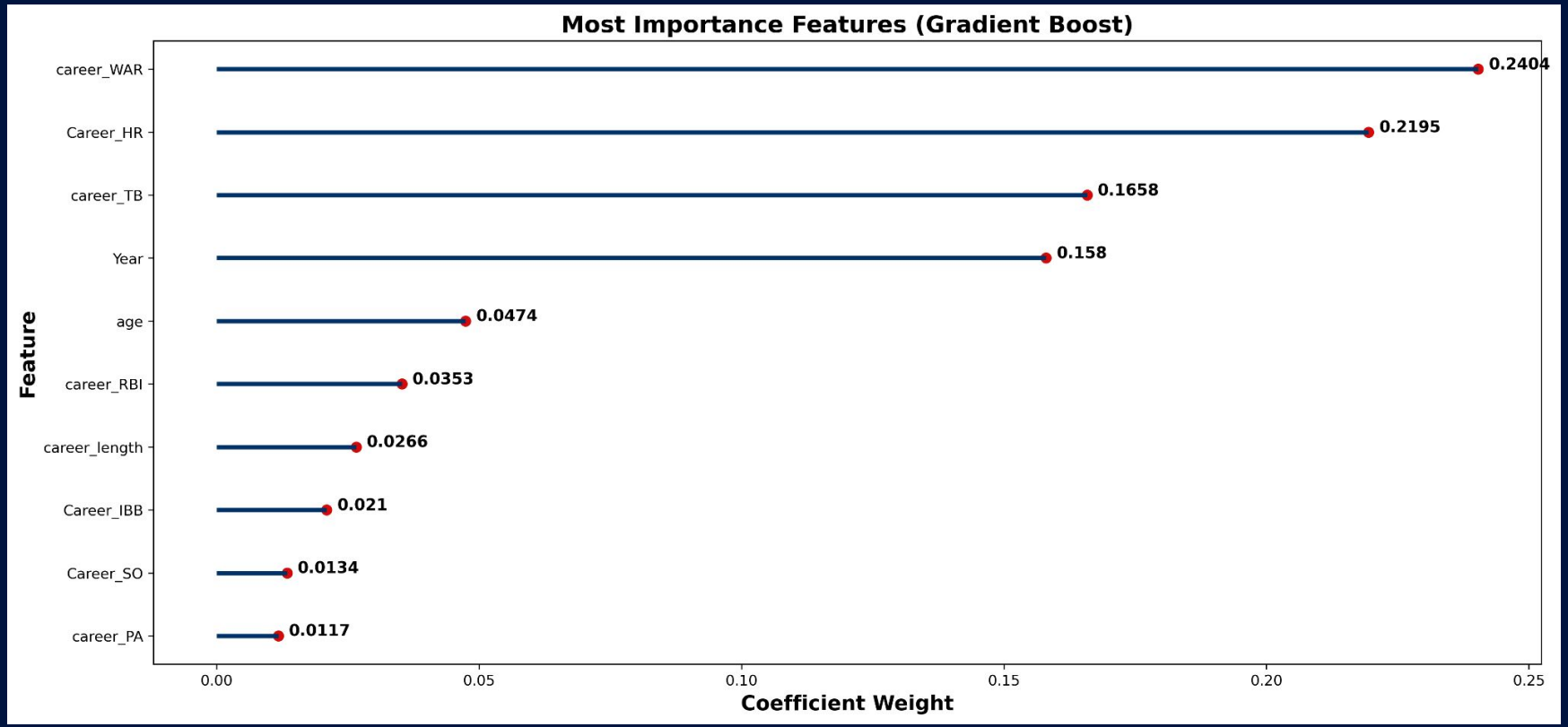
- LINEAR REGRESSION
  - R2 = 0.62
  - RMSE = ±$3,222,420.69
- DECISION TREE
  - R2 = 0.785
  - RMSE = ±$2,422,400.47
- RANDOM FORESTS
  - R2 = 0.83
  - RMSE = ±$2,180,027.92
- **GRADIENT BOOSTING REGRESSION (MVP)**
  - **R2 = 0.84**
  - **RMSE = ±$2,080,887.37**

Gradient Boosting Regressor





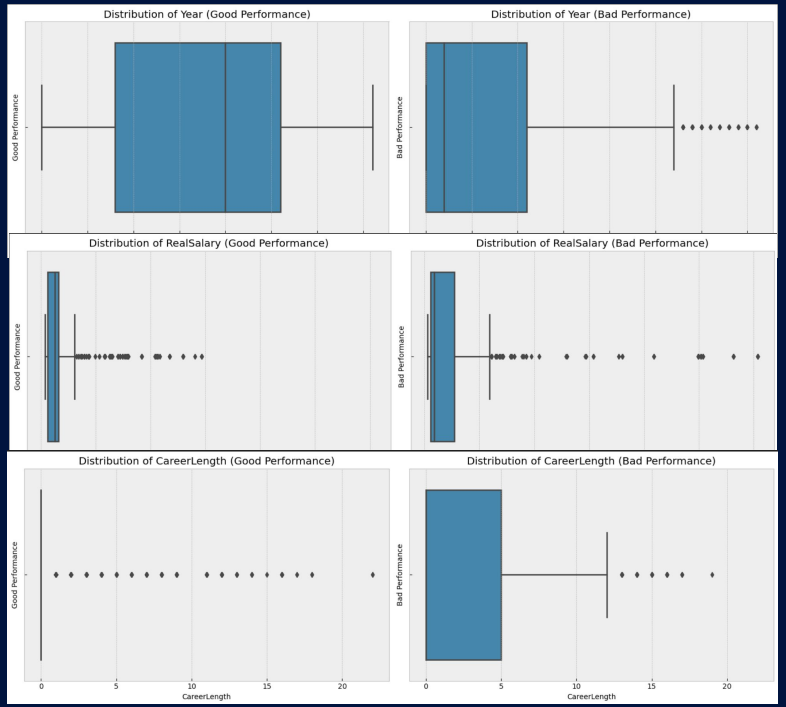
Most Important Features - Gradient Boosting Regressor



Model Testing, performance and behavior

Good Performance Threshold - 10% (Prediction ±10% from Real Salary)

- 57% of Players



- Performs better with more recent data
- More accurate for younger players with shorter careers
- Struggles with players having extensive career statistics
- Predicts lower real salaries more accurately
- Tends to underestimate players with higher Career WAR values (Super Stars)

Model Testing, performance and behavior

Testing on players from 2022 Season (Unseen Data)

- Trusting our model on players from the 2022 MLB season to determine underpaid or overpaid status

	PlayerID	Name	RealSalary	PredictedSalary	ErrorPercentage	Value
62	stottbr01	Bryson Stott	763000.0	749370	1.79	Fairly-Valued
72	biggica01	Cavan Biggio	2313525.0	1161575	49.79	Over-Valued
304	hoernni01	Nico Hoerner	763000.0	857249	12.35	Under-Valued
143	stantmi03	Giancarlo Stanton	31610000.0	29516857	6.62	Fairly-Valued
146	torregl01	Gleyber Torres	6812500.0	5516420	19.03	Over-Valued

Conclusion, Limitations and Future Steps

The prediction models are more proficient with players in the early stages of their careers or approaching their first years of arbitration. They tend to struggle with predicting the salaries of superstar outliers and older players who have already secured large contracts. The models demonstrate increased effectiveness for those playing in the initial phases of their careers.

Limitations

- Non-Statistical Factors Ignored: Marketability, leadership, fan appeal not considered.
- Salary Structure Limitations: Doesn't include contracts, arbitration, bonuses, etc.
- Outliers Impact: Superstars or injured players can skew results.

Future Steps

- Incorporate Advanced Metrics: For deeper insights.
- Analyze Pitchers: For a complete MLB salary picture.
- Investigate Market Influences: Understanding non-performance-related factors.
- Time-Series Analysis: Refined view of trends.
- Robust Handling of Outliers: Wider applicability.

Questions?

Thank You!

Email: g.guerreroja@gmail.com
GitHub: https://github.com/MemoGJ/Hector_G_Capstone
LinkedIn: https://www.linkedin.com/in/hector-guillermo-guerrero-jauregui/