**Zev Kaplan**

**Coral Harrison**

**Darin Young**

**Introduction**

**Problem Statement and Interested Parties**

In this report we want to explore the relationship between performance metrics and salary in Major League Baseball (MLB). Our dataset contains detailed player statistics and salary information, so our goal is to identify which performance factors most strongly influence salary outcomes. We will analyze the data to uncover trends and other causal factors that will give us insight into how player performance correlates to financial compensation. Our report will be valuable to teams, agents, and the players in understanding the factors that affect financial dynamics in the MLB.

**Importance**

Understanding the factors that impact a player’s salary in Major League Baseball (MLB) is pivotable for multiple stakeholders. For teams, assists in making decisions for efficient allocation of resources and helps ensure competitive balance. For players and their agents, it is critical to know the most valuable performance metrics to assist in contract negotiations and additionally in professional growth. These dynamics will clarify how sports organizations value talent and illuminate the decision-making process. This study is compelling as data-driven decisions are increasingly shaping professional sports.

**Source**

The dataset used for this analysis was sourced from Kaggle, a web platform that is a community for data scientists and machine learning professionals. Here is the link to our dataset: ([The Most Cost-Effective MLB Hitters).](https://www.kaggle.com/datasets/thedevastator/uncovering-the-most-cost-effective-mlb-hitters-o) Kaggle provides access to a wide range of publicly available datasets, including sports statistics, making it an invaluable resource. The MLB dataset utilized here includes comprehensive player statistics and salary information, enabling a thorough investigation of the association between key performance indicators and financial compensation.

**Usefulness**

This dataset is ideal for solving the problem because it includes detailed player performance metrics along with their salaries. Key statistics like home runs, runs batted in, and stolen bases allow us to analyze how various aspects of a player's performance influence their pay. With data spanning multiple seasons and teams, the dataset provides a broad and reliable foundation for uncovering patterns and identifying which factors have the biggest impact on MLB salaries.

**Proposal**

**Models**

We will attempt a combination of modeling approaches.

1. Linear regression will serve as a starting point, offering a simple and interpretable framework to quantify how individual performance factors, such as home runs or RBIs, affect salary.
2. Next, we will use a stepwise regression to refine the analysis by systematically selecting the most significant predictors, ensuring the model remains efficient and focused.
3. For more complex relationships, decision trees and random forests may be used to uncover non-linear patterns and interactions between variables, with random forests providing variable importance scores to highlight the most influential metrics.
4. Finally, we may use forecasting models by predicting future salaries based on historical trends and projected performance metrics.

Together, these methods provide a powerful toolkit for understanding and predicting the impact on the salaries of MLB players.

**Evaluation**

Due to the nature of our research, we will be interested in ensuring that the model's performance is as good as it can be. In addition to building our models, we also plan to rigorously assess and evaluate the performance of our model(s). To ensure we have a good balance between test and training data, we will use the Scikit Learn library’s train\_test\_split() function with 80% of the data allocated for training and 20% allocated for testing. Metrics that will be important are accuracy of the regression models (accuracy of testing), error rates (MSE, RMSE, MEA, MAPE), and variance coverage (R2, COD). To ensure the best results, hyperparameter tuning might be required and the use of scalars to normalize the data.

**Lesson Learned**

We hope to reveal all the metrics that most significantly influence a player’s salary and understand the impact on decisions that go into establishing capital budgeting decisions. Linear and stepwise regression will help us pinpoint the most relevant predictors and their importance; decision tress and random forests will drive our exploration into non-linear correlations and interactions amongst these metrics to understand the combined effects. Forecasting models will help in predicting future salary trends from historical patterns and anticipated performance.

**Concerns**

Some of the normal risks of this type of report include oversimplifying the factors influencing salaries because there are many non-performance-related issues that are factors. Things like market conditions, team budgets, and player popularity. Ethical concerns in this type of analysis may find that potential misuse of findings of team dynamics are less quantifiable and perpetuate biases in salary decisions. Also, there is a concern for breach of privacy as sensitive data about players is included in the analysis. We certainly plan to manage this data responsibly to ensure that our findings are objectively presented and promote fair interpretation and evaluation.

**Contingency**

In case our models discover that salaries do not correlate to actual player metrics, we plan to use an additional dataset that contains the popularity of current baseball players, and with a script, pull the lifetime metrics for those players. With the added datasets, we anticipate the discovery of additional variables that can correlate with player salaries.

**Additional Information**

We will also take model validation and testing into account as it is essential to ensure that the findings are solidly built, broad based, and not overly influenced by any peculiarities in the dataset.