BAX 452 Machine Learning

Instructor:
Jörn Boenhke

Machine Learning Final Project

Unveiling The Statistics that Drive MLB Salaries with Machine Learning

A Collaborative Project by:
Anurag Vedagiri
Kuan-Yu (Leo) Liao
Rishikesan (Rishi) Ravichandran

University of California, Davis
Master of Science in Business Analytics

Table of Contents

Executive Summary

As more and more athletes signed large contracts with sports organizations, we are curious how the teams set budgets when signing free agents, and what are some major components that impact the contract. This report analyzes the impact of various baseball metrics on player salaries in Major League Baseball. As a team, we utilized machine learning models to identify the key performance indicators that most significantly influence player compensation. The analysis employed a comprehensive dataset encompassing essential baseball metrics such as Earned Run Average, Batting Average, Runs Scored, Hits, Home Runs, Stolen Bases, and Strikeouts, among others.

Some limitations within the data include:

- **Player Versatility:** We addressed players potentially switching positions within a year by combining their overall statistics. This, however, limits our ability to capture the unique salary influences of each position.
- **Team Impact:** Players who change teams during a season can have their performance metrics impacted. We combined their seasonal data which overlooks the effect of different team environments.
- **Award Representation:** Multiple awards were transformed into dummy variables and this approach may overlook variations in award prestige that could affect salaries.

Leveraging an array of machine learning models, including Lasso regression, Decision Trees, and Random Forests, our analysis successfully unveils hidden patterns and relationships between baseball statistics and player salaries It delves into the methodology used for data collection, exploration, and model development, ultimately revealing key findings and insights. Major League Baseball teams can gain a significant advantage in player acquisition, contract negotiations, and roster management by understanding which statistics most heavily influence salaries. This information can be used by passionate baseball fans as well to make more informed decisions when supporting their favorite teams and sports betting companies to leverage these insights to refine their analytical models and potentially offer more accurate odds.

Background

Major League Baseball has long relied on statistics to evaluate player performance and inform decision-making. However, the explosion of data available in recent years, coupled with the rise of machine learning, has opened exciting new possibilities for analyzing player value and predicting performance. Traditionally, baseball analytics have focused on simple metrics like batting average, on-base percentage, and earned run average. While these measures offer valuable insights, they only tell part of the story. Today, advanced metrics delve deeper, capturing a more nuanced picture of player performance across various aspects of the game. This report leverages machine learning models to go beyond traditional metrics and uncover the hidden patterns within this vast dataset of baseball statistics. By analyzing these deeper relationships, we aim to identify the key performance indicators that most significantly influence player salaries in MLB. This information empowers teams to make data-driven decisions regarding player acquisition, contract negotiations, and roster management, ultimately optimizing their competitive advantage.

Problem Statement

The problem at hand is the obscure relationship between a player's performance metrics and their salary, a clarity that is essential for MLB teams in terms of strategic player selection, conducting fair salary negotiations, and creating balanced rosters. This lack of transparency hinders data-driven decision-making. This project confronts this challenge by applying various machine learning models – specifically Lasso regression, Decision Trees, and Random Forests – to translate the KPIs that influence MLB salaries into a tool of value for MLB teams, the everyday MLB fan, and sports betting companies.

Introduction

Our dataset originates from the Lahman Baseball Database, which includes Major League Baseball statistics from 1871 to 2022. The database consists of 28 different CSV files, each containing a table of data in different aspects of the game. From the 28, we chose the most relevant 7 files to build our model:

- **People:** stores basic information of all the players who ever played in the MLB.
- **Salaries:** stores salary for each player each season, up to 2016
- **AwardsPlayers:** award winner each season
- **Batting:** stores batting stats for each player
- **Fielding:** stores fielding stats for each player
- **Pitching:** stores pitching stats for each player
- **Teams:** stores basic information for each team

Data Handling Before Merging

After importing these 7 CSV files into Python as data frames, we conducted a preliminary handling of the data, which included removing unnecessary columns for each file and filtering data. Since our objective is to predict salaries and its available data only goes up to 2016, and we want the data from a span of 5 seasons to train and test our model, our target of data is from 2012 to 2016.

At this stage, we also use one-hot encoding to generate dummy variables for the AwardsPlayers data frame. This process creates a separate column for each of the awards, ensuring that each player per year remains in a single row even if they receive multiple awards.

Next, to prevent duplicate rows from occurring from players featuring in different fielding positions or playing for multiple teams within a single season, we combined each player’s statistics. This includes summing most of the columns, averaging some of the variables and manually calculating ERA for pitchers. Then, we merged the 7 data frames into one that we wanted to proceed with.

Data Handling After Merging

After merging the data, we decided to add some columns that might help predict the salary but were not explicitly included in the original data. At this stage, we added:

- **Age**
- **Years of experience in the league**
- **Pitched for 9+ innings during a single season**

Age is one of the most common factors when predicting player salary. Athletes typically peak in their late 20s, after which performance tends to decline. This decline, along with factors like injury probability, impacts a player's market value. Therefore, we want to add this value to the dataset.

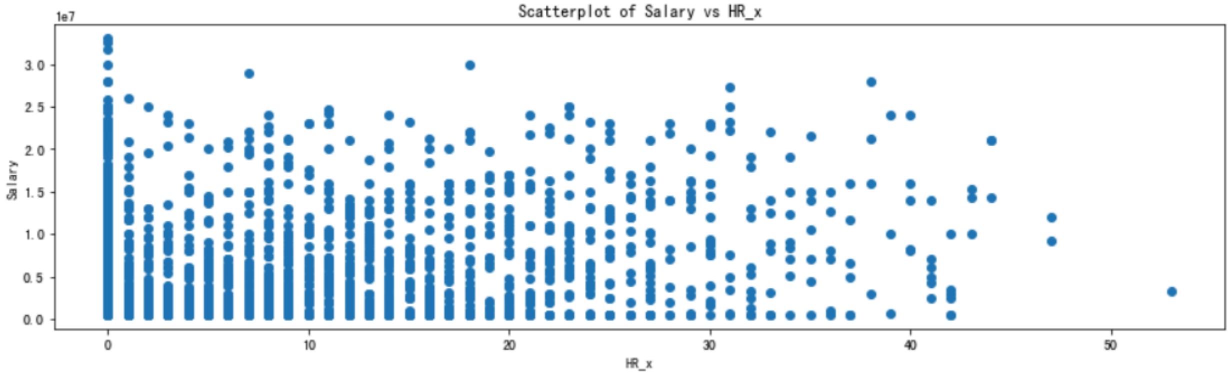
Experience also plays a significant role in salary predictions. MLB players with more experience often possessed a deeper understanding of the game and demonstrated their ability to survive at the highest level. Baseball is ultimately a mind game, so mentality is as important as physicality. Thus, while talent is crucial, charisma and expertise can greatly influence a player's value in the league.

We also wanted to differentiate between pitchers and position players as this can potentially impact a player’s value. However, some position players occasionally pitch in games, so relying only on the presence of pitching statistics isn't sufficient. Therefore, following research and leveraging our domain expertise, we set a threshold of 9 innings pitched to distinguish pitchers and position players.

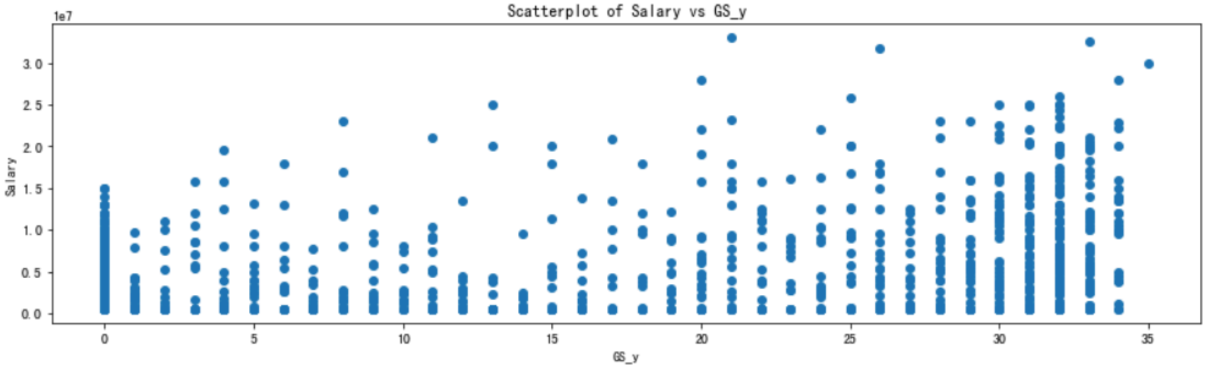
Exploratory Data Analysis

Following data processing and merging the data frames, we proceeded with exploratory data analysis. Our main approach involved the use of data visualization to see if there are explicit patterns within the data. We plotted histograms to visualize the distribution of predictor variables' counts. We also utilized scatterplots to investigate potential correlations between our predictor variables and salary.

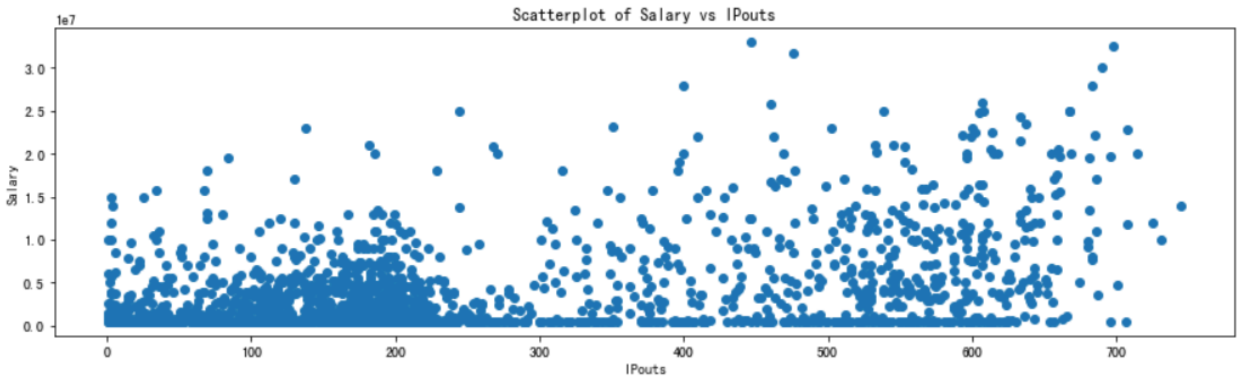
While we do not see much pattern from the histogram, we can find some interesting scatterplots.



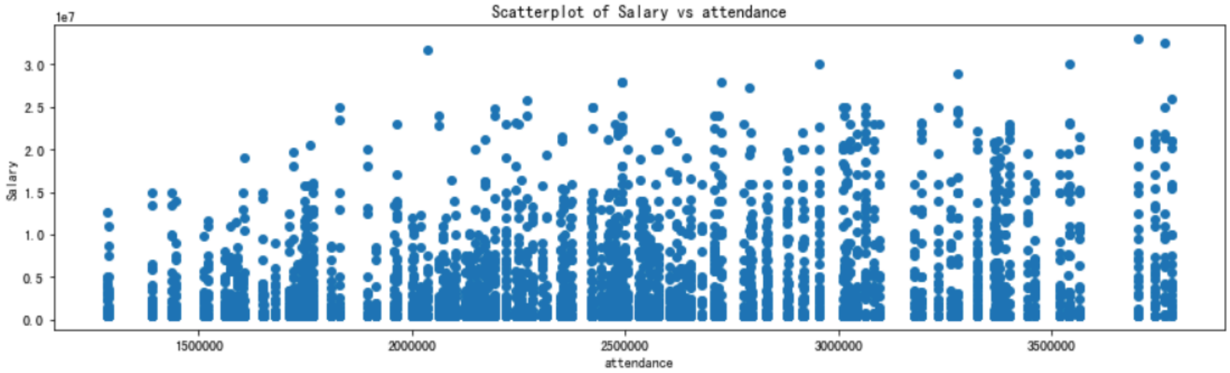
As hitters, more home runs do not translate to higher pay.



Starting pitchers tend to have higher salaries among pitchers.



As pitchers, more innings pitched is correlated to salary.



Attendance of each team seems to be correlated with salary.

After visualizing, we utilized both one-hot and label encoding to deal with categorical variables. Subsequently, we handle missing values by filling in with zeros. At this stage, we have a finalized dataset ready for model training and testing purposes.

Approach

In this phase, we implement 3 methods learned during the quarter to predict player salaries. We defined X variables as all columns in the data frame except for salary, while the Y variable represents the salary. In all approaches, we split the dataset into training and testing sets with an 80:20 ratio, and we further scaled the data before training lasso and random forest models.

Lasso Regression

In the Lasso approach, our first task was to determine the optimal alpha parameter, which controls the penalty magnitude. We use a 5-fold cross-validation to identify the best alpha. Then, using this optimal alpha and the scaled data, we trained the Lasso regression model. Finally, we use the X_test data and the trained model to make predictions.

Decision Tree

We use the DecisionTreeRegressor function in the scikit-learn Python package to build a decision tree, and we manually set the maximum depth of the tree to 3 levels. Unlike the previous approach, we did not scale our data here. Instead, we trained the decision tree regressor model using X and Y training data and made predictions using X testing data.

Random Forest

We utilize the RandomForestRegressor function in the scikit-learn Python package to build a random forest model. Similar to the first approach, we fitted the random forest regressor model with scaled X and Y training data and then made predictions using scaled X testing data.

Results

Lasso Regression

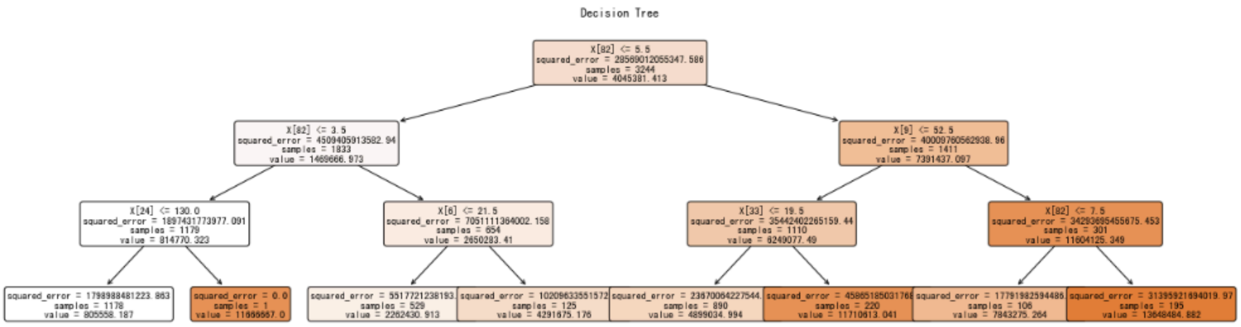
- **Best alpha:** 1.0
- **MSE:** 12895104459223.4

- **RMSE:** 3590975.4189110515
- **R-Squared:** 0.49388791773995677

The optimal alpha value was determined to be 1.0 via Cross-Validation. The Mean Squared Error and Root Mean Squared Error are considerably high, which might indicate the presence of significant noise within the dataset or that our model's simplicity does not capture the complexities of the data. The R-squared value suggests that approximately 49.4% of the variance in MLB player salaries is explained by the model, which is a moderate fit.

Decision Tree

- **MSE:** 14088761023307.635
- **RMSE:** 3753499.8365935273
- **R-Squared:** 0.44703882000194406



- **X[82]:** years_in_league
- **X[9]:** RBI
- **X[24]:** DP
- **X[6]:** 2B

- **X[33]:** GS_y

The Decision Tree model, with a controlled depth, displayed an RMSE that is comparable to the Lasso model but with a slightly lower R-squared value, indicating a similar level of prediction accuracy but with potentially higher variance in the residuals.

Random Forest

- **MSE:** 9534609658387.262
- **RMSE:** 3087816.325234916
- **R-Squared:** 0.6257819265441054

The Random Forest regressor provided an improvement in the MSE and RMSE, indicating a better fit to the data compared to the previous models. The R-squared value improved to approximately 62.6%, suggesting that this model has a better capability in explaining the variations in player salaries.

Recommendations

1. An in-depth analysis of the feature importance derived from the Random Forest model could provide insights into which variables are the most influential when it comes to determining MLB salaries.
2. Consider expanding the dataset to include more recent years or additional metrics that could have an impact on player salaries, such as social media presence, fan base, and marketability factors.
3. Future work should focus on refining the models, perhaps through hyperparameter tuning, feature engineering, or trying different machine learning techniques that may capture the nonlinear relationships within the data more effectively.

Conclusion

This project investigated the key statistical factors influencing Major League Baseball (MLB) player salaries. We employed machine learning models, including Lasso Regression, Decision Trees, and Random Forests, to analyze a comprehensive dataset encompassing the 2012-2016 seasons from the Lahman Baseball Database. The analysis focused on how various metrics, such as batting statistics, fielding numbers, pitching performance, player age, experience, and award recognition, impact player compensation.

The Random Forest model emerged as the most effective approach, achieving an R-squared value of 0.626, indicating it could explain 62.6% of the variance in salaries based on predictor variables like years of experience in the league, runs batted in (RBI), games started as a pitcher, and innings pitched. This suggests a strong correlation between these specific metrics and player salaries. However, the Lasso and Decision Tree models exhibited lower R-squared values (0.494 and 0.447, respectively), highlighting potential hidden complexities in the salary determination process that these models did not fully capture. The Decision Tree identified years in the league runs batted in, double plays, and doubles hit as the top splitting predictors, offering additional insights into player valuation.

While these findings are promising, incorporating data beyond on-field performance, such as fan popularity metrics, social media presence, and marketability factors, may enhance prediction accuracy in future iterations. MLB teams can leverage this knowledge for optimized roster management, data-driven contract decisions, and acquiring players with the best value proposition. Additionally, fans gain the ability to assess if salaries align with production, while sports betting companies can refine over/under wager odds on impactful seasonal stats. By continuously refining these machine learning models and integrating advanced metrics with traditional statistics, we can deepen the understanding of the sophisticated relationship between performance and earnings in baseball and ultimately enhance decision-making across the sport through data.