# Forecasting Major League Baseball Statistics with Machine Learning

**Authors: Nick Babcock, Amui Gayle, Kunal Chhabria**
College of Computing and Informatics, Drexel University

## Abstract

This paper presents a machine learning approach to predicting final season-ending statistics for Major League Baseball players. The increasing use of analytics in baseball has led to a greater focus on statistical analysis, and our project aims to leverage this trend by building accurate prediction models. We use recorded statistics and performance data from previous seasons to train our models, which are then used to forecast the statistics for the 2021 season. By comparing our predictions against the actual statistics for the 2021 season, we are able to evaluate the performance and effectiveness of our models. Our findings demonstrate the potential of machine learning techniques for improving decision-making in baseball and enhancing teams' competitive advantage through data-driven strategies.

## 1 Introduction

The game of baseball is a statistics driven sport and it is only becoming more analytical as our technology continues to advance and allow us to build better prediction models. There are a lot of different decisions that go into managing the game of baseball, and using analytics to make these decisions has completely changed the game [1]. Analytics not only help teams gain competitive advantages and develop strategies against opposing teams and their players, but it is also used to place a dollar value on a player by forecasting their future performance when deciding whether to sign a player to long term contract or not. Given all the statistics that are created in each and every game that's played, there is a lot of data for teams to use at their disposal.

A baseball season consists of 162 games which is about seven months long, and each player's statistics are updated every time they play in a game. In this project, we have built two prediction models using Linear Regession and XGBoost to predict a player's 2021 final season ending statistics of certain categories. This was done by evaluating each player's performance and recorded statistics from previous seasons. These statistics were used to train our models and ultimately predict the outcome of each players' 2021 season statistics. We then utilized the developed models to forecast the 2021 season statistics and compared them with the actual values, allowing us to thoroughly evaluate and validate the efficacy of our models.

## 2 Dataset

Our dataset was obtained from Baseball Savant [2], a comprehensive resource that provides detailed statistical data on every pitch thrown in Major League Baseball games. This website captures high-resolution images and video of each pitch, along with a range of other performance metrics such as pitch speed, spin rate, and launch angle. However, our dataset was customized with their search tool to include the recommended batting statistics that are used to evaluate a player's batting performance. By utilizing this rich dataset, our project leverages the latest advances in machine learning technology to develop accurate and effective prediction models for Major League Baseball statistics. `https://baseballsavant.mlb.com/statcast_search`

### 2.1 Data Dictionary

Expanding on the importance of having a data dictionary, it is crucial to ensure that all individuals analyzing the dataset have an understanding of the data being used. This allows for more accurate and consistent analysis, regardless of one's familiarity with the specific sport or statistic. By providing a comprehensive data dictionary, we aim to make our dataset more accessible and user-friendly to a wider audience. The data dictionary located in Table 1 below contains the titles of each player performance statistic, or the dataset features, paired with a short description of their meanings.

Table 1: Data dictionary for MLB statistics

| Feature | Description |
|---|---|
| Season | The year of the season the accompanied statistics are from |
| Name | Name of each individual player |
| Team | Name of the team a player is on |
| Age | Age of the player in years |
| PA | Number of player's plate appearances |
| AB | Number of player's at bats |
| H | Number of player's hits |
| HR | Number of player's home runs |
| R | Number of player's runs scored |
| RBI | Number of player's runs batted |
| SB | Number of player's stolen bases |
| BB% | Percentage of player's at bats resulting in a base-on-balls |
| K% | Percentage of player's at bats resulting in a strike out |
| ISO | Measure of the raw power of a player |
| BABIP | Player's batting average on balls in play |
| AVG | Player's batting average |
| OBP | Player's on base percentage |
| SLG | Player's slugging percentage |
| LD% | Player's percentage of line drives per ball where contact was made |
| GB% | Player's percentage of ground balls per ball where contact was made |
| FB% | Player's percentage of fly balls per ball where contact was made |
| IFFB% | Player's percentage of infield fly balls per ball where contact was made |
| HR/FB | Player's homerun to flyball rate |
| O-Swing% | Player's percentage of swings at pitches outside the strike zone |
| Z-Swing% | Player's percentage of swings at pitches inside the strike zone |
| Swing% | Player's percentage of swings per pitch |
| O-Contact% | Player's percentage of swings that made contact on pitches outside the strike zone |
| Z-Contact% | Player's percentage of swings that made contact on pitches inside the strike zone |
| Contact% | Player's percentage of swings that made contact with the ball |

## 2.2 Detailed Explanation of the Data

The following detailed description of the data provides a more comprehensive overview of the dataset, supplementing the information already presented in the data dictionary. These definitions were sourced from MLB's website glossary of standard statistics [3].

- **PA** - Plate appearances / when a player completes their batting turn, regardless of the result.
- **AB** - Any plate appearance that results in a hit, error, fielder's choice, or a non-sacrifice out
- **H** - A hit occurs when a batter strikes the baseball into fair territory and reaches base without doing so via an error or a fielder's choice.
- **HR** - A home run occurs when a batter hits a fair ball and scores on the play without being put out or without the benefit of an error.
- **R** - A player is awarded a run if he crosses the plate to score his team a run.
- **RBI** - A batter is credited with an RBI in most cases where the result of his plate appearance is a run being scored.
- **SB** - A stolen base occurs when a baserunner advances by taking a base to which he isn't entitled.
- **BB** - A walk (or base on balls) occurs when a pitcher throws four pitches out of the strike zone, none of which are swung at by the hitter.
- **K** - Percentage of player's at bats resulting in a strike out
- **ISO** - ISO measures the raw power of a hitter by taking only extra-base hits – and the type of extra-base hit – into account.
- **BABIP** - BABIP measures a player's batting average exclusively on balls hit into the field of play, removing outcomes not affected by the opposing defense (namely home runs and strikeouts).
- **AVG** - Batting average is determined by dividing a player's hits by his total at-bats for a number between zero (shown as .000) and one (1.000).
- **OBP** - OBP refers to how frequently a batter reaches base per plate appearance.
- **SLG** - Slugging percentage represents the total number of bases a player records per at-bat.
- **LD%** - Line-drive rate represents the percentage of balls hit into the field of play that are characterized as line drives.
- **GB%** - Ground-ball rate represents the percentage of balls hit into the field of play that are characterized as ground balls.
- **FB%** - Fly-ball rate represents the percentage of balls hit into the field of play that are characterized as fly balls.
- **IFFB%** - Infield-fly-ball rate represents the percentage of balls hit into the infield that are characterized as fly balls.
- **HR/FB** - Home-run-to-fly-ball rate is the rate at which home runs are hit by a batter for every fly ball he hits.
- **O-Swing%** - O-Swing% represents the rate of the number of swings at pitches outside the zone per pitches thrown outside the zone
- **Z-Swing%** - Z-Swing% represents the rate of the number of swings at pitches inside the zone per pitches thrown inside the zone
- **Swing%** - Swing% represents the player's rate of swings per pitch that is thrown
- **O-Contact%** - O-Contact% represents the rate of the number of swings at pitches

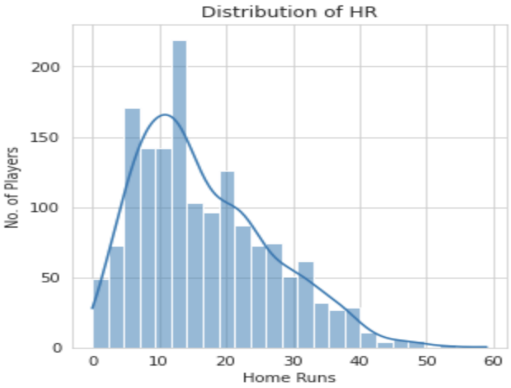

Figure 1: Distribution of Home Runs (HR).

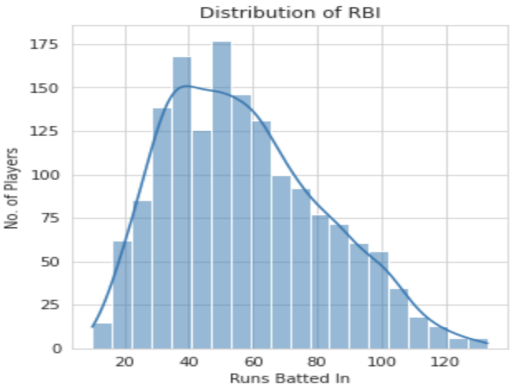

Figure 2: Distribution of Runs-Batted-In (RBI).

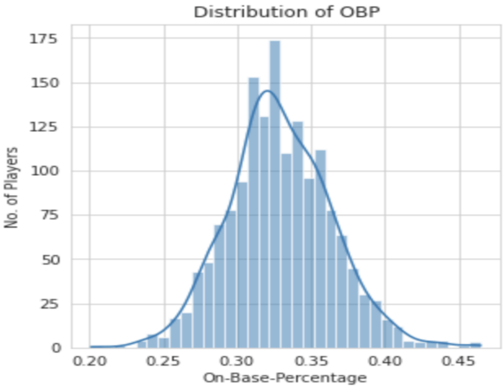

Figure 3: Distribution of On-Base-Percentage (OBP).

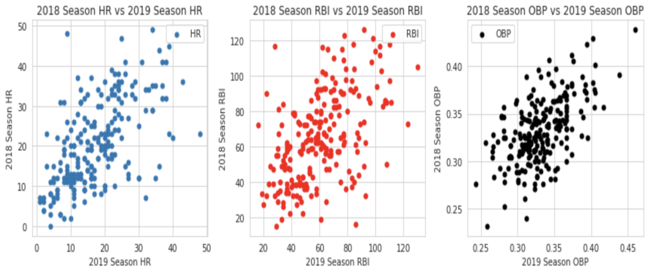

Figure 4: Player statistics from 2018 compared to 2019.

## 3 Exploratory Data Analysis

Our analysis aims to investigate the impact of a player's past performance on their year to year performance by specifically evaluating their relations to the 2021 season. To achieve this, we first looked at the distribution of player statistics across the league (**Figure 1, Figure 2, Figure 3**). The plots shows that most players have relatively low HR and RBI counts, while OBP follows a normal distribution.

To further investigate the impact of past performance on a player's year to year statistics, we analyzed the distribution of the 2018 season statistics of HR, RBI, and OBP against the distribution from the 2019 season for players who have played both years (**Figure 4**). The plot shows a consistent correlation from year to year for each player's statistic of HR, RBI, and OBP with the exception of some outliers. Home Runs (HR) showed the least consistency, however it is difficult to truly value this information as there are many reasons as to why a player may not perform the same in any given season. Although it is out of the scope of this study, some of proven factors to effect player performance are player injuries, a player losing the starting job, or a player joining a new team in a new city.

Next, we wanted to explore the distribution of player ages (**Figure 5**), which shows a relatively normal distribution but with most players being in their mid 20s. We also wanted to show the relationship between a player's age and their plate discipline (**Figure 6**). Plate discipline defines a player's ability to judge which pitches to swing at and it can be measured by statistics like O-Swing%, Z-Swing%, and O-Contact%. We found that, on average, younger players tend to have worse plate discipline than older players, as evidenced by the higher O-Swing% and lower O-Contact% values for batters in their 20s compared to those in their 30s. We can also see that players tend to perform worse in these categories around the age of 35, which happens to be around the age where most players decide to retire. It's important to note that the scope of **Figure 6** is limited to players aged between 23 to 36 years, as it is less common for a player beyond this age range to be an every day starter for their team.

## 4 Methodology

The methodology adopted in this study involved employing various pre-processing techniques on the dataset to optimize its readiness for training machine learning models. Following this, the processed data was utilized to develop machine learning models that could accurately

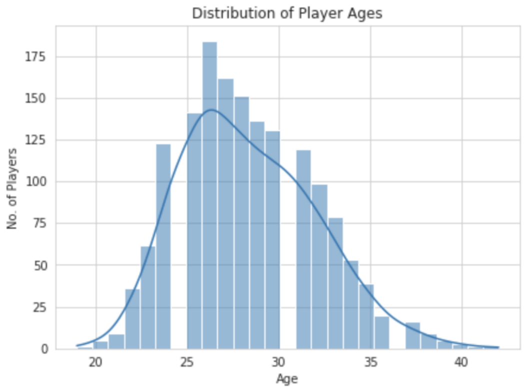

Figure 5: Distribution of player ages.

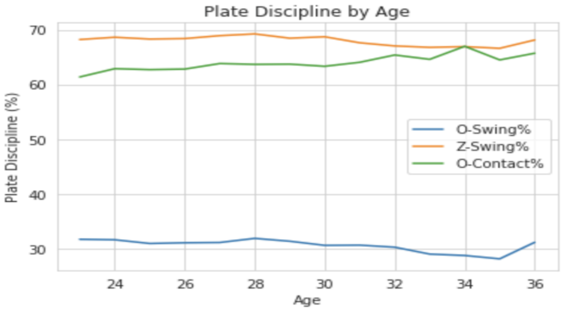

Figure 6: Average player plate discipline metrics.

predict the desired player statistic. This approach highlights the significance of data preparation in machine learning and demonstrates the effectiveness of using different regression models to predict numerical data.

## 4.1 Data Pre-processing

The dataset was fairly clean due to the realiable source it was provided from, however there were a few cleaning tasks required. Initially, all null values were eliminated from the dataset, including null columns resulting from the dataset generator [2].

The dataset contains several percentage-based features with string representations that required the removal of the % operator sign as well as a conversion to float data types in order to be used in our machine learning models.

Before training the mode, machine learning, we transformed the player statistics to be used as predictors into a standardized feature vector using VectorAssembler, and similarly transformed the target variable data into a target vector. Scaling the feature vector with StandardScaler ensured all features were in a similar scale and ultimaely improved model accuracy. Finally the data was split into random 70%-30% training and testing datasets so our models can be trained and then evaluated on unseen data to prevent overfitting. Overall, cleaning your data using these preprocessing steps can greatly improve the accuracy and robustness of your machine learning models.

## 4.2 Machine Learning Modeling

The predictions in this study leveraged three separate PySpark pipelines, including Linear Regressionand XGBoost. Our models were all trained with the training data that was first split into feature and target vectors using PySpark's VectorAssembler tool and then randomly splitting the data using a 70-30 test-train split. We evaluated our models using metrics such as RMSE, MSE, MAE, and $R^2$, to evaluate their performance. Our machine learning pipeline exemplified the potential of machine learning techniques for aiding decision-making professionals in Major League Baseball.

# 5 Results and Discussion

Linear Regression and XGBoost have proved to be sufficient models to predict the final season-ending statistics for Major League Baseball players. The Linear Regression model achieved an $R^2$ score of 0.9248 while the XGBoost model outperformed the Linear Regression model with an $R^2$ score of 0.9598. Additionally, the linear regression model had $R^2$ scores ranging from 0.8895 to 0.9732, and the XG Boost model ranging from 0.8470 to 0.9643. This shows that both models can effectively explain the variance in the dependent variables, making them both accurate and reliable in predicting the final season-ending statistics for MLB players. Of course these people are human beings and there are many factors than can disrupt these models such as a season-ending injury, which is something we would like to quantify and explore in future studies.

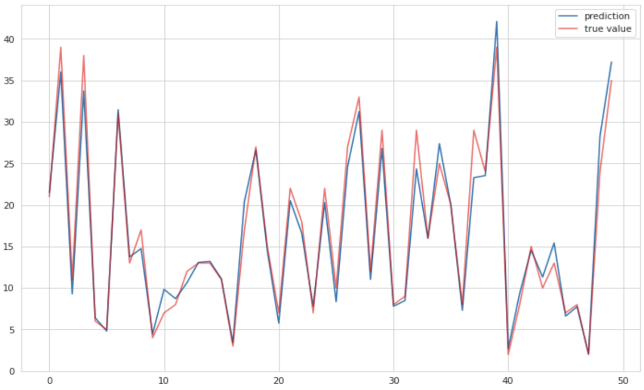

Figure 7: Home Runs (HR) were best predicted by the XGBoost model, receiving an $R^2$ score of 0.96 in comparison to its Linear Regression score of 0.95.

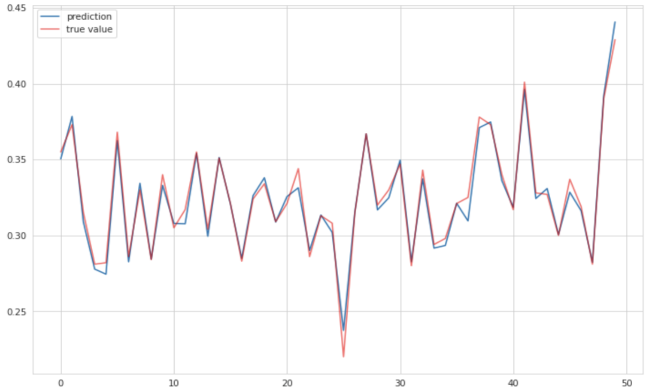

Figure 8: On-Base-Percentage was best predicted by the Linear Regression model, receiving an $R^2$ score of 0.97 in comparison to its XG Boost score of 0.93.

## 6 Conclusion and Future Work

In this study, we introduced a machine learning method for forecasting Major League Baseball players' final season-ending stats. To forecast the statistics for the 2021 season, our models used performance information from prior seasons as well as recorded statistics. We demonstrated the potential of machine learning techniques for improving teams' competitive advantage through data-driven strategies through a rigorous evaluation of our forecasts against the actual statistics for the 2021 season. Our results indicate that our models were able to achieve high accuracy in predicting the final season-ending statistics for most of the categories considered. We found that the linear regression and random forest regression models were the most effective at predicting the outcomes of the various performance metrics we examined.

In future work, we could focus on adding more features to the dataset, such as injury history and team composition to improve the accuracy of the predictions. Additionally, we could explore different machine learning models or even apply our models to other sports with similar statistics. Ultimately, the ability to forecast player performance can have significant implications for teams looking to make data-driven decisions about player contracts, trades, and game strategies.

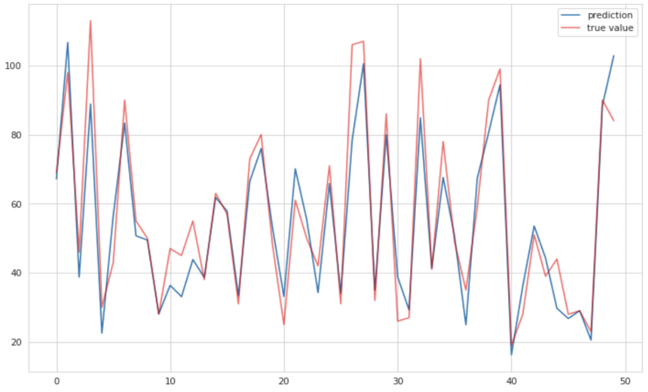

Figure 9: Runs Batted In (RBI) were best predicted by the Linear Regression model, receiving an $R^2$ score of 0.89 in comparison to its XG Boost score of 0.85.

## References

[1] Bradbury, J. C. (2021). Artificial intelligence, machine learning, and the bright future of baseball. SABR Baseball Research Journal, 50(2), 41-49. Retrieved February 13, 2023, from https://sabr.org/journal/article/artificial-intelligence-machine-learning-and-the-brig

[2] MLB Baseball Savant. (n.d.). Statcast Search. Retrieved February 1, 2023, from https://baseballsavant.mlb.com/statcast_search

[3] MLB. (n.d.). Glossary. Retrieved March 13, 2023, from https://www.mlb.com/glossary