Data science in Baseball

Kiran Ingle
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
kiran.ingle@vit.edu

Ankur Raut
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
ankur.raut20@vit.edu

Vithika Pungliya
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
vithika.pungliya20@vit.edu

Atharva Purohit
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
atharva.purohit20@vit.edu

Roshita Bhonsle
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
roshita.bhonsle20@vit.edu

Sakshi Suryawanshi
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
sakshi.suryawanshi20@vit.edu

Abstract — **Baseball is one of the richest data -driven sports, where a seemingly innumerable number of metrics are available for calculating player performance. Major League Baseball (MLB) stands for "national entertainment" focused on data-driven analysis not only the fan base and the decisions of the franchise staff but also orthopaedic and sports books. In this study we have used the Lahman Package in R which provides the tables from the 'Sean Lahman Baseball Database' as a set of R data frames. We have analysed players batting and fielding performances and based on this data we have traced their career trajectories of top 10 batters and pitchers. We have also used Multiple linear regression model to predict the wins in MLB. The paper also explores the relationship between how many runs a team score and the number of the wins they have.**

Keywords — Data Analysis, Baseball, Lahman Dataset.

I. INTRODUCTION

Since the foundation of the Society of Baseball Research in 1971, an explosion of new measures has been developed for understanding offensive and defensive contributions of players in baseball. Over the years, there has been a rise in the level of competition in professional sports leagues. More and more organizations are turning to data science to develop a competitive edge.

There has been a recent rapid growth in the development and application of R, the programming language and software environment for statistical computing and graphics. R is now widely used in academic research, education, and industry. It is constantly growing, with new versions of the core software released regularly and more than 4,000 packages available. For our study we will be mainly be focussing the Lahman Package. The package contains database of pitching, batting, fielding, salary stats and many more statistics for Major League Baseball from 1871 to 2021.

Baseball and statistics go hand-in-hand. Analysis of these statistics has become an important part of MLB and also helps teams better understand the talents that their players possess. Homeruns are among the most popular aspects of baseball as it can change the outcome of an entire game in just one hit. Analysing the most important metric of batting, i.e., home runs, does gives us valuable results. By analysing the home runs per game played, we will find out the top ten batters according to the Lahman dataset and also plot their career trajectories to find out at what time in their career their skills were at peak and when did they start to decline. Similarly, the pitcher's dataset is analysed and the top ten pitchers are found out and their career trajectory are plotted. The study also focuses on predicting wins by teams in Major League Baseball by using the Multiple Linear Regression models. The Teams table, which contains yearly statistics and standings for teams, is used as a dataset for training the model. Lastly, we found out a relation between the how many runs the team has scored and the number of wins by them. The Team's table was used for training the Linear Regression and Decision tree Regression model.
The scope of this study will be:

1. Predicting the wins by teams in Major League Baseball by using the Multiple Linear Regression model
2. Analysis of career performance of top batters and pitchers based on the average homeruns and strikeouts by the players individually
3. Exploring the relationship between how many runs a team score and the number of the wins they have.

II. LITERATURE REVIEW

In this paper, the authors attempt to predict home run hitting performance of Major League Baseball players using a Bayesian semiparametric model. They estimate performance curves for each player using orthonormal quartic polynomials and also use a Dirichlet process prior on the unknown distribution for the coefficients of the polynomials, and parametric priors for the other effects. The model is trained on the data from 1871 to 2008. The data from 2009 to 2016 is used to test the predictive ability of the model. A parametric model is also made fit to compare the predictive performance of the models. The authors use 'pure performance' curves to predict future performance for 22 players. The nonparametric method provided superior predictive performance. Finally, an RMSE of 0.01485 for the HB model and 0.01482 for NPB model was observed [1].

In this paper the authors propose the use of publicly available Statcast data and PECOTA (Player Empirical Comparison and Optimization Test Algorithm) forecasts to help predict the batting averages for MLB. They introduce a "luck" component to the batting that will not be repeated in future seasons. The "luck" component is established by analysing the characteristics of the detailed player at-bats using the Statcast Data. Using the 2015 Statcast data for all players, a logistic regression model is trained to estimate the probability of a hit. From the obtained probabilities, Statcast batting averages for 2016 are obtained. Now using the 2016 Statcast Predictions, 2016 PECOTA predictions & the actual 2016 batting averages - combined 2017 predictions are obtained. Finally, a comparison is performed [2]

In this paper the authors consider the problem of estimating a Major League Baseball player's batting average in the second half of a season based on his performance in the first half. Two linear regression models are made to fit to the player's averages from each half of the 2004 season and were used to predict batting averages in the latter half of the 2005 season. The first model used only batting averages from the first period of play to predict the performance while the second included the player's number of at-bats in the first period along with his batting average. The predictions were made from the First half of 2005 and also from all 2004 and first half of 2005. A comparison between the proposed models and the three Bayesian estimators proposed by Brown (2008) is conducted systematically. The proposed models outperform the estimators proposed by Brown. The paper predicts the second-half performance using 1.5 seasons' worth of data as opposed to brown that uses only 2005 seasons' data. All comparisons were made using four measures of error prediction - two total squared errors, RMSE and Mean absolute difference [3].

In this paper, the authors discuss a method to aid in predicting winners in Major League Baseball. They determine three main factors - team strength (which includes the past performance of the two teams), the Batting ability and the starting pitchers. Along with these factors they add a 'home-field advantage' variable to form a two-stage Bayesian model and use the Markov Chain Monte Carlo algorithm to carry out Bayesian inference that will help predict outcomes for the games played in the future. The authors make an important observation that a majority of teams win more often at home than on the road, hence suggesting the usefulness of the home field advantage variable. They deduce that the relative strength of each team is based on three ratios - ratio of winning percentages between two teams, ratio of overall team batting averages and the ratio of the ERAs (earned run averages) between two starting pitchers. In the first stage of the proposed model, it is assumed that the probability that a given team wins is a random sample from a beta distribution with parameters based on the relative strength variable and the home field advantage variable. In the second stage, the outcome (win or loss) is a random sample from a Bernoulli distribution with this winning probability. MCMC sampling is used to simulate the outcomes of future games [4].

In this study, the authors propose various deep learning and machine learning models to predict the outcomes of MLB matches. The match data of 30 teams during the 2019 MLB season with only the starting pitcher or with all pitchers in the pitcher category is collected to compare the prediction accuracy. A one-dimensional convolutional neural network (1DCNN), a traditional machine learning artificial neural network (ANN), and a support vector machine (SVM) were used to predict match outcomes with fivefold cross-validation to evaluate model performance. The highest prediction accuracies were 93.4%, 93.91%, and 93.90% with the 1DCNN, ANN, SVM models, respectively, before feature selection; after feature selection, the highest accuracies obtained were 94.18% and 94.16% with the ANN and SVM models, respectively [5].

[6] book series reflects the recent rapid growth in the development and application of R, the programming language and software environment for statistical computing and graphics. R is now widely used in academic research, education, and industry. It is constantly growing, with new versions of the core software released regularly and more than 4,000 packages available. It is difficult for the documentation to keep pace with the expansion of the software, and this vital book series provides a forum for the publication of books covering many aspects of the development and application of R.
[2] considers the problem of estimating a Major League Baseball player's batting average in the second half of a season based on his performance in the first half. We fit two linear regression models to players' averages from each half of the 2004 season, use these models to predict batting averages in the latter half of 2005 and compare the results to those achieved by three Bayesian estimators considered by Brown (2008) [3].

III. METHODOLOGY

A) Dataset

First, we obtained data from one of the most popular and free of cost MLB data sources: The Lahman Database. This database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2021.The dataset contains __ tables out of which will mainly be focussing on teams, batting and pitching for analysis.

B) Predicting wins by a team

1. Data Pre-processing

A new column called run game was created to find out the number of runs a team scored per game. For that runs column was divided by the sum of wins and losses. The mean runs per year was also calculated which was also plotted using Ggplot library. For better analysis only the data collected after 2010 was considered for further analysis. Finally, the dataset was split into a training and testing set.

2. Model Training

For predicting the wins by a team, the wins column was chosen as the target variable on which a linear regression model was applied. The columns with low correlation with the target variable are removed. The model is trained again on the training set.

3. Testing and visualization of result

Model is tested on the testing set and a new column is created containing the predictions. For comparison of actual and predicted results Ggplot is used for plot a smooth and a point graph.

C) Runs vs Wins

The second analysis in our study involves visualizing the Runs versus the Wins prediction. The Linear regression as well as the Decision Tree Regression Algorithms have been used for analytical modelling.

1. Data Pre-processing

The Teams table of the Lahman Dataset has been used for this part of the study. Only the important columns such as the TeamID, YearID, LeagueID, Games Played, Wins, Losses, Runs Scored, Runs scored by Opponents are extracted from the Teams table. Values of years greater than 2000 have been used for the analysis. A new column for Run Differential and Winning Percentage have been created. The formulas for calculating both these variables are given below. A scatterplot can be used to visually display the relationship between Run Differential and Winning Percentage.

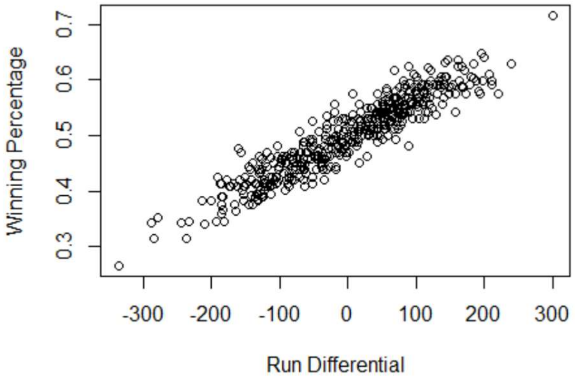

Fig 1. Graph of Winning Percentage vs Run Differential

As we can see from Fig 1, there is a strong positive, linear association, or correlation between the 2 variables. It is rather intuitive that in order to win more games, a team should limit the amount of runs they allow.

2. Model Training

The 2 variables i.e., Winning Percentage and Runs Differential are fitted into 2 models, Linear Regression and Decision Tree Regression Models.

After fitting the Linear Regression model, the equation that we get is Wpct= 0.49999918 +0.0006287 *RD. The 0.4999918 value from this equation means that if a team has a run differential of 0, our model predicts that, on average, they will win about 49.99% of their games. Likewise, the 0.0006287 value suggests that for every 1 additional run-in run differential, our model predicts a 0.0006287 increase in winning percentage on average. We can see how well this equation predicts values by looking at its residuals.

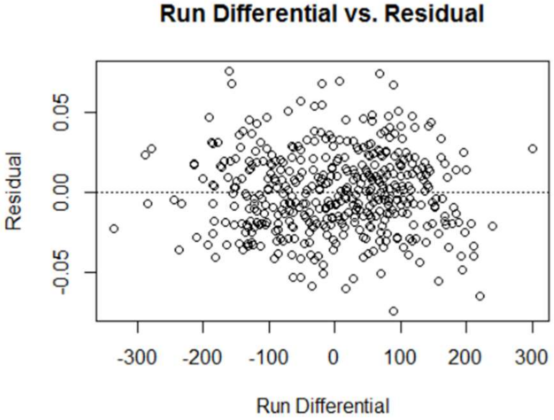

Fig 2. Graph of Residual vs Run Differential

As visible from Fig 2., the residuals are randomly dispersed around 0, we have evidence that using a linear model for this data is appropriate.

After fitting the Decision Tree regression Model, the following tree diagram is obtained as visible is Fig 3

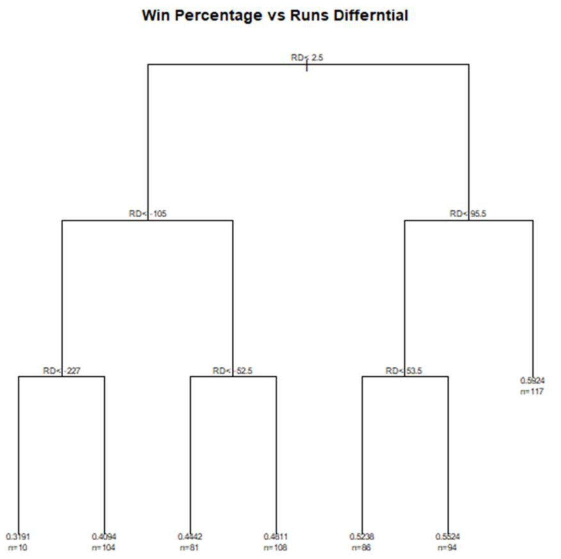

Fig 3. Graph of Residual vs Run Differential

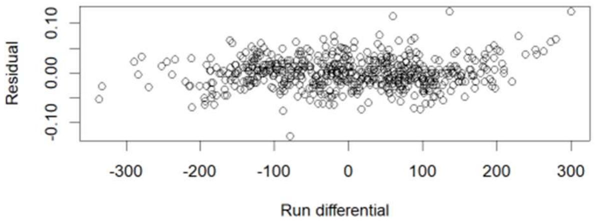

Fig 4. Residual vs Run Differential

The graph for Residuals versus the Run Differential, in the Decision Tree Regression models is as above.

D) Home Runs versus the Career Trajectories

The third part of our study does a career performance analysis based on the number of homeruns hit. The career trajectories of top ten players are plotted by using the facet plot.

1. Data pre-processing and Visualization

Being a Hall of Fame player is based on a whole career of performance. The data are given per year and per player. To get career-long home run statistics we need to group the data by player and analyse total statistics for each player's sub- data frame. We restricted our attention to players that were in the league for at least 10 years. We merged this data into the Batting data frame. We used Ddply to group the data by player, we added the career year variable calculated for this player, then joined all the data frames together into one. To get deeper information about top career players it'll be helpful to analyse the most recent players and Hall of Fame players from the past. To this end, we'll segregate players that started before 1940 from those that started after 1940. This requires tagging each player with the starting year. We added a column to the data. frame with this information. Now we select subframes of players with separate starting years. We formed a new data frame that just has the total home runs per player. We sorted this data frame by total home runs in descending order. We viewed the performance of all ten players before 1940 and after 1940 using a facet plot.

E) Strikes versus Career Trajectories

This last part of our study focuses on analysing the career performances based on the number of strikeouts by the players. The career trajectories of top ten players are plotted by using the facet plot.

2. Data pre-processing and Visualization

Pitching table from the Lahman dataset is used to analyse the career performance of pitchers. The most important parameters in pitching will be the strikeouts and the number of years has played. To get career-long strike outs statistics we need to group the data by player and analyse total statistics for each player's sub-data frame. We have restricted our study to players that were in the league for at least 10 years. We merged this data into the Pitching data frame. We used Ddply to group the data by player, we added the career year variable calculated for this player, then joined all the data frames together into one. To analyse the data more accurately we segregated players that started before 1940 and those after 1940. This requires tagging each player with the starting year. We added a column to the data frame with this information. Now we select subframes of players with separate starting years. We formed a new data frame that just has the total strikeouts per player. We sorted this data frame by total strikeouts in descending order. We viewed the performance of all ten players before 1940 and after 1940 using a facet plot.

IV. RESULTS AND DISCUSSION

A) Predicting wins by a team

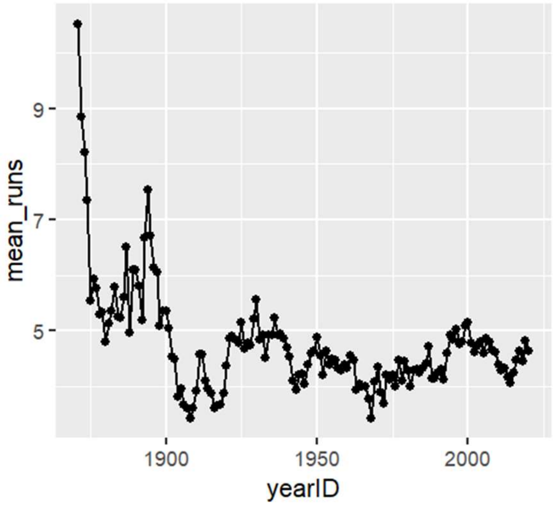

Fig 5. Graph of Mean Runs vs Year

The graph above depicts the overall trend of average runs over the years, and as we can see, the average runs have dropped slightly as the years have passed.

After the model is trained it is tested on the testing set and predicted results were added as a new column pred.

Comparing the wins column and pred columns we can observe that all the values predicted by our model are very similar to the actual values.

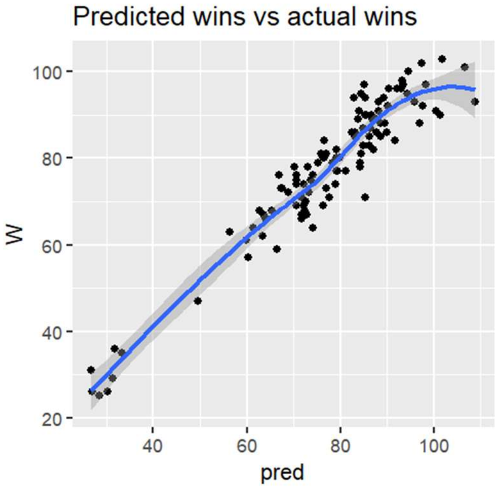

Fig 6. Graph of Predicted Wins vs Actual Wins

In the above graph, the line represents the predicted wins and points represent the actual wins. As we can see the line is fitting the data almost perfectly which further proves the reliability of the model.

By visualizing the results, we can conclude that the linear model is a good choice for performing regression on the teams table to predict the wins by a team.

B) Runs VS Wins

As visible from Fig 3, 4 the graphs of Residual vs Run Differentials are similar and randomly dispersed around 0, thus showing that both the models are a very good fit for predicting the win percentage.

C) Home Runs versus the Career Trajectories

For analysing the batting performance of the top batters, we have used the Batting table of the Lahman dataset.

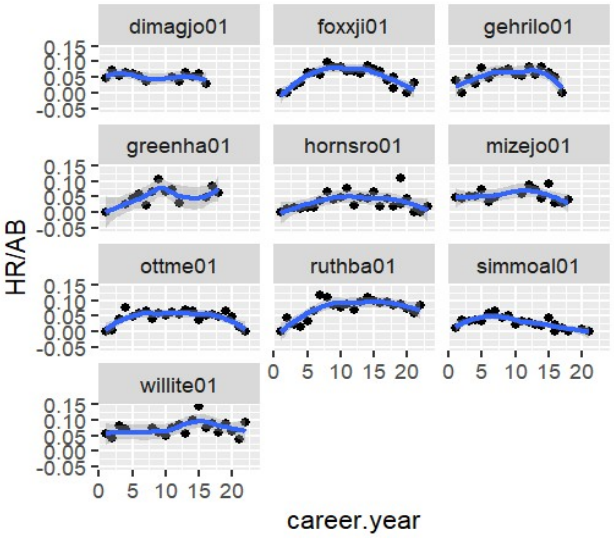

Fig 7. Career Trajectories of Batters before 1940

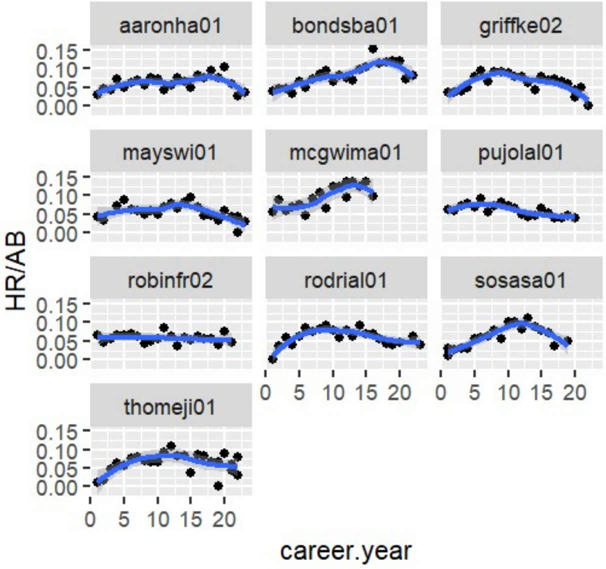

Fig 8. Career Trajectories of Batters after 1940

In most players their skills declined after about 10-12 years or stayed flat. Also, they tend to climb to a plateau early (7-8 years).

D) Strikes versus Career Trajectories

For analysing the pitching performance of the top pitchers, we have used the Pitching table of the Lahman dataset.

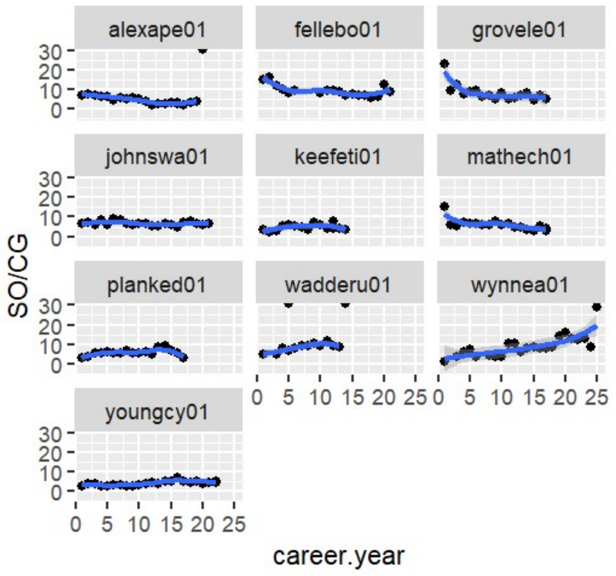

Fig 9. Career Trajectories of Pitchers before 1940

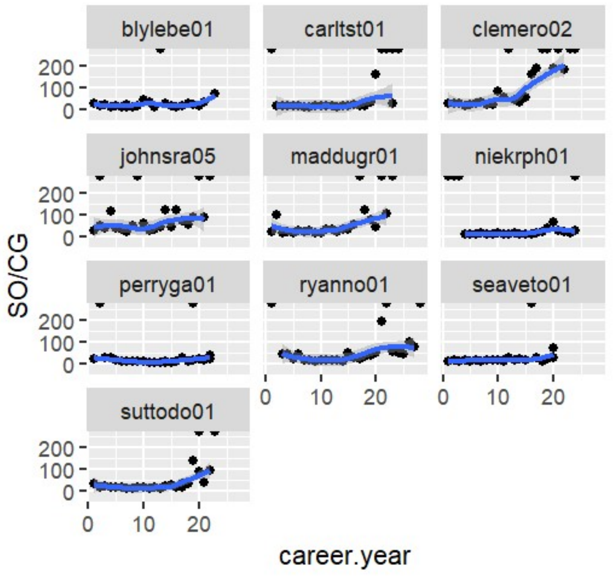

Fig 10. Career Trajectories of Pitchers after 1940

In most players their skills started improving after about 10-12 years and the stayed flat. Most players have consistent covers and there is rare decline in skills as observed from the plots.

V. CONCLUSION

This paper focuses on four different parts of MLB. Two of those focus of the batting and pitching analysis of the players. By using the batting and pitching datasets the carrier trajectories of top 10 players are found out based on their batting and pitching performance in MLB. This study also affirms the potential of ML in the prediction of winners based on the Lahman Teams dataset by using the Multiple regression model.

REFERENCES

[1] Karnuta, Jaret M., et al. "Machine learning outperforms regression analysis to predict next-season major league baseball player injuries: epidemiology and validation of 13,982 player-years from performance and injury profile trends, 2000-2017." Orthopedic journal of sports medicine 8.11 (2020): 2325967120963046.

[2] Baseball and Machine Learning: A Data Science Approach to 2021 Hitting Projections | by John Pette | Towards Data Science

[3] Kim, Jaemin, et al. "Data analytics and performance: The moderating role of intuition-based HR management in major league baseball." Journal of Business Research 122 (2021): 204-216.

[4] Analyzing the performance of the Major League Baseball Teams by using the Data Envelopment Analysis Bi, Yanzhi.Business & Entrepreneurship Journal; Christchurch Vol. 10, Iss. 1, (2021).

[5] Sensors | Free Full-Text | Individualized Ball Speed Prediction in Baseball Pitching Based on IMU Data

[6] Elitzur, Ramy. "Data analytics effects in major league baseball." Omega 90 (2020): 102001

[7] Crotin, Ryan L., Toshimasa Yanai, Peter Chalmers, Kenneth B. Smale, Brandon J. Erickson, Koji Kaneoka, and Masaya Ishii. "Analysis of injuries and pitching performance between major league baseball and nippon professional baseball: A 2-team comparison between 2015 to 2019." Orthopaedic Journal of Sports Medicine 9, no. 5 (2021): 23259671211008810.

[8] Platt, Brooks N., Timothy L. Uhl, Aaron D. Sciascia, Anthony J. Zacharias, Nicole G. Lemaster, and Austin V. Stone. "Injury rates in Major League Baseball during the 2020 COVID-19 season." Orthopaedic journal of sports medicine 9, no. 3 (2021): 2325967121999646

[9] Yang, Tae & Swartz, Tim. (2004). A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball. Journal of Data Science. 2. 61-73

[10] Neal, Dan, Tan, James, Hao, Feng and Wu, Samuel S. "Simply Better: Using Regression Models to Estimate Major League Batting

Averages" Journal of Quantitative Analysis in Sports, vol. 6, no. 3, 2010.

[11] Bailey, S.R.; Loeppky, J.; Swartz, T.B. The Prediction of Batting Averages in Major League Baseball. Stats 2020, 3, 84-93. https://doi.org/10.3390/stats3020008

[12] Kaan Koseler, Matthew Stephan, ” Machine Learning Applications in Baseball: A Systematic Literature Review”(2018), Applied Artificial Intelligence, 31:9-10, 745-763, DOI: 10.1080/08839514.2018.1442991

[13] Gabriel B. Costa, Michael R. Huber, John T. Saccoman, “Understanding Sabermetrics: An Introduction to the Science of Baseball Statistics, second edition”, US, McFarland & Co Inc, 5th June 2019