ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis [Program Data Science and Marketing Analytics]

# Homerun Characteristics and Generalized Batting Performance Prediction for Players in Major League Baseball (MLB)

Name: Ziyi Lin

Student Number: 536661

Supervisor: Dr. Andreas Alfons

Second assessor: Dr Michel van Crombrugge

Date Final Version: 22/08/2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam

# Table of Contents

## Glossary[1]

At-bat (AB): When a batter reaches base without being hit by pitch, base on balls and sacrifice himself.

Ball: A ball thrown outside the strike zone without getting hit

Batting Average (BA): Hits divided by the at-bats.

Catcher Interference (CI): A catcher interferes with the batter to swing.

Double (2B): A Batter hits the ball into fair territories and goes to second base safely.

Hit (H): When a batter hits the ball into the fair territory without getting fielded out.

Hit by pitch (HBP): Batter (Other than the bat) got hit by the ball thrown by the pitcher, the batter batting goes to first base.

Home Run (HR): A Batter hits the ball and goes to all three bases and returns to home plate safely.

Innings (INN): Game Progression indicator.

On base Percentage (OBP): A measurement with how frequently a batter goes onto the base.

Out (O): Batter who is battling and/or baserunning retired by the team in the field.

Plate Appearance (PA): A batter completes a battling turn regardless of the outcome of the turn.

Sacrifice Bunt (SH): Batter bunt and got fielded out but allows other runners to score a run.

Sacrifice Fly (SF): Batter hit a flyball and got fielded out but allows other runners to score a run.

Single (1B): A Batter hits the ball into fair territories and directly goes to the first base safely without

Slugging Percentage (SLG): Total number of base hits per at bats, where different hits value differently.

Strike Out (SO): A Pitcher throws any combination of three strikes to the batter.

Strikes: A ball thrown inside the strike zone without getting hit (First two foul balls also count as a strike).

Triple (3B): A Batter hits the ball and goes to the third base safely.

Walk (BB): Pitcher throws four balls, and the batter batting goes to first base.

[1] All the definitions from the word are retrieved from the MLB official website and the official rulebook in 2022 (MLB, 2022)

## Abstract

This paper explores the possibility to predict homeruns and one generalized batting performance indicator, specifically the slugging percentage (SLG) for batters by using machine learning approaches in Major League Baseball (MLB) with game log data from 2022 regular season from Savant Statcast. In this paper the logistic regression and eXtreme Gradient Boosting (XGBoost) is compared for the predictive performance from the model side as well as the applicability of predicting the performance for batters and their characteristics when predicting homerun. The result shows that XGBoost performs somewhat better than the logistic regression for homerun prediction but still with a slightly high under and over evaluation. This paper also finds out that it is possible to extract some insights about batters' performance by predicting the batting outcomes and transform it into SLG indicator and some characteristics from each batter can also be potentially shown in this process.

## Introduction

From time to time, the sports industry has grown larger than ever before and some teams and franchises in certain sports are earning a large amount profit. At the same time, it results in a generous number of views and creates a huge fan base for each team and franchise. Meanwhile, the athletes and their performances are also an important part of it because it is tied to their team performance and for each team, bumping up the team performance and aiming to have an excellent result. Therefore, for teams, using the proper metrics to access the performance from their own athletes and analyzing their opponents will become their priority. Different sports have different matrices to evaluate individual and team performance. In this paper, I will be focus on analyzing the performance of baseball players in Major League Baseball (MLB) in the United States. In baseball, the athletes' performance and the decision making will be crucial for the games' result. Especially for the batters, getting home runs is one of the factors for accessing the athletes' performance.

Some potential factors, including the launch angle, exit velocity and travelled distances, will be examined in this paper. In addition, some other dynamic factors, such as the characteristics of the ball thrown by the pitcher and the progression of the game can also

play a part in hitting a homerun. In this situation, there are two groups of main factors, namely the physical factors and the dynamic factors. Since the motion of baseball should obey classical mechanics, more precisely the Second Law of Motion of Newton (1833), and the launch angle and exit velocity would subsequently give the final travelled distance. In addition, in the paper of Sawicki, Hubbard, and Stronge(2003), the authors also mentioned the optimal way to hit baseball in a theoretical way. In addition, pitchers' action is also important. Since the pitcher decides which ball to throw, if the pitcher throw a slow ball rather than a fast ball and the batter use the same launch angle and the force, it will not be likely to hit a home run which means the pitcher's characteristics, such as the release speed and spin rate of the ball thrown is potentially a factor of hitting a homerun. Besides that, as Cross (1998) has mentioned in his paper that there are certain areas which are the sweet spots for hitting a baseball on the baseball. In this case, it can be inferred that the hitting zones can also be a factor since in some zones batters are not likely to hit in the sweet spot on the bat. It is also to be noted that there might also be other physical variables that can affect this such as the weather conditions and the pitch dimensions.

Besides the physical factors, some in-game dynamic variables can also influence the outcome of the swing. In some game situations, such as the team is behind the score, or the pitcher has two strikes. This can influence the batters because they are performing under pressure which will subsequentially affect their swings, which Baumeister and Showers (1986) also discussed this effect on performance. Therefore, the in- game dynamics can also potentially affect if the ball is homerun.

There is still another thing which needs to be mentioned. Since the environment when playing baseball is changing, some factors, such as the temperature and humidity, might also play a role of hitting a home run. In addition, some stadiums have a fixed roof or retractable roof, such as Tropicana Field for Tempa Bay Rays, which is totally an indoor stadium. In addition, according to one report from Fox Weather (2023), Miami Marlins has 78 out of 81 games played in the regular season with their retractable roof closed. This means that the weather factor will not be effective if the match is played indoors. At the same time, in the official MLB baseball rule book (2023), it only states the minimum dimension of the field. In this case, the hitting distance does not guarantee if a hit can be a

homerun or not. It also depends on which way the ball goes; some trajectories would have less distance and some trajectories need more distances to travel to get a homerun. From the game dynamic wise, playing away games would also put some sort of pressure on the away team athletes which can also affect athletes' hitting performance. Therefore, the variable of home team, specifically playing the home match, should also be added into all three hypotheses since it plays both physical factors and dynamic factors.

Based on the situation above, it is interesting to study how batters swing optimally to get the homerun and how the game dynamics affect their batting performance. Therefore, the following central question will be formulated:

*"How do we predict hit ball to be a home run and how well batters achieve it under different game circumstances? "*

It is scientific relevant to do the research because knowing how players hit the home run will be a useful information for the team to understand the home run which can leads to the team identifying strong and weak batters in terms of hitting, not only for the franchise itself, but for all opponent teams as well. It can also help the team to find some sort of strategies when facing strong batters against the opponents because the team of MLB would like to mainly focus on their results. In addition, even though there are some research about sports analytics, comparing to other issues, sports analytics such. as this, to some extents are somehow ignored by the academia. Among the previous research of sports analytics in baseball, it is related to evaluate the performance level of the players in general or analyzing in physical and in life science perspective. In addition, some in-game dynamics, such as the opponent pitchers ball thrown can also play a crucial part in the batters' actions. Therefore, it is way more important to convey this research to better understanding the science behind each specific sport, where in this paper – baseball.

It is also socially relevant to do this research because first, sports are part of people's life and baseball has some popularity in the US and MLB is one of the biggest associations. Meanwhile, home runs can result in better fans involvement in the game since when the home run occurs, fans watching the game inside the stadium can try to catch the ball and

keep the ball which gives fans better interaction. At the same time, it will retain attract new fans which can potentially be helpful for the athletes and the team.

In this paper, I will analyze hits from the batters and their corresponding characteristics of this hit and then analyze the link between hitting a home run and the characteristics of the hit. This will be done by using some advanced regression models and some tree-based algorithms. For the following section, I will discuss the related literature of sports analytics in general and specifically for baseball, as well as some previous research in baseball analytics. In addition, how the people analyze the performance for the players among different sports will also be discussed.

## Literature review

Even though sports analytics are, to some extent, being ignored by the academia, there are still plenty of papers which have been done in this field, especially in the era of "Big Data". Morgulev, Azar, and Lidor (2018) define that sports analytics investigates the performances in sports by using some scientific techniques. It consists of Information gathering, Data Management, Data Analysis and then give the final decisions from the decision makers. In the so called "Big Data" era, it transforms how the game will be played strategically. In the article, the authors gave some examples of the use proper indicators to identify the valuable player and the player development. They use examples in NBA, where Boston Celtics successfully drafted Rajon Rondo who turned out to be an all-star point guard because the scout from Celtics successfully identified that the rebound ability for a point guard is vital. Comparable situation goes to Seattle Supersonics where the team drafted Russel Westbrook since he has great shooting skill at that time for a point guard. The authors also give a counter example from football where Sir Alex Ferguson wrongly sold Jaap Stam because his tackling statistics decreased. Morgulev et al (2018) also pointed out that by analyzing the sports data, it can help the team make decisions and learn the human behaviors among these sports data.

Sarlis and Tjortjis (2020) evaluate the performance of the basketball players and teams and they are trying to optimize the rating system and find out what are the most essential characteristics for the players to get the Most Valuable Player (MVP) and Defender of the

Year. In the article, they created two different measurements, the Aggregate Performance Indicator (API) and Defensive Performance Indicator, and they correctly predict the MPV by using the data from the specific season with API. Barrow, Drayer, Elliott, Gaut and Osting (2013) focus on the various kinds of ranking in different sports and their predictive power. They found out that for each method applied, the predictive power will vary among different sports.

In the previous paragraph several findings of sports analytics from different sports are presented. It also has a certain number of applications for baseball. Mizels, Erickson, and Chalmers (2022) describe that in the current situation for data analytics in baseball, it allows people to use Statcast database to evaluate and predict the performance of different teams and players with different performance metrics. During the game, it is also possible to analyze the motion of the ball with kinematic data. The article also states that by using machine learning and artificial intelligence it is also possible to predict players injury.
Chu and Wang (2019) study relationship between implementation of sports analytics and the team performance for MLB teams. The authors have found out that for the teams who do not believe that sports analytics is useful, it is less likely that the corresponding team is going to manage to make it into the playoffs. Furthermore, teams with medium-sized research staff perform more consistently than other groups. However, it is still hard to predict if the team will be successful in the playoffs. Sports analytics in baseball also has a name of Sabermetrics, where Albert (2017) called it as scientific study of baseball. In this case, it can also be seen as sports analytics in baseball by using mathematical and statistical methods.

At the early age of baseball games, James (2010) gives the simple run creation formula, the foundation of analyzing the game. Lindsey (1963) first uses probability theory to find out the strategies in baseball. The author first investigates the winning game strategies, and he has found out that as the game progresses, getting more runs becomes more essential to subsequentially win the game. Besides of runs creations, Lindsey also examines some useful strategies in different circumstances, such as intentional walk, fielding options, stealing bases and intention of double play. In addition, the author gives a measure of batting

performance which is helpful for the run creating and hitting. The article also concludes that a desirable batting performance is a good indicator of the value of the batter.

Courneya and Chelladurai (1991) try to find an optimal model for measuring the performance of baseball players. They categorize different measures into primary, secondary, and tertiary measures and they use the sample data from National Collegiate Athletic Association (NCAA) baseball teams. The article shows that the mean correlation between primary and tertiary measures is lower than the mean correlation between secondary and tertiary measures. Variance wise, the secondary measures explain the most variance compared to the primary and tertiary measurements on average but for the hitting and pitching measures, most variances are captured in the primary measurements. For the tertiary measures specifically, the authors discover that run differentials correlate the highest with both primary and secondary measurements. They give some explanations such as the close performance level and difficulties of maintaining the performance level for both teams. At the same time, famous sabrematrician McCraken (2001) argues that pitchers do not have so much influence in terms of defense and they cannot prevent hits at all. He also gives some explanations such as the scouting issue and the dynamic between the batter and the pitchers.

Besides the typical performance measurements, the prediction of the game and various kinds of result is also important. Ganeshapillai and Guttag (2012) predict that for the next pitch whether the pitcher will throw a fast ball. They built a model with historical data from both regular season matches and playoffs from 2008 and 2009 season by using the supported vector machine. They listed out some important classifiers, such as the opposite batter from the prior round and the count of the current round. They also state that the pitchers will be more predictable in less favorable counts compared to favorable counts. In addition, starting pitchers in general tend to choose throwing more various balls than a reliever. Hoang (2015) also tries to predict the next pitch for the pitcher by using different methods. He first divided the groups with dynamic features and returned to three groups, namely the batter favored, neutral and pitcher favored. He finds out that for overall results, Linear Discriminant Analysis (LDA) has the best accuracy. For the count analysis, predicting accuracy with batter favored counts is higher than the neutral and pitcher favored counts.

For the individual pitcher's analysis, the author discovers that the accuracy from different individuals vary. This conclusion is in line with the finding from Bock (2015), where he argues that for different pitchers, the predictability of the next ball thrown is different. In addition, Boch (2015) also states that the predictability of the pitcher can somehow project their long-term performances. Knowing what will come from the next pitch is important, not only for the opposite batters, but also managers from both teams.

Gartheeban and Guttag (2013) also study the in-game decision making the baseball games, namely if the manger will choose to relive the pitcher by modeling with the Pitcher's Total Bases. They evaluated the data from 2006 to 2010 and they showed that their model with regularized regression has better accuracy compared to the manager's model, which only included pitch count and the score. They use the example from one case game between Milwaukee Brewers and Pittsburgh Pirates game in July 2010 as well, showing that keeping the started pitcher play in the $6^{th}$ inning was a wrong decision from the manager. Nakahara, Takeda and Fujii (2023) evaluates the pitching strategies by using the propensity score and trying to find a relatively optimal pitching strategy for the pitcher. They illustrate that for the pitchers, when the ratio of the pitching inside to outside is balanced, throwing outside can be an effective strategy again the batter, especially throwing a fastball and throwing inside can be risky. Healey (2015) assesses the strike out rate from batter and pitcher pairs by using the game play data. To do this, the author builds up log5 model, where it evaluates the 1 vs 1 winning probability from two agents, as well as a generalized model to match the batter and pitchers and assesses the hand they are using for throwing or swinging in the game. He uses the descriptor reliability to select the proper variables into the exploratory ones. He concludes that variances strike out probability is lean on batters' hitting ability and their characteristics for most of times. By using a similar method, he also predicts the probability of a batter hitting a ground ball. He also mentioned that hitting from a ground ball depends on the strikeout rate from the pitcher and the ground ball rate from the pitcher, which can have some effects on team performances (Healey, 2017).

After explaining the previous works in baseball. In the next section, I will explain the data sources of this research, followed by the methodologies and the analysis.

# Data

## Data Descriptions

The performance data from athletes will be collected from Savant, which is the database dedicated to sports analytics specifically for baseball. It contains all the data from each team. Including the in-game behaviors from the pitchers, batters, and field. It also gives the results for each pitch and bats in every game team and player have played from different seasons. This means all the game logs are included. Table 1 shows part of variables in the Savant dataset according to the documentation of Savant (2023):

Table 1: Part of the Variable description of the Savant Data Set.

| Variable name | Description |
|---|---|
| pitch_type | The type of pitch derived from Statcast. |
| release_speed | All velocities from 2017 and beyond are Statcast, which are reported out-of-hand. |
| player_name | Player's name tied to the event of the search |
| batter | MLB Player Id tied to the play event. |
| pitcher | MLB Player Id tied to the play event. |
| p_throws | Hand pitcher throws with |
| home_team | Abbreviation of home team. |
| balls | Pre-pitch number of balls in count. |
| strikes | Pre-pitch number of strikes in count. |
| inning | Pre-pitch inning number. |
| hit_distance | Projected hit distance of the batted ball in feet(ft). |
| launch_speed | Exit velocity of the batted ball |
| launch_angle | Launch angle of the batted ball |
| release_spin | Spin rate of pitch tracked by Statcast in revolutions per minute(rpm) |
| bat_score | Pre-pitch bat team score |
| fld_score | Pre-pitch field team score |

*Note.* All the speed are in mile per hour(mph)

Besides the identification variables like batter and pitcher, some variables, such as the hit distance and launch speed are variable related to the physics of the bat and the ball, and other variables such as the bat and filed score, inning, balls, and strikes are the game dynamic variables. In addition, strike, balls and out counts will be transformed to categorical variable instead of the numerical variable and I will create another variable called the score differential which is bat team score minus the field team score.

In this situation, I will be focusing the data from the batters, specifically in 2022 season and for the matches played in regular season and play-offs because these games matter the most for every franchise since it has the direct effect on the final ranking for the. To measure which hit can be a homerun, I will select the full game log data with every pitch thrown. This raw dataset gives a grand total of 720272 data points. This allows me to predict home since it has all the corresponding variables for all the hypotheses because all the required variables are inside this dataset.

## Descriptive Analysis

In total, in the 2022 regular season, all 30 teams recorded 5215 homeruns. This means that for each game it will have on average 1.073 homerun per match and 0.007 homerun per pitch, given that there are 708540 pitches played among all the batters. According to the dataset, Aaron Judge from New York Yankees scores the most homerun with 62 homeruns, followed by Kyle Schwarber with 46 homeruns from Philadelphia Phillies and Mike Trout with 40 homeruns from Los Angeles Angel. If I account for the stadium on which one get the most homerun, the first place has gone to the Great American Ball Park in Cincinnati, where it received 217 homeruns, followed by American Family Field in Milwaukee which it has 215 homeruns and Yankee Stadium in New York with 204 homeruns, which means that different ball park can have different effects  Table 2 and 3 shows the result of top five total homeruns in the circumstances.

Table 2: Top five most homerun scored players.

| Players | Homeruns |
|---|---|
| Aaron Judge | 62 |
| Kyle Schwarber | 46 |
| Mike Trout | 40 |
| Peter Alonso | 40 |
| Riley Austin | 38 |

Table 3: Top five most homerun scored stadiums.

| Venue | Homeruns |
|---|---|
| Great American Ball Park | 217 |
| American Family Field | 215 |
| Yankee Stadium | 204 |
| Rogers Center | 204 |
| National Park | 200 |

If I use the percentage of homerun scored of total pitches for the players, there are some changes in the result. Khahill Lee ranks first with 33.3% homerun however he only had 3 pitches played for the whole season. Aaron Judge still rank the second and the third will be Joe Dunand but he only played 47 pitches. Table 4 and 5 show the result about percentage of homerun over all pitchers by players and stadiums.

Table 4: Top five most homerun percentage players.

| Player | Percentage of Homerun | Total Pitches |
|---|---|---|
| Khahill Lee | 33.33 | 3 |
| Aaron Judge | 2.13 | 2906 |
| Joe Dunand | 2.12 | 47 |
| Nick Plummer | 2.08 | 96 |
| Matt Carpenter | 2.07 | 723 |

Figures 1,2,3 show the homerun scored in different in game situations. According to Figure 1, in the last inning there are less homerun scored compared to the other inning except the extra innings since the extra innings do not happen very often, which leads to less homerun scored. Figure 2 and Figure 3 show the homerun scored from different Outs and Strikes. From both figures we can see that when the outs are increased, the homerun score decreases. The same trend applies to the different strikes. This might be that under these circumstances, the batter might have more pressure to hit a ball and get a good hit.

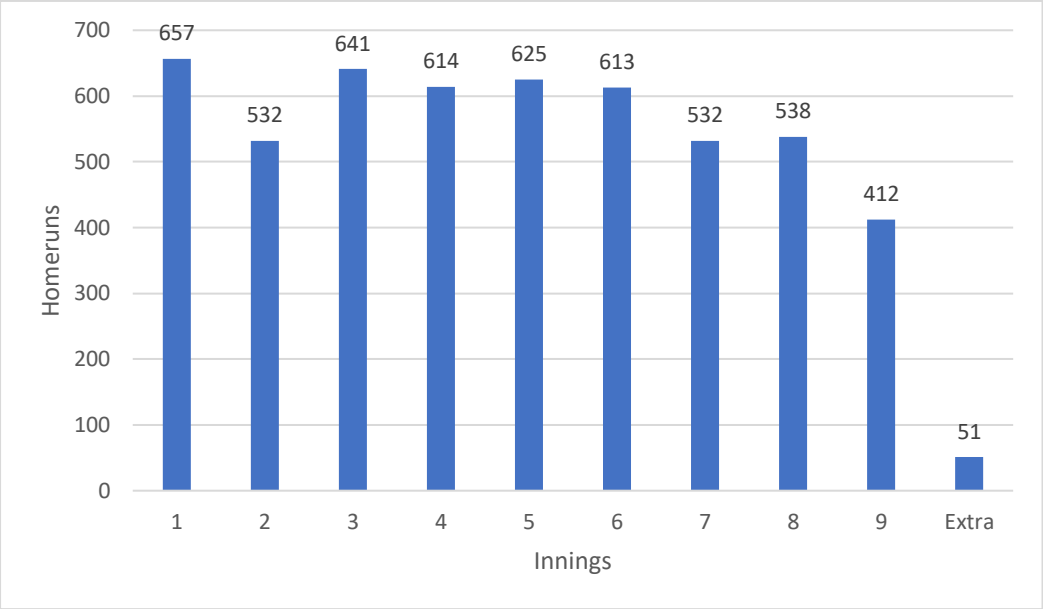


Figure 1: Homerun Scored from each innings.

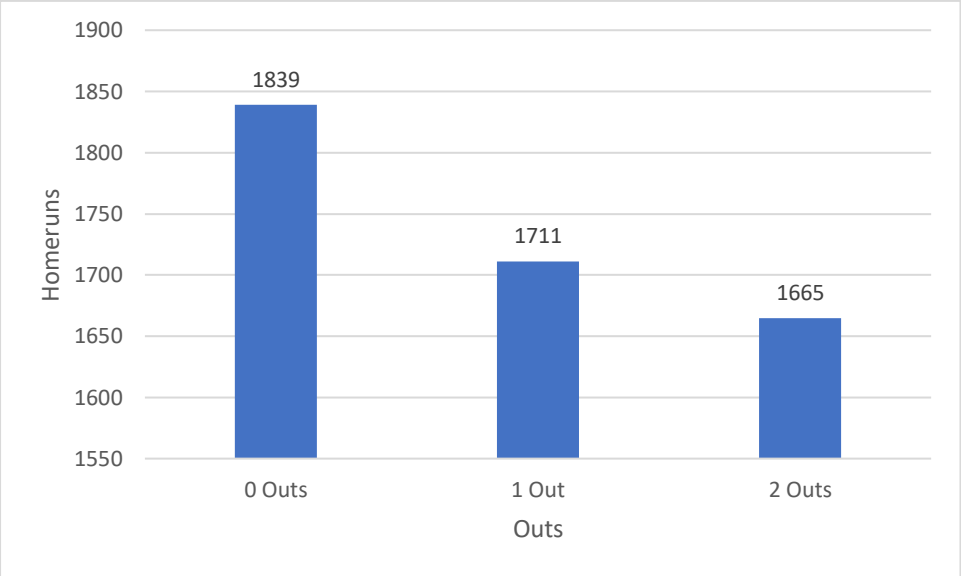


Figure 2: Homerun scored from different Outs.

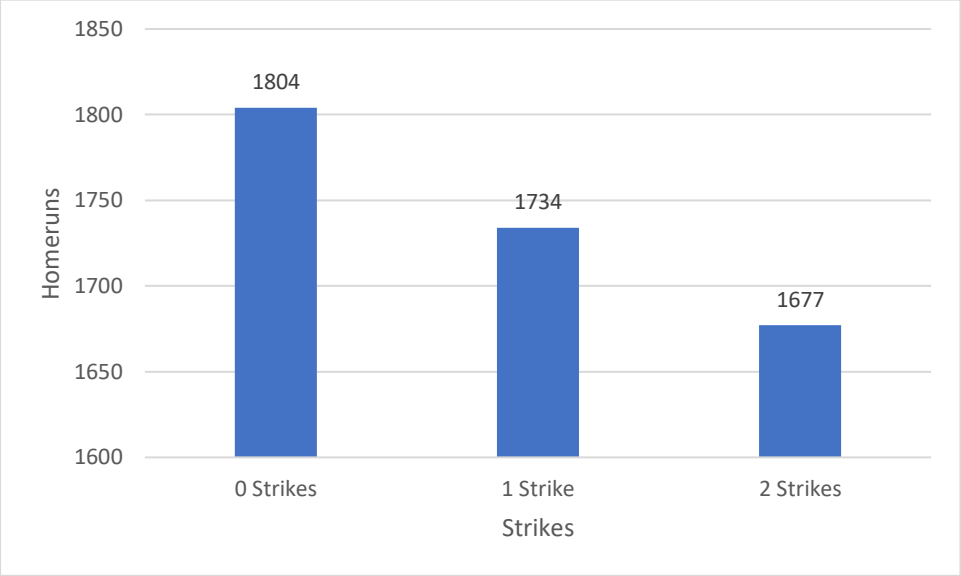


Figure 3: Homerun Scored from different Counts/Strikes.

## Methodologies

This paper aims to predict whether a swing from the batter will be a home run. In this context, the dependent variable is a binary variable, and it is a classification problem. In this case. Two different methods will be applied, namely logistic regression and one machine learning method.

### Logistic Regression

I will first use logistic regression first to look at the insights of this, which here the dependent variable is whether a swing becomes homerun, which is the binary variable. This gives some initial insights about the prediction itself and what kind of factors can influence the outcome of being a homerun. It evaluates the probability of the event happening. In logistic regression, it uses the log odds and then converts the log odds into the probability via the logistic function. For each coefficient of every independent variable, it will be estimated by the maximum likelihood estimation (MLE). This means that it maximizes the corresponding likelihood function to find the optimal value of different parameters.

Therefore, in this scenario, the logistic regression can be written as following:

$$P(Homerun) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}.$$

Where $\beta_0$ is the intercept, $x_1, x_2 \ldots . x_i$ are the independent variables in this specific regression and $\beta_1, \beta_2, \ldots \beta_i$ are the coefficients of corresponding independent variables. To get the linear relationship to have a better interpretation of coefficient, this can be transformed by writing it in the form of log-odds which is the following:

$$\ln\left(\frac{P(Homerun)}{1-P(Homerun)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n.$$

However, there is a potential problem regarding the batters' decision. In some situations, the batter would not decide to swing and in other occasions, the batter misses the swing for some reason such as bad swing timing, which is not captured in the dataset. This makes the ball a swing strike. In short, there will be only two outcomes, namely. hit or not hit the ball. In this situation, I suggest first to identify which ball gets hit with the pitchers' ball characteristics, which in this situation, the release speed, in game situation and potentially the zone as independent variable and the dependent variable as the pitchers' ball thrown. After identifying which balls are likely to be hit by the batter, I can start to predict and classify the home run again. The identification of the ball get hit can also be written like the following:

$$P(Hit\ the\ Ball) = \frac{1}{1+e^{-(c_0+c_1 m_1+c_2 m_2+\cdots+c_n m_n)}}.$$

Where $c_0$ is the intercept, $c_1, c_2 \ldots . c_i$ are the independent variables in this specific regression and $m_1, m_2, \ldots m_i$ are the coefficients of each independent variable in this regression. Similarly, this can also be written in the format of log-odds:

$$\ln\left(\frac{P(Homerun)}{1-P(Homerun)}\right) = c_0 + c_1 m_1 + c_2 m_2 + \cdots + c_n m_n.$$

## Machine Learning Method

Because simple logistic regression with large data set can give noisy results, besides the logistic regression, I will also use a proper machine learning method which can help me to answer the central question better. In the article of "Machine learning applications in baseball: A systematic literature review" by Koseler and Stephan (2017), besides regression, they suggest that in a binary classification problem, a Support Vector Machine (SVM) method can be viable method since it gives a viable prediction accuracy. But in this situation, I would like to use a tree-based method such as decision tree to do the prediction since decision tree itself supports well with non-linearities.

However, there are some drawbacks with decision trees such as relative low prediction accuracy and non-robust results. In this case, boosting will be a decent choice. This method is based on a question by Kearns & Valiant (1989) about transforming weak learners into strong learners and Schapire(1990) proved that the transformation is possible, which developed the boosting method.

There are various kinds of forms of boosting. I will use gradient boosting method for this research which will help answer the research question because firstly, it is intuitive to have a tree-based method since the batter makes decisions of swing which is linked to the outcome of the ball. Secondly, it also can work with some non-linear data and a better predictive power compared to decision tree. In addition, it can also help me to grab some useful patterns and relationship in the given data which can subsequentially improve the prediction power of homerun and help the team to identify the strong batters and weak batters which can help the team to understand more about it and make proper strategies with certain batters. It can have a purpose of helping weaker batters to improve their hitting skills.

I will also combine both physical and dynamic factors in the gradient boosting model to provide a bigger picture. In this situation, a classification model of gradient boosting will be built to predict the homerun. Boosting decreases the variance. In the gradient boosting method, it increases the prediction accuracy by using the residual weight. Suppose there is a dataset $D = \{(t_i, y_i)\}_{i=1}^{n}$ and each datapoint has a loss function of $L = (y_i, F(t_i))$, with weak learner $h$, a gradient boosting for classification has the following procedure (Hastie, T., Tibshirani, R., & Friedman, J. (2009):

1) Calculate the best initial value for predicted outcome, this will be done by calculating $F_0(t) = arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$
2) $\forall\ K \in N^+$(K is total number of iteration), for $k = 1$ to $k = K$, calculate the following: $r_{ik} = -\left[\frac{\partial L(y_i, F(t_i))}{\partial F(t_i)}\right]_{F_k(t)=F_{k-1}(t)}$, where $r_{ik}$ is the negative gradient, in another word the pseudo residuals. The aim of this is to find the steepest descent.
3) Fit the weak learner $h_k$ and find the optimal weak learner in accordance with $r_{ik}$

4) Add the optimal weak learner and apply a multiplier $\gamma$ by doing the one-dimensional optimization across all the observations $\gamma_k = arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{k-1}(t_i) + \gamma h_k(t_i))$ since gradient boosting learns sequentially.
5) Update the model by doing $F_k(t) = F_{k-1}(t) + \gamma_k h_k(t)$
6) Repeat the procedures 2),3),4) and 5) till $K$ and after those outputs of $F_k(t)$ are given.

Besides the normal gradient boosting, one can also apply eXtreme gradient boosting (XGBoost) developed by Chen & Guestrin (2016) for the prediction. It is a more regularized form of gradient boosting method, and the model is constructed by using the gradient boosting decision tree. In this situation, the loss function contains a regularization term. This can be written by $R(\theta) = L(\theta) + \Omega(\theta)$, where $\Omega(\theta)$ is a regularization term. Since it is a classification problem, the loss function will be a logistic one, which can be written as $L(\theta) = \sum_{i=1}^{n}[y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i)\ln(1 + e^{\hat{y}_i})]$. In the result and analysis section, I will apply the XGBoost to conduct the analysis.

There are still a few adjustments which need to be made in this case. Since not all the team played the playoff and only two teams from 2022, namely the Houston Astros and Philadelphia Phillies has played until the final World Series, I will only use the data from the regular season where all the teams must play the regular season match. Each batter from different teams will be trained and evaluated separately from both training and test set to predict every batter. In other words, I will train the data on an individual level and use this to predict the corresponding individual. This will indirectly allow interpersonal comparison from the prediction result on each player which can be helpful to identify strong and weak batters in terms of their batting performance and keep an eye on them according to their own batting habits. This can be down by checking the variables importance, pattern, and relationships inside each set of data from individual across different players. More importantly, since the prediction is done at a micro-level, it can be useful for the manager to do the in-game decision making. Precisely when they can ask the pitcher to let the batter go onto the base intentionally. In addition, it can also show the predictive power for the model

on different individual batters by using the evaluation metrics such as accuracy and precision.

## Pre-Processing and Hyperparameter Tuning

After the descriptive analysis, I am going to do analyze the data obtained on Savant Statcast and make predictions on homeruns for each batter in the regular season. This will be done in two different stages. In the first stage, I am going to predict whether the batter will hit the ball and on the second stage, I will predict if this hit becomes a homerun. In this case, I will use two different methods to execute the analysis, namely the logistic regression and gradient boosting as mentioned before, which can give me some useful insights with it.

Before I do the prediction, the raw data from savant needs to be pre-processed and separated from training and testing data. According to the descriptive analysis, it is super rare to have a homerun given that only 0.7% of pitches in regular season transformed into homerun. In this case, I will use the batter who has over 800 pitches to do the prediction. I will also separate the training and testing data differently since the match is game based and the traditional 80% and 20% split will not be applied because it will potentially separate the match and damage the data structure. In this case, I will use the MLB trade deadline, namely August 2nd, 2022, as the split data from afterwards, the rooster becomes fixed since the trade will be stopped after that which the team will be more stable. To avoid the perfect separation error, each player must record at least one homerun in both training and testing dataset.

For gradient boosting methods, hyperparameters, such as learning rate, maximum depths and numbers of estimator need to be tuned. This is done using k-fold cross validation and a grid search to find the best hyperparameters. Therefore, I will use 5-fold cross validation and the corresponding grid search. Some hyperparameters, such learning rate, maximum tree depths, number of estimators, gamma, lambda, and minimum child weight, are advisable for tuning. In this case, I will tune the learning rate, number of estimators and number of estimators with a proper grid. Table 5 shows all the hyperparameters that I will use for the grid search and the corresponding value for the grid search.

Table 5: Hyperparameters used for XGBoost tuning.

| Hyperparameter | Grid | Note |
|---|---|---|
| Learning Rate | {0.01,0.1,0.3} | Lower learning rate makes the model slower and more conservative. Also called shrinkage factor |
| Maximum tree depths | {3,4,5} | Maximum splits for each decision tree made |
| Number of estimators | {50,100,150,200} | Number of trees in the model |
| Minimum Child Weight | {1} | Minimum sum of the weight of having a child from a leaf node, Default value is one |
| Gamma | {0} | Also called minimum spilt loss. Specify the minimum reduction required to make a split. Default value is zero |
| Lambda | {1} | L2 regularization term, Default value equals to one |

# Result & Analysis

## Regression Analysis

In my baseline setting, I will use logistic regression to predict the homerun. In the first stage, the independent variables are release speed, score differences, home game, innings, strike count, out count, and zone where ball was thrown with the dependent variable of hit the ball. The following results show the top five and last five in terms of accuracy on predicting

power on different players. Table 6 shows the top five players in terms of accuracy in stage one.

Table 6: Top five players in terms of the accuracy score from the model in stage 1

| Player | Accuracy |
|---|---|
| Sam Hilliard | 0.821 |
| Steven Kwan | 0.819 |
| Yasmani Grandal | 0.800 |
| Issac Paredes | 0.795 |
| Carlson Kelly | 0.795 |

We can also see some different characteristics among different players according to their coefficients from each independent variable. The coefficient of the ball is thrown in the inside zone is mostly positive and significant for different players, meaning that this increases the chance of the batter hitting the ball. We take Austin Hedges as an example and its result is shown in Table 7

Table 7: Regression result from Steven Kwan

| | Coefficient |
|---|---|
| release_speed | -0.041*** |
| | (0.003) |
| score_diff | -0.001 |
| | (0.022) |
| home | -0.555*** |
| | (0.137) |
| inning | -0.042 |
| | (0.026) |
| strike_1 | 2.031*** |
| | (0.175) |
| strike_2 | 3.523*** |
| | (0.209) |
| out_1 | 0.106 |
| | (0.162) |
| out_2 | 0.101 |
| | (0.170) |
| Inside_zone | 2.592*** |
| | (0.166) |
| Number of observations | 1580 |

*Notes:* This is an estimation of probability of hit the ball from the player Steven Kwan given certain independent variables. The dependent variable 'hit the ball' is given in probability. The values given in the brackets are the standard deviation of each independent variables. 1 Star means a statistical significance of p-value<0.05, 2 stars mean a statistical significance of p-value<0.01 and 3 stars mean a statistical significance of p-value<0.001.

According to this table, the coefficient of release speed, home, strike count 1, strike count 2 and throw in the inside zone are significant, which means that all the variables have a significant effect on this player hit the ball, where release speed and home game have a negative effect on the probability of hitting the ball. Here, we can also say that for Steven Kwan, he will be less likely to hit the ball when it is a fast ball, and it is a home game.

In the second stage, the independent variable for estimating the will be the launch angle, launch speed, score differences, innings, out counts, strike counts and zone thrown, Table 8 shows the top five players' prediction in terms of accuracy.

Table 8: Top five players in terms of the accuracy score from the model in stage 2

| Player | Accuracy |
|---|---|
| Jose Trevino | 0.991 |
| Carson Kelly | 0.991 |
| Owen Miller | 0.989 |
| Santiago Espinal | 0.989 |
| Christian Vázquez | 0.989 |

From this table, we can see that all of them are around 99%. It might have good prediction power. However, this can be explained by the following: The probability of a swing which hit the ball becomes a hit is usually not high and homerun is even rarer. Some players barely get a homerun. Therefore, it makes the prediction generally harder even though the accuracy seems super high. In this case, it is also hard to interpret the result for a lot of players. Here we use Mike Trout, who has the biggest contract in the league currently, as an example, and the result is shown in Table 9.

Table 9: Regression result from Mike Trout

| | Coefficient |
|---|---|
| launch_speed | 0.007 |
| | (0.009) |
| launch_angle | -0.008 |
| | (0.010) |
| score_diff | 0.172* |
| | (0.086) |
| home | -0.428 |
| | (0.446) |
| inning | -0.128 |
| | (0.085) |
| strike_1 | -1.164 |
| | (0.515) |
| strike_2 | -1.527*** |
| | (0.571) |
| out_1 | -0.581 |
| | (0.538) |
| out_2 | -0.147 |
| | (0.545) |
| Inside_zone | -0.282 |
| | (0.621) |
| Number of observations | 174 |

*Notes:* This is an estimation of probability of getting Homerun from the player Mike Trout given certain independent variables. The dependent variable 'Home Run' is given in probability. The values given in the brackets are the standard deviation of each independent variables. 1 Star means a statistical significance of p-value<0.05, 2 stars mean a statistical significance of p-value<0.01 and 3 stars mean a statistical significance of p-value<0.001.

In this example, for Mike Trout, the coefficient of score differential and strike count in two are significant both in 5% level, where score differential has a positive magnitude and strike count two has a negative magnitude. In this case, for Mike Trout, one more score differential gives 18.8% of chance getting homerun on average and when it is in strike count

two, it decreases the chance of getting homerun on average by 78.3%. This means that for Mike Trout, when the team is leading on the score, he can have a better chance getting homerun since it might give this player more confidence of hitting and potentially less pressure when leading the score. Similarly, when the strike count is on two, it might be that he will be under some pressure, thus reducing his chance of hitting a homerun.

Table 10: Predicted Top five players hitting homeruns by using logit model.

| Player | Predicted Homerun | Actual Homerun | Accuracy |
|---|---|---|---|
| Aristides Aquino | 22 | 7 | 0.773 |
| Donovan Solano | 12 | 7 | 0.885 |
| William Contreras | 11 | 6 | 0.864 |
| Nelson Velazquez | 9 | 2 | 0.828 |
| Danny Jansen | 8 | 1 | 0.875 |

According to Table 10, we see that in the normal logistic regression prediction, we see that even though the prediction accuracy is high, it has a huge misclassification and some big understate and overstate for the predicted homerun. In this situation, it is hard to tell which players are good at hitting home runs. Since all the samples from all the players are highly imbalanced because homeruns happen not so often, it is hard to achieve an optimal prediction result from the logistic regression.

In short, it is somehow possible to predict the homerun with logit model. However, it should be also noted that it can be noisy. In addition, if the data is skewed, in this case many samples are not getting the homerun, it will not give the ideal result. In the next section, I will use the Gradient Boosting method to predict homeruns.

## eXtreme Gradient Boosting

In this section, I will use gradient boosting to predict homerun. As mentioned before in the methodology section, I have chosen the eXtreme gradient boosting (XGBoost). Therefore, Table 11 shows the top accuracy of the model in the first stage, alongside AUC score and optimal hyperparameters.

Table 11: Top five players in terms of the accuracy score from the model in stage 1 alongside the optimal hyperparameters

| Player | Accuracy | AUC score | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|---|---|
| Steven Kwan | 0.828 | 0.865 | 0.010 | 3.000 | 200.000 |
| Jorge Polanco | 0.816 | 0.853 | 0.010 | 4.000 | 200.000 |
| Sam Hilliard | 0.810 | 0.872 | 0.010 | 3.000 | 100.000 |
| Brandon Belt | 0.808 | 0.840 | 0.010 | 3.000 | 50.000 |
| Carson Kelly | 0.800 | 0.810 | 0.010 | 3.000 | 50.000 |

Same as before, we can also see how different situations affect different players through the feature importance. We again take Steven Kwan as an example and Figure 4 gives the feature importance of him.

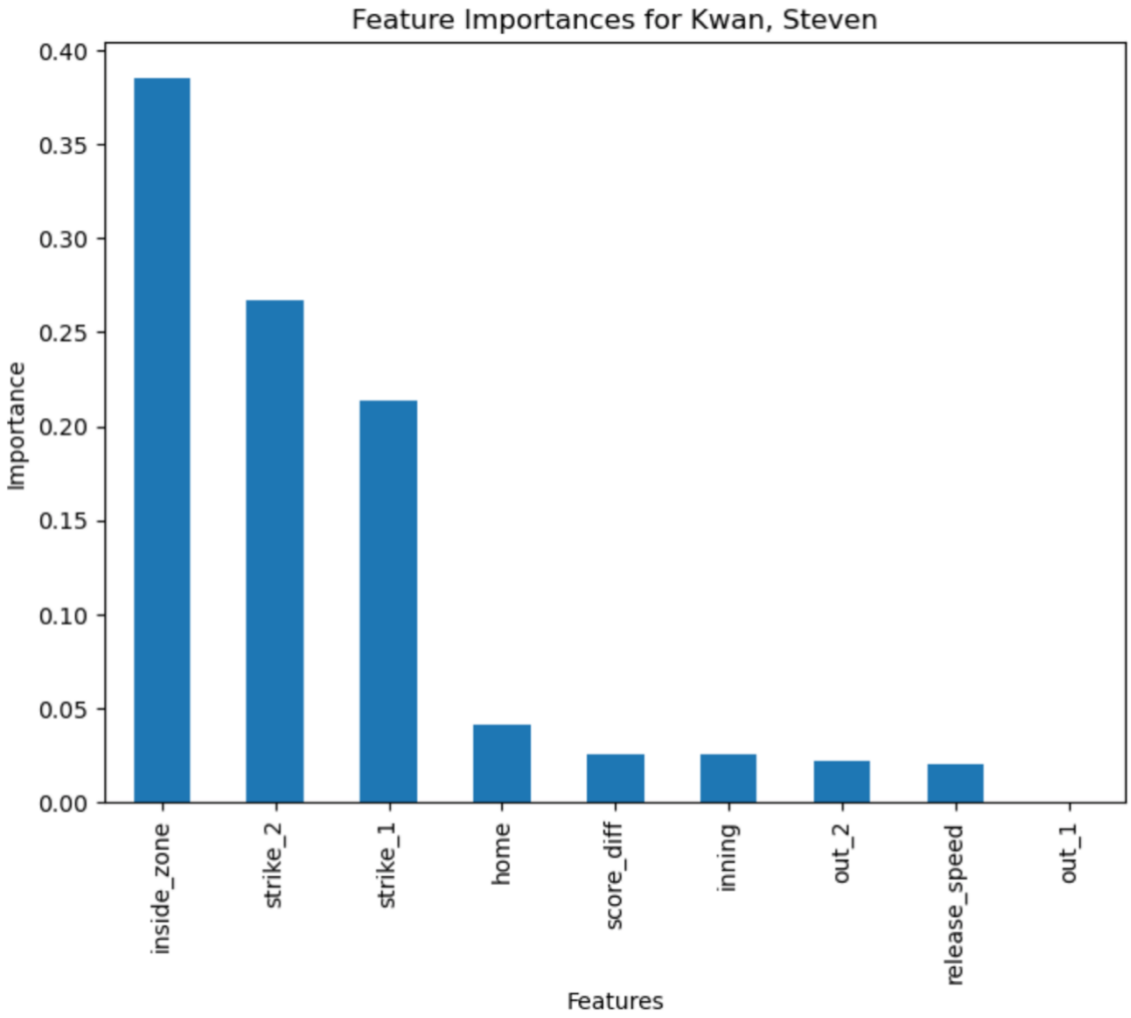


Figure 4: Feature Importance of Steven Kwan in stage 1

According to Figure 4, For Steven Kwan, it affects him more of hit the ball when the ball is thrown in the inside zone and the strikes count compared to other variables which means that Steven Kwan is sensitive to the strike count and ball thrown in inside the zone.

In the second stage, by using the same independent variable from the last section, Table 12 shows the top five accuracy of the model and its corresponding AUC score for the players excluding seven players with 100% alongside the best hyperparameters.

Table 12: Top five players in terms of the non 100% accuracy score in stage 2 with the optimal hyperparameters

| Player | Accuracy | AUC score | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|---|---|
| Rafael Devers | 0.993 | 0.993 | 0.100 | 4.000 | 100.000 |
| Joey Wendle | 0.992 | 0.500 | 0.010 | 3.000 | 50.000 |
| Manuel Margot | 0.991 | 0.470 | 0.010 | 3.000 | 50.000 |
| Carson Kelly | 0.991 | 0.913 | 0.010 | 3.000 | 150.000 |
| Raimel Tapia | 0.991 | 0.995 | 0.010 | 3.000 | 150.000 |

The prediction accuracy is around 99%. However, as I mentioned in the last paragraph, there are seven players who get 100% accuracy. This might have the following reasons: Firstly, it might have some overfitting problems. Secondly, for each pitch played, getting a hit can be hard and getting a homerun is rare, meaning that many players do not hit a lot of homeruns in the given timespan. This makes data skewed since a lot of hits are not homerun which can make prediction difficult in this case.

Despite that, we can also find the characteristic of a player when by looking at feature importance. Using the example from Mike Trout, Figure 5 shows the feature importance of the player in stage two.

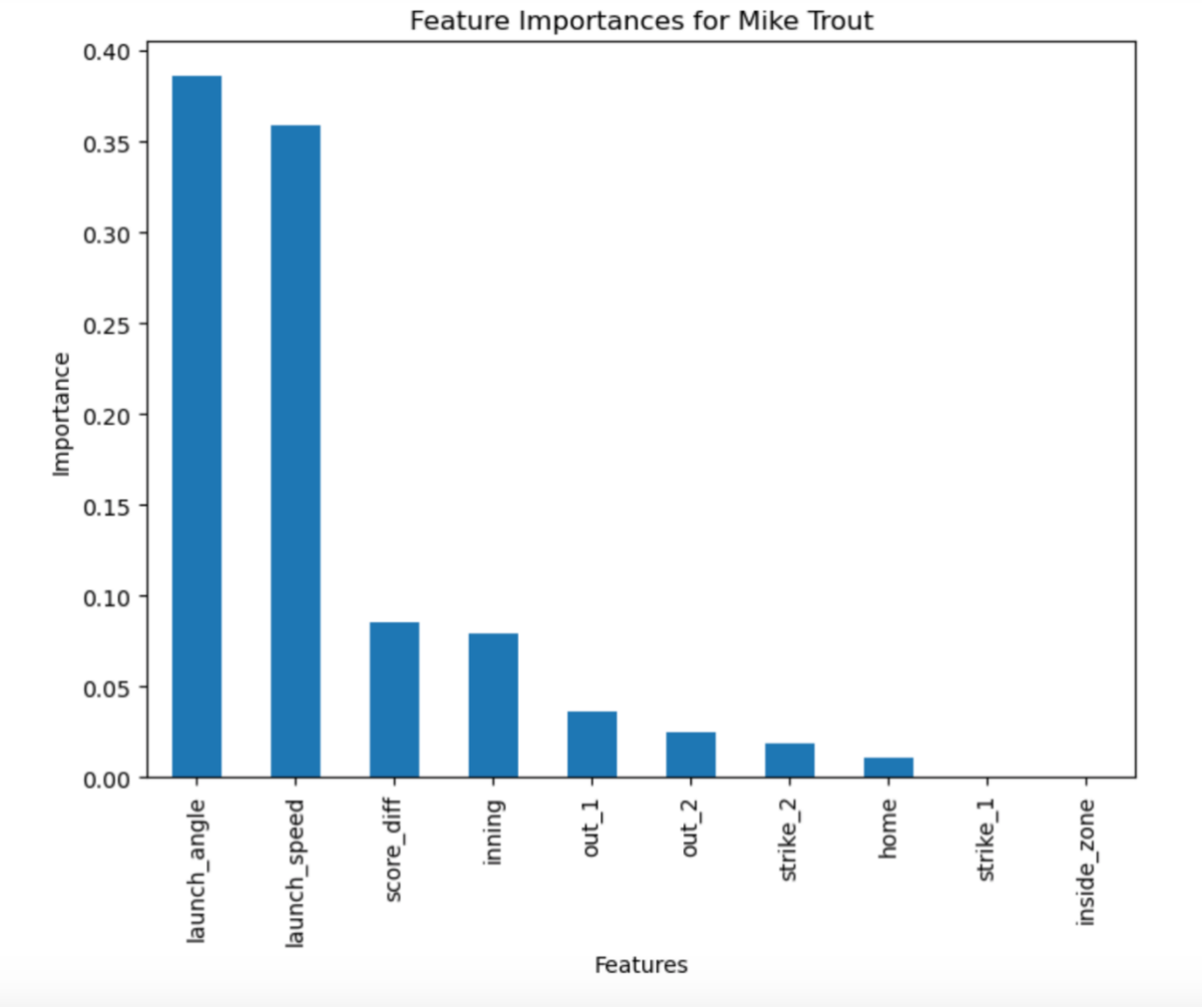


Figure 5: Feature Importance of Mike Trout in stage 2

According to the feature importance of Mike Trout, the launch speed and the launch angle played a huge part in him hitting the homerun, so do most of the other players. It can also be seen that score difference and innings can affect him for hitting a homerun. At the same time, out count, as well as strike count and playing home match has a minor effect on hitting a homerun for him. Table 13 shows the predicted top five players hitting homeruns and Table 14 shows their optimal hyperparameters.

Table 13: Predicted Top five players hitting homeruns by using XGBoost model.

| Player | Predicted Homerun | Actual Homerun | Accuracy | AUC Score |
|---|---|---|---|---|
| Aaron Judge | 27 | 20 | 0.913 | 0.935 |
| Lars Nootbaar | 21 | 9 | 0.898 | 0.883 |
| Matt Olson | 20 | 13 | 0.930 | 0.952 |
| Yordan Alvarez | 17 | 7 | 0.927 | 0.964 |
| Ryan Moutcastle | 16 | 8 | 0.910 | 0.950 |

Table 14: Optimal Hyperparameters of Predicted Top five players hitting homeruns.

| Player | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|
| Aaron Judge | 0.010 | 5.000 | 150.000 |
| Lars Nootbaar | 0.010 | 3.000 | 150.000 |
| Matt Olson | 0.010 | 3.000 | 150.000 |
| Yordan Alvarez | 0.300 | 3.000 | 50.000 |
| Ryan Moutcastle | 0.010 | 3.000 | 100.000 |

According to Table 14, compared to logistic regression, the accuracy score is higher when using XGBoost. However, a slightly big overvalue and undervalue still exists which means there is still a certain amount of misclassification. This might have the same problem with the prediction in logistic regression before such as the imbalance with the data. It is still hard to say that Lars Nootbaar has better batting performance than Matt Olson. Therefore, it should be noted that additional metrics for the performance measurement should be considered.

## Transforming Battling Outcome into Performance

After all the prediction, one question still needs to be answered is how to determine the performances from a single player. Using homeruns from a single player could be a useful indicator. However, from the descriptive data we know that hitting a homerun is difficult,

which means that it will not happen often. At the same time, some other aspect of batting performance for the individual batter might be ignored. Even though baseball has more aspects to individual performances, it cannot take away the fact that it is still a team sport and there are still some aspects to collaboration. Therefore, it is hard to find a good indicator about individual performance, especially for the batter when they are attacking. There are a few performance indicators about the performance of the batter, one of them is the batting average (BA). It can be written as $BA = \frac{H}{AB}$, which is the ratio of the hits and at-bats.

However, battling average has some drawbacks about this indicator, such as it does not indicate how powerful a batter is because it makes all the hits the same. In addition, some tactical sacrifice can also affect the statistics of BA since the batter gets fielded out in this case, which lowers their BA stats but in favor of the team result. In addition, these different performance indicators are related to plate appearance (PA), which means that only pitches played with a definite outcome will be considered. This can also be written as the following formula: $PA = 1B + 2B + 3B + HR + BB + K + HBP + SF + SH + CI + E + DFO$. This includes all the outcomes like all the hits, (single (1B), double (2B), triple (3B) and homerun (HR)), all the outs (strike out(K) and defensive out (DFO)), all the batter sacrifices (Sacrifice fly (SF) and Sacrifice bunt (SH)) and all the errors (Walk (BB), Hit by pitch (HBP), Catcher Interference (CI) and fielding error(E)). From the plate appearance we can also get the at-bat statistics, which are the plate appearances excluding sacrifices of batters and errors from the pitchers. The at-bat statistics can also be written as the following formula: $AB = PA - BB - HBP - SH - SF - CI.$

In this case, some indicators, such as slugging percentage (SLG) and on-base percentage (OBP) are a better indicator. SLG is calculated by all the different hits divided by at-bats, where different hits are weighted differently. It can be written in the following formula: $SLG = \frac{1B+2\times 2B+3\times 3B+4\times HR}{AB}$, in which homerun gets the highest weight on that hit.

OBP evaluates how often the batter becomes a runner on the base. It can be calculated by the sum of all the hits, walk and hit by the pitch divided by the sum of at-bats, walk, hit by the pitch, and sacrifice fly. The formula can be written as follows: $OBP = \frac{H+BB+HBP}{AB+BB+HBP+SF}$.

As we can see, SLG gives more weight when a batter gives a better hit and OBP gives an overview of how frequently a batter goes to the base. In this situation, I will choose SLG as the indicator of performance measurement from the individual because SLG focused more on the batting perspectives for batters. Previously, I have predicted homerun, which is part of the SLG indicator. Therefore, in this section, I will also predict single, double, triple homerun and bats get field out for each player all together by using the XGBoost method. This means that it can transform the prediction for the batting outcome to the prediction of the batter performance indirectly after I predict all the outcomes for the players. Once I know how many single, double, triple and homerun for each player in predicted value, I can calculate the predicted value of SLG for each player. In addition, since in this case the plate appearance data are more important. Therefore, the data processing in this case will also be different from the previous section, where the pitch data with definite result will be selected since pitch data in this situation will not be suitable.

After the data corresponding data got selected, there are still a few things to explain in the prediction of batting outcome. Firstly, events that grant a batter automatically onto the first base because of the fault of the pitcher, such as hit by pitch or walk should be excluded in the batting prediction because the batter does not get a bat. Secondly, if the team decided to sacrifice the batter by different means, these data will not be included in the predicting the outcome of a bat because these calls are mostly tactical and will not hinder the batting performance from a batter. Thirdly, if events are related to the stolen base, then the data will not be included in the prediction part of batting outcome since the batter is the runner in this case and has nothing to do with batting. In the end, I will not include the strike out data in this case since the characteristics of strike out are mixed. On the one hand, it is about the performance of the pitchers and not so much about the batter. On the other hand, some failed tactical sacrifices can be noted as strike out since these bats are bunts,

which is not the indication of the batters' performance. This setting is also consistent with the setting of Burch (2020) has stated.

However, these excluded data will still be included into the final calculations related to the statistics with plate appearances and at-bat data to calculate the SLG indicator for each player if they are inside the calculation of the plate appearance and at-bat. To summarize this, the data available for doing the prediction of the batting outcome will include the events with single, double, triple, homerun, and field out and the final predicted SLG statistics calculation will include the predicted single, double, triple and homerun and the initial at bat and plate appearances data. In this case, I will still using the gradient boosting method, but here it will be only single stage because here only the plate appearance will be considered, and incomplete pitch data, which pitches does not give a definite outcome, are not inside the calculation of SLG indicator. In addition, I will also compare the result with training all the data altogether.

Therefore, I obtained the following result based on the previous paragraph. Table 15 shows the overall impression for the actual data and predicted data when all the data are trained together in the testing data set. In this case, the optimal hyperparameters are Learning Rate = 0.1, Maximum Depth = 5 and Number of estimators = 100

Table 15: Overall description from actual and predicted data

| | Actual | Predicted |
|---|---|---|
| Single | 7925 | 7451 |
| Double | 2395 | 729 |
| Triple | 171 | 0 |
| Home Run | 1631 | 1470 |
| Field Out | 22255 | 24610 |
| Accuracy Score | / | 0.754 |

In this situation, we can see that the overall accuracy score in this case is 75.4%. However, it predicts less doubles and more outs than the actual and slightly less singles and homeruns. In addition, the model did not predict any triples.

Table 16 shows the top five predicted SLG indicators and its actual SLG indicator from the test dataset.

Table 16: the top five player in predicted SLG indicators with the actual SLG data (Trained Together)

| Player | Predicted | Actual |
|---|---|---|
| Bryan De La Cruz | 0.725 | 0.725 |
| Aaron Judge | 0.682 | 0.679 |
| Yordan Alvarez | 0.601 | 0.595 |
| Mike Trout | 0.574 | 0.574 |
| Jesus Sanchez | 0.563 | 0.563 |

However, under MLB official rules (2022), a batter must record 502 PA in the whole regular season to be eligible to be in the statistical rankings for the corresponding award. In other words, players who have lower than 502 PA numbers still get the statistics, but with no rankings since they are ineligible. In this case, since my test data contains only the data from the beginning of August till the end of first week of October and usually every team has around 60 matches to play afterwards. Therefore, the required PA number will be adjusted to 186 PA. Table 17 presented the adjusted top five player in predicted SLG indicators with the SLG data in accordance with the official MLB rules.

Table 17: the top five player in predicted SLG indicators with the actual SLG data adjusted with minimum PA requirement (Trained Data Together)

| Player | Predicted | Actual |
|---|---|---|
| Aaron Judge | 0.682 | 0.679 |
| Yordan Alvarez | 0.601 | 0.595 |
| Ryan Mountcastle | 0.560 | 0.560 |
| Teoscar Hernandez | 0.535 | 0.535 |
| Kyle Schwarber | 0.506 | 0.504 |

According to Table 17, after adjusted with the minimum PA requirement imposed, we can say that in this case, under the circumstances that all the data are trained together, Aaron Judge will be the best batter in terms of SLG indicator with the predicted value of 0.682, followed by Yordan Alvarez and Ryan Mountcastle, who has the predicted value of 0.601 and 0.560, respectively.

In the following part, I will show the result where the data are trained individually. It is also interesting to see the accuracy score from the players and Table 18 and Table 19 shows the top and bottom five accuracy of different players alongside the optimal hyperparameters.

Table 18: Top five players in terms of the accuracy score

| Player | Accuracy | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|---|
| Kole Calhoun | 0.848 | 0.010 | 4.000 | 50.000 |
| Brandon Lowe | 0.813 | 0.010 | 5.000 | 150.000 |
| Connor Joe | 0.811 | 0.100 | 3.000 | 150.000 |
| Josh H. Smith | 0.809 | 0.010 | 3.000 | 50.000 |
| Marwin Gonzalez | 0.800 | 0.300 | 5.000 | 100.000 |

Table 19: Bottom five players in terms of the accuracy score

| Player | Accuracy | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|---|
| Sergio Alcantara | 0.460 | 0.100 | 4.000 | 150.000 |
| Jesus Sanchez | 0.500 | 0.010 | 4.000 | 50.000 |
| Bobby Dalbec | 0.518 | 0.300 | 5.000 | 200.000 |
| Rodolfo Castro | 0.530 | 0.010 | 3.000 | 50.000 |
| Trevor Story | 0.535 | 0.100 | 3.000 | 100.000 |

Here, we can see that the accuracy from different players is different and the predicted accuracy ranges across the players. Therefore, it can potentially underestimate or overestimate the stats. We can also use the minimum requirement of PA to select the players in the ranking. Table 20 and Table 21 show the top and bottom accuracy score after adjusted and Figure 8 is the histogram of the accuracy number distribution.

Table 20: Top five players in terms of the accuracy score with minimum requirement of PA

| Player | Accuracy | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|---|
| Rowdy Tellez | 0.800 | 0.010 | 4.000 | 50.000 |
| Taylor Ward | 0.789 | 0.010 | 4.000 | 200.000 |
| Rhys Hoskins | 0.785 | 0.010 | 5.000 | 150.000 |
| Mike Yastrezemski | 0.771 | 0.010 | 5.000 | 50.000 |
| Alejandro Kirk | 0.762 | 0.010 | 3.000 | 50.000 |

Table 21: Bottom five players in terms of the accuracy score with minimum requirement of PA

| Player | Accuracy | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|---|
| Rodolfo Castro | 0.530 | 0.010 | 3.000 | 50.000 |
| Paul Goldschmidt | 0.578 | 0.010 | 3.000 | 50.000 |
| Jake Fraley | 0.580 | 0.100 | 3.000 | 150.000 |
| Adolis Garcia | 0.582 | 0.010 | 4.000 | 50.000 |
| Eloy Jimenez | 0.583 | 0.010 | 4.000 | 100.000 |

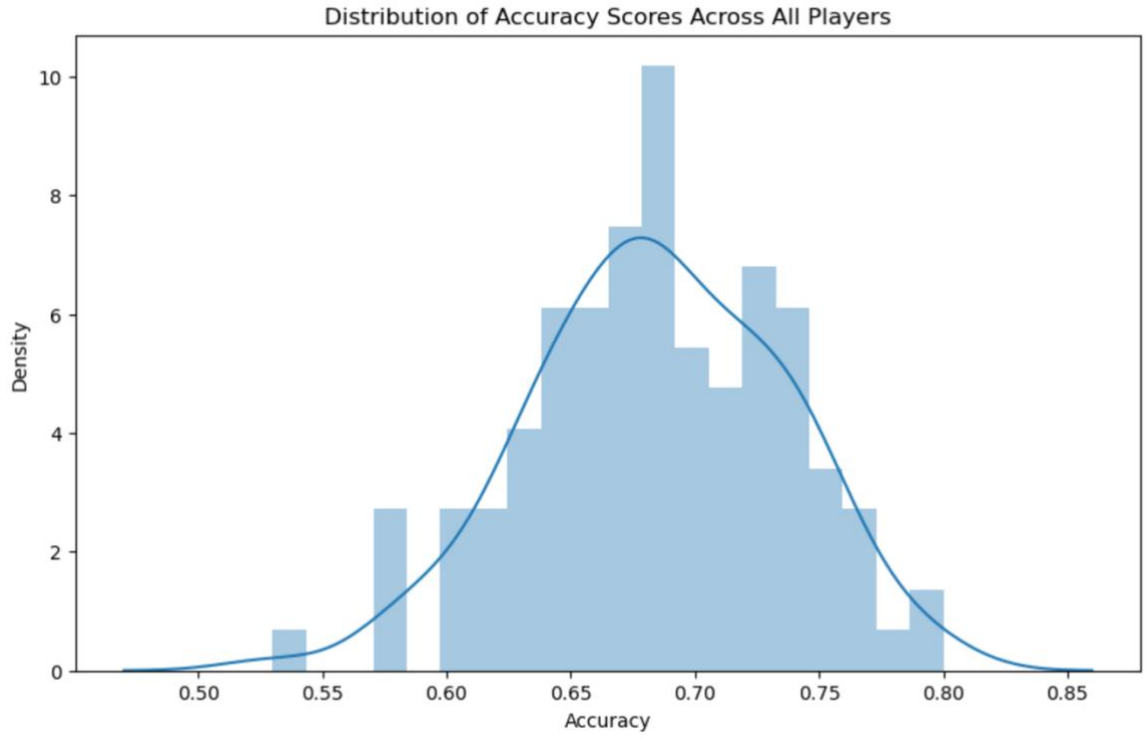


Figure 6: Histogram of prediction accuracy number distribution with the minimum requirement of PA.

From Table 18 and Table 19, under the minimum PA rule, the range of predicted accuracy number across the players is 0.270, which is smaller than without the minimum PA rule of

0.313. According to the histogram from Figure 6, most players' predicted accuracy is around 0.65 to 0.75, meaning that there can be some comparison of the batting performance from different players. However, it will still have some overestimation and underestimations.

I can also check the feature importance from these players about their characteristics, in this situation, I will compare Nick Castellanos (Accuracy value of 0.759) and Mookie Betts (Accuracy value of 0.746) and Figure 7 and 8 shows their feature importance, respectively.

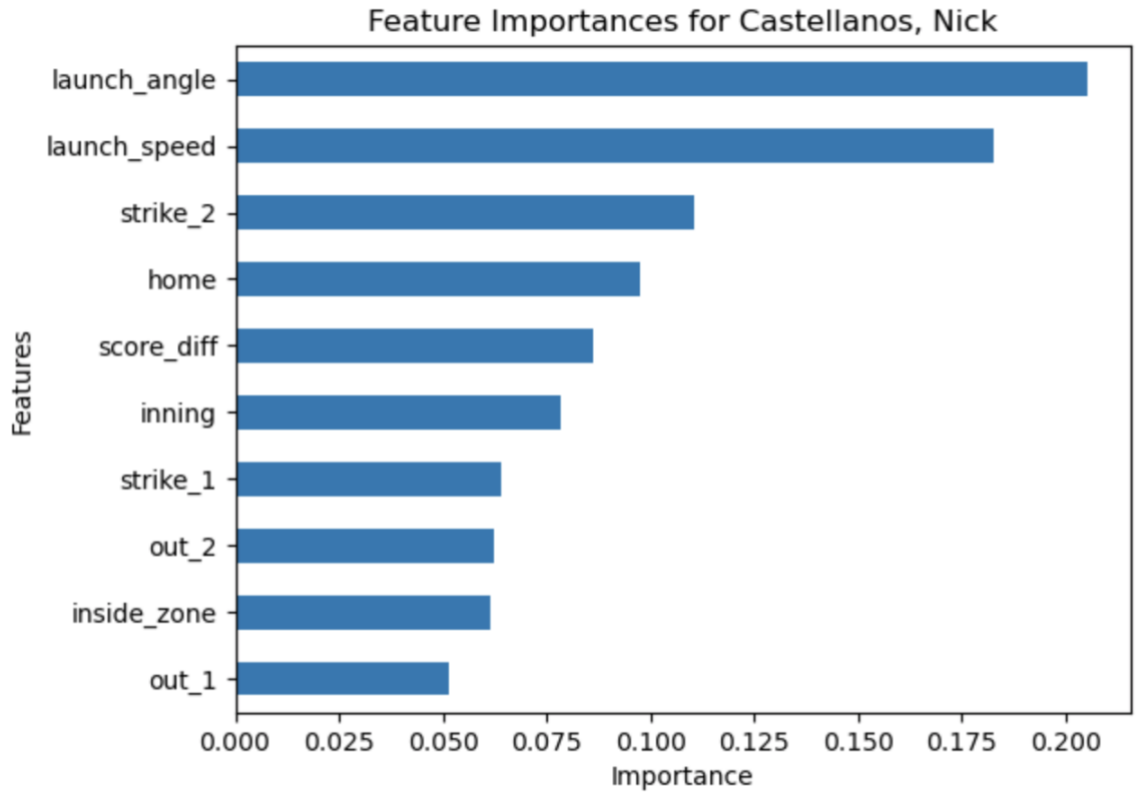


Figure 7: Feature Importance of Nick Castellanos of batting

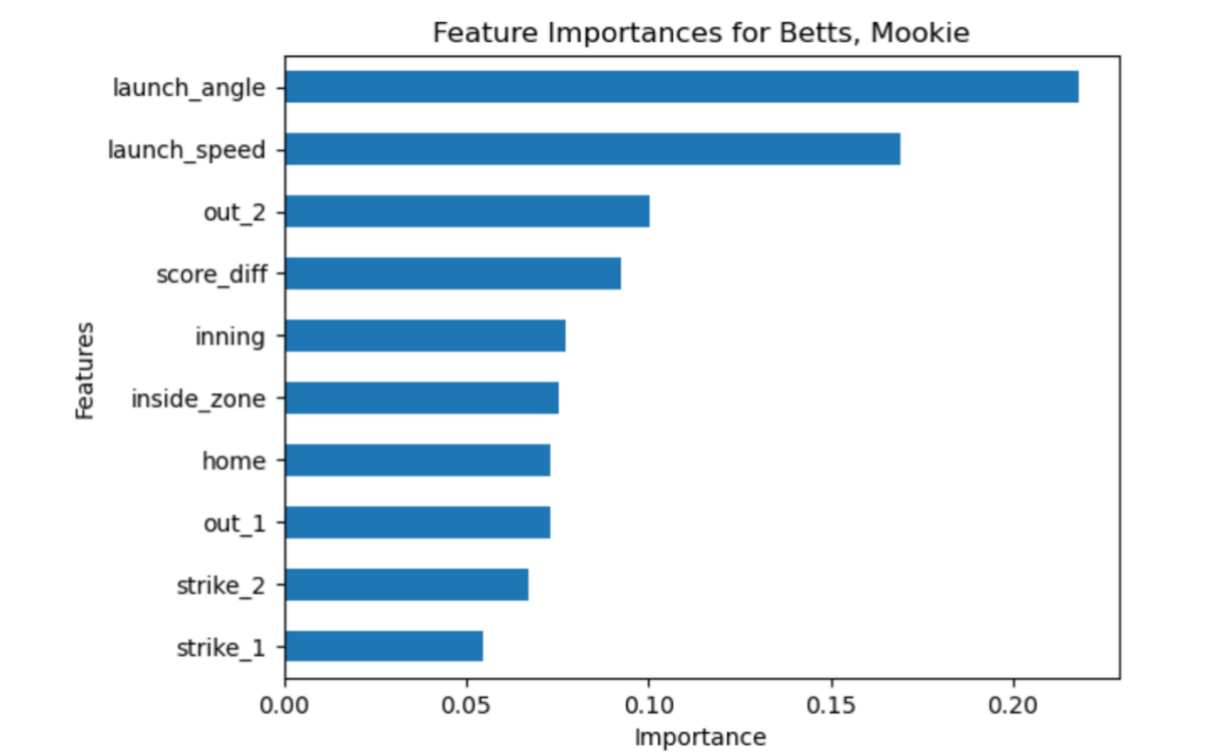


Figure 8: Feature Importance of Mookie Betts of batting

According to both Figure 7 and Figure 8, both players' batting performance are sensitive to the launch speed and launch angle. For Mookie Betts's, his batting performance is more sensitive to having two outs in the inning and difference of the score than Nick Castellanos. For Nick Castellanos however, his batting performance is more sensitive in the and play home games than Mookie Betts' batting performance.

In this case, I will also calculate the SLG statistics for each player after the minimum PA requirement is imposed. Table 22 shows the top 5 players with their corresponding SLG indicators and Table 23 shows SLG statistics from Nick Castellanos and Mookie Betts with their corresponding optimal hyperparameters.

Table 22: the top five player in predicted SLG indicators with the actual SLG data adjusted with minimum PA requirement (Trained Data Individually)

| Player | Predicted | Actual | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|---|---|
| Aaron Judge | 0.682 | 0.679 | 0.010 | 3.000 | 150.000 |
| Yordan Alvarez | 0.587 | 0.595 | 0.010 | 3.000 | 200.000 |
| Paul Goldschmidt | 0.560 | 0.442 | 0.010 | 3.000 | 50.000 |
| Matt Olson | 0.538 | 0.480 | 0.010 | 3.000 | 200.000 |
| Nathaniel Lowe | 0.528 | 0.444 | 0.010 | 3.000 | 200.000 |

Table 23: predicted SLG indicators with the actual SLG data of Mookie Betts and Nick Castellanos

| Player | Predicted | Actual | Optimal Learning Rate | Optimal Maximum depth | Optimal Number of Estimators |
|---|---|---|---|---|---|
| Mookie Betts | 0.420 | 0.302 | 0.100 | 5.000 | 200.000 |
| Nick Castellanos | 0.281 | 0.297 | 0.010 | 5.000 | 200.000 |

According to Table 23, we can say that Aaron Judge and Yordan Alvarez rank first and second. the predicted value of SLG indicator when the data is trained individually. However, rank three and five are more overestimated in this case, with rank four slightly overestimated. This can also be seen that in Table 24, where the predicted value of Mookie Betts is overestimated, and Nick Castellanos is slightly underestimated. To understand this, I will use the confusion matrix to find the misclassification with Mookie Betts, Matt Olson, Nathaniel Lowe, and Paul Goldschmidt. Figure 9 shows the confusion matrix for all four players.

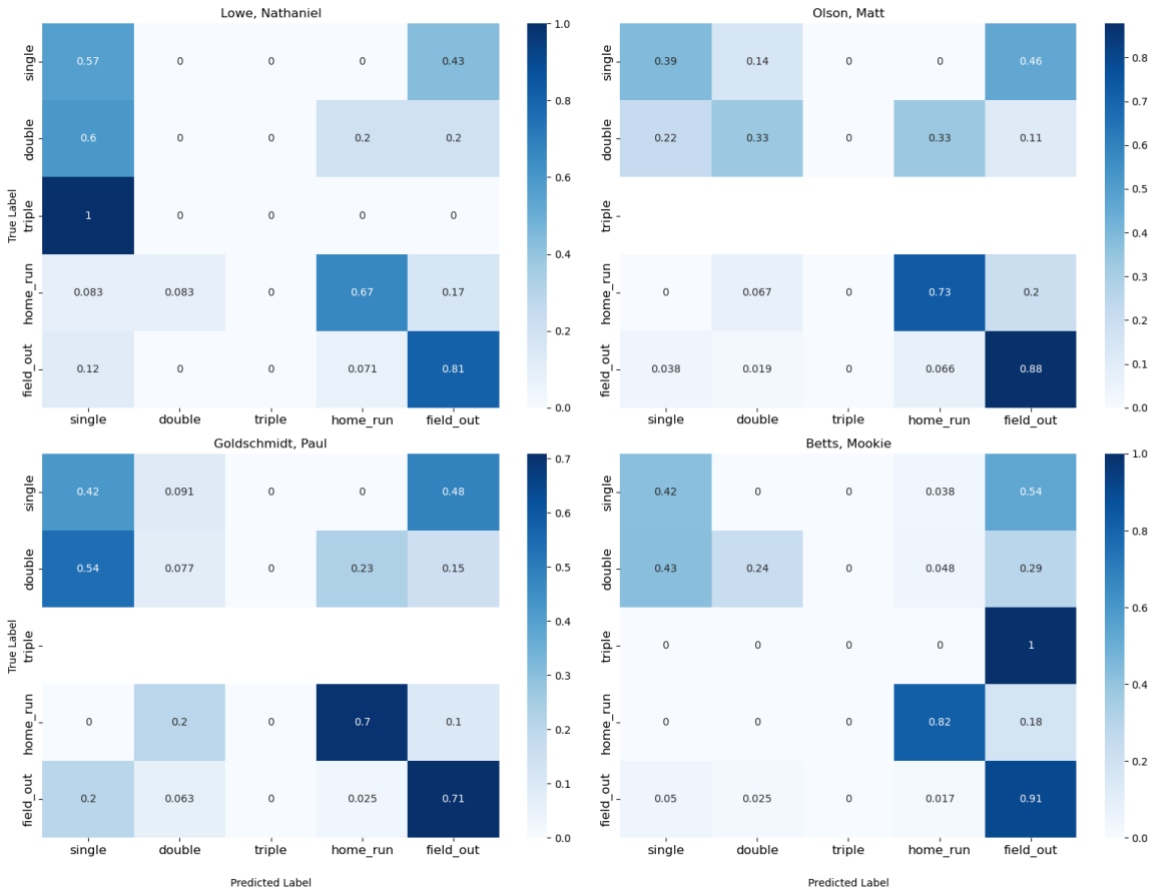


Figure 9: Confusion Matrices for given four players.

From the confusion matrices, all four players identify outs well and for homerun, Nathaniel Lowe is slightly less accurate compared to other three players. However, there are some miss classifications here. For example, from the model of Paul Goldschmidt and Nathaniel Lowe, a lot of doubles get misclassified as homerun. In addition, all four players have a fair number of singles misclassified as outs. Therefore, because the formula of SLG weights the most on homerun, this means that the predicted SLG value will be overstated. This will be the same for all the players. If some hits are misclassified to homerun, the final SLG statistics will be overestimated. This problem arises because of the following potential reasons and one of them is event though the minimum PA rule has enforced in this case, there is not enough data point for each player, which increases the chances of over and underestimation of the players performances.

To summarize, I will say that it might be useful to predict the batter's performances by transforming the batting outcome into SLG indicators. This can be either done by training the Data individually and altogether since in the end, the SLG indicators will be an individual

indicator for both methods and people can potentially use that for understanding the performance of an individual player and identifying which player are strong in terms of their batting performances. From the outcome prediction result, we can see that if we train the data altogether, the final prediction of the SLG indicator will be in line with the actual SLG value. It can have some slight under and overestimation for players. However, when training the data altogether, we cannot get the insight of individual specific characteristics related to their batting outcome, especially some specific in game scenarios. On the contrary, training the data individually can give me more insights about specific characteristics for specific related to their batting. For instance, it can see if the players are sensitive for their batting performance in some scenarios such as innings and strike numbers which is helpful for the team. However, in terms of prediction, the model can behave differently across the players which can make the performance comparison difficult, but possible. This is because there is way less training data input compared to training all the data together and everyone has different PA data. At the same time, this problem can be minimized if we use more data points related to the plate appearance for each player, such as two- or three-years regular season plate appearance data. In all, it can also potentially be an alternative to understanding the hitting outcome characteristics for each player and their performance predictions.

## Discussions and Conclusion

To conclude, this paper shows that for the homeruns, the outcome is more determined by the kinetic characteristics, such as the launch speed and launch angle. Other characteristics, such as the strikes count and out counts are not important for homeruns, for most of the players. However, having a homerun is rare and it can be difficult to do the prediction. For the general hits, not only the kinetic characteristics, but also the in-game scenario can play a certain role for batters' batting outcome. For the performance indicators, it can be predicted by training the data altogether or individually. However, by training the data individually the model will behave differently across the players, and it can have larger variations, overestimation and underestimation than train the data all together. At the same time, it gives the individual insights of each player about the batting outcome characteristics

which cannot be sufficiently provided when training the data altogether. Managers can use both ways to keep in check the players' performance and/or understanding the player's characteristics to improve their players.

Homerun, or even other hits are an integral part of the baseball game. It is worthwhile to note that when a player hits a homerun, it can have a certain effect on the income from the players. For the players, as Dollar (2015) states, for every home run appeared, it is on average worth 45572 dollars in their salaries. Therefore, for the players they can theoretically increases their personal income by trying to hit more homeruns. It also influences the team side. By hitting the homerun, it gives the fans interaction and more involvement into the game since the fans can catch the homerun ball and keep it for themselves. The franchise can also use that as their advertising to retain and gain fans by using video footage on social media. This can increase the interaction with the fans and the team. It also leads to better home results for the team. In the research of Smith and Groetzinger(2010), they found out if more fans attend the match in the stadium, it increases various stats for the home team which leads to higher chance of winning the home match. In addition, there are also some social effects on hitting the homerun since the ball itself can create a huge amount of value. According to this report of Fox 4 (2022), Barry Bonds 73rd homerun ball was sold in an auction with an amount of 517500 dollar which can add more "flavor" into the sport.

Besides homerun, it is also vital to predict other batting outcome, such as single, double, and triple since it is hard to evaluate a batter's performance sorely on homerun because it happens not often, even if it is a good individual indicator about batting. From the individual point of view, other batting outcomes can be essential to determine batters' batting abilities by transforming predicted outcomes into predicted performance indicators such as BA, SLG and OBP indicators (In this paper SLG indicator). In addition, creating runs will be beneficial since the player can potentially score a run which helps the team.

In a perspective of team itself, scoring the points, putting the batter onto the base is more important to score as many points as possible. Therefore, evaluating the performance with other batting outcome is necessary since you need to put the player onto the base to score

points and if the batter is performing well in their batting performance, they can also create some extra runs with help the team to win the game. A great example is some singles can create more than one run scored which is favorable for the team to get good result. In addition, according to Einolf (2004) who compares the structure of the income between National Football League (NFL) and MLB. He concludes that for MLB, winning is everything to make the team profitable and get higher overall value of the team because the contract of a baseball player is longer and larger which means to attract great players for getting better team result. Otherwise, there is no point in investing such a high amount into signing valuable players. In other words, the best marketing for the MLB teams is to get better team results, with certain helps of excellent players.

However, this research also has some limitations. In the model perspective. When training the data for everyone, the model behaves differently for each player. This means it can understate or overstate the player's performance in predicting the batting outcome. In addition, I only used one-year regular season data with the test data to be post transfer deadline and the training data to be before the transfer deadline. This can have some problems such as not enough plate appearance data for doing the training and the prediction for some individuals. In addition, I also observe that the model cannot evaluate some players and do the prediction for them. This might have two explanations: first, the player hit an outcome in the test dataset, such as triple. However, in the training set they hit zero triples. This means that the model cannot identify triple which is problematic. Another explanation for this is that since the training and testing data set is split by time, some players play less after the spilt time due to some reasons, such as injury, which can also be a problem because it gives less data as usual, make it harder to predict the outcome.

On the other hand, once batters launch the ball and then run to the base or run to another base, they also become the runner. In this situation, the running speed of batters can also influence the hitting outcome since in some cases, if a runner has a high running speed and with a decent hit, a homerun can be theoretically achieved without the ball going over the fence. This has an implication in general hit prediction. This can give more insight into the run creation, which is an important part of a team. In this case, it is also important to check for batters about their run creation because in some situations, even though the batter was

out, the team can still score a run or place one of the players into a favorable base. In the end, it is still a team sport, and the aim is to win the game. Therefore, some aspects from the game, such as the sequence of the batter batting and the tactical calls in the game are interesting to check, such as when to sacrifice the batter to achieve optimal result.

Besides performance, some other topics, such as players' contract value relative to their performance can also give some insights about players whether their performance level is worthwhile for certain amount of contract. This can also be a noteworthy part to explore if certain players are undervalued or overvalued based on their performance.

In addition, this paper is mainly about the batter performance in batting from the game play point of view. It also needs to be complimented from the sports science or life science point of view from the batter in the training to improve their batting in training, such as checking the corresponding muscle for batting and other indicators of athleticism.

## References


Albert, J. (2014). Sabermetrics. *Wiley StatsRef: Statistics Reference Online*, 1-6.

Barrow, D., Drayer, I., Elliott, P., Gaut, G., & Osting, B. (2013). Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, *9*(2), 187-202.

Baumeister, R. F., & Showers, C. J. (1986). A review of paradoxical performance effects: Choking under pressure in sports and mental tests. *European Journal of Social Psychology*, *16*(4), 361-383.

Beneventano, P., Berger, P. D., & Weinberg, B. D. (2012). Predicting run production and run prevention in baseball: the impact of Sabermetrics. *Int J Bus Humanit Technol*, *2*(4), 67-75.

Bock, J. R. (2015). Pitch sequence complexity and long-term pitcher performance. *Sports*, *3*(1), 40-55.

Burch, T.J., (2020) *Classifying MLB Hit Outcomes - Part 1: Model Selection.* Retrieved from Taylor James Burch: http://tylerjamesburch.com/blog/baseball/hit-classifier-1

Burroughs, B. (2020). Statistics and baseball fandom: Sabermetric infrastructure of expertise. *Games and Culture*, *15*(3), 248-265.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chu, D. P., & Wang, C. W. (2019). Empirical study on relationship between sports analytics and success in regular season and postseason in Major League Baseball. Journal of Sports Analytics, 5(3), 205-222.

Clark, J. F., Ellis, J. K., Bench, J., Khoury, J., & Graman, P. (2012). High-performance vision training improves batting statistics for University of Cincinnati baseball players. *PloS one*, *7*(1), e29109.

Courneya, K. S., & Chelladurai, P. (1991). A Model of Performance Measures in Baseball. *Journal of Sport & Exercise Psychology*, *13*(1).

Cross, R. (1998). The sweet spot of a baseball bat. American Journal of Physics, 66(9), 772-779.

Dollar, S., (2015) *Modeling Salary Arbitration: Stat Components*. Retrieved from FanGraphs: https://blogs.fangraphs.com/modeling-salary-arbitration-stat-components/

Einolf, K. W. (2004). Is winning everything? A data envelopment analysis of Major League Baseball and the National Football League. *Journal of Sports Economics*, *5*(2), 127-151.

Farrell, B. (2019). Machine Learning Algorithm for Predicting Major League Baseball Team Wins.

Fox 4 (2022). *How much is Aaron Judge's 62nd home run ball worth? ?* Retrieved from Fox 4 News: https://www.fox4news.com/news/how-much-is-aaron-judges-62nd-home-run-ball-worth

Ganeshapillai, G., & Guttag, J. (2012). Predicting the next pitch. In *Sloan Sports Analytics Conference*.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R. Spinger.*

Gartheeban, G., & Guttag, J. (2013). A data-driven method for in-game decision making in mlb: when to pull a starting pitcher. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 973-979).

Gerrard, W. J. (2016). Beyond Moneyball: Using Data Analytics to Improve Performance in Elite Team Sports. *Sport and Entertainment Review*, *2*(1).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. *The elements of statistical learning: data mining, inference, and prediction*, 337-387.

Healey, G. (2015). Modeling the probability of a strikeout for a batter/pitcher matchup. *IEEE Transactions on Knowledge and Data Engineering*, *27*(9), 2415-2423.

Healey, G. (2017). Matchup models for the probability of a ground ball and a ground ball hit. *Journal of Sports Analytics*, *3*(1), 21-35.

Hoang, P. (2015). *Supervised learning in baseball pitch prediction and Hepatitis C Diagnosis*. North Carolina State University.

James, B. (2010). *The new Bill James historical baseball abstract*. Simon and Schuster.

Koseler, K., & Stephan, M. (2017). Machine learning applications in baseball: A systematic literature review. *Applied Artificial Intelligence*, *31*(9-10), 745-763.

Lindsey, G. R. (1963). An investigation of strategies in baseball. *Operations Research*, *11*(4), 477-501.

McCracken, V. (2001). Pitching and defense: How much control do hurlers have?. *Baseball Prospectus*, *23*.

Middleton, J., Murphy-Hill, E., & Stolee, K. T. (2020). Data analysts and their software practices: A profile of the sabermetrics community and beyond. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1-27.

Mizels, J., Erickson, B., & Chalmers, P. (2022). Current state of data and analytics research in baseball. *Current reviews in musculoskeletal medicine*, *15*(4), 283-290.

MLB (2022): *Official Baseball Rules*. Retrieved from MLB Official information: https://img.mlbstatic.com/mlb-images/image/upload/mlb/hhvryxqioipb87os1puw.pdf

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, *5*, 213-222.

Nakahara, H., Takeda, K., & Fujii, K. (2023). Pitching strategy evaluation via stratified analysis using propensity score. *Journal of Quantitative Analysis in Sports*, (0).

Newton, I. (1833). *Philosophiae naturalis principia mathematica* (Vol. 1). G. Brookman.

Petersen, A. M., Jung, W. S., & Stanley, H. E. (2008). On the distribution of career longevity and the evolution of home-run prowess in professional baseball. *Europhysics Letters*, *83*(5), 50010.

Puerzer, R. J. (2002). From scientific baseball to sabermetrics: Professional baseball as a reflection of engineering and management in society. *NINE: A Journal of Baseball History and Culture*, *11*(1), 34-48.

Sarlis, V., & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, *93*, 101562.

Sawicki, G. S., Hubbard, M., & Stronge, W. J. (2003). How to hit home runs: Optimum baseball bat swing parameters for maximum range trajectories. *American Journal of Physics*, *71*(11), 1152-1162.

Smith, E. E., & Groetzinger, J. D. (2010). Do fans matter? The effect of attendance on the outcomes of major league baseball games. *Journal of Quantitative Analysis in Sports*, *6*(1).

Valero, C. S. (2016). Predicting Win-Loss outcomes in MLB regular season games–A comparative study using data mining methods. *International Journal of Computer Science in Sport*, *15*(2), 91-112.

Yang, T. Y., & Swartz, T. (2004). A two-stage Bayesian model for predicting winners in major league baseball. *Journal of Data Science*, *2*(1), 61-73.