Case-Study: Predicting Batter Performance Based on Pitcher Matchup From the 2023 MLB Season

Sean P. Finlon

Department of Mathematics & Statistics, University of Wisconsin-La Crosse

DS 785: Capstone

Prof. Tracy Bibelnieks

December 10, 2023

Abstract

The ability to predict the performance of individual matchups between Major League Baseball (MLB) batters and pitchers can be extremely useful information to have for team general managers when building a roster, coaches when setting a lineup or making in-game decisions, and even sportsbooks when placing odds on an event. With the recent introduction of the wealth of sensor data thanks to Statcast, previously unmeasured attributes of the game are now archived and allow for analysis for further insights to be understood. The development of expected statistics, such as the expected weighted on-base average (xwOBA) of a batted ball given its exit velocity and launch angle, is now possible with this data, which have proved to be more reflective of performance value. This study aims to predict the xwOBA of a batter for a specific pitcher matchup from the 2023 MLB season, while also identifying the most impactful factors regarding pitchers and their pitch behavior for batter performance. According to the neural network model built for the study, the most important variables include the pitch position when crossing home plate and the vertical pitch movement, while other traditionally assumed high-impact variables, such as handedness and pitch velocity, are not found to be as comparatively impactful.

Keywords: baseball, MLB, batter performance, batter matchup, Statcast, xwOBA, neural network

Table of Contents

List of Tables

List of Figures

Introduction

Ever since the breakout novel *Moneyball: The Art of Winning an Unfair Game* by Michael Lewis recounted the groundbreaking analytics strategy of the 2002 Oakland Athletics led by their General Manager Billy Beane, organizations around Major League Baseball (MLB) have been in an analytics arms race to maximize their talent per dollar spent. The collection of in-game data is the fuel for these behind-the-scenes evaluation advancements, and they saw a major upgrade in the 2015 season when Statcast was installed in all MLB ballparks. Statcast is a system using Doppler radar and stereoscopic video to track the movement of the baseball and each player for every play of every game throughout the season to gather up to a massive seven terabytes of data per game (Healey, 2017). This new wealth of sensor data available to organizations, along with portions available to the public, has allowed developments in player evaluation that were impossible a mere couple of decades ago.

Background

In the pre-analytics and sensor data MLB world, batter performance was evaluated by a limited number of stats recorded during a game and the inherently biased perceived skill level assigned by management or scouts. The traditional stats recorded during a game, like the outcome of a plate appearance being a single, flyout, or strikeout, for example, can be dependent on many other variables historically untracked, such as defensive skill, pitch movement, or even defensive positioning luck (Healey, 2017). The majority of these stats were outcome-based and rarely considered the cause and repeatability of an event. To better illustrate the flaws of the outcome-based traditional stats, consider two scenarios: a batted ball that was hit extremely hard into the outfield gap that resulted in a double versus a batted ball that was hit very softly but fell perfectly between three fielders on the opposing team that resulted in a double. Both hits had the

same result in the box score, but the first scenario of the hard-hit ball would result in a double far more often than the softly hit, yet perfectly placed, ball if the event was repeated. From a player performance evaluation perspective, if one did not witness the two hits live and only had the box score to analyze, the evaluator may rate them the same while in reality the first scenario should be valued higher than the second. The granular sensor data provided by the Statcast system could theoretically solve exactly this problem and now allows the incorporation of many of these previously untracked variables in analytical studies to gain a greater understanding of player performance.

Purpose

One aspect that is somewhat unique to the game of baseball is the isolated individual matchup possibilities between two athletes: a pitcher and a batter. Other popular team sports that are currently popular in America, like American football, basketball, or ice hockey, are extremely fluid, and it can be difficult to isolate the impact of an individual athlete accurately. The segmented nature of baseball, along with not having a running clock, allows for the ability to gather very consistent data across a season and even years.

Throughout baseball history, it has not been uncommon to have a manager set a unique lineup for many games straight, dependent on the opposing team's personnel. Being a baseball manager is essentially a large optimization problem, in that they are attempting to find the combination of resources to produce the best outcome possible, and organizations are desperately trying to discover insights to have an edge on their opponents. Previously, studies have used traditional stats and handedness to predict batter performance based on pitcher matchup or to predict a specific event outcome, like a strikeout, but few have included the newly introduced variables now gathered and provided by Statcast (Doo & Kim, 2018; Healey, 2015).

The purpose of this project is to incorporate the Statcast data into a model produced using a sophisticated data science technique to predict batter performance in the MLB based on the pitcher's profile they are matched up against.

Objectives

The objectives of this project include:

1. Determine which model-building and validation techniques are appropriate for predicting batter performance based on pitcher matchup.
2. Analyze various models for accuracy and performance to identify the most effective model.
3. Use the model built to predict batter performance based on pitching profile.
4. Identify the most important variables within the model that best describe the performance of batters.

The scope of this project was framed from an individual batter versus pitcher scenario, but this concept has the potential to be expanded upon in the future.

Significance

This project could be significant for MLB teams to gain an edge in a competitive league, as previously mentioned, by allowing organizations to place batters and pitchers in beneficial situations. It could assist in assembling a roster as well, potentially identifying favorable matchups to fill weaknesses or identifying the weaknesses themselves on a team. It could even be a potential selling point for a player during free agency or seeking an extension if it highlights a specialty of theirs that may not be obviously apparent. The predicted performance of batters based on pitcher matchups could also benefit the currently booming industry of sportsbooks, and similarly, fantasy sports providers. Sportsbooks could use the insights gained to more accurately

place odds on batter events, which could be applied to fantasy baseball for projecting points or stats for individual games as well.

Assumptions, Limitations, and Delimitations

While the intricate sensor data of Statcast can provide details not measured before, there are still plenty of assumptions and limitations that come with the dataset. There are a variety of variables that are not reflected within the dataset that can alter outcomes within a game, including fatigue, weather, or other environmental factors. The pitching profile data provided is fairly complete, but batter swing mechanics data is not included, limiting the insights available to retool a batter's approach. Biological influences, such as batter eyesight and other health factors, are also not considered and could be significant within the context of a single matchup (Laby et al., 2019). Additionally, the pitching profile for an individual pitcher represents a general representation based on their previous repertoire. It does not account for real-time strategizing for a matchup or the alterations the pitcher may make for a specific matchup, like adding a new pitch or releasing the ball from a new arm position.

Organization

Following the introduction, a literature review was conducted to describe previous research in the field and where this project fits within the baseball analytics landscape. After that, a methodology chapter was included to describe the steps taken to collect and clean the data, the considerations and techniques used to include or exclude certain data, an exploration of different techniques to create the model and validation for why a technique was chosen, and the testing of the models. Next, the results and conclusion, or analysis, chapter presented the final model built and its results, while also commenting on whether they were expected or not, along with their

potential implications. Finally, a next steps and conclusion chapter was included to identify areas of future work or research and summarize the project.

Conclusion

Baseball is a game rich with traditions, many of which have been very difficult to break in the new analytics landscape. Incorporating groundbreaking new sensor data gathered from around the MLB and predictive model-building techniques gave further insights into a difficult game at the expense of challenging many traditional managerial practices and decisions. The ability to accurately predict the performance of a batter dependent on their pitcher matchup can prove to be extremely valuable in certain circumstances, as described further in this project.

Literature Review

Accurately predicting the performance of a batter at the plate has been a focal point throughout much of the baseball analytics boom. Previous efforts were investigated to gain a better understanding of the techniques and concepts utilized while exploring this aspect of the MLB or, in several cases, baseball on a world stage. The themes examined were the overall evolution of performance metrics, the budding uses for player evaluation with the newly introduced sensor data, and previously developed batter and pitcher performance models to better inform the decisions of this project. The tools used to assist in the search for these articles and academic journals were the University of Wisconsin-La Crosse's Murphy Library database and Google Scholar.

Performance Metrics Evolution

The longstanding, traditional batter performance statistic that has been used throughout the history of baseball was batting average. Batting average is the percentage of times a batter reaches safely on a hit per at-bat. Regardless of its mainstay status of the game, its role in player

evaluation has been questioned and largely proven to not be very effective. As addressed by F.C. Lane (1916) in the extremely early years of the game in *Baseball Magazine*, batting average fails to place a higher weight on more valuable hitting events, while simultaneously neglecting other valuable events that contribute to a batter's performance and ignoring opposing defensive abilities. Despite this early observation by Lane, batting average proved to be a cornerstone of batter performance evaluation for numerous decades to follow.

The next principal statistic used in batter performance tracking was on-base percentage (OBP), or the percentage of times a batter reaches base per plate appearance. In their study modeling the events of a specific batter and pitcher matchup, Doo and Kim (2018) utilized OBP for their performance statistic of choice for a few reasons. One of these reasons was it was more descriptive than batting average, in that it also accounted for a batter's plate discipline by incorporating walks and hit-by-pitches into its formula. While being the focal point of the flourishing baseball analytics community in the early 2000s, OBP still has its limitations. Similar to batting average, it fails to differentiate between the value of hits and still neglects the defensive component of the game (Umemura et al., 2021). Slugging percentage, and subsequently, on-base plus slugging (OPS), is an attempt at a statistic to properly value extra-base hits, but it still falls short of weighting the events accurately (Slowinski, 2010).

The latest iteration of batting performance statistics to be wildly adopted is called weighted on-base average (wOBA). Developed by Tom Tango, wOBA attempts to weigh each batter event proportionately to their actual run value (Slowinski, 2010). The resulting statistic is a much better reflection of offensive performance for a batter, perhaps at the expense of not being as intuitive in its calculation as its predecessors. A further extension of this statistic is the expected weighted on-base average (xwOBA), which assigns a value corresponding to the

expected wOBA of a batted ball given its exit velocity and launch angle as opposed to the actual batting result (Umemura et al., 2021). While the exact calculation for this statistic is proprietary, it largely corrects for defensive contributions, and it is provided by the MLB in their Statcast dataset (“Statcast Search CSV Documentation”, n.d.). Given these corrections on the traditional performance stats, several projects have utilized xwOBA as their metric of choice when evaluating various aspects of the game, including Umemura et al. (2021) analyzing the performance of different pitch types in the Japanese professional league called Nippon Professional Baseball (NPB) and Brill et al. (2023) analyzing the performance of batters while investigating the through the lineup penalty. These studies made compelling arguments for future research to apply the xwOBA metric when measuring the performance of batters.

New Sensor Data

The modern baseball stadiums outfitted with sensors and cameras have provided a new source of data to be mined by analysts. The Statcast system is the most comprehensive example of this new revolution, which used Doppler radar via TrackMan to track the pitched and batted ball and two arrays of optical video sensors and stereo vision techniques to track player movement (Healey, 2017). Similarly, several Nippon Professional Baseball (NPB), the professional league in Japan, teams have employed the TrackMan Doppler radar technology starting in 2014 to gather pitch-tracking data (Umemura et al., 2021). In the study by Umemura et al. (2021), they worked to quantitatively classify TrackMan data of pitches in the NPB into their various pitch types and then analyzed the pitch type characteristics and the resulting batter performance for each category. Both a pre-pitch declaration by the pitchers themselves and a technique called the Variational Bayesian Gaussian Mixture Models were employed in a two-step approach to sort the pitches into their respective pitch types (Umemura et al., 2021).

Likewise, Statcast derives the pitch type that is provided in their publicly accessible dataset to make it possible to analyze the various arsenal pitch usage (“Statcast Search CSV Documentation”, n.d.).

In his article in the *Proceedings of the IEEE* journal, Glenn Healey (2017) outlined several different skills that are now quantifiable using sensor data, including pitch framing by catchers, player reaction time, and player route efficiency, along with supporting existing skills already quantifiable with higher accuracy. He also demonstrated the improved measurability of batter performance given the batted ball’s initial speed, horizontal launch angle, and vertical launch angle with the ability to produce a three-dimensional plot he called the wOBA cube (Healey, 2017). From this cube, a given batted ball speed, horizontal and vertical launch angle can provide an expected value of the aforementioned wOBA metric. The article states that such statistics derived from sensor data have been shown to have higher repeatability than outcome-based statistics, supporting the use of such data for future player evaluation purposes (Healey, 2017).

In a separate study, Healey (2020) expands on this three-dimensional wOBA cube by adding a fourth dimension, the batter’s running speed, measured by the Statcast system’s optical measurement components. From the sensor data, he mapped the intrinsic value of wOBA for a given value of batted ball speed, horizontal launch angle, vertical launch angle, and running speed for both right-handed and left-handed batters. Incorporating the speed factor of the batter led to a significant increase in accuracy for a batter’s observed wOBA on batted ball events (wOBAcon) compared to the previous model (Healey, 2020). The results of this study display the merit of using sensor data and how the additional measurable skills can add value when using

models to evaluate player performance. It also illustrates the few limitations of the wOBA metric, in that it does not incorporate player speed in its calculation.

Performance Model Analysis

Several efforts have previously been made to produce models to predict specific outcomes related to player performance. In the paper by Fellingham and Fisher (2018), a Bayesian semiparametric model was utilized to predict the home run production of MLB players. The inputs into this model included the total number of home runs hit, the total at-bats, age, team and season played per player, and also adjusted for era and home ballpark. The results of the model were fairly accurate in predicting the home run per at-bat per season and allowed the authors to identify the best “pure” home run hitters throughout history once the ballpark, season, and era effects were stripped away (Fellingham & Fisher, 2018). The conclusions of this study displayed the importance of less obvious variables, like the era played or ballpark effects, have on the performance of players that can be hidden in observed statistics.

Other more granular efforts have been conducted for specific batter and pitcher matchup models for performance as well. In another project led by Glenn Healy (2015), the probability of a strikeout in a particular batter and pitcher matchup was modeled. The technique used by Healy is called a log5 model, which is a special case of a logistic regression model (Healy, 2015). Four general models were created based on the different handedness combinations of the batters and pitchers in an attempt to increase accuracy while still having a large sample size, and the only inputs initially were the batter’s strikeout rate in the matchup, the pitcher’s strikeout rate in the matchup, and the league average strikeout rate for that handedness combination (Healy, 2015). A groundball rate component was later added mirroring the previous strikeout format and was found to be largely significant. A notable takeaway from the study was that batters were

responsible for more variance than pitchers related to strikeout probability (Healy, 2015). A limitation of this study was its broad scope by clustering all matchups by handedness. While it has a significant impact, it may lose the nuance of particular matchups of interest and become overly influenced by the large sample sizes of the groups.

A study by Doo and Kim (2018) also utilized and expanded upon a log5 model, but instead of strikeout probability for a specific MLB batter and pitcher matchup, they used it to find the probability a batter would reach base in a given matchup applied to the Korea Baseball Organization (KBO). Using the inputs of the batter's OBP, the pitcher's opponent OBP, and the league average OBP, they developed a Bayesian hierarchical log5 model to use a smaller sample size of previous outcomes for a matchup, and the results found their method to be very effective compared to the traditional log5 model (Doo & Kim, 2018). Later in the study, a defensive index parameter was introduced into the model and was also found to be significant (Doo & Kim, 2018). A limitation of this study is the use of OBP for the performance of a specific plate appearance between a batter and a pitcher. As previously noted, OBP is a well-respected metric, but there are other more descriptive metrics available, such as wOBA. A noteworthy observation between the two matchup-based studies reviewed is they shared the common theme of using logistic models, in that the event in question was binary. Either a batter strikes out or they do not, and likewise for reaching base. A more descriptive performance metric that properly weights batter events based on their run value, such as wOBA, is not binary, and therefore could not be utilized in the same fashion as the output of a log5 model.

Conclusion

Previous research surrounding the performance of batters and pitchers at the professional level of baseball has evolved greatly in recent times and produced many important insights that

have changed the way the game is played. This evolution is perhaps best illustrated in the metrics used to evaluate player performance, developing from the problematic batting average statistic to xwOBA as a better reflection of a batter's worth at the plate. The modernized stadiums equipped with sensors and cameras have produced a wealth of data that has paved the way for skills to finally be quantifiable and input into evaluation models. Such evaluation models, which can be as granular as predicting various events of specific batter and pitcher matchups, have proved to be extremely useful for team management and learning valuable insights for player usage. Overall, the use of sensor data to predict the performance of a batter and pitcher matchup in the form of modern metrics was supported by this literature review, along with justifying the utilization of the xwOBA metric to measure the performance in a unique matchup between a pitcher and a batter. Drawing from the conclusions of these previous works, a foundation was set for the methodology of the batter and pitcher matchup regression model measuring xwOBA to be further explored.

Methodology

In the process of building a model to predict the performance of a batter against a specific pitcher matchup, this chapter describes the methodology of the project broken into four main sections:

1. The data collection section provides an overview of the origin of the dataset utilized in the model-building process.
2. The data preparation section goes in-depth into how the dataset was examined, manipulated, and cleaned to be useable for future analysis.
3. The model generation and evaluation section thoroughly describes the various approaches taken while building the different types of models and how they were evaluated.

4. The final model section briefly comments on the model that was selected.

Data Collection

The dataset used to complete this project was obtained from Statcast for the 2023 Major League Baseball season, from opening day on March 30, 2023, to the last game of the World Series on November 1, 2023. As previously mentioned, the Statcast system used Doppler radar to track the pitched and batted ball, along with two arrays of optical video sensors and stereo vision techniques to track player movement (Healey, 2017). This modern data collection system allowed for the gathering of new variables previously unobtainable, along with some more traditional ones. The dataset contained 92 variables, including descriptive variables like the player identification number and team played for, along with variables such as pitch type, release speed, the horizontal and vertical release positions, the pitcher and batter handedness, the horizontal and vertical movement of the pitch, the horizontal and vertical positions of the ball when it crossed the plate, the velocity and acceleration of the pitch in the x-, y-, and z-dimensions, the spin rate of the pitch, the release extension of the pitcher, the effective speed of the pitched ball derived from the speed of the pitch and the pitcher's extension, the weighted on-base average (wOBA) value of the play, and the estimated wOBA of an event based on the batter's launch angle and exit velocity ("Statcast Search CSV Documentation", n.d.). All relevant variables were framed from the catcher's perspective. The dataset contained 656 unique batters and 863 unique pitchers that made an appearance throughout the 2023 MLB season, including over 968,000 individual pitches recorded. Each individual's exclusive dataset was gathered utilizing the baseballr library package in R to be further manipulated.

Data Preparation

There were several steps involved in the preparation of data for building the model, including exploratory data analysis, filtering the dataset, and creating a pitcher profile for each pitcher.

Exploratory Data Analysis

The dataset from Statcast from the 2023 MLB season previously described was examined and several variables were considered for the model. Many were simply describing the scenario and not within the scope of this project, such as the player identification number of each fielder, or whether or not there were baserunners during the plate appearance. In total, 20 variables were selected for the next exploratory phase that described the pitch a batter saw and their subsequent result of the plate appearance, as given in Table 1.

Table 1

Statcast Variables and Descriptions

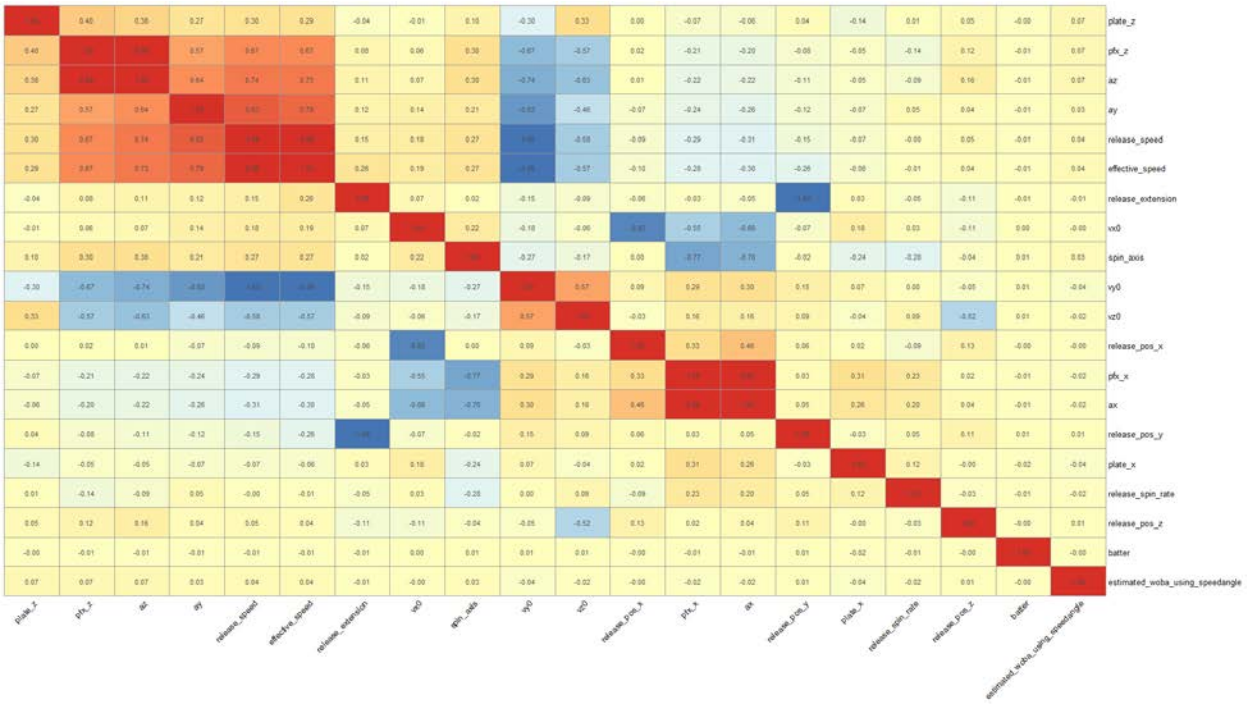
Variable name	Description
release_speed	Pitch velocity out-of-hand
release_pos_x	Horizontal release position of the ball measured in ft from catcher's perspective
release_pos_z	Vertical release position of the ball measured in ft from catcher's perspective
p_throws	Hand pitcher throws with
pfx_x	Horizontal movement in ft from catcher's perspective
pfx_z	Vertical movement in ft from catcher's perspective
plate_x	Horizontal position of the ball when it crosses home plate from the catcher's perspective
plate_z	Vertical position of the ball when it crosses home plate from the catcher's perspective
vx0	Velocity of the pitch, in ft/s, in x-dimension, determined at y=50 ft
vy0	Velocity of the pitch, in ft/s, in y-dimension, determined at y=50 ft
vz0	Velocity of the pitch, in ft/s, in z-dimension, determined at y=50 ft
ax	Acceleration of the pitch, in ft/s, in x-dimension, determined at y=50 ft
ay	Acceleration of the pitch, in ft/s, in y-dimension, determined at y=50 ft
az	Acceleration of the pitch, in ft/s, in z-dimension, determined at y=50 ft
effective_speed	Derived speed based on the extension of the pitchere's release
release_spin_rate	Spin rate of pitch
release_extension	Release extension of pitch in ft
release_pos_y	Release position of pitch measured in feet from the catcher's perspective
spin_axis	The Spin Axis in the 2D X-Z plane in degrees from 0 to 360, such that 180 represents a pure backspin fastball and 0 degrees represents a pure topspin (12-6) curveball
estimated_woba_using_speedangle	Estimated wOBA based on launch angle and exit velocity

Note. Adapted from https://baseballsavant.mlb.com/csv-docs. Copyright by Baseball Savant.

After identifying these potential variables, they were explored to find their relationship with one another. The entire dataset was used for this exploration segment. The correlation coefficient was found for each quantitative variable, as displayed in the heatmap of Figure 1, to identify how each was linearly related. As shown in the figure, several variables were directly correlated with each other. After further investigation, these variables were derived from one another, such as the horizontal and vertical movement of a pitch being derived from the corresponding acceleration components of the pitch, the various speed measurements in the y-coordinate, and the release extension and the release position y-coordinate. Given this revelation, the variables of release_speed, pfx_x, pfx_z, effective_speed, and release_extension were removed.

Figure 1

Correlation of Statcast Variables Heatmap

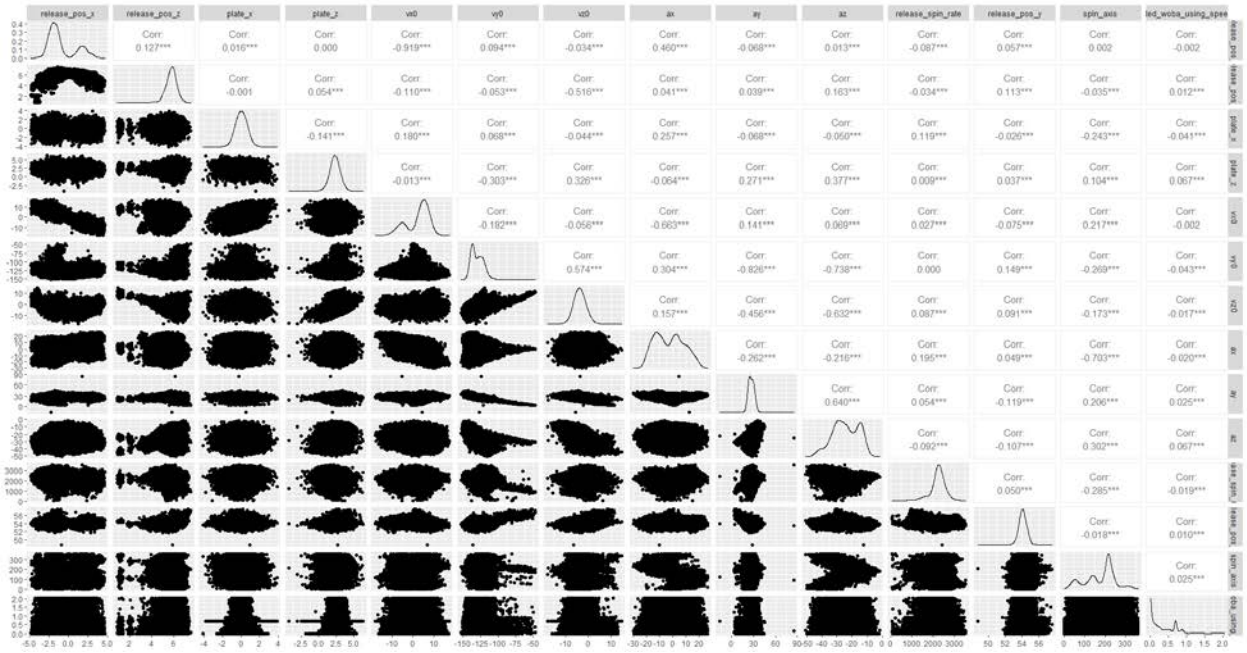


The distributions of the remaining variables were observed, as shown in Figure 2, to determine if any transformations or manipulation of the data was necessary. A few variables,

such as the horizontal release position of a pitch, displayed bimodal distribution, which can be explained by the different handedness of pitchers. Due to this clear relationship between the horizontal release point of a pitch and their handedness, the categorical variable of p_throw was also removed. The only variable determined to warrant a transformation was the response variable of the estimated wOBA. The distribution for this variable had a severe right skew, which conceptually made sense, as the majority of plate appearances result in a low wOBA, including strikeouts, while there was also another peak at around 0.7, which corresponds to the value of walks. Given a large portion of this variable had a value of 0, a $\log(x + 1)$ transformation was applied in an attempt to increase normality. The variable after the transformation was signified as xwoba_log.

Figure 2

Distributions of Selected Statcast Variables for Model Building



Filter and Imputation

Of the over 968,000 individual pitches recorded in the Statcast dataset during the 2023 MLB season, the project was only concerned with those events that ended a plate appearance with a measurable wOBA value. Since the estimated wOBA value in the dataset was only applied on a batted ball with a launch angle and exit velocity, any plate appearance that ended without a batted ball was not assigned a value for this variable. To simulate a batter's strikeout and walk tendency based on certain pitches, those plate appearances that ended with one of these outcomes were assigned their respective values in the estimated wOBA variable. This would correspond to 0 for a strikeout and 0.7 for a walk. While this manipulation of data fundamentally changes the meaning of the estimated wOBA variable, the value gained from simulating a batter's approach was deemed beneficial to the scope of the project. An alternative approach would be to only include plate appearances that resulted in a batted ball in play, but such a method would lose the plate discipline factor of a batter and thus not paint as complete a picture of their value.

Additionally, an arbitrary cutoff of 250 occurrences was chosen for a batter to qualify for a model for a few reasons. First, it was assumed that a batter's traditional statistics did not stabilize until around 120 plate appearances (Salfino, 2020). A more standard cutoff for such rate statistics related to MLB batters throughout a season would be to use what the league defines as a qualified batter, or a batter that tallies at least 3.1 plate appearances per game for the season. This would equate to 502 plate appearances to qualify for an entire 162-game season ("Qualifier", 2010). As a middle ground between the assumed amount needed for a batter's approach to stabilize and the traditional qualified batter, the 250 amount was settled upon to maintain a sizable batter sample. Furthermore, this filtering reduced the number of batter models

required from 656 to 337. Creating this many individual models was computationally expensive, so this great reduction was beneficial to the scope of the project.

Pitcher Profile

Since the main focus of this project was on direct matchups between batters and pitchers, a pitcher profile needed to be developed for each pitcher to subsequently be input into each batter model produced. A pitcher's profile contains several different pitch types that can act very differently and they throw each pitch type at different frequencies. The dataset provided the pitch type for each ball thrown during every plate appearance, so a method was needed to create an accurate representation of a pitcher's arsenal. The process that was chosen involved averaging each quantitative variable for each pitcher's respective pitch type and finding the frequency at which the pitcher throws that specific pitch type. An example of the before and after of this process can be seen in Figures 3 and 4. Figure 3 entails the pitch arsenal for the 2023 National League Cy Young recipient, or the most valuable pitcher, Blake Snell provided by Statcast, which displays a heatmap of each pitch type, where red signifies the most often pitch location. This representation of the data shows how much variance is present in the pitch data, especially concerning location. Figure 4 then displays the averaged location of each pitch type from Blake Snell, along with their label and the percentage of total pitches that pitch type composed. The size of the circle of each corresponding pitch correlates to the frequency the pitch is thrown. In addition, the pitch release locations for each pitch type were included in the figure, symbolized by a square, and a dashed line of the corresponding color connects the two to better visualize the path of the pitch. The dashed line does not properly trace the movement of a pitch once it leaves the pitcher's hand, but it is a helpful aid nonetheless.

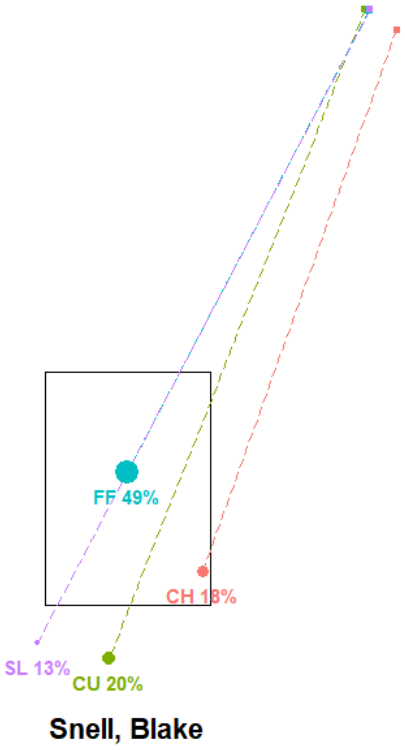
Figure 3

Blake Snell Statcast Pitch Arsenal Heatmap

Note. From Blake Snell's 2023 season page on Baseball Savant from https://baseballsavant.mlb.com/savant-player/blake-snell-605483?stats=statcast-r-pitching-mlb&season=2023. Copyright by Baseball Savant.

Figure 4

Blake Snell Average Pitching Profile Visual from Catcher's Perspective



Each averaged pitch type would then be input into the batter model and weighted by the frequency of occurrence per pitcher in an attempt to simulate a plate appearance. While a plate appearance can be heavily dependent on previous scouting for a batter and would not necessarily adhere to the same measurements as this developed pitcher profile, it was thought to provide insight into how a batter would respond to a normal approach from the respective pitcher.

Model Generation and Evaluation

Once the data was collected, cleaned, and prepared, the model-building process began. Two separate data science techniques were utilized for the model generation portion of the project: the generalized linear model (linear regression) and artificial neural networks. These two methods were chosen to explore both linear and nonlinear relationships between the covariates and the dependent variable. The models created were evaluated against each other to identify the strongest candidate for the final model to input the newly created pitcher profiles.

Generalized Linear Model

Despite some concerns related to the linearity assumptions required, as a baseline, a generalized linear model was developed for each batter utilizing the caret library in R using the glm method. Linear regression models are regarded as a simpler modeling technique that are beneficial in identifying easily interpretable linear relationships between the covariates and the dependent variable (Götze et al., 2023). The general formula for this model for a single response variable is given by:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i \quad (1)$$

where y is the response, or dependent, variable, β_0 is the intercept, β_i are the coefficients, and x_i are the independent variables. The response variable, in this case, was the log-transformed xwOBA, while the independent variables were the other 13 listed above. To standardize the

quantitative variables, they were scaled to the original dataset to have an overall mean of 0 and a standard deviation of 1. In an attempt to reduce bias and overfitting, the cross-validation function within the same caret library was utilized. Cross-validation involves splitting the dataset into different folds of training and testing data, and iterating the model-building method through these partitions of training and testing data to avoid the bias and overfitting that may arise from a single training dataset alternative. A 10-fold cross-validation was chosen while building the generalized linear model.

When considering the evaluation of the models, the method used was to find the average of the root mean squared error (RMSE) measurements across every individual model and use that for comparison. The formula for RMSE is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (2)$$

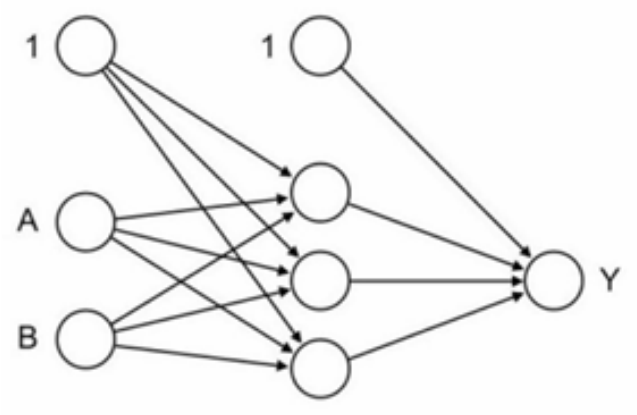
where $\hat{y}_i$ are the predicted values, y_i are the actual values, and n is the number of observations. This was thought to be a more streamlined approach to simplify the interpretation of the models, as opposed to finding the parameters that optimized each batter model individually. With this in mind, the 10-fold cross-validation generalized linear model had an average RMSE of 1.001422132 across all 337 batter models.

Neural Network Model

Similarly, a single-hidden-layer neural network model was developed utilizing the caret library in R using the nnet method. The most basic form of this type of model is visualized in Figure 5.

Figure 5

Single-Hidden-Layer Neural Network Example



Note. From "neuralnet: Training of Neural Networks," by F. Günther & S. Fritsch, 2010, *The R Journal*, *2*(1), 30–38. Copyright 2010 by Frauke Günther and Stefan Fritsch.

In this example, one can observe the input variable nodes of A and B, the three hidden nodes in the middle, and the output variable node of Y. There are also two nodes labeled with the 1 label that signifies a bias term that gets added to the respective nodes of that row. Additionally, each arrow denotes a weight value term. In its simplest form of only containing the input variables layer and a single output neuron without any hidden node layer, the formula is essentially identical to the generalized linear model described in the previous section:

$$z = f(b_0 + \sum_{i=1}^{n} w_i x_i) \tag{3}$$

where z is the output, b_0 is the bias unit, or intercept, w_i are the weight values, x_i are the input variables, and n is the number of input variables (Günther & Fritsch, 2010). With the introduction of a single hidden layer of nodes, the formula is expanded to the following:

$$z = f\left(b_0 + \sum_{j=1}^{m} w_j \cdot f\left(b_{0j} + \sum_{i=1}^{n} w_{ij} x_i\right)\right) \tag{4}$$

where z is the output, b_0 is the bias unit, or intercept, for the output neuron, w_j are the weights of the specified hidden node connected to the output neuron, b_{0j} are the bias units, or intercepts, of

the specified hidden node, w_{ij} are the weights leading from the input layer to the specified hidden node, x_i are the input variables, n is the number of input variables, and m is the number of nodes within the single hidden layer (Günther & Fritsch, 2010). The weights are initially assigned at random and are used to find an output that is subsequently compared to the actual value. The weights are then updated to improve the accuracy and this process, known as backpropagation, is continued until the model converges or reaches a predetermined maximum number of iterations. To better control this backpropagation process and reduce overfitting, a regularization term can be introduced to the cost function that is used to determine how the weights are updated. This can be shown in the following formula:

$$C = C_0 + \frac{\lambda}{2n}\sum_{i=1}^{m} w_i \quad (5)$$

where C is the cost function, C_0 is the original cost function before the regularization term, λ is the weight decay parameter, n is the number of observations in training data, w_i are the weight values, and m is the number of weights (Nielsen, 2015).

The two parameters that were tuned through when building the model were the number of nodes in the single hidden layer and the weight decay value. The number of hidden nodes that were tested was between 2 and 5, while the weight decay value of between 0 and 20 by increments of 1 was also tested, with the maximum iteration limit set to 2000. Additionally, a 10-fold cross-validation method was also utilized to further reduce overfitting and bias, mirroring the process from the generalized linear model section. The parameters that yielded the best result for the neural network were 2 nodes and a weight decay of 11 to give a RSME of 0.9902194. The results of every combination of test parameters between the decay value of 2 to 20 can be observed in Figure 6 with the optimal decay signified by the black dashed line. A visualization of

the optimal neural network model can be seen in Figure 7 with the model input variables and the output variable of log transformation of xwOBA.

Figure 6

Optimized 10-fold Cross-Validation Neural Network Size and Decay

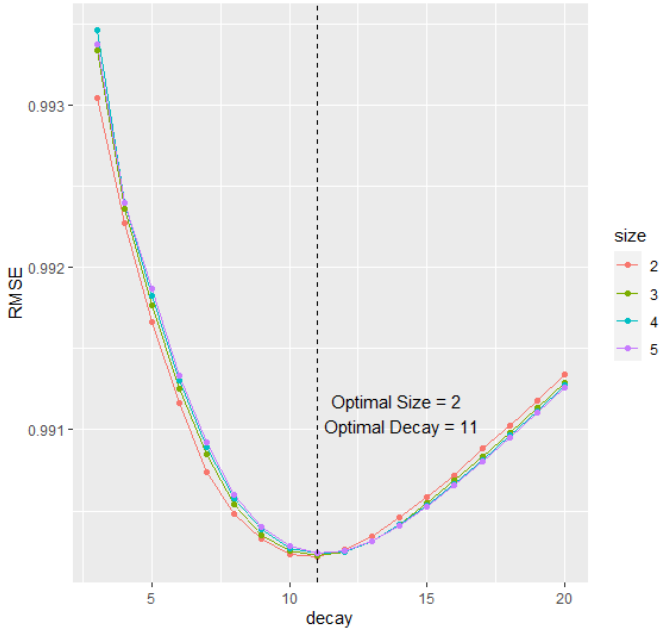
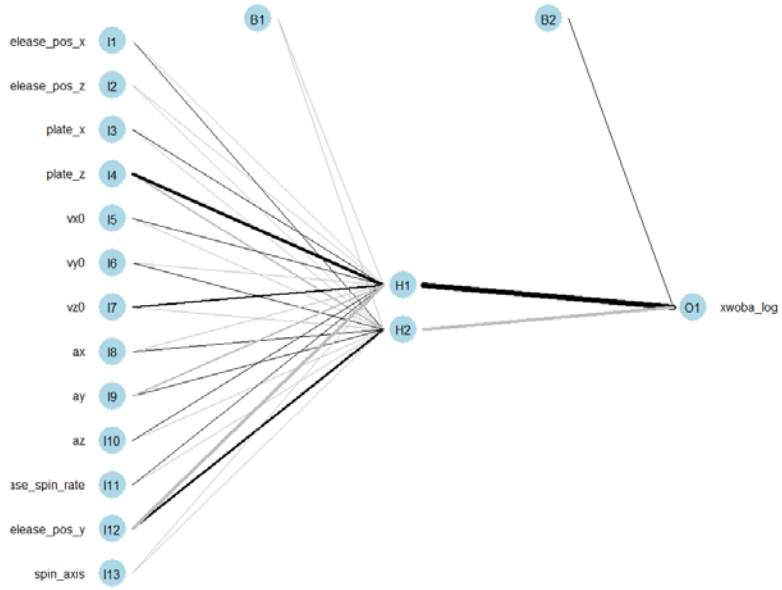


Figure 7

Batter Neural Network Model Visualization



The mean RMSE of the models from the varying neural network parameters and the linear regression model were justified in the use of performance comparison due to the substantial sample size of 337, and its distribution being reasonably normal, as per the Central Limit Theorem (Kwak & Kim, 2017). The RMSE distribution for the model of 2 nodes with a weight decay value of 11 can be observed in Figure 8, with the dashed black line representing the mean of the distribution. A quartile-quartile, or Q-Q, plot of the RMSE distribution, which plots the quantiles of the distribution against each other, can be viewed in Figure 9 as well (Ghasemi & Zahediasl, 2012). A straight, diagonal line formed from the data would represent a perfect normal distribution in the Q-Q plot. Additionally, a Shapiro-Wilk test was conducted on these RMSE results, which is a test based on the correlation between the results and a normally distributed dataset with the same mean and standard deviation. The resulting p-value from this test was 0.4135, which easily surpasses the 0.05 threshold to commonly refer to normality. The two visual checks, plus the additional normality test, left little doubt that the RMSE distribution was normal and the use of the RMSE mean to evaluate the performance of the various model parameters was supported.

Figure 8

Neural Network Individual Batter Model RMSE Distribution: 2 Node & 11 Weight Decay Value

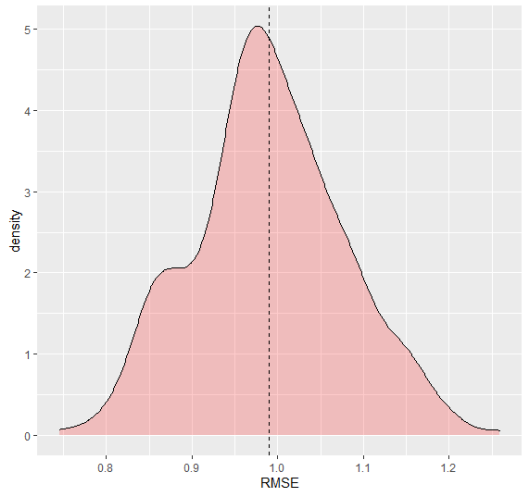
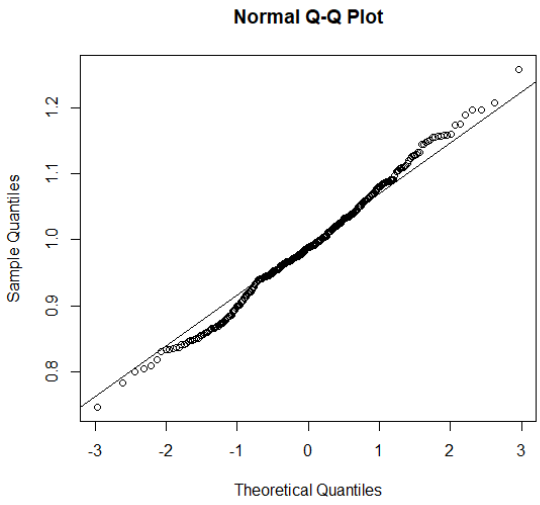


Figure 9

Neural Network Individual Batter Model RMSE Q-Q Plot: 2 Node & 11 Weight Decay Value



Final Model

After the best generalized linear model and neural network model were chosen from their respective model-building technique, the two models were compared to one another. The models performed surprisingly similarly, but the 2 hidden node neural network with a weight decay of 11 slightly outperformed the generalized linear model. The neural network had a lower average RMSE and a lower average mean absolute error (MAE). While the latter metric was not used in evaluating the models against each other, it was reassuring that the neural network outperformed the linear model to leave little doubt.

Conclusion

The extensive dataset describing the 2023 MLB season was collected via Statcast and subsequently examined, cleaned, and filtered to prepare to build a model that predicts the performance of a batter given a pitcher matchup. Two different model-building techniques were utilized, and a neural network model of 2 hidden nodes and a weight decay value of 11 was deemed to be the most effective after the aggregate evaluation. Next, all 337 batter models newly

created were subjected to each of the unique pitcher profiles, and the results were further analyzed.

Results and Analysis

After successfully creating the individual models to predict the performance of a batter in a specific pitcher matchup utilizing the methodology previously described, the results were examined in full detail in this chapter. The analysis was divided into 4 main sections:

1. The model discussion section describes and analyzes the selected neural network model in further detail.
2. The variable importance section thoroughly explores the measured impact of the different input variables displayed.
3. The comparison to traditional judgment section compares the findings of the analysis to the longstanding thought processes of baseball.
4. The model shortcomings section reflects upon aspects of the model that could limit its effectiveness.

Model Discussion

As mentioned in the prior methodology section, the best performing and selected single hidden layer neural network model with 2 nodes and a decay value of 11 possessed a RMSE of 0.9901935. The model gave a R^2 value of 0.03019299, which is a direct result of the little correlation present in the covariates to the dependent variable of the expected wOBA as observed in the methodology section. This value indicates that about 3% of the variability observed in the log transformation of the expected wOBA variable is explained by the model. This would suggest that the model produced may not be the best equipped to accurately predict the exact xwOBA for a specific matchup between a batter and a pitcher from the variables provided, but

important insights can still be gathered for a specific batter that can help place them in their most advantageous situation.

The unique pitcher profiles were input into each batter model to produce a corresponding log_xwOBA value for each matchup. Similar to the filtering subjected to the batter data described in the methodology chapter, only pitchers with at least 250 recorded pitches in the 2023 season were included in the analysis to remove pitchers with limited data and occurrences where position players pitched in blowout games. This reduced the number of pitchers to 604, giving the total number of 203,548 unique matchups with 337 batters. The results were transformed back to represent the expected wOBA value after the matchups were run. The mean of the results of the top 10 performing batters and bottom 10 performing batters are displayed in Table 2.

Table 2

Top and Bottom 10 xwOBA Batter Averages from Simulated Pitcher Matchups

Top Performers		Bottom Performers	
Player	**Mean xwOBA**	**Player**	**Mean xwOBA**
Acuña Jr., Ronald	0.3940506	Escobar, Eduardo	0.2102636
Judge, Aaron	0.3778258	Doyle, Brenton	0.2164171
Alvarez, Yordan	0.3768425	Castro, Rodolfo	0.2188547
Seager, Corey	0.3721947	Haase, Eric	0.2226361
Harper, Bryce	0.3544648	Bae, Ji Hwan	0.2250685
Betts, Mookie	0.3529547	Maldonado, Martín	0.2259227
Freeman, Freddie	0.3526318	Wendle, Joey	0.2261910
Soto, Juan	0.3524710	Schmitt, Casey	0.2266187
Ohtani, Shohei	0.3496875	Allen, Nick	0.2267894
Tucker, Kyle	0.3399607	Bethancourt, Christian	0.2269661

The top performers mirror perennial Most Valuable Player (MVP) candidates and players who are tracking for the Hall of Fame, which is a good indicator for the model's representation of comparative batter performance. Ronald Acuña, Jr., the winner of the National League MVP award in 2023, was the best average performer in the model by a fairly large margin. As

displayed in Table 2, the output values for xwOBA are slightly lower than what was observed in the 2023 season. For comparison, according to Statcast from the 2023 MLB season with a minimum of 250 plate appearances, the maximum observed xwOBA measured was Aaron Judge at 0.461, with the minimum xwOBA belonging to Eduardo Escobar at 0.244 (“Baseball Savant”, 2023). When interpreting these results, it is important to keep in mind the makeup of the pitcher profiles used as the input, as they would likely change depending on the batter matchup. It is not surprising that the model values would skew lower, as the costliest pitching mistakes are masked by the average profile methodology utilized. One mistake pitch towards the middle of the plate is all it can take during a plate appearance to inflate the xwOBA from a strikeout, a xwOBA value of 0, to a homerun, a xwOBA value of about 2, and many of those types of pitches are probably not well represented for pitchers when averaging all of their pitches across a season. Additionally, the average pitch qualities for a pitcher would almost certainly trend toward a pitcher’s strengths along the edges of the strike zone.

The highest recorded matchup was the batter Corey Seager versus the pitcher Tyler Rogers with the xwOBA value of 0.5540474, while the lowest recorded matchup was the batter Shea Langeliers versus the pitcher Tim Hill with the xwOBA value of 0.1220927. These results are interesting because both pitchers have very distinct pitching motions that would make them outliers in various release point variables. Tyler Rogers’ and Tim Hill’s pitching profiles were visualized in Figures 10 and 11, respectively.

Figure 10

Tyler Rogers Average Pitching Profile Visual from Catcher's Perspective

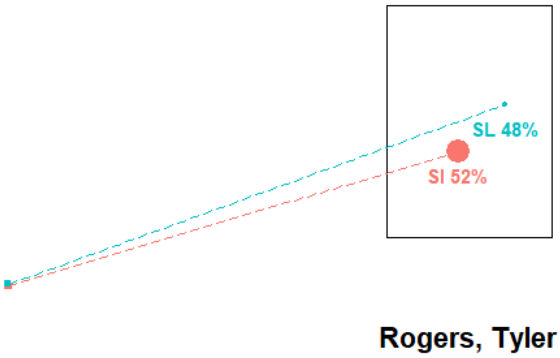
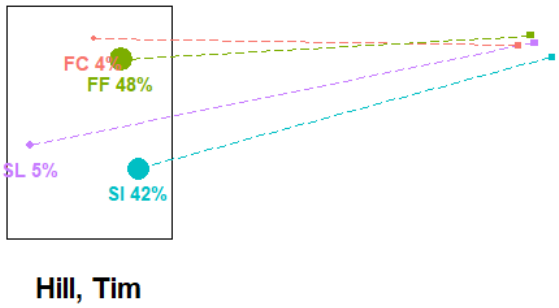


Figure 11

Tim Hill Average Pitching Profile Visual from Catcher's Perspective



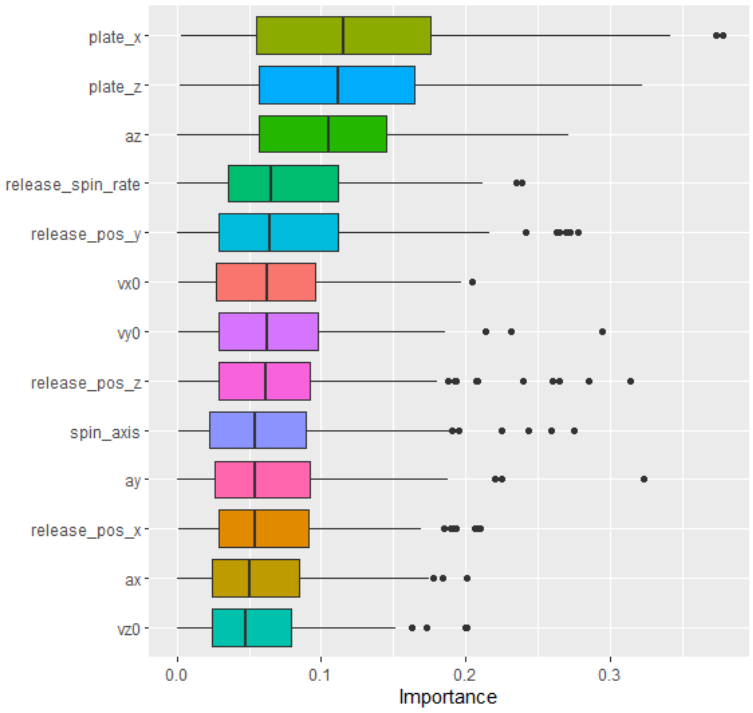
Rogers, in traditional baseball terminology, is considered a submarine pitcher, in that his release point is below his waist when he pitches, while Hill's release point is extremely far towards the first base side as a left-handed pitcher, and the ball is essentially released behind where a left-handed batter would be standing at the plate in the batter's box. Few other pitchers in the current game have similar release profiles to these two, which could be an explanation for their extreme results within the model.

Variable Importance

Among the most important insights gathered was a comparative importance measurement of the variables input into the multiple models. The interpretability of the different variable impacts within a neural network is not as straightforward as it is with linear regression. In this analysis, the variable importance of each batter neural network model was found via Garson's formula using the garson function from the NeuralNetTools library in R. Using this method, the relative importance of a predictor is determined by dissecting the weights within the neural network, and then subsequently pooling and scaling the weights associated with each variable to generate a value between 0 and 1 that represents their relative importance when compared to one another (Zhang et al., 2018). These values were aggregated and compared to one another to observe each variable's relative importance for a batter's performance. The results of this Garson's formula analysis are visualized in Figure 12 below.

Figure 12

Variable Importance Boxplot for Batter Neural Network Models



As observed from the figure, two of the three variables that were clearly the highest indicator of batter performance throughout the numerous models were related to the position of the pitch when it crossed home plate: plate_x and plate_z. The next variable that was of comparable importance was associated with the acceleration of the pitch in the z-dimension determined at the 50 ft mark between the mound and home plate with az. The remaining variables were comparable in importance, with the least important variable being identified as the velocity of the pitch in the z-dimension, or vz0.

As mentioned, these variable importance findings indicate that the highest predictors of batting performance appear to be the position of the pitch when it crosses home plate, followed by the movement of a pitch. This stresses the importance of plate discipline for a batter while stressing the importance of pitch placement and movement for a pitcher. This conclusion makes logical sense, as poor pitch placement, especially in the center of the strike zone, has long been associated with very good batting performance (Vock & Vock, 2018). Similarly, pitches with a large amount of movement, or break, would logically make it more difficult to generate quality contact for a batter, but the higher importance of movement in the z-dimension can be insightful for pitchers when developing their pitches. Alternatively, this analysis surprisingly shows that the pitch speed, indicated by the variable vy0, does not have very much impact on batter performance compared to the top three variables.

Comparison to Traditional Judgment

The traditional baseball viewpoint on handedness is that a matchup of opposite handedness between a pitcher and a batter would benefit the batter and a matchup of corresponding handedness would benefit the pitcher, as illustrated by Glenn Healey's study (2015) of modeling the probability of a strikeout for a batter when he split them into their various

handedness matchup scenarios. If this assumption were to be supported by the findings of the performance models developed in this project, one would expect the pitcher release position variable in the x-plane, release_pos_x, to be of high comparative importance. While there may be outliers, the analysis of the various models would generally point to low importance being assigned to the handedness of a pitcher across all of the batter models evaluated as indicated by the release_pos_x variable in the previously observed variable importance figure, Figure 12. This challenges the traditional thinking within baseball management of seeking out matchups based on handedness to optimize performance in their favor. A possible explanation for these findings is that many of the players used in platoon roles, or specializing against a specific pitcher handedness, did not have enough plate appearances to qualify for the model. Two players who have a history of large splits in performance between the handedness of pitchers are Jorge Soler, a right-handed batter who traditionally performs better against left-handed pitchers, and Kyle Schwarber, a left-handed batter who traditionally performs better against right-handed pitchers. The distribution of Jorge Soler's predicted xwOBA from the model is represented in Figure 13, and his unique garson chart indicating the variable importance in his model is represented in Figure 14. A similar collection of charts is shown in Figures 15 and 16 for Kyle Schwarber.

Figure 13

Jorge Soler 2023 Predicted Pitcher Matchup xwOBA Density Plot

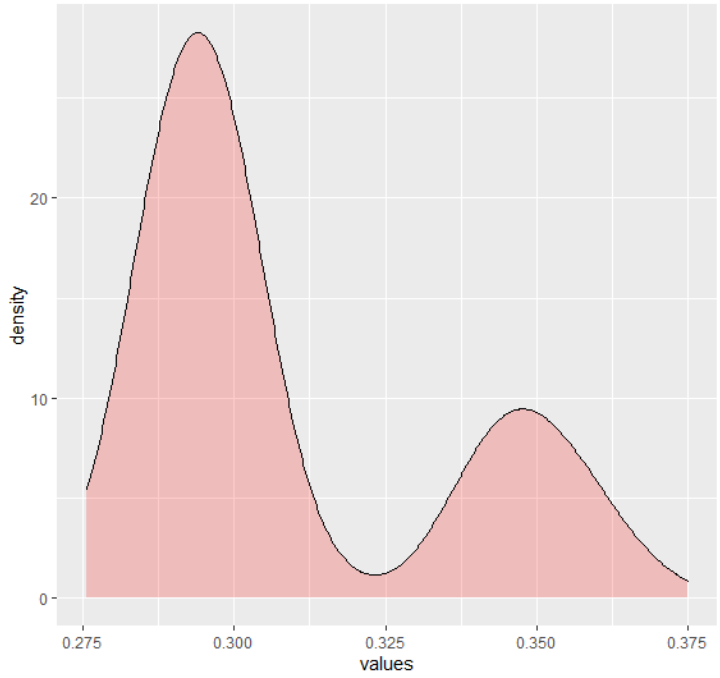


Figure 14

Jorge Soler 2023 Neural Network Model Garson Chart

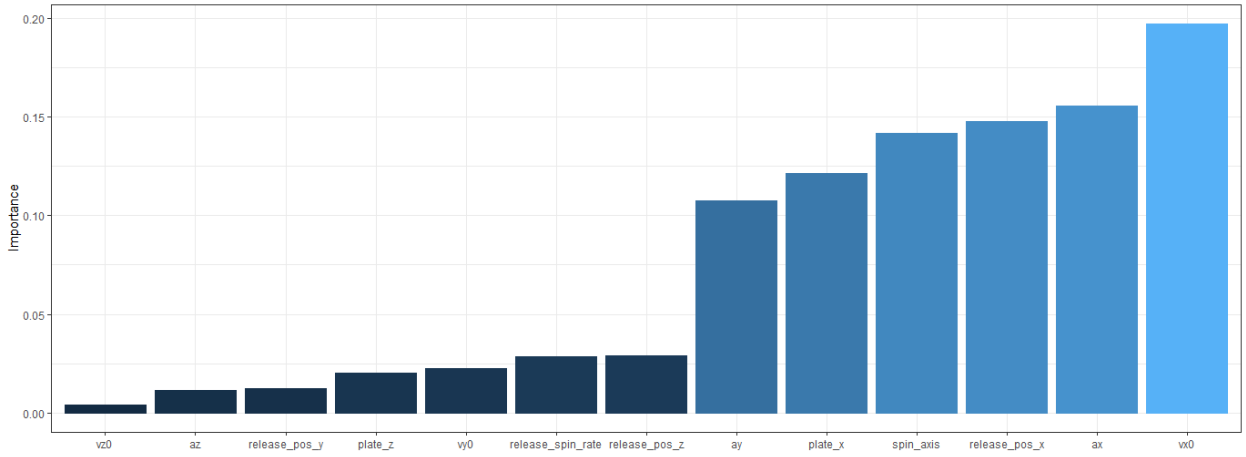


Figure 15

Kyle Schwarber 2023 Predicted Pitcher Matchup xwOBA Density Plot

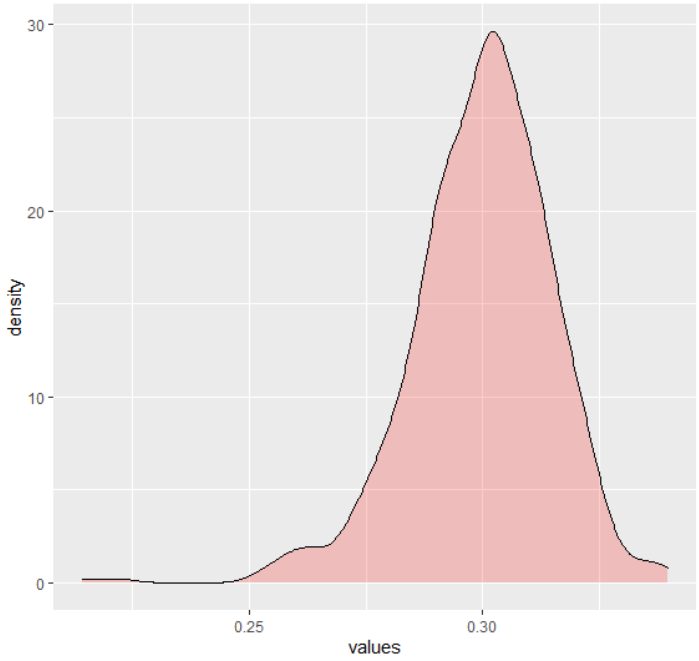
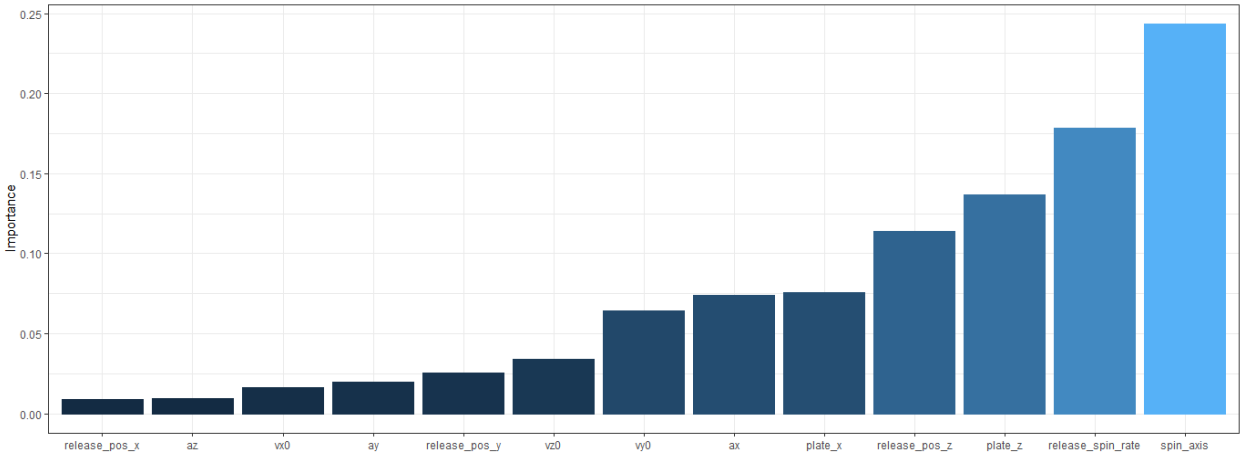


Figure 16

Kyle Schwarber 2023 Neural Network Model Garson Chart



The influence of pitching handedness is extremely apparent in Jorge Soler's performance distribution in the representation of the bimodal peaks. The smaller group represents the left-handed pitchers, and it is clear from the density plot, and by the release_pos_x variable being of high importance, that the traditional baseball thinking of platoon players is displayed in players

like Jorge Soler. The same cannot be said about Kyle Schwaber, as there are not two separate groups apparent in his distribution of predicted xwOBA, and the release_pos_x variable is the lowest weighted variable in his garson chart. The results do not entirely refute the importance of handedness in batting matchups in general, but these discrepancies muddy the concept of platoon performance for players who are talented enough to receive at least 250 plate appearances throughout a season and suggest that other factors should be considered first in many of those instances.

The longstanding assumption regarding pitch location is that pitches that cross home plate within the strike zone, especially around the center of the plate, are correlated with high batter performance, and this performance decreases as the pitch location moves away from the center in any direction. The assumption regarding the vertical movement of a pitch is that an increase in pitch movement in the vertical dimension leads to a decrease in batter performance. As previously touched on in the variable importance section, the high importance of pitch location and acceleration in the z-dimension, or vertical movement of a pitch, is consistent with traditional baseball experiences. The standout discrepancy from the traditional way of thinking is the lower importance of pitch speed. The average velocity of pitchers has been creeping upwards since essentially the origins of the game, but it would appear that players become well-adjusted and it may suggest its importance has become overstated in the current game.

Model Shortcomings

The model created in this project successfully uncovered insights into the performance of batters in the 2023 MLB season based on pitcher matchups, but it was not without its shortcomings. First, the independent variables used were not closely correlated to the performance metric chosen, xwOBA. This made the usefulness of the model limited when

forecasting future performance. Next, to reiterate, the profile of a pitcher used is only a snapshot of what occurs during a game. Extensive scouting curates a specific plan of attack for nearly every pitcher and batter matchup in the modern age of the MLB that may look drastically different from the average pitching profiles used for this analysis. This plan has the potential to deviate significantly from their average profile if they decide to throw a new pitch they were developing or decide to release the pitch from a different position. Plus, it fails to grasp the importance of pitch sequencing. Additionally, limiting the amount of data to one season's worth perhaps did not provide enough detail for each individual model. The scope of the project was large in creating a model for every unique batter, but once the data was divided between the 337 models, there may not have been enough to reach the highest performance potential.

Furthermore, while the model appeared to represent the comparative batter performance well compared to what was observed in the 2023 season, the same cannot said about the pitchers. The worst-performing pitchers on average in the models were among the best-performing relief pitchers throughout the 2023 season, as shown in Table 3. The two largest discrepancies came from the power throwers Félix Bautista and Josh Hader, who are widely regarded as two of the best in the game at this position at the conclusion of the 2023 season.

Table 3

Worst Performing Pitchers from Model Simulated Matchups Versus 2023 MLB Statcast Results

Pitcher	Model Avg xwOBA	2023 Statcast xwOBA	2023 Avg Fastball Velocity (MPH)	2023 Fastball Velocity Percentile
Bautista, Félix	0.292665	0.224	99.5	99
Jackson, Zach	0.289520	0.293	93.1	34
Estévez, Carlos	0.289038	0.311	97.1	92
Stanek, Ryne	0.288340	0.310	98.2	96
Tapia, Domingo	0.287773	0.307	97.1	92
Finnegan, Kyle	0.287116	0.333	97.3	93
Graterol, Brusdar	0.287080	0.271	98.6	98
Cueto, Johnny	0.286765	0.345	92.2	22
McGough, Scott	0.286665	0.321	93.5	40
Harvey, Hunter	0.286458	0.283	98.3	97
Tarnok, Freddy	0.286110	0.400	95.2	71
Hader, Josh	0.285960	0.238	96.1	85

Note. 2023 Statcast xwOBA, average fastball velocity, and fastball velocity percentile adapted from baseballsavant.mlb.com. Copyright by Baseball Savant.

The limitations likely stem from these pitchers frequently throwing high-velocity strikes with little vertical break in traditionally batter-friendly areas of the strike zone, as well as pitching fewer innings and providing fewer data points than a starter would. This can be observed in Figures 17 and 18, which display the Statcast heatmap pitching arsenal of Felix Bautista, along with the averaged pitching profile input into the model. A similar pair of visuals are also displayed for Josh Hader in Figures 19 and 20. As previously observed, the three most impactful variables observed throughout the models were largely related to the pitch position when it crossed home plate and the acceleration in the vertical dimension. This would suggest that the models do not properly weigh the whiff, or swing-and-miss, potential of these high-velocity pitch outliers even when they are in traditionally high-performing batting areas of the plate, as the observed analysis would support a different conclusion. A potential explanation

could be that there are fewer occurrences of these extreme velocity pitches represented in the dataset due to their apparent low batted ball rate.

Figure 17

Félix Bautista Statcast Pitch Arsenal Heatmap

Note. From *Félix Bautista*'s 2023 season page on Baseball Savant from https://baseballsavant.mlb.com/savant-player/felix-bautista-642585?stats=statcast-r-pitching-mlb&season=2023. Copyright by Baseball Savant.

Figure 18

Félix Bautista Average Pitching Profile Visual from Catcher's Perspective

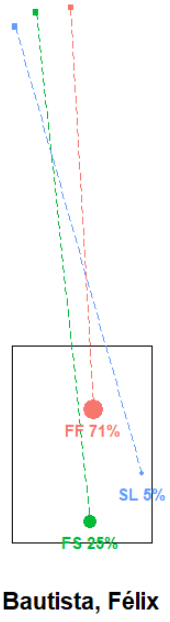


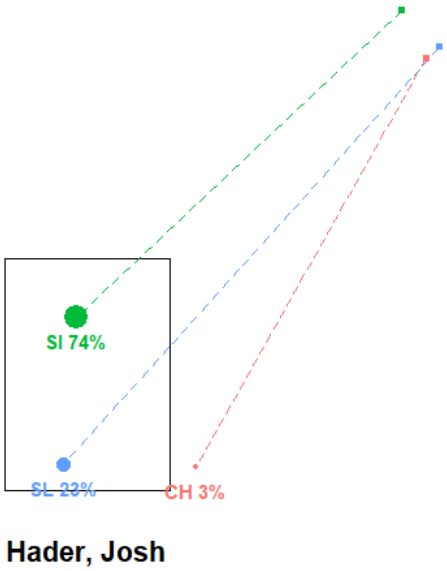
Figure 19

Josh Hader Statcast Pitch Arsenal Heatmap

Note. From Josh Hader's 2023 season page on Baseball Savant from https://baseballsavant.mlb.com/savant-player/josh-hader-623352?stats=statcast-r-pitching-mlb&season=2023. Copyright by Baseball Savant.

Figure 20

Josh Hader Average Pitching Profile Visual from Catcher's Perspective



Conclusion

A regression model for each batter from the 2023 MLB season of at least 250 plate appearances to predict their performance in a specific pitcher matchup was successfully built using the method of artificial neural networks after it was identified as the most effective model-building technique. Their results were analyzed and discussed after the qualified pitcher profiles were applied. While the forecasting usefulness was found to be not entirely effective due to the low correlation between the covariates and the output variable, several interesting insights were discovered. The comparative variable importance revealed the general importance of pitch location selection and the vertical movement of a pitch above the other variables. The insights gained from the variable importance analysis reinforced several traditional concepts in baseball, like plate discipline and pitch movement, but muddied other concepts, like the importance of handedness and pitch velocity. Although the methodology was highly individualized in the sense of building a unique model for each batter, the generalness of the approach highlighted several apparent shortcomings. Despite these revelations, the project successfully completed the defined objectives.

Next Steps and Conclusion

Next Steps

Numerous areas of interest were presented throughout this project that could warrant further investigation in the future. As alluded to in the results and analysis chapter, expanding the amount of data to multiple seasons' worth of Statcast data could improve the model's average performance. Using a similar set of models but simulating a plate appearance pitch-by-pitch instead of observing the average outcome approach could be an alternative methodology as well. Also, the addition of swing tracking data for each batter combined with the Statcast data could

allow for a single, more generalized model to be developed as opposed to individual ones to attempt to predict the performance of a batter given a specific pitcher matchup. Additionally, combining the methods explored in this project with the type of logistic regression models studied during the literature review could potentially forecast a specific pitcher and batter matchup with higher accuracy (Healy, 2015; Doo & Kim, 2018). Using the models developed in this project to isolate the quality of contact of batted balls in play, or xwOBABIP, of a certain pitch profile, and then combining it with a predicted strikeout and walk probability could improve upon the methodology. Finally, the current models can be combined to simulate the performance of teams matching up against each other or could be utilized for roster-building purposes to identify pitching profiles a team may struggle against. Similar analyses have been conducted in the past to predict team outcomes from various performance statistics, and the incorporation of this model could improve upon these approaches (Huang & Li, 2021).

Conclusion

In the pursuit of developing a model to predict the performance of a specific batter and pitcher outcome, the process began with reviewing the history of several statistics to describe batter performance in baseball, the potential impact of new sensor data used to track the game, and previous efforts conducted to predict various batter performance metrics. After this review, the performance statistic of xwOBA was chosen as the dependent variable in the prospective model. Once the 2023 MLB season sensor data collected from Statcast was examined, cleaned, and filtered, two different model-building techniques were explored to produce the 337 individualized batter models: linear regression and artificial neural networks. The subsequent models were evaluated against one another, and the most effective model, a neural network of 2 hidden nodes and a weight decay value of 11, was chosen. The 604 average pitcher profiles

constructed were applied to this final model to predict the performance of 203,548 unique batter and pitcher scenarios. While the predictive effectiveness of the final models appeared to be lacking, important insights were able to be gathered from the variable importance analysis. Some of these insights aligned with traditional assumptions within the game of baseball, like the importance of pitch location and vertical movement in batter performance, but others challenged the assumptions, like those related to handedness and pitch speed advantages. Overall, the models produced in this project led to a greater understanding of various batter's approaches and can be a powerful tool for team management.

References

Baseball Savant: Statcast. Baseball Savant. (2023). https://baseballsavant.mlb.com/

Brill, R. S., Deshpande, S. K. & Wyner, A. J. (2022). A Bayesian analysis of the time through the order penalty in baseball. *Journal of Quantitative Analysis in Sports*. https://doi.org/10.1515/jqas-2022-0116

Doo, W., & Kim, H. (2018). Modeling the Probability of a Batter/Pitcher Matchup Event: A Bayesian Approach. *PloS One*, *13*(10), e0204874–e0204874. https://doi.org/10.1371/journal.pone.0204874

Fellingham, G. W., & Fisher, J. D. (2018). Predicting Home Run Production in Major League Baseball Using a Bayesian Semiparametric Model. *The American Statistician*, *72*(3), 253–264. https://doi.org/10.1080/00031305.2017.1401959

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Götze, T., Gürtler, M., & Witowski, E. (2023). Forecasting Accuracy of Machine Learning and Linear Regression: Evidence from the Secondary Cat Bond Market. *Journal of Business Economics*, *93*(9), 1629–1660. https://doi.org/10.1007/s11573-023-01138-8

Günther, F., & Fritsch, S. (2010). Neuralnet: Training of Neural Networks. *The R Journal*, *2*(1), 30–38. https://doi.org/10.32614/rj-2010-006

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. https://doi.org/10.1109/TKDE.2015.2416735

Healey, G. (2017). The New Moneyball: How Ballpark Sensors Are Changing Baseball. *Proceedings of the IEEE*, *105*(11), 1999–2002. https://doi.org/10.1109/JPROC.2017.2756740

Healey, G. (2020). Combining Radar and Optical Sensor Data to Measure Player Value in Baseball. *Sensors (Basel, Switzerland)*, *21*(1), 64–. https://doi.org/10.3390/s21010064

Huang, M.-L., & Li, Y.-Z. (2021). Use of Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches. *Applied Sciences*, *11*(10), 4499–. https://doi.org/10.3390/app11104499

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, *70*(2), 144–156. https://doi.org/10.4097/kjae.2017.70.2.144

Laby, D.M., Kirschen, D.G., Govindarajulu, U. *et al.* The Effect of Visual Function on the Batting Performance of Professional Baseball Players. *Sci Rep* **9**, 16847 (2019). https://doi.org/10.1038/s41598-019-52546-2

Lane, F. C. (1916). Why the System of Batting Averages Should Be Changed. *Baseball Magazine*, *16*, 41–47.

Nielsen, M. A. (2015). Improving the way neural networks learn. In *Neural Networks and Deep Learning*. essay, Determination Press.

Qualifier. (2010). Baseball Reference. Retrieved November 26, 2023, from https://www.baseball-reference.com/bullpen/Qualifier

Salfino, M. (2020). Stabilizing rates: When can we start trusting stats in a 60-game season? *The Athletic*. Retrieved November 26, 2023, from

https://theathletic.com/1917950/2020/07/09/which-statistics-should-fantasy-gms-lean-on-in-a-short-season/?access_token=1093576&redirected=1

Slowinski, P. (2010, February 15). *wOBA*. Fangraphs. https://library.fangraphs.com/offense/woba/

Statcast Search CSV Documentation. (n.d.). Baseball Savant. Retrieved October 28, 2023, from https://baseballsavant.mlb.com/csv-docs

Umemura, K., Yanai, T. & Nagata, Y. (2021). Application of VBGMM for pitch type classification: analysis of TrackMan's pitch tracking data. *Japanese Journal of Statistics and Data Science*, *4*, 41–71. https://doi.org/10.1007/s42081-020-00079-8

Vock, D. & Vock, L. (2018). Estimating the effect of plate discipline using a causal inference framework: an application of the G-computation algorithm. *Journal of Quantitative Analysis in Sports*, *14*(2), 37-56. https://doi.org/10.1515/jqas-2016-0029

Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., & Goyal, H. (2018). Opening the Black Box of Neural Networks: Methods for Interpreting Neural Network Models in Clinical Applications. *Annals of Translational Medicine*, *6*(11), 216–216. https://doi.org/10.21037/atm.2018.05.32

Appendix

The R source code written to complete this project can be found in the corresponding Capstone repository at https://github.com/ayyfinn/DS785_Capstone/.