PREDICTION OF A MAJOR LEAGUE BASEBALL STRIKE USING SITUATIONAL FACTORS

By

TOM EVERT AND SOPHIA TORRES

CAPSTONE PROJECT

Submitted in partial satisfaction of the requirements for the degree of

BACHELOR OF SCIENCE

in

STATISTICS

in the

COLLEGE OF SCIENCE

at

CALIFORNIA STATE UNIVERSITY,
MONTEREY BAY

Approved:

(Advisor)
Dr. Steven Kim
Department of Mathematics and Statistics

Spring 2022

Acknowledgements

We would like to thank the Math and Statistics faculty and staff for making this research and analysis possible in completion of our undergraduate education. Specifically, we want to thank Dr. Alison Lynch, Dr. Judith Canner, and Dr. Steven Kim. These professors provided a well-balanced yet challenging education that gives us the confidence to get through future education and professional goals.

Abstract

Pitch prediction has become one of the most common uses for strategic statistics in baseball as analytics rise in popularity. There is limited research on outcome pitch prediction (strike or ball prediction), but pitch type prediction has been the source of much baseball-based analysis. The primary objective of this analysis was to construct an accurate predictive model to determine outcome of a pitch before it is thrown. In doing so, we determined if this outcome is affected by several situational factors evaluated. Over 2 million pitches over the course of four seasons were evaluated and 11 situational variables were used as potential predictors. Note that many of the available potential variables were excluded from analysis because they occur too close to the time of the event. However, we were still able to consider several situational factors that could be recorded before hand and still impact pitch outcome. Many models were compared against each other for accuracy of prediction including 7 logistic regression models and two mixed effects models. The mixed effects models were not found to have an substantial impact on accuracy of prediction when accounting for pitcher variability, so we proceeded with the simplest and most accurate logistic regression model. We used the count of the at bat (number of balls and strikes thrown to that batter) and the hand dominance of the batter to predict the outcome of the pitch. After cross-validating and testing, we found that our most accurate model was about 54% accurate. While we were discouraged with our low accuracy levels, we can attribute this to a professional pitchers intentions to be unpredictable. We still found that our model can be more accurate than a coin toss for pitch prediction, and can be a good building block for further strategic analysis. We recommend a pitch prediction model for both offensive and defensive strategies, and from the data and tested models, further analysis could provide specific data on individual pitchers, as well as their performance under the pressure of different situations.

Contents

Acknowledgements ii

Abstract iii

1 Introduction 1

2 Methods 3

2.1 Data 3

2.2 Model Building Strategies 4

2.2.1 Logistic Regression 4

2.2.2 Cross-Validation and Model Accuracy Testing 5

2.2.3 General Linear Mixed-Effects Modeling (GLME) 6

3 Results 7

3.1 Descriptive Statistics 7

3.2 Cross-Validation Model Accuracy Testing 9

3.3 GLME Modeling Accuracy Testing 10

3.4 Model Parameter Estimates 11

4 Discussion 12

4.1 Practical Applications 15

4.2 Future Considerations 16

5 Conclusion 17

1 Introduction

The strike zone of Major League Baseball has become the source of much controversy with digital strike zones superimposed on the TV of every American watching a ballgame at home. According to a Boston University observational study, 34,294 calls were missed from behind home plate in the 2018 season (Williams, 2019). There are already numerous software used in baseball analytics to determine the path of a pitch and where it ends up after it is thrown. However, the purpose of this analysis is to determine that pitch's outcome before a ball leaves a pitcher's hand. Pitch prediction has also seen its fair share of the spotlight with the Houston Astros infamous cheating scandal in 2017, where the team used cameras to spot opposing team hand signals, and forms of noise communication to relay what pitch was coming to the batter. This form of pitch prediction is far more specific, and illegal, than the analysis of given factors, available to all, to predict the outcome of a pitch.

From the perspective of a coach and a batter, it would be beneficial to have some idea of whether the next pitch will be a strike, or a ball. The concern is not to predict whether the ball will be hit, but the aim is to predict the location of the next pitch based on situational factors. The location of the pitch will determine if it will be a strike or a ball. This prediction can be of benefit to both the team hitting the ball and the team fielding the ball. Pitch location is a valuable tool defensively and offensively, and can be used in a legal sportsmanlike manner by using readily available information to make a prediction (Haque, 2020). Since a pitch thrown in the strike zone is usually easier to hit, players in the infield and outfield can make the appropriate adjustments to get ready for a hit. On the other hand, the prediction can help the team hitting the ball by aiding in their decision to swing the bat or not, depending on if it predicts strike or ball. Statistics are on the rise year by year, with player specific defensive shifts increasing from 12.1% in 2017 to 34.1% in 2020 (Ishii, 2021). Literature suggests that a binary predicted outcome for a pitch is an entirely possible goal, with predictions such as fast-ball or off-speed, strike or ball, etc. represented as possible outcomes (Plunkett, 1998). While baseball might be a game of bias and unknown factors, statistics can be an extremely helpful tool in creating an even more strategic game.

Increased knowledge and feedback of situational strategy has been shown to create impact neurological focus, according to a study done at Michigan State and Illinois Wesleyan Universities (Themanson, Bing, Sheese, & Pontifex, 2019). Baseball is “the leader in applying quantitative methods to the sporting world for the past [twenty] years,” and statistical analysis like pitch prediction can be extremely useful in gameplay and strategy (Kim, 2020). With more information available, pitch prediction can become a part of an already strategic game in the form of immediate feedback to the batter. Studies have shown that ”advance information about the next pitch type decreased the timing error” which can improve accuracy of the batters timing and decision (Kidokoro, Matsuzaki, & Akagi, 2020). A pitch prediction model has been proven to be useful for different defensive strategies, but it is difficult to find the best one. We hypothesized that the probability of a strike can be explained by several situational factors, but that this type of pitch prediction will be very difficult. With that being said, the objective of this data analysis is to produce a best predictive model that determines whether a pitch will be a strike or a ball based on different situational factors. The secondary objective of this analysis is to create a functional application that will receive the situational factors that precede the pitch and then output the predicted outcome.

2 Methods

2.1 Data

The data used in our analysis was retrieved from a Major League Baseball database imported into Kaggle's data warehouse. The data holds every individual pitch thrown in the Major League regular season from the years 2015 to 2018. There were about 2.1 million pitches observed from 1,331 pitchers across 51 variables recorded in the data set. Player identification and at-bat data were matched by the at-bat ID, and from this individual ID number, data regarding the situation prior to the pitch, as well as the pitch itself, were recorded for analysis. To address our main objective, we only used data describing the situation pre-pitch, including the hand dominance of the pitcher and batter, the number of balls and strikes recorded at the time of the pitch, and whether a runner was on first base, second base, third base, or any combination of the three. All categorical variables were transformed into binary numeric variables for ease of analysis, including hand dominance and pitch type. The variable "type" was then recorded after the pitch was thrown, acting as the result of the pitch. An "X" was recorded for a hit, a "B" for a ball, and "S" for a strike. To create the necessary binary outcome of the pitch, the data for balls hit into play was manipulated to match pitches located in the strike zone to a strike, and pitches located out of the strike zone to a ball. The umpire's call will still stand for called balls and strike in order to account for pitches the batter swung at outside of the strike-zone, which still indicate a strike. Once the variable inconsistency was rectified, the data set was split into two groups: 2015-2017 data was renamed as training data, and 2018 data was renamed as testing data to be used for model validation down the line.

2.2 Model Building Strategies

To begin our analysis, we used the training data set which consisted of 15 randomly selected pitchers of different skill levels. Skill level was evaluated using pitch count over the whole season, as well as the pitcher's earned run average, or ERA. From each skill level, high, medium, and low, 5 pitchers were randomly selected using a random number generator. Using logic and prior knowledge of the game, 9 different predictive models overall were created, to be evaluated for each of the 15 pitchers. The models were evaluated using an accuracy calculation, to determine which out of the 9 choices had the highest accuracy. We define accuracy as the proportion that the predicted binary outcome and the actual outcome match. The models were then cross validated across 50 randomly selected pitchers, taking care to evaluate that each pitcher had pitched in all 4 seasons recorded. The first seven models were logistic regression models for prediction of a binary outcome and the final three were mixed effects models.

2.2.1 Logistic Regression

For notation purposes, we let π_{Mi} denote the probability of a strike under model M_i, for $i = 1, 2, ..., 7$ (i.e., `type` = 1). We say, `inning` represents the inning of the game. Let, `b_score` denote the score of the team at bat at time of pitch, `b_count` represent the number of balls called during the at bat at time of pitch, `s_count` represent the number of strikes called during the at bat, and `pitch_num` denote the number pitches thrown during the at bat. Also let, `on_1b` indicate if there is a runner on first base (i.e., 1 if a runner is on base and 0 otherwise), `on_2b` for a runner on second, and `on_3b` for a runner on third. Finally, let `outs` be the number of outs achieved during that inning before the pitch, `p_throws` denote the dominant hand of the pitcher and `stand` denote the dominant hand of the batter (i.e., 1 if a player is right-hand dominant, 0 otherwise). Given the aforementioned variable names, the seven logistic regressions are given below:

$$
\begin{aligned}
\pi_{M1} &= \frac{e^{\beta_0+\beta_1\texttt{b\_score}+\beta_2\texttt{b\_count}+\beta_3\texttt{s\_count}+{}_4\texttt{pitch\_num}}}{1+e^{\beta_0+\beta_1\texttt{b\_score}+\beta_1\texttt{b\_count}+\beta_3\texttt{s\_count}+{}_4\texttt{pitch\_num}}}\\
\pi_{M2} &= \frac{e^{\beta_0+\beta_1\texttt{on\_1b}+\beta_2\texttt{on\_2b}+\beta_3\texttt{on\_3b}+{}_4\texttt{outs}}}{1+e^{\beta_0+\beta_1\texttt{on\_1b}+\beta_2\texttt{on\_2b}+\beta_3\texttt{on\_3b}+{}_4\texttt{outs}}}\\
\pi_{M3} &= \frac{e^{\beta_0+\beta_1\texttt{stand}}}{1+e^{\beta_0+\beta_1\texttt{stand}}}\\
\pi_{M4} &= \frac{e^{\beta_0+\beta_1\texttt{stand}+\beta_2\texttt{on\_1b}+\beta_3\texttt{on\_2b}+\beta_4\texttt{on\_3b}+{}_5\texttt{outs}+\beta_6\texttt{b\_score}+\beta_7\texttt{b\_count}+\beta_8\texttt{s\_count}+{}_9\texttt{pitch\_num}}}{1+e^{\beta_0+\beta_1\texttt{stand}+\beta_2\texttt{on\_1b}+\beta_3\texttt{on\_2b}+\beta_4\texttt{on\_3b}+{}_5\texttt{outs}+\beta_6\texttt{b\_score}+\beta_7\texttt{b\_count}+\beta_8\texttt{s\_count}+{}_9\texttt{pitch\_num}}}\\
\pi_{M5} &= \frac{e^{\beta_0+\beta_1\texttt{s\_count}+\beta_2\texttt{b\_count}+\beta_3\texttt{stand}}}{1+e^{\beta_0+\beta_1\texttt{s\_count}+\beta_2\texttt{b\_count}+\beta_3\texttt{stand}}}\\
\pi_{M6} &= \frac{e^{\beta_0+\beta_1\texttt{on\_3b}+\beta_2\texttt{outs}+\beta_3\texttt{pitch\_num}}}{1+e^{\beta_0+\beta_1\texttt{on\_3b}+\beta_2\texttt{outs}+\beta_3\texttt{pitch\_num}}}\\
\pi_{M7} &= \frac{e^{\beta_0+\beta_1\texttt{inning}+\beta_1\texttt{on\_3b}+\beta_3\texttt{on\_2b}+{}_4\texttt{on\_1b}+\beta_5\texttt{outs}+\beta_6\texttt{pitch}}}{1+e^{\beta_0+\beta_1\texttt{inning}+\beta_1\texttt{on\_3b}+\beta_3\texttt{on\_2b}+{}_4\texttt{on\_1b}+\beta_5\texttt{outs}+\beta_6\texttt{pitch}}}.
\end{aligned}
$$

Under each model, π_{M_i}, where $i \in \{1,2,3,4,5,6,7\}$ indicates the referenced model, represents the probability of a strike for a particular pitch, as evaluated by the different situational factors in each model. Each of the models went through an accuracy test, evaluating the fitted values against the observed values found in the training data. Then each model further underwent cross validation against 50 randomly selected pitchers. The pitchers were randomly picked from a pool of pitchers that had played in all seasons of both the testing and training data. Once cross validation was performed, the three most accurate models were validated using the testing data set.

2.2.2 Cross-Validation and Model Accuracy Testing

To begin validating our developed models we will need to test them on a subset of the population (i.e., 50 randomly selected pitchers from 2016). Then we will build the models using this subset of the population however, we only use data from 2015, 2016, and 2017. Data from 2018 will be used to test the formed models. This process of model building is known as cross-validation.

After we have build models, we need a way to test how accurate the model predicts a strike. Note that, this is a logistic regression which estimates values of probability $0 < \hat{\pi} < 1$. The goal is to predict the Bernoulli random variable, `type = 1`, which indicates a strike. So, we need a rule that will call our predicted value a strike (`type = 1`) when $\hat{\pi} > \textit{threshold}$. The threshold is the smallest value of $\hat{\pi}$ for which we will call the predicted probability a strike. By applying this rule we can compare our predicted pitch outcome to the actual pitch outcome observed in the cross-validation data subset.

Additionally, using computational software like R, we can perform these accuracy tests over many possible threshold values. Then, we will select the threshold value that returns the largest accuracy as the final threshold. Tabulating these results for each model will aid us in selecting the more desirable models. While the accuracy measures the comparison between predicted and observed values, the threshold at which the highest accuracy occurs is valuable information as well. This value tell us the range of possible values of the predicted strike probability ($\hat{\pi}$) for which we feel safe calling a strike (`type = 1`).

2.2.3 General Linear Mixed-Effects Modeling (GLME)

Using the same notation, we composed the final 3 models as general linear mixed-effects models. Let π_i be the probability of a strike for the ith pitcher given situational factors. The GLME models can be expressed as,

$$\pi_{M8_i} = \frac{e^{\beta_0 + b_{0_i} + \beta_1 \texttt{b\_score} + \beta_2 \texttt{b\_count} + \beta_3 \texttt{s\_count} + {}_4 \texttt{pitch\_num}}}{1 + e^{\beta_0 + b_{0_i} + \beta_1 \texttt{b\_score} + \beta_1 \texttt{b\_count} + \beta_3 \texttt{s\_count} + {}_4 \texttt{pitch\_num}}}$$

$$\pi_{M9_i} = \frac{e^{\beta_0 + b_{0_i} + \beta_1 \texttt{stand} + \beta_2 \texttt{on\_1b} + \beta_3 \texttt{on\_2b} + \beta_4 \texttt{on\_3b} + {}_5 \texttt{outs} + \beta_6 \texttt{b\_score} + \beta_7 \texttt{b\_count} + \beta_8 \texttt{s\_count} + {}_9 \texttt{pitch\_num}}}{1 + e^{\beta_0 + b_{0_i} + \beta_1 \texttt{stand} + \beta_2 \texttt{on\_1b} + \beta_3 \texttt{on\_2b} + \beta_4 \texttt{on\_3b} + {}_5 \texttt{outs} + \beta_6 \texttt{b\_score} + \beta_7 \texttt{b\_count} + \beta_8 \texttt{s\_count} + {}_9 \texttt{pitch\_num}}}$$

$$\pi_{M10_i} = \frac{e^{\beta_0 + b_{0_i} + \beta_1 \texttt{s\_count} + \beta_2 \texttt{b\_count} + \beta_3 \texttt{stand}}}{1 + e^{\beta_0 + b_{0_i} + \beta_1 \texttt{s\_count} + \beta_2 \texttt{b\_count} + \beta_3 \texttt{stand}}}$$

where b_{0_i} is assumed to follow $N(0, \tau)$ and represents the intercept parameter adjustment for each respective ith pitcher.

The GLME model allows us to operate under the idea that not all observations in the data are independent. Each at-bat ID is an individual, and a given pitcher throws multiple pitches which are captured in many different at-bat ID numbers. In using this model, we can account for individual pitcher-level variability. Once again, both models were evaluated using an accuracy threshold, and further cross validated against a random group of 50 pitchers from the data set (see Section 2.2.3). Once the accuracy of the GLME models are determined, we can asses its usefulness and evaluate on the 2018 subset of our original data, if needed.

3 Results

3.1 Descriptive Statistics

During the initial assessment of our large data set that covered every pitch thrown in Major League Baseball between 2015 and 2018 we found some interesting descriptive statistics. Because the variable that indicates whether was pitch was a strike or ball (i.e., `type`) has been converted into 1's and 0's we can find strike probabilities by simply calculating proportions. We found that within the entire data set the overall strike probability is, $\pi = 0.4576$. The strike and ball count variables (i.e., `s_count` and `b_count`, respectively) are used in a majority of our models and the conditional probabilities are presented in Table 3.1 given the number of strikes and balls.

	Strikes = 0	Strikes = 1	Strikes = 2
Balls = 0	0.4969	0.4095	0.3714
Balls = 1	0.4893	0.4444	0.4096
Balls = 2	0.5008	0.4718	0.4482
Balls = 3	0.5965	0.4809	0.4731

Table 3.1: Conditional strike probability given the at-bat count

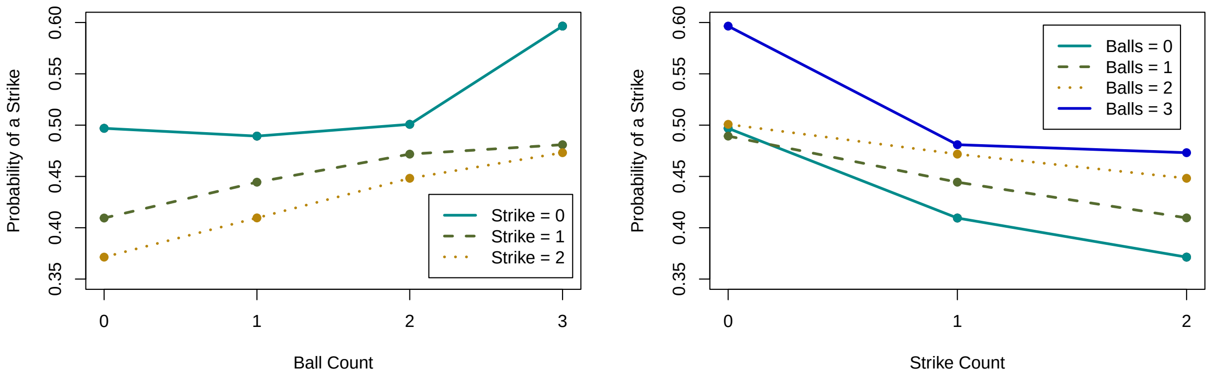

Figure 3.1: Strike probability given the at-bat count

In Table 3.1, at the beginning of an at-bat, when there are 0 balls and 0 strikes in the count, the probability of a strike is $\pi_{0,0} = 0.4969$. The highest probability is found when there are 3 balls and 0 strikes at $\pi_{3,0} = 0.5965$ and, the lowest probability is when there are 0 balls and 2 strikes at $\pi_{0,2} = 0.3714$. Note that, both of these scenarios are at both extremes of the at-bat sequence. There may be a more intuitive and strategical reason for this that we will discuss later. Furthermore, these conditional probability trends can be seen graphically in Figure 3.1.

We also looked into each pitcher's specific strike probability and wanted to compare that to the overall strike probability which does not factor for individual pitchers. This turns out to be another conditional probability of a strike given the pitcher's id number, $P(\texttt{type = 1}|\texttt{pitcher\_id})$. Below in Figure 3.2 we can see a histogram of all the MLB pitchers and their strike probability from 2015-2018. Notice that the histogram resembles a distribution centered around the mean (π_{avg}). In this case we estimate the average pitcher-specific strike probability to be $\pi_{p.avg} = 0.4388$. Recall that regardless of the pitcher, the strike probability is $\pi = 0.4576$. Hence, the difference is $\pi_{p.avg} - \pi = -0.0188$.

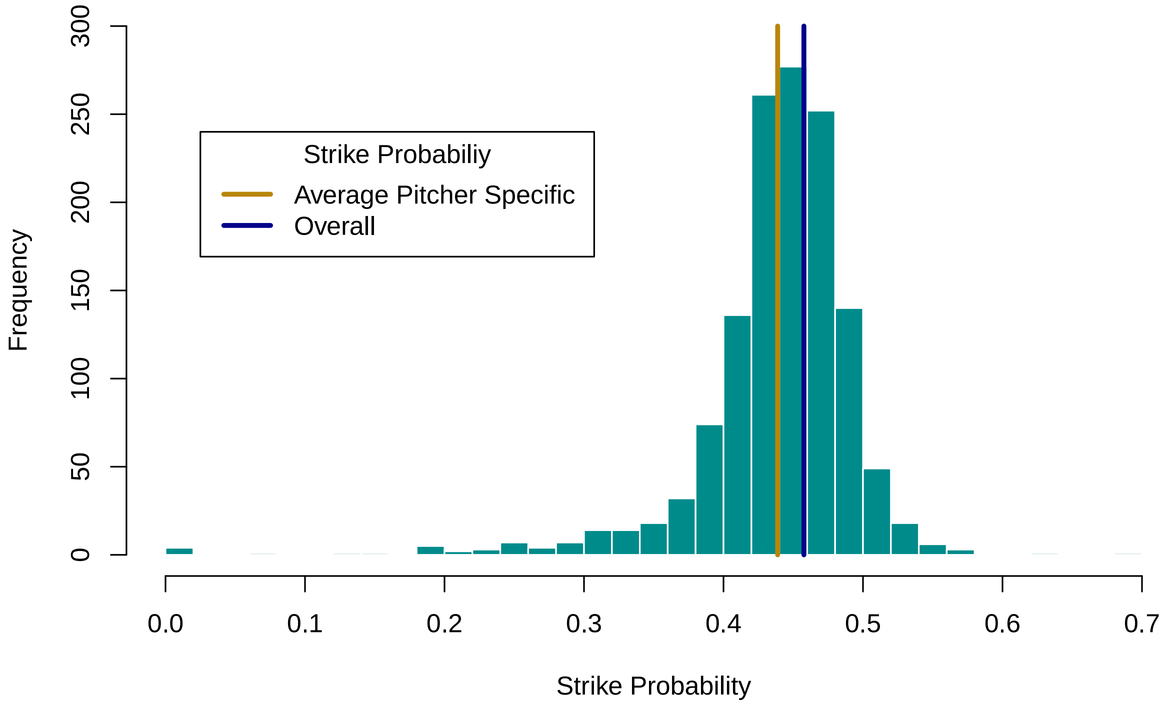

Figure 3.2: Histogram of pitcher specific strike probabilities within the MLB from 2015-2018

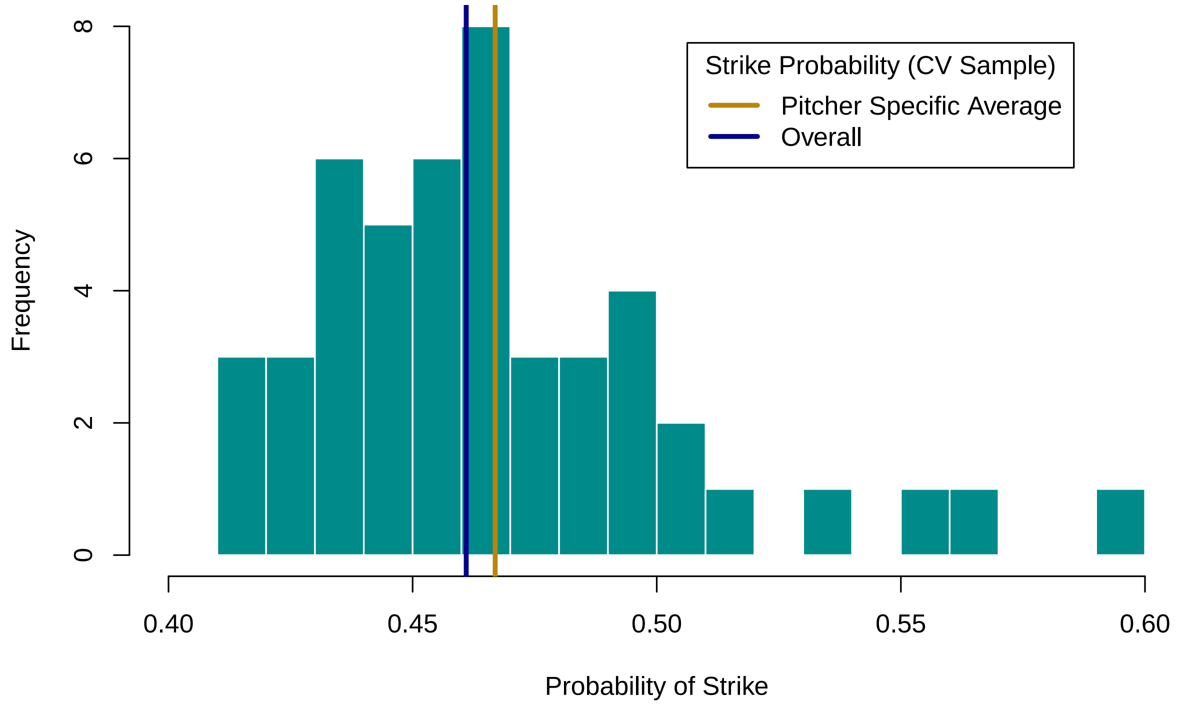

Figure 3.3: Histogram of pitcher specific strike probabilities within the cross-validation Sample

3.2 Cross-Validation Model Accuracy Testing

After selecting the cross-validation sample, we initially wanted to look at a histogram of these selected pitcher-specific strike probabilities in order to compare them to the overall cross-validation sample's strike probability regardless of the pitcher. We found that the average pitcher-specific strike probability is $\pi_{p.avg} = 0.4669$ and the same sample's strike probability regardless of the pitcher is $\pi_{CV} = 0.4610$. Hence, the difference is $\pi_{p.avg} - \pi_{CV} = 0.0059$ which is very close. Although this is a much smaller sample, we can see in Figure 3.3 that this sample's distribution is also centered near the mean.

In evaluating our models we wanted to take into account the AIC, MSE, and the accuracy of the model's predicted values. The accuracy of each model was found comparing predicted values to observed values. In Table 3.2 we can see that Models 1, 4, and 5 all have high accuracy in the cross-validation training sets but the degree of superiority is very small ($< 1\%$) when compared to the other models.

	Variables	AIC	Accuracy	Threshold
π_{M1}	4	338243	0.5415	0.63
π_{M2}	4	339654	0.5390	0.62
π_{M3}	1	339670	0.5390	0.62
π_{M4}	9	338147	0.5441	0.62
π_{M5}	3	338217	0.5417	0.63
π_{M6}	3	339580	0.5390	0.62
π_{M7}	6	339501	0.5391	0.62

Table 3.2: Prediction model assessment

	C-V Accuracy	2018 Accuracy	Threshold
π_{M1}	0.5415	0.5415	0.63
π_{M4}	0.5441	0.5436	0.62
π_{M5}	0.5417	0.5414	0.63

Table 3.3: Evaluation of selected models on 2018 data

Moving forward, we tested these models on the subset of the original data that represents all pitches in 2018. In Table 3.3 we can see that although model 4 has a slightly higher accuracy, and the accuracy test was done at a lower threshold at 0.62 compared to 0.63 for the others. In other words, model 4 predicts a strike when the predicted $\hat{\pi} > 0.62$. It is worth noting that although model 4 yielded the highest accuracy, model 5 is more desirable. From looking at both Tables 3.2 and 3.3 we can see that model 5's accuracy in cross-validation is only $0.5417 - 0.5441 = -0.0024$ and the change in mean square-error is $0.2469 - 0.2468 = 0.0001$. With such a large data set, the cross-validation and the actual validation with the 2018 data have very similar results. Considering, the simplicity of model 5, these differences mean we should expect fairly similar results and calculation under model 5 should be faster since there are less variables to input.

3.3 GLME Modeling Accuracy Testing

	C-V Accuracy	C-V Threshold	ME Accuracy	ME Threshold
Model 1	0.5415	0.63	0.5398	0.56
Model 4	0.5441	0.62	0.5407	0.54
Model 5	0.5417	0.63	0.5400	0.56

Table 3.4: Comparison to general linear mixed-effects (GLME) model

Since the objective of this study is to simply find the best prediction model, we tabulated our results from the general linear mixed-effects models which we applied to the better performing Models 1, 4, and 5 as seen in Table 3.4. Accounting for pitcher to pitcher differences, the GLME version of model 4 yielded the highest accuracy of 0.5407 at a threshold of 0.54. So, the mixed-effects model is predicting a strike when the predicted probability is greater than 0.54 ($\hat{\pi_{M4}} > 0.54$), compared to 0.62 under the original model 4. Also, the accuracy under the mixed-effects model is lower than the original ($0.5407 < 0.5441$). Although, we can predict strikes more often under the GLME model due to a lower threshold, we lose accuracy. Therefore, for prediction purposes, these models are not desirable.

3.4 Model Parameter Estimates

Because the GLME model results were not helpful, we move forward to asses the parameter estimates of our original logistic regression models 1, 4, and 5. These estimates along with their respective standard errors and p-values are found in Table 3.5. Immediately, we see that the larger model 4, whose accuracy was the highest. had p-values for parameter estimates that fairly high. In this regression $\hat{\beta}_2$, $\hat{\beta}_5$, and $\hat{\beta}_6$ had p-values of 0.1555, 0.2436, and 0.2734, respectively. Of the two simpler models, model 1 also had a parameter with a very high p-value. The parameter estimate for the score of the batter's team (i.e. `b_score`, $\hat{\beta}_1$) had a p-value of 0.7294. Fortunately, all of the parameter estimates for model 5's variables had p-values less than 0.0001, giving them all statistical significance.

	Variable	Parameter	Estimate	Standard Error	P-Value
Model 1	b_score	β_1	0.0002	0.0005	0.7294
	b_count	β_2	0.0181	0.0027	< 0.0001
	s_count	β_3	-0.0590	0.0030	< 0.0001
	pitch_num	β_4	0.0082	0.0023	0.0003
Model 4	stand	β_1	0.0129	0.0021	< 0.0001
	on_1b	β_2	-0.0033	0.0023	0.1555
	on_2b	β_3	-0.0160	0.0028	< 0.0001
	on_3b	β_4	-0.0158	0.0037	< 0.0001
	outs	β_5	0.0015	0.0013	0.2436
	b_score	β_6	0.0005	0.0005	0.2734
	b_count	β_7	0.0182	0.0027	< 0.0001
	s_count	β_8	-0.0596	0.0030	< 0.0001
	pitch_num	β_9	0.0084	0.0023	0.0003
Model 5	s_count	β_1	-0.0494	0.0013	< 0.0001
	b_count	β_2	0.0270	0.0012	< 0.0001
	stand	β_3	0.0127	0.0021	< 0.0001

Table 3.5: Model parameter estimates

4 Discussion

The results of our analysis confirm our suspicions that finding an accurate predictive model will be extremely difficult because of the level of the pitchers. A pitchers goal is to be precise and unpredictable. If a pitcher is good at their job, it will make our analysis difficult, and as we saw with our lower accuracy levels overall, perfect or near perfect prediction of a MLB pitch is unattainable. However, with predictive models we developed, we were able to achieve the best fit model using some logical situational factors that would aid in pitch outcome prediction.

Looking at Figure 3.1 and Table 3.1 we can make some useful inference regarding the strike probability given the at-bat strike count. We see the highest strike probability when there are 3 balls and no strikes (i.e. $\pi_{3,0} = 0.5965$). In this situation the pitcher is under pressure to throw a strike. If the following pitch is a ball, the batter advances to first base. To get the batter out, the next pitch must be a strike, which helps explain why this conditional probability is higher. Similarly, we see that the lowest strike probability is when there are no balls and 2 strikes (i.e. $\pi_{0,2} = 0.3714$). Here, the pitcher has more room to be creative since he is under less pressure. There is no immediate need for a strike, and a pitcher can thrown different types of pitches in order to get the batter to swing for the pitch. Different types of pitches will have different velocities, release angles, and final locations relative to the batter's strike zone. Professional pitchers can make the ball move and appear like a strike when in fact it will be a ball (e.g. "Curve Ball", "Slider", etc.). The lowered pressure can help make the pitcher more creative in their pitch decisions, but professional batters will also recognize a ball more often at this level of the sport.

Even though the data set included 51 variables, we were careful to select only the ones based on situational factors. We only considered variables that, from the coach's perspective, can be recorded prior to the next pitch being thrown so that it could be a functional coaching tool. Factors like speed, trajectory, or takeoff point were not taken into consideration. It is important to note that these other factors recorded while the pitch is in the air may yield different prediction modeling results. We speculate that if these other pitch factors improve predictability, we would need to rely on the skill of the batter to read and react to these input variables in milliseconds (the average pitch from a major league pitcher takes between 375-500 milliseconds to get to the catcher). So, in choosing our variables, we opted for models, both simple and complex, which included only pre-pitch situational factors. The main variables used in the analysis are depicted in Section 2. According to Occam's Razor, "when you have two competing theories that make exactly the same predictions, the simpler one is the better," (Gibbs & Hiroshi, 1996). This proved true in our analysis. As seen in Section 3, we observe that the simpler models are fairly accurate. The three models using the simplest combinations of variables, formed using logical observations of related factors provided the highest accuracy rating and were the ones we proceeded with in evaluation.

Recalling the histogram depicted in Figure 3.2, notice that the average of the pitcher-specific strike probabilities differs from the overall MLB strike probability by -0.0188. Hence, given the situational factors we have discussed, a coach could do a decent job of estimating an individual pitcher's strike probability ($\pi_{p.avg}$) by just assuming the overall strike probability (π). As surprising as it may seem this histogram shows us that a majority of MLB pitchers have a strike probability between 0.40 and 0.50 making this a reasonable estimation range for any unknown pitcher from the coach's perspective regardless of any factors. There is not enough variability from pitcher to pitcher to make the mixed effects model useful, so we proceed with the logistic regression models.

When we assess the results in Table 3.2 we found that Models 1, 4, and 5 all had the highest accuracy (i.e. 0.5415, 0.5441, and 0.5417, respectively). When we compare the number of variables involved in each model we see that model 4 which has the highest accuracy considers 9 variables whereas model 5 only considers 3. Allowing the coach to only consider 3 factors is much more desirable than 9. However, if we look further at Table 3.2 we can see that Model 3 only considers 1 variable, the stance of the batter (`stand`). This model also had one of the lowest accuracy, 0.5390 which is not much different than 0.5417 (i.e. Model 5). Therefore, it may not be practically beneficial to consider many factors, given the results in this table show that they are all roughly similar.

One more concern with the results found in Tables 3.2 and 3.3 is with the threshold level at which our model predicted a strike (i.e. `type` = 1). All of the models predicted a strike when the predicted value $\hat{\pi}$ was greater than 0.62 or 0.63. Having a wider threshold may help accuracy by capturing more possible predicted probabilities ($\hat{\pi}$). But, if it is too wide we would be considering lower predicted values (i.e. $\hat{\pi} < 0.5$). These predictions are closer to being a ball (`type` $= 0$) than a strike (`type` $= 1$). Given that all of our selected models had an accuracy slightly greater than 0.50 and that, in general we are considering all predicted probabilities ($\hat{\pi}$) greater than 0.62 to be a strike, it feels like we could perform just as well with the flip of a coin. Considering that Figure 3.2 shows that the overall strike probability regardless of pitcher is 0.4568, there does not seem to be much value in using the situational factors we considered for this study.

When we considered the results from the GLME Model, we were looking at adjusting the model parameters slightly for each individual pitcher. Notice in Table 3.4 that, for all the selected models, the measured accuracy is slightly lower than when we do not account for pitcher-to-pitcher differences. The threshold window did become larger with model 4 predicting strikes when values of $\hat{\pi}$ are between 0.54 and 1.0. But, because the accuracy did not change much, we believe that according to the prediction models we developed, there is not much difference in pitcher-specific strike probabilities given the situational factors we have discussed. And, when we refer to the histograms in Figures 3.2 and 3.3, we can also see that the actual observed overall strike probability is not much different than the average pitcher-specific strike probability ($\mid \pi_{p.avg} - \pi \mid < \pm 0.02$). Therefore, given the parameters of our developed models we have no evidence to support using the mixed-effects models. Estimation of pitcher-specific strike probabilities is more time consuming while producing similar results.

Finally, the model parameter results seen in Table 3.5, give us further evidence to support Model 5 as a best-fit model in terms of accuracy and simplicity, but also the low p-values of estimated parameter values. The p-value in a logistic regression tests the null hypothesis, $\beta_i = 0$ against the alternative, $\beta_i \neq 0$. Since all parameter estimates had p-values less than 0.0001, we say that we have less than a 0.01% chance of observing the test statistics for each parameter or more extreme, if the null hypothesis were true. Therefore the parameter estimates in Model 5 have strong statistical significance.

Furthermore, interpretation of these parameter estimates tell us that for $e^{\hat{\beta}_2}$ we see that the odds of a strike increases by 2.7% (i.e. 0.0270 times) for each pitch called a ball given that the strike-count and stance of the batter remain constant. We can also see evidence of this trend depicted in Figure 3.1. For each strike, as more balls are thrown to the batter, the strike probability increases. Similarly, we estimate for $e^{\hat{\beta}_1}$ that when the ball-count and stance of the batter remain the same we see the strike probability decrease by 4.94% (i.e. -0.0494 times) for each pitch called a strike. Again, this trend can be seen in Figure 3.1 where we see that when the ball count is constant, the probability of a strike decreases as more strikes are thrown. To a fan of baseball this may seem intuitive. Since the pitcher's goal is throw three strikes and get the batter out, as more pitches are called balls, the importance of throwing a strike increases. Once four balls are thrown, the batter can move to first base and the batter's team is closer to scoring a run. Inversely, as more pitches are called strike, the importance of throwing a called strike decreases. The pitcher can be more creative with their pitch selection since they are getting closer to a strike-out. Even though we have shown that these models are not relatively accurate, based on our assessment of accuracy and simplicity Model 5 is our best-fit model whose parameter estimates are all statistically significant.

Although these conclusion and results may seem discouraging it only strengthens the respect we must have for professional athletes. These pitchers on average, are able to distract themselves from the pressures of the game and focus on each pitch moment by moment. For the most part they do not let the at-bat strike count, the score, or even whose on base affect their poise and focus. They were selected to play professionally for these reasons. Maintaining mental focus is beneficial to maximizing one's athletic potential. For those who watch baseball on TV, or listen to it on the radio, notice how the announcers seem to illustrate the pressure these pitchers are on. What this study shows is that while the pressure of the situation may be enormous, the probability of throwing is strike should still be expected to be between 0.40 and 0.50.

4.1 Practical Applications

It is worth noting that there is not much existing literature on this topic. However, there are multitudes of articles on the other form of pitch prediction, predicting the type of pitch thrown (i.e., fastball, curveball, off-speed, etc.) or speed of a pitch. Most literature found on these types of pitch prediction prove that baseball is increasingly becoming more statistically and analytically based (Kim, 2020). With analytics coming to the forefront of baseball, a model like our predictive models can become a huge part of in-game strategy during baseball games. When turned into a function, our models can be transformed into a application in which a coach, player, or fan can input the expected factors: b_score, b_count, s_count, and pitch_num, or s_count, b_count, and stand for models 1 and 5 respectively. After these 3 to 4 variables have been determined, they are all we need to make a final prediction. It can be useful for both teams to predict. Often times, during baseball games, announcers aim to excite the audience by emphasizing high pressure situations, such as when the bases are loaded or it is the bottom of the ninth inning. However, our findings show that pressure like this is actually not statistically significant in determining whether the pitch is a ball or a strike. Something like our model application can somewhat undermine the announcers job of exciting the crowd, but it could be very interesting to see that theory in action. It might be useful to use the model in live scenarios to see if the pressure is affecting pitchers in the future. As analytics have become such a large part of baseball, it could benefit in the announcers booth as well to have this information.

4.2 Future Considerations

To take this study further we suggest that statisticians look into the other variables in the data set that take into account the physical characteristics of the pitch. As aforementioned, they are evaluated too close to the time of the event. We hypothesize that a more accurate predictive model might be obtained if we consider factors such as, speeds and accelerations in the x, y, and z directions, the release location and trajectory angle of the pitch, the spin rate of the ball, or even the spin direction of the thrown ball. There are so many factors to consider for which this great data set has recorded. However, when we set out to conduct this study we had hopes of developing some sort of computer application for which a coach could manually enter the pre-pitch situational factors we have discussed. The application could produce a predicted strike probability which could aid a coach's decisions. Factors that we suggest looking into further for more accurate models are happening in real-time, as the ball is in the air. A coach could develop predicted strike probabilities with the real-time pitch variables before the game and be able to aide the players in future games. Additionally, our model and data can be used to determine individual pitcher statistics. For example, this data can be used to evaluate which pitcher is best under high pressure situations (e.g. a full count), and equally as important, who is worst under that pressure. It could be reasonably evaluated which pitcher has the highest strike probability in future studies or analyses.

5 Conclusion

The final results of our analysis shows not only the normality of such a large data set, but also that major league baseball pitchers have been successful at what they do, disguising a pitch and maintaining poise . With a data set of about 2.1 million observations, we can imagine that we have caught most of the possible pitching scenarios. Note that the large sample size was helpful to detect the association, but it was not helpful to improve the prediction using situational factors for any practical significance. We observed that there is no completely accurate model, and that the accuracy of the models actually go down as complexity increases. Our analysis proved to validate our initial thoughts that the predictability of a strike would be challenging and the accuracy may be low. We recognize that this is because of the heightened skill of professional pitchers. Overall, although we were not able to come up with the most accurate predictive model which considers only situational factors, we were successful in proving that prediction will be difficult across the entire MLB. We know now that professional athletes can maintain focus and poise which makes it difficult to predict a strike based solely of the game-time situational factors. Showing this through the testing methods that we put each of our proposed models through was itself a good discovery from our research.

In the future, it may prove useful to use the same large data set and look at the type of pitch thrown, so that we may be able to align with the literature on pitch prediction, and create a more useful discovery. However, we can anticipate that even regardless of the skill level of the pitcher, pitch prediction will be extremely difficult. Made more difficult if one desires to exclude real-time pitch characteristics. In that case, providing a tool for coaches to use and update during the game may be unreasonable as data collection will cause challenges. Regardless of the methods used to make future models, if the pitcher is at a high level and extremely accurate, they will aim to be unpredictable with their movements and patterns. Similarly, if a pitcher is less skilled and more inconsistent, that inconsistency could create unaccounted for variability in the data. It follows then that, pitch prediction of any kind is extremely difficult, but even a model that can be over 50% accurate (as our best models were) can be beneficial in baseball strategy, as statistical analysis becomes more involved in the sport.

References

Gibbs, P., & Hiroshi, S. (1996). What is occam's razor? *University of California Riverside*. https://math.ucr.edu/home/baez/physics/General/occam.html

Haque, S. (2020). Baseball pitch prediction. *Towards Ai*, 1. https://towardsai.net/p/machine-learning/baseball-pitch-prediction

Ishii, B. (2021). Using pitch tipping for baseball pitch prediction. *California Polytechnic University, San Louis Obispo Department of Computer Science*, 1-59. https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=3865&context=theses

Kidokoro, S., Matsuzaki, Y., & Akagi, R. (2020). Does the combination of different pitches and the absence of pitch type information influence timing control during batting in baseball? *Public Library of Science*, *15*(3), 1. https://go-gale-com.csumb.idm.oclc.org/ps/i.do?p=AONE&u=csumb_main&id=GALE|A617767818&v=2.1&it=r

Kim, D. (2020). Predicting the next pitch uing artificial neural networks. *Towards Data Science*, 1. https://towardsdatascience.com/predicting-the-next-pitch-using-artificial-neural-networks-fc464383f53d

Plunkett, R. (1998). Pitch type prediction in major league baseball. *Harvard Library Office for Scholarly Communication*, 1151. https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364634

Themanson, J. R., Bing, N. J., Sheese, B. E., & Pontifex, M. B. (2019). The influence of pitch-by-pitch feedback on neural activity and pitch perception in baseball. *Journal of Sport Exercise Psychology*, *41*, 66-68. https://web-s-ebscohost-com.csumb.idm.oclc.org/ehost/pdfviewer/pdfviewer?vid=0&sid=58d2088e-29f5-41af-ab07-68af61fa9a89%40redis

Williams, M. T. (2019). Mlb umpires missed 34,294 ball-strike calls in 2018. bring on robo-umps? *BU Today*. https://www.bu.edu/articles/2019/mlb-umpires-strike-zone-accuracy/