SMU Data Science Review

Volume 8
Number 1 *Spring 2024*

Article 9

Baseball Decision-Making: Optimizing At-bat Simulations

Varun Gopal
Southern Methodist University, varung@mail.smu.edu

Krithika Kondakindi
Southern Methodist University, kkondakindi@mail.smu.edu

Nibhrat Lohia
Southern Methodist University, nlohia@mail.smu.edu

Morgan Williams
Southern Methodist University, the.mcw18@gmail.com

Baseball Decision-Making: Optimizing At-bat Simulations

Varun Gopal, Krithika Kondakindi, Christopher Williams, Nibhrat Lohia
Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
varung@mail.smu.edu, kkondakindi@mail.smu.edu, chriswilliams@mail.smu.edu

1 Introduction

The integration of analytics, statistics, and data science into baseball represents a profound shift in how the game is understood, played, and strategized. This journey, spanning over several decades, has evolved from rudimentary statistical records to the sophisticated application of machine learning and artificial intelligence. The inception of baseball analytics can be traced back to the early 20th century, with the pioneering work of individuals like Henry Chadwick, who introduced fundamental baseball statistics such as batting average and earned run average (ERA). However, it was not until the latter part of the century that the field of sabermetrics emerged, fundamentally changing the sport's analytical landscape.

Sabermetrics, a term coined by Bill James in the late 1970s, refers to the empirical analysis of baseball through statistics that measure in-game activity. James and other sabermetricians sought to answer complex questions about baseball not addressed by traditional statistics, thus laying the groundwork for the data-driven approach that characterizes the sport today. The publication of James's "Baseball Abstracts" throughout the 1980s brought sabermetrics into the mainstream, challenging long-held beliefs and strategies with statistical evidence.

The Moneyball era of the early 2000s, popularized by Michael Lewis's book "Moneyball: The Art of Winning an Unfair Game," marked a turning point in the acceptance and application of analytics in baseball. The Oakland Athletics' successful adoption of data-driven strategies to assemble a competitive team on a limited budget demonstrated the power of analytics to identify undervalued talent and optimize team performance. This approach has since been adopted and expanded upon by numerous Major League Baseball (MLB) teams, leading to an arms race in analytics departments seeking competitive advantages.

The advent of advanced tracking technologies in the late 2000s and early 2010s, such as PITCHf/x, Statcast, and Hawk-Eye, ushered in a new era of baseball analytics. These systems provide a wealth of data on every pitch and play, including ball trajectory, spin rate, player movement, and more, enabling a level of analysis that was previously impossible. The granularity and volume of this data have facilitated the development of sophisticated metrics and models.

Against this backdrop, the proposed development of a feedforward neural network-based simulation environment for modeling pitch outcomes represents the latest advancement in the ongoing integration of data science into baseball. This environment aims to harness the vast array of pitch data now available, applying neural network architectures to predict outcomes with high precision. By simulating various pitch types and conditions, this model will provide a foundational tool for further research, particularly in the application of reinforcement learning (RL) and deep learning to pitching strategies.

The feedforward neural network, characterized by its layer-by-layer processing of inputs to outputs without backloops, is particularly suited for this task due to its ability to model complex, nonlinear relationships in data. In the context of pitch outcome simulation, the network will learn to identify patterns and correlations within the pitch data, using this information to predict the likelihood of each possible outcome. This capability is critical for the subsequent application of RL algorithms, which require accurate, detailed simulations to learn and optimize strategies effectively.

The envisioned future of this research involves integrating the simulation environment with advanced RL and deep learning models. These models will iteratively interact with the simulated environment, learning to adjust pitching strategies to maximize favorable outcomes. This process not only promises to enhance the strategic depth of pitching but also represents a significant step forward in the application of artificial intelligence in sports.

The development of a neural network-based simulation environment for pitch outcomes is both a continuation of and a significant leap forward in the rich history of baseball analytics. From the early days of simple statistical records to the sophisticated data science applications of today, the field has consistently sought to deepen our understanding of the game. By leveraging the latest in machine learning and artificial intelligence, this research aims to further revolutionize baseball strategy, offering teams and coaches new tools to refine their approach to pitching in an increasingly competitive and data-driven sport.

2 Literature Review

2.1 Evolution of Data and Analytics in Baseball

Statcast Data. In recent years, Major League Baseball has embraced advanced technologies to revolutionize the way the game is played and analyzed. One significant innovation is the integration of the Statcast system across all MLB stadiums by the 2015 season. Statcast employees sophisticated Trackman phased-array Doppler radar technology suited behind the home plate to track the ball's path, spin rate, and velocity, along with detailed player movements on the field. Additionally, Stereoscopic optical video arrays positioned along the third base line provide precise player tracking data, enabling the quantification of performance metrics such as defender reaction time, route efficiency, and speed. This rich

data source not only enhances the fan experience with real-time information but also presents baseball organizations with new opportunities to refine their strategies and decision-making processes. Consequently, many clubs have invested in teams of statisticians to harness this wealth of data effectively (Mizels, Erickson, & Chalmers, 2022). The widespread adoption of Statcast data underscores the importance of technological innovation in sports analytics, highlighting the value of real-time data in enhancing player evaluation and game strategy.

2.2 Machine Learning Applications in Baseball

Machine Learning Systematic Literature Review. The systematic literature review (SLR) conducted by Koseler and Stephen (2017) delves into the growing intersection of machine learning and baseball analytics. The research team meticulously categorized and summarized the applications of ML in baseball, shedding light on the diverse methodologies employed in analyzing player and team performance. By framing their research questions within the context of the PICOC (Population, Intervention, Comparison, Outcomes, Context) criteria, the team focused on highlighting the population of baseball analysts, the intervention of ML techniques, the outcomes related to performance evaluation, and the broader context encompassing both academic research and industrial practice. This methodical approach ensured the accuracy and reproducibility of the study, providing a comprehensive overview of the current state of ML applications in baseball analytics.

Drawing from a pool of 145 candidate articles, Koseler and Stephen (2017) identified 32 articles that met their inclusion criteria, reflecting the growing interest in ML driven approaches in baseball analytics. Notably, Regression tasks emerged as the most dominant among the problem classes explored in the literature, with Support Vector Machines (SVMs) and K-nearest neighbors (KNNs) also emerging as leading algorithms. Bayesian inference was also featured prominently, particularly in Regression tasks. The team's synthesis of the extracted data revealed a diverse landscape of ML applications in baseball, ranging from predictive modeling of player performance to strategic decision making by teams. This comprehensive analysis serves as a valuable source for researchers seeking to navigate the complex terrain of ML in baseball analytics.

Looking ahead, Koseler and Stephan (2017) anticipate the growing influence of neural networks in baseball analytics, leveraging advancements in ML frameworks like TensorFlow and Torch. While SVMs and KNNs have been beneficial in handling discrete and data heavy nature of baseball datasets, the rise of neural networks holds promise for unlocking deeper insights and predictive capabilities.

2.3 Sabermetrics

Pitch Selection. Sabermetrics has significantly contributed to the evaluation of pitch selection strategies in the sport. Beneventano, Berger, and Weinberg (2012) explored the impact of sabermetrics on predicting run production and run prevention in baseball. This study sheds light on the importance of data-driven decision-making in pitch selection, emphasizing the need for pitchers to consider advanced metrics when choosing between inside and outside pitches.

In another study, Otremba (2022) introduced "SmartPitch," an applied machine learning approach aimed at optimizing professional baseball pitching strategy. While not directly discussed in the provided citation, this research highlights the growing trend of utilizing advanced analytics and machine learning techniques to enhance pitch selection strategies in contemporary baseball.

Additionally, Nakahara, Takeda, and Fujii (2023) conducted research on pitching strategy evaluation using stratified analysis and propensity score methods. Although the primary focus of this study is on assessing pitching strategies, the findings have implications for pitch selection. This study underscores the importance of balancing pitch types and locations to maximize effectiveness while avoiding predictability.

Plate Discipline. In recent baseball analytics trends, there has been a notable resurgence in exploring the impact of plate discipline on batter performance, as highlighted by Vock and Vock (2018). Plate discipline, often characterized by a batter's tendency to swing at pitches within or outside the strike zone, presents a nuanced yet crucial aspect of offensive play. However, existing metrics have primarily focused on simplistic binary classifications of pitches without fully capturing the subtleties of plate discipline. To address this gap, researchers have turned into causal inference frameworks to estimate the effect of altered plate discipline on batter outcomes. By employing methodologies such as the G-computation algorithm within a potential outcome framework, analysts can isolate the impact of plate discipline independent of batter's inherent hitting ability or the pitches encountered during plate appearances. This approach allows for a more accurate modeling of batter tendencies, enabling researchers to dive deeper into the nuanced relationship between plate discipline and offensive performance.

An illustrative application of this methodology, as demonstrated by Vock and Vock (2018), involves estimating the hypothetical performance of a batter under different plate discipline scenarios. By leveraging data collected through advanced tracking systems like the PITCHf/x, researchers can assess the potential improvement in offensive production if a batter were to adopt the plate discipline of another player. For instance, the study explores how Starlin Castro's offensive metrics might have differed had he demonstrated the plate discipline of Andrew McCutchen during the 2012–2014 seasons instead. Through careful modeling and analysis, researchers estimate the anticipated changes in batting average, on-base percentage, and slugging percentage, shedding light on the nuanced relationship

between plate discipline and batter performance. This methodological framework not only improves our understanding of individual player dynamics but also offers valuable insights into the broader factors influencing offensive outcomes in baseball.

In summary, sabermetrics and advanced statistical methods have greatly influenced the evaluation of pitch selection and plate discipline in baseball. These studies collectively emphasize the importance of data-driven decision-making for pitchers and batters, ultimately contributing to the ongoing evolution of the sport.

2.4 Methods

Neural Networks. Lee (2022) presents a novel approach to predicting pitch type and location in baseball using an ensemble model of deep neural networks. The methodology begins with an extensive data collection process, encompassing diverse game situations such as score differentials, count scenarios, and baserunner positions. This rich dataset serves as the foundation for training the prediction models, enabling the recognition of intricate patterns and correlations between various game contexts and pitch outcomes.

Subsequently, the model building phase entails the development of deep neural networks with multiple hidden layers, aimed at capturing the intricate relationships existing in the data. Lee utilizes advanced techniques like ReLU activation functions for hidden layers and logistic functions for output layers, alongside methodologies such as Xavier initialization to set initial weights effectively. Moreover, to mitigate the risk of overfitting, Lee implements strategies like repeated random sub-sampling validation (Monte Carlo cross-validation), ensuring thorough assessment of model performance and detection of overfitting instances. This iterative refinement process enhances the robustness and generalizability of the ensemble models, enabling accurate prediction of pitch types and locations across diverse game scenarios.

In essence, Lee's methodological framework highlights the importance of leveraging cutting-edge machine learning techniques, together with comprehensive datasets and thorough validation protocols, to develop precise and reliable prediction models in sports analytics. By integrating deep neural networks and ensemble modeling methodologies, Lee's research contributes to the advancement of predictive modeling capabilities in baseball, offering valuable insights for enhancing performance analysis and decision-making in the sport.

Markov Decision Process. In the context of baseball, a Markov Decision Process (MDP) is used to model the pitcher’s decision making during an at-bat. Each state in the MDP represents the pitcher’s current pitch selection, based on prior information such as the count,

previous pitch selection, batter's action, and pitch result. Transition probabilities in the MDP, estimated from pitch-by-pitch data, outputs the optimal batting strategies against specific pitchers. Reinforcement Learning algorithms can then be applied to find and evaluate these optimal strategies, considering both rewards and future outcomes. The practical use case for these strategies is demonstrated through simulations and analysis of real-world baseball data, suggesting that utilizing pitcher-specific behaviors can yield advantages.

The literature reviews provide a comprehensive overview of the evolution of data and analytics in baseball, tracing from its development from simple and straight forward statistical records to complex machine learning applications. Studies on sabermetrics and plate discipline emphasize the importance of data-driven decision making for pitchers and batters. Our research builds on this foundation by developing a neural network-based simulation environment for modeling pitch outcomes, aiming to enhance predictive modeling capabilities and decision-making in baseball strategy.

3 Methods

3.1 Data Sourcing

Play-by-play data was sourced from Statcast covering the 2017 – 2019 MLB seasons using MLB's Baseball Savant API. The data set is extensive with over 700,000 pitches for each full season. It includes everything from each player involved in the play, pitch count, and pre- and post-pitch scores to pitch release point, pitch velocity, and launch angle.

Table 1. Range of variables representing the batter's performance.

Variable Name	Description
xba	Expected batting average
xslg	Average expected slugging percentage
woba	Weighted on-base average
xwoba	Average expected woba
xobp	Expected on-base-percentage
xiso	The difference between the average expected slugging percentage and expected batting average

wobacon	woba given that the ball was in play
xwobacon	xwoba given that the ball was in play
bacon	Batting Average given that the ball was in play
xbacon	xba given that the ball was in play
exit_velocity_avg	Average exit velocity
launch_angle_avg	Average launch angle
z_swing_percent	Percentage of swings given the ball was in the strike zone
oz_swing_percent	Percentage of swings given the ball was outside the strike zone
iz_contact_percent	Percentage of contacts with a ball in the strike zone
oz_contact_percent	Percentage of contacts with a ball outside the strike zone
whiff_percent	Percentage of swings where the batter missed the ball

The MLB Stats API was also utilized to get the probable starting pitchers and lineups for all games in the seasons of interest. When selecting starting pitchers, only the names and player ID were needed for this analysis. Starting batters were selected with batter name and player ID and several other variables (Table 1) representing the batter's average performance over the season.

3.2 Dataset Creation and Modification

After merging the pitch-by-pitch data, pitcher data, and batter data by player ID, the data was further modified and filtered. The reported strike-zone for each batter varied from pitch to pitch. Therefore, for consistency in selecting adequate pitch locations, new features were created to represent the mean for both the top and bottom of the strike zone for each batter.

To represent the outcomes that could move the at-bat's ball-strike counter forward in an appropriate manner, a revised description variable was created. This feature categorized the outcome of a pitch into four distinct levels: ball, foul, strike, and into play.

Table 2. Pitch variables in latest dataset following feature engineering.

Variable Name	Description
release_pos_x	Horizontal (left-right) release point from the batter's point of view
release_pos_y	Horizontal (forward-backward) release point from the batter's point of view

release_pos_z	Vertical release point from the batter's point of view
pfx_x	Horizontal pitch movement in feet
pfx_z	Vertical pitch movement in feet
vx0	Average horizontal (left-right) pitch velocity in feet per second
vy0	Average horizontal (forward-backward) pitch velocity in feet per second
vz0	Average vertical pitch velocity in feet per second
ax	Average horizontal (left-right) pitch acceleration in feet per second
ay	Average horizontal (forward-backward) pitch acceleration in feet per second
az	Average vertical pitch acceleration in feet per second
effective_speed	Pitch speed in mph
release_spin_rate	The ball's spin rate after release
release_extension	Forward pitcher extension in feet
spin_axis	Axis of rotation for the ball in the x-z plane (left-right and up-down) from 0 to 360

After feature engineering, a new dataset was created only including the pitch location variables plate_x and plate_z, a one-hot-encoded representation of the number of balls and strikes prior to the current pitch, 15 pitch variables (Table 2), the batter variables, and the revised description variable. In this new dataset, all null values were deleted. In addition, all continuous variables were scaled by subtracting from the mean and dividing by the standard deviation of the given variable. The revised description feature was then one-hot-encoded and used as a target.

3.3 Outcome Modeling

The model, visualized in **Figure 1**, employs a feedforward architecture comprised of two hidden layers, each consisting of 128 neurons (Otremba 2022). The selection of features for the model includes a comprehensive set of pitch characteristics, the ball-strike count, pitch location, and batter performance variables, these were used to predict the revised description of the pitch's outcome, encoded as a one-hot vector. The mathematical foundation of the model is encapsulated in a succinct overall formula (1).

$$f(\boldsymbol{x}) = g \circ f_2 \circ f_1(\boldsymbol{x}) \tag{1}$$

Here, f_1 (2) and f_2 (3) represent the activations of the two hidden layers using the ReLU (Rectified Linear Unit) function, ensuring non-linearity in the model's learning process.

$$f_1(\boldsymbol{x}) = \boldsymbol{a}_1 = \max\left(0, \boldsymbol{W}_1 \cdot \boldsymbol{x} + \boldsymbol{b}_1\right) \tag{2}$$

$$f_2(\boldsymbol{a}_1) = \boldsymbol{a}_2 = \max(0, \boldsymbol{W}_2 \cdot \boldsymbol{a}_1 + \boldsymbol{b}_2) \quad (3)$$

The output layer, processed through a softmax function g (4), transforms the neural network's raw scores into probabilities, offering a probabilistic interpretation of each potential pitch outcome.

$$g(\boldsymbol{a}_2)_i = \frac{e^{\boldsymbol{w}_3^i \cdot \boldsymbol{a}_2^i + \boldsymbol{b}_3^i}}{\sum_{j=1}^{4} e^{\boldsymbol{w}_3^j \cdot \boldsymbol{a}_2^j + \boldsymbol{b}_3^j}} \quad (4)$$

This approach equips the model with the capability to inject an element of randomness into the simulation environment, a feature that proves instrumental when developing a reinforcement learning model utilizing this environment as its foundation.

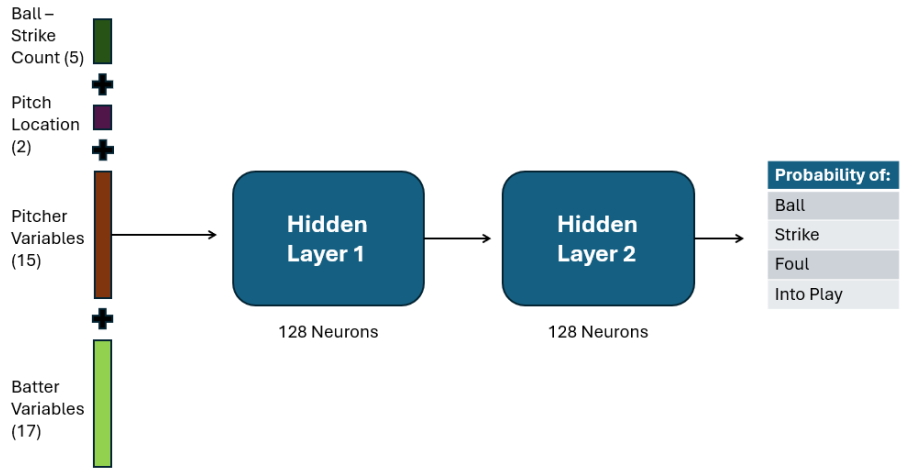

Fig. 1. The conceptual architecture of the neural network model.

The training process involves a 70/30 split of the data into training and test sets, ensuring the model's performance is validated on unseen data. By adopting this neural network architecture, the simulation environment provides a nuanced and predictive framework for analyzing and predicting the outcomes of pitches in baseball.

3.4 Simulation

Different pitchers have different arsenals, or pitch types, that they can be expected to throw in a game. In addition, a pitch type may be expressed differently from pitcher to pitcher. After selecting a pitcher and a given pitch type, pitches were simulated by sampling from a multivariate distribution fitted to all pitch variables of that pitch type thrown by the pitcher. This array was then scaled by subtracting from the average of the original data and dividing by the standard deviation for each pitch variable.

The game of baseball could theoretically be viewed through the lens of a Markov Decision Process (Otremba 2022). In this case, the states are represented by the ball and strike count, and actions are represented by the pitch's type and location relative to the strike zone. The initial state will always be represented as [0,0], and the act of selecting a pitch type creates a hypothetical pitch.

Pitch locations, at least in the action input, are represented as spaces on a grid. This grid was fitted to the strike zone of the selected batter, and then translated to a real pitch-location (Otremba 2022). In our case, the default grid consisted of a strike zone 6 spaces in width and 8 spaces in height with extensions outside the strike zone: 3 spaces on the left and right of the strike zone and 4 spaces on the top and bottom of the strike zone

Once the hypothetical pitch and true location were determined, they could be concatenated with a given batter's variables, and the ball and state count was passed through the neural network. The ball-strike counter would move up based on the pitch's outcome, changing the state: Balls increased the ball counter by 1, strikes increased the strike counter by 1, fouls increased the strike counter by 1 until the strike counter reached 2 (after which point the foul had roughly a 5% chance to increase the strike counter further), and hits into play ended the at bat. This process would repeat until either the ball counter reached 4 or the strike counter reached 3. Two possible terminal states for example, are [4, 0] and [0, 3].

Rewards were calculated based on the season's RE288 values – RE288 being the run expectancy conditional to the base loadout, number of outs, and the ball-strike count– when moving from one non-terminal state to another (Otremba 2022). A random out and base-loadout configuration was chosen for each at-bat, leaving the ball-strike counter to determine the RE288 value for each state. The reward for non-terminal states was then calculated by taking the difference of the RE288 value corresponding to the current state and the RE288 value corresponding to the next state. The reward for terminal states was one of two options: the batter's xWOBA if the terminal state was a strikeout or a terminal foul or the batter's negative xWOBA if the terminal state was a hit-into-play or a walk.

4 Results

To examine the model's effectiveness, the predicted outcomes must be compared to the actual outcomes in a few ways. If the predicted distribution of outcomes matches that of the

tested distribution, the assumption can be made that the model will predict a realistic distribution given random pitcher-batter matchups. By examining precision, recall, F1 score, and confusion matrices of the model, the predictive ability of the model can be assessed.

The predicted distribution of outcomes bears some similarity to the distribution of outcomes present in the test set. This similarity can be seen in figures 2, 3, and 4 as well as table 3. Initially, this suggests a frequency of predicted outcomes close to that of the real world. This would also suggest a realistic frequency of predicted outcomes for unknown pitcher-batter matchups.

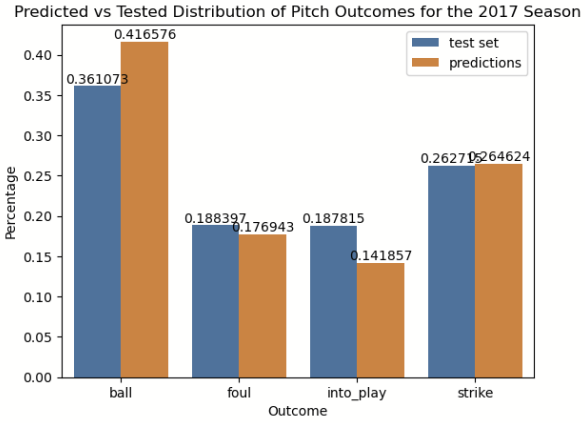

Fig. 2. Distribution of predicted outcomes compared to tested outcomes for the 2017 season.

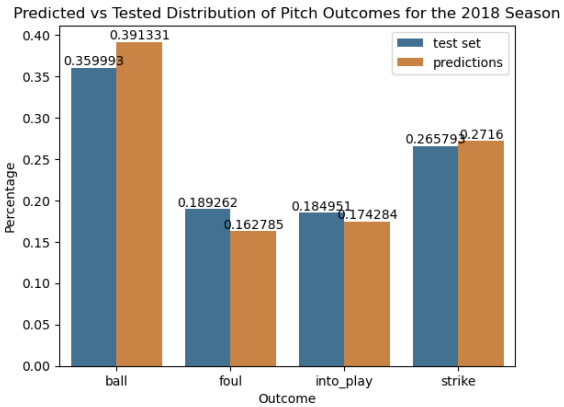

Fig. 3. Distribution of predicted outcomes compared to tested outcomes for the 2018 season.

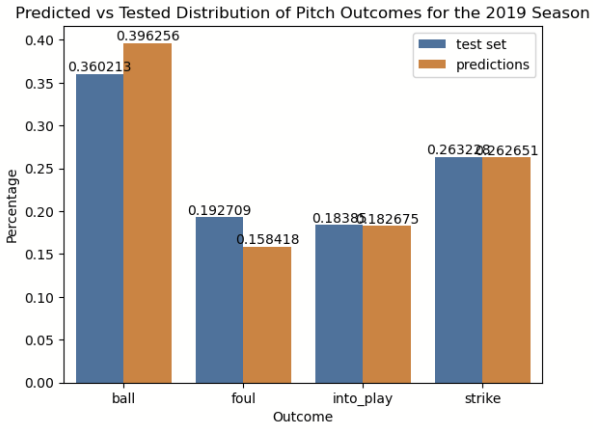

Fig. 4. Distribution of predicted outcomes compared to tested outcomes for the 2019 season.

The percentage difference between predicted and actual outcomes for the 2017, 2018, and 2019 seasons shows that the model tends to overpredict ball and strikes while underpredicting fouls and balls into play. While initially, predicted and test distributions may have seemed similar, the variation in outcomes must also be noted.

Table 3. Percent Difference between Real and Predicted Outcome Distributions

	Outcome			
Season	Ball	Foul	Into Play	Strike
2017	0.153717	-0.0608	-0.2447	0.007266
2018	0.087052	-0.139896	-0.05767	0.021848
2019	0.107442	-0.169	-0.00856	-0.01637

The precision and recall of the model show that for all seasons, balls were classified best, strikes were classified better than fouls and hits-into-play, and hits-into-play were classified better than fouls. The confusion matrices of true vs predicted outcomes support this conclusion as well. This data can be seen in tables 4, 5, and 6 as well as figures 5, 6, and 7.

Balls and strikes being the most well predicted outcomes makes sense, as these two can often be attributed to location relative to the strike zone: balls were likely calls if the

pitch was outside of the strike zone. Likewise, strikes occur much more often than hits into play and fouls, so if a ball is within the strike zone, it is likely that the pitch will be called a strike. Fouls and hits into play could be confused with each other and strikes because of their similarity in characteristics; two pitches may have similar variables and pitch location paired with batter contact and may have different outcomes.

Table 4. Classification Report of Predicted vs Tested Outcomes for the 2017 Season

	Precision	Recall	F1-Score	Support
Ball	0.75	0.87	0.81	35370
Foul	0.38	0.36	0.37	18455
Into Play	0.43	0.33	0.37	18398
Strike	0.53	0.53	0.53	25735
Accuracy			0.58	97958
Macro Avg	0.52	0.52	0.52	97958
Weighted Avg	0.56	0.58	0.57	97958

Table 5. Classification Report of Predicted vs Tested Outcomes for the 2018 Season

	Precision	Recall	F1-Score	Support
Ball	0.78	0.84	0.81	34658
Foul	0.38	0.33	0.35	18221
Into Play	0.42	0.39	0.40	17806
Strike	0.53	0.54	0.54	25589
Accuracy			0.58	96274
Macro Avg	0.53	0.53	0.53	96274

Weighted Avg	0.57	0.58	0.58	96274

Table 6. Classification Report of Predicted vs Tested Outcomes for the 2019 Season

	Precision	Recall	F1-Score	Support
Ball	0.78	0.86	0.82	32994
Foul	0.39	0.33	0.36	17769
Into Play	0.42	0.41	0.41	16812
Strike	0.52	0.51	0.52	24315
Accuracy			0.58	91890
Macro Avg	0.53	0.53	0.53	91890
Weighted Avg	0.57	0.58	0.57	91890

As stated earlier, the precision and recall of the model's predictions suggest greater effectiveness predicting balls than strikes and greater effectiveness predicting strikes than fouls or hits into play. The confusion matrices below (Figures 4, 5, and 6) support this as well. When the model predicted balls, it was unlikely that the actual outcome was anything else. When the model predicted strikes, the actual outcome was most likely to be a strike, almost equally likely to be a hit into play or foul ball, and least likely to be a ball. The model did, however, tend to confuse hits into play and fouls. Again, this was not unexpected, as the model inputs for a hit into play and a foul would have likely been very similar.

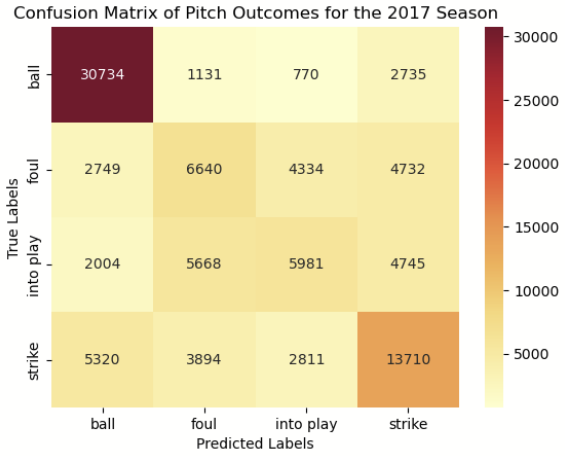

Fig. 5. Confusion Matrix of Tested vs Predicted Pitch Outcomes for the 2017 Season

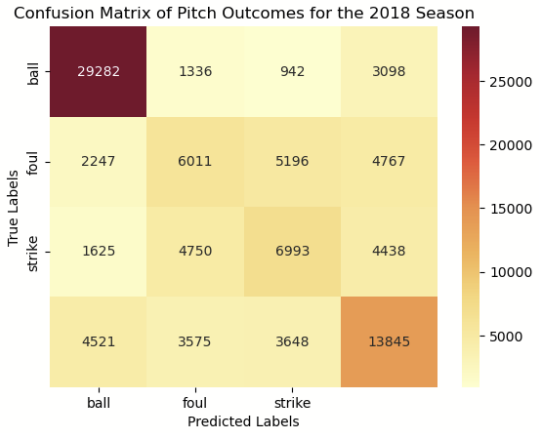

Fig. 6. Confusion Matrix of Tested vs Predicted Pitch Outcomes for the 2018 Season

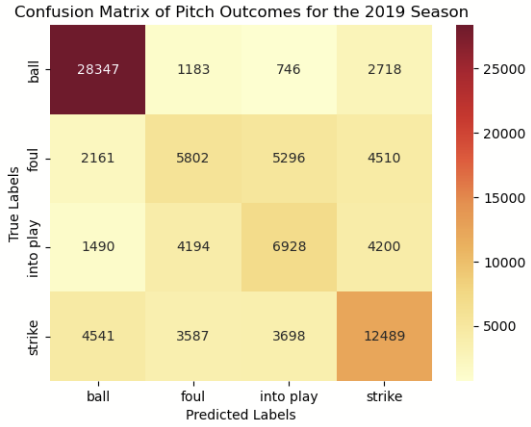

Fig. 7. Confusion Matrix of Tested vs Predicted Pitch Outcomes for the 2019 Season

5 Discussion

The integration of advanced statistical methods and machine learning techniques in baseball has transformed the sport's landscape, offering teams unprecedented insights into player performance and game strategies. This research contributes to this ongoing evolution by proposing a unique approach to enhancing batting strategies. However, there are ethical considerations that must be addressed.

One significant concern is the potential overreliance on statistical models, which may overlook the nuanced human elements of the game. While machine learning models can provide valuable insights, they should not replace the expertise and intuition that coaches and players bring to the table. It is essential to strike a balance between data-driven decision making and traditional coaching and strategy methods to ensure the complete development of players and the integrity of the game.

Additionally, there are broader ethical implications regarding data privacy and player consent that goes beyond just the game of baseball. As the use of advanced tracking technologies becomes more prevalent in sports, questions are raised on the ownership and control of player data. Players should have some level of agency over how their data is collected, stored, and used, with protocols in place to protect their privacy and rights. Furthermore, teams and researchers must uphold ethical standards in data analysis and

interpretation, avoiding biased and discriminatory practices that may perpetuate inequities within the sport.

While advancements in data science offer exciting opportunities for innovation in baseball analytics, it is important to approach these developments with careful consideration of ethical practices. By prioritizing transparency and respecting players' rights, teams can leverage the power of data-driven insights while maintaining the authenticity and originality of the sport.

6 Conclusion

This paper embarks on a comprehensive journey through the integration of advanced statistical methods, machine learning techniques, and sabermetrics into the realm of baseball, particularly focusing on pitch events. This exploration was set against a backdrop of an evolving sports analytics landscape, where the historical progression from basic statistical records to the application of sophisticated data science and artificial intelligence has revolutionized our understanding and strategic approach to baseball. Central to this study was the development and implementation of a feedforward neural network-based simulation environment. This innovative framework was meticulously designed to predict pitch outcomes with remarkable precision, leveraging an extensive dataset encompassing a vast array of pitch characteristics. Through this endeavor, we delved deep into the granularities of pitch behavior and player interactions, uncovering nuanced insights that could potentially alter traditional baseball tactics and strategies. The methodology employed in this research represents a significant leap forward in baseball analytics. This facilitated a more refined analysis of player interactions and pitch outcomes but also sets the stage for the application of reinforcement learning models. These models, through iterative interaction with the simulation environment, promise to refine pitching strategies by learning to maximize favorable outcomes and minimize loss.

Our findings underscore the substantial impact of data-driven decision-making in baseball. The predictive models developed through this study have shown a capacity to replicate the distribution of outcomes observed in real-world datasets, thus validating the effectiveness of our simulation environment. This is a testament to the potential of integrating machine learning and neural network methodologies in sports analytics, offering a new lens through which teams and coaches can evaluate and enhance pitching strategies. Navigating the intricate landscape of baseball analytics, this paper underscores the vital integration of data-driven insights with the essential human elements of intuition and expertise within the sport. As it delves into the ethical realms of data privacy and player consent, the research emphasizes the need for a balanced approach, where advanced statistical models complement rather than replace the nuanced judgment of coaches and

players. The exploration through neural network-based simulations and machine learning techniques highlights the importance of maintaining the integrity of the game while embracing technological advancements. Marking a significant step forward, this study not only enriches the historical narrative of baseball analytics but also paves the way for future explorations in sports analytics. The advent of a new era, characterized by the strategic application of artificial intelligence, promises to revolutionize strategic decision-making in baseball. Standing at the precipice of this evolving frontier, the potential for uncovering deeper insights and transformative strategies seems boundless, heralding a future where data-driven innovations coalesce with traditional baseball wisdom to redefine the strategic contours of the game.

7 References

Baumer, B., & Zimbalist, A. (2014). The sabermetric revolution: Assessing the growth of analytics in baseball. Philadelphia, PA: University of Pennsylvania Press.

Beneventano, J., Berger, L. R., & Weinberg, J. (2012). Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics. International Journal of Business, Humanities and Technology, 2(4).

Cui, A. Y. (2020). Forecasting outcomes of Major League Baseball games using machine learning (EAS 499 Senior Capstone Thesis). University of Pennsylvania. https://fisher.wharton.upenn.edu/wp-content/uploads/2020/09/Thesis_Andrew-Cui.pdf

Elitzur, R. (2020). Data analytics effects in Major League Baseball. Omega, 90, 102001. https://doi.org/10.1016/j.omega.2018.11.010

Hoang, P., Hamilton, M., Murray, J., Stafford, C., & Tran, H. (2015). A dynamic feature selection based LDA approach to baseball pitch prediction. In X. L. Li, T. Cao, E. P. Lim, Z. H. Zhou, T. B. Ho, & D. Cheung (Eds.), Trends and Applications in

Knowledge Discovery and Data Mining (Vol. 9441, pp. 152-164). Springer, Cham. https://doi.org/10.1007/978-3-319-25660-3_11

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

Lee, J. S. (2022). Prediction of pitch type and location in baseball using ensemble model of deep neural networks. Journal of Sports Analytics, 8(2), 115–126.

Mizels, J., Erickson, B., & Chalmers, P. (2022). Current state of data and analytics research in baseball. Current Reviews in Musculoskeletal Medicine, 15, 283–290. https://doi.org/10.1007/s12178-022-09763-6

McShane, B. B., Braunstein, A., Piette, J., & Jensen, S. T. (2011). A hierarchical bayesian variable selection approach to Major League Baseball Hitting Metrics. Journal of Quantitative Analysis in Sports, 7(4). https://doi.org/10.2202/1559-0410.1323

Nakahara, H., Takeda, K., & Fujii, K. (2023). Pitching strategy evaluation via stratified analysis using propensity score. Journal of Quantitative Analysis in Sports, 19(2), 91–102. https://doi.org/10.1515/jqas-2021-0060

Otremba Jr., Stephen Eugene. (2022) "SmartPitch: Applied Machine Learning for Professional Baseball Pitching Strategy." Master's thesis, Massachusetts Institute of Technology

Plunkett, R. (2019). Pitch type prediction in Major League Baseball. Bachelor's thesis, Harvard College. https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364634

Severini, T. A. (2020). Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports (2nd ed.). New York: Chapman and Hall/CRC. https://doi.org/10.1201/9780367252090

Sidhu, G., & Caffo, B. (2014). Moneybarl: Exploiting pitcher decision-making using reinforcement learning. The Annals of Applied Statistics, 8(2). https://doi.org/10.1214/13-aoas712

Sidle, G., & Tran, H. (2018). Using multi-class classification methods to predict baseball pitch types. Journal of Sports Analytics, 4(1), 85-93. https://doi.org/10.3233/JSA-170171

Sun, H. C., Lin, T. Y., & Tsai, Y. L. (2023). Performance prediction in Major League Baseball by long short-term memory networks. International Journal of Data Science and Analysis, 15, 93–104. https://doi.org/10.1007/s41060-022-00313-4

Tango, T. M., Lichtman, M. G., & Dolphin, A. E. (2014). The book: Playing the percentages in baseball. CreateSpace.

Vock, D. M., & Vock, L. F. (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

Watkins, C. (2020). Novel Statistical and Machine Learning Methods for the Forecasting and Analysis of Major League Baseball Player Performance (Order No. 27964705). Available from ProQuest One Academic. (2406520816). http://proxy.libraries.smu.edu/login?url=https://www.proquest.com/dissertations-theses/novel-statistical-machine-learning-methods/docview/2406520816/se-2

Yee, R., & Deshpande, S. K. (2023). Evaluating plate discipline in Major League Baseball with bayesian additive regression trees. Journal of Quantitative Analysis in Sports, 0(0). https://doi.org/10.1515/jqas-2023-0048