New model could have 'Moneyball'-like impact on baseball players' value

April 6 2022, by Jessica Hallman

Credit: CC0 Public Domain

In the movie "Moneyball," a young economics graduate and a cash-strapped Major League Baseball coach introduce a new way to evaluate baseball players' value. Their innovative idea to compute players' statistical data and salaries enabled the Oakland A's to recruit quality talent overlooked by other teams—completely revitalizing the team without exceeding budget.

New research at the Penn State College of Information Sciences and Technology could make a similar impact on the sport. The team has developed a machine learning model that could better measure baseball players' and teams' short- and long-term performance, compared to existing statistical analysis methods for the sport. Drawing on recent advances in natural language processing and computer vision, their approach would completely change, and could enhance, the way the state of a game and a player's impact on the game is measured.

According to Connor Heaton, doctoral candidate in the College of IST, the existing family of methods, known as sabermetrics, rely upon the number of times a player or team achieves a discrete event—such as hitting a double or home run. However, it doesn't consider the surrounding context of each action.

"Think about a scenario in which a player recorded a single in his last plate appearance," said Heaton. "He could have hit a dribbler down the third base line, advancing a runner from first to second and beat the throw to first, or hit a ball to deep left field and reached first base comfortably but didn't have the speed to push for a double. Describing both situations as resulting in 'a single' is accurate but does not tell the whole story."

Heaton's model instead learns the meaning of in-game events based on the impact they have on the game and the context in which they occur, then outputs numerical representations of how players impact the game by viewing the game as a sequence of events.

"We often talk about baseball in terms of 'this player had two singles and a double yesterday," or 'he went one for four," said Heaton. "A lot of the ways in which we talk about the game just summarize the events with one summary statistic. Our work is trying to take a more holistic picture of the game and to get a more nuanced, computational description of how players impact the game."

In Heaton's novel method, he leverages sequential modeling techniques used in natural language processing to help computers learn the role or meaning of different words. He applied that approach to teach his model the role or meaning of different events in a baseball game—for example, when a batter hits a single. Then, he modeled the game as a sequence of events to offer new insight on existing statistics.

"The impact of this work is the framework that is proposed for what I like to call 'interrogating the game,'" said Heaton. "We're viewing it as a sequence in this whole computational scaffolding to model a game."

The model's output can effectively describe a player's influence on the game over the short term, or their form. Displayed as 64-element vectors—obtained by adapting work from computer vision—these form embeddings capture a player's in-game influence and can effectively be used to describe their impact in the short term, such as the span of 15 plate appearances, or averaged together to analyze longer time periods, such as over the course of the player's career. Additionally, when combined with traditional sabermetrics, the form embeddings can predict the winner of a game with over 59% accuracy.

Heaton described how embeddings created by both his method and the traditional sabermetrics method plot the same data. When viewed over time, sabermetric-based representations of player impact can be somewhat sporadic, changing significantly from one game to the next. Heaton's method helps "smooth out" the way players are described over time, while still allowing for fluctuation in player performance.

"Both embeddings can help differentiate good players from bad players," said Heaton. "But ours provides much more nuance into the exact way in which the good players impact the game."

To train their model, the researchers used data previously collected from systems installed at major league stadiums that track detailed information on every pitch thrown, such as player positioning in the field, base occupancy, and pitch velocity and rotation. They focused on two types of data: pitch-by-pitch data, to analyze information such as pitch type and launch angle; and season-by-season data, to investigate position-specific information such as walks and hits per inning pitched for pitchers and on-base-plus-slugging percentage for batters.

Each pitch in the collected dataset has three identifying features: the game in which it took place, the at-bat number within the game and the pitch number within the at-bat. By using these three pieces of information, the researchers were able to completely reconstruct the sequence of events that constitute an MLB game.

The researchers then identified 325 possible game changes that could occur when a pitch is thrown, such as changes in the ball-strike count and base occupancy. They combined this information with existing pitch-by-pitch data that describes the thrown pitch and at-bat action, then input player records from sabermetrics to be able to describe what happened, how it happened, and who was involved with each play.

The work blends Heaton's research focus of natural language processing with his interest in the historical statistical analysis of baseball.

"There's this whole ecosystem built up around modeling language and the sequence of words," said Heaton. "It seems like there was potential for it to be adopted to model sequences of other things; to just generalize it a little bit. I started thinking about sports analytics and it just seemed like there was a lot that could be done to improve both our understanding of the game and how the game is modeled computationally."

The researchers hope that their work will serve as a strong starting point toward a new way of describing how athletes in baseball and other sports impact the course of play.

"This work has the potential to significantly advance the state of the art in sabermetrics," said Prasenjit Mitra, professor of information sciences and technology and co-author on the paper. "To the best of our knowledge, ours is the first to capture and represent a nuanced state of the game and utilize this information as the context to evaluate the individual events that are counted by traditional statistics—for example, by automatically building a model that understands key moments and clutch events."

Heaton and Mitra presented their paper, "Using Machine Learning to Describe How Players Impact the Game in the MLB," was one of seven finalists in the 2022 Research Paper competition at the MIT Sloan Sports Analytics Conference earlier this month.

More information on the competition, as well as links to the paper and its opensource code and data can be found at www.sloansportsconference.com/research-paper-competition.

Provided by Pennsylvania State University

Citation: New model could have 'Moneyball'-like impact on baseball players' value (2022, April 6) retrieved 19 April 2024 from https://phys.org/news/2022-04-moneyball-like-impact-baseball-players.html