Determining Success and It's Value: A Baseball Problem

Shane Sarnac, Zak Collins, Ryan Smith (All CSCI 4502)

Motivation

Analytics in sports is becoming more prevalent than ever. Teams seem to be more and more open to using stats in their day to day decisions. From who they hire to the strategies they employ. And in general the MLB is ahead of the curve; they have more teams classified as being "all in" than any other sport (see ESPN's Great Analytics Ranking). Analytics in baseball is everywhere; from the outfield shifts that are used to when what pitches are thrown. Of the 10 MLB playoff teams in 2015, 9 are at least classified as being "believers" in analytics, 5 of those 9 are classified as being "all in".

Statistical analysis in baseball is an invaluable aspect to the success of teams. Many teams are starting to shift the position of their players based on the statistics of where the ball is usually hit in the current situation. We believe that statistics in baseball can be used for much more than that. Having a method to predict how successful a player will be would give teams an advantage when deciding whether or not to keep a player, or when trading players. Knowing what college and minor league stats are best at predicting a player's success in the major leagues can make teams much more effective at drafting quality players, and being aware of any correlation between a player's height, weight, or handedness, and their ability to become a good pitcher or hitter could be an important factor when deciding which players to draft.

Baseball is expensive. Without any kind of salary cap, the contracts given to the best players in baseball are increasing in value at a very rapid rate, both in yearly value and in the number of years given. While a player may live up to that value in the first years of a contract, it becomes increasingly important to be able to predict how well that player may perform towards the end of the contract. On the other side of that, finding player strengths that are undervalued in the open market are also growing in value. As a result, the ability to accurately determine a player's value directly impacts how much talent a team can acquire given its budget. Teams that accumulate the most talent win the most games, and in baseball that is all that matters.

Previous Work

In the early 2000's, Michael Lewis popularized a data-driven approach to evaluating baseball games through his book *Moneyball*, in which he followed Oakland Athletics' General Manager Billy Beane as he attempted to make a very financially poor team a very successful one by using advanced metrics to find market inefficiencies in the player market. The origin of this data-driven approach, known as Sabermetrics, is generally attributed to Bill James, who annually published a book called *The Bill James Baseball Abstract* starting in 1977, in which he used self-collected data to test baseball myths while also constructing stories about players and teams that may not have otherwise been noticed.

Today, many different organizations attempt to use data to make predictions about the game. Prediction systems like those developed by Fangraphs, Bleacher Report, USA Today, and the Society for American Baseball Research strive to determine how players will perform and how those performances will affect the overall standings at the end of the season. All teams

have developed data analysis divisions in their organizations to some degree in an attempt to improve their edge in the game. This has led to some controversies as data becomes more important in the game, highlighted by an attempted data theft by the St. Louis Cardinals of the Houston Astros' databases in recent years.

Proposed Work

Our main goal is to predict future performance for a given player. For instance for batters we plan to predict relevant MLB stats like BA, OPS, RBI's, or WAR, given a player's stats for previous years. We can boil relevant data down to 3 categories: General, Batting, and Advanced. General describes basic information about the player such as Name, Team, Year, Draft Position, Age, Position, etc. Batting is all basic stats that are relevant to batting such as BA, Hits, RBIs, Hits, BBs, HRs, OPS, SOs, etc. And lastly we will look at Advanced; derived metrics like WAR, WPA, and ISO. While performance prediction is key, we would be remiss to ignore other relevant questions this data can answer. Some other uses include: Player Comparisons: i.e. given a player's stats for previous years, can we map their career trajectory to a player with a similar trajectory? Stat Correlations: i.e. what is the relationship of Age to a specific batting stat like BA, OPS, and/or SLG?

Another aim of the project is to determine the most valuable players in Major League Baseball. Batting, pitching, and other historical metrics will be used to develop a model to predict the production of a given player in upcoming seasons. Using the assumption that better players increase the chances that a team will win more games in a season, we look to determine how much of an impact a given player has on team wins, and by extension, how valuable that player is given current markets.

For our data we will primarily use the Lahman Database, but pull other data we might need from Retrosheet Players Data, and Fangraphs Database. We will use several data mining techniques including regressions, clustering, and any other applicable techniques we learn in class, and compare our results to find the best method.

Evaluation

The player performance model can easily be tested by comparing the model's output for a given set of players over several seasons and comparing them to the actual results of the following seasons. Each milestone will increase the number of statistics looked at as well as decrease the threshold for what constitutes a successful prediction. Hitters and pitchers will be evaluated separately, as the definition of success varies greatly between the two types of players.

While the model develops, players that sustained career-altering injuries or missed significant parts of seasons due to injury will be excluded from the data set. This includes players whom received suspensions or missed significant amounts of time because of activities unrelated to individual performance.

Provided the time, the player performance prediction model will be expanded to attempt to predict the overall performance of a given team provided the player set for a given year. The predictions for the team's performance will be compared to historical data and a success will be defined as meeting a certain threshold of success, which will be raised as milestones are met.

Milestones

Note: Because we are developing different models based on position, the specific statistics used for each milestone will depend on the relationships we find between the variables as the model grows. For now, the following milestones will only list a tentative number of statistics being looked at as well as general accuracy levels. Also note that all milestones are subject to change.

1. 3 basic statistics, using players with at least three years of experience (100 games played per season minimum), with accuracy threshold within 80% of real values for following year.
2. 5 basic statistics, using players with at least three years of experience (100 games played per season minimum), with accuracy threshold within 85% of real values for following year and 80% of real values for the year after.
3. 7 basic or higher level statistics, with accuracy threshold within 90% of real values for first season after selected years, 85% of real values for second season after base set of years, and 80% of real values for third season after base set.
4. 9 or more statistics, with accuracy threshold within 90% of actual statistics for up to 3 years beyond base set of years and within 80% for the first 6 years after base set.
5. Apply model for team success based on the predicted performance of all players on the Opening Day roster and players who played the most innings for the team that year (to compensate for major injury). Win prediction accuracy within 5 games for all teams.

Peer Review Session

We met with Rohit, Kevin, Akash, and Lucas who are doing Predictive Analytics in the NBA. Our discussion with them was very helpful. Previously we were only considering one question to answer with our project, but after talking to them it became clear that we should have at least 3 questions if not more, so that if the question we chose was infeasible for whatever reason, our project could still be successful. This also allows each person in our group to choose a question that they are most interested in. In talking with them, we got the idea to have multiple levels for some of our questions, and that we should start small and gradually make our mining algorithm more intricate.

Since their project is very similar to ours, we were able to refine some of the questions we were interested in, and make sure that it would be possible to come to a conclusion given the data we have access to. How to deal with missing data was another topic of discussion. Both of our teams are looking into predicting a player's stats based off of their previous stats. We hadn't considered the specific method we would use to do this before our discussion with them, but they had several good ideas including clustering similar players together, and using regressions.

Evaluation of our results was another aspect of our project that we hadn't considered thoroughly, but gained some valuable insights from the other team. One of the biggest was what amount of error was acceptable for us to consider our predictions accurate, and we came to the conclusion that it would depend on the prediction, but it should typically be within one standard

deviation and most likely should be decreased as we improve our algorithms and add more factors for our model to consider.

References

- "Great Analytics Ranking." ESPN. ESPN The Magazine, n.d. Web. <http://espn.go.com/espn/feature/story/_/id/12331388/the-great-analytics-rankings>.
- Birnbaum, Phil. "A Guide to Sabermetric Research." Sabr.org. N.p., n.d. Web. <http://sabr.org/sabermetrics>.
- Silver, Nate. "We're Predicting The Career Of Every NBA Player. Here's How." FiveThirtyEight. N.p., 09 Oct. 2015. Web. 25 Feb. 2016. <http://fivethirtyeight.com/features/how-were-predicting-nba-player-career/>.