Baseball Player Performance and Its Factors

Vail Dorchester and Jackson Mediavilla

Project Statement/Motivation

We wish to examine the correlations between numerous factors and player performance in high stake situations. Data mining game events may reveal correlations that are not expected or easily identifiable. We will investigate which factors contribute most to player performance.

We intend to focus on "outside" factors such as weather, stadium attendance, and closeness of the game. We may or may not find interesting correlations between player performance and other factors. Based on the correlation results, we may be able to infer causation on player performance.

The results of this study may reveal interesting information regarding player performance. This information could be used for several reasons including player recruitment, predicting individual player performance given situational information, and predicting the outcome of a game given a current game state.

Literature survey

In his paper, *The Statistical Mirage of Clutch Hitting*, Harold Brooks discusses the statistics behind the phenomenon of clutch hitting, or more colloquially, when a player hits better when under pressure as opposed to normal circumstances. In this paper, Brooks defines being 'under pressure' as being tied or behind by four or fewer runs in the seventh inning, including some more detailed cases of when the bases are loaded. Brooks argues that in order for a model that predicts clutch hitting to be accurate, it must not only show strong correlations between performance and pressure, but it must be consistent throughout multiple seasons as well and not a single-season-anomaly. When following this model, Brooks argues that one can define successful models for clutch hitting in singular seasons, but that these models show little to no correlation with each other across multiple seasons and are therefore inconsistent and inaccurate. He also argues that it is necessary to first define how well the average player does in a clutch situation, and then compare if a certain player or group of players consistently plays above that average across multiple seasons. For this value to be significant, he argues that it must be greater than one standard deviation from the mean in either direction, and that certain players would fall into this statistically significant category season after season, or even that the clutch rating of players would be consistent throughout seasons. He then goes to show that neither of these cases are true; clutch rating tends to vary from season to season, and individual players tend to vary their performance and rarely consistently bat better under pressure. From this, Brooks argues that clutch hitting is irrelevant and "a mirage at best" (Brooks, H, 1988).

In another article, Scott Lindholm discusses the effect of temperature on batting performance. Interestingly, Lindholm notes that there is a strong positive correlation between temperature and batting performance, but he also notes that this could be due to many, many factors, and that temperature may not necessarily be a causal factor in increased batting performance, but merely a related factor. Still, though, this data would prove useful in building models to predict when a player would perform at their best (Lindholm, S, 2014).

Win Expectancy will play a large role in the analysis of player performance. According to the *Win Expectancy* article on the FanGraphs website (https://www.fangraphs.com/library/misc/we/), win expectancy is the percent chance a team will win based on situational factors such as score, inning, runners on, and outs. To calculate win expectancy, on must identify all similar game state situations in the past and calculate the percentage of teams which found themselves in those situations and won. This method of win expectancy calculation gives both teams a 50 percent chance of winning at the start of the game. We will be utilizing previous research and calculations performed by Tangotiger in our analysis of player performance. The figure below shows a small portion of the table used for determining win expectancy based on a game state. The entire table can be found at http://www.tangotiger.net/welist.html

Inning	Top/Bottom	Score	Outs	1B	2B	3B	WE
7	Bottom	-1	0				0.353
7	Bottom	-1	0	1st			0.431
7	Bottom	-1	0	1st	2nd		0.545
7	Bottom	-1	0	1st	2nd	3rd	0.687
7	Bottom	-1	0	1st		3rd	0.612
7	Bottom	-1	0		2nd		0.487
7	Bottom	-1	0		2nd	3rd	0.656
7	Bottom	-1	0			3rd	0.545

Proposed Work

In our literature survey, we found that there has been a rather extensive amount of research over evaluating individual player, and this makes sense. Producing that data and selling it to recruiters or scouts proves to be quite valuable, and in fact, most professional teams employ a professional statistician to do reports and research in this exact field. However, less research has been conducted regarding general player performance (performance of all players, not just one), and the factors that go into that. We found an article on the topic of clutch hitting, and they discussed that a trend of correlations should hold across multiple seasons and players, and this is an insight we will most certainly make use of.

With all of this in mind, a plan of our proposed work begins to take shape. First, it will be necessary to download the data itself. Most of the data is split into categories such as All-Star game score, postseason games, and regular season events (meaning every single event in every single game of the regular season), which is one of the sets we are interested in. This data is then further divided into seasons or collections of seasons, so it will be necessary to join the datasets into one large set, and then to clean and normalize the data. This will likely include writing scripts to scrape file systems for data given that each decade contains upwards of 400 files. The cleaning will involve making sure the format is uniform across these files, and accounting for missing values or choosing which values to ignore. For example, one of the factors we are interested in is weather, yet many of the files list wind speed, precipitation, and sky conditions as unknown. Because of this, it will be necessary to decide if there exist enough points to make use of those attributes or if they should be discarded. There is also a rather significant amount of superfluous data that will be culled down such as who the umpires were, who reported the game, who kept score, who translated the game, and so on. Again, this will likely involve writing scripts given the massive amounts of data that would be required to sift through.

Once this has been done, we will still need to roll the data into more concise cubes of information; these cubes will likely vary but could contain dimensions such as season, team, and batting average for example. After the data has been properly imported, cleaned, and sorted, we will likely begin by calculating and examining summary statistics for the attributes that we are interested in — sky conditions, crowd turnout, temperature, home team vs away team, etc. Once some of the interesting attributes have been identified, will create or reuse some popular models for player performance including batting average, isolated power, slugging power, or other models of our choice. With these models, we will need to again calculate some summary statistics including the standard deviation, the mean, and the quartile ranges. With this, it will be possible to identify whether a player's performance is comparatively high enough to be statistically significant, i.e., one standard deviation above the mean. Once this has been done, we can calculate correlation coefficients for these performance models and some of the outside factors or combinations of factors. From this, we will gain some insight and knowledge into what factors tend to correlate to higher performance, and we can begin to group these factors and see how a combination of factors plays in to performance. From this, we hope to create a reasonably accurate model that can accurately predict under which conditions baseball players tend to perform better.

Data Set

We are using the Retrosheet event dataset, which can be downloaded from retrosheet.org/game.htm. The data is currently downloaded on both Jackson and Vail's computers.

The dataset contains every play in the included games. Each entry in the dataset reports both the event and the game state after the play. There are more than 95 different fields which may apply to a single event. The fields provide situational information such as score, identity of batter, number of outs, handedness of batter and pitcher, etc.

Evaluation Methods

We can perform basic statistical analysis to discover relationships between a range of factors and player performance. This will require defining which situations qualify as "high stake" and a method to evaluate player performance. We will use Brooks' definition of "under pressure" to define high stake situations in games. These situations occur when a team is tied or behind by four or less runs during innings seven or greater. We are currently developing a method for evaluating player performance that assigns each team a probability of winning given the current game state. Once finalized, we will use this function to analyze how specific game events change the team's probability to win.

Tools

- Python
- Numpy
- Pandas
- Matplotlib
- Sklearn
- Scipy
- Jupyter Notebook
- Excel
- Git, GitHub
- Bash, sed, Regex

Milestones

Milestones Completed

This report constitutes Part 3 - Project Progress Report. This file will have been uploaded to the GitHub repository by April 10th at 11:00 a.m. This submission is an updated, extended version of the initial report submitted on March 6th. We have

reviewed the sections *Motivation, Proposed Work, Tools, Evaluation,* and *Milestones*. We have removed the section *Summary of peer review session.* Two subsections have been added to *Milestones - Milestones Completed and Milestones Todo*. And finally, a new section, *Results so far*, has been added at the end.

Milestones To-do

We will have completed the remaining parts of the project by 11:00 a.m. on Tuesday May 1. Part 4 is the Project Final Report. The submission for this milestone will be an updated, extended version of this Part 3 Report. The sections to be included are *Abstract, Introduction, Related Work, Data Set, Main Techniques Applied, Key Results, Applications,* and possibly *Visualization*. Part 5 is the Project Code and Descriptions. The submissions for this milestone include all the source code on GitHub, a five-minute video discussing the project and the results, and a README to display on the GitHub main page. Part 6 is the Project Presentation. On one of the last three days of classes, we will present our project to the class. We will also submit our slide deck to our GitHub. Part 7 is Peer Evaluation and Interview Questions. The submission for this part of the project will be the peer evaluation form submitted individually to Moodle.

Results so far

Our work so far has been mainly focused on data aggregation, cleaning, and preprocessing. The data we are using includes information from games starting in 1950 and ending in 2017. The data was downloaded as zip files from the Retrosheet website. Each zip file contained the data for one decade and was separated between roughly 350 files per decade. These files ranged from team rosters to league rosters and individual event files. The data we are interested in exists in the event files, which are marked by a .EVX file extension. However, the decade-folders each contained many other extraneous files. To solve this, we wrote a bash script that utilized Regex and *sed* commands to create a file containing the paths of every event file across all decades from 1950-2017. In conjunction with this path file, we used the executable file BEVENT.EXE (also downloaded from the Retrosheet website) to convert the files to comma separated (CSV) files for easy importation to dataframes. Because the command can only handle one file at a time and there were hundreds of files to process, each located in one of ten separate directories, we created a Jupyter Notebook to iterate over the path file, which would then pass each path to the BEVENT program and generate a .csv for each event file. The result of executing the notebook was hundreds of CSVs, each containing one team's data from one year. These CSV files were then moved into a separate directory, and we created additional python scripts to import all of this data into hundreds of Pandas dataframes — one for each team and year — as well as one aggregate dataframe. Because the CSV files did not contain column information, the next step in preprocessing the data was to add field names for the 97 possible columns. This was done by reading the documentation of the BEVENT.EXE program and copying the attribute descriptions into a text file, which was then used in python to add the column descriptions to the dataframes. At this point, the data is nearly ready for statistical analysis.

With about 8.7 million rows of data, complete iteration takes a significant amount of time. Thus, we have created some Python functions that take as an argument the number of smaller CSV files to include and return a smaller concatenated dataframe. This allows us to work with smaller datasets at a time to develop our analytics code. Once we are satisfied with the results using a small dataset, we will run the entire dataset through our program.

In baseball, clutch measures how well a player performs in high stakes situations. Given this definition, one of our first tasks was to define what exactly a constituted a high stakes situation. One of the definitions we came up with was a play that took place in the 9th inning when there were already two outs. In this situation, there is a significant amount of pressure on the batter to perform well, because it is often their teams last at-bat. When looking at this situation, we chose to analyze the relationship between the number of outs before the play occurred and the event value. Here, the event value is simply a number indicating what happened that play with 0 being no event, 1 being the batter was called out, 2 being a single, 3 a double, 4 a triple, and 5 a home run. The data is plotted below in Figure 1. When looking at this data, we see that the vast majority of times in this situation, the player is marked out and the play ends. Furthermore, singles are the most common hit value, followed by doubles, home runs, and finally triples. This slight increase in home run frequency may indicate an underlying clutch factor, but more analysis is needed to say whether this is the case. In Figure 2, we see the same data plotted for any inning of the game as opposed to the ninth inning only. We see here that in any time of the game, when there are two outs, the distribution of event values is extremely similar. This leads us to believe that the inning has no effect on the performance of a player at bat with two outs.

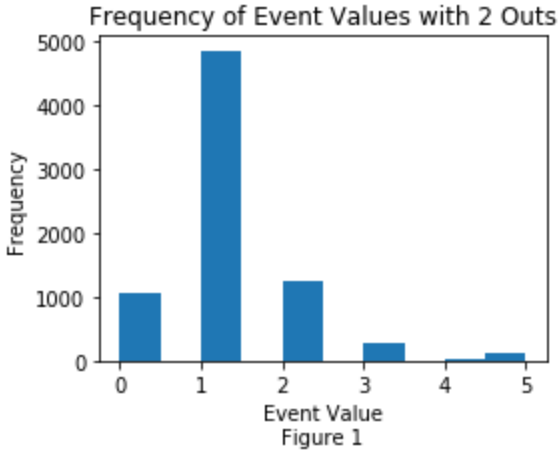

Figure 1

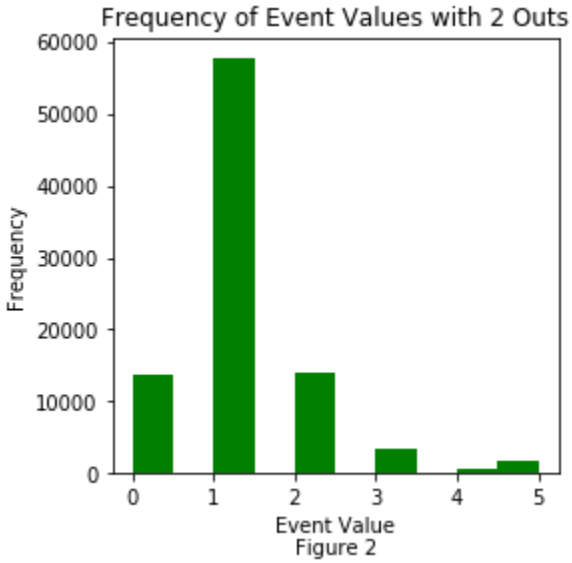

Figure 2

Citations

Brooks, H. (1988). The Statistical Mirage of Clutch Hitting. *Baseball Research Journal*. Retrieved from http://research.sabr.org/journals/the-statistical-mirage-of-clutch-hitting

Lindholm, S. (2014, March 27). Hitting and temperature. Retrieved from https://www.beyondtheboxscore.com/2014/3/27/5546952/hitting-temperature-baseball