Final Project

DAT-520

November 5, 2014

**ABSTRACT**

Major league baseball ownership and management teams agonize over whether hitting or pitching will carry their team to success, as understanding the keys to victory will outline the clearest path to developing a winning team. In an attempt to determine whether hitting or pitching performance provides greater impact to overall team winning percentage, more than half a century’s worth of team statistics for hitting (batting average, slugging percentage) and pitching (walks plus hits per inning, earned run average) were examined and incorporated into a bottom-up CART decision tree. Walks plus hits per inning was the most influential statistic of the four, as the model determined that there was a 74% probability of a losing season if a team produced a higher than average WHIP for a given year. It appeared that the pitching statistics chosen were significantly more influential in predicting a winning or losing team, and a management/ownership team armed with this decision support model and results should be weighing pitching statistics more heavily in individual scouting, recruiting, and signing efforts when building a team.

**TABLE OF CONTENTS**

Introduction…………………………………………………………….…………….……………4

Research Question..…………………………………………………….…………………………4

Data Appraisal….…………………………………………………………………………………5

Measurable Utilities……………………………………………………………………………… 5

Methods……………………………………………………………………………………………6

Analytical Approach………………………………………………………………………………7

Bias and Confounding……………………………………………………………………………..8

Flexibility……….…………………………………………………………………………………9

Ethical Considerations……………………………………………………….…………………..10

Data Model………………………………………………………………………………………10

Conclusion..…………………………………………………………………...…………………14

References.………………………………………………………………………………………16

**INTRODUCTION**

Baseball is generally thought of as the most statistically developed sport, but with many schools of thought on what produces winning teams. While use of advanced analytics has been become more prevalent in identifying undervalued talent by way of less popular individualized baseball metrics, evaluating the relative worth of players for signing should also be based on the importance of pitching and hitting success in reference to the team’s success. In 1978, SABR (Society for American Baseball Research) conducted a study that polled 50 baseball experts among 26 major league teams, and 44 of these 50 experts responded that pitching was the single most important factor in winning (Skipper, n.d.).

A CART bottom-up model using both pitching and hitting input variables demonstrates that predictor pitching variables play a greater role than hitting variables in predicting team winning percentage.

**RESEARCH QUESTION**

Are winning teams built based on pitching or hitting? A winning team is a team that wins more than half of its games in a given season (winning percentage > .500), and the objective of the analysis is to use relevant performance measures of pitching and hitting performance as predictor variables to try to determine whether a winning record is more dependent on pitching or hitting performance.

**DATA APPRAISAL**

The dataset used for this analysis was one table from a relational database of historical baseball data. While there were a vast number of tables available in the Lahman database, the dataset chosen was the “Teams.csv” file, which contained a single record for every major league baseball team and its year-end statistics for every year from 1871 through 2011 (Lahman, n.d.). Some of the statistics used for the model were manually derived, but all counts used for calculations were taken from this single data source.

**MEASURABLE UTILITIES**

**Input Predictor Variables**

* **BA (Batting Average), categorical:** records classified as either above or equal to/less than historical average of .259
* **SLG (Slugging Percentage), categorical:** records classified as either above or equal to/less than historical average of .396
* **ERA (Earned Run Average), categorical:** records classified as either above or equal to/less than historical average of 4.00
* **WHIP (Walks Hits per Innings Pitched), categorical:** records classified as either above or equal to/less than historical average of 1.36

**Target Variable**

* **Winning, categorical:** Success “Winning” is defined as a team with a winning percentage above .500, or winning more than half of its games through a full season.

**METHODS**

While there were no missing values in this dataset, the entire dataset was not used, as it was assumed that very old statistics (pre-1950) were not as statistically relevant in predicting a team’s success in the modern era, due to rule changes and other factors; therefore, statistics for the years of 1871 through 1949 were omitted entirely.

Additionally, several calculations had to be performed on the dataset in order to derive some of the necessary metrics for the study, including the target variable winning percentage “Winning”, WHIP, and SLG. A more detailed legend of how these variables were derived from this dataset is listed below:

* Winning = W / (W + L), or Wins / (Wins + Losses) *[derived manually]*
* BA = Hits / At Bats
* SLG = TB / AB, or Total Bases / At Bats *[derived manually]*
  + Where Total Bases = 1\*(1B) + 2\*(2B) + 3\*(3B), and 4\*(HR),
    - Singles (1B) = H – (2B + 3B + HR)
* WHIP = (HA + BBA) / IP, or (Hits Allowed + Bases on Ball Allowed) / Innings Pitched
  + Where IP = IPouts / 3 *[derived manually]*
* ERA = Earned Run Average per 9 innings, or (Earned Runs Allowed / Total Innings Pitched) \* 9

Random samples of the manually derived calculated values within the dataset were verified against a secondary data source (http://www.baseball-reference.com) to ensure accuracy.

The input variables were chosen specifically with recursive partitioning in mind, and therefore there were exactly two pitching and two hitting statistics chosen as inputs to the team success model. If three variables were chosen and two were hitting statistics, then this would naturally skew the model to hint that hitting statistics play a more significant role in predicting team success. Each of the four metrics in the dataset was a ratio variable, but these were converted into categorical variables, with a value describing each team’s placement above or below the historical average for that metric over the 61-year span in the dataset. Historical averages for each of the four metrics were derived directly from averaging each field within the dataset. It was assumed that expressing each team’s performance as simply “above or below” the statistic’s historical average would be an appropriate way of normalizing this dataset for the context of this decision support model, since the measure of success itself is a classification above the average (above a winning percentage of .500).

**ANALYTIC APPROACH**

In predicting winning percentage or a similar team outcome (will this team make the postseason, will this team win the division/World Series, etc.), many studies confined research to utilize run scoring (hitting) and run prevention (pitching), since producing more runs than an opponent in a given game actually determines which team wins a given game. The research of Bill James, whose work included trying to predict baseball team winning percentage, assumed that offense and defense are independent of one another, which is important when applying variables to the model and determining which will contribute more to a winning percentage (Miller, 2006). The same assumption is made for this model, since a given team and player is either hitting or pitching during a given inning.

However, runs scored and runs allowed are team metrics for the most part (runs scored can be an individual statistic but is more of a function of a player’s on base percentage, placement in the order, and maybe base stealing ability), and are typically not used to determine player’s values directly. Due to this, other individual statistics must be used, which is why the four individual performance statistics (two for hitting, two for pitching) were used in the model. For hitting, including both batting average and slugging will not only indicate the frequency of hits, but also how valuable each hit is (single versus home run). For pitching, WHIP will indicate how frequently a pitcher will put hitters on base (they cannot score unless first reach base), and ERA will describe how frequently earned runs actually score off the pitcher. While it is clear that there is no widely accepted set of statistics that purely translates to hitting or pitching superiority, incorporating more than one statistical hitting and pitching measure each increases the chances that offensive and pitching performance will be properly represented with this model and exercise.

**BIAS AND CONFOUNDING**

The greatest potential for bias in the model setup lies with my own choice of the four input variables used to predict the target variable. Are these the “right” variables to use, or is this model misrepresenting how a team should build its team and strive for a winning record? While some research would agree these metrics are important performance indicators (and good team performance predictors), many other experts might disagree on the choice to use these four specific statistics. From a statistical standpoint, ERA might be the statistic that most closely ties run scoring or run prevention at an individual level (run prevention, of course), and there may not be any hitting statistic that can match this kind of relevance. As the model was developed, ERA turned out not to be the most influential variable, but there could be a number of reasonable arguments against using the BA, SLG, ERA, and WHIP. Weaknesses in the model include the omission of other potentially relevant categorical variables, such as whether or not each team in the dataset was in the American League or the National League. The rule difference between leagues, where the designated hitter is used in the AL, generally results in higher overall hitting performance and degraded pitching measures, and this could be perceived as a major factor that could influence how management might build their team. Adding or replacing some of the input variables could potentially improve the model, specifically by minimizing its error rate in predicting winning percentages.

**FLEXIBILITY**

While the potential limitations and biases have been discussed, there is still flexibility in this model in that it could be applied to any team, and it could fit to other datasets, assuming complete team level statistics are used. However, the model is inflexible in that it splits based on placement above or below averages for each metric, but there is no consideration to *how much* above or below the averages a given team’s statistics fall.

A 2011 study of MLB teams over the decade prior concluded that in the context of successful teams, the teams with relative elite pitching (runs allowed) had a higher winning percentage than those successful teams with relatively high hitting (runs scored) performance (Petti, 2011). What this means is that managers must be cognizant of the deltas between the their own team’s statistics and the historical averages, and not to chase percentage points thinking it will win more games for the team simply it crosses a threshold within the model to classify their team as “above average”.

**ETHICAL CONSIDERATIONS**

As with any study or analysis, there are legal and ethical considerations that must be made with regard to data management. Fortunately, in a decision support exercise utilizing Major League Baseball statistics, there are not many concerns from this perspective, as all data is free and publicly available. If anything, utilizing more emphasis of statistical performance to make personnel decisions should increase the fairness of these business transactions. Even if the model performs poorly, the intent is to base player merit on relevant performance measures that will help the team win the most games possible, as opposed to other motivating factors, such as signing popular players who may increase ticket and merchandise sales.

**DATA MODEL**

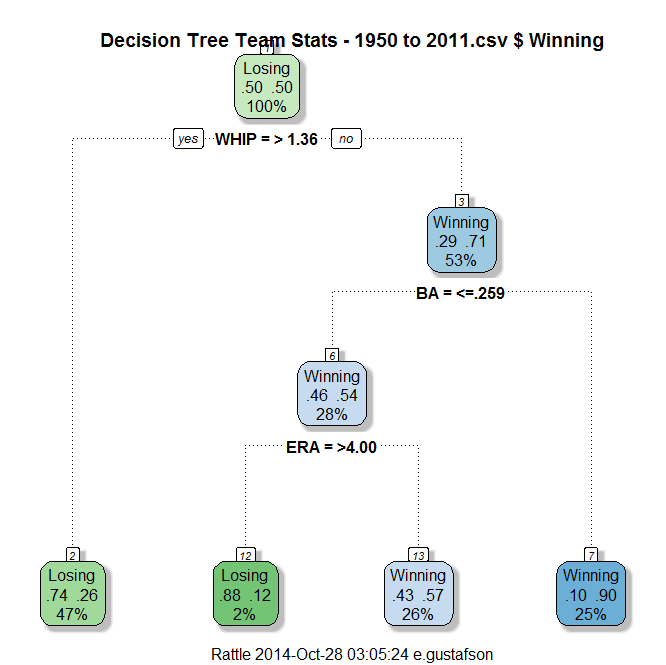
Below is the CART model developed for the decision support task of predicting a team’s success over a given season. The model was developed using the Rattle GUI within R using the following code:

library(“rattle”)

rattle()

[.csv file was loaded into R from the GUI once initialized]

**FIGURE 1 (next page).** Bottom-up decision tree using Team Stats dataset using Rattle. The top label at the top of each box lists the classification at each node (“Winning” or “Losing”), the decimals in the middle indicate the probability of a losing team (left decimal) or winning team (right decimal) at each node, and the bottom percentages indicate the percentage of the prior node that sample represents.



**TABLE 1:** Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 1051

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 1051 523 Losing (0.50237869 0.49762131)

2) WHIP=> 1.36 495 130 Losing (0.73737374 0.26262626) \*

3) WHIP=<= 1.36 556 163 Winning (0.29316547 0.70683453)

6) BA=<=.259 298 138 Winning (0.46308725 0.53691275)

12) ERA=>4.00 24 3 Losing (0.87500000 0.12500000) \*

13) ERA=<=4.00 274 117 Winning (0.42700730 0.57299270) \*

7) BA=>.259 258 25 Winning (0.09689922 0.90310078) \*

Classification tree:

rpart(formula = Winning ~ ., data = crs$dataset[crs$train, c(crs$input,

crs$target)], method = "class", parms = list(split = "information"),

control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:

[1] BA ERA WHIP

Root node error: 523/1051 = 0.49762

n= 1051

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CP | nsplit | rel | error | xerror | xstd |
| 1 | 0.439771 | 0 | 1.00000 | 1.06692 | 0.030934 |
| 2 | 0.017208 | 1 | 0.56023 | 0.56023 | 0.027795 |
| 3 | 0.010000 | 3 | 0.52581 | 0.54302 | 0.027527 |

Time taken: 0.02 secs

**TABLE 2.** Error matrix for the Decision Tree model on Team Stats - 1950 to 2011.csv [validate] (counts):

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Actual | Losing | Winning |
| Losing | 85 | 33 |
| Winning | 25 | 82 |

**TABLE 3.** Error matrix for the Decision Tree model on Team Stats - 1950 to 2011.csv [validate] (proportions):

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Actual | Losing | Winning |
| Losing | 0.38 | 0.15 |
| Winning | 0.11 | 0.36 |

Overall error: 0.2577778, Averaged class error: 0.2571146

Starting at the root node, there is virtually a perfect split of team records, with a 50% probability of having a winning or losing record any given year within the dataset. This is expected, since historically over time, each game must be won by one team and lost by another, so this probability should be 50% if all teams are included in the dataset, which is the case. At the root node, the first classification split occurs with the WHIP pitching metric: 47% of all teams had a WHIP above the historical average of 1.36 (the lower the WHIP, the better the performance), and 53% featured a WHIP above 1.36 (worse performance). For the teams with a WHIP above 1.36, this branch ends in a leaf node, where there is a 74% probability of a losing season based on WHIP alone.

The second node level splits team outcomes based on the batting average metric (the higher the better): if the WHIP is less than or equal to 1.36 and the same team features a batting average above .259 (25% of all teams), then these teams are predicted to have a 90% probability of a winning record. It should not be surprising that balanced teams who can pitch and hit will achieve winning records most frequently. For the teams that have a WHIP below 1.36 but also a batting average below .259, the next classification split occurs with ERA, another pitching statistic (the lower the better). For this third level node, if the team ERA is less than or equal to 4.00 (26% of all teams), then there will be a 57% probability of a winning season. Conversely, if the team ERA is worse than 4.00, then the model predicts only a 12% probability of a winning season. However, this last pool contains only 2% of teams in the dataset. It should be noted that SLG (slugging percentage) was not utilized in the model output for any classification split.

The error matrix included with the tree output gives an indication of how well the model performed over the dataset. Overall, the error rate was approximately 26%, 11% of which was teams that the model predicted to be losers but finished winners, and the other 15% comprised of teams projected to be winners based on the input variables but finished below .500, as losing teams. The false negative outcome from this model, where the team is predicted to be a winner but finishes a loser, would be the most concerning to a manager or team relying on this model as a guide for choosing players, since the team would have underperformed against preseason expectations.

**CONCLUSION**

The structure of the tree indicates that the pitching statistic WHIP is the most influential variable, and aside from a second level classification split on BA, the pitching statistics drive the entire model, since SLG was completely omitted in the tree. Despite the high error rate and high chance for misclassification at the root node for WHIP (root node error = .49762), it still appears that team pitching statistics will play more of a role in determining a team’s ability to win games over the course of a season. Therefore, a user could use this decision support model to better evaluate individual players and how those contributions to team statistics may or may not improve the team’s overall chances of a winning season, and if all else is equal, opt for elite pitching talent. Aside from the model data, it might intuitively make sense that a better than average pitcher will impact winning percentage over the course of a season, since a pitcher, especially a starting pitcher, is influencing the majority of outs made against the opposition, whereas a hitter can only impact offensive outcomes one-ninth of the time. This is different from offensive and defensive performance for other sports, such as basketball or football, where the best offensive players for a team can be involved in every offensive play. Given the categorical nature of the input variables, a given player’s statistics “above” or “below” should not be considered alone in determining potential contributions to a future team’s success, since considering how much above or below the historical average a given player’s statistics are (a hitter whose batting average is 35 percentage points above the average is likely going to contribute more than a pitcher whose WHIP is .05 lower than the historical average). Additional courses of action recommended might include altering the timeframe of the data and rerunning the model to incorporate different periods, such as more recent years only, or by including new input variables such as AL/NL.

References

Lahman, Sean. (n.d.). SeanLahman.com. Retrieved October 16, 2014 from http://www.seanlahman.com/baseball-archive/statistics/

Skipper, Jr., J. (n.d.). Is Pitching 75% of Baseball? Expert Opinions. Retrieved October 27, 2014, from http://research.sabr.org/journals/is-pitching-75-of-baseball

Petti, B. (2011). Is it Better to be an Elite Run Producing or Run Preventing Team? Retrieved November 2, 2014, from http://www.beyondtheboxscore.com/2011/2/22/1994723/is-it-better-to-be-an-elite-run-producing-or-run-preventing-team

Miller, S. (2006). A Derivation of the Pythagorean Won-Loss Formula in Baseball. Retrieved November 2, 2014, from http://web.williams.edu/Mathematics/sjmiller/public\_html/399/handouts/PythagWonLoss\_Paper.pdf