Colorado State University
University Honors Program

A Whole New Ballgame? How the Best MLB Players are Chosen In the Age of Sabermetrics

A Statistical Investigation into the MLB Network Top 100 Players List

Jonathan Olds

Honors Thesis
Fall 2020

Contents

Abstract

In each spring since 2010, baseball analysts affiliated with the cable television station MLB Network have released a list of the 100 best players heading into the upcoming season. The first ten iterations of this list are situated in a unique period of baseball history, in which analysts have a collection of data and player evaluation measures available to them that is larger than at any other time in the sport. These measures fall into two broad categories - (1) conventional statistics, which are usually defined as 'counting statistics' and have been a part of baseball from the sport's inception in the mid 1800's, and (2) sabermetrics, whose metrics are designed to account for context to provide a more objective examination of the sport and its' players. This paper serves as an exploration into the past decade of baseball analytics and aims to determine whether analysts have chosen to utilize conventional statistics or sabermetrics, or a combination of the two, in their evaluation of players. We fit logistic regression models while using principal component analysis and lasso penalization as two variable selection methods with the goal of accurately predicting the players who would make the Top 100 lists. Models built with sabermetric statistics produced an average of 13.8 percentage points difference in sensitivity when predicting pitchers than models with conventional statistics, indicating that analysts likely weight sabermetrics more heavily in their evaluation of pitchers. Models built with sabermetric statistics produced an 8.5 percentage points difference on average in sensitivity when predicting batters than models with conventional statistics, but models built with both statistics added an additional 1.0 percentage point on average in sensitivity. These results indicate that analysts use a more balanced combination of conventional statistics and sabermetrics in their evaluation of batters.

1 Introduction

Among the major American professional sports leagues, Major League Baseball (MLB) teams trail only National Football League (NFL) teams in roster size. Each MLB team may have up to 40 players on their active roster; this roster is described in the league operating book as being a list of all players that the club "anticipates participating" during the season [1]. Even while not accounting for the potential movement of different players on and off the roster throughout a 162 game schedule, this equates to roughly 1200 athletes taking the field for a major league team in any given season.

In order to accurately assess the skill of 1200 players, analysts must have access to the best measures of performance available. However, 'best' is naturally a subjective term, and in this case is even more so, because the history of baseball statistics has created two distinct groups of metrics - conventional statistics and sabermetrics. Though these two groups were introduced over different time periods, they have been forced to coexist in the twentieth century. A brief summary of that process is described in the following paragraphs.

1.1 History of Baseball Statistics

The rules of baseball were formalized in the 1830's, just as various statistical societies were being formed, such as the American Statistical Association in 1839. Although numerical data had been gathered before in other sports such as cricket, baseball record-keepers quickly maintained the most extensive records of any sport. Even before professional leagues existed, teams kept records of statistics, believing that they held the key to identifying the best players and strategies [2].

Some of the earliest conventional statistics came from writer and baseball propagandist Henry Chadwick. Chadwick began awarding batters who reached base but didn't score a 'base hit' in 1867, and then divided hits by at-bats to create a new statistic a few years later, which quickly became known as batting average. [3].

The first half of the 1900's saw the invention of more conventional statistics that quickly became standard. For example, earned run average (ERA) was invented in 1912, and the run batted in (RBI) was made official in 1920. Meanwhile, the unprecedented success of players like Babe Ruth and Ted Williams sparked an even greater interest in individual statistics, rather than the scores of the games themselves [3].

In 1947, the Brooklyn Dodgers hired Allen Roth, making him the first statistician employed full-time by a major league baseball club. Roth collected data on new statistics, including an early form of on-base percentage, batting average with runners in scoring position, and performance in different ball/strike counts. Throughout the second half of the twentieth century, team and player statistics became more accessible to teams and the public, first through encyclopedias and baseball cards and then by way of computers.[3].

By this point in history, most conventional baseball statistics used in player evaluation had been invented and were utilized regularly. For batters, the most popular of these were batting average, home runs and runs batted in, while wins, losses and earned run average were most frequently utilized when discussing pitchers. Fielding errors were one of the only measures used to compare defensive ability [4]. These statistics were embedded into the minds of most players, coaches, analysts, and fans as the best standard of evaluation, primarily because of their longevity and accessibility.

One of the first minds to think differently was a statistician named Bill James. He was the first to coin the term 'sabermetrics' in 1980 as a combination of the first letters of the Society of American Baseball Research, to which he belonged, and 'metrics' [4]. He defined the term as "the search for objective knowledge about baseball" and was intent on finding and making accessible better methods of player evaluation. [5].

James published the first edition of his annual Baseball Abstract book in 1977; these books included James' statistical insights that decried many of the conventional statistics and methods utilized in teams' decision-making process and suggested new alternatives [5, 6]. For example, James created a metric called runs created (RC) in the late 1970's to measure a batter's ability to help his teams score runs, based on the perception that runs were more valuable to a team offensively than hits [7]. While the books received some criticism by the baseball community, they reached mainstream popularity by the mid-1980's [5].

Inspired by James, new statistics such as on-base percentage + slugging percentage (OPS) were utilized in box scores. Also inspired by James, new organizations and sabermetricians released research devoted to more new evaluation measures, primarily concerned with assessing a player's overall value. The leading metrics in this area were Value Over Replacement Player, first published in 1996 by Baseball Prospectus and Defense Independent Pitching Stats, created by Voros McCracken in 2001 [3].

Yet, while sabermetrics had made it to the fringes of baseball relevance by the turn of the millennium, it was Billy Beane's decision-making process, inspired by Bill James and rooted in analytics, that showed the world all at once that sabermetrics could work at the macro level. Beane's approach as general manager of the Oakland Athletics, a small-market team with an equally small payroll, sparked a run of unexpected success, including an American League West division title in 2002. [8]. This success came because Beane recognized that analytics could be used to identify statistics that were undervalued, such as on-base percentage. The Athletics then were able to acquire players skilled in those areas while remaining financially efficient [9]. Oakland's methods and subsequent success were described in detail in Michael Lewis' book Moneyball in 2003, which spread analytical insights rapidly throughout the league and into dozens of other industries over the next decade.

At the same time, baseball database websites such as Baseball Reference and FanGraphs began to publish their own sabermetric measures of skill and value, such as Wins Above Replacement (WAR) which was created to estimate how many wins a player is worth above a replacement level player, or Skill Interactive Earned Run Average, which was built as an ERA estimate for pitchers that attempts to more accurately isolate a pitcher's performance from outside influences. [10, 11]

By the 2010's, sabermetrics could no longer be ignored - they had proven themselves to be useful and accessible. As a result, over the past decade, baseball teams began to commit to the use of sabermetrics. By 2013, 23 of 30 teams were using techniques inspired by the Oakland Athletics and by 2019, every team had created their own analytics departments, ranging in size from four to twenty employees. [12]

While baseball teams were showing their enthusiasm in embracing sabermetrics, baseball media showed less ability to adapt to the change. A study of written media conducted in 2017 found that, while sabermetrics have been mentioned more frequently since the analytical success of

Oakland in 2002, conventional statistics still outnumber sabermetrics in mentions by a value of 25:1. Theories on this trend include that casual fans are resistant to sabermetrics, or unable to understand them, so media has catered to them with stories backed by conventional statistics [4].

Facts and research show that major league clubs have come to fully embrace sabermetrics within the last decade, and that the baseball media has not shown the same ability or desire. However, it is not clear where baseball analysts, that is, those not employed by major league teams, stand on this issue. Thus, it is the goal of this paper to place this subsection of the greater baseball community on the analytics spectrum. In other words, have non-team-affiliated baseball analysts bought into the value of sabermetrics, or do they still rely primarily on conventional statistics?

2 Data

We chose the Top 100 players list released by MLB Network as our tool for investigating this problem for a couple of reasons. First, the analysts who create these lists are not affiliated with or employed by any specific team, which should significantly reduce any team or market bias, and also means that they are able to make analytical decisions not constrained by a club's goals or resources. While these analysts are employed by a member of the media, they are not reporters, and thus are not restrained by the needs of fans in their decisions either. There may be writers with the network who communicate the decisions of these analysts in ways that cater to fans, but it is likely that the analysts themselves are free to use a full spectrum of analytical methods in their selections.

Secondly, in creating these lists, the analysts in question are narrowing down roughly the best 8% (100/1200) of players in the league for any given season. To separate the best 8% from the other 92%, there must be a great deal of analytical effort and a broad range of factors considered. Thus, it is fair to say that the analytical scope required to build the Top 100 Players list is significant enough to draw relevant results about the tendencies of baseball analysts as a group.

2.1 The Top 100 Players Lists

The first edition of the Top 100 Players List was released by MLB Network in spring of 2011. It was designed to rank, from 1 (best) - 100 (worst), the best players in the game, based off of players' performance prior to 2011, with emphasis on the past three seasons, and projected performance in the upcoming 2011 season. This list was repeated the following spring in 2012 with the same design, and each season afterwards through 2020.

The iterations of the list after 2017 were accessible through the MLB.com website, while lists from the older seasons were available on other various baseball websites.

These lists, however, provided only the name, position and rank of the players on the lists, and did not include any information on other players that did not make the list. As such, data were gathered on all players from each season from 2008 to 2020 on numerous variables of interest.

However, it was decided that it is unlikely that, in the making of a Top 100 Players list, analysts were examining the statistics of each player from that season, because a large percentage of those players did not contribute enough to their team to warrant consideration. For example, in 2019, 52.1% of the eligible batters who participated in a major league game that season played less than 30 games, and only 25% of the players played in half of their team's games (81) [10]

Based on this principle, the data were narrowed down to only include players who made the list, and then all batters who averaged more than 251 at-bats per season over the past three years and pitchers who averaged more than 41 innings pitched in the same time span. Rookies or other players who had yet to reach three seasons of playing time were included if their at-bats/innings pitched values met the necessary requirements in their only season or averaged together to meet those requirements in their only two seasons. These values equate to roughly half of the plate-appearances required for a batter to qualify for a batting title, and a quarter of the innings necessary for a pitcher to qualify for an ERA title. Thus, it was determined that any player who averaged that mark over the previous three seasons would have contributed significantly to their team and would warrant consideration for making a Top 100 list. This left sample sizes of 3,537 batters and 4,774 pitchers deemed eligible for the Top 100 List across the 10 seasons.

It is important to note that the data were gathered for each individual player separated by season, meaning that in many instances, one batter had multiple seasons of data in the sample. For example, a batter that played from 2008-2016 would have nine seasons worth of data initially, but if the batter only averaged more than 251 plate appearances in the past three years in four of those seasons, then only those four seasons would be included in the 3,537 number.

2.2 Player Statistics

Baseball Reference and Fangraphs are among the leading sources of statistical baseball information, and served as the primary sources of data. The cumulative effort of these two databases provided dozens of statistics within the conventional and sabermetric categories. These websites also provided important descriptive variables of each player, including age, team, league, season and a unique player identifier.

The data gathered through Baseball Reference and Fangraphs were initially collected by season from 2008 to 2020 and by category (batting, pitching, fielding). When all scraping was completed, there were eight tables of various statistics within each season that were merged together to result in one larger table for each season including all relevant batting, pitching and fielding statistics. After adjusting for minor merging errors found in the data, the data were merged again to form one large dataset that included season as a new column. It was from this point that the dataset was redivided to the 3,357-observation batters data and the 4,774-observation pitchers data, based on the conclusion that batters and pitchers do not have enough in common to be compared directly in the same analytical approach.

The batters data initially contained 80 relevant batting and fielding metrics, as well as data on awards such as All-Star berths, Gold Glove, Silver Slugger, Most Valuable Player and Rookie of the Year. Several measures of base-running ability were included among the 80 metrics as well, because speed and skill on the bases are additional aspects with which batters can contribute value to their team.

The pitchers data contained 67 relevant pitching statistics at first, as well as data on the same awards (plus Cy Young Award). Measures of hitting and base-running were not included, because an examination of the data revealed that very few pitchers contribute meaningfully to their team with their bat or on the bases.

Some statistics were removed from both datasets for a variety of reasons. First, there were some statistics that we perceived to be irrelevant or not useful for player evaluation (ex: strikeouts per win). Other statistics in the dataset were counting statistics that could be replaced with rate statistics that were still conventional in nature but served as more accurate measures of performance (ex: home runs allowed vs. home runs allowed per nine innings). In these cases the former statistic was dropped in favor of the latter. Finally, the most recent or advanced sabermetric statistic was chosen in situations where multiple editions existed in the data. An example of this is the selection of weighted runs created plus (wRC+) over weighted runs created (wRC). In other similar instances, some sabermetric statistics offered player values in terms of both runs added and wins added; in these cases, the statistics that gave values in terms of wins were selected based on our belief that wins are ultimately more important to a team than runs. These omissions left us with 58 batting statistics and 41 pitching statistics. It was necessary for us to categorize each of these non-awards variables as a conventional statistic or a sabermetric statistic. In determining the categorization of these variables, several factors were considered: the age of the statistic, whether it served as a counting variable, a function of a counting variable(s) or a context-dependent variable, as well as what categorization it was generally given by the baseball community. Table 1 and Table 2 below show the categorizations of these variables for batters and pitchers.

Based off our understanding that analysts considered the previous three years of statistics in their evaluation of players, each of these statistics were transformed into three year averages where available, with each season equally weighted 0.33. If a player only had two seasons of data available to that point, the data were weighted 0.5 each.

Categorization of Statistics for Batters

Conventional Statistics for Batters	**Sabermetric Statistics for Batters**
Hits (H)	Batting Average on Balls in Play (BABIP)
Walks (BB)	Line Drive Percentage (LD%)
Strikeouts (SO)	Ground Ball Percentage (GB%)
Doubles (2B)	Fly Ball Percentage (FB%)
Triples (3B)	Home Runs per Fly Ball (HR/FB)
Home Runs (HR)	Pull Percent (Pull%)
Total Bases (TB)	Center Percent (Center %)
Runs Batted In (RBI)	Opposite Field Percent (Oppo%)
Batting Average (BA)	Soft Contact Percent (Soft %)
On Base Percentage (OBP)	Medium Contact Percent (Med %)
Slugging Percentage (SLG)	Hard Contact Percent (Hard %)
On Base + Slugging (OPS)	Walks per Strikeout (BB/SO)
Stolen Bases (SB)	Weighted On Base Average (wOBA)
Caught Stealing (CS)	On Base + Slugging Plus (OPS+)
Runs Created per Game (RC)	Weighted Runs Created Plus (wRC+)
Putouts (PO)	Baseball Reference Batting Wins (Batting$_{BR}$)
Assists (A)	Fangraphs Batting Wins (Batting$_{FG}$)
Errors (E)	Baseball Reference Wins Above Replacement (WAR$_{BR}$)
Double Plays (DP)	Fangraphs Wins Above Replacement (WAR$_{FG}$)
Fielding Percentage (FLD%)	Baseball Reference Offensive WAR (oWAR$_{BR}$)
Range Factor per 9 Innings (RF/9)	Fangraphs Offensive WAR (oWAR$_{FG}$)
	Wins Above Average (WAA)
	Run Expectancy Wins (REW)
	Win Probability Added (WPA)
	Player Leverage Index (pLI)
	Win Probability Added per Leverage (WPAperLI)
	Clutch Hitting (Clutch)
	Offensive Winning Percentage (oWN%)
	Ultimate Base Running (UBR)
	Weighted Grounded into Double Plays (wGDP)
	Weighted Stolen Bases (wSB)
	Base Running Runs (BsR)
	Runs Saved on Good Plays (Rgood)
	Baseball Reference Defensive Runs Saved (DRS$_{BR}$)
	Fangraphs Defensive Runs Saved (DRS$_{FG}$)
	Baseball Reference Defensive WAR (dWAR$_{BR}$)
	Fangraphs Defensive WAR (dWAR$_{FG}$)

Table 1: We categorized the above statistics relevant to a batters' performance into two groups - conventional and sabermetric. Each group contains metrics related to batting, fielding and base-running because these are the three primary ways in which batters contribute to their team. In situations where both Fangraphs and Baseball Reference contributed versions of the same sabermetric statistics (ex: WAR), both were included due to variations in their respective formulas for calculating those statistics. These are denoted with $_{BR}$ and $_{FG}$ subscripts in the table.

Categorization of Statistics for Pitchers

Conventional Statistics for Pitchers	**Sabermetric Statistics for Pitchers**
Wins (W)	Left on Base Percentage (LOB%)
Losses (L)	Line Drive Percentage (LD%)
Complete Games (CG)	Ground Ball Percentage (GB%)
Shutouts (SO)	Fly Ball Percentage (FB%)
Saves (SV)	Soft Contact Percent (Soft %)
Hits (H)	Medium Contact Percent (Med %)
Runs (R)	Hard Contact Percent (Hard %)
Earned Runs (ER)	Batting Average on Balls in Play (BABIP)
Earned Run Average (ERA)	Earned Run Average Plus (ERA+)
Walks + Hits per Inning Pitched (WHIP)	Fielding Independent Pitching (FIP)
Run Support per 9 Innings (RS/9)	Expected Fielding Independent Pitching (xFIP)
Home Runs per 9 Innings (HR/9)	Skill Interactive Earned Run Average (SIERA)
Walks per 9 Innings (BB/9)	Baseball Reference Wins Above Replacement (WAR$_{BR}$)
Strikeouts per 9 Innings (SO/9)	Fangraphs Wins Above Replacement (WAR$_{FG}$)
Runs Against per 9 Innings (RS/9)	Game Leverage (gLI)
	Wins Above Average (WAA)
	Wins Above Average Adjusted (WAAadj)
	Wins Above Average Win-Loss Percentage (waaWL%)
	Full Season Win-Loss Percentage (FullSeasonWL%)
	Win Probability Added (WPA)
	Run Expectancy Wins (REW)
	player Leverage Index (pLI)
	Win Probability Added per Leverage (WPAperLI)
	Clutch Pitching (Clutch)
	Shutdowns (SD)
	Meltdowns (MD)

Table 2: We categorized the above statistics relevant to a pitchers' performance into two groups - conventional and sabermetric. Only pitching metrics were included, because it is rare that a pitcher contributes meaningfully to his team throughout a full season in any other way. In situations where both Fangraphs and Baseball Reference contributed versions of the same sabermetric statistics (ex: WAR), both were included due to variations in their respective formulas for calculating those statistics. These are denoted with $_{BR}$ and $_{FG}$ subscripts in the table.

3 Methods

3.1 Logistic Regression Approach

After gathering and cleaning the data, we used logistic regression through the caret package in R to predict player membership on the Top 100 lists [13]. At a broad level, logistic regression models were chosen as the primary analytical method because our response variable was a binary indicator, with a value of 0 if a particular batter did not make the Top 100 list that season, and a value of 1 if the batter did make the list. Thus, the goal of our logistic regression models was to determine how well a specific subset (eg: conventional/sabermetric) of the variables predicted players on the list. This goal was based on the assumption that a model with a higher prediction accuracy meant that the variables in that model were more important to the MLB Network analysts in their criteria for selecting the Top 100 players.

Prior to conducting any modeling, the relevant data were split, such that models would be fit based on 70% of the unique players in the data. The other 30% of the players were held back, reserved for testing prediction accuracy of the models after fitting.

However, even after eliminating unnecessary variables through the process described in the data section above, the data still contained a large number of covariates. This presented concerns about overfitting and potential high correlation between variables. As such, two variable selection methods were utilized to construct useful logistic regression models - principal component analysis and lasso penalization.

3.2 Principal Component Analysis

Principal component analysis (PCA) is useful in situations with large data sets, because it reduces the dimensions of the dataset by creating linear combinations of the original variables. These combinations, called components, are uncorrelated with each other and are ordered in such a way that so that the first components explain most of the variability in the original data [14].

PCA was applied to six different sets of predictor variables, three of which were related to the batters data and three to the pitchers data. These variable sets consisted of (1) the conventional statistics for batters, which make up the left side of Table 1, (2) the sabermetric statistics for batters, which make up the right side of Table 1, (3) the combination of all conventional statistics and sabermetric statistics for batters, (4) the conventional statistics for pitchers, which make up the left side of Table 2, (5) the sabermetric statistics for pitchers, which make up the right side of Table 2, (6) and the combination of all conventional statistics and sabermetric statistics for pitchers.

After conducting PCA on each group of statistics, the number of components were chosen within each group. Specifically, the minimum number of components were chosen such that 95% of the variance in the original variables was cumulatively explained by the combination of those components. Those selected PCA components were then used as covariates in eight individual logistic models, labeled as follows: (1) PCA components of conventional batters statistics, (2) PCA components of conventional pitchers statistic, (3) PCA components of sabermetric batters statistics, (4) PCA components of sabermetric pitchers statistics, (5) Combination of the PCA components used in models 1 and 3, (6), Combination of the PCA components used in models 2 and 4, (7) PCA components of all batters statistics, and (8) PCA components of all pitchers statistics. Below is an example of the model structure that was consistent across each model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{PCA Batter Conventional 1}+$$
$$\beta_2 \cdot \text{PCA Batter Conventional 2...}$$

Where p is the probability of a batter making the Top 100 list in a particular season.

It must be noted that in five of the eight models, the full range of PCA components that would be necessary to explain 95% of the variability within the original variables caused separation issues in the data. To account for this issue, some components were dropped in those models. The final number of components included in each model are shown in Table 3

Subset of Data Modeled	**Original # of PCA Components**	**Final # of PCA Components**
Conventional Batters Statistics	16	16
Conventional Pitchers Statistics	16	2
Sabermetric Batters Statistics	23	14
Sabermetric Pitchers Statistics	21	11
Conventional & Sabermetric Batters PCA Combination	39	39
Conventional & Sabermetric Pitchers PCA Combination	37	31
Conventional & Sabermetric Batters Statistics	29	29
Conventional & Sabermetric Pitchers Statistics	26	21

Table 3: In five of the eight logistic regression models containing PCA components, the original number of components required to reach 95% variability explained caused separation issues within the data. Thus, these models were reduced further to include only the components that would not induce this error.

3.3 Penalized Logistic Regression

LASSO penalization (abbreviated as PLR), when utilized in logistic regression models, provides an alternative approach to variable selection and dimension reduction. Rather than creating linear combinations of the original variables, lasso penalizes, or regularizes the coefficients in the model itself, pushing all of their values closer to zero, and forcing some to equal zero exactly, thus eliminating those coefficients' effect from the model. [15].

Rather than utilizing PCA components in the logistic regression models, the variables themselves were included. Six models were built, following in the same structure as the PCA models - (1) All conventional batters statistics, which make up the left side of Table 1, (2) All conventional pitchers statistics, which make up the left side of Table 2, (3) All sabermetric batters statistics, which make up the right side of Table 1, (4) All sabermetric pitchers statistics, which make up the left side of 2), (5) All batters statistics, and (6) All pitchers statistics. 10-fold cross validation was used to select the regularization parameter, λ, in each model.

3.4 Prediction Methodology

The remaining 30% of the data not initially put into the models were used, in conjunction with the model coefficients, to return a vector of probabilities corresponding to the likelihood that each player was assigned to the Top 100 list in that season. These probabilities were used to assign membership or non-membership to the list. Specifically, for each player in question, if the predicted probability was greater than 0.5, then that player would be assigned a 1, indicating they would make the list, and if the predicted probability was less than 0.5, the player would be assigned a 0, indicating that they would not make the list. These values were juxtaposed with the actual selections made by the MLB network analysts in confusion matrices to assess model prediction accuracy and performance.

Specifically, each confusion matrix offered five primary measures of prediction accuracy. Accuracy is an overall assessment of a model's ability to differentiate positive and negative cases correctly. In the context of this project, accuracy is calculated as the number of Top-100 players and non-Top-100 players correctly identified divided by the total number of players in the 30% of data used for prediction. Sensitivity refers to the proportion of positive cases identified correctly. For this project, sensitivity is computed as the number of players correctly identified as being Top-100 divided by the total number of Top-100 players in the prediction dataset. Specificity is the proportion of negative cases identified correctly, or stated in terms of this project, the number of non-Top-100 players identified correctly divided by the total number of non-Top-100 players in the prediction dataset [16]. Positive predictive value is defined as the proportion of cases predicted as positive that are actually positive. In the context of this project, the positive predictive value of a model is the proportion of players predicted to be Top-100 players that were actually Top-100 players. In contrast, negative predictive value is defined as the proportion of cases predicted as negative that are actually negative. In the context of this project, the negative predictive value of a model is the proportion of players predicted to be non-Top-100 players that were actually non-Top-100 players [17].

4 Results

4.1 PCA Results

The confusion matrix for the model built with PCA components made up of conventional batters statistics is included in Figure 1 below, followed by a chart (Table 4) summarizing important accuracy values between all models built on PCA components.

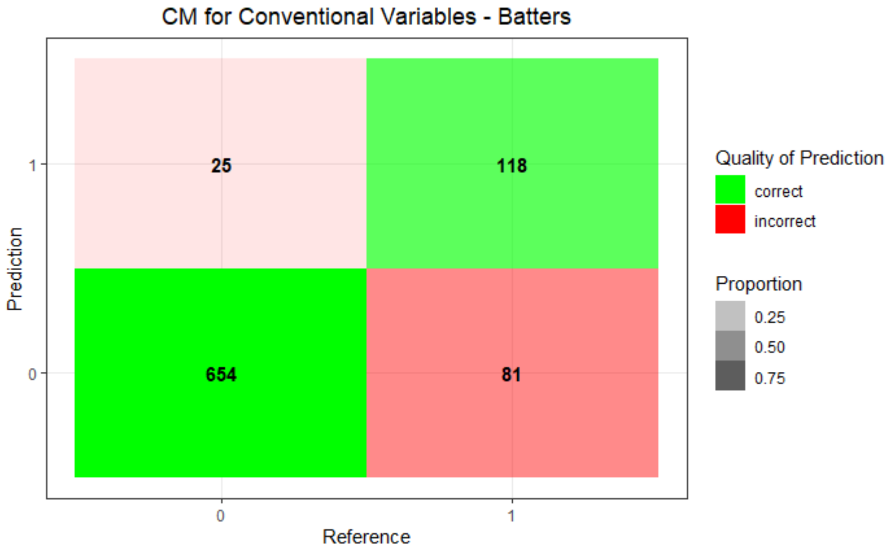

Figure 1: The prediction accuracy of the logistic regression model built with principal components of all conventional batters statistics summarized in a confusion matrix.

The values in Figure 1 indicate that the model correctly predicted 118 players who made the list, while incorrectly classifying 81 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 654 players who missed the list, while predicting that 25 players would make the list who actually were not selected.

Players	**Variables**	**Accuracy**	**Sensitivity**	**Specificity**	**Positive Pred Value**	**Negative Pred Value**
Batters	Conventional	0.879	0.593	0.963	0.825	0.890
Batters	Sabermetric	0.880	0.678	0.940	0.767	0.909
Batters	Conventional & Sabermetric PCA	0.879	0.678	0.938	0.763	0.909
Batters	Conventional & Sabermetric Statistics	0.885	0.698	0.940	0.772	0.914
Pitchers	Conventional	0.950	0.600	0.986	0.821	0.959
Pitchers	Sabermetric	0.961	0.730	0.985	0.832	0.972
Pitchers	Conventional & Sabermetric PCA	0.942	0.574	0.980	0.750	0.957
Pitchers	Conventional & Sabermetric Statistics	0.950	0.617	0.985	0.807	0.961

Table 4: The five accuracy measures of each logistic regression model built with PCA components.

Comparison of the various accuracy rates revealed that sensitivity was a good indicator of model performance. As seen in Table 4, the model built with conventional batting statistics resulted in a 0.593 sensitivity rate, or the proportion of batters who made the list that were selected by the model. The model built with sabermetric batting statistics gave a 0.678 sensitivity rate, an increase of almost 10 percentage points; this indicates that sabermetric statistics are notably more important to the analysts in their selection of batters on the Top 100 list. The final batters model, built from principal components of both conventional and sabermetric statistics, returned a sensitivity rate of 0.698, a small improvement over the model with only sabermetrics. This suggests that some conventional statistics may provide additional information about batters when utilizing sabermetric statistics that are relevant for determining those that make the list.

For pitchers, the model with conventional statistics returned a sensitivity rate of 0.60, while the model with sabermetric statistics returned a much higher rate of 0.73. This change is even larger than the difference between batting models and indicates that sabermetrics may have a more significant impact for the MLB Network analysts on deciding the best pitchers in the game than on deciding the best batters. The pitching model that utilized PCA components of all variables gave a 0.617 sensitivity rate, which was marginally better than the model with just conventional statistics and considerably worse than a model with just sabermetric statistics. This implies that sabermetrics alone are the preferred analytical measures of analysts in their evaluation of pitchers. For both batters and pitchers, sabermetric statistics resulted in models with higher sensitivity rates of 0.678 to 0.593 and 0.73 to 0.60 respectively when compared to models with just conventional statistics. This indicates that these models selected a greater proportion of players who actually made the list.

4.2 PLR Results

The confusion matrix for the LASSO penalized model built on conventional batters statistics is included in Figure 2, followed by a chart summarizing important accuracy values between the models.

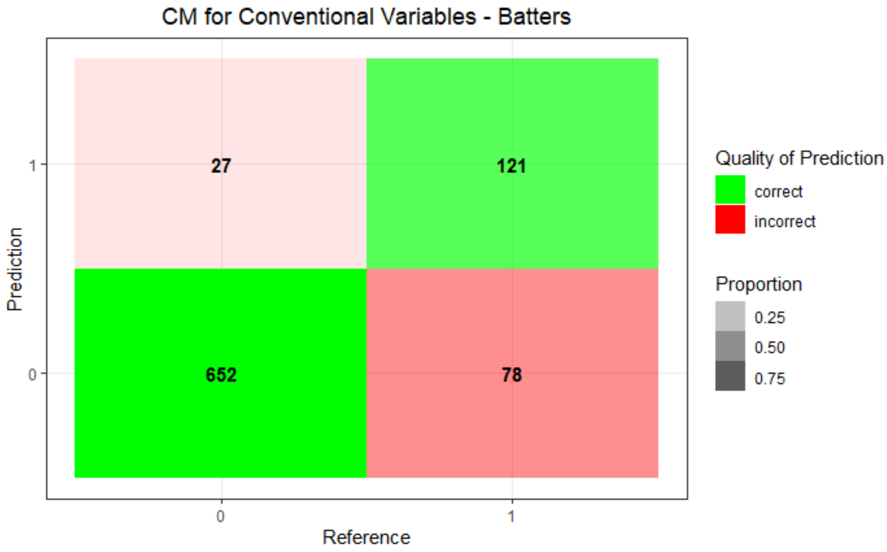

Figure 2: The prediction accuracy of the logistic regression model built with penalized lasso regression conducted on all conventional batters statistics, summarized in a confusion matrix.

The values in Figure 2 indicate that the model correctly predicted 121 players who made the list, while incorrectly classifying 78 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 652 players who missed the list, while predicting that 27 players would make the list who actually were not selected.

Players	**Variables**	**Accuracy**	**Sensitivity**	**Specificity**	**Positive Pred Value**	**Negative Pred Value**
Batters	Conventional	0.880	0.608	0.960	0.818	0.893
Batters	Sabermetric	0.884	0.693	0.940	0.771	0.913
Batters	Conventional & Sabermetric	0.883	0.693	0.938	0.767	0.913
Pitchers	Conventional	0.947	0.583	0.985	0.798	0.958
Pitchers	Sabermetric	0.954	0.696	0.981	0.792	0.969
Pitchers	Conventional & Sabermetric	0.951	0.670	0.980	0.778	0.966

Table 5: The five accuracy rates of each LASSO penalized logistic regression model.

The comparison of the various accuracy rates revealed that sensitivity was again a good indicator of penalized logistic regression model performance. The sensitivity rate of the penalized logistic regression model built with sabermetric batting statistics was 0.693, which was almost 10 percentage points higher than that of the model with conventional statistics at 0.608. This serves as more evidence that sabermetric statistics are given greater weight in the selection of batters by the MLB Network analysts.

The results of the PLR models do not indicate that additional value could be given by including conventional statistics as complementary to sabermetrics, as the sensitivity rate of the model combining all statistics is exactly the same as that of the sabermetric model.

For the evaluation of pitchers, the sensitivity rate of the conventional model was notably lower, at 0.583, than that of the sabermetric model, which was 0.696. In addition, the sensitivity rate of the model including both statistics is just 0.670. These results support the hypothesis that analysts consider sabermetrics to be considerably more important in their evaluation of pitchers.

5 Discussion

5.1 Comparison of Methods

The LASSO penalized logistic regression models that were built with batters statistics reveal similar patterns overall as those models built with PCA components. These trends are visualized in Figure 3 and summarized in the paragraphs that follow.

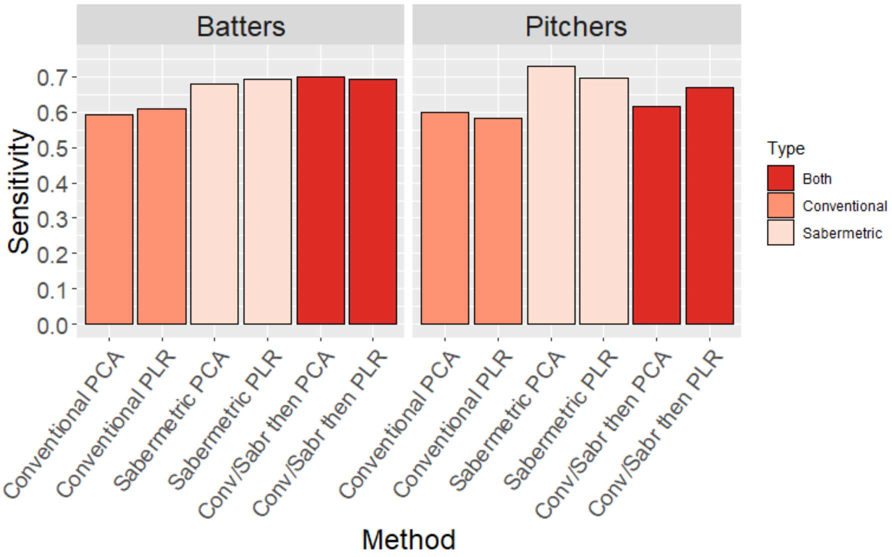

Figure 3: The sensitivity rate was seen to be the most productive method of comparison between models because it identifies the proportion of players on the list selected by the model. The plot reveals that the use of sabermetrics in some capacity to the evaluation method increases the sensitivity rate regardless of position. There appears to be a stronger indication that sabermetrics alone are better than a combination of sabermetrics and conventional statistics for the selection of pitchers, while the two variable selection methods present contradictory results regarding the same conclusion for batters.

When utilized to predict list membership among the test data set, both models built with the initial sabermetric set of variables gave considerably higher proportions of players on the list who were correctly identified than their comparable models built with the initial conventional set. The similar results between variable selection models lends strength to the conclusion that sabermetric variables are notably more favored by analysts in their evaluation of batters.

However, while the results of the models built with principal components indicated that additional value could be given by including conventional statistics as complementary to sabermetrics, the same indication was not found in the results of the penalized logistic regression models. Specifically, the PCA model built with all statistics performed slightly better in the proportion of players on the list who were correctly identified than the model built with just sabermetrics. However, this rate was exactly identical between the same two PLR models, which suggests that there may in fact be no additional value to adding conventional statistics to player evaluation that utilizes only sabermetrics.

For pitchers, the conclusions drawn from the two kinds of models support each other more holistically. In each variable selection method, the models built with the initial sabermetric set of statistics resulted in sensitivity rates that were more than 10 percentage points higher than those found from the models with the initial conventional set. These two differences in sensitivity rates are the largest differences found between any two models, and specifically are larger than the differences in the same models when applied to batting statistics. This suggests that the sabermetrics are not only more important than conventional statistics when evaluating pitchers, but that sabermetrics may be weighted even more heavily over conventional statistics when assessing pitchers than when assessing batters.

The models also agree in their estimation of the value of conventional statistics when added to sabermetrics in pitcher evaluation. In both variable selection methods, the models based on the initial combination of all statistics performed worse in their ability to correctly predict players on the list than those that just used sabermetric statistics.

5.2 Comparison of Conventional & Sabermetric

This study served as an broad investigation into which type of statistics - conventional or sabermetric - are most valued by analysts in their evaluation of players. Specifically, two variable selection methods, principal components analysis and lasso penalized logistic regression, were utilized in the construction of logistic regression models to measure which type of statistic resulted in better prediction of the players who were selected to the MLB Network Top 100 Players list. The results are intriguing, and present interesting future questions.

Specifically, sabermetrics appear to be valued above conventional statistics in a vacuum when assessing both batters and pitchers. This is not surprising, given the amount of new and effective sabermetric statistics that have been invented in the past two decades, as well as the increase in accessibility to those statistics and the marketing of sabermetrics as a whole. For example, statistics such as Wins Above Replacement and Fielding Independent Pitching were created prior to 2010, which gave enough time for these statistics to be proven effective, or at least intriguing, prior to the time period examined in the study [18]. Furthermore, websites such as Baseball Reference and Fangraphs have increased the accessibility to and understanding of sabermetrics for the general baseball community. Sabermetrics have also been assisted by the marketing efforts of writers such as Keith Law, who wrote *Smart Baseball* in 2017 to make fans aware of flaws in conventional statistics that were at least partially resolved by sabermetrics. All of these factors likely have contributed to sabermetrics being weighted over conventional statistics for analysts in the 2010's.

However, between batters and pitchers, the models present different conclusions regarding the complementary value of conventional statistics. Although there is conflicting evidence whether a combination of both groups may be the most preferred method of analysis for batters, the results were overwhelmingly in favor of a sabermetric-only method of evaluation for pitchers.

This may be partially explained by greater momentum to the movement that conventional statistics hold less value for pitchers than for batters. For example, as early as the 1990's, it became accepted that a pitcher was not directly responsible for the number of hits or runs they allowed, due to factors like their team's defense [18]. As metrics such as Fielding Independent Pitching were created to better isolate a pitcher's performance, even basic statistics such as win-loss record were seen as inaccurate. [19]. While some conventional batting statistics, such as batting average, have lost their value in similar ways, there remain a large number of conventional statistics that give information that many still interpret as valuable. For example, home runs are still seen as a fairly interpretable indicator of a batter's power, and walks can be viewed as a solid indicator of a batter's discipline. As such, the increased attempts by the baseball community to use sabermetrics to isolate a pitcher's performance from many of the older conventional statistics may explain why analysts are more apt to select sabermetric statistics alone in their evaluation of pitchers, and perhaps have not yet reached that analytical strategy for batters.

These conclusions present interesting consequences for the future of baseball from an analytical standpoint, consequences which may be thrilling to sabermetricians, and harrowing for the baseball purists. The analytics community appears to be more committed to sabermetrics than ever before. If the trends noted in this study continue into the future, this will undoubtedly increase the hold that sabermetrics have on the way the game is observed and analyzed. As such, an important question must be asked. Will the conventional statistics of a century prior grow obsolete and unused, or will they still be seen as valuable? Or will they simply hold a place in the game for the sake of tradition and nostalgia? This is likely a more complex question than it may appear to be on the surface, given the interwoven dynamic between fan desires and needs, the media's presentation of the game, and the performance of players and teams. However, a primary factor in the answer to this question may be the influence that analysts have on the system as a whole. If the indications of this study are correct and analysts have placed themselves on the sabermetric side of the baseball analytics spectrum, do they have enough power to convert parties that have generally been more prone to rely on conventional statistics, such as the media and fans? While the long term answer to this question is unknown, it is clear that sabermetrics have become a large part of the 21st century game.

5.3 Limitations

This study was limited in several ways that must be acknowledged. Primarily, the nature of the experimental design meant that the two distinct groups of statistics - conventional and sabermetrics - were examined together in large groups. As such, the conclusions from this study are fairly restricted to the groups of statistics as a whole. It could only really be said that one set of statistics appear to be more valuable than the other, rather than identifying any one specific statistic as more valuable than any other.

Furthermore, there were variables known to be relevant factors to the MLB analysts in the selection of the players on the list that were not able to be included in this study. They include intangibles and defensive position. Intangibles are inherently not measurable and could not have been included, but adding a more holistic position adjustment beyond simply batters and pitchers could have increased prediction accuracy of the models and better isolated the effect of conventional and sabermetric statistics.

It should be noted that the statistical techniques of MLB analysts are unknown. It may be possible that our logistic regression modeling procedure was not the most accurate method for replicating the analytical process or the relationships between various statistics that were used to select the Top 100 players.

While the best effort was made to accurately replicate the pool of players from which analysts would be selecting the top 100 players from each season, this result was still imperfect. First, there were several players who, because of injury, new entry to the league, or other factors, were included on the list but had not yet met the requirements to be kept in the dataset. As such, the simple at-bats or innings-pitched requirement was not a perfect system. In addition, the requirements used likely still left too many players in the pool who were never considered by the analysts, thus adding noise to the model that was unnecessary and acting as an overall hindrance to prediction accuracy. There may have been other factors besides at-bats or plate-appearances that would have better created an accurate selection pool.

5.4 Conclusions

This study was designed as an initial exploration into the types of statistics considered most important for player evaluation by baseball analysts. Due to the time constraints present with this analysis, there were a couple of additional angles of investigation that went unexplored. These could be relevant and intriguing in future study. One such addition to this project might be the addition of positional adjustments to the models beyond batters and pitchers. There may very well be different statistics that are more important for some positions than others, so these adjustments might produce more specific results. For example, first basemen and catchers are much more commonly known as power hitters than second basemen, so perhaps metrics of power are more relevant for evaluating first basemen than second basemen. A second potential approach to explore this question could revolve around limiting the dataset to just the players on the

Top 100 list, and using modeling techniques to determine which statistics are most important for analysts in ordering the players from 1-100, rather than selecting those 100 players as we focused on with this study.

To this point though, the key implication of this study is that baseball analysts appear to value sabermetrics over conventional statistics in their evaluation of players, especially when considering pitchers. How players are evaluated, not only by analysts, but also by teams, the media, and fans will play a crucial role in baseball's future, a future that this project has began to shed a light. We encourage further study into the ever-changing analytical state of baseball as the sport continues into a new decade.

References

[1] *What is a 40-man Roster?: Glossary.* URL: http://m.mlb.com/glossary/transactions/40-man-roster (page 1).

[2] Christopher J. Phillips. “The Bases of Data”. In: *Harvard Data Science Review* 1.2 (Nov. 1, 2019). https://hdsr.mitpress.mit.edu/pub/3adoxb26. DOI: 10.1162/99608f92.5c483119. URL: https://hdsr.mitpress.mit.edu/pub/3adoxb26 (page 2).

[3] Alan Schwarz. *A Numbers Revolution.* July 2004. URL: https://www.espn.com/mlb/columns/story?columnist=schwarz_alan&amp;id=1835745 (pages 2, 3).

[4] Joseph Abisaid and William Cassidy. “Traditional baseball statistics still dominate news stories”. In: *Newspaper Research Journal* 38.2 (2017), pp. 158–171 (pages 2, 4).

[5] *A Guide to Sabermetric Research.* URL: https://sabr.org/sabermetrics (pages 2, 3).

[6] *Sabermetrics: Baseball Analytics and the Science of Winning [Infographic].* Oct. 2015. URL: https://onlinegrad.syracuse.edu/blog/sabermetrics-baseball-analytics-the-science-of-winning/ (page 3).

[7] *A Primer on Statistics.* URL: https://sabr.org/sabermetrics/statistics (page 3).

[8] Cliff Corcoran. *Cliff Corcoran: How important was Moneyball to the success of the 2002 A's?* Sept. 2011. URL: https://www.si.com/more-sports/2011/09/22/moneyball-impact (page 3).

[9] Mike Boylan. *"Moneyball" and the Oakland A's: How Has It Been so Misunderstood?* Oct. 2017. URL: https://bleacherreport.com/articles/679950-revisiting-moneyball-and-the-oakland-as-how-has-it-been-so-misunderstood (page 3).

[10] *Reference.com WAR Explained.* URL: https://www.baseball-reference.com/about/war_explained.shtml (pages 3, 5).

[11] Neil Weinberg. *Complete List (Pitching).* URL: https://library.fangraphs.com/pitching/complete-list-pitching/ (page 3).

[12] Chad Raines. *Prior to COVID-19, MLB Front Offices were growing their analytics departments, as they should continue to do going forward.* July 2020. URL: https://baseballcloud.blog/2020/07/02/prior-to-covid-19-mlb-front-offices-were-growing-their-analytics-departments-as-they-should-continue-to-do-going-forward/ (page 3).

[13] Max Kuhn. *caret: Classification and Regression Training.* R package version 6.0-86. 2020. URL: https://CRAN.R-project.org/package=caret (page 8).

[14] Aaron Schlegel. *Principal Component Analysis with R Example.* Jan. 2017. URL: https://aaronschlegel.me/principal-component-analysis-r-example.html (page 9).

[15] Alboukadel Kassambara. *Penalized Regression Essentials: Ridge, Lasso, Elastic Net*. Mar. 2018. URL: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/ (page 11).

[16] Alireza Baratloo, Mostafa Hosseini, Ahmed Negida, and Gehad El Ashal. "Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity". In: *Emergency* 3.2 (2015), pp. 48–49 (page 11).

[17] David Felson. *Screening for Disease*. 2020. URL: https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_screening/ep713_screening5.html (page 12).

[18] *The Many Flavors of DIPS: A History and an Overview*. May 2010. URL: https://sabr.org/journal/article/the-many-flavors-of-dips-a-history-and-an-overview/ (pages 17, 18).

[19] Lisa Gray. *Why Pitchers' Win-Loss Records Have Lost Significance*. Nov. 2009. URL: https://bleacherreport.com/articles/294793-why-pitchers-win-loss-records-have-lost-significance (page 18).

A Appendix

A.1 List of Figures

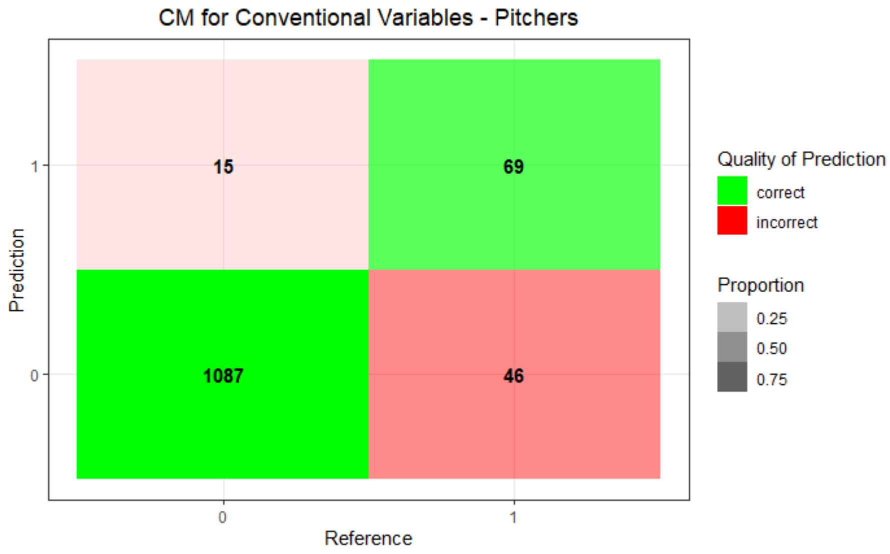

Figure 4: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with principal components of all conventional pitcher statistics. The values indicate that the model correctly predicted 69 players who made the list, while incorrectly classifying 46 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 1087 players who missed the list, while predicting that 15 players would make the list who actually were not selected.

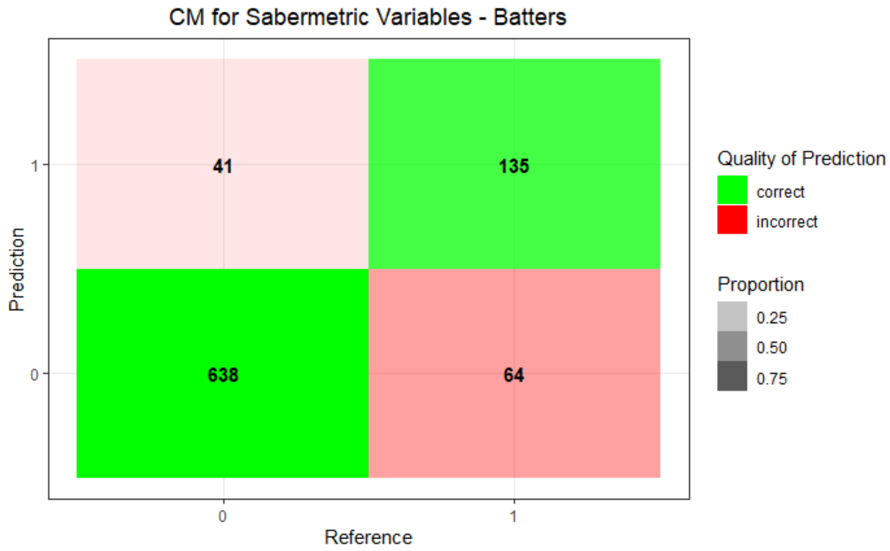

Figure 5: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with principal components of all sabermetric batter statistics. The values indicate that the model correctly predicted 135 players who made the list, while incorrectly classifying 64 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 638 players who missed the list, while predicting that 41 players would make the list who actually were not selected.

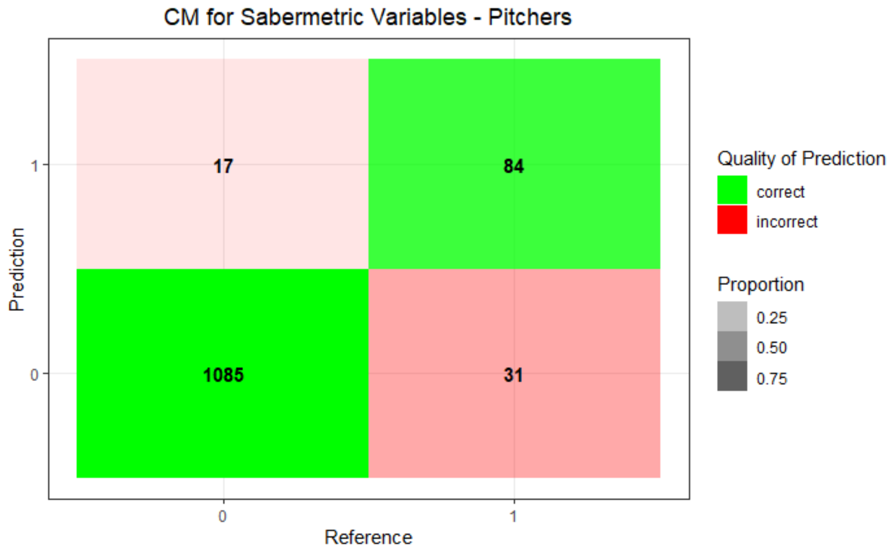

Figure 6: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with principal components of all sabermetric pitcher statistics. The values indicate that the model correctly predicted 84 players who made the list, while incorrectly classifying 31 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 1085 players who missed the list, while predicting that 17 players would make the list who actually were not selected.

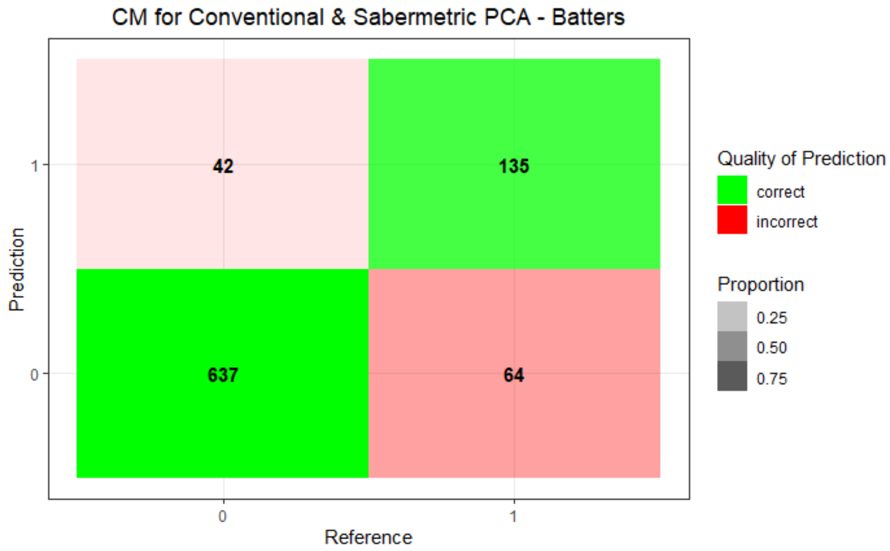

Figure 7: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with the combination of principal components of all conventional and sabermetric batter statistics. The values indicate that the model correctly predicted 135 players who made the list, while incorrectly classifying 64 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 637 players who missed the list, while predicting that 42 players would make the list who actually were not selected.

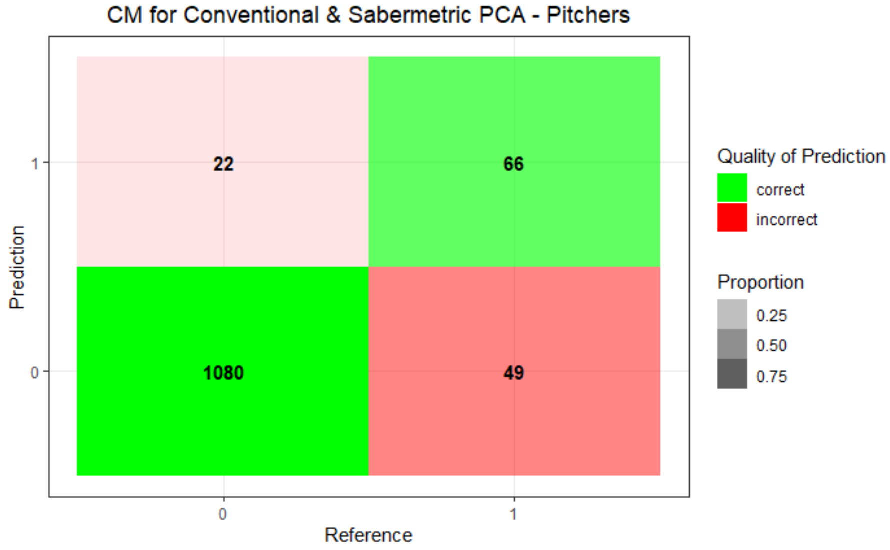

Figure 8: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with the combination of principal components of all conventional and sabermetric pitcher statistics. The values indicate that the model correctly predicted 66 players who made the list, while incorrectly classifying 49 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 1080 players who missed the list, while predicting that 22 players would make the list who actually were not selected.

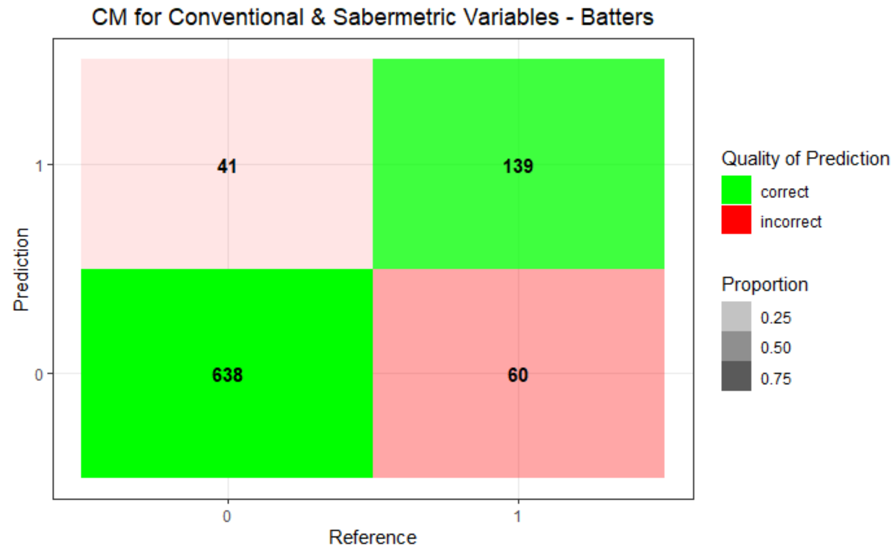

Figure 9: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with principal components of all conventional and sabermetric batter statistics. The values indicate that the model correctly predicted 139 players who made the list, while incorrectly classifying 60 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 638 players who missed the list, while predicting that 41 players would make the list who actually were not selected.

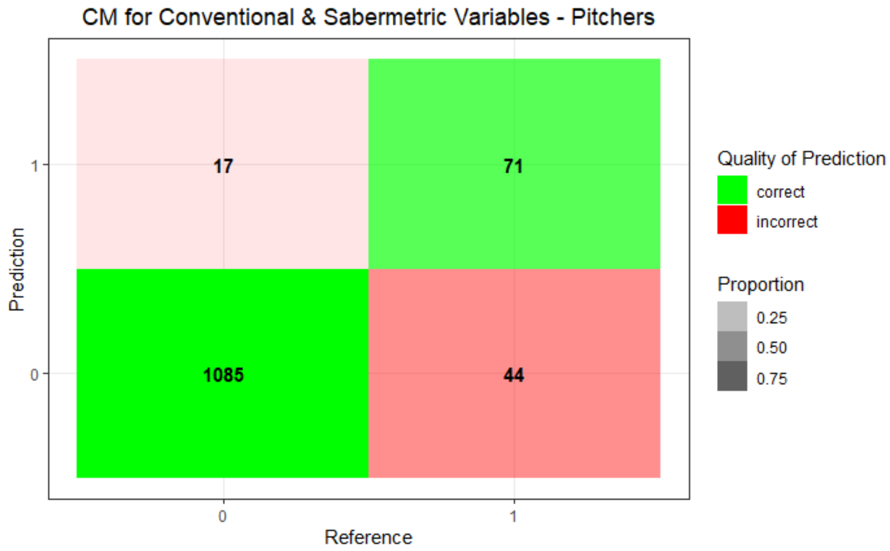

Figure 10: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with principal components of all conventional and sabermetric pitcher statistics. The values indicate that the model correctly predicted 71 players who made the list, while incorrectly classifying 44 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 1085 players who missed the list, while predicting that 17 players would make the list who actually were not selected.

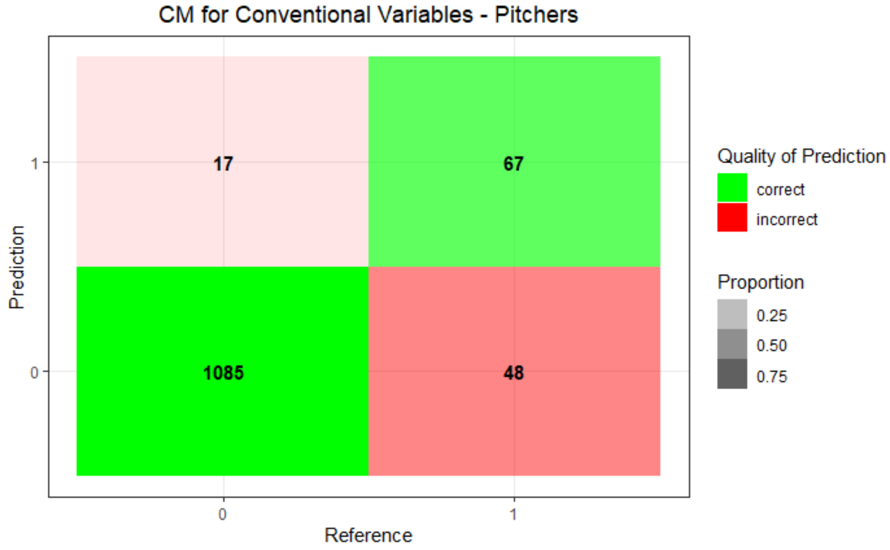

Figure 11: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with penalized lasso regression conducted on all conventional pitcher statistics. The values indicate that the model correctly predicted 67 players who made the list, while incorrectly classifying 48 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 1085 players who missed the list, while predicting that 17 players would make the list who actually were not selected.

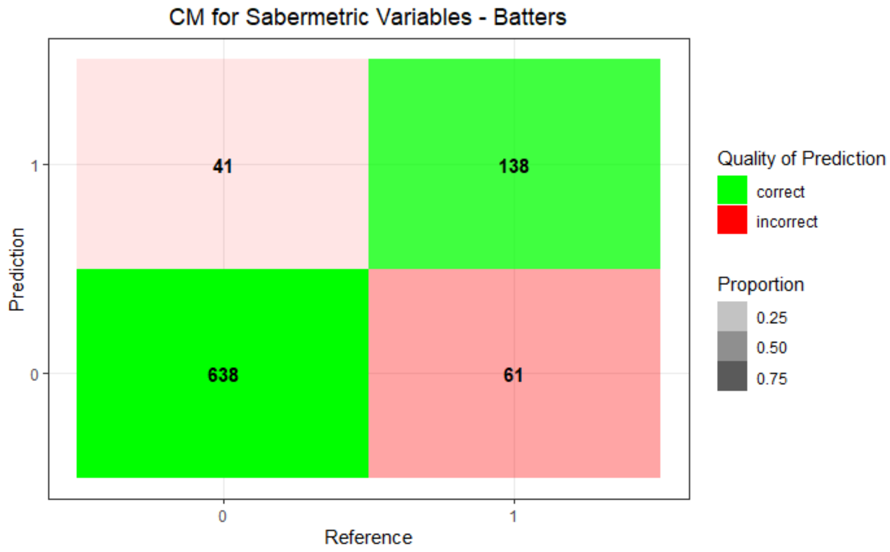

Figure 12: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with penalized lasso regression conducted on all sabermetric batter statistics. The values indicate that the model correctly predicted 138 players who made the list, while incorrectly classifying 61 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 638 players who missed the list, while predicting that 41 players would make the list who actually were not selected.

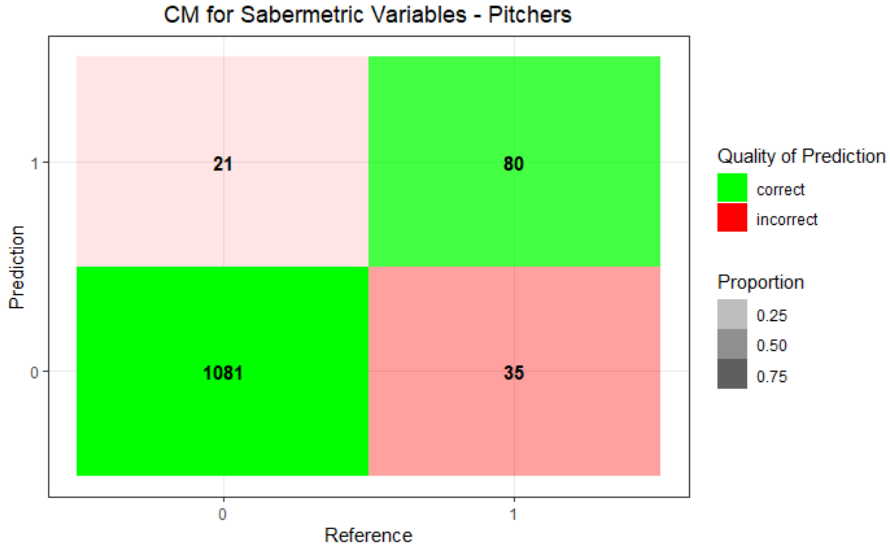

Figure 13: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with penalized lasso regression conducted on all sabermetric pitcher statistics. The values indicate that the model correctly predicted 80 players who made the list, while incorrectly classifying 35 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 1081 players who missed the list, while predicting that 21 players would make the list who actually were not selected.

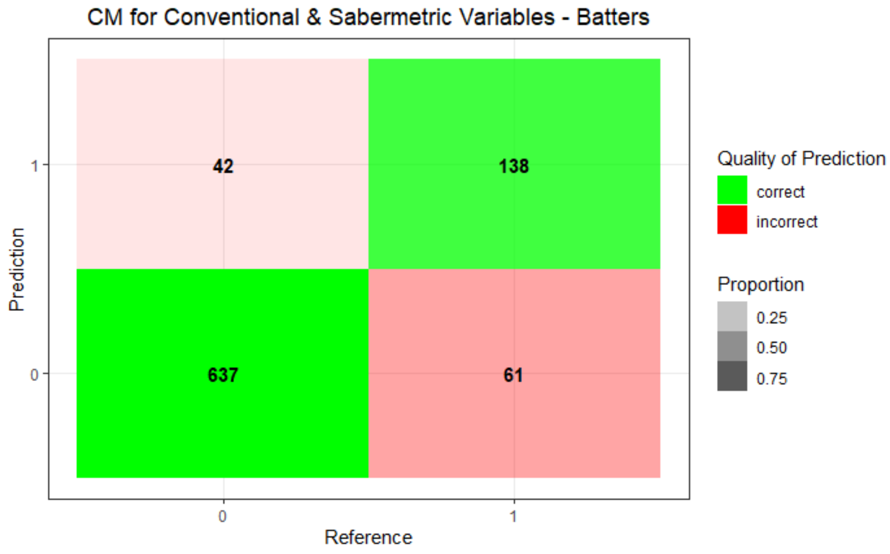

Figure 14: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with penalized lasso regression conducted on all conventional and advanced batter statistics. The values indicate that the model correctly predicted 138 players who made the list, while incorrectly classifying 61 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 637 players who missed the list, while predicting that 42 players would make the list who actually were not selected.

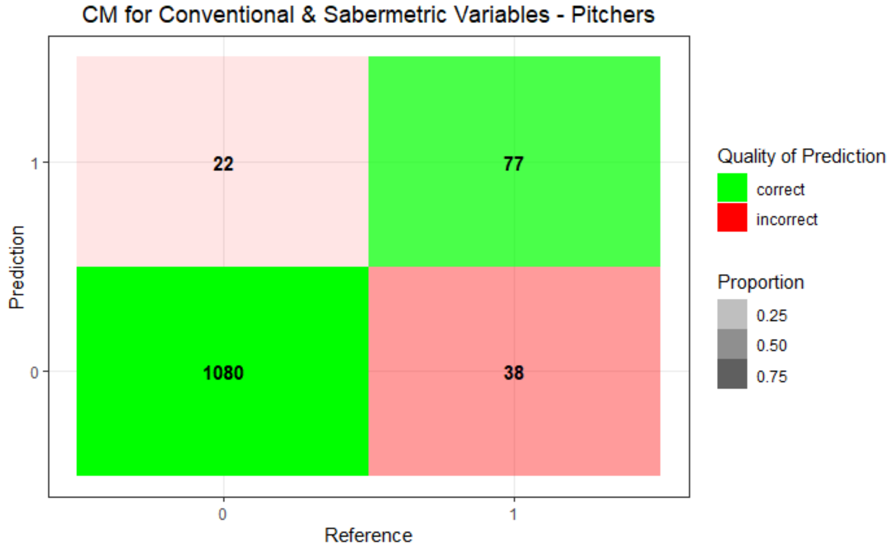

Figure 15: The confusion matrix above demonstrates the prediction accuracy of the logistic regression model built with penalized lasso regression conducted on all conventional and sabermetric pitchers statistics. The values indicate that the model correctly predicted 77 players who made the list, while incorrectly classifying 38 players as missing the list who actually were selected by MLB Network analysts. Inversely, the model correctly assigned 1080 players who missed the list, while predicting that 22 players would make the list who actually were not selected.