This is an Accepted Manuscript of an article published by Taylor & Francis in Applied Artificial Intelligence on February 26 2018, available online: https://doi.org/10.1080/08839514.2018.1442991

Machine Learning Applications in Baseball: A Systematic Literature Review

Kaan Koseler (koselekt@miamioh.edu) and

Matthew Stephan* (stephamd@miamioh.edu)

Miami University

Department of Computer Science and Software Engineering

205 Benton Hall

510 E. High St.

Oxford, OH 45056

Abstract

Statistical analysis of baseball has long been popular, albeit only in limited capacity until relatively recently. In particular, analysts can now apply machine learning algorithms to large baseball data sets to derive meaningful insights into player and team performance. In the interest of stimulating new research and serving as a go-to resource for academic and industrial analysts, we perform a systematic literature review of machine learning applications in baseball analytics. The approaches employed in literature fall mainly under three problem class umbrellas: Regression, Binary Classification, and Multiclass Classification. We categorize these approaches, provide our insights on possible future applications, and conclude with a summary our findings. We find two algorithms dominate the literature: 1) Support Vector Machines for classification problems and 2) k-Nearest Neighbors for both classification and Regression problems. We postulate that recent proliferation of neural networks in general machine learning research will soon carry over into baseball analytics.

keywords: baseball, machine learning, systematic literature review, classification, regression

1 Introduction

Baseball analytics has experienced tremendous growth in the past two decades. Often referred to as "sabermetrics", a term popularized by Bill James, it has become a critical part of professional baseball leagues worldwide (Costa, Huber, and Saccoman 2007; James 1987). All teams in Major League Baseball (MLB) have their own departments dedicated to such analysis and millions of dollars invested (Sawchik 2015). Popular websites such as Fangraphs[1] and Baseballsavant[2] further exemplify this popularity. There is also a growing body of academic literature investigating baseball analytics.

While analyzing baseball data is nothing new, analytics incorporating machine learning (ML) techniques are emerging. Machine learning is particularly suited to the data-heavy, discrete nature of baseball (Schumaker, Solieman, and Chen 2010). Professional baseball teams now collect data on nearly every aspect of the game. For example, the PITCHf/x system[3] generates large amounts of data by tracking pitched balls, which is particularly useful. Machine learning allows teams and other stakeholders to glean insights that are not readily apparent from human analysis.

In this article we perform a systematic literature review to categorize and summarize the applications of machine learning to baseball. Our goal is to establish the state of the art, help practitioners discover existing techniques, and guide future research. We categorize the approaches into the three problem classes defined by Smola & Vishwanathan (Smola and Vishwanathan 2008): Binary Classification, Multiclass Classification, and Regression. Additionally, we include representative examples for each category and speculate on potential future applications. While we perform an exhaustive survey of the publicly available literature, an important caveat to consider is that the public literature is inherently incomplete. Baseball analytics is a multi-million dollar and competitive industry. Professional organizations have a strong motivation to keep their work proprietary.

We begin in Section 2 with background information on baseball analytics to help

[1] http://www.fangraphs.com/
[2] https://baseballsavant.mlb.com/
[3] http://www.sportvision.com/baseball/pitchfx

establish the context of this paper. In Section 3, we outline the protocol we use for our systematic literature review based on established guidelines. We organize Section 4 by the three problem classes. It presents our findings on existing work, examples, and speculative potential applications. We summarize our results and conclude in Section 5.

2 Background

We assume that readers have a sufficient understanding of machine learning (Kelleher, Mac Namee, and D'Arcy 2015), including the three problem classes we use to categorize applications (Bishop 2006). Thus, we include only a brief primer on baseball analytics.

2.1 Baseball Analytics

Due to its wealth of data and discrete nature, baseball lends itself to statistical analysis more than any other sport. Many books have been written on the subject, and baseball teams have prominently embraced data-driven and statistical analysis (Lewis 2004; Costa, Huber, and Saccoman 2012; Baumer and Zimbalist 2013). Although the machine learning problem classes are known to these organizations, they are not often referred to as formally in mainstream literature and culture.

There are forms of statistical analysis applied to baseball unrelated to machine learning. Even simple statistics, such as batting average or a pitcher's win-loss record, have use in determining the success of a player or team. Prior to Bill James's popularization of more complex analysis in the 1980s, these simple metrics served as the statistical foundation of baseball for decades (Lewis 2004). Bill James is credited with popularizing the usage of *Sabermetrics*, although no precise definition for the term exists. In practice, *Sabermetrics* refers to any statistical analysis beyond the basic descriptive statistics (Costa, Huber, and Saccoman 2007).

One example of beyond-basic analysis is Bill James's Pythagorean expectation, which we present in Equation 1. It is a relatively simple formula, but goes beyond a basic win-loss ratio to calculate the expected number of wins for a team given their runs scored

and runs allowed. The formula is as follows (Miller 2005),

$$ExpectedWinRatio = RunsScored^2/(RunsScored^2 + RunsAllowed^2) \quad (1)$$

This Pythagorean expectation might be appropriate for a sports website or for amateur fans and analysts. The machine learning analyses that we present are likely more appropriate for professional analysts, enthusiasts with a mathematical/scientific background, and academics interested in the field.

3 SLR Protocol

We followed the established systematic literature review (SLR) protocol as defined by Keele et al. (Keele et al. 2007). This is to ensure our methodology is traceable and repeatable.

3.1 Research Questions

We used Pettigrew and Roberts' PICOC criteria to frame our research questions (Petticrew and Roberts 2008):

- **Population:** Baseball Analysts (academic or industry) and others interested in the intersection of machine learning and baseball analytics.
- **Intervention:** Machine learning techniques applied to baseball analytics. Specifically, machine learning techniques used for statistical analysis of performance.
- **Comparison:** Not applicable. We are interested in all techniques and classifying them.
- **Outcomes:** We look for techniques that evaluate past performance and/or inform future decisions.
- **Context:** Approaches from both academic research and publicly available industrial practice.

This allowed us to form the two research questions of this survey:

1. What are all the different ways machine learning has been applied to baseball?
2. What is the distribution of these applications across the machine learning problem classes?

3.2 Search Strategy

Our search strategy included all major online libraries relevant to this domain. This included IEEExplore, ACM Digital Library, Google Scholar, Citeseer Library, Inspec, ScienceDirect, Ei Compendex, Journal of Sports Analytics, International Journal of Computer Science in Sport, Sloan Sports Conference, and the Fan Graphs and Baseball Savant websites.

Our search strings involved finding the union of the word "Baseball" and a set of terms derived from the Bishop textbook on machine learning (Bishop 2006). Formally, our search string had the form of [*Baseball* & (Term1 OR Term2 ... OR TermN)]. The N terms included *machine learning, Binary Classification, Multiclass classification, Regression, supervised learning, unsupervised learning, novelty detection, statistical analysis, data analysis, data mining, prediction, analysis, models, evaluation, big data, inferring, inference, predict, stats, statistical, mining, data, model, modeling, neural net, markov, bayes, Bayesian, svm, support vector machine, hyperplane, expectation propagation, categorical, tuple, feature vector, feature, error rate, data cleaning, cross-validation, induction, regressor, decision tree, deep learning, reinforcement learning.*

3.3 Study Selection Criteria

Our selection of study criteria was 1) the study must address some aspect of the statistical analysis of baseball, and 2) the study must utilize a machine learning approach. We further include any study that describes using machine learning techniques in any form to any level of baseball. We exclude any study that does not focus on the statistical analysis

of performance. For example, highlight extraction from raw video, or business/profit analysis is not included.

3.4 Study Selection Procedure

Kaan Koseler performed the study extraction and collection using the search terms. Inclusion/Exclusion was performed through collaboration between Kaan Koseler and Matthew Stephan. Disputes to this effect were resolved through discussion.

3.5 Study Quality Assessment Procedure

Due to exploratory nature of this survey, we disregard study quality. We used several public, non-academic articles, which were not peer reviewed. We note these explicitly.

3.6 Data extraction strategy and Synthesis

The data extraction for each approach and article involved manually determining the problem class being explored and what the approach is being used for. We examined each of the papers in detail to extract out the results of their work.

Our synthesis of the extracted data was not a formal meta-analysis as we were not evaluating the success of some particular intervention. Rather, we were simply compiling an exhaustive list of studies and categorizing them by problem class. The synthesis consisted of determining the problem class(es) explored by the article and categorizing it accordingly.

3.7 Results of Protocol

In total, we found 145 articles using our search terms and strategy. Of those 145, 32 articles met our inclusion criteria and 117 were excluded. This high exclusion rate is explained by both the inclusion and the exclusion criteria. Many of the articles employed statistical analysis of performance without using a machine learning technique. This is in large part due to the search term "Regression" having two slightly different meanings

between traditional statistical analysis and machine learning. There are many applications of a Regression problem which involve analyzing the correlations between variables. This is not a machine learning problem as there is no prediction. There were also several studies found using machine learning focusing on analyzing ticket sales or other financial matters instead of performance.

4 Machine Learning Applied To Baseball

Machine learning's predictive power has led to its use in baseball for both practical and research applications. Machine learning analysis improves with increasing numbers of observations. This is readily illustrated when one considers extremes. Consider predicting if a pitcher's next pitch will be a fastball or not. If there are only a couple observations of the pitcher's past pitches, it will be nearly impossible to predict the next pitch accurately. However, if there are one hundred thousand observations the accuracy of the prediction will be fairly high. For example, Major League Baseball has the opportunity for a relatively large number of observations featuring seasons of 162 games per season with 30 teams. Additionally, the technology at this level is much more advanced than the amateur levels, providing access to more types and granularity of data.

We organize our presentation by each of the machine learning problem classes. We do this to explicate the distribution of applications, as expressed in our second research question. For each application, we firstly summarize each and include any related/extended versions of their use. We describe notable and illustrative examples in detail. Secondly, we present our insights on how each respective technique can be leveraged in this field in the future.

4.1 Binary Classification

4.1.1 Existing Work

Consider our earlier example of predicting whether a pitcher's next pitch will be a fastball or not. This is a Binary Classification problem, in that there are two classes: fastball and

non-fastball. Ganeshapillai and Guttag demonstrated excellent predictive improvements when using machine learning for this exact problem (Ganeshapillai and Guttag 2012; Hoang et al. 2015). They use a linear Support Vector Machine (SVM) to classify pitches based on data from the 2008 season and predict the pitches of the 2009 season. They consider 2008 pitching data as training data containing labeled examples of the pitch types thrown by pitchers. They achieved a significant performance improvement over a naive classifier, which is a simple classification based on probability such that, if a pitcher in 2008 used a fastball greater than 50% of the time, the naive classifier predicts every pitch in 2009 would be a fastball. The Ganeshapillai and Guttag model was able to correctly predict the next pitch 70% of the time, whereas the naive classifier was able to correctly predict the next pitch 59% of the time. Thus, they achieved an 18% improvement (59*1.18) over the naive classifier. In their study it was difficult to improve prediction for pitchers who overwhelmingly utilized one pitch, such as Mariano Rivera. Rather, their SVM approach holds more promise when a pitcher has a more diverse pitch repertoire.

Hoang et al. also studied the problem of classifying pitches into fastball and non-fastball categories (Hoang et al. 2015; Hoang 2015). Their work compares different Binary Classification algorithms' prediction accuracy. Their four algorithms included k-nearest neighbors (kNN), Support Vector Machine with linear kernel (SVM-L), Support Vector Machine with Gaussian kernel (SVM-G), and Linear Discriminant Analysis (LDA). The SVM approach fared the worst among the approaches they employed. LDA performed the best. LDA achieved roughly 78% accuracy, while SVM achieved 75% accuracy. This work was repeated by Hamilton et al. (Hamilton et al. 2014). They focused their work on improving feature selection and achieved a modest improvement in prediction accuracy by considering dominant features only.

Soto Valero compared different machine learning algorithms to predict the winner of a baseball game for all 30 teams in the MLB using 10 years of historical data as training data (Soto Valero 2016). They compared four algorithms : 1) k-nearest neighbors, 2) artificial neural networks/Multi-Layer Perceptron, 3) Decision Trees, and 4) SVM using

the SMO implementation. They indicate the SVM approach was the most successful, with prediction accuracy of just under 60%.

Work by Jang et al. proposed an approach to predict whether a given player in South Korea would be asked to join the South Korean national baseball team (Jang, Nasridinov, and Park 2014). They collected data on nine players that were on the national team, and used five candidate players as test data. A kNN algorithm was used, but no results were presented. This work was merely a proposal into exploring the feasibility of the kNN algorithm.

4.1.2 Discussion: Potential Applications

There are many potential future baseball applications for Binary Classification. Simple examples include classifying a match up between two teams as a win or a loss, deciding if a player will bunt, and classifying whether a certain team will choose to intentionally walk a player at bat. In the first example, observations might consist of a vector of players and their individual performances using statistics such as on-base percentage or more advanced statistics. For the bunting and intentional walk examples, situational data like outs, game score, and a player or manager's tendencies can be considered by analysts.

Considering the results demonstrated in the literature, we recommend starting with either an SVM approach or an LDA. Both algorithms have concrete evidence demonstrating good results in prediction accuracy. Although Hoang et al. suggest using LDA, the improvement they witness is marginal compared against the SVM approach. SVMs are fairly easy to construct and manipulate. Each analyst should determine through experimentation which algorithm is best for their problem domain and interests.

4.2 Multiclass Classification

4.2.1 Existing Work

A simple example of Multiclass Classification in baseball is classifying non-fastball pitches by pitch type, such as curve balls, change ups, or sliders. Even the "Fastball" category can be further subdivided into cutters, two-seam fastballs, four-seam fastballs, et cetera.

An analyst may want to classify pitches into one of more than two different classes.

Sidle's graduate work is the most expansive. They apply three different methods to classify pitches into seven different types. Sidle employs LDA, SVM, and bagged random forests of Classification Trees to classify pitches (Sidle 2017). They achieve improvements in prediction accuracy over a simple naive classifier. Their results demonstrate the forest of Classification Trees is the superior prediction method, with LDA second, and SVM behind. However, Sidle notes that LDA is both more efficient and more consistent than the forest of Classification Trees, with efficiency measured by computation time. Sidle found that using a naive guess, 51% of pitches thrown by starters were accurately predicted, compared to 57% accuracy for relievers (Sidle 2017). The major issue with pitch prediction is that there are different types of pitchers: starters and relievers. This difference in prediction accuracy is likely due to the need for starting pitchers to utilize a larger arsenal of pitches, whereas relievers may rely on a smaller number of pitches. Starting pitchers play more innings and throw more pitches than relievers, so a starter with only a few pitch types will quickly be exploited by the opposition. Because of this difference, models produced by different methods may show consistency differences in prediction. There are even differences between pitchers of the same type. Even among starters, some are easier to predict than others. Sidle shows that using a naive classifier, they can predict 88% of R.A. Dickey's pitches, as their pitches are almost exclusively knuckleballs. However, Juan Nicasio is harder to pin down with their pitches being predicted with 75% accuracy.

Bock performed similar work with Multiclass Classification of pitch type (Bock 2015). They use predictions to build a model of the pitcher's long-term performance. Specifically, they used both multinomial logistic regression and SVM algorithms to train their models. They derived an overall prediction accuracy of 74.5%, which was better than Sidle's prediction accuracy of roughly 65% across all three methods. From these predictions, they derived a pitch sequence predictability measure, termed "predictability index". They further used this index in a linear regression analysis in an attempt to predict long-term earned run average (ERA). Their analysis revealed that a pitcher's predictability index was not correlated with long-term ERA.

Attarian et al. used data from the PITCHf/x system to classify pitches thrown by certain pitchers into different pitch types (Attarian et al. 2013; Attarian et al. 2014). These types were based on different characteristics like spin rate and velocity of the pitch. They used a kNN algorithm and compared it against a naive Bayes classifier. The kNN algorithm achieved a 4% average improvement over the naive Bayes classifier. Another analysis they performed was using LDA to reduce the number of features for predicting pitch type. They reduced the features down to 4 dominant predictors: spin rate, spin direction, break angle, and start speed. However, using these 4 predictors provided only an 1.68% improvement in prediction accuracy.

For a university project, Ishii used clustering algorithms to determine undervalued players and classify them based on pitch type and repertoire (Ishii 2016). They used both k-means clustering and hierarchical clustering in their analysis, seeking to find players whose ERA was higher than their cluster ERA, which represents an average ERA for players of that skill level. Players who fit this criteria were deemed to be undervalued. Ishii found no difference in classification based on pitch type or repertoire. Both were equally effective in determining undervalued players when using clustering algorithms.

Tolbert and Trafalis used an SVM to determine the winners for the American League Championship Series, the National League Championship Series, and the World Series in the MLB (Tolbert and Trafalis 2016). Their analysis used different kernels for the SVM. They assessed the accuracy and examined which features were best in making a prediction. The SVM using a Gaussian Radial Basis Function kernel was the most accurate, with *wins* and *double plays turned* as the most predictive features.

One public and popular source employing Multiclass Classification is the Baseball-Savant website[4]. They use "hang-time", that is, the time that a batted ball spends in the air, and ball travel distance to categorize the "catchability" of a ball in play (Baseball Savant). This classification is based on the probability of a defender catching the ball. The classes themselves are explicitly delineated. For example,the $50-75\%$ catch probability class. Batted balls in this class are more likely than not to be caught, due to

[4] `https://baseballsavant.mlb.com/`

either a long hang time or short travel distance. These two features together determine the categorization of the batted ball.

4.2.2 Discussion: Potential Applications

As is the case for Binary Classification, the SVM approach is well-studied and effective for Multiclass Classification. One possible future approach is to group players into undervalued, correctly valued, and overvalued groups based on performance and salary. This would be of use to general managers and other personnel in a baseball team's front office. Rather than using traditional metrics, analysts could use advanced batting metrics to cluster batters and determine which "value" group they belong to. This can be combined with a salary analysis to allow baseball teams to determine players' worth and to assist in decisions to sign or cut players from the team. It can also be used to evaluate the performance of young players in the minor leagues and determine if they should be promoted.

Ishii has demonstrated that clustering using the k-means algorithm is effective for identifying undervalued players (Ishii 2016). This is what we recommend as a starting point for others looking to cluster players. Attarian et al. demonstrated that the kNN algorithm is effective for classifying pitch types, and combining this with LDA for feature selection is even more effective (Attarian et al. 2014). Baseball teams might utilize a similar approach when preparing for their opponents and studying their pitching habits.

4.3 Regression

4.3.1 Existing Work

Fichman and Fichman measure and show a decline in batting average over a player's career (Fichman and Fichman 2012). They use age as a feature vector, x, to derive a batting average, y. The results align with common knowledge that athletes tend to perform at a lower level as they age.

Healey (Healey 2015) uses the log5 model to assess the probability of a strikeout given a specific match up between a pitcher and a batter. The log5 model is an analogous

calculation to that of the Elo rating (Elo 1978), which is used in chess to predict the probability of a win in a match between two players. Additionally, log5 has commonly been used to evaluate a match between two teams to estimate winning percentages. We present its canonical usage in Equation 2 and 3. Let A be the winning percentage of team A and B be the winning percentage of team B. To know the probability, P, that team A will win against team B, one can use log5 as follows,

$$P = \frac{A - A * B}{A + B - 2 * A * B} \tag{2}$$

Healey modifies this basic formula to model the probability of a strikeout E^* as follows:

$$E^* = \frac{(BP)/L}{(BP)/L + (1 - B)(1 - P)/(1 - L)} \tag{3}$$

where B is the batter's strikeout rate, P is the pitcher's strikeout rate, and L is the average league strikeout rate. Healey further incorporates ground ball rates into the model (Healey 2017) and investigates their impact on the E^* formula.

Barnes and Bjarnadottir used Regression models to assess free agent performance by identifying undervalued and overvalued players (Barnes and Bjarnadóttir 2016). They used linear Regression, linear Regression with feature selection, Regression trees, and gradient-boosted trees. They determined linear Regression with feature selection models had the greatest potential for identifying highly overvalued or highly undervalued players. Feature selection indicated that wins above replacement (WAR) was the best statistic for predicting future performance. This was measured by calculating a surplus value, indicating performance greater than that predicted by the Regression model.

Das and Das used a neural network to analyze which aspects of a ball in flight contribute most to a fielder's ability to catch it (Das and Das 1994). By continually feeding velocity and elevation angle of balls in air, the neural network was able to predict the proper coordinates to position itself to catch the ball. Their final results indicated that towards the end the ball's flight, ball velocity becomes more impactful to catch probability than elevation angle, which is still a large contributor.

Everman created his own statistic, referred to as Calculated Aggregate Value (CAV), to predict the winner of a playoff series (Everman 2015). We present their algorithm in Equation 4. In evaluating a matchup, the team with the higher CAV was predicted to win. Everman evaluated the predictions of CAV and other common metrics using the 2004 MLB playoff season and states only that the CAV made the correct prediction nearly every time (Everman 2015). The author posits that the CAV statistic can be used as an excellent predictor in future research.

$$CalculatedAggregateValue = AdjustedProduction * WinningPercentage + FieldingAverage * WinningPercentage \quad (4)$$

Tung developed their own statistic attempting to measure a player's performance. They refer to it as the Offensive Player Grade (OPG) (Tung 2012). This metric measures a player's offensive performance while ignoring their defensive statistics. Tung used Principal Components Analysis (PCA) to develop this metric, analyzing the various offensive statistics of a player to determine which ones are relevant and to assign appropriate weights. They use k-means to cluster these players into groups based on their OPG score. Baseball analysts might find this metric of use when assessing a player's offensive value.

Freiman demonstrated the feasibility of using Random Forests to predict a player's election to the Baseball Hall of Fame (Freiman 2010). Results indicated that Freiman achieved 75% prediction accuracy using the Random Forests. Only 1% of the players who were actually elected were predicted not to be elected by Freiman's approach. They state the most important individual statistic to predict Hall of Fame election was the number of home runs throughout the player's career.

Ganeshapillai and Guttag developed a model to determine when a starting pitcher should be pulled from the game. They first trained a manager model that would pull the pitcher based on data that would be available to a manager during the game, and actual decisions made by major league managers as training data. They then built a regularized

linear Regression model to predict whether a pitcher, if not pulled, would surrender a run in the next inning (Ganeshapillai and Guttag 2014). This model disagreed with the manager model 48% of the time, achieving notable improvement in accuracy over the manager model. This accuracy was measured by whether or not the pitcher surrendered a run in the next inning. Of note, Ganeshapillai and Guttag's model accuracy and improvement over the manager model increases for each subsequent inning.

Herrlin explored fantasy baseball roster optimization (Herrlin 2015). They used a Bayesian approach to build models for both pitchers and batters. These models were used to build Regression trees that would be able to predict outcomes for the player throughout the rest of the season. They also explored batting order optimization using the results returned by the Regression trees. There was no single statistic used for this optimization, but rather different Regression trees modeling different statistics, such as batting average or ERA.

Huddleston used Bayesian machine learning to predict future performance in fantasy baseball using the single statistic of fantasy points to compare players (Huddleston 2012). Another difference between this analysis and Herrlin's is that Huddleston creates three different models that vary in their treatment of hitters and pitchers. The first model does not differentiate between the two, the second model distinguishes between hitters and pitchers only, and the third model further distinguishes starting and relief pitchers. These models are all trained by using prior distributions of points scored from previous seasons as training data. The results indicate the second model fits the data best.

Jensen et al. developed a detail Bayesian model to assess the evolution of hitting performance over a player's career (Jensen, McShane, Wyner, et al. 2009). They utilize several different techniques in building their model, including 1) hidden Markov Chains to model movement between "elite" and "non-elite" status, and 2) Gibbs sampling (George and McCulloch 1993) to estimate posterior distributions of home run rates. They describe the difference in home run rates between combinations of elite and non-elite designated hitters and shortstops. Beyond the common sense trend of declining performance with age, the authors also show that "elite" players have steeper declines than "non-elite"

players. Similar work has also been done by Stevens in attempting to model a pitcher's strikeouts and walks as they age (Stevens 2013). Using a logistic Regression model of the past 100 hundred years of historical pitcher data, they illustrate that Elite players suffer from greater declines in performance than non-elite players, but tend to maintain excellent performance until well into their 30s. The curves also show that peak performance tends to occur around age 25.

Jiang and Zhang used Bayesian methods to predict a player's batting average in the 2006 MLB season (Jiang, Zhang, et al. 2010). They used batting averages from the 2005 season as training data. Their main goal was to show the feasibility of Empirical Bayes over a simpler least-squares predictor (Hardy 1977). The results indicated that Empirical Bayes "may capture a great portion of the effects of missing covariables in the linear model" (Jiang, Zhang, et al. 2010). This leads to the recommendation that analysts consider using empirical Bayesian methods rather than a simple linear Regression least squares model.

Yang and Swartz employ a Bayesian approach to calculate the probability of a team winning a certain game (Yang and Swartz 2004). They combine home field advantage with past performance, batting ability, and starting pitchers in a two-stage Bayesian model. The first stage assumes that the probability of a team winning is a "random sample from a beta distribution with parameters based on the relative strength variable and the home field advantage variable." The second stage is a random sample from a Bernoulli distribution of the first stage's probability. This is combined with Gibbs sampling from a Monte-Carlo Markov Chain to make predictions.

Lyle uses a variety of techniques including SVMs, artificial neural networks, and model trees to predict several different offensive statistics (Lyle 2007). These are typical statistics used to evaluate a player's offensive prowess, such as runs, RBIs, hits, triples, and doubles. Lyle compared these techniques against existing baseball prediction systems such as the Player Empirical Comparison and Optimization Test Algorithm (PECOTA) (Silver 2003), which uses a nearest neighbor search comparing players to other players, and the Szymborski Projection System (ZiPS) (Prospectus 2012). The results showed that

Lyle's predictors were only able to outperform the existing PECOTA and ZiPS systems on the triples statistic. On all other statistics, the existing systems were superior. Both PECOTA and ZiPS are proprietary systems.

Panda uses penalized Regression models to reduce the number of features required to make player predictions (Panda 2014). Beginning with a total pool of 31 different offensive and five defensive metrics, Panda demonstrates using penalized Regression models results in a reduction to seven offensive metrics and two defensive metrics. These reduced metrics are what "distinguish a player over time" and were determined by analyzing their "signal". They indicate what percentage of players differed from the overall mean. Those metrics with a high signal were deemed to be worthy of inclusion in the model.

An unimplemented proposal by Reeves posits that the k-NN algorithm can be used to predict player performance through clustering (Reeves 2010). This is similar to the PECOTA system, but Reeves gives two major differences. Rather than assess only a three-year window of a player's career, Reeves assesses their entire career. The difference is that only one performance statistic will be generated, compared against the seven stat lines from PECOTA, each with their own confidence interval.

We discovered a public source that used Bayesian Regression to develop a model that would assess a player's batting ability, the LingPipe blog (Lingpipe Blog 2009). It is based primarily on a player's batting average, with the author measuring batting ability as a number between 0 and 1. Greater weight is given to consistent performance over increasing batting opportunities. For example, using their statistic classifies a player with 40 hits and 100 opportunities, a .400 batting average, as inferior to a player with 175 hits out of 500 opportunities, a .350 batting average. Another public resource describes importing data from EVN files into Microsoft Excel and running statistical analysis on a pitcher's "score"(Sidran 2005). This "score" metric is a basic evaluation of the pitcher and may be a useful tool for analysts looking for a simple but effective pitching metric. For example, balls and walks are worth -1; stikes, balls in play, and foul balls are worth +1; and singles, doubles, triples, and home runs are worth -1,-2,-3, and -4, respectively. This metric can be used to predict when a pitcher should be pulled from the game, in the

form of a probability that a pitcher will "falter" by falling below a certain running score threshold. This decision is based on past data indicating the point at which a pitcher has faltered in the past.

4.3.2 Discussion: Potential Applications

There are plentiful potential Regression problems in baseball analytics. For instance, an analyst might be interested in predicting a player's batting average for the season and use data collected from past seasons as training data to make that prediction. A variety of algorithms can be applied. In the literature we presented, various forms of Bayesian Inference were used often and achieved good performance. An analyst might also utilize a Linear Regression model, which is easier to implement, but this might result in decreased prediction accuracy.

Analysts wishing to predict a pitcher's ERA using data collected from past seasons can train a Bayesian model. Future research could employ applying Artificial Neural Networks for such tasks. Given the increasing popularity of deep learning libraries like TensorFlow and Torch, these might offer better performance and ease of implementation than the techniques employed in the articles we presented. Although few examples of neural networks were found during our survey, it is impossible to ignore their current domination of the machine learning field as a whole.

5 Summary and Conclusion

In total, we found 5 articles exploring Binary Classification, 8 articles for Multiclass classification, and 19 articles for Regression. These 32 articles were drawn from a pool of 145 candidate articles, of which 115 were excluded for either failing to meet inclusion criteria or for meeting the exclusion criteria. We summarize the articles from our systematic literature review and their respective problem classes in Table 1.

In reviewing the literature, we noticed several algorithms were used frequently. In particular, SVMs, KNNs, and Bayesian inference were popular approaches. Of the articles

we reviewed, Bayesian inference was the most common approach for Regression tasks, which are themselves the most common in the literature. Some of the articles we reviewed made use of existing machine learning software like WEKA or R to run their analyses. However, many researchers chose to implement their algorithms manually. This demonstrates the relative potential and need for implementation of these approaches. This helps illustrate that future researchers or analysts need not limit themselves to working with existing software. We present a ranking of the popularity of the approaches in Table 2 and summarize it in Figure 1. As shown, SVM and KNN approaches were used most often, each appearing at least 25% of the time. Despite the current domination of Artificial Neural Networks in general machine learning literature, they were used in only 9% of the articles meeting our inclusion/exclusion criteria. We anticipate that this will change.

Considering the recent growth in machine learning and the popularity of baseball, there are bound to be future researchers who study this intersection of baseball analytics and machine learning. In our review, we discovered that SVMs were the most popular method of classification, while Bayesian Inference mixed with Linear Regression was the most popular method for Regression tasks. It should be noted that SVMs can be used for classification tasks only. Bayesian Inference, however, can be used for both classification and Regression. The articles we presented used Bayesian Inference for Regression tasks only. We anticipate this will change in the coming years. There is currently dominance of neural networks in the machine learning literature and a proliferation of libraries like TensorFlow and Torch. These libraries allow users to quickly build and train neural networks. Neural networks also have the advantage of being useful for both classification and Regression tasks. It is our prediction that baseball analytic research will catch up to general machine learning research and begin employing neural networks for analysis.

Our hope is that this report will serve as a go-to resource for those interested in learning about the intersection of machine learning and baseball. We also hope to help facilitate those pursuing further research and baseball analysis.

References

Attarian, A, et al. 2013. "A comparison of feature selection and classification algorithms in identifying baseball pitches". In *International MultiConference of Engineers and Computer Scientists*, 263–268.

Attarian, A, et al. 2014. "Baseball pitch classification: a Bayesian method and dimension reduction investigation". In *IAENG Transactions on Engineering Sciences: Special Issue of the International MultiConference of Engineers and Computer Scientists 2013 and World Congress on Engineering 2013*, 392–399. CRC Press.

Barnes, Sean L, and Margrét V Bjarnadóttir. 2016. "Great expectations: An analysis of major league baseball free agent performance". *Statistical Analysis and Data Mining: The ASA Data Science Journal* 9 (5): 295–309.

Baseball Savant. *Statcast Catch Rates*. (Accessed January 8, 2018) . `http://baseballsavant.mlb.com/statcast_catch_probability`.

Baumer, Benjamin, and Andrew Zimbalist. 2013. *The sabermetric revolution: Assessing the growth of analytics in baseball*. University of Pennsylvania Press.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Springer.

Bock, Joel R. 2015. "Pitch Sequence Complexity and Long-Term Pitcher Performance". *Sports* 3 (1): 40–55.

Costa, Gabriel B, Michael R Huber, and John T Saccoman. 2012. *Reasoning with Sabermetrics: Applying Statistical Science to Baseball's Tough Questions*. McFarland.

— . 2007. *Understanding sabermetrics: An introduction to the science of baseball statistics*. McFarland.

Das, Rajarshi, and Sreerupa Das. 1994. "Catching a baseball: a reinforcement learning perspective using a neural network". In *AAAI*, 688–693.

Elo, Arpad E. 1978. *The rating of chessplayers, past and present*. Arco Publishing.

Everman, Brad. 2015. *Analyzing Baseball Statistics Using Data Mining*. (Accessed January 9, 2018). `http://truculent.org/papers/DB%20Paper.pdf`.

Fichman, Mark, and Michael A Fichman. 2012. "From Darwin to the Diamond: How Baseball and Billy Beane Arrived at Moneyball". (Accessed January 9, 2018). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2112109.

Firsick, Zachary. 2013. "Predicting Major League Baseball Playoff Outcomes Through Multiple Linear Regression". PhD thesis, University of South Dakota.

Freiman, Michael H. 2010. "Using random forests and simulated annealing to predict probabilities of election to the Baseball Hall of Fame". *Journal of Quantitative Analysis in Sports* 6 (2): 1–35.

Ganeshapillai, Gartheeban, and John Guttag. 2014. "A Data-driven Method for In-game Decision Making in MLB".

— . 2012. "Predicting the next pitch". In *Sloan Sports Analytics Conference*.

George, Edward I, and Robert E McCulloch. 1993. "Variable selection via Gibbs sampling". *Journal of the American Statistical Association* 88 (423): 881–889.

Hamilton, Michael, et al. 2014. "Applying machine learning techniques to baseball pitch prediction". In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods*, 520–527. SCITEPRESS-Science and Technology Publications, Lda.

Hammons, Christopher. 2006. "A Bayesian Approach to Markov Chain Baseball Analysis". PhD thesis, Georgetown College.

Hardy, Rolland L. 1977. "Least squares prediction". *Photogrammetric Engineering and Remote Sensing* 43 (4): 1905–1915.

Healey, Glenn. 2017. "Matchup Models for the Probability of a Ground Ball and a Ground Ball Hit". *Journal of Sports Analytics* 3 (1): 21–35.

— . 2015. "Modeling the probability of a strikeout for a batter/pitcher matchup". *IEEE Transactions on Knowledge and Data Engineering* 27 (9): 2415–2423.

Herrlin, Daniel Luke. 2015. "Forecasting MLB performane utilizing a Bayesian approach in order to optimize a fantasy baseball draft". PhD thesis, San Diego State University.

Hoang, Phuong. 2015. *Supervised Learning in Baseball Pitch Prediction and Hepatitis C Diagnosis*. North Carolina State University.

Hoang, Phuong, et al. 2015. "A Dynamic Feature Selection Based LDA Approach to Baseball Pitch Prediction". In *Trends and Applications in Knowledge Discovery and Data Mining*, 125–137. Springer.

Huddleston, Scott D. 2012. "Hitters vs. Pitchers: A Comparison of Fantasy Baseball Player Performances Using Hierarchical Bayesian Models". PhD thesis, Brigham Young University-Provo.

Ishii, Tatsuya. 2016. *Using Machine Learning Algorithms to Identify Undervalued Baseball Players*. (Accessed January 9, 2018). `http://cs229.stanford.edu/proj2016/report/Ishii-UsingMachineLearningAlgorithmsToIdentifyUndervaluedBaseballPlayers-report.pdf`.

James, Bill. 1987. *The Bill James Baseball Abstract 1987*. Ballantine Books.

Jang, Wu-In, Aziz Nasridinov, and Young-Ho Park. 2014. "Analyzing and Predicting Patterns in Baseball Data using Machine Learning Techniques". *Advanced Science and Technology Letters* 62:37–40.

Jensen, Shane T, Blakeley B McShane, Abraham J Wyner, et al. 2009. "Hierarchical Bayesian modeling of hitting performance in baseball". *Bayesian Analysis* 4 (4): 631–652.

Jiang, Wenhua, Cun-Hui Zhang, et al. 2010. "Empirical Bayes in-season prediction of baseball batting averages". In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, 263–273. Institute of Mathematical Statistics.

Keele, Staffs, et al. 2007. "Guidelines for performing systematic literature reviews in software engineering". In *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. sn.

Kelleher, John D, Brian Mac Namee, and Aoife D'Arcy. 2015. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press.

Lewis, Michael. 2004. *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Lingpipe Blog. 2009. *Bayesian Estimators for the Beta-Binomial Model of Batting Ability*. (Accessed January 8, 2018). https://lingpipe-blog.com/2009/09/23/bayesian-estimators-for-the-beta-binomial-model-of-batting-ability/.

Lyle, Arlo. 2007. "Baseball prediction using ensemble learning". PhD thesis, University of Georgia.

Miller, S. J. 2005. "A Derivation of the Pythagorean Win-Loss Formula in Baseball". *ArXiv Mathematics e-prints*.

Moy, D. 2006. "Regression Planes to Improve the Pythagorean Percentage: A regression model using common baseball statistics to project offensive and defensive efficiency". Master's thesis, University of California, Berkeley.

Panda, Mushimie Lona. 2014. "Penalized Regression Models for Major League Baseball Metrics". PhD thesis, University of Georgia.

Petticrew, Mark, and Helen Roberts. 2008. *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons.

Prospectus, Baseball. 2012. *Baseball Think Factory*.

Reeves, Jason. 2010. *Major League Baseball Performance Prediction*. (Accessed January 9, 2018). http://www.cs.dartmouth.edu/~lorenzo/teaching/cs134/Archive/Spring2010/proposal/cs134Proposal2/cs134.html.

Sawchik, Travis. 2015. *Big Data Baseball: Math, Miracles, and the End of a 20-Year Losing Streak*. Flatiron.

Schumaker, Robert P, Osama K Solieman, and Hsinchun Chen. 2010. *Sports data mining methodology*. Springer.

Sidle, Glenn Daniel. 2017. "Using Multi-Class Machine Learning Methods to Predict Major League Baseball Pitches". PhD thesis, North Carolina State University.

Sidran, D Ezra. 2005. *A Method of Analyzing a Baseball Pitcher's Performance Based on Statistical Data Mining*. (Accessed January 9, 2018). University of Iowa. https://www.researchgate.net/publication/267918769_A_Method_of_Analyzing_a_Baseball_Pitcher's_Performance_Based_on_Statistical_Data_Mining.

Silver, Nate. 2003. "Introducing pecota". *Baseball Prospectus* 2003:507–514.

Smola, Alex, and S.V.N Vishwanathan. 2008. *Introduction to Machine Learning*. Cambridge University Press.

Soto Valero, C. 2016. "Predicting Win-Loss outcomes in MLB regular season games–A comparative study using data mining methods". *International Journal of Computer Science in Sport* 15 (2): 91–112.

Stevens, Guy. 2013. "Bayesian Statistics and Baseball". PhD thesis, Pomona College.

Swan, Gregory, and Anthony Scime. 2010. "Winning Baseball Through Data Mining." In *DMIN*, 151–157.

Tolbert, Brandon, and Theodore Trafalis. 2016. "Predicting Major League Baseball Championship Winners through Data Mining". *Athens Journal of Sports* 3 (4): 239–252.

Tung, David D. 2012. *Data Mining Career Batting Performances in Baseball*. (Accessed on January 9, 2018). http://vixra.org/pdf/1205.0104v1.pdf.

Yang, Tae Young, and Tim Swartz. 2004. "A two-stage Bayesian model for predicting winners in major league baseball". *Journal of Data Science* 2 (1): 61–73.

Table 1: Included Articles Categorized by Problem Class

Problem Class	Title	Source
Binary Classification	Predicting the Next Pitch	Ganeshapillai and Guttag 2012
	Supervised learning in Baseball Pitch Prediction and Hepatitis C Diagnosis	Hoang 2015
	A Dynamic Feature Selection Based LDA Approach to Baseball Pitch Prediction	Hoang et al. 2015
	Analyzing and Predicting Patterns in Baseball Data using Machine Learning Techniques	Jang, Nasridinov, and Park 2014
	Predicting Win-Loss outcomes in MLB regular season games	Soto Valero 2016
Multiclass classification	Baseball pitch classification: a Bayesian method and dimension reduction investigation	Attarian et al. 2014
	A comparison of feature selection and classification algorithms in identifying baseball pitches	Attarian et al. 2013
	Applying machine learning techniques to baseball pitch prediction	Hamilton et al. 2014
	Winning Baseball Through Data Mining.	Swan and Scime 2010
	Using Machine Learning Algorithms to Identify Undervalued Baseball Players	Ishii 2016
	Using Multi-Class Machine Learning Methods to Predict Major League Baseball Pitches.	Sidle 2017
	Predicting Major League Baseball Championship Winners through Data Mining	Tolbert and Trafalis 2016
Regression	Great expectations: An analysis of major league baseball free agent performance	Barnes and Bjarnadóttir 2016
	Catching a baseball: a reinforcement learning perspective using a neural network	Das and Das 1994
	Analyzing Baseball Statistics Using Data Mining	Everman 2015
	Predicting Major League Baseball Playoff Chances Through Multiple Linear Regression	Firsick 2013
	Using random forests and simulated annealing to predict probabilities of election to the Baseball Hall of Fame	Freiman 2010
	A Data-driven Method for In-game Decision Making in MLB	Ganeshapillai and Guttag 2014
	A Bayesian Approach to Markov-Chain Baseball Analysis	Hammons 2006
	Forecasting MLB performance utilizing a Bayesian approach in order to optimize a fantasy baseball draft	Herrlin 2015
	Hitters vs. Pitchers: A Comparison of Fantasy Baseball Player Performances Using Hierarchical Bayesian Models	Huddleston 2012
	Hierarchical Bayesian modeling of hitting performance in baseball	Jensen, McShane, Wyner, et al. 2009
	Empirical Bayes in-season prediction of baseball batting averages	Jiang, Zhang, et al. 2010
	Baseball prediction using ensemble learning	Lyle 2007
	Penalized Regression Models for Major League Baseball Metrics	Panda 2014
	Regression planes to improve the pythagorean percentage	Moy 2006
	Major League Baseball Performance Prediction	Reeves 2010
	A Method of Analyzing a Baseball Pitcher's Performance Based on Statistical Data Mining	Sidran 2005
	Bayesian Statistics and Baseball	Stevens 2013
	Data Mining Career Batting Performances in Baseball	Tung 2012
	A two-stage Bayesian model for predicting winners in major league baseball	Yang and Swartz 2004

Table 2: Approaches Ranked By Prevalence

Approach	Included Articles Using Approach
K-nearest neighbors	9/32 = 28.1%
Support Vector Machine	8/32 = 25%
Linear Regression	7/32 = 21.8%
Tree-based Methods	6/32 = 18.75%
Linear Discriminant Analysis	5/32 = 15.6%
Bayesian Inference	5/32 = 15.6%
Artificial Neural Network	3/32 = 9.4%
K-means	2/32 = 6.25%
Principal Component Analysis	1/32 = 3.13%
Logistic Regression	1/32 = 3.13%

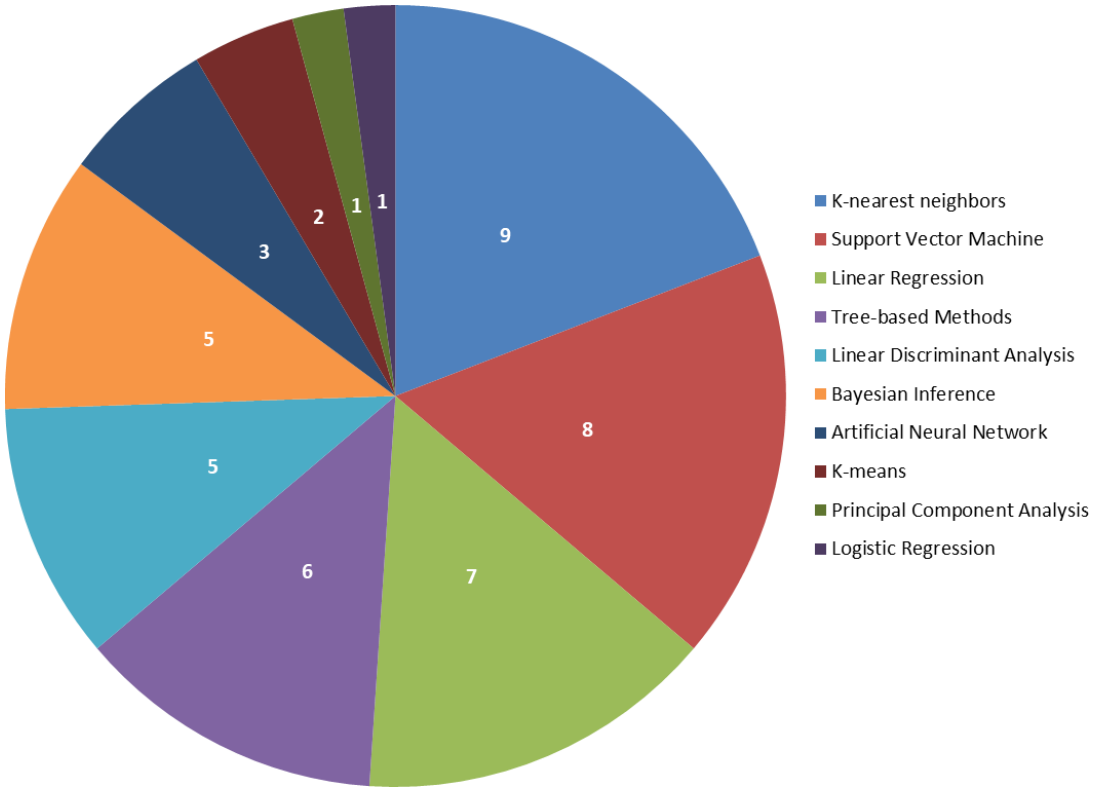

Figure 1: Frequency of algorithm approaches