Predicting MLB All-Stars from Prior Yearly Statistics

John McCarthy

August 3, 2020

Abstract

All-Stars in baseball can be labeled the best of the best that baseball has to offer in a given year. It is proven that All-Stars improve a team's chances of winning [7]. If we can predict All-Star caliber players earlier, it allows MLB teams to scout and trade more efficiently. Using the Lahman Database for batting, pitching, and fielding statistics, we train logistic regression and Support Vector Machines (SVM) models. Three separate levels of models were implemented: the original dataset, after manually excluding features that did not distinguish between All-Stars and non All-Stars through feature distributions, and after adding new percentage features that drive baseball research that were not listed in the original database. AUC scores and F1 scores were used to analyze the model's effectiveness. The best AUC scores of the logistic regression models were 0.790, 0.832, 0.743 for batting, pitching, and fielding respectively. The SVM model was only ran after excluding features and before adding new percentage features. The results were 0.739, 0.809, and 0.712 for batting, pitching, and fielding. Pitching performed the best for all iterations and fielding performed the worst. The batting percentage features were not weighted by the model in an intuitive fashion, Batting average and on-base percentage were negatively weighted while having a heavy positive influence on performance in the MLB. With the results, we can conclude that hits for batters and strikeouts for pitchers are strong positive indicators of All-Star status throughout Major League Baseball.

–

1 Introduction

Baseball has a great number of opportunities for analysis on tons of available statistics and player records. One major need for a baseball organization is predicting which players are going to produce and give them the best shot to win the World Series and build a dynasty. Professional baseball teams have bought in on the statistical revolution in baseball with every major league franchise establishing a department dedicated to statistical analysis [10]. Machine learning techniques are great for predictive analyses and there are a number of models that are applicable to baseball research.

In this paper, we explored predicting MLB All-Star selections from past performances and determining what the most meaningful features are for predicting whether a player will become an All-Star. We utilized the Lahman database to combine All-Star information with batting, pitching, and fielding data. With logistic regression as the model, we attained decent results in terms of ROC/AUC scores. The pitching models performed the best, while fielding models performed the worst.

With many predictive models created for team success, individual player success, and hall of fame voting, there has surprisingly been a lack of evidence of models predicting All-Star voting. All-Star status is a relatively effective way of determining top talent in Major League Baseball and is why it is worth predicting to provide teams with greater insight of which players will make a meaningful impact to the success of a MLB franchise.

2 Literature Review

Starting in the 1970s, baseball began to add advanced statistics to further quantify their game. These statistics and all baseball analysis deeper than the standard descriptive stats were coined as sabermetrics by Bill James, founder of the Society of American Baseball Research (SABR), hence the name, sabermetrics [2]. An example of an early sabermetric James defined was Runs Created (RC).

$$RC = \frac{(H + BB) * TB}{(AB + BB)}$$

This formula calculates the expected amount of runs a player will contribute to their team, beyond the traditional stats, Runs Batted In (RBI) and Runs (R). Even though baseball already had ways to calculate how players create runs for a team, James was not satisfied with the current measurement, so he created his own. More recently, researchers added to RC by creating the stat wRC+, which makes extra base hits more valuable than singles and assigns a numerical value with 100 being league average. In addition, wRC+ accounts for ballpark factors, acknowledging that every stadium isn't built the same and some heavily favor pitchers or batters [1]. The main point of going through these stats is to show that baseball researchers are dedicated to improving how they quantify baseball.

As technological advancements are made, the possibilities and applications for machine learning research and baseball do as well. For example, machine learning techniques have stormed sports analytics with models that can predict which players or which teams will outperform others [4] [5] [9].

Different prediction tasks target different audiences in the baseball community. For example, predicting match outcomes favors the fans of the game when participating in fantasy leagues and in betting practices, whereas predicting player outcomes also benefits the engaged fan, but is mostly oriented with team management, giving the general manager of a MLB organization more data to represent their players. Several models have been explored for questions in both these areas. Most often, neural networks, linear regression, or logistic regression are implemented when predicting match outcomes or player performance. A student researching at California Polytechnic State University, Bryce Farrell used a binary classification neural network to predict if MLB teams would reach their desired win total using team stats. His model predicted well with very poor and very good teams and struggled with middle of the pack teams [4]. At a lower level, an algorithm can be used to predict each individual MLB game. Collecting data from 2002 up to 2018, Roger Pharr, a machine learning researcher, predicted the winning team of an MLB game with 5 percent more accuracy than bookwriters in the field. Using extreme gradient boosting models, Pharr discovered than historical win spread was the most determining factor in predicting a winning team with pitching data close behind [9]. In the figure below, a larger value denotes a larger influence for predicting the team who will win a certain game in Pharr's study.

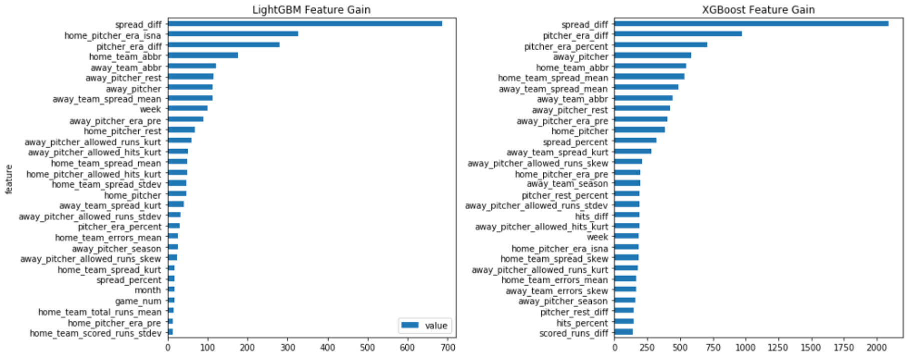

Figure 1: Total feature gains for predicting baseball game winners [9]

Transitioning to predicting player statistics, which helps team management and front office strategies, it is worth noting that MLB organizations spent a combined $492B on acquiring young talent [5]. In a section of his PhD dissertation at Chapman University, Christopher Watins sought out to predict a player's future wOBA (weighted on-base average). wOBA calculates overall hitting performance by assigning weights to different types of outcomes such as singles, doubles, triples, home runs, and walks. Watkins used 49 variables as features to his algorithms anywhere from simple identifiers such as age and name, to advanced sabermetrics such as wRC+. He discovered that from linear regression, ridge regression, elastic net regression, extreme gradient boosting, and neural networks that the regression techniques were the most effective in determining future wOBA in regards to mean absolute error (MAE) [11] [Table 1]

In addition to predicting stats for Major League Baseball players, researchers have also used machine learning to identify outstanding players. One example

Table 1: Weighted On Base Average (wOBA) measured in Mean Absolute Error (MAE) across different models [11]

Method	MAE
Linear Regression	0.0240
Ridge Regression	0.0241
Elastic Net Regression	0.0242
Extreme Gradient Boosting	0.0265
Neural Network	0.0243

is predicting who will be inducted into the Baseball Hall of Fame, in which a gradient boosting classifier was most successful for baseball data scientist, Micah Melling [6]. The features that were most important were All-Star appearances and batting average. The gradient boosting classifier as the most successful for this project is interesting because of the claim made by researchers, Mills and Salaga at the University of Michigan in 2011. They modelled Hall of Fame voting with random forests and found this method worked the best [8]. This denotes that the two studies had different intentions, which is true. The former study sought to predict future hall of fame players, whereas the latter study predicted hall of fame voting patterns. Nevertheless, the hall of fame is an exclusive highly prestigious club for any sport and admittance indicate a very successful MLB career.

Not only do All-Star appearances have a significance for Hall of Fame voting, but for team success as well. Through a multiple linear regression model, a researcher at Southern Utah University discovered that an additional All-Star would increase the win total for a team by an average 2.471 games, the probability of making the playoffs by 8.7 percent, and the chances of winning the World Series by 1 percent [7]. This is significant when teams are considering purchasing an All-Star caliber player's contract in free agency or through trades. The same research showed that 8 - 9 All-Stars on a team were optimal for winning games and playoff probability while 6 - 7 All-Stars was the most significant interval for World Series probability [7].

All-Star appearances, having the most importance in hall of fame prediction and significance in increasing team success should be investigated and predicted. However, it seems that the current research has yet to cover this area. To the best of our knowledge, the research closest to our inquiry in All-Star voting has covered the question of racial discrimination in fan voting and found that discrimination was not significant in fan voting patterns. In fact, at least in the 1990s, voters actually favored minorities such as African Americans and Latinos [3]. Similarly, many studies have sought to predict Hall of Fame candidacy and admissions, but few to none have explored predicting future All-Stars. With an overwhelming impact of finding prospective talent and all-star appearances on the success of a professional baseball career, it is a good idea to use prior MLB statistics to predict if a player has made an All-Star team or will in the future.

3 Purpose

1. Predict All-Stars

2. Elevate talent search technology and team management practices in professional baseball

3. Display the most important statistics in predicting All-Star status

4 Methods

4.1 Dataset

The data was collected from the Lahman Database, where Sean Lahman, a dedicated member of SABR, has created a number of dataframes covering all aspects of baseball, from player statistics as expected, all the way to how many games a player appeared at each position. For this analysis, four tables from the database were used including the batting stats, pitching stats, fielding stats, as well as one containing the all-star information. These tables are updated every year with new data from the previous MLB season and the data coverage starts in 1876.

To narrow down the volume of information to a manageable size, the information from the tables were only taken from the year 2000 on. This helps with regulating variables such as season length and the All-Star roster size for each league which were last modified in 1961 and 1998 respectively. Initially, each dataframe: batting, pitching, and fielding;were merged with the all-star data to differentiate between All-Star and non-All-Star players by adding a column to the original data for All-Star status. All-Stars were given a 1 in this column, whereas non-All-Stars were given a 0. The approach taken was one to continually upgrade model features as results were recorded at different levels. First, all of the numerical statistics included in the original dataframes were included. Then, we run experiments with non-distinguishing features removed. We select these by plotting feature distributions of each feature and remove the features which have distributions that look too similar to harbor a significant difference between All-Stars and non-All-Stars. After the exclusion of features, percentage features were added to attempt to explain at a deeper level, meaningful connections between the existing features in the dataset. Features that could be constructed from the original dataset were added to the data columns for the linear regression models. For the batting data, there were four new columns added: batting average (AVG), on-base percentage (OBP), slugging percentage (SLG), and on-base plus slugging percentage (OPS). For the pitching data, only one feature was added to the data, walks plus hits over innings pitched (WHIP), which explains the average number of baserunners a pitcher allows in one inning. Fielding data did not receive any new features because of the inability to create new columns for the prominent percentage features due to a lack of data. The percentage features that were added were chosen because they carry the most weight in player examination and are easily interpretable.

4.2 Models

The first model used was a logistic regression model as this is well suited for our binary classification task of predicting All-Star versus non-All-Star status. Logistic regression is a linear model that weights different features in order to predict a certain label that has two outcomes. When fitting the model on the data, the model determines a threshold value that is equal to the sum of the weighted features when the probability of the label being one class or the other is 50 percent. After a threshold has been established any probabilities calculated from the sum of the weighted features that is over 0.5 is predicted as the 1 class (All-Star) and any value below 0.5 is predicted as the 0 class (non-All-Star). To get the probability of the correctness of the prediction, the value of the weighted features' sum is plugged in this formula as t:

$$f(t) = \frac{1}{1 + e^{-t}}$$

$f(t)$ is the probability of the 1 class encompassing the datapoint in question.

The other model that was used in this study were kernel support vector machines (SVM). We used SVM to assess our output with a model that could capture non-linearities. SVM can capture non-linear relationships between variables whereas our vanilla logistic regression does not. Intuitively, when presented with multi-dimensional data, SVM attempt to create hyperplanes which separate data classes. When creating the decision boundary, SVM can be thought of as maximizing the margin between the two classes of data or minimizing the hinge loss function which penalizes incorrectly classified predictions. Prior to running SVM, we exclude handpicked features to reduce runtime. The features were selected by examining feature distributions and by applying baseball knowledge.

4.3 Metrics

Receiver Operator Characteristics/Area Under Curve (ROC/AUC) describes the relationship between true positive findings for the logistic regression model against the false positive rates. This goes deeper than accuracy, analyzing how the model actually does with positive cases, in this example, All-Star players. AUC scores can account for class imbalances, such as the high amount of non-All-Stars compared to the low amount of All-Stars better than accuracy.

The other metric that was used to evaluate the models was F1 score. This metric takes precision and recall for the model and computes a percentage that emphasizes the lower value to describe how well the model is performing. Precision is the amount of correct positive (All-Star) predictions divided by the total amount of predicted positives, and recall follows similar logic but with the negative (non-All-Star) predictions. The formula for F1 score is the following:

$$\frac{2 * (precision) * (recall)}{(precision) + (recall)}$$

5 Results

5.1 Original Dataset

For all the numerical features included in the unaltered dataset, the AUC scores for the batting data, pitching data, and fielding data for logistic regression are 0.744, 0.830, and 0.743 respectively. The F1 scores for the batting, pitching, and fielding data are 0.189, 0.332, and 0.144 respectively.

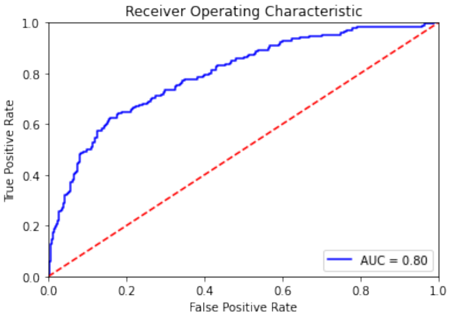

Figure 2: ROC curve for batting data on original dataset

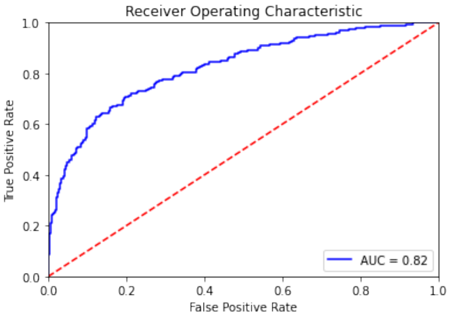

Figure 3: ROC curve for pitching data on original dataset

5.2 Excluded Features

After manually selecting and excluding features, the AUC scores for the batting, pitching, and fielding data are 0.790, 0.819, and 0.704. Additionally, the F1 scores were 0.473, 0.452, 0.081 respectively. As for the SVM results the AUC scores for the batting, pitching, fielding are 0.739, 0.809, and 0.712. No F1 scores were recorded for the SVM model.

5.3 Adding New Features

In addition to the previous edition of the logistic regression model, the model including the new features warranted an AUC score of 0.679 and an F1 score of 0.195 for the batting data.With the addition of the walks plus hits divided by

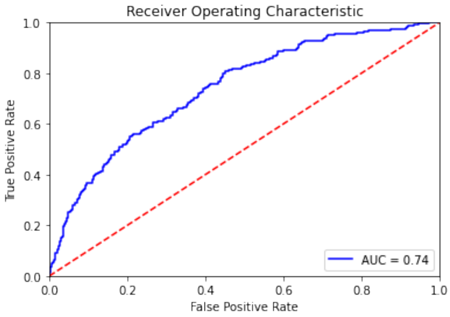

Figure 4: ROC curve for fielding data on original dataset

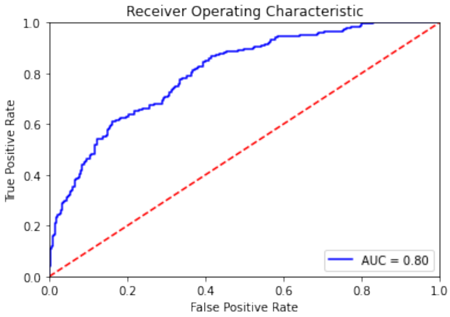

Figure 5: ROC curve for batting data after excluding features

innings pitched (WHIP) column, the new pitching model gave an AUC score of 0.832 and an F1 score of 0.484.

6 Discussion

Each dataset: batting, pitching, and fielding; had their best results in different models. Fielding data had better results with the original dataset, batting data had its best results after excluding features, and the pitching model performed best with its added percentage feature, walks plus hits over innings pitched (WHIP). Initially we expected each model would improve with each change to the datasets, but this was not the case for batting and fielding mod-

Data Type	Original Dataset	Excluded Features	Percentage Features
Batting	0.744	**0.790**	0.679
Pitching	*0.830*	*0.819*	***0.832***
Fielding	**0.743**	0.704	N/A

Table 2: Logistic Regression AUC Scores. *Italicized text* refers to the best score for iteration of dataset/ column. **Bold text** refers to the best score for a datatype/row

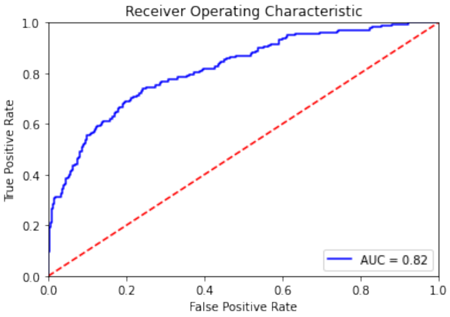

Figure 6: ROC curve for pitching data after excluding features

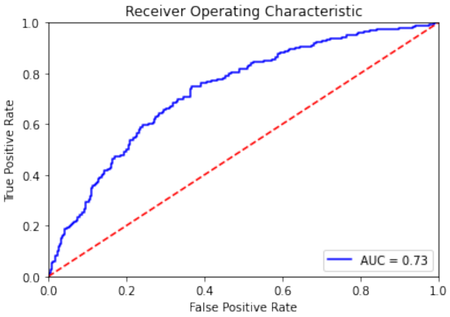

Figure 7: ROC curve for fielding data after excluding features

els. The weights learned by the models were both intuitive and not intuitive in the batting and pitching data. It was found that the fielding data has little conclusive evidence to offer.

6.1 Batting Model Discussion

For the batting data, hits (H) was the most positively weighted coefficient for all three models and stood out more than any other feature with walks (BB) and slugging percentage (SLG) with the next most positive weights [Figure 10]. Intuitively, these results make sense because hits and walks are the main factors in explaining how often a hitter reaches base and slugging percentage explains the quality and power in the hits a player achieves.

This raises an interesting point about the percentage feature on-base percentage (OBP), which has a strong negative weight in predicting All-Star status. OBP should go hand in hand with hit (H) totals and walks (BB) because OBP is just the rate at which players get on-base, or when broken down, accumulate hits and walks. Many features in the batting data after adding percentage features had weights that contradicted their influence in baseball. OBP, home runs (HR), at-bats (AB), and games played (G) had negative weights when these stats actually have positive effects in the MLB and grounding into double plays (GIDP) had a positive weight when the action is not positive in a baseball game.

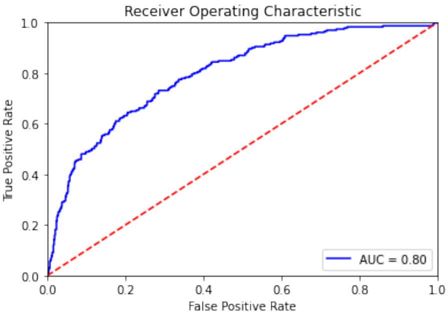

Figure 8: ROC curve for batting data after adding new features

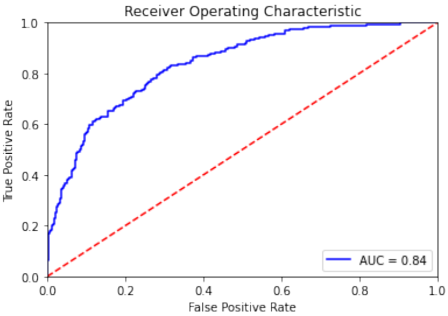

Figure 9: ROC curve for pitching data after adding new features

6.2 Pitching Model Discussion

For pitching data, the feature weights were also interesting. Strikeouts (SO) was the most positively weighted feature, while games started (GS) was overwhelmingly the most negative. As expected, reputable statistics for pitchers such as earned run average (ERA), walks plus hits divided by innings pitched (WHIP), and batting average of opponents (BAOpp) were negatively weighted [Figure 11]. These metrics measure how well the opposing lineup performs against a certain pitcher, so the lower numbers for ERA, WHIP, and BAOpp mean better results for a pitcher and a better chance to make the All-Star Game. It is also worth noting, GS being extremely negative could be because of the mix of relief pitchers and starters being in the same data. Relievers do not start many games if any, but less pitchers of this distinction make the All-Star Game.

6.3 Fielding Model Discussion

Training the model with fielding data performed the worst overall. Fielding data had the lowest AUC and F1 scores. This makes sense because fielding has been proven to have the least amount of variance between all MLB players out of batting, pitching, and fielding [12]. Most of the provided stats for fielding data is circumstantial. Certain positions such as center fielders (CF), shortstops (SS), and first basemen (1B) will have more put-outs (PO) and assists (A) then other positions, but does not necessarily mean they are superior fielders.

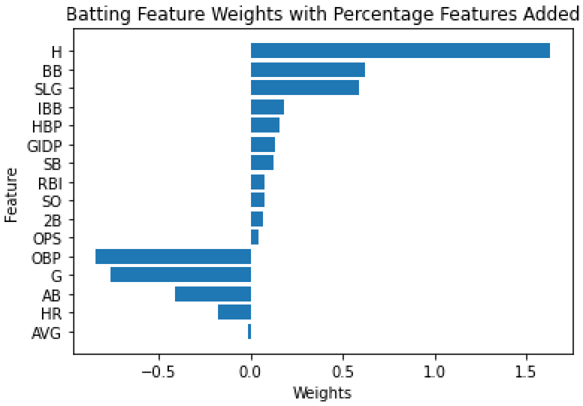

Figure 10: Batting model feature weights after adding new percentage features

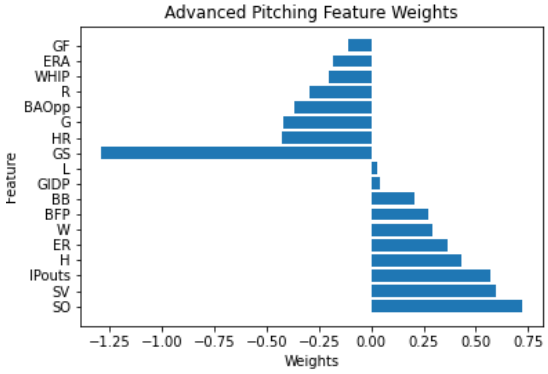

Figure 11: Pitching model feature weights after adding new percentage features

Fielding does have more descriptive statistics such as zone rating, and fielding percentage, but these were unable to be generated from the dataset. Again G has a strong negative weight [Figure 12]. Games played for all three datasets had a negative weight in the model. This does not make much baseball sense because All-Stars play more games on average than non-All-Stars as seen in the feature distributions for games played. A possible explanation could be that the model is attempting to artificially mold that percentage notion penalizing players who have more hits simply because of playing more games and having more at-bats. At-bats is also a negative weight in the batting data.

7 Conclusion

Baseball teams need their players to perform well and being selected to the All-Star Game is a good way of qualitatively determining success in the MLB. Prior studies have explored MLB Hall of Fame voting and have shown that All-Star appearances have the greatest effect on Hall of Fame status. With that being said, this paper addressed the gap by creating a logistic regression

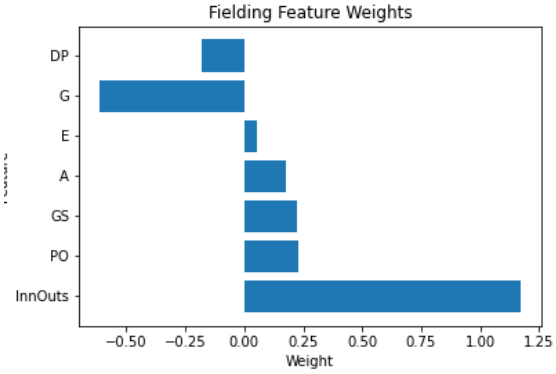

Figure 12: Fielding model feature weights after excluding handpicked features

model to predict All-Star appearances. Knowing which player stats influence All-Stars the greatest can allow teams to identify talent earlier in a players career. We found that hits for batters and strikeouts for pitchers were the most positively weighted features in the models which made baseball sense. However, the weights for many features were not intuitive such as home runs, on-base percentage, batting average and games played. The batting model after adding manual features did not perform well as the model did not receive the new percentage features well and weighted all of them against what baseball enthusiasts would, except for slugging percentage which was weighted positively as intuition would expect.

Future studies can explore the data in a time series fashion. This study was conducted in a way where a player was classified as an All-Star even in years when they did not make the team. However, since they made at least one appearance, they were labelled as an All-Star for their whole career. This was done to try to find differences between All-Stars and other MLB players even when a certain player had a down year. Time series data will allow for teams to tell which individual stats influences All-Star voting the most.

Another path to explore could be voting practices by fanbases. Does one team's players receive significantly more votes consistently over a stretch of seasons? This analysis could also potentially be used to quantify fan loyalty, an important factor for an organization to consider when they market their team. Having All-Star talent on a team is crucial for success in the MLB and should have the analysis and attention of baseball front offices for the foreseeable future.

References

[1] What is logistic regression.

[2] Runs created, 2014.

[3] Craig A Depken II and Jon M Ford. Customer-based discrimination against major league baseball players: Additional evidence from all-star ballots. *The Journal of Socio-Economics*, 35(6):1061–1077, 2006.

[4] Bryce Farrell. Machine learning algorithm for predicting major league baseball team wins. 2019.

[5] Alexander Gow. Using machine learning to predict mlb success based on milb performance. 2019.

[6] Micah Melling. Using machine learning to predict baseball hall of famers, 2017.

[7] Cassidy Mickelson-Carter. The worth of an mlb all-star: Are mlb all-star players the key to wins, the playoffs, and the world series?

[8] Brian M Mills and Steven Salaga. Using tree ensembles to analyze national baseball hall of fame voting patterns: an application to discrimination in bbwaa voting. *Journal of Quantitative Analysis in Sports*, 7(4), 2011.

[9] Roger Pharr. Predicting mlb game outcomes with machine learning, 2019.

[10] Travis Sawchik. *Big data baseball: Math, miracles, and the end of a 20-year losing streak*. Macmillan, 2015.

[11] Christopher Watkins. Novel statistical and machine learning methods for the forecasting and analysis of major league baseball player performance. 2020.

[12] Mitchell T Woltring, Jim K Rost, and Colby B Jubenville. Examining perceptions of baseball's eras: A statistical comparison. *Sport Journal*, 2018.