“Predicting Major League Baseball Player Statistics”

Corey Yesavage

Coreycyesavage@lewisu.edu

DATA-59000-001-SU21

Data Science Project

Lewis University

ABSTRACT

Predictive analytics has become a major player in professional sports. While most sports are successfully using analytics, the sport that has seen the greatest impact has been baseball. Baseball has always been a sport that revolved around statistics, so it was only a matter of time before predictive analytics took over. In this project, we harness the power of machine learning to predict player statistics for the Major League Baseball team, the San Diego Padres. While leagues, teams, and outside organizations all have their own predictive modeling algorithms, this study uses linear regression, polynomial regression, support vector regression, and random forest regression to predict player statistics for the 2021 season. The predicted statistics were then fed into an artificial neural network that predicted the chances of the Padres making the MLB playoffs.

1. INTRODUCTION

Major League Baseball is one of the few American professional sports leagues that does not have a salary cap. That means that teams can spend as much money as they choose on players. While this does bring some advantages to the sport, it also brings it share of disadvantages. One major disadvantage is the difference in financial means between teams. Big market teams such as the New York Yankees or Los Angeles Dodgers can spend up to 3-5 times as much on players than small market teams such as the Oakland Athletics or Baltimore Orioles [1]. This puts small market teams at a major disadvantage before the season even starts. This financial discrepancy eventually led to a change in how teams would go about and build their rosters. Most teams had to start taking an in-depth look at the numbers and try to find statistical bargains.

One of the leaders of the analytical movement in baseball was the Oakland Athletics general manager, Billy Beane. Billy put it this way, “It's all about evaluating skills and putting a price on them. Thirty years ago, stockbrokers used to buy stock strictly by feel. Let's put it this way: Anyone in the game with a 401-K has a choice. They can choose a fund manager who manages their retirement by gut instinct, or one who chooses by research and analysis. I know which way I'd choose. [2]". Billy, and people like Billy, started to turn baseball into a game of numbers. The analytical movement that started out as foolish talk has turned into the new norm. While the extent does vary, every team in Major League Baseball is currently using analytics in some way.

Many facets of baseball use predictive analytics, but the most important must be predicting how a player will perform. Accurately predicting a player’s statistics for the upcoming season has many benefits. With the ability to accurately predict how a player will perform, teams are able to make more informed decisions. Informed decisions such as roster building or what type of contracts to hand out. While teams are currently imploring their own models, there is always a need for better accuracy with modeling. This need for accurate predictive modeling is the basis for this project. My intent is to create machine learning models that baseball teams or outside organizations can use to predict future statistics for players.

1. RELATED WORK

While there is related work/research when it comes to using machine learning to predict something baseball related, almost all available work revolves around predicting game outcomes. There is some available work revolving around predicting player statistics, but it is very limited. Most (if not all) Major League Baseball (MLB) teams currently have analytical departments. These departments build and implement their own models for predictive analytics. These models are kept in-house, as they are too valuable to share with other teams or outside organizations. Sports is all about your team having an advantage, so you would not want to do anything that helps your opponent.

Besides sports teams, other organizations such as fantasy sports platforms/websites and sports betting platforms/websites implore their own predictive models. Just like with the professional sports teams, these organizations are built on creating an advantage for their own organization. The greater the accuracy with their modeling, the greater their advantage to make money.

1. PROJECT DESCRIPTION

For this assignment, I used machine learning to predict player statistics for the 2021 San Diego Padres baseball team. While the models can be implemented for any Major League Baseball player, I decided to focus specifically on the Padres players. The data used for this assignment was collected from three different sources. Statistics for individual players were collected from a baseball data website called Fangraphs [3]. Team statistics were collected from the websites Kaggle [4] and Baseball Reference [5].

The statistics that were collected for batting include batting average (AVG), slugging percentage (SLG), on-base percentage (OBP), on-base plus slugging (OPS), games played (G), plate appearances (PA), homeruns (HR), runs batted-in (RBI), runs scored (R), stolen bases (SB), doubles (2B), triples (3B), walks (BB), and strikeouts (SO). The statistics that were collected for pitching include games started (GS), innings-pitched (IP), earned run average (ERA), walks/hits per inning (WHIP), strikeouts per 9-innings (K/9), and walks per 9-innings (BB/9). The statistics that were collected for teams include what year/season (Season), number of games played in that season(G), runs scored (RS), runs against (RA), team on-base percentage (OBP), team slugging percentage (SLG), team batting average (BA), opponents’ on-base percentage (OOBP), opponents’ slugging percentage (OSLG), opponents’ batting average (OAVG) and if that team made the playoffs.

For the players on the San Diego Padres, I used all their previous seasons. The Padres have an assorted roster when it comes to years of experience, ranging from zero to 12 years. The Padres that had zero years of experience had to be left out of the analysis, as my models were not suited to handle without previous statistics. This will be one area of concern to delve in to with future work. For the team statistics, I used statistics for all teams for every season since 1990.

The decision to use these statistics to determine team success was taken from a previous assignment I completed for DATA-51000: Data Mining and Analytics. For that assignment, I used various machine learning methods to determine what statistics a team need to reach to make the playoffs. These methods included random forest, adaboost, KNN, and decision trees. I used the same Kaggle dataset for both assignments. Kaggle does compile the data; however, the data is taken from Baseball Reference website. While that assignment gave some inspiration, this current assignment took a different direction with implementing an artificial neural network to make predictions.

1. METHODOLOGY

For this assignment, the programs Jupyter Notebooks and Excel were used. Excel was used for formatting and preprocessing the data while Jupyter Notebooks was used for the machine learning algorithms and artificial neural network. After compiling the data, the first step was to format the data. Major League Baseball seasons consist of 162 regular season games. With the COVID pandemic, the 2020 season was shortened to 60 regular season games. While the percentage/ratio type categories are not affected by the number of games, the total categories are. The total categories are G, PA, HR, RBI, R, SB, 2B, 3B, BB, SO, GS, and IP. The pitching categories (GS and IP) were multiplied by 2.7 to reflect production over a full 162 game schedule. The assumption with that is that they would stay healthy and stay on their current start and inning pace. While teams might have used pitchers different because of the shortened season, it is impossible to know or account for. The batting categories (G, PA, HR, RBI, R, SB, 2B, 3B, BB, and SO) were adjusted to a per game basis to account for the shortened seasoned. PA, HR, RBI, R, SB, 2B, 3B, BB, and SO were divided by the number of games the player played. The resulting number gives a per game total which is not affected by how many games are played. The G category was divided by the number of games that team played, resulting in the percentage of games played category (%G).

Instead of predicting batting statistics for all pitchers individually, I lumped the pitchers together. This was done to account for half of the pitchers in the league getting little to no bats each season. The American League uses a designated hitter (DH) to bat for pitchers, so anyone that pitches for an American League team has little to zero at bats per season. The San Diego Padres do have several pitchers that have pitched for or are coming over from American League teams. It is impossible to accurately project statistics off just a few at bats. Also, pinch hitters are often used to bat for relief pitchers, making it hard to account for that when predicting number of at bats for relief pitchers. Too many complications come into play when trying to predict batting statistics for individual pitchers. Outside of one MLB pitcher, pitchers perform extremely poor when it comes to batting. When comparing the best and worst hitting pitching staffs, the variance is low. With a low variance and difficulty of predicting individual player statistics, I thought it was best to just use team pitcher batting statistics. I used the pitcher team totals for batting statistics for each of the seasons from 2010 through 2019. The 2020 season was left out because batters did not bat that season as the entire league used a DH to bat for pitchers.

For each player category, I ran the numbers through a linear regression, polynomial regression, support vector regression, and a random forest regression. Feature scaling was used for each algorithm. Since there were so few data points for each player, the test set included all data points. The mean squared error was used to determine the line of best fit. While the percentage/ratio categories provided exact predictions, the total categories needed to be computed over the entire season. This was done by multiplying %G by 162 to get the total number of expected games for that player. After finding the expected number of games you then multiply that by the per game numbers for HR, RBI, R, SB, 2B, 3B, BB, and SO, resulting in the season totals. For the pitcher batting statistics, the per game stats did not need to be calculated since the statistics excluded the 2020 season. Their batting statistics were all calculated based on season totals.

Figure 1: Machine Learning Process Flow

For the ANN, the first step after importing the data was to split the data into training and test sets. After splitting the sets (80% training, 20% test), feature scaling was performed on the data. The next step is to initialize the ANN and add the input layer and hidden layers. For this assignment, I created two hidden layers that consisted of seven nods each. Different amounts of hidden layers and nodes were tested; however, two hidden layers with seven nodes in both layers produced the best results.

The activation function that was used for the hidden layers was the rectified linear unit (ReLU). After the second hidden layer comes the output layer that consists of one unit and a sigmoid activation function. The ANN was then compiled with an Adam optimizer and a binary cross entropy loss function. The network used the default learning rate of 0.001 for the Adam optimizer. The ANN was then trained on the training set and evaluated on accuracy. At 17 epochs the accuracy plateaued out at around 90-92% accuracy.

Figure 2: Artificial Neural Network Process Flow

1. RESULTS AND DISCUSSION

Tables 1-3 show algorithm results for both batting and pitching statistics. While the tables do not show every algorithm for every player, it does show every algorithm for a few players. Linear regression and polynomial regression showed early on that they were very poor predictors of player statistics. Linear regression showed a constant increase, decrease, or stagnate regression line and that is just not the case with player statistics. Players almost always have upward climb at the beginning of their careers followed by a decrease in the later half. Linear regression is not able to account for a player having a rise and fall in production. Polynomial regression tended to be too dramatic with its rise and fall. It often predicted values that would either smash current MLB records or values that were impossible. While both regression lines did show a mean squared error of zero for players with two years or less, they are not used as the predicted values are just not accurate. Supervised Vector Regression (SVR) and Random Forest showed themselves to be the two best algorithms, so those results were provided for every player that had played a minimum of two seasons.

Timeline

Description automatically generated with low confidence

Table 1A: Batting Algorithm Results

A picture containing timeline

Description automatically generated

Table 1B: Batting Algorithm Results

Timeline

Description automatically generated

Table 2: Relief Pitcher Algorithm Results

Table

Description automatically generated

Table 3: Starting Pitcher Algorithm Results

When looking over the results, there are a few areas that need addressing. The first and most important is making sure that there are the correct number of games played by batters (1458), correct number of games started by pitchers (162), and the correct number of innings pitched by pitchers (1,458). Seven of the Padres’ games this year will be played in an American League stadium, so for seven games the pitcher will not be batting, and a designated hitter will be. That situation is the reason that pitchers are only batting in 155 games. Predictions for these three categories do not add up and need to be adjusted.

The number of games for batters adds up to 1,483, leaving me 25 games too many. My models were not able to fix this issue, so I had to adjust the number of games by hand. I reduced the number of games played for Jurickson Profar from 147 games to 122 games. Jurickson is a versatile player that backs up players at several positions. While Jurickson does play a lot, his new roll does have him playing as a backup, so I feel comfortable reducing his number of games. Since batting statistics were predicted on a per game basis, adjusting his season totals were simple.

The model predicted the number of pitching games starts at 148 and innings pitched at 1,255. These predictions leave the Padres at 14 games and 203 innings too short. Just like with the batting, the best option to resolve this problem is manually adjusting games started and innings pitched. Both original and adjusted pitching predictions are given in Table 4.

Being the best pitcher on the team, I needed more than the 27 predicted starts for Yu Darvish, so I upped his starts to 29 games. Two of Yu’s previous seasons were cut short by injury, which slightly skews his predicted number of starts for 2021, so expecting two extra starts is very reasonable. Joe Musgrove was signed in the off season to be the third starter in the rotation. His best regression model had him starting 24 games this coming season. Joe’s GS and IP were continuing to climb as he was transitioning from a relief pitcher to a full-blown starter over the last few years. Joe’s 2020 season was affected because of injuries, so expecting a fully healthy Joe to pitch more than 24 starts is almost a guarantee. As a third man in a rotation, barring injury, you are always expected to complete more than 24 starts. Joe’s expected GS was increased from 24 to 29 games. In his two seasons, Chris Paddack has shown himself to be a workhorse. With his young age and strong arm, his expected GS was increased from 30 to 32 games. Dinelson Lamet has also been a victim of one injury short season effecting his predicted starts for 2021. Without the 2019 season being affected by injury, Dinelson Lamet would be over 30 expected GS for the 2021 season. With his continued growth and 2020 bounce back from injury, I have increased his expected GS from 25 to 30. Extra innings were added to each of the starters to complete the required number of innings for the season. While manually making these adjustments are not ideal, they are also the best solution currently. Additional modeling can be performed to account for the injuries that have impacted the model. This solution will continue to be researched and developed in future work.

Table

Description automatically generated

Table 4: Original Predicted GS and IP compared to Adjusted GS and IP for starting pitching staff

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **G** | **PA** | **R** | **2B** | **3B** | **HR** | **RBI** | **SB** | **BB** | **SO** | **AVG** | **OBP** | **SLG** | **OPS** |
| Fernando Tatis Jr. | 154 | 676 | 121 | 26 | 8 | 42 | 107 | 29 | 63 | 180 | 0.297 | 0.372 | 0.580 | 0.953 |
| Manny Machado | 154 | 655 | 92 | 26 | 3 | 35 | 100 | 9 | 63 | 112 | 0.276 | 0.346 | 0.510 | 0.856 |
| Trent Grisham | 154 | 606 | 91 | 20 | 7 | 22 | 70 | 15 | 71 | 156 | 0.241 | 0.340 | 0.433 | 0.774 |
| Eric Hosmer | 120 | 494 | 67 | 19 | 0 | 25 | 102 | 8 | 29 | 98 | 0.280 | 0.326 | 0.489 | 0.816 |
| Jake Cronenworth | 146 | 519 | 70 | 41 | 8 | 11 | 54 | 8 | 49 | 81 | 0.285 | 0.354 | 0.477 | 0.831 |
| Victor Caratini | 98 | 274 | 22 | 15 | 0 | 3 | 30 | 0 | 22 | 62 | 0.243 | 0.330 | 0.337 | 0.668 |
| Austin Nola | 105 | 379 | 51 | 18 | 2 | 14 | 51 | 1 | 35 | 79 | 0.271 | 0.348 | 0.463 | 0.811 |
| Wil Myers | 138 | 519 | 72 | 32 | 3 | 28 | 77 | 11 | 46 | 146 | 0.266 | 0.336 | 0.518 | 0.854 |
| Tommy Pham | 112 | 469 | 55 | 13 | 1 | 13 | 47 | 20 | 57 | 99 | 0.236 | 0.335 | 0.369 | 0.703 |
| Jurickson Profar | 122 | 450 | 61 | 17 | 1 | 16 | 57 | 12 | 37 | 63 | 0.258 | 0.330 | 0.426 | 0.755 |
| Combined Pitchers | 155 | 288 | 10 | 3 | 0 | 1 | 10 | 1 | 9 | 132 | 0.104 | 0.136 | 0.122 | 0.258 |
| Team Total | 1458 | 5329 | 712 | 230 | 33 | 210 | 705 | 114 | 481 | 1208 | 0.26 | 0.332 | 0.453 | 0.785 |

Table 5: Final Batting Projections

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | GS | IP | ERA | WHIP | AVG | K/9 | BB/9 |
| Blake Snell | 29 | 160 | 3.32 | 1.19 | 0.224 | 11.5 | 3.26 |
| Yu Darvish | 29 | 200 | 3.19 | 1.10 | 0.215 | 11.2 | 2.62 |
| Joe Musgrove | 29 | 200 | 4.00 | 1.23 | 0.231 | 11.2 | 3.13 |
| Chris Paddack | 32 | 180 | 4.04 | 1.10 | 0.231 | 9.3 | 1.9 |
| Dinelson Lamet | 30 | 172 | 2.88 | 1.02 | 0.185 | 11.9 | 3.38 |
| Adrian Morejon | 6 | 30 | 7.35 | 1.74 | 0.325 | 10.9 | 2.61 |
| Drew Pomeranz | 6 | 66 | 2.47 | 1.14 | 0.178 | 13.3 | 4.52 |
| Emilio Pagan | 0 | 58 | 4.27 | 1.04 | 0.191 | 9.7 | 3.39 |
| Austin Adams | 1 | 12 | 3.99 | 1.51 | 0.203 | 14.1 | 6.75 |
| Pierce Johnson | 0 | 50 | 3.84 | 1.27 | 0.219 | 10.3 | 4.24 |
| Miguel Diaz | 0 | 18 | 6.89 | 1.59 | 0.291 | 7.3 | 3.98 |
| Mark Melancon | 0 | 59 | 2.99 | 1.32 | 0.263 | 6.4 | 2.75 |
| Craig Stammen | 0 | 71 | 4.64 | 1.23 | 0.260 | 7.9 | 1.59 |
| Nabil Crismatt | 0 | 66 | 3.24 | 0.84 | 0.200 | 8.6 | 1.08 |
| Tim Hill | 0 | 46 | 4.40 | 1.27 | 0.240 | 9.2 | 2.93 |
| Nick Ramirez | 0 | 53 | 5.01 | 1.26 | 0.219 | 8.8 | 3.66 |
| Keone Kela | 0 | 17 | 3.66 | 1.61 | 0.273 | 12.3 | 4.03 |
| Team Total | 162 | 1458 | 3.73 | 1.17 | 0.222 | 10.4 | 2.9 |

Table 6: Final Pitching Projections

Jake Cronenworth and Nabil Crismatt had only played one previous season. With only one datapoint per category, my models are unable to provide any type of regression. Without any type of regression line, predictions are impossible. Cronenworth’s 2020 statistics were replicated into a full 162 game season and those numbers were used for his 2021 projection. Nabil Crismatt’s number of innings pitched were increased as he has more experience, but his 2020 ratios were replicated for his 2021 numbers.

Tables 7 and 8 compare some of my predicted statistics with some of the leading expert websites in the industry. While teams will not divulge their predictions for players, fantasy websites and sports statistic websites do put out their predictions. Their results have been widely used and accepted as expert analysis. Even though expert analysis for predicting player statistics is wrong all the time, it should still be used as a measuring gauge to see if my models are predicting similar results.

The websites that I used to compare my results are Steamer Projections [6], ESPN [7], and Roto Grinders [8]. While there are some differences, my batting predictions match up well compared to these expert websites. My pitching predictions do have some similarities and some differences. The biggest difference is with my un-adjusted predictions for GS and IP, they tend to underpredict. Pitchers often have their workload progressed as they gain experience or have their roles changed based on team needs, so GS and IP tend to be categories that are hard to predict using only previous career stats. The expert websites tend to have a bigger variance with ERA predictions, but my model predictions tend to fall within that variance. WHIP, K/9, and BB/9 are very comparable between the predictions.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Fernando Tatis Jr.** | **PA** | **R** | **2B** | **3B** | **HR** | **RBI** | **SB** | **BB** | **SO** | **AVG** | **OBP** | **SLG** | **OPS** |
| Mine | 676 | 121 | 26 | 8 | 42 | 107 | 29 | 63 | 180 | 0.297 | 0.372 | 0.580 | 0.953 |
| ESPN | 644 | 119 | N/A | N/A | 36 | 96 | 29 | 59 | 176 | 0.284 | 0.357 | 0.538 | 0.896 |
| STEAMER | 677 | 113 | 29 | 5 | 39 | 98 | 26 | 68 | 161 | 0.286 | 0.365 | 0.551 | 0.916 |
| RotoGrinders | 675 | 105 | 32 | 5 | 37 | 102 | 24 | 64 | 191 | 0.272 | 0.349 | 0.530 | 0.879 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Manny Machado** | **PA** | **R** | **2B** | **3B** | **HR** | **RBI** | **SB** | **BB** | **SO** | **AVG** | **OBP** | **SLG** | **OPS** |
| Mine | 655 | 92 | 26 | 3 | 35 | 100 | 9 | 63 | 112 | 0.276 | 0.346 | 0.510 | 0.856 |
| ESPN | 664 | 93 | N/A | N/A | 34 | 101 | 11 | 68 | 113 | 0.280 | 0.354 | 0.508 | 0.862 |
| STEAMER | 661 | 97 | 28 | 2 | 36 | 104 | 9 | 68 | 116 | 0.273 | 0.350 | 0.515 | 0.865 |
| RotoGrinders | 689 | 93 | 30 | 2 | 39 | 93 | 9 | 68 | 121 | 0.272 | 0.345 | 0.516 | 0.862 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Trent Grisham** | **PA** | **R** | **2B** | **3B** | **HR** | **RBI** | **SB** | **BB** | **SO** | **AVG** | **OBP** | **SLG** | **OPS** |
| Mine | 606 | 91 | 20 | 7 | 22 | 70 | 15 | 71 | 156 | 0.241 | 0.340 | 0.433 | 0.774 |
| ESPN | 640 | 95 | N/A | N/A | 25 | 78 | 18 | 90 | 141 | 0.260 | 0.368 | 0.471 | 0.839 |
| STEAMER | 604 | 85 | 23 | 4 | 22 | 69 | 14 | 75 | 134 | 0.249 | 0.349 | 0.441 | 0.790 |
| RotoGrinders | 610 | 79 | 26 | 4 | 22 | 72 | 12 | 72 | 153 | 0.240 | 0.336 | 0.429 | 0.766 |

Table 7: Comparing projections for Fernando Tatis Jr., Manny Machado, and Trent Grisham

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Blake Snell** | **GS** | **IP** | **ERA** | **WHIP** | **AVG** | **K/9** | **BB/9** |
| Mine | 29 | 134 | 3.32 | 1.19 | 0.224 | 11.5 | 3.3 |
| ESPN | 28 | 166 | 3.47 | 1.25 | N/A | 10.8 | N/A |
| STEAMER | 28 | 151 | 3.35 | 1.20 | N/A | 11.0 | 3.4 |
| RotoGrinders | 28 | 153 | 3.53 | 1.18 | N/A | 11.4 | 3.2 |
|  |  |  |  |  |  |  |  |
| **Yu Darvish** | **GS** | **IP** | **ERA** | **WHIP** | **AVG** | **K/9** | **BB/9** |
| Mine | 27 | 164 | 3.19 | 1.10 | 0.215 | 11.2 | 2.6 |
| ESPN | 31 | 189 | 2.90 | 1.07 | N/A | 11.0 | N/A |
| STEAMER | 31 | 189 | 3.42 | 1.12 | N/A | 11.2 | 2.7 |
| RotoGrinders | 31 | 176 | 3.57 | 1.12 | N/A | 11.3 | 2.9 |
|  |  |  |  |  |  |  |  |
| **Joe Musgrove** | **GS** | **IP** | **ERA** | **WHIP** | **AVG** | **K/9** | **BB/9** |
| Mine | 24 | 120 | 4.00 | 1.23 | 0.231 | 11.2 | 3.1 |
| ESPN | 28 | 148 | 3.83 | 1.21 | N/A | 9.5 | N/A |
| STEAMER | 29 | 169 | 3.85 | 1.22 | N/A | 9.0 | 2.6 |
| RotoGrinders | 29 | 162 | 3.75 | 1.16 | N/A | 9.4 | 2.4 |

Table 8: Comparing projections for Blake Snell, Yu Darvish, and Joe Musgrove

After all the player statistics were predicted, the data was put into an artificial neural network to predict if this team would make the playoffs. The categories that were used for the ANN were runs scored, runs against, team batting average, batting average against, team on-base percentage, on-base percentage against, team slugging percentage, and slugging percentage against. After adjusting the network, the best model produced a 92% accuracy. Figure 3 gives the confusion matrix results for the network. Along with the 92% accuracy, the model gives my 2021 San Diego Padres a 97% chance to make the playoffs.

Text

Description automatically generated with medium confidence

Figure 3: ANN Confusion Matrix

1. FUTURE WORK

While I am pleased with the results of this project, there are a few areas of this project that will need further attention. The first area that is going to need attention is developing a model that will work for first- and second-year players. My models were unable to properly handle these players. This is not an ideal situation as teams want to know what to expect from first- and second- year players. One possible way to address this would build a model that intakes minor league statistics and compares that to how other first- and second- years performed with similar minor league statistics. This type of modeling will take extensive work but might just prove to be very valuable.

Another area that needs further attention is the issue with predicting the correct amount of G, GS, and IP. For this project, I had to slightly alter all three categories as my model predicted too many G for batters and not enough GS and IP for pitchers. While career statistics play a major factor in future predictions, a player’s role with the team should come into play. Players change roles throughout their careers, especially when they change teams. Being able to account for a player’s role and place some type of value on that would benefit the modeling. Accounting for past injuries should also help this situation. Statistics can be skewed when a player has had a season cut short because of injuries. A model that can address injuries without letting it skew results will be very beneficial.

These issues along with implementing some ensembles methods might prove to be very valuable. These areas will be further researched and addressed to help create the most accurate models possible.

1. CONCLUSION

Analytics has changed the game of baseball forever. It has changed everything from roster building to on-field strategy to financial decisions. While some like the change and others hate it, everyone can see that it is not slowing down. The game is no longer a game won in the dirt; it is a game won with algorithms. As this change continues, it becomes a necessity to continue to improve modeling. Machine learning and deep learning will make this possible.

The purpose of this project was to test the predictive power of machine learning and artificial neural networks. I wanted to see if machine learning could build a model that accurately predicts baseball player statistics for the 2021 season. After that I wanted to see if those statistics translated into playoff baseball for the San Diego Padres. These models fulfilled their purpose and provided very accurate results. They proved that the power of machine learning and artificial neural networks is tremendous and only continuing to grow.

VIII. REFERENCES

[1] “MLB 2021 Payroll Tracker,” *Spotrac.com*. [Online]. Available: https://www.spotrac.com/mlb/payroll/. [Accessed: 22-Jun-2021].

[2] H. Bryant, “The Revolutionary,” *ESPN*, 26-Jul-2009. [Online]. Available: https://www.espn.com/espn/otl/columns/story?columnist=bryant\_howard&id=4357166. [Accessed: 22-Jun-2021].

[3] “FanGraphs Baseball: Baseball Statistics and Analysis,” *FanGraphs Baseball | Baseball Statistics and Analysis*. [Online]. Available: https://www.fangraphs.com/. [Accessed: 24-May-2021].

[4] WesDuckett, “Moneyball,” *Kaggle*, 19-Oct-2017. [Online]. Available: https://www.kaggle.com/wduckett/moneyball-mlb-stats-19622012. [Accessed: 18-Jun-2021].

[5] “MLB Stats, Scores, History, & Records,” *Baseball*. [Online]. Available: https://www.baseball-reference.com/. [Accessed: 24-May-2021].

[6]"Steamer Projections - Home", *Steamerprojections.com*, 2021. [Online]. Available: http://www.steamerprojections.com. [Accessed: 01- Jun- 2021].

[7]"2021 Fantasy Baseball Ranking & Projections - ESPN", *ESPN.com*, 2021. [Online]. Available: https://fantasy.espn.com/baseball/players/projections. [Accessed: 01- Jun- 2021].

[8]"RotoGrinders: The Daily Fantasy Sports Authority", *Rotogrinders.com*, 2021. [Online]. Available: https://rotogrinders.com. [Accessed: 01- Jun- 2021].