Austin Amann Vaughan 10101667 Michael Spiering

CMPE452  Final Report

MLB Statistic prediction using neural networks

# Problem Statement

The purpose of this project is to use a backpropagation neural network to predict offensive statistics for professional baseball players in the MLB. Since baseball is rich with statistics only batting average, home runs, runs batted it, on base percentage and hits will be looked at. This is because these statistics were determined to be the most valuable statistics when summarizing a players offensive performance. Baseball was chosen because analytics are a popular tool that teams use in order to improve the performance of their players and teams. Statistics have been kept in the MLB since the game was first played professionally in the late 1800s. Since there is such a large amount of data only the previous 5 years of statistics will be used in order to minimize storage and execution time when training the neural network. K means clustering and K nearest neighbour classification will be used in order to provide context to the initial and predicted data.

# Research

## Michael’s Research

Jim Albert’s 2016 paper “Improved Component Predictions of Batting Measures” breaks down standard stats for analyzing players such as batting average and on-base percentage into components such as strikeout rates and home runs. Though not using Neural Networks, his approach and results were comparable to our 5 input 1 output backpropagation networks. Albert looked for correlation between components, as well as applied various techniques of randomization and variance to predict a players offensive output. Our backpropagation networks did the same. His conclusions showed that many stats are based significantly on year to year luck, and its near impossible to predict when players are going to be outliers while there tends to be much lower variance in middle of the pack players. In the same way, we saw a very gaussian average in our output. The trained networks could almost never predict accurately players on the high and low end of the spectrum. Both our group and Albert saw in results that career trajectories are near impossible to predict because of such dramatic season to season change.

Sarah Reid Bailey’s paper “Forecasting Batting Averages in MLB” goes very in depth on all the factors that could predict a player’s chance to get a hit on each at bat. Bailey went very in depth into things such as the opposing pitcher, average ball trajectory speed/angle, and the type of hits a player gets. After combining her results with other models, the variation that she concluded ended up quite low (as low as 0.05). Though just as we did, and almost all models have found, the league average always looks gaussian around about a 0.260 batting average. This paper showed a lot of things that could be considered in a future expansion of our implementation. Bailey also found, in the same way as Jim Albert, that players on the high and low end of batting average ended up having very large variances. Her section on “Players with Large Absolute Errors” showed interesting insight into the types of players that should be “ignored” to increase accuracy of a model. Applying her findings to the end of our solution would be a great next step to try for a better model.

## Austin’s Research

In order to gain a better understanding of the project, background research was performed. Two research papers were analyzed in order to broaden understanding of statistical analysis in professional baseball.

The first research paper, Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network, implemented and ANN to determine if a player would be elected to the baseball hall of fame based on their career statistics. This paper was chosen to see how other teams approach the prediction problem with the use of a neural network. Although this paper was published in 2008 the approach the research team took should still be applicable today. The hall of fame is a collection of the all time best players in the MLB. Players are elected yearly after they have retired from the sport and only have a limited number of years of eligibility before they are removed from the ballot. This paper is limited to analyzing position players, excluding pitchers. Each players offensive and defensive statistics are used in order to determine if they are hall of fame caliber. The model also considers if the player has won any individual performance awards or World Series championships. The model analyzed the current members of the hall of fame in order to gain understanding into what constitutes a hall of fame player. Similarly, to this project the researchers provided the network with a set number of offensive and defensive statistics that were determined to be the most important for predicting the desired output. As expected, the research team provided their model with a larger variety of statistics. This information could help improve this project if the scope of the project was ever expanded in the future. The research team’s data preprocessing method was also similar to this project. The team applied a series of conditions including number of games played and player position in order to provide the most accurate input data for their neural network. The researches also applied K means clustering in order to cluster their data before sending it to their ANN. Once the data was processed the research team explored a variety of approaches with their ANN. They analyzed the performance when different activation functions and learning methods were applied. The final ANN used 2 hidden layers as the researchers determined that maximized the prediction accuracy for their dataset. Their model was able to achieve a 95% prediction accuracy for hall of fame induction. The information gathered from this research paper will assist the team in determining the best approach to take when solving the prediction problem laid out in the introduction.

The 2nd research paper, In-Season Prediction of Batting Averages: A Field Test of Empirical Bayes and Bayes Methodologies, analyzes techniques to predict a player’s batting average over the course of a season. The system analyzes the first 3 months of a batter’s season in order to establish a performance metric that will allow the system to predict their overall batting performance at the end of the season. An empirical and hierarchical Bayes prediction model were chosen by the research team in order to achieve their prediction goals.

Similarly to the last paper, this system focuses only on position players in order to predict their batting performance at the end of the season. The research team pulled full seasons of batting data from similar sources as this project. Preprocessing was performed in order to isolate the players of interest as well as their game by game statistics in the 1st 3 months of a season. Once the dataset was processed it was passed onto the Bayes prediction model. The research team also analyzed passing more and less data to the prediction system in order to see if the results were affected. It was determined that players often exhibit streaky performance that can skew the output of the prediction model. Using smaller chunks of the season when predicting the final batting average proved to improve prediction accuracy. The prediction model achieved accuracy between 90 and 95% depending on the amount of input data provided. This paper provided insight into the amount of data to provide to a prediction model in order to achieve higher prediction accuracy. It also solidified our assertion that batting average is an extremely important statistic that determines a players overall performance.

# Dataset

The statistics used in the project will be sourced directly from the MLB. Specifically, the MLB Advanced Media API will be used. The MLB provides this API in order to simplify the data acquisition process for anyone interested in analyzing yearly player statistics. This API was chosen because it was easy to access and pull all the required data into a .csv file in order to be used in this project. Other sources of data that were considered include Baseball-Reference which is a statistics website that houses similar data to the official MLB API. No simple way to access the data was provided by Baseball-Reference and manually acquiring all the required statistics proved to be too time consuming.

In order to minimize outliers only players that had a minimum of 50 at bats were used. The data was then modified to only include players that appeared in every season of interest. This ensured consistent data and would allow for specific players statistics to be projected for the following season. Once the selected statistics were acquired from the MLB, the data was plotted in Matlab. Various statistics were plotted against eachother. A sample of these plots can be found in Figure 1 - Initial Data and Figure 2 - Initial Data. Plotting the data in Matlab allowed for an opportunity to identify any obvious trends. It also allowed for any missed outliers to be identified and possibly removed.

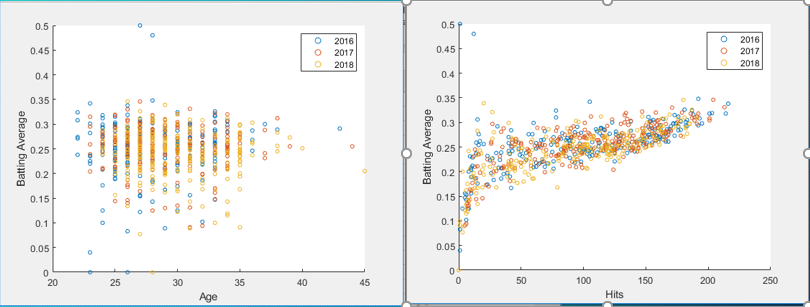


Figure 1 - Initial Data

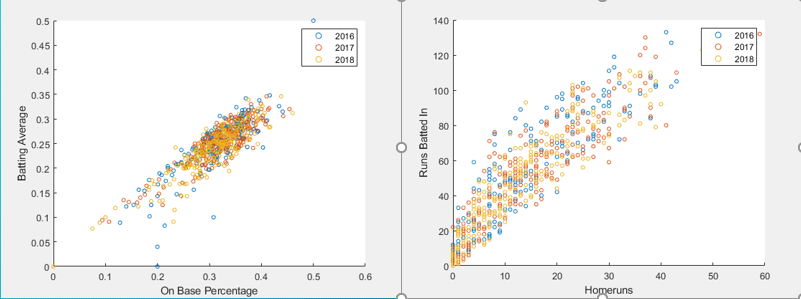


Figure 2 - Initial Data

Once an initial understanding of the data was achieved the data was passed to a K means clustering program. The players were clustered into four categories, below average, average or replacement level, above average and elite. These clusters are displayed from left to right in Figure 3 - Clustered Data. Clustering the data provided context to an individual players performance and would allow for predictions to be categorized.

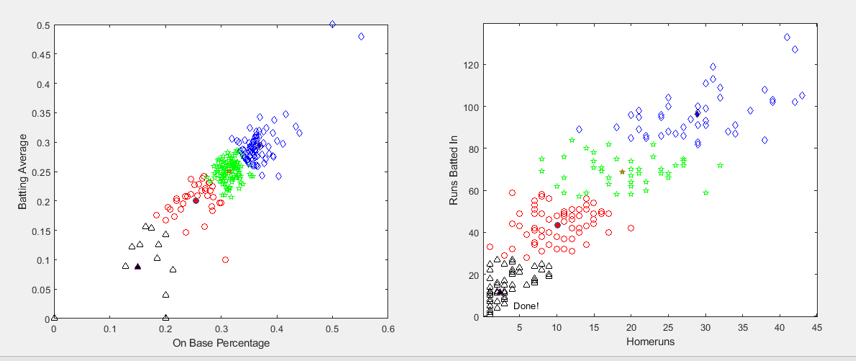


Figure 3 - Clustered Data

After the data was clustered and displayed, the preprocessing of the dataset was complete. The dataset could then be passed into the backpropagation neural network for training and prediction.

# Solution

## Michael’s Approach

## 

Figure 4 Design Philosophy Diagram

The overall design plan shown above shows the entire process I took from getting the initial list of active players to final outputs. All code was written in python, all data stored in csv files. As mentioned in the Dataset section, the first step was acquiring a master list of all active players with a unique mlb\_id needed to pull from the MLB API.

|  |  |
| --- | --- |
| mlb\_id | mlb\_name |
| 592091 | A.J. Achter |
| 595918 | A.J. Cole |
| 454560 | A.J. Ellis |
| 456167 | A.J. Griffin |
| 543362 | A.J. Jimenez |
| 621345 | A.J. Minter |
| 571963 | A.J. Morris |
| 572041 | A.J. Pollock |
| 519263 | A.J. Schugel |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Figure 5 Example of initial entries

These Id’s were then used to pull from the API the stats we wanted year by year with enough at bats to be considered relevant (we chose 50). The data was then sorted and balanced into yearly datasheets with matching players who played in every year.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Id | rbi | hr | obp | avg | hits | age |
| 572041 | 76 | 20 | 0.367 | 0.315 | 192 | 1987 |
| 571437 | 22 | 5 | 0.338 | 0.241 | 33 | 1991 |
| 543305 | 33 | 11 | 0.323 | 0.256 | 90 | 1989 |
| 594809 | 56 | 14 | 0.361 | 0.287 | 175 | 1988 |
| 430945 | 82 | 27 | 0.308 | 0.269 | 147 | 1985 |
| 608365 | 54 | 13 | 0.307 | 0.242 | 115 | 1994 |
| 134181 | 83 | 18 | 0.334 | 0.287 | 163 | 1979 |

Figure 6 Example of intermediate data, for 2015

The data now being organized, the next step was to design ANN models. As a group we had decided that feedforward networks with backpropagation error correction would be a good solution, and that is what I mainly focused on. We were looking for correlation between any stats from a previous year to the next, and I had to try many different input/output and network connection combinations. The general goal was to find data that correlated closer than just predicting the same result in the next year. First to prove the validity of the network algorithm a very simple net was built. This network had 2 inputs of batting averages, 2 hidden nodes, and an output prediction for the next years data. This first network converged to roughly the average of the two inputs, and a leaguewide gaussian distribution. This was an expected result, confirmed the network was working, and forecasted what was to come.

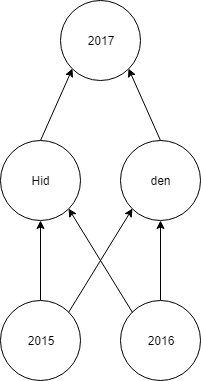


Figure 7 First simple network

The process was now to design increasing complex networks with inputs and outputs shuffled looking for strong correlation. The next move was expanding the first network.

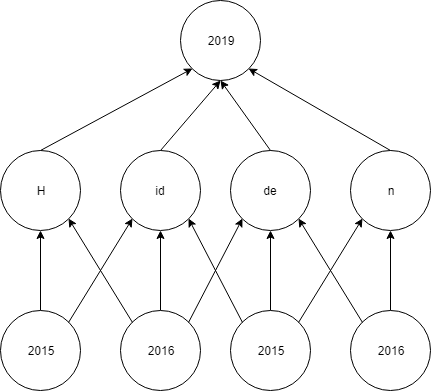


Figure 8 Expanded simple network

This expanded network gave roughly the same results, with slightly more overall accuracy. This network ended up being one of the best overall average predicter getting within 5%. However, the average standard deviation was over 10%. This is because the network converged quite close to the league average of 0.260 and in general, as background research confirmed, year to year variance is in general quite high. This network was used for both batting average and on-base percentage.

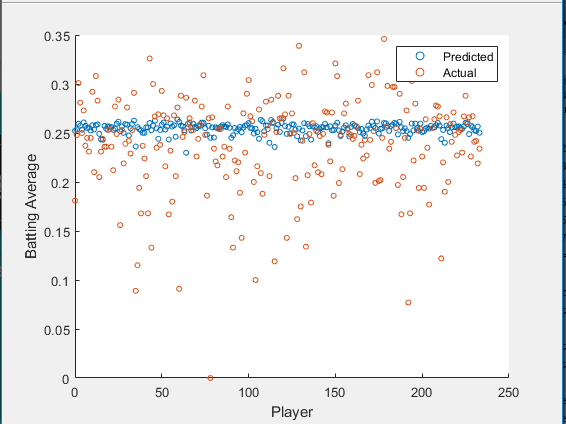


Figure 9b Results for batting average

Many iterations of networks were made attempting to find strong correlation. These ranged from 2-5 inputs and 2-5 outputs., and heavily increasing the weight of one component.

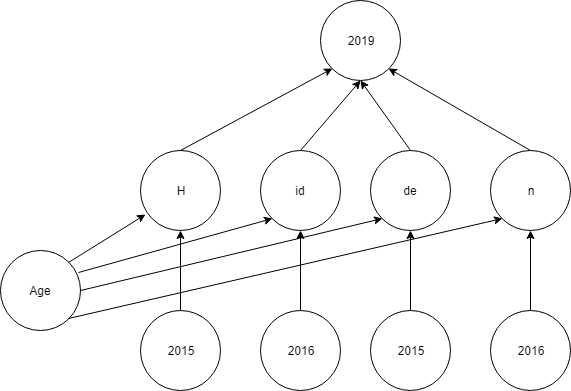


Figure 10 Network to test correlation of one component (in this case age) vs yearly data of another

Dozens of networks were tried, and eventually it was determined that on-base percentage had a higher correlation year to year than any other stat. A 5-input node 1-output network was built, and the results came out accurate enough to be satisfied with. This network took in the data from the previous year for a player and predicted their on-base percentage for the next year.

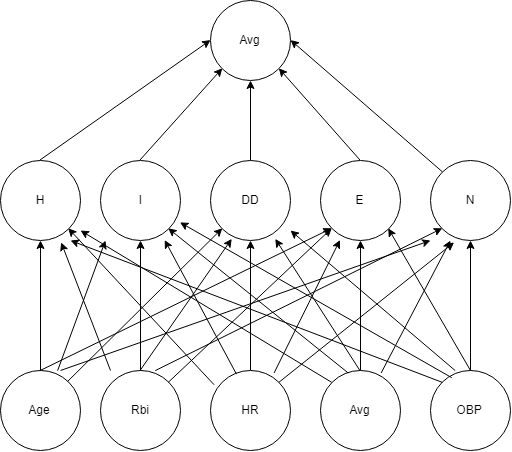


Figure 11 5 input on-base percentage predicting network

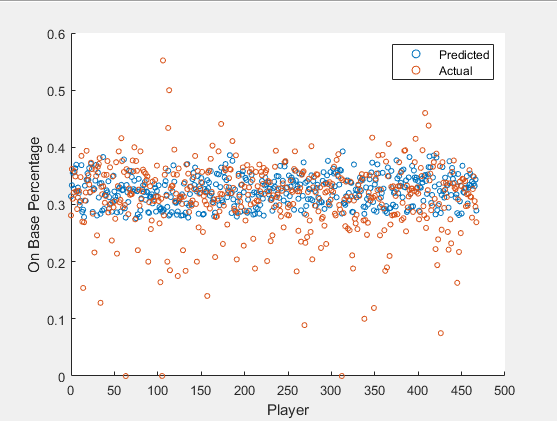


Figure 12 Output of 5-1 network for on-base percentage

## Austin’s Approach

In order to provide context to our predicted statistics K-means clustering was used. This allowed players to be categorized into 4 categories based on their performance. Once the network made a prediction K-nearest neighbour was used in order to classify the predicted data into the proper cluster.

### K-Means Clustering

My focus for this project was on the data preprocessing and classification. K-means clustering was chosen as a method to categorize players performance year to year. Players were categorized as below average, average/replacement level, above average and elite. K-means clustering allows the data to be split into k clusters. This will allow for the number of clusters to be increased if the scope of the project is expanded and more insight into a player’s performance is needed. The cluster centres were initiated in 4 areas of the provided data. The data was then classified into corresponding clusters by computing the Euclidean distance to the various cluster centres. After each pass the cluster centres are updated to be the mean of each cluster. The process is then repeated until the final clusters are found. A sample of clustered data can be viewed below in Figure 13 - K-means clustering output for various input data.

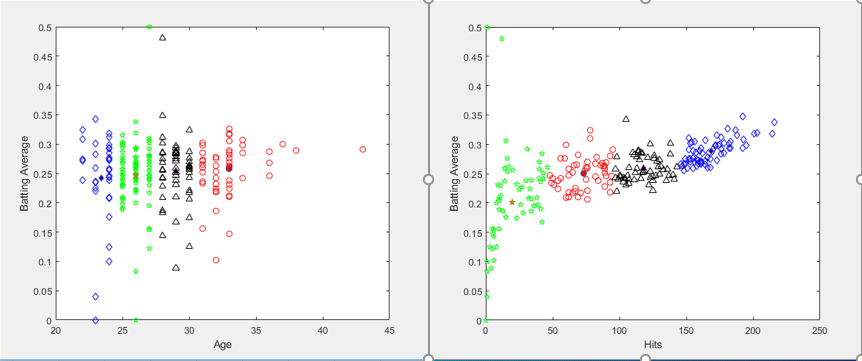


Figure 13 - K-means clustering output for various input data

Clustered data allowed us to further split our data if needed when providing the processed data to the neural network. It will also allow for the predicted data to be classified once the neural networks training and prediction is completed.

### K-Nearest Neighbour Classification

K-Nearest Neighbour was used in order to classify the predicted data from the neural network. Using the clusters determined in the previous section, the predictions were classified into 4 categories. K-Nearest Neighbour works similarly to K-Means Clustering. Euclidean distance is calculated in order to determine the distance between the predicted data and the k closest points. The cluster that has the majority of the k closest points is then chosen and the prediction is classified. The use of classification allows for teams to see how their players performance compares to previous years.

# Summary

As mentioned before, the main goal was to find a network that could predict better than a model that just predicts no change in a player’s stats. Though year to year variance tends to be quite high for individuals overall for the entire league, the variance year to year for on-base percentage was calculated to be 0.042 points for our data set. This 5-1 network ended up having a variance of 0.039 points. Which we felt was about as good as we were going to get with our time and resources available. While running through all the networks, at no point was there a drastically amazing correlation found between components. As research told us, the variance year to year is just simply too high to predict far off the average.

Going forward, I personally feel that the best way we could have improved our results is researching more into current cutting-edge models that analysts are using and optimizing those. It was interesting to go from just basic stats and determine that the networks matched the general averages, but I’m sure better results could come from increasing complex inputs and networks. I also considered creating some PCA’s to parse through many different components before creating my networks which may have given some clear directions to focus. Overall the results were satisfactory. The network performed about as well as intuition would, but absolutely had room for improvement.

# Next Steps

Statistical analysis in professional baseball is an ever-expanding industry. Since the introduction of personal computers in the 1990s statistical analysis techniques have improved in sophistication and scale. Baseball has a large amount of tracked statistics that allow for the game to be analytically analyzed. The neural network implemented in this project performed adequately for the relatively small amount of data we provided it. In order to improve the network’s precision and accuracy the dataset can be expanded to include more years of statistics. This new larger dataset can be analysed in order to identify trends that emerge and fade out over time. The data can then be split into groups of years where the game is statistically similar. The new dataset can then be used to better train the neural network and hopefully improve its overall performance. The clustering of the data can also be improved by adding more clusters. This will allow teams to have a more accurate understanding of the type of player they are analyzing. Possible clusters include, smaller differences in the performance clusters, clustering the players by position and clustering the players by years of experience in the MLB. These new clustering methods will provide more context to the teams when analyzing players. When classifying the predictions provided by the neural network, a more efficient classification method can be used. Decision trees or Naïve Bayes classifiers could be used. Using a more efficient classification method will allow for the size of the dataset to be increased at the users discretion. Since this project only focused on offensive statistics, the solution could be expanded to predict defensive and pitching statistics. Expanding the prediction ability of the network will allow teams the ability to predict and analyze all facets of their team. This will assist the teams in deciding which players to keep on the team, which players to cut and which players they can acquire from other teams in order to improve their team. The expanded network will hopefully assist teams in achieving their goals of winning MLB championships year after year.

# References

Albert, J. (2016). Improved Component Predictions of Batting Measures. *Bowling Green State University*. Retrieved from arxiv.org/pdf/1505.05557.pdf

Bailey, S. R. (2015). Forecasting Batting Averages in MLB. *University of the Pacific*. Retrieved from summit.sfu.ca/system/files/iritems1/17740/etd10438\_SBailey.pdf.

Young, W. A., Holland, W. S., & Weckman, G. R. (2008). Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network. Journal of Quantitative Analysis in Sports, 4(4). doi:10.2202/1559-0410.1131

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics,2*(1), 113-152. doi:10.1214/07-aoas138