**Baseball Player Position Predictor**

By,

Deni Avinash Aalla Balaji,

Rohit Ramna,

Aaqil M Rahman

Master of Professional Studies 2018-2019

Penn State University

**Abstract:**

Sports is one of the biggest entertainments and business in the world. There are billions of money being invested and spent by both the management and fans to build and watch their teams perform. In this case use of data analytics becomes a vital part to deliver a positive result. In this era, Sports teams are said to be having competitive disadvantage if they don’t use data to make decisions. Baseball is one of the biggest sport in North America and it is more popular in United States. Our goal is to find a way in which we can position our fielders in baseball such that we reduce the numbers that is being conceded by the team to the opponents. We will experiment this solution using Artificial Neural Networks.

**Introduction:**

**Aim:**

To Analyze the baseball fielding dataset and identify best suite positions for a player based on his attributes.

**Motivation:**

As sports enthusiasts we were more fascinated on how data analytics works in the field of sports. We took baseball as our sport of interest as it is one the costliest and popular sport in the United States.

**Data:**

**Data Source:** https://www.baseball-reference.com/leagues/MLB/2018-standard-fielding.shtml

We obtained our data from **Baseball-Reference.com** which is a website providing [baseball statistics](https://en.wikipedia.org/wiki/Baseball_statistics) for every player in [Major League Baseball](https://en.wikipedia.org/wiki/Major_League_Baseball) history. The data is derived from sabermetrics which is the hub of baseball science. The site is often used by major media organizations and baseball broadcasters as a source for statistics. It offers a variety of advanced baseball [sabermetrics](https://en.wikipedia.org/wiki/Sabermetrics) in addition to traditional baseball counting stats.

**Label Description:**

PO -- Putouts

A -- Assists

E -- Errors Committed

DP -- Double Plays Turned

Fld% -- **Fielding Percentage**  
(Putouts + Assists) / (Putouts + Assists + Errors)

Rtot -- **Total Zone Total Fielding Runs Above Avg**  
The number of runs above or below average the player was worth based on the number of plays made.  
This number combines the R*tz*, R*dp*, R*of*, R*catch* numbers into a total defensive contribution. See the glossary section for a more complete explanation.  
*Provided by BaseballProjection.com*

Rtot/yr -- **Total Zone Total Fielding Runs Above Avg per 1,200 Inn**  
The number of runs above or below average the fielder was worth per 1,200 Innings (approx 135 games). This number combines the R*tz*, R*dp*, R*of*, R*catch* numbers into a total defensive contribution.  
See the glossary section for a more complete explanation. *Provided by BaseballProjection.com*

Rdrs -- **BIS Defensive Runs Saved Above Avg**  
The number of runs above or below average the player was worth based on the number of plays made. This number combines the R*pm*, R*bdp*, R*bof*, R*bcatch* numbers into a total defensive contribution.  
*Provided by Baseball Info Solutions*

Rdrs/yr -- **BIS Defensive Runs Saved Above Avg per 1,200 Inn**  
The number of runs above or below average the fielder was worth per 1,200 Innings (approx 135 games). This number combines the R*pm*, R*bdp*, R*bof*, R*bcatch* numbers into a total defensive contribution. For pitchers, this is set to 200 Innings. *Provided by Baseball Info Solutions*

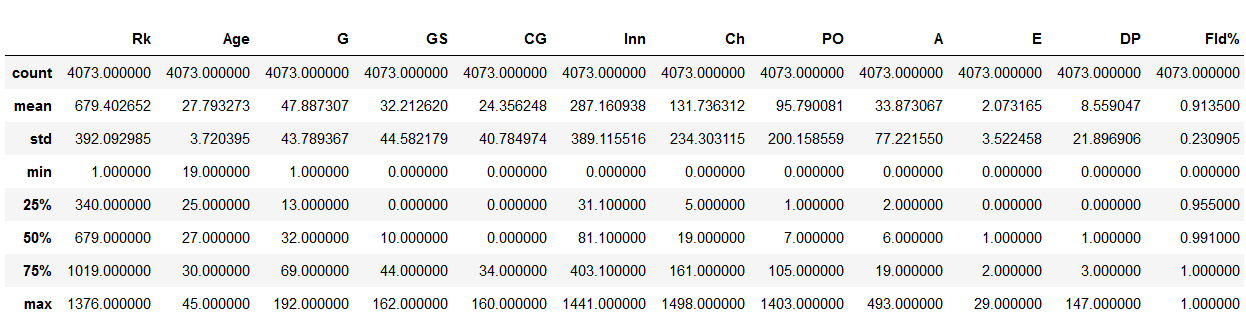
RF/9 -- **Range Factor per 9 Inn**  
9 \* (Putouts + Assists) / Innings Played

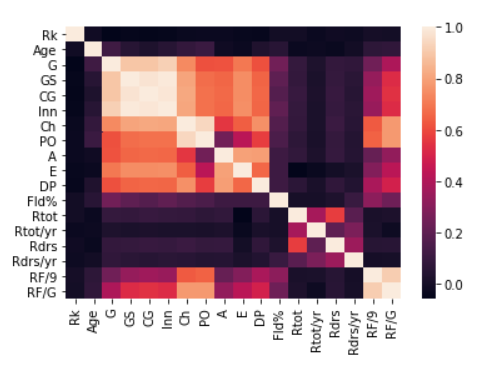
RF/G -- **Range Factor per Game**  
(Putouts + Assists) / Games Played

Pos Summary -- **Positions Played**  
The positions either followed by the games played at that position or in order of games or innings played. For a single season, **\*** indicates they played at least 2/3rds of the team games there Positions after / indicate less than ten games played at those positions. For career, a + sign means more than 300 games at that position and a - sign means less than 30 games.

**Summary Statistics:**

We did some descriptive statistics to get an overview idea of the data we have. We found the quartiles for each attribute, standard deviation, mean and median. We also created a heatmap to identify the correlation between each attribute.

****

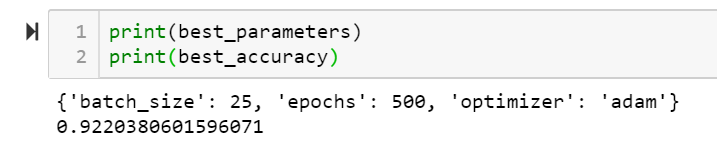


**Data Preprocessing:**

**Hyperparameter Optimization:**

We needed to find the appropriate hyperparameter to be used so we thought of using Grid search algorithm. Grid search algorithm uses all the parameters and creates a combination of models. It calculates the error in each model after running the dataset for each model created and identifies the error in each. It gives us the result of the hyperparameter combination that had least error.

We tired with batch size as 25 and 32, epochs as 100 and 500 and finally Optimizers as Adam and Rmsprop. In the end the result from Gris search algorithm was:

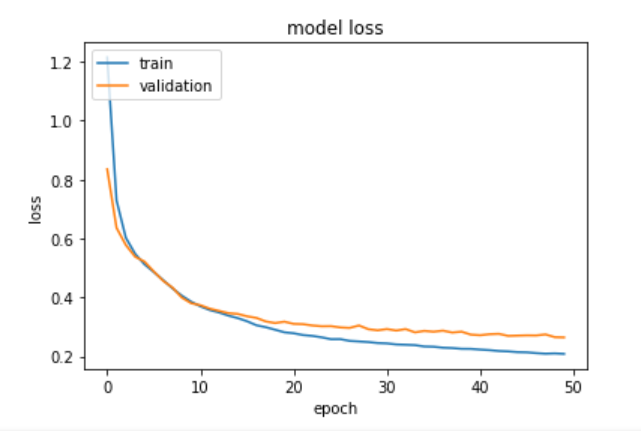


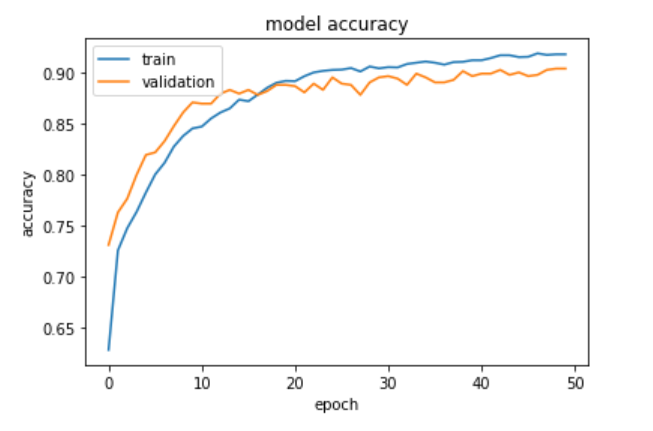
**Neural Network:**

Finally, we created our Neural Network using the following parameters:

* Layers – 3 layers
* Activation functions – Relu and Softmax
* Optimizer – Adam
* Epochs – 500
* Batch size - 25
* Loss – categorical cross entropy

**Results:**

We found that the model worked quite well with both training accuracy and validation accuracy were around 90-95%. The loss was also quite less in both Training and Validation part.



**Conclusion** :

We found that our model with neural network works fine. Given the historical attributes of a player it can be used by the scouting team to choose the right player to buy and help the manager or coach to position the players in the right place to prevent the team from conceding runs and eventually win games.