# Machine Learning Engineer Nanodegree

## Capstone Proposal

John K. Hancock
July 7, 2018

## Proposal

### Domain Background

My interest in Data Science is largely inspired by my interest in sports. My favorite sport, Major League Baseball ("MLB"), has been a pioneer in using advanced analytics to challenge pre-conceived notions of how to evaluate major league talent through the use of innovative statistics. Bill James[1], a statistician and founding member of the Society for American Baseball Research (SABR) pioneered the term Sabrmetrics which uses empirical data and statistics to determine why teams win and lose and the value of a player. Today, most MLB teams use advanced statistics to build a team that they hope can maximize their chances of winning.

This project will use Machine Learning tools and techniques to build a model that can predict the number of wins a team can achieve using the vast array of baseball statistical data. Using standard Supervise Learning algorithms along with Deep Neural Networks, this project will determine how a General Manager of a MLB can build a team that can win the most games.

### Problem Statement

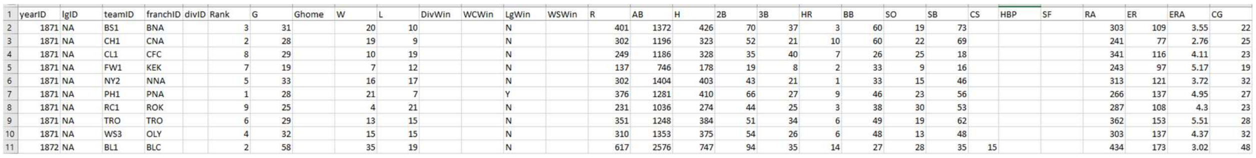
An MLB General Manager of a baseball team is tasked with building a team that can make the playoffs. In order to do that, s/he must win more games than the other teams. The GM assigns his top Data Scientist to build a model that can predict how many wins a team will have based on historical statistical data from winning teams and list the key statistical features that winning team can have. In the end, the GM will use this model to find players that fit these key statistical features. For example, if On-Base Percentage ("OBP") shows that leads to more wins, the GM will look for players that have demonstrated high OBP.

[1] Bill James, https://en.wikipedia.org/wiki/Bill_James

# Datasets and Inputs

Baseball has a long history of meticulous record keeping which laid the groundwork for data exploration and analysis. There are several sources of data for this project. The primary datasets that I will use are the Sean Lanham Baseball Archive and Baseball Reference.com. Sean "founded the website to create a repository for baseball stats and historical information, and that work eventually led to [him] serving as an editor for Total Baseball: The Official Encyclopedia of Major League Baseball and other well-respected sports reference books."[2] Baseball reference contains "baseball statistics from 1871 to the present for major league players, teams, and leagues".[3]

The dataset from the Lanham Baseball archive contains over 2800 records of yearly team data dating back to 1871. There are over 45 features for each team, some categorical, but most are continuous. See Tab A for a description of the features in this dataset.



| 1 | yearID | lgID | teamID | franchID | divID | Rank | G | Ghome | W | L | DivWin | WCWin | LgWin | WSWin | R | AB | H | 2B | 3B | HR | BB | SO | SB | CS | HBP | SF | RA | ER | ERA | CG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1871 | NA | BS1 | BNA |  | 3 | 31 |  | 20 | 10 |  |  | N |  | 401 | 1372 | 426 | 70 | 37 | 3 | 60 | 19 | 73 |  |  |  | 303 | 109 | 3.55 | 22 |
| 3 | 1871 | NA | CH1 | CNA |  | 2 | 28 |  | 19 | 9 |  |  | N |  | 302 | 1196 | 323 | 52 | 21 | 10 | 60 | 22 | 69 |  |  |  | 241 | 77 | 2.76 | 25 |
| 4 | 1871 | NA | CL1 | CFC |  | 8 | 29 |  | 10 | 19 |  |  | N |  | 249 | 1186 | 328 | 35 | 40 | 7 | 26 | 25 | 18 |  |  |  | 341 | 116 | 4.11 | 23 |
| 5 | 1871 | NA | FW1 | KEK |  | 7 | 19 |  | 7 | 12 |  |  | N |  | 137 | 746 | 178 | 19 | 8 | 2 | 33 | 9 | 16 |  |  |  | 243 | 97 | 5.17 | 19 |
| 6 | 1871 | NA | NY2 | NNA |  | 5 | 33 |  | 16 | 17 |  |  | N |  | 302 | 1404 | 403 | 43 | 21 | 1 | 33 | 15 | 46 |  |  |  | 313 | 121 | 3.72 | 32 |
| 7 | 1871 | NA | PH1 | PNA |  | 1 | 28 |  | 21 | 7 |  |  | Y |  | 376 | 1281 | 410 | 66 | 27 | 9 | 46 | 23 | 56 |  |  |  | 266 | 137 | 4.95 | 27 |
| 8 | 1871 | NA | RC1 | ROK |  | 9 | 25 |  | 4 | 21 |  |  | N |  | 231 | 1036 | 274 | 44 | 25 | 3 | 38 | 30 | 53 |  |  |  | 287 | 108 | 4.3 | 23 |
| 9 | 1871 | NA | TRO | TRO |  | 6 | 29 |  | 13 | 15 |  |  | N |  | 351 | 1248 | 384 | 51 | 34 | 6 | 49 | 19 | 62 |  |  |  | 362 | 153 | 5.51 | 28 |
| 10 | 1871 | NA | WS3 | OLY |  | 4 | 32 |  | 15 | 15 |  |  | N |  | 310 | 1353 | 375 | 54 | 26 | 6 | 48 | 13 | 48 |  |  |  | 303 | 137 | 4.37 | 32 |
| 11 | 1872 | NA | BL1 | BLC |  | 2 | 58 |  | 35 | 19 |  |  | N |  | 617 | 2576 | 747 | 94 | 35 | 14 | 27 | 28 | 35 | 15 |  |  | 434 | 173 | 3.02 | 48 |

Baseball Reference.com also compile yearly team statistics. However, it has more advanced statistics such as Runs Allowed per Game and ERA+ which is adjusted to a team's ballpark. A sample pitching csv from 2017 is attached.

[**UPDATE:** For this project, I will be using statistics compiled by the website, FanGrpahs.com. From Wikipedia: "FanGraphs.com is a website run by Fangraphs Inc., located in Arlington, Virginia, and created and owned by David Appelman that provides statistics for every player in Major League Baseball history." (https://en.wikipedia.org/wiki/Fangraphs). FanGraphs compiles basic to highly advanced datasets for every MLB team. I am a member of the site and have access to this data.]

# Solution Statement

To predict wins, the Data Scientist will need to build a mapping function from input variables (historical team statistics for the past ten years) to a continuous output variable (the number of wins). For this step, s/he will use the standard Linear Regression classification algorithm from Sci-kit Learn. Additionally, the Data Scientist will also use Deep Learning Neural Networks to perform Logistic Regression. [4] The simple neural network will map continuous variables inputs (e.g. Home Runs, OBP, Earned Run Averages, Strikeouts, etc.) to a continuous output variable (the number of wins). For this deep neural network, s/he will remove the final sigmoid unit from the network and just return the value from the network of inputs, weights, nodes, and ReLu functions. The end value will be the weighted sum

[2] Read more about Sean Lanham Baseball Archive here: www.seanlahman.com

[3] Baseball reference, www.sports-reference.com/

[4] https://skymind.ai/wiki/logistic-regression

of the outputs from the previous layer. We would still train the network using back propagation to minimize the Mean Squared Error (“MSE”). Finally, the Data Scientist will analyze the results of the predictive models and use unsupervised learning models to cluster current players that have the features that the models say give the team the best chance to win.

## Benchmark Model

The Data Scientist will compare the results from both the linear regression and the deep neural network models will reveal which team statistics are the better predictors of overall success. After evaluation the model will be used to make predictions.

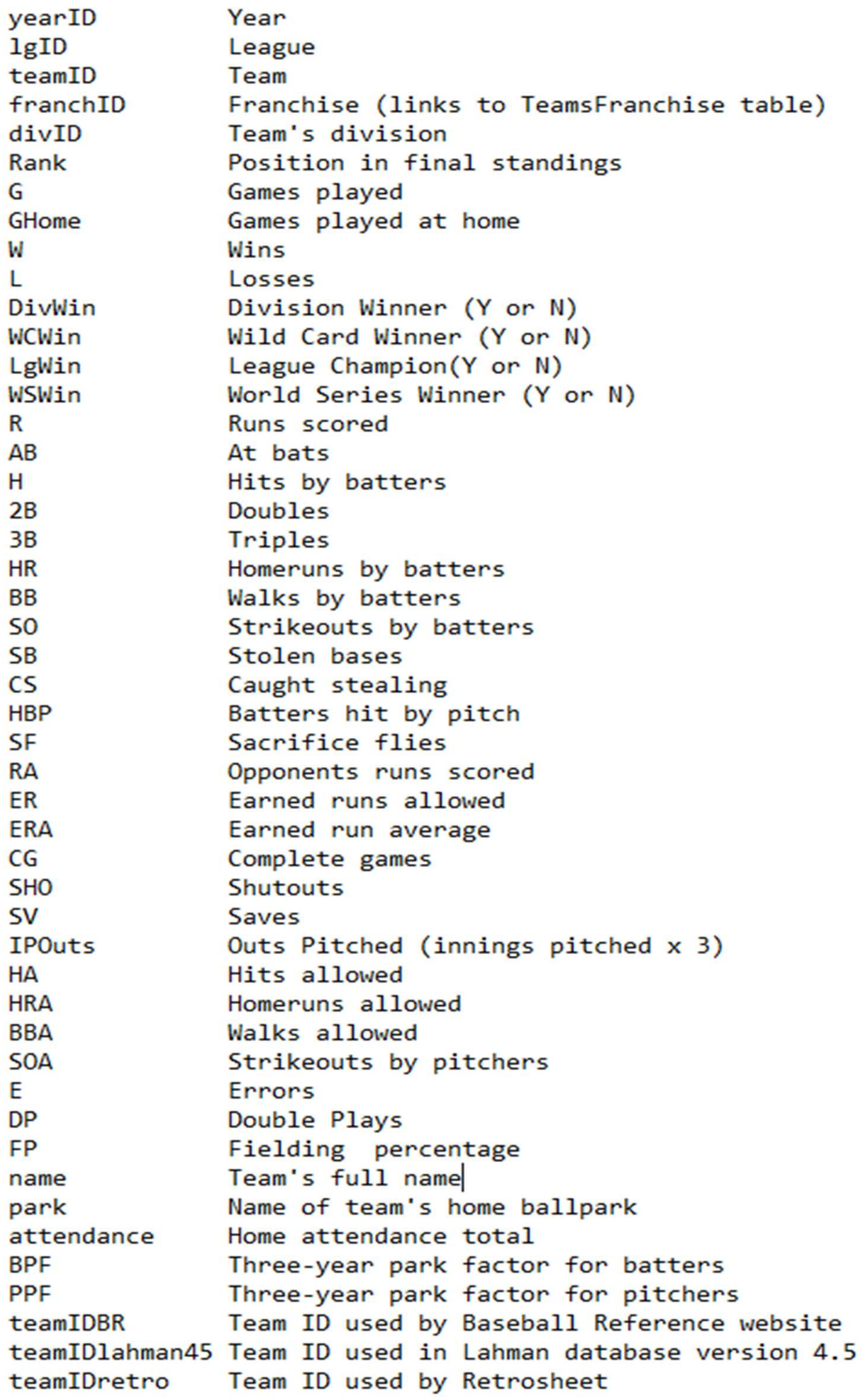
## Evaluation Metrics

To evaluate the regression models, the Data Scientist will used Mean Squared Error (“MSE”) function and calculate the $R^2$ score from sklearn for the simple linear regression model. If the $R^2$ score is close to one then the model is a good one. For the Deep Neural Network, the Keras.model score function will be used to evaluate the accuracy of the model.

## Project Design

- Collect data from the baseball archive from the Lanham baseball website as well as additional statistics from baseball reference website.
- Perform exploratory analysis of the collected data, visualize features, and pre-process the data, and possibly scale or one-hot encode the data
- Apply the Sklearn regression model and performance tune it if need be
- Visualize the results
- Evaluate the model
- Perform a Deep Neural Network Linear Regression analysis
- Visualize the results
- Evaluate the model
- Perform a segment analysis of current MLB players
- Feature scale the player data and remove outliers
- Transform features with Principal Component Analysis (PCA)
- Create player segments based on the features from the linear regression and deep neural network models
- Visualize the segments

## TAB A – TEAMS.csv



| | |
|---|---|
| yearID | Year |
| lgID | League |
| teamID | Team |
| franchID | Franchise (links to TeamsFranchise table) |
| divID | Team's division |
| Rank | Position in final standings |
| G | Games played |
| GHome | Games played at home |
| W | Wins |
| L | Losses |
| DivWin | Division Winner (Y or N) |
| WCWin | Wild Card Winner (Y or N) |
| LgWin | League Champion(Y or N) |
| WSWin | World Series Winner (Y or N) |
| R | Runs scored |
| AB | At bats |
| H | Hits by batters |
| 2B | Doubles |
| 3B | Triples |
| HR | Homeruns by batters |
| BB | Walks by batters |
| SO | Strikeouts by batters |
| SB | Stolen bases |
| CS | Caught stealing |
| HBP | Batters hit by pitch |
| SF | Sacrifice flies |
| RA | Opponents runs scored |
| ER | Earned runs allowed |
| ERA | Earned run average |
| CG | Complete games |
| SHO | Shutouts |
| SV | Saves |
| IPOuts | Outs Pitched (innings pitched x 3) |
| HA | Hits allowed |
| HRA | Homeruns allowed |
| BBA | Walks allowed |
| SOA | Strikeouts by pitchers |
| E | Errors |
| DP | Double Plays |
| FP | Fielding percentage |
| name | Team's full name |
| park | Name of team's home ballpark |
| attendance | Home attendance total |
| BPF | Three-year park factor for batters |
| PPF | Three-year park factor for pitchers |
| teamIDBR | Team ID used by Baseball Reference website |
| teamIDlahman45 | Team ID used in Lahman database version 4.5 |
| teamIDretro | Team ID used by Retrosheet |