Neural Networks: Predicting MLB Runs Scored

Miguel Corona

Sep 2020

Abstract

Baseball is deemed an unfair game due to the wealth imbalance that places a disadvantage to small market teams. The 2002 Oakland A's pioneered a methodology to compete with wealthier teams through Moneyball that emphasized statistics instead of traditional baseball ideas. The results of the organization's experiment revolutionized the manner MLB rosters are constructed. The objective of this paper is to document the efforts of recreating the results from the Oakland A's through the use of neural networks.

1 Introduction

Baseball is a sport in which the winner is determined by the team that has more runs at or after nine innings of play. The team with the more talented roster is deemed the favorite for each series and the quality of the roster may be influenced by the team's payroll. Major League Baseball differs from other American sports leagues in that it does not impose a salary cap on teams to prevent wealthier teams from acquiring the best players on the market. The unfair distribution of wealth across the league forced small market teams to innovate in order to compete with the wealthier teams that resulted in the birth of Moneyball. The key concept in Moneyball is to recreate the best players' output as an aggregate of multiple, less expensive players' output via underappreciated baseball statistics. Peter Brand, a character from the film Moneyball, provided his insight as, "Your goal shouldn't be to buy players. Your goal should be to buy wins. In order to buy wins, you need to buy runs"[1].

Bill James, a baseball statistician, developed a formula to acquire the winning percentage of a team based on the runs scored and runs allowed. An approximation for the winning percentage of a team can be approximated by the Pythagorean Theorem of Baseball[3]

$$W\% = \frac{RS^2}{RS^2 + RA^2} \tag{1}$$

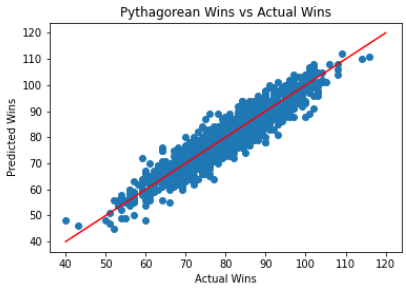

Figure 1: Pythagorean Approximation

The pythagorean expectation can be multiplied by the number of games within the season to acquire the expected wins. The relationship between the number of wins and the pythagorean expectation are depicted in Figure 1. The formula places an emphasis on scoring runs which is typically the most feasible adjustable parameter for most teams. Reducing the runs allowed, RA, is difficult due to the pitching requirements that tend to be more expensive and not as readily available as offensive options. The objective of the experiment is to create and train a neural network capable of predicting the number of runs a team can score. The predictions shall be made with the usage of the statistics: batting average which is the ratio of hits per at-bat, on-base percentage which measures the percentage a batter reaches base per plate appearance, and slugging percentage which measures the expected number of bases a batter acquires per at-bat that ranges from zero to four where four indicates a homerun. At-bats and plate appearances are distinct as at-bats are only counted when the batter attempts to reach base and it excludes walks, hit-by-pitches, and sacrifice plays. Plate appearances on the other hand count for each instance a batter appears to bat.

2 Data Analysis

2.1 Dataset

The dataset for the experiment was acquired from Kaggle as the "Moneyball Dataset"[2]. It contains statistics of interest for all MLB teams from the 1969 season through the end of the 2012 season for a total of 1232 entries. The number of entries is not a multiple of the 30 current teams due to the expansion eras that occurred that resulted in additional teams such as the Arizona Diamondbacks and the Tampa Bay Rays. The statistics contained within the dataset are: Runs Scored (RS), Runs Allowed (RA), Wins, On-base Percentage (OBP), Slugging Percentage (SLG), Batting Average (BA), and Playoffs. The dataset contains more data than is required for the experiment as the features of interest will be subset of the available attributes.

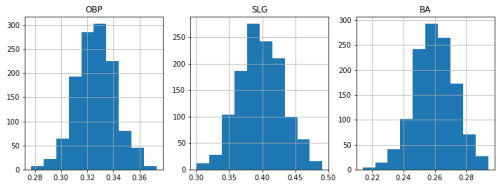

Figure 2: Input Data Distribution Histogram

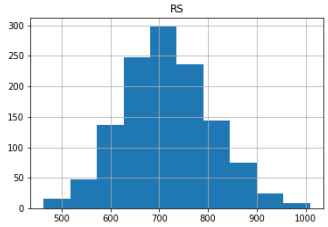

Figure 3: Output Data Distribution Histogram

2.2 Input Data: BA, OBP, SLG

The experiment shall utilize the statistics OBP, SLG, and BA from the dataset to predict the runs scored. The distribution of the input data are depicted in Figure 2 and the pertinent information is contained within Table 1. The statistics of interest are centered about a mean and tail off towards the ends simulating a bell-like distribution. The data is void of any noticeable biases based on the distribution of the inputs.

2.3 Output Data: Runs Scored

The metric of interest that the Neural Network shall attempt to predict is the Runs Scored. The distribution of the output is depicted in Figure 3 and the pertinent information is contained within Table 2. The output appears to be void of any noticeable bias due to the bell-like distribution centered about a mean.

Table 1: Input Data Statistics

Metric	Mean	Std	Min	Max
OBP	0.326	0.015	0.277	0.373
SLG	0.397	0.0333	0.301	0.491
BA	0.259	0.0129	0.214	0.294

Table 2: Output Data Statistics

Metric	Mean	Std	Min	Max
RS	715.082	91.534	463	1009

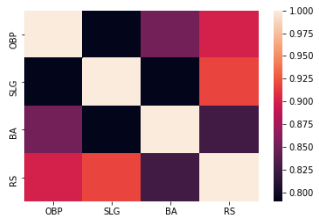

Figure 4: Input/Output Correlation

2.4 Initial Assessment

An initial assessment was performed by observing the correlation between the input data and the output data. The resulting heat map is depicted in Figure 4. The heat map indicates that there is a stronger correlation between the Runs Scored and the metrics OBP and SLG than with BA. The correlation results suggest that a neural network may emphasize the OBP and SLG in training.

2.5 Data Normalization

The input data was normalized through a Z-Score calculation for each element x within X. The Z-Score calculation is performed via

$$\frac{x - \mu}{\sigma} \tag{2}$$

The rationale for selecting Z-Score for normalization was to minimize deviations in which the game was played such as the offensive surge in the steroid era. The steroid era is a blurry timeline between the late 80's and early 2000's that did not penalize players using performance enhancing drugs which resulted in a surge in offensive metrics. The usage of these drugs led to an increased number of runs scored which may skew the statistics with max-mean normalization.

References

[1] Moneyball. Dir. Bennett Miller. Columbia Pictures, 2011. Film

[2] WesDuckett. Moneyball Dataset, 2017

[3] "Pythagorean Theorem of Baseball", Baseball Reference, Sports Reference LLC, https://www.baseball-reference.com/bullpen/Pythagorean_Theorem_of_Baseball