Batting Above Average

Group 12

Brendan O'Connell – 16319077

Jack Mac Namara – 16321927

Motivation

We plan on making an Application Project. We will use machine learning to predict scores of baseball games given season and game statistics.

This project is very interesting to us because baseball is very statistic-heavy sport. There are many raw offensive and defensive statistics available to use in predicting the outcomes of games. The resource we plan on using to collect the data also includes some calculated aggregate statistics which we can consider. For example, WAR (wins above replacement) is a non-standardized baseball statistic that represents the number of additional wins a team has achieved as a result of an individual's effort. It may be interesting to see which of these statistics provide better predictions.

Dataset

The data we plan to collect is simply baseball statistics from the 2018 season. We plan on testing different offensive and defensive statistics in order to predict the scores of games. At the moment, the offensive statistic we are considering is batting average, and the defensive statistic is average number of runs allowed per 9 innings. We believe that these two statistics have a significant impact on the number of runs scored by a team in a game.

We plan on collecting this data from a baseball statistics website called https://www.baseball-reference.com/. The team-wide statistics will be pulled manually from the website, but the individual rosters of each game will be pulled using a web scraper implemented in Python using Selenium.

Method

The problem we are tackling is a regression problem, as we intend to predict scores. Therefore, we will be using several different regression methods e.g. linear regression, logistic regression, lasso regression. We will also attempt to implement a deep neural network to predict the scores. We will use cross-validation to calculate the optimal hyperparameters for each model.

Intended Experiment

We plan on analysing the 2018 MLB season and building a model using the offensive and defensive statistics collected. Our primary goal is to test data (which has been split into training and testing sets) for the 2018 season. To measure fluctuations across different seasons, we will then test our model against a few games in the 2019 season to see if there is in fact a difference between seasons, or if our model will remain consistent in predicting scores. We plan to evaluate our regressors using several methods, including mean squared error, mean absolute error, and R2 score. We will investigate which method is the best for our problem and evaluate models using our selection.