Fall 2023 Erdos Project Executive Summary

Team Pythagorean Expectation as a Prediction in Baseball: Steven Creech, Mohsen Mohaghegh, Joe Schulze, Weiqi Want, Billy Warner

Github: https://github.com/schulze61/erdos_baseball_project

Overview

Our goal was to develop a predictive model for a baseball season using the notion of pythagorean expectation which is a statistic that accurately predicts the win percentage of a team over a season. We built two models and compared them to an Elo based model created by 538. Our metric for the success of our models is given by the Brier score which is a number between 0 and 1 that measures how good a prediction is to the actual outcome of an event with a score of 0 meaning that there was no error in the prediction.

Approach

We developed two different approaches to this problem the first was based off simulation and the second was a regression model.

- The simulation model used the fact that if one assumes that the Pythagorean expectation formula is true, then the runs scored and allowed can be modeled by random variables drawn from a Weibull distribution. We had four different methods of determining the parameters for our simulations.
- For the regression model, we modelled the probabilities of victor for a team via a logistic regression on a single feature. We focused on Pythagorean expectation for one feature, but we also constructed a new statistic that we called the runs scored adjusted for opponent that takes into account the opponents defense compared to our offense.

Results

As for our results, we found that the simulation model was not as effective as the 538 model in predicting the winner of a game. Furthermore, of our four different methods to get the parameter of the Weibull distribution, we found that the method that took into account the opponents defense was the most effective of our four models.

As for the regression model, we found that when we focused on the pythagorean expectation, it did not our perform the 538 model. However, for the opponent adjusted statistic we did achieve a lower Brier score on average in most seasons. Furthermore, we observed that if we looked only towards the end of the season, each of these regressions tend to converge. This should make sense since the Pythagorean expectation should be a better predictor at the end of the season when we have more information of how the team performed on average. Thus, we found that our opponent adjusted model tends to be a good predictor while the Pythagorean expectation tends to be a better predictor at the end of the season.

Future Iterations

After observing the convergence in our logistic regression model, in the future, we would like to see if we restrict our simulation only to the end of the season if that improves the simulation model to around the levels of the 538 model.