Predicting Baseball Season Final Season Records using ML

Liam San Pedro, Joel Rodriguez

liam.sanpedro01@utrgv.edu, joel.a.rodriguez03@utrgv.edu

September 26, 2019

Summary

Using Machine Learning, the goal of our project is to predict what final records each baseball team (in the MLB) will have by the end of the regular season. Since the MLB season just ended, we are using data from the 2018 season, and validating our data by comparing it to the 2019 baseball season. Effectively, since the 2019 season just finished, we'll be able to see how accurate our methods were.

Background

We want to use data from previous, more recent seasons (the latest being 2018). The goal of this project is to predict the win rate of the 2019 season. Since the season is already over, we want to merely validate our information in hopes that we can predict the upcoming 2020 season.

We know that we have parameters from previous seasons, such as how well they did at home games v. on the road, what their team's fielding percentage was, their run differential, etc. There are also intangibles -- we also know that there is an off-season, where the team makes changes to their coaching, their team's roster, etc. We can combine these pieces of information to gather whether a team is expected to do better than their previous season, or potentially, worse.

The idea for our project is to predict how much they improved, or worsened, compared to previous seasons. Using a past season, we know that this is a good basis for letting our model learn with raw, concrete data. We can assume that with 162 more games, they should perform at about the same rate, given there are no changes. But using machine learning, we want to add in the idea of intangibles making a difference to the team's overall record for the next season. We should be able to assume that a team who spent $200 million in offseason signings -- whether it be a new coach or new player -- is expected to perform at a better rate than a team who might've spent only $30 million during the offseason. This is the idea of declaring weights to intangibles, since there is no definitive way to track how a team became better or worse.

The model should not return a value, like 0 or 1, because we are predicting team ranking with relation to the other teams in the league. To do this, we want to use a data model that is common with predicting based off of previous data: Artificial Neural Networks. By using the ANN model, we can train a data set which has components (in this case, SOS, SRS, etc., explained further in the Data/Methods section) and attach weights to those components. By attaching certain weights to those components, we can continuously change the weights to match high levels of predictive accuracy. Eventually, we want to check what weights made the model most accurate, and attach those weights to the 2019 season, and see how our data turns out.

Goal and Objectives

The goal of the study is to help find a pattern using a team's previous seasons statistics and adding the effects of their offseason. We want to eventually train a model to find definitive weights between each individual statistic. Thereafter, we want to add how much money a team spent during the offseason, using actual monetary values. These values can help determine how a team *should* improve, given they are spending money to do just that. By combining both previous statistics and money used to improve, we should be able to predict how a team fares for the 2019 season.

Data and Methods

Using data up to the 2018 season, we plan to use the following parameters for our study:

- Strength of Schedule (SOS): The number of runs per game their opponents are better/worse than the average team
- Simple Rating System (SRS): The number of runs per game they are better/worse than the average team
- Pythagorean Win-Loss (pythWL): Expected win/loss record based on the number of runs scored and allowed by the team
- Money Spent for Players (MSP): Money spent used specifically for signing, or trades between teams; generally known in sports as salary
- MSPD: Difference between a team's MSP from season before
- ERA Average for team: The total earned runs allowed per nine innings.
- Home Runs of the previous season
- Batting and on base percentage of the top 12 players of each team.

- MSFD: Difference between a team's MSF from season before
- Money Spent for Coaching (MSC): Money spent for a team's coaching
- MSCD: Difference between a team's MSC from season before

Using the data, we plan to implement the method of classification. This is where we do reinforced learning to determine which characteristics of a data set are most important to determine the outcome of the team's record.

References

baseball-reference.com