Data Mining - Proposal (Problem)
Using Data Mining to predict the MLB 2018 Cy Young winners

Mick Tuit

November 2018

1 Introduction

Every year in the MLB (Major League Baseball) two pitchers are voted to be the winners of the 'Cy Young' awards. These pitchers are considered to be the best performing pitchers of the season. The MLB consists of two leagues, ie. National League (NL) and American League (AL) and the pitchers correspond to the leagues. Since the winners are decided by votes from members of the Baseball Writers' Association of America, this award is somewhat subjective and interesting to predict.

2 Machine Learning software

To solve this problem I am planning to use the XGBoost open-source software library (Python package) which uses gradient boosting, which uses decision trees. I will use this to create a model. Using 10-fold cross validation I want to validate the model and I can use XGBoost to tune the parameters. To evaluate the model I plan to use AUC metric and accuracy. Then I can use the model to predict the winners for this year.

3 Literature

I used chapter 4 of TSK for the information about classification. I also used this website for inspiration of the idea.