ML Final Report

Mr. O'clock insanely practices baseball to join MLB at 103.

December 21, 2024

1 Introduction

This final project is aim to use machine learning to predict HTMLB outcome. We have explore through many algorithms and data preprocess, and result in 0.58989 accuracy in predicting same season outcome and 0.55473 in predicting a whole new season.We are going to break it down what have we done with every model and preprocess and what are the results.

2 Data Preprocess

Team Attribute and Pitcher Attribute

We try to add 60 boolean columns: is_home_team_{team name} and is_away_team_{team name} which indicate that if home team or away team equal to team name. Same method can be used on pitcher name.

However we try it on random forest an observed that these attribute may have low importance. We guess that perhaps this kind of data is too scatter to make it important

Alternatively, we can labelize string columns, this may work in algorithm based on decision tree.

Drop std and skew

Since we think std and skew has little relationship with winning or lose. We try to drop all of those column.

The result is that there are no difference in accuracy. However, training speed is faster using this data.

3 Models

KNN

PLA

Linear Regression

Logistic Regression

SVM

Random Forest

CATboost

Catboost is a algorithm based on Gradient Boosting Tree

Blending and Bagging

4 Conclusion