MLB Injury Prediction Analysis

Colby Reichenbach

2024-03-22

Major League Baseball (MLB) is faced with the task of balancing player performance and health with success and financial stability. As a result, predictive algorithms have become increasingly important and popular but have not been resistant to ethical dilemmas. The study done by Karnuta et al. (2020), "Machine Learning Outperforms Regression Analysis to Predict Next-Season Major League Baseball Player Injuries' ', showcases the potential of machine learning (ML) to forecast injuries with high accuracy. This paper will serve as an attempt to validate the findings, and compare the efficiency and accuracy of ML models over traditional logistic regression in predicting the likelihood and location of player injuries, as well as touching on the ethics of using such models in this context.

In this study, data was drawn from four publicly accessible online baseball databases. Player information was confined to data from the 2000 - 2017 totaling 17 seasons of data. This encompassed data from 1931 position players and 1245 pitchers. The data included variables such as age, performance metrics, and detailed injury history. This data was used to train 84 different ML algorithms under 6 different models: logistic regression (LR), random forest, k-nearest neighbors (KNN), Naïve Bayes, XGBoost, and a top 3 ensemble classifier. The top 3 ensemble classifier is formed from a combination of the top 3 performing models, though it is not mentioned explicitly in the paper which 3 models these were the paper does give statistics for the overall accuracy of each model type. Nevertheless, the top 3 ensemble classifier was built using soft voting, prompting the model to decide to classify a point as "yes injury" or "no injury". Furthermore, each model required 90% of the data to train the model, while the other 10% was used to test the model. By repeating this 10 total times, and using a separate 10% of data each time, all of the data is eventually used to test the model without being used to train the data. Overall these models aimed to predict general next-season injuries and specific anatomical injury locations. Data was split up based on player roles, pitchers and position players, and ran separately through the model to accommodate for the distinct injury risks specific to each role.

To validate the results from the models, the area under the receiver operating characteristic curve (AUC) was used to determine the validation. Accuracy %, F1 score, and Brier score loss were also used. Results varied based on the model. For predicting next season injuries for position players, the top 3 ensemble model was the highest AUC, while also boasting the highest accuracy. However, for pitchers random forests had the same AUC as the top 3 ensemble, though the top 3 ensemble had a higher accuracy. We also see a slight decrease in accuracy in injury prediction when evaluating pitchers, this suggests pitchers have a higher complexity when predicting injuries and may require more specialized models. As far as predicting injuries to specific anatomical locations, Top 3 ensemble was also the best predictive model for position players for every region with the exception of the elbow. Logistic regression predicted injuries to the elbow region with the highest accuracy of the models at 63%. For pitchers, top 3 ensemble was the best predictive model. Overall the trend showed through the results shows the ML models outperformed logistic regression. This follows the overarching conclusion that ML models outperforms logistic regression, as seen in 13 of 14 cases, offering a potential to implement ML to revolutionize the prediction of injuries in professional sports.

Although this study shows a remarkable accuracy in injury prediction, it raises multiple ethical concerns. The use of sensitive health information within the models brings privacy and consent concerns for the players. In order for the model to predict injuries with a high accuracy, sensitive health data must be included in the model, data that some players may not want shared with front offices of the organization. Furthermore, ML models are susceptible to learning biases learned through their training data. If the data the model was trained on reflects bias in areas such as race or age, the model might inaccurately classify players as a result.

However, arguably the biggest concern is the impact on players' careers. Labeling players as injury prone has many consequences such as limiting games played, decreased market value, and challenges in contract negotiations. Front offices could use the models to make personal decisions on signing players, extending them, or cutting them could be deemed as exploiting them. For example, take a 35 year old all star player who is seeking a new lucrative contract that is labeled as injury prone by the model because of learned biases by the model that inaccurately predict older players as injury prone. Would it be ethical for the front office to use that information to offer the player less money than he is asking for because of fear of injury? It is a difficult question, and one that would be the main topic of discussion if these models were implemented for use within the sport.

In all, ML's growing presence in Major League Baseball offers a potential revolutionary way to predict injuries, as well as opening the door to create models to predict other variables in the sport, such as performance. Nevertheless, it is surrounded by ethical concerns over player privacy and consent, as well as its overall impact on player's careers. In order to incorporate these ML models within not just the MLB but all professional sports, there needs to be proper guidelines that protect players and front offices from the potential exploitative outcomes that could come from the results. If properly incorporated, there is no doubt machine learning would help to advance MLB, as well as other sports, in many aspects and offer benefits to both front offices and the players.