Analyze and clean MLB data and apply machine learning tools to hazard classification tables

Wei-Cheng Su

Department of financial engineering and actuarial mathematics, Soochow University

Abstract

In this paper, we explore the issues related to the lack of professional athletes in Taiwan's insurance industry and used the open data in mlb official website, fangraphs and baseball reference. The detail of the construction for our goal be given as follows: First, we discuss and analyze the relevant insurance products in different countries. We evaluating players and the risks they faced, and collected the data of player which is for the injury rate, re-injury rate, salary, number of injuries days, number of injuries and other contents of baseball players. Therefore, we obtain the average cost of injury. Secondly, we used the decision tree and some machine learning algorithm to discuss the accuracy of predicting the player's injury and the depth parameters in the decision tree. Moreover, we use method of model classification to group MLB players, and then use the injury rate to develop a risk rating table. Finally, the random decision forests method is used to pick up the important of the input feature for the prediction of injury, and provide reference for in the future.

Keywords: Sports Injury Insurance, Injury rate, Injury risk classification method, decision tree, random forest