Keegan and Subhan
Multivariate Analysis
Professor Hartlaub

Background

Like all athletes, elite runners are subject to a lot of injuries. However, in most sports, there are always surprise injuries. These consist of injuries that are not common and are unanticipated. As far as elite runners are concerned, they do not encounter a lot of these 'surprise injuries.' Most of their injuries happen because of overuse training meaning it should be easier to predict when to compared to other sports that have more randomeness. It is because of this reason we decided to predict injuries in elite runners. In addition to having a model with high predictive power, we also want to identify the most important variables in predicting injury.

Data

Our data was sourced from Kaggle and is made up of a team of 74 elite athletes that run 800m to the marathon. However, there were certain things that we did to improve our models. The first thing that we did was create a window that showed the past six days for each independent variable using the lag function. We also took the time to remove any observations where an athlete had been injured in the past 3 weeks to avoid capturing reinjuries that had not yet healed.

Findings

We started off by constructing a logistic regression model. The variables that we used in this baseline model were selected using backward elimination. We found four of the nine variables to be statistically significant, and the AUC we recorded was about 0.60. Although this model was better at guessing,we wanted to look into more complex models to see if they could do better. Therefore, we decided to move on to a much stronger machine-learning technique - Gradient Boosting. Since our dataset was very unbalanced we were able to get an extremely high AUC of .99 and an accuracy of 90% however this is because our model was very bias to the majority class making it a useless model. We then adjusted the model by testing on balanced data by oversampling the minority which lowered the accuracy to 58% but gaves us a much more balanced model. The model ended up being balanced but not that useful so we decided to focus on the most important variables which were strength training 4 days ago, km in z3/4 6 days ago, no run sessions(XT), total KM 4 days before, km in Z5, threshold 1 or 2 two days ago, and perceived recovery 5 days ago. While our model fails to be very useful it does point to some possibly important variables.

Future Work

For future research, the one thing we would definitely want to do is find a much larger dataset. Our dataset goes back only seven years hence we think that a larger dataset would definitely improve our models. We would also like to do a time series analysis to see if there are any differences in the types of injuries runners have had to deal with over the past 50 years. Also, we would like to explore other types of models like Deep Gaussian Neural Networks. There appears to be a decent amount of data available for college runners. Therefore we would like to make similar models for them and compare them to the models for elite/Olympic level runners and see the differences in injury types and frequency.