Data Preparation for Injury Prediction

Mladen Jovanovic [1]

[1]Faculty of Sport and Physical Education. University of Belgrade.

Injury Prev | Team sport | Data analysis | R

Headline

Before commencing any type of data analysis task (be it descriptive, predictive or causal inference) (1) complex phenomena of training needs to be represented in a format that is appropriate for a given analysis and questions at hand. This process involves numerous simplifications and assumptions, which become part of the statistical model itself. (2)

When it comes to injury predictions, practitioners are interested in predicting over-use soft-tissue injuries (i.e. hamstring, quads and groin pulls) that results from training load. Training load is usually monitored and represented using GPS (i.e. Total Distance, High Speed Distance, etc), heart rate (i.e. TRIMP score, time over 90% HRmax, etc) and subjective ratings (i.e. session RPE). More complex models involves moderation of external (GPS) and internal (heart rate and session RPE) training load effects on injury likelihood by athlete readiness (i.e. wellness questionnaire, CMJ analysis, grip strength, etc) and athlete characteristics (i.e. age, height, previous injury, etc). (2)

Currently there is no consensus on how these complex data and relationships should be represented for injury prediction tasks. This technical note aims to explain one particular approach of data representation and preparation for injury prediction tasks. Generated sample data in this technical note involve season long day-to-day collection of [1] session RPE, [2] Total Distance and [3] High Speed Distance for three athletes who suffered over-use soft tissue injuries. The accompanying video details data preparation and features engineering (creating new variables from existing ones) in R-Studio (3) (which is IDE for R language) (4). R packages used in this technical note are plyr (5), dplyr (6), reshape2 (7), ggplot2 (8), TTR (9) and zoo (10).

The following techniques are explained in the accompanying video:

- Data representation using day-to-day approach
- Exponential Moving Averages
- Acute to Chronic Workload Ratio (ACWR)
- Injury Lead Tags
- Lag variables

After preparing the data and engineering new features using the approach explained in the accompanying video, practitioners can proceed with prediction tasks using multitude of classification methods (11). Expanding on these techniques is beyond the scope of this technical note and interested readers are directed to provided references.

Accompanying video

https://vimeo.com/279226723/0299ff1237

Accompanying dataset

R script and dataset used in accompanying video is available at GitHub repository: https://github.com/mladenjovanovic/Data-Preparation-for-Injury-Prediction

Twitter: Follow Mladen Jovanovic @physical_prep

References

1. Hernán MA, Hsu J, Healy B. Data science is science's second chance to get causal inference right: A classification of data science tasks. ArXiv e-prints. 2018 Apr; Retrieved from: https://arxiv.org/abs/1804.10846v4

2. Jovanović M. Uncertainty, heuristics and injury prediction. Aspetar Sports Med J. 2017 Feb;6:18-24. Retrieved from: http://www.aspetar.com/journal/viewarticle.aspx?id=353.W0icuNgzYWp

3. RStudio Team. RStudio: Integrated Development Environment for R [Internet]. Boston, MA: RStudio, Inc.; 2016. Retrieved from: http://www.rstudio.com/

4. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Retrieved from: https://www.R-project.org/

5. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software [Internet]. 2011;40(1):1–29. Retrieved from: http://www.jstatsoft.org/v40/i01/

6. Wickham H, FrançoisR, Henry L, Müller K. dplyr: A Grammar of Data Manipulation [Internet]. 2018. Retrieved from: https://CRAN.R-project.org/package=dplyr

7. Wickham H. Reshaping Data with the reshape Package. Journal of Statistical Software [Internet]. 2007;21(12):1–20. Retrieved from: http://www.jstatsoft.org/v21/i12/

8. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2009. Retrieved from: http://ggplot2.org

9. Ulrich J. TTR: Technical Trading Rules [Internet]. 2018. Retrieved from: https://CRAN.R-project.org/package=TTR

10. Zeileis A, Grothendieck G. zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software. 2005;14(6):1–27.

11. Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer Science & Business Media; 2013. 616 p.

Copyright: The articles published on Science Performance and Science Reports are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.