Public Project Summary: A Data-based Approach to Prevent Lower Body Muscle Strains in Soccer

Author: Tamás Koncz

Supervisor: Gábor Békés

Program: Master of Science in Business Analytics

Date: August 21, 2018

Table of Contents

Introduction 1
Injury prediction 2
Important factors in injury risk 3
Closing remarks 4
References 5

List of Figures and Tables

FIGURE 1: LOST PLAYER DAYS 1
FIGURE 2: COST-BENEFIT ANALYSIS OF MODELING APPROACH 2

Introduction

The aim of this work is to identify direct links between injury risk of professional soccer players and certain relatable factors.

Specifically, the injuries in focus are muscle strains happening in the lower body. To grasp the seriousness of this problem, consider the below charts:

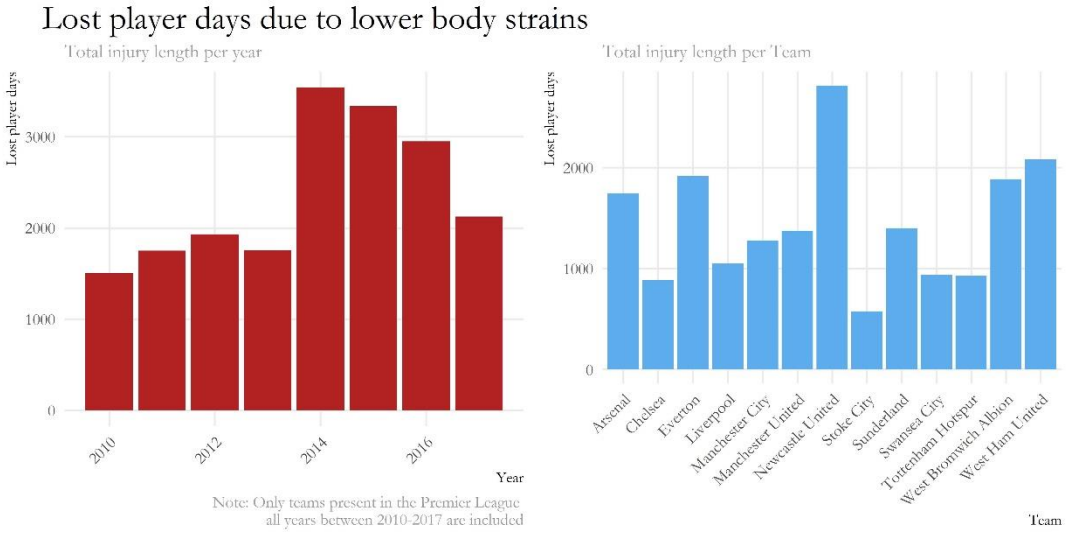

Figure 1: Lost player days

Many matches are missed by players due to injuries - a costly experience for teams, and an unpleasant one for both players and soccer fans.

A wide collection of player level information was used to explore the existing relationship of many variables and injury risk, using visual and statistical modeling methods.

While it was not possible to establish an accurate prediction system for individual injuries, the analysis demonstrates that there are other valuable insights that could be gathered. Among the most important ones are the impact of past injuries on future risk.

There are believed to be two roadblocks to high-accuracy predictions: 1. Injuries are rare events, which makes it hard to model them from a technical perspective 2. Any dataset of historical player injuries is inherently biased, as coaches and teams already apply their own mitigation methods.

The last section contains suggestions for future work, among them an interesting concept of investigating injuries in multiple-day time windows, rather than treating them as in-game occurrences.

Injury prediction

The primary goal of this project was to establish a statistical link between soccer injuries and other explanatory factors.

These factors include but are not limited to the players physical attributes (age, height, weight), play-load (minutes and games played in recent time periods), and injury history.

With the use of state of the art machine learning methods, a predictive relationship could be established, however with limited accuracy. A major technical challenge was the (thankfully) rare occurrence of lower body strains (even if they are the most frequent injuries in soccer), which amounts to roughly 1 injury per 100 games per player.

A cost-benefit analysis was performed to calculate the added value of making player rest decisions based on models, compared to the status quo (decision is made by the coaching staff).

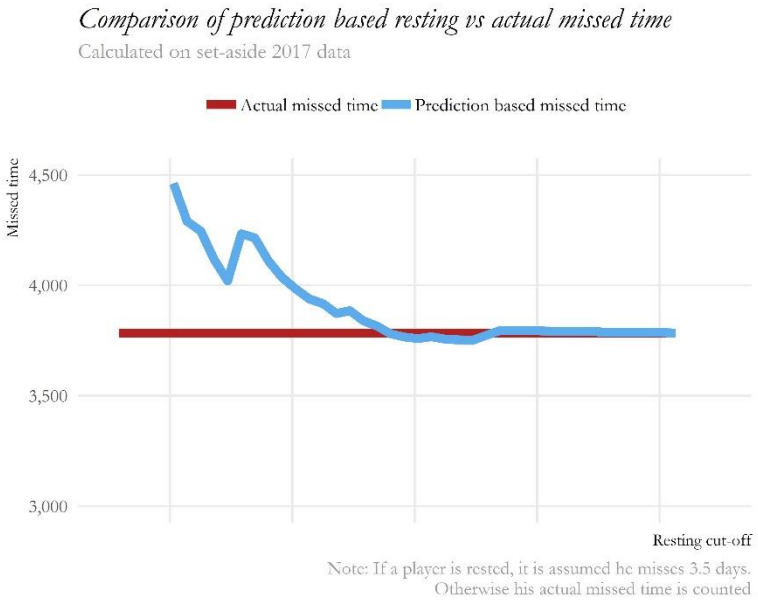

Figure 2: Cost-benefit analysis of modeling approach

Figure 2 benchmarks the use of a model, with different conservativeness levels, against the total missed time by players. This shows that even based on historical data, virtually no reduction in the total missed time can be achieved.

The reason is that if a player is rested, that also incurs a cost, at least one missed game. Even if injury events can be modeled with the use of public data, the accuracy of predictions is not good

enough to avoid resting players who are not in high-risk of injury, and small 1 game rests add up, if they happen every now and then.

A likely challenge is that prediction of injuries is not a controlled scientific experiment – teams already do their best to avoid player injuries, hence we have no "what-if" data on what would have happened if they employed no mitigation of their own.

Important factors in injury risk

Even if a method could not be established that adds value in individual players' resting decisions, this does not render the analysis work useless. Models good establish good-enough relationships between injuries and explanatory factors on the aggregate, enabling us to investigate which ones are the most likely to lead up to injuries.

Amongst them the most important is a player's previous injury history – as Figure 3 shows, players who had been injured in their career before, are 2.5 times more likely to get reinjured (a 1.2 percentage point absolute difference) in any given game.

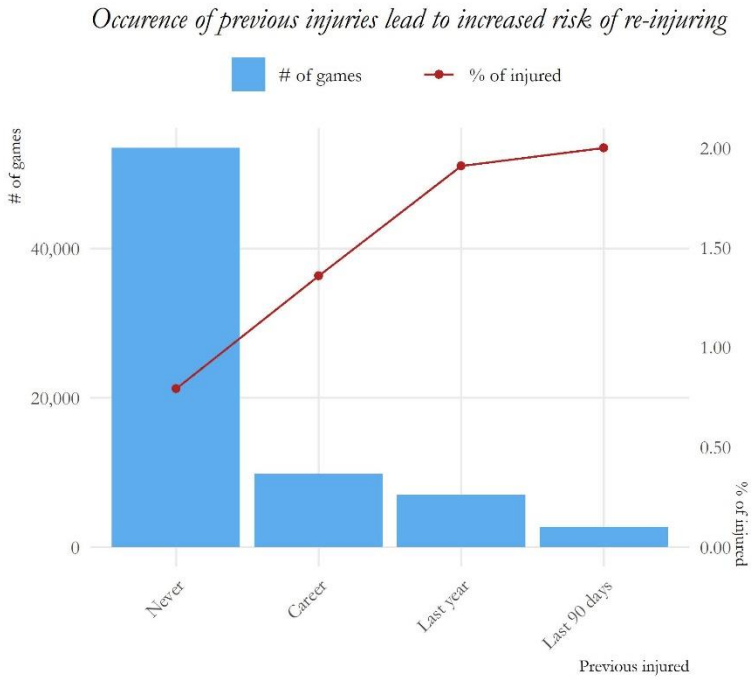

Figure 3: Injury risk against past injuries

Other important factors included age (the older the more likely to get injured), and amount of play in the recent season (which, somewhat surprisingly, shows a reverse link to injuries).

Closing remarks

These results should not be looked at without context – it is very important that analyzed data was from one professional league, covering the last 8 available years.

To understand whether they hold generally true, one needs to consider structural components – type of play, workload, natural conditions (weather) and player composition, which might be all specific to the field of application.

While this analysis was done with high emphasis on robustness and reproducibility of results, it is encouraged to revisit this problem with new methods or data. An interesting concept was raised in a recent paper (Talukder & Vincent, 2016), which intends to predict NBA injuries in a forthcoming multi-day window. Applying this to the soccer injury problem, it was possible to significantly increase predictive accuracy, however still not to the extent to make it unquestionably superior compared to status-quo of resting decisions.

References

Talukder, H., & Vincent, T. (2016). Preventing in-game injuries for NBA players. Boston: MIT SLOAN Sports Analytics Conference.

EMPTY PAGE

EMPTY PAGE