Jared Smith

May 2, 2023

Human Performance Analytics - Professor Barron

**Introduction**

In this project I looked to investigate the problem of injury risk in relation to a stringent training schedule. The dataset contains incredibly detailed training logs that could be significant in outline certain metrics that are related to injuries. This is important to me, because I want to build a similar model using personal training data in the future, so this problem would be a good case study for future work in the field.

**Related Work**

There is related work towards this dataset through the citation of this project [1]. I felt their model was pretty elementary and did not reproduce an appropriate accuracy for the depth of data that they had. The model for weekly data produces an AUC of 0.678 [1], which is a decently accurate model for real world data. In a model for their daily dataset they stated their model can determine “around 30% of the injuries can be detected with a false positive rate of only 10%” [1]. Overall, the study produced findings that the sports analytics field can say are significant, but let’s attempt our own test on the data and try to outperform their model.

**Data Description**

The dataset utilized tracked data from middle and long distance members of a Dutch Running Team. Overall, 74 national and international level runners who participated in middle and long-distance events were included (i.e 800m to marathon). There were 27 female and 47 male runners within the dataset. However, gender and the identity of these runners were kept anonymous. The data contained 42,798 total observations over the 9 year span. The data was separated into daily and weekly logs which the coaches tracked. The daily logs were important for determining short term changes, while the weekly logs could track the importance of longer-term activities leading up to injuries. Important features in the logs were: Number of sessions, number of rest days, total KM run, summary statistics for an athlete’s perceived exertion/recovery/training success, as well as a binary column for injury. There were other variables that included intensity during certain zones of training. I did not truly understand this feature, but they were not found to be important in the final model. I also included derived metrics for an athlete's performance in previous weeks leading up to the week of an injury. These variables were noted as ae\_7\_21 and er\_7\_21, which is the athletes average exertion in the current week, weighted by their previous two weeks of average exertion and average recovery during training. Rec\_MaxExer\_7\_21, which factors the athletes current recovery based on the max exertion during previous weeks. Lastly, Exer\_v\_rec\_7\_21, which is the athletes max exertion in the current week based on the recovery in previous two weeks. For the final model, 75% of these observations were split into training data and 25% into test data.

**Methods/Results**

I first used summary statistics within the dataset to highlight certain athletes who are more injured than others. I then employed basic visualization techniques to obtain initial insights into the data. I used density plots of features to find general insights that could lead to deeper findings. With a factored injury variable, I was able to gather analysis on certain metric values that are significant in leading to injury (Appendix 1). These methods were applicable to gather initial insights on injury risk, but hold some limitations. Obviously, analyzing strictly from visuals does not express any direct correlation between variables. Also, this method I employed did not employ previous weeks of injury, thus I could have missed several important insights. Otherwise, it was a good start to the bulk of the analytics of this project.

I then performed a search and used an XGBoost model to identify the key features associated with runners' injuries in the dataset. This method was applicable in applying classification, as our representation of injuries were 0’s and 1’s. The search applied the previous use of visuals to find what I deemed important variables. I then used logistic regression to determine if there were any significant variables that could lead to injury. From these original searches, I was able to gather that regular training session metrics (sessions, total kms, perceived metrics), as well as my derived variables could be important factors leading to injury. I also attempted to use intensity metrics were were outlined in the .txt file associated with the dataset, but these overfit my model. Then I applied an XGBoost model with these important variables and split the data into 75% training and 25% testing data. I did this in order to train the model on a sufficient amount of injuries in the training dataset. I also had to slow the learning rate (eta), to improve my model. In the beginning, the model would teach itself on the data too quickly and provided a low accuracy of predicting injury. I used my predicted values on my Test Data and inputted into a correlation matrix to investigate the specificity and sensitivity. From the XGBoost model I was able to plot its AUC value as well as create SHAP charts to weigh the overall importance of each variable in relation to injury.

From this setup, my overall results were incredibly promising! The initial visuals demonstrated that for athletes with high injury risk, there are certain metrics which are higher during weeks where they sustain an injury. The applications of my methods created a model with 0.713 AUC and an accuracy of 75.08%. In this context, AUC (Area under Curve) and the accuracy means that the model has a moderate ability to predict injury. My model when applied to the testing was able to predict and highlight the majority of injuries within the testing set (Appendix Figure 2). The confusion matrix demonstrated that I am marking a large number of uninjured athletes as injured, and missing a few athletes who were actually injured. The medium-high sensitivity value of 0.5878, means that we are missing a few of the injured observations, but there could be many reasons for this. Since the data is weekly, when the injury occurs could hold a great influence on the data. If one was to get injured early in the week they would have very low values versus someone later in the week. However, I think this analysis can be used as actionable for the coaching staff. I can most likely use these results to create a “Vulnerable to Injury” variable for those I falsely predicted to be injured. I believe that since there were only 575 injured observations within this dataset it made it hard to understand exactly, since there was not a larger population of injuries. I think a future focus could include the athletes with the most amount of injuries to reduce the size of the dataset. Some other challenges and issues I dealt with were optimizing the model as I spoke of previously with the eta metric.

The SHAP analysis resulted in the most important metrics in relation to injury (Appendix Figure 3).. I was able to gather that during weeks of injury max/avg exertion and the number of sessions during that session are incredibly important. My derived variables, which incorporate previous training weeks, also found that the level of recovery in a week of injury, weighted by max exertion in previous weeks has a significant impact. These derived variables could be used in the future to develop a threshold variable for coaches to look at to possibly reduce injury risk.

**Actions**

This analysis can be presented to running coaches to highlight important risks for injury. First, from the confusion matrix, we can create a variable for those that were predicted to be injured. There is a combination of the metrics that has led our model to designate these athletes at risk of injury. From the new variable we can assign a “Vulnerability to Injury” variable for athletes at the beginning and end of each week. Personally, I think the Recovery based on an athletes previous two weeks of Max Exertion is a good variable to base this on. If this value crosses a certain threshold, coaches can assign a rest day or lighten the workload to begin the week. This is very similar to Whoop’ strain analysis that they currently use for their sports band, but we have more insight with the data we are tracking. Our main goal is to prevent as many injuries as possible, while maintaining a high level of training.

Second, the team needs to include some other variables within the dataset for more insights. They need to associate the events in which each athlete participates within the datasets. Athletes participate in different events, so their needs are vastly different. Obviously, one can gather that some athletes run more than others and make inferences on the data. But it is important to begin grouping these athletes so coaches can garner more insights. This is similar to the soccer example in class, where based on position, some players ran more or got injured more than others. This can be used to group runners and possibly provide a deeper insight in the future. The data also needs to contain injured days/weeks of an athlete, as well as method of injury. This will show coaches how much time they are losing to injury. As well as if injuries were activity related or something entirely else. Overall, there is a lack of transparency within this dataset, where new insights could be seen.

I would use Change Management, to emphasize the importance of my new “Vulnerability to Injury” variable and how this could change performance in competition and prevent injury. I would like to establish the importance of load management throughout the season. This would be a great way to direct the coaching staff and athletes towards my mindset. In this scenario the “Elephant” would be the training staff, athletes, and coaching staff. It is the entire makeup of the running team. Thus, I would have to get the training staff and athletes involved in the process of load management. I would first educate the athletes on the importance of load management and how this new variable can be used to monitor their personal health. This can be done safely by privately providing each athlete's information to the training staff and athletes alone. If “Vulnerability to Injury” is above a certain threshold training staff can request days off or decreased workload. I believe that providing data to the athletes will also make them mindful of themselves and their physical feelings. The current process of daily data collection, already is a great habit that will shape their path through the process. I think this methodology is a great framework:

A. Demonstrating importance of Injury Risk to coaching staff

B. Educating Athletes and training Staff on the “Vulnerability to Injury”

i. Annual team meetings on data collection and past injuries

C. Habitual data tracking and management

i. Additional daily/weekly presentation of performance to athletes

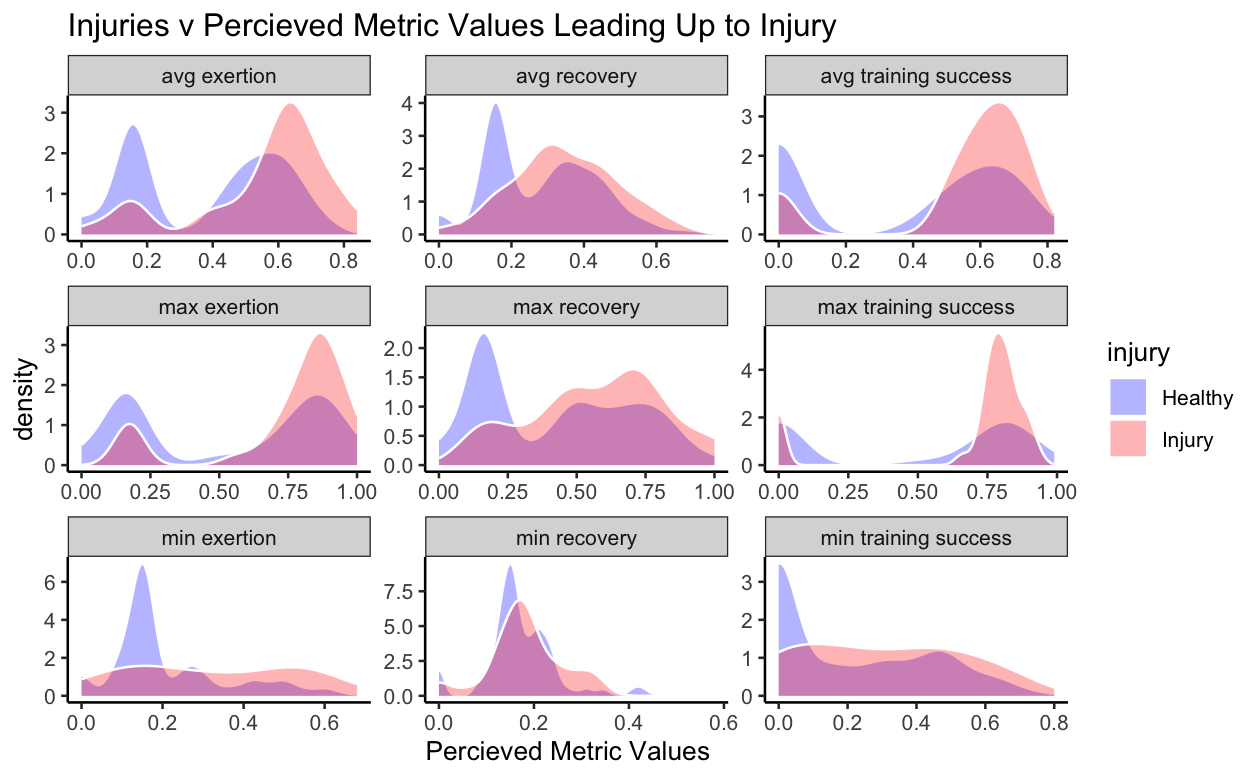
ii. Weekly check ins by training staff

D. Monitor athletic performance and injuries throughout next season

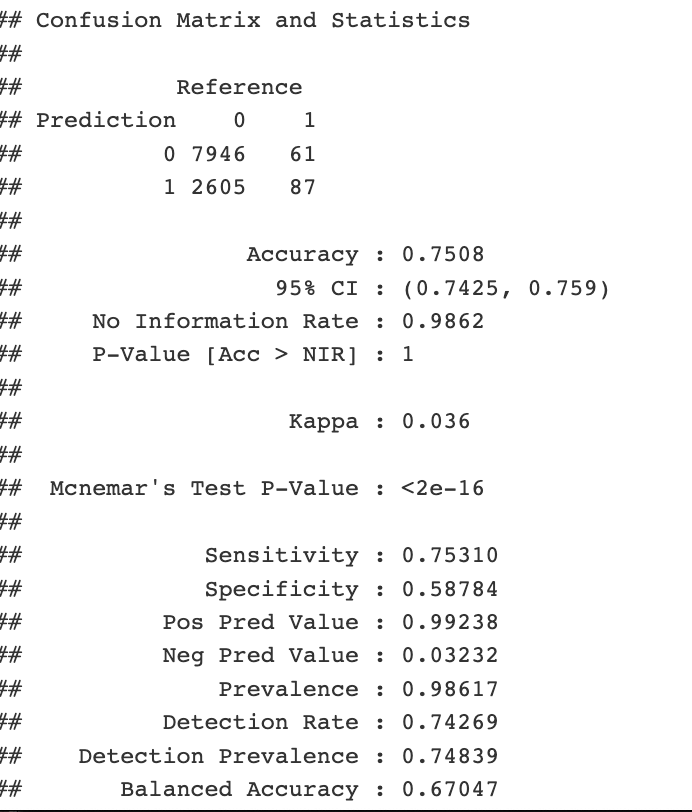
**Conclusion and Future Work (0.25-0.5 pages)**

Overall, this project investigated injury risk in association with several variables tracked during training by a Dutch Running Team. I was able to create a moderately accurate model for predicting injury risk, using skills developed during class. From these insights I was able to prepare an actionable plan for coaches, staff, and athletes to utilize to lower injury risk and become more mindful. Next steps for this project would be to analyze the athletes who are injured the most to find if there are more significant trends between those who are repetitively injured and those who are not. Another next action is to further develop the “Vulnerability to Injury” variable to a point where it can be presented to the team. We can also apply a dashboard, where athletes can input their perceived training success and automatically see their recovery and performance.

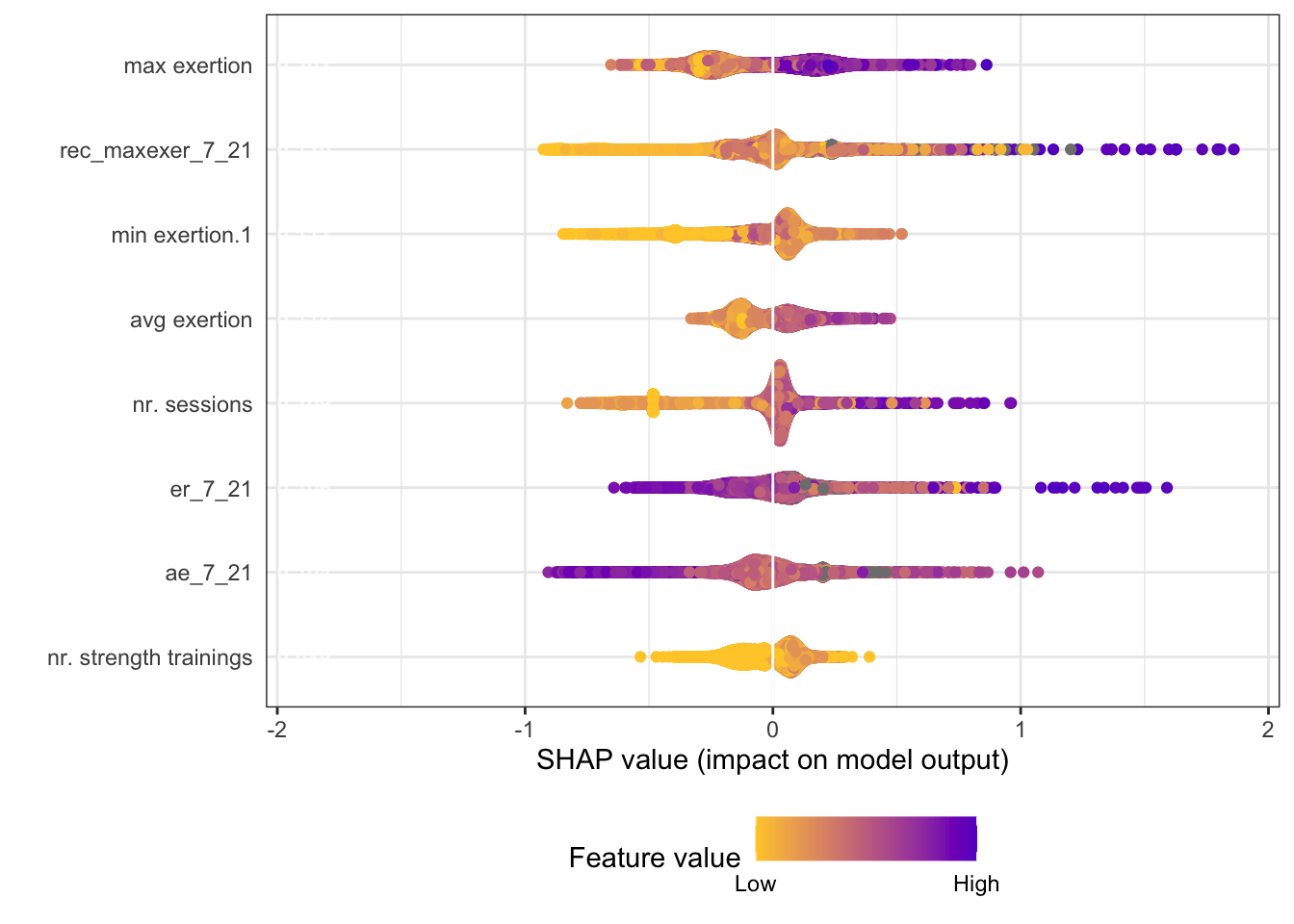
**Appendix**

****

**Figure 1. Athlete ID #26 Injuries and Metric Values.**

****

**Figure 2. Confusion Matrix**

**​​**

**Figure 3. SHAP Value chart weighted against injury.**

**Sources**

[1] S. Lovdal, Ruud J. R. Den Hartigh, G. Azzopardi, "Injury Prediction in Competitive Runners with Machine Learning", International Journal of Sports Physiology and Performance, 2020, in press. [https://doi.org/10.34894/UWU9PV](https://doi.org/10.34894/UWU9PV%7D)