Model Info Sheet for Detecting and Preventing Leakage in ML-based Science

**Section 1: Information about paper or report**

1) Author(s): Jeffrey A. Rothschild, Tom Stewart, Andrew E. Kilding, Daniel J. Plews

2) Title of the paper or report which introduces the model

**Predicting daily recovery during long-term endurance training using Machine Learning analysis**

3) DOI or permanent link to the paper or report (for example, link to arxiv.org webpage)

4) License: Under which license(s) are the data and/or model shared?

5) Email address of the corresponding author

Jeffrey.Rothschild@aut.ac.nz

**Section 2: Scientific claim(s) of interest**

6) Does your paper make a generalizable claim based on the ML model?

Our models aid the prediction of day-to-day recovery status among endurance athletes

7) Is the scientific claim made about a distribution or population from which you can sample?

Yes

If yes: (a) what is the population or distribution about which the scientific claim is being made?

Endurance athletes training ≥ 6 hours per week

(b) What is the sample used for the study?

Male and female endurance athletes aged 18 or older who train at least seven hours per week and use a smartphone app to track their dietary intake, heart rate variability (HRV), and sleep.

8) Does the scientific claim only apply to certain subsets of the distribution mentioned in Q6?

 Our model might not generalize to athletes not currently tracking the various metrics.

**Section 3: Train-test split is maintained across all steps in creating the model**

9) Train-test split type: How was the dataset split into train and test sets?

For group models, data were split into a training set (75%) and a testing set (25%). To avoid data leakage, all observations from a given participant were assigned to either the training or testing set, and preprocessing steps such as standardization and calculations for removal of highly correlated variables were performed only using the training set.

For individual models there were not enough data points for a test-train split, so cross-validation was used for hyperparameter tuning and accuracy metrics were calculated using 500 bootstrap resamples.

10) Are there duplicates in the dataset? If yes, explain how duplicates are handled to ensure the train-test split.

No.

11) In case the dataset has dependencies (e.g., multiple rows of data from the same patient), describe how the dependencies were addressed (for example, using block-cross validation).

All data splits (test-train and cross-validation) were performed so that a given participant could not be in both the training and testing sets. To address potential issues with autocorrelation, a process of Markov unfolding was used (described in the manuscript).

12) List all the pre-processing steps used in creating your model. For example, imputing missing data, normalizing feature values, selecting a subset of rows from the dataset for building the model.

Missing data was imputed at the individual level, prior to joining data into a grouped dataset. Multiple linear regression and nearest neighbor algorithms were used for diet and training measures, and median values were used for other variables

Normalization, the removal of variables with zero variance, and the removal of highly correlated variables was performed on the training data, using the recipes R package in the Tidymodels framework.

13) How was the train-test split maintained during each pre-processing step? If applicable, use a separate line for each step mentioned in Q14.

The test-train split is maintained throughout the process in the Tidymodels framework.

14) List all the modeling steps used in creating your model. For example, feature selection, parameter tuning, model selection.

* Highly correlated variables were removed from the training set
* Parameter tuning was performed on the 9 different algorithms using a workflow set in the workflowsets R package.
* Models were selected in two main ways:
	+ Group models – the tuned models were fed into the Stacks R package, which then created an ensemble using bootstrap resampling
	+ Individual models – the model (and hyperparameter set) with the lowest RMSE was selected as the chosen model

15) How is the train-test split is observed during each modeling step? If applicable, use a separate line for each step mentioned in Q16.

The test-train split is maintained throughout the process in the Tidymodels framework.

16) List all the evaluation steps used in evaluating model performance. For example, cross-validation, out-of-sample testing.

Group models – Accuracy metrics were established using the previously unseen test dataset

Individual models – Accuracy metrics were established using 500 Bootstrap resamples

17) How is the train-test split observed during each evaluation step? If applicable, use a separate line for each step mentioned in Q18.

Group models – Accuracy metrics were established using the previously unseen test dataset.

Individual models – Bootstrap resamples were used in lieu of a test-train split.

**Section 4: Test set is drawn from the distribution of scientific interest.**

18) Why is your test set representative of the population or distribution about which you are making your scientific claims?

Data were collected and analyzed using 40 athletes from this population.

19) Explain the process for selecting the test set and why this does not introduce selection bias in the learning process.

The test set was randomly selected from the initial dataset.

20) In case your model is used to predict a future outcome of interest using past data, detail how data in the training set is always from a date earlier than the data in the test set.

Although lagged values were included in the test set, the test-train split was performed at the participant level (meaning no participants were in both sets).

**Section 5:** **Each feature used in the model is legitimate for the task**

21) List the features used in the model, alongside an argument for their legitimacy. A legitimate feature is one that would be available when the model is used in the real world and is not a proxy of the outcome being predicted. You can also include this list in an appendix and reference the relevant section of your Appendix here.

|  |  |
| --- | --- |
| **Category**  | **Variables** |
| **Training, dietary intake, sleep duration** | These are variables that are routinely tracked by athletes and coaches and would be expected to have the largest influence on day-to-day recovery |
| **Engineered features** | We created 7-d moving averages of several factors relating to training load and dietary intake, with the assumption that accumulated training load (and/or energy deficit/surplus) could influence day-to-day recovery. We also calculated a sleep index (sleep duration x quality) based on other published findings. |
| **Subjective measures** | Athletes daily perceptions of soreness, life stress, and sleep quality were considered an important reflection of real-world practices by coaches. |
| **Non-exercise** | As objective recovery measures, we included resting HRV and resting HR (daily, change from previous day, and 7-d moving averages of each) |
| **Subject characteristics** | Participant ID, age, HRV app, sleep app, percentage of missing data, competitive level, primary sport, training age, body weight, day of the week. These could help to explain variation between athletes. |
| For outcome measures recorded in the morning (AM PRS, HRV change), any variables that occurred later the same day were excluded from the modeling (e.g., how someone felt during exercise that day, or their dietary intake that day). |