

Estimating and predicting athlete performance using Machine Learning

Keywords: performance estimation, performance prediction, modelling, machine learning

A great amount of aspects of cycling could benefit from the revolutionary solutions available nowadays thanks to data science. This advanced computational discipline could be used to model sorely complex physiological processes, such as that of the human body, thus might have a key role in physical performance. In this article, we introduce how Machine Learning algorithms are used to estimate the current physical abilities of an athlete (i.e. his maximal power output) or predict his performance along the next few weeks.

As pointed out onto Figure 1, the various features required for our Machine learning model are determined from already existing labelled data during a phase called *training*, based on an archive containing any cycling activity from a given rider. For each session, we consider several computed average metrics (power, heart rate, cadence...), as well as the ones quantifying the session workload and the level of shape of the athlete. Once the model is said to be trained, it eventually makes it possible to estimate the rider's current power profile.

Regarding *prediction*, the model is first trained over a dataset containing an archive of past sessions, compiled with the upcoming ones, planned by the athlete training program. This allows us to predict the peak power output of the athlete over the next few weeks.

In order to validate our approach, we used our algorithms to estimate and predict the maximal 10-minute peak power of a world-class athlete over 3 months. The results we obtained through our models are displayed on Figure 2. These results should be compared with the all-out test values (red dots) produced by the rider every week. The solid lines show power values estimations generated by two different algorithms, calculating the current shape of the rider (i.e. with models trained over past sessions only). The two models are based on different estimations of the workload and the level of shape. The basic estimation model (fuchsia line) used the Training Stress Score and the Training Stress Balance while the advanced estimation model (light blue line) is based on metrics normalized to the athlete. Therefore, those metrics vary between 0% and 100%. As can be observed on Figure 2, the advanced model is significantly more accurate than the basic one in approaching the values extracted from the all-out tests, standard deviations obtained with these models being quantified respectively at 8W and 22 W. Regarding predictions (blue dashed line), we take the sessions originally planned by the rider's training program into account in order to train our model. The figures obtained thanks to the advanced predicting model still match the values from the all-out tests pretty well, as pointed out by a tiny 10W standard deviation.

As shown above, the results obtained in this article highlight the accuracy and reliability achievable thanks to advanced data science, which proves to be the current best tool to estimate any current peak power output of a given athlete, and to calculate his performance during the near future. This approach sets the path of a great amount of applications, especially for determining and optimizing a custom training plan, allowing any athlete to reach his maximal level of performance at the very day he/she chose to.

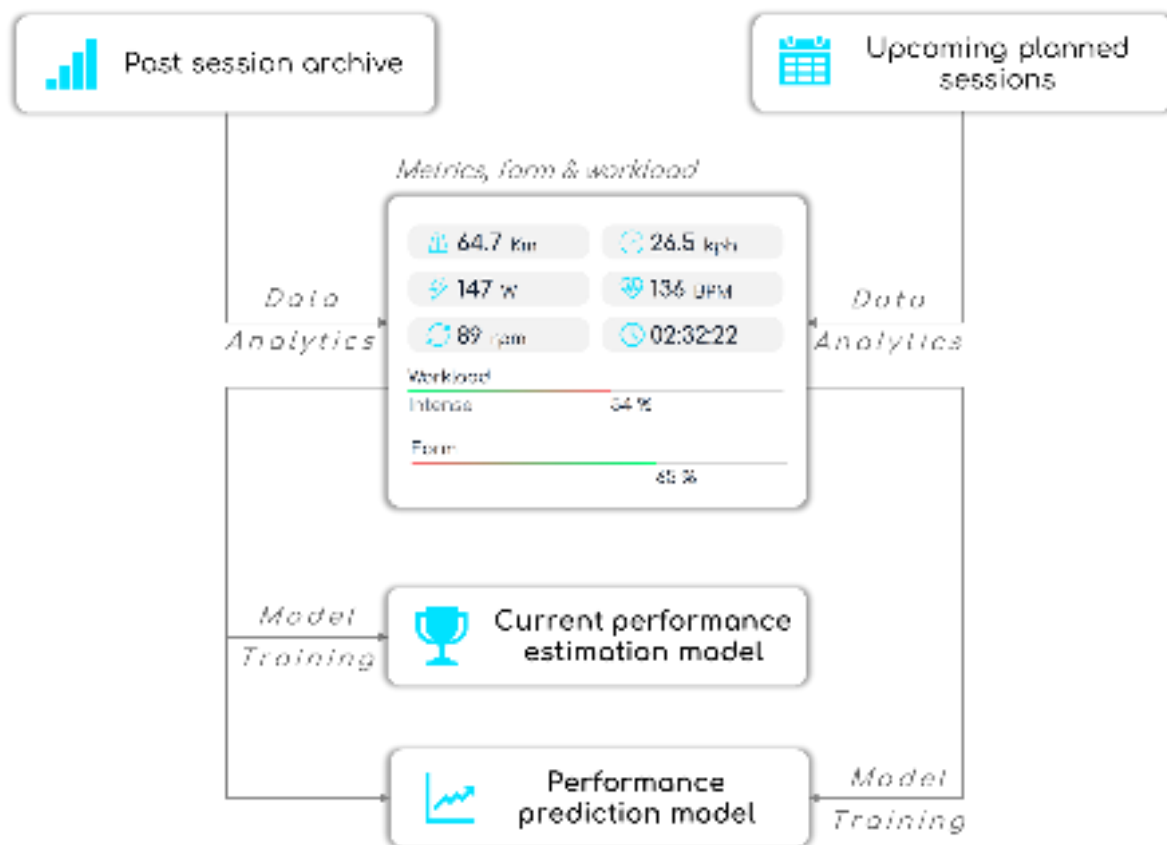


Figure 1 : Scheme detailing the main steps of the Machine learning models used to estimate and predict the level of performance of a given rider

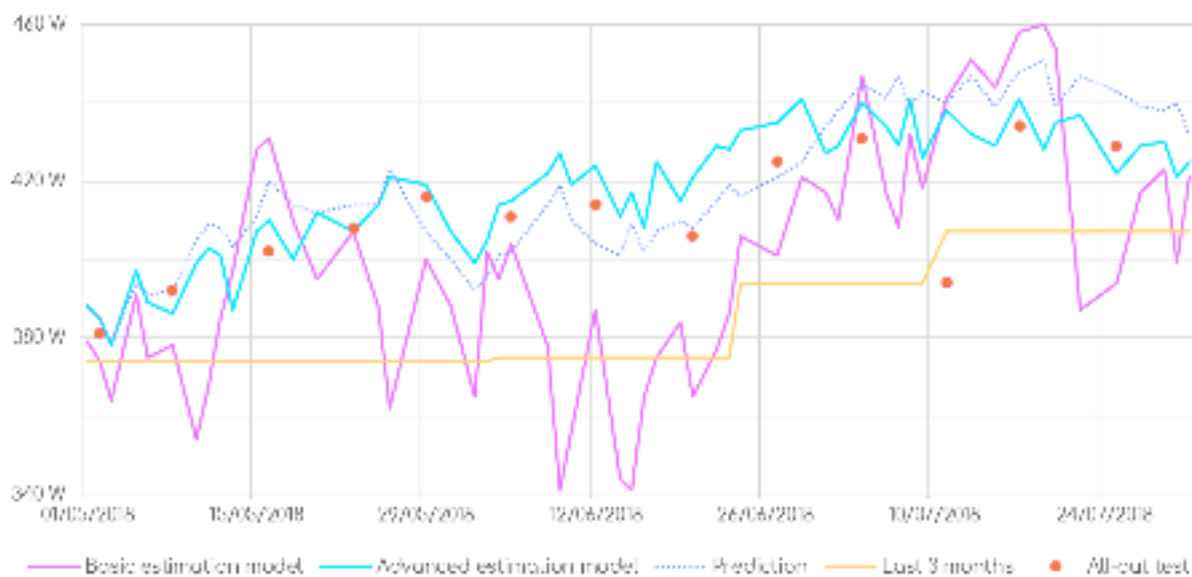


Figure 2: 10-min maximal power output measured and computed with different Machine Learning algorithms.