The International Trail Running Association (ITRA) is the world's largest trail running association. Its website contains records of more than 1.6m runners in races such as UTMB, Lavaredo Ultra Trail, Western States, Sierre-Zinal...

We develop ScrapITRA, a Python package for scraping, downloading, and formatting data of both runners and races from the website of ITRA. ScrapITRA is available at https://github.com/ricfog/ScrapITRA. How does ScrapITRA work? The package leverages Selenium and BeautifulSoup for dynamic scraping.

What do you obtain with ScrapITRA? Get data at
• runner level: demographics and results;
• race level: runners’ results and details on the trail.

Why? You can now analyze performance of more than 1m runners over the last 15 years.

Ultra-Trail du Mont-Blanc (UTMB) is the “holy grail” of ultra trail running. Starting in 2003, 2500 runners from over 100 nations gather in Chamonix (France) in the last week of August for a tough challenge: 171 km with more than 10000 m of elevation gain passing through France, Italy, and Switzerland (Figure 1). While elite runners typically complete the race in ~20 hours, most runners cross the finish line in more than 40 hours, right before the time barrier of 46.5 hours. Due to its toughness and thanks to the beautiful landscape, UTMB is seen by many as the pinnacle of a career in trail running.

Qualification to UTMB is based on a draft (last year in 3 chances of getting in) and a minimum ITRA score is required to get into the draft. The extreme conditions of the race make prediction tasks challenging.

Our models target three quantities:
(I) probability of dropping out (logistic regression, random forest, XGBoost);
(II) expected passage time (LASSO, random forest, XGBoost);
(III) prediction interval for passage time (random forest).

The models are fitted using four different sets of features (compared against intercept-only model):

Modeling Results

(I) Best model: XGBoost with Ckp+ITRA+Lag1 & 2 - LOYO CV AUC: 0.88, liftAUC: 3.33.

(II) Best model: Random forest with Ckp+ITRA+Lag1 & 2 - LOYO CV RMSE: 15.

Considerable improvement in model performance by including ITRA runner level information for tree-based models, capturing nonlinear interactions between runners and checkpoints.