TRAP: A Predictive Framework for Trail Running Assessment of Performance

Carnegie Mellon University
Statistics & Data Science

Riccardo Fogliato, Natalia Lombardi Oliveira, and Ronald Yurko

Carnegie Mellon University Department of Statistics & Data Science

International Trail Running Association

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...

Date	Name of the race	Distance / D+	Time	Rank.	Rank. M	Score
2018-08-30	Utmb® 2018 - Ooc	56km / 3450m+	07:31:38	133 / 1478	114	630
2019-07-27	Tushare Mountain Runs 2019 - Marathon	41.7km / 2250m+	06:47:01	18 / 74	16	562

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.

UTMB

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.

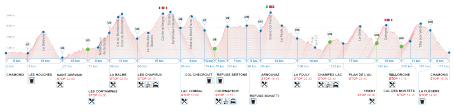
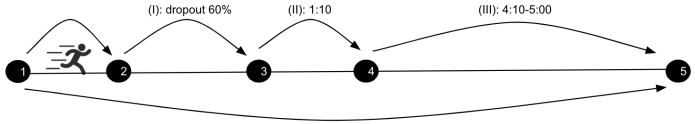


Figure: UTMB 2017 course map.

Qualification to UTMB is based on a draft (last year 1 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.

Overview of Modeling

We model the runner's performance at the checkpoint level: for each station, we output a prediction both for the following station and for the end of the race in Chamonix.



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):

Checkpoint-level:	ITRA information:	Lag1:	Lag2:
Distance, altitude, elevation, variation, food +	Gender, # races, mean ranking, mean distance +	Speed and time last checkpoint	Speed and time second to last checkpoint

We evaluate models with leave one-year-out (LOYO) cross validation (CV) for UTMB 2017-2018.

Modeling Results

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

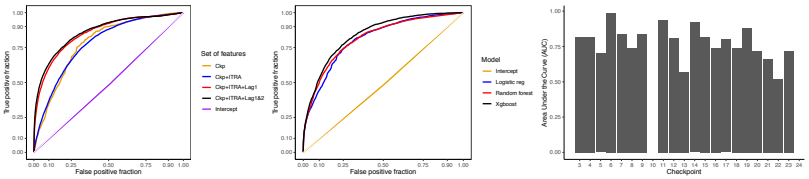


Figure: Left and center: comparison of ROC curves for features and models with features Ckp+ITRA+Lag1& 2. Right: AUC by checkpoint for xgboost and features Ckp+ITRA+Lag1& 2.

(II) Best model: Random forest with Ckpt+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.

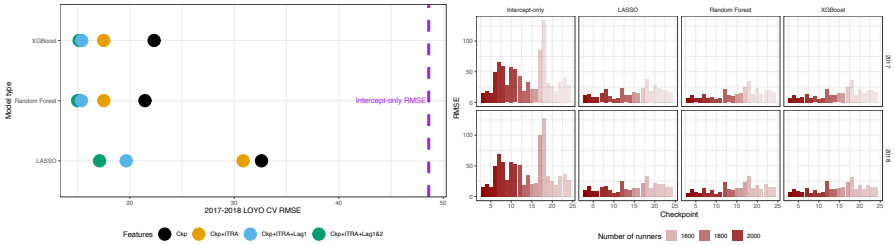


Figure: Left: comparison of LOYO CV RMSE curves for features and models. Right: LOYO CV RMSE by checkpoint and year for models using Ckp+ITRA+Lag1& 2.

EDA for UTMB 2017-2018 races

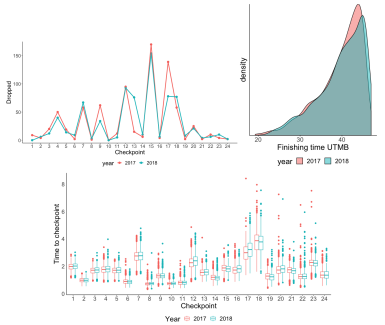


Figure: Top left: Number of runners who dropped out of the race at each checkpoint for 2017 and 2018 UTMB. Top right: Finishing time distribution among runners who finished 2017 and 2018 UTMB. Bottom: Time to checkpoint for 2017 and 2018 UTMB.

(III) Prediction interval results

Quantile regression via random forest

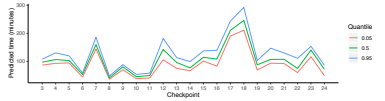


Figure: Prediction interval for 5%, median, and 95% quantiles for a randomly selected runner.

Discussion and Future Work

- dynamically adjust between checkpoint
- integrate models (I) and (III) for classification
- explore methodology for intervals, eg conformal
- **propose alternative to the ITRA score**
- test general framework on other races.

We hope that our seminal work might help the Data Science community gain interest in the (still unexplored) world of trail running. For this reason, we plan to release the ITRA data set on Kaggle.

References

[1] Chen et al. *xgboost: Extreme Gradient Boosting*, 2019. R package version 0.81.0.1.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, New York, 2009.

Corresponding Author: Riccardo Fogliato. Email: `rfogliat@andrew.cmu.edu`