Machine Learning of Pacing Patterns for Half Marathon

Johan Atterfors, Johan Lamm and Moa Johansson*

Chalmers University of Technology, Gothenburg, Sweden
{atjohan,lammj,jomoa}@chalmers.se

Abstract. Every year over 40 000 runners participate in Gothenburg Half Marathon, one of the world's largest half-marathons. Based on publicly available results data (423 496 entries) for ten years (2010 – 2019)[1], we investigate machine learning models for two tasks: prediction of finishing times and identification of runners risking hitting the wall. Our models improve results over the current baseline on finish time prediction and manage to correctly identify many of the runners who later hit the wall, although it also misclassifies many who do not.

Keywords: Pacing Pattern, Half Marathon, Machine Learning, Performance Analysis.

1 Introduction

Gothenburg Half Marathon is one of the worlds largest and attracts over 40 000 participants per year, most of which are recreational runners of varying fitness, age and ability. In prior work, we found that most of these runners could benefit from pacing themselves more evenly, and most lose quite a bit of time (on average 4 minutes) on the second half of the race [6]. Furthermore, between 8-9% of entrants will experience a dramatic slowdown, "hitting the wall", somewhere during the race, which also puts them at risk of collapse or injury, especially in hot weather [5]. Men seem to be twice as likely as women to hit the wall and were more negatively affected by high temperatures. Younger runners also ran a higher risk than the middle aged for both men and women. Here, we investigate if machine learning can be used to identify and, in the longer run, assist recreational runners adjust their pace depending on their fitness level and the temperature of the day.

2 Data and Methods

The dataset consists of 423 496 entries from runners (140 409 female, 283 087 male) who completed the race between the years 2010 - 2019, and includes in

* Johan Atterfors and Johan Lamm contributed equally to this work and are joint first authors. Moa Johansson supervised the work.

[1] https://reg.goteborgsvarvet.se/sok/resultatlista.aspx

addition to finishing time, sex, age and 5 km splits. Many runners are repeat participants, so the dataset contains 184 890 unique individuals. In addition we also include information about temperature of the day from the Swedish Meteorological and Hydrological Institute (SMHI) [2].

Our machine learning models were developed in Python using libraries SKLearn and TensorFlow [1, 3]. The source code developed is available online. [3]

Definition of Hitting the Wall. We adopt a pace-based operational definition of hitting the wall (HTW) [8], simply as a dramatic slowdown of over 25% on a 5 km segment, compared to a previous base-pace. The base pace defined as the average pace over the first 0-5km and 5-10km segments. Using this definition, 8.6% of all starts results in the runner hitting the wall (10.2% for men, 5,4% for women).

3 Predicting Finishing Times

Three methods for predicting finish time from passing intermediate split times are compared: first, the *baseline model* is currently used for Gothenburg Half Marathon live results, and simply predicts that runners will maintain their most recent 5km pace for the remainder of the race. Second, we fit a model that uses multiple *linear regression* to make the prediction according to least squares fit on historical data, and third, a small feed forward *neural network* consisting of one hidden layer with 40 nodes. For evaluation, we computed the mean absolute error (MAE), using 5-fold cross validation with an 80-20% random split averaging over all splits. Models were tested with each 5km intermediate split times as input, simulating how the runner progress through the race. We expect more accurate predictions as the runner approaches the goal. Results are shown in Fig. 1. Both the linear regression model and the neural network outperform the baseline at all intermediate splits with the neural network slightly outperforming the linear regression model at the 10 and 15km splits. Largest absolute improvements over baseline is found at the 10 km split where the the neural network has a 2:09 minute lower MAE than baseline.

Adding Additional Features. To further improve results, we experimented with adding additional input features to the neural network, based on factors seen to influence pace in prior work; age, sex and daytime temperature, and also prior finishing time and pacing for returning runners [6]. To avoid data leakage, we only predict performance for the 2019 race. Thus the resulting prediction values can not be directly compared to the method comparison analysis. Results are shown in Fig. 2. Adding more features slightly improve performance, especially for predictions early in the race, where combining all features results in a decrease of MAE by 33 seconds.

[2] https://www.smhi.se/en/services/open-data/search-smhi-s-open-data-1.81004
[3] https://github.com/JohanAtterforsStudent/PacingProject/tree/main/MachineLearningofPacingPatternsforHalfMarathon

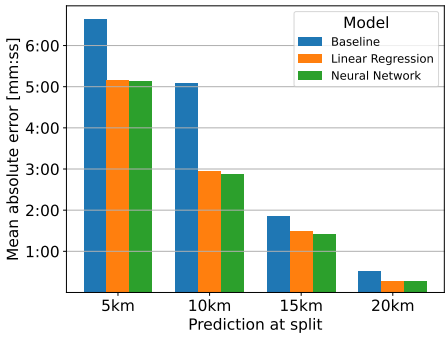

Fig. 1. Mean absolute error for the Baseline, Linear Regression and Neural Network model at all intermediate splits.

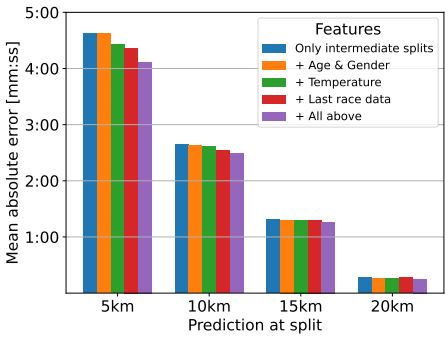

Fig. 2. Mean absolute error for neural network models with and without additional features

4 Who Will Hit the Wall?

We want to investigate if it is possible to predict which runners will hit the wall, before they do it. Based on features describing the runner, their history and how they have paced initially (age, sex, first two split paces, previous finishing time and number of previous races, split ratio and temperature), we compiled a subset of 238 567 data points, including only runners participating several times. This was divided into 90% for training and 10% for testing. We evaluated several standard machine learning algorithms got best performance from a Balanced Random Forest model [7], which is designed for imbalanced datasets such as this (recall, only 8,6% of runners HTW). The experimental results are shown in table 1. The model does a reasonable job: it identified 76% of the runners who actually hit the wall (recall 0.76), but at the cost of also labelling many who do not HTW as such (precision 0.19).

	Precision	Recall	f1-score
Avoid HTW	0.97	0.73	0.84
HTW	0.19	0.76	0.30

Table 1. Precision/recall analysis for trying to predict which runners will hit the wall

By computing the permutation feature importance score, we can measure which features were most important for the model. Unsurprisingly, the 5 and 10 km paces were the most important, followed by the previous finishing time, and third temperature. The rest of the features was given relatively little importance.

5 Discussion and Conclusion

Even simple linear regression and neural network models improve over the naive baseline used in many live results applications today. We get a large improve-

ment, especially for the first half of the race, which can be further improved by adding information about the runner (in addition to split times), such as the temperature, age, sex, and last year performance.

We have made a first small step towards developing a model for pacing assistance, to warn runners risking hitting the wall. While the model over predicts how many runners will HTW, we can get some insight in what additional features may be useful. As previous finish time, which approximate a runner's ability, was important, there is an argument to try to include historical training data, also lack thereof was identified as a risk in factor [4]. Temperature was also an important feature here, which is known to influence the proportion of runners hitting the wall [9, 6]. Surprisingly, the model did not place much importance on sex, despite results showing to that men are at increased risk of HTW [9, 4, 6]. This could possibly be because the share of female runners is already smaller (about 33%) and fewer of them HTW, so under-sampling the non-HTW runners reduce the share of females in the training data even further. Ideally, we would also want to include some information about the physiological state, such as heart rate of the runner. To measure this, data from a classical exercise watch would be useful, as it could also include more fine-grained pacing data, as in [2]. One may also reconsider the problem and try to apply a regression model to predict the slowdown instead of static class membership.

References

1. Abadi, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Berndsen, J., Lawlor, A., Smyth, B.: Exploring the wall in marathon running. Journal of Sports Analytics **6**(3), 173 – 186 (2020)
3. Buitinck, L., et al.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122 (2013)
4. Buman, M.P., Brewer, B.W., Cornelius, A.E.: A discrete-time hazard model of hitting the wall in recreational marathon runners. Psychology of Sport and Exercise **10**(6), 662–666 (2009). https://doi.org/10.1016/j.psychsport.2009.04.004
5. Carlström, E., Börjesson, M., Palm, G., Khorram-Manesh, A., Lindberg, F., Holmer, B., Berner, A., Örninge, P., Luning, H., Nilson, F., Gelang, C., Thorsson, S.: Medical emergencies during a half marathon race - the influence of weather. Int. Journal of Sports Medicine (2019). https://doi.org/10.1055/a-0835-6063
6. Johansson, M., Atterfors, J., Lamm, J.: Pacing patterns of half-marathon runners: An analysis of ten years of results from Gothenburg half marathon. Preprint: https://research.chalmers.se/en/publication/532586 (2022)
7. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research **18**(17), 1–5 (2017), http://jmlr.org/papers/v18/16-365.html
8. Smyth, B.: How recreational marathon runners hit the wall: A large-scale data analysis of late-race pacing collapse in the marathon. PLOS ONE **16**(5), 1–25 (05 2021). https://doi.org/10.1371/journal.pone.0251513
9. Trubee, N.W.: The effects of age, sex, heat stress, and finish time on pacing in the marathon. Ph.D. thesis, University of Dayton (2011)