Using Machine Learning to Build a Better Fitness App to Help Runners to Run a Faster Marathon

Barry Smyth
Insight Centre for Data Analytics
School of Computer Science
University College Dublin, Ireland
barry.smyth@ucd.ie

Abstract
We explore using machine learning to help marathoners achieve a personal best for an upcoming race, by helping them to select a goal-time and a pacing plan. We evaluate several representational alternatives, and algorithms, using real-world race data, to highlight the performance implications of different types of marathon histories and landmark races, concluding that richer representations do not always deliver better prediction performance.

1 Introduction
Mobile and wearable technologies help capture data about our activities, habits, and lifestyles, often seducing us with the potential to live healthier and more productive lives [13]. Certainly the application of AI to personal health and healthcare is not new [3, 4, 22, 23] but the always-on nature of smartphones and wearables is creating an even greater opportunity for novel preventative, proactive, and personalised interventions [8, 9, 11, 14, 15].

Mobile apps like Strava and RunKeeper record our daily activities, but they remain largely silent when it comes to proactively assisting users as they train, recover, and compete. This offers some exciting opportunities for AI, to support users with personalized training plans [5, 20], injury prevention advice, route/race/event recommendation, performance prediction and race planning [2, 6, 10], among others.

In this paper we focus on performance prediction and race planning for marathoners. We describe a technique for determining an achievable personal best (PB) time for an upcoming race and generate a tailored pacing plan to achieve this time. This is related to the general problem of race-time prediction, which plays an important role in many endurance sports. For example, [2] uses a finite mixture approach with partial race data to model the performance and strategy of runners in a 24-hour ultra race, to identify clusters of runners who differ in their speed and propensity to stop; see also [21].

Most relevant to this work is recent research by [17, 18], which also looked at PB prediction and pacing, using a single past race as the basis of prediction. We build on this approach, by harnessing extended race histories, involving multiple races. We evaluate certain types of landmark races, as key representational features, to show how some races can help prediction while others can be harmful; in fact we show that richer representations (involving more races) are not always beneficial. We evaluate our approach using race data from 25 major marathons in 3 different cities.

2 Defining the Problem
While the work of [17, 18] is best suited for novice marathoners, with just one recent race, in this work we will target more experienced marathoners, those with at least 4 previous races. We do this for two reasons: (1) novice marathoners are often focused on finishing the race, rather than achieving a PB; and (2) more experienced runners have a richer race history to exploit for prediction and planning. In this work we target two separate but related tasks: (1) the prediction of a suitable (achievable) goal-time and (2) the recommendation of an appropriate pacing plan to achieve this time.

2.1 Task 1: Predicting an Achievable PB Time
For a marathon runner, determining a challenging but achievable PB time is an important pre-race task. Choosing a time that is too conservative will leave the runner feeling unfulfilled, while an overly ambitious goal may ruin their race, increasing the likelihood that they will ‘hit the wall’. Thus, predicting a best achievable finish-time is non-trivial and getting it wrong can have a disastrous effect on race-day.

2.2 Task 2: Recommending Pacing Plans
Marathoners need to translate their goal-time into a pacing plan, for different segments of the race. Figure 1 shows a 5km segment paces for a 4 hour 13 minute finisher.
from a pacing plan that is matched to their goal-time, personal fitness level, and the course conditions and terrain.

2.3 An Example Use-Session

Figure 2 presents our app prototype, designed to provide PB advice and race planning. It shows the runner providing their race history (Figure 2(a)) to obtain their goal-time prediction and race plan (Figure 2(b)). Then, during the race itself the app provides real-time feedback by comparing the runner’s current pace versus their goal-pace (Figure 2(c & d)).

3 Recommending a Personal Best

Next, we describe how to generate PB predictions and pacing recommendations by extending the work of [17, 18] to harness multi-race histories of more experienced runners.

3.1 From Races to Cases

Our starting point is a marathon race-record, with a runner’s id, gender, city, date, finish-time, and the 5km segment-times/paces as per Equation (1). We use 5km segments because these data are usually provided by big city marathons. Segment paces are stored as relative paces – relative difference between the segment pace and the mean race-pace (MRP), see Equation (2), so that a relative pace of -0.1 means that the runner completed that segment 10% faster than their MRP.

\[ m_i = (r, gender, city, date, time, rp_5, ..., rp_40, rp_{Final}) \]  
\[ \text{relPace}(p, MRP) = \frac{MRP - p}{MRP} \]  

A runner r has a history of races, \( H(r) \) (Equation 3), and [17, 18] described one way to transform these race-records into PB cases by pairing a runner’s most recent race-record (MR(\( r \))) with a designated PB race-record, PB(\( r \)); that is, the race-record with the fastest finish-time, as per Equations 4.5 and 6. This allows PB cases to be used as training instances in an ML setting, for example, by using the MR(\( r \)) features to predict PB finish-times and pacing profiles in PB(\( r \)).

\[ H(r) = \{m_1, ..., m_n\} \]  
\[ \text{PB}(r) = \arg \min_i H_i(r).\text{time} \]  
\[ \text{MR}(r) = \arg \max_i H_i(r).\text{date} \text{ if } i \neq \text{PB}(r) \]  
\[ c(r) = \{\text{MR}(r), \text{PB}(r)\} \]

3.2 Identifying Landmark Races

Where [17, 18] stop short, is any consideration of richer race histories with multiple past races. From a representational perspective, extending single-races to variable length race histories is non-trivial. Here we adopt a race representation from a fixed set of specific landmark races. These races are chosen because they are likely to influence PB prediction and race planning. For the purpose of this work we identify the following landmark races (excluding their PB race):

- \( MR \) – The most recent (last) race in the runner’s history.
- \( LR \) – The least recent (first) race in the runner’s history.
- \( MV \) – The runner’s most varied; the race with the highest coefficient of variation of the segment paces.
- \( LV \) – The least varied race in the runner’s history; that is the race with the lowest pacing variation.
- \( PPB \) – The previous PB race; the runner’s fastest race in their race history, prior to the current PB.
• **PW** – The personal worst race; the slowest race in the runner’s history.
• **MNPB** – a pseudo race-record based on the mean of the runner’s non-PB races.

Each of these races corresponds to a specific race-record in the runner’s history, adding the usual set of features (year, finish-time, segment paces) to a PB case. For example, Equation (7) corresponds to a representation involving all of these landmark races, plus the PB. It is worth pointing out that multiple landmark races may correspond to the same race-record for a runner if, for example, the PW race is also the MV race.

\[ c(r) = \{MR, LR, MV, LV, PPB, PW, MNPB, PB\} \]  

(7)

Later, we consider representations using different combinations of landmark races, to determine whether some are more or less useful than others, from a ML perspective.

### 3.3 PB Prediction & Pacing Recommendation

We treat the task of determining a challenging but achievable PB time as a prediction problem, using the PB cases as training examples, with their PB time as the prediction target. In Section 4.2 we consider 3 common ML approaches and the resulting models can be used to make PB predictions for unseen runners.

Next, given a new PB prediction, we recommend a suitable pacing plan by selecting the \( k (k = 20) \) PB cases whose PB times are closest to the predicted PB time. These cases correspond to runners who managed to achieve a similar PB, to the one predicted for the current runner, and the assumption is that their PB pacing profiles provide a basis for the new pacing plan. This plan is generated from the mean relative paces of these \( k \) cases.

### 4 Evaluation

The key contribution in this work is to extend the ideas presented in [17][18] with enriched marathon representations. In particular we evaluate whether these representations help or hinder our ability to make accurate PB predictions and recommend high quality race plans and how do any improvements in performance manifest when it comes to runners with different levels of ability or expectations?

#### 4.1 Datasets

In this evaluation we use public race records from three marathon majors, the Berlin Marathon, the London Marathon, and New York City Marathon; see Table 1. From these data we can identify 170,000 runners who have completed at least 4 races (in a given city) to act as our target dataset of PB cases.

#### 4.2 Methodology

As mentioned, each PB case contains a PB part, using the features from the fastest race for the runner. This acts as the ‘solution’ to the problem we wish to solve. Each case also contains a ‘problem’ part. In [17][18] the problem part was limited to the features of a single (most recent) race (MR), which we refer to here as the baseline representation. We wish to evaluate using different landmark races in our representations and so each representation include this MR race plus one or more landmark races. In total there are 64 unique combinations of MR plus one of the 6 additional landmark races (LR, LV, MV, PPB, PW, MNPB).

We test 3 goal-time prediction algorithms – \( kNN (k = 20) \), Linear Regression (LR), and Elastic Nets (EN) — and for each algorithm-dataset combination we use a standard 10-fold cross-validation methodology to generate and test our PB predictions and pacing recommendations.

For each test instance, its problem part is used to generate a PB prediction, which is compared to the actual PB time of the test instance to compute a prediction error. Similarly, the recommended pacing plan (based on this prediction) is compared, segment by segment, to the actual pacing the runner ran during their actual PB race, to compute the similarity between the recommended and actual pacing.

In what follows, we will compare the prediction error and pacing similarity results for different combinations of algorithms and representations.

#### 4.3 Prediction Error vs. Representation Richness

First we will consider how the different landmark races used in our representations influence the prediction error. Are more races better than fewer races? Are some races more powerful predictors than others?

### Are Richer Representations Better Representations?

To test whether richer representations (more races) produce better predictions we compute the average prediction error for representations containing \( 1 \leq n \leq 7 \) landmark races. There is only 1 representation involving a single landmark race (MR) and there is only a single representation involving all 7 landmark races (\( MR_{LR\_MNPB\_LV\_MV\_PW\_PPB} \)). But there are multiple representations for other values of \( n \).

The results are shown in Figure 3 as mean prediction error versus the number of landmark races in a given representation. On average, richer representations enjoy better prediction accuracy. Also, female runners enjoy slightly lower error rates than their male counterparts; this is consistent with studies (e.g. [7][12][19]) which show female runners tend to have more consistent (and predictable) pacing than males.

### An Analysis of Landmark Races

Although richer representations tend to produce more accurate predictions, this does not mean that a specific representation with \( n \) landmark races will always beat a representation containing \( < n \) landmark races. For example, the best performing representation we found was \( MR_{LR\_PPB} \) (\( n = 3 \)) as shown in Figure 3(a)-(c), which presents the average prediction error (for all runners) for all 64 representations but
ordered by decreasing error; for reasons of clarity, we only label a subset of the representations on the x-axis.

Error rates vary from 6% (the baseline) to about 2%. By including additional landmark races we often see an improvement in prediction accuracy, but not always.

**Which Races Help the Most?**

Clearly, not all landmark races are created equally. To explore this, we compare the error rates for representations with a given landmark race to those without it. For example, to evaluate the utility of \( LV \) we calculate the error for all representations with \( LV \) (\( MR_{LV}, MR_{LV_{MNPB}}, ..., MR_{LR_{LV_{MNPB_MV_PPB_PW}}}, ..., MR_{MR_{MNPB_MV_PPB_PW}} \)) and compare this to the error of representations without \( LV \) (\( MR, MR_{MNPB}, ..., MR_{MR_{MNPB_MV_PPB_PW}} \)). Then, we can calculate a relative change in the error due to the presence of \( LV \), the **benefit**, such that a **positive benefit** means including \( LV \) tends to reduce/improve the error compared to excluding \( LV \).

Figure 3 presents the average benefit scores for each of the 6 optional landmark races (\( LR, LV, MNPB, MV, PPB, PW \)). The results are broadly similar, regardless of algorithm or city, with some minor differences between male and female runners. The \( PPB \) landmark race (the runner’s prior personal best race) stands out as the most useful to include in a PB case; \( PPB \) tends to improve the error by at least 30% and sometimes by up to almost 50%.

Conversely, races such as \( LV \) (least varied) and \( PW \) (personal worst) do not appear to help prediction. This is probably because these types of races are unlikely to be representative of the runner’s true ability. For example, a runner’s worst race (\( PW \)) will typically be an outlier. The explanation for \( LV \) races it less obvious, since the conventional wisdom asserts that well-trained and disciplined runners tend pace their races evenly [1]. However, this doesn’t mean all evenly paced races are run by well trained, disciplined runners. Runners who are “taking it easy” on the day, or running a “tune-up” race, will often run a more a more evenly paced race, which is not representative of their true ability, and so doesn’t help when it comes to prediction.

**4.4 Prediction & Recommendation in Practice**

In this section we present a side-by-side comparison between the baseline \( MR \) representation used in \[17, 18\] and our ‘best’ performing \( MR_{LR_PPB} \) representation, to determine how both perform in terms of prediction error and pacing similarity with respect to runner ability and PB improvement levels.

**On Runner Ability**

The first two columns of Figure 6 show the prediction error and pacing similarity for the baseline \( MR \) representation used in \[17, 18\] and our ‘best’ performing \( MR_{LR_PPB} \) representation, to determine how both perform in terms of prediction error and pacing similarity with respect to runner ability and PB improvement levels.
Figure 4: The average PB prediction error for different representations and algorithms for Berlin, London and New York.

Figure 5: The benefit for landmark races for Berlin, London, and New York and for kNN, linear regression, and elastic net algorithms.
Figure 6: The errors and pacing similarities for Berlin, London, and New York. In each graph, results for male and female runners are shown for the baseline (MR), in grey, and the best performing MR LR PPB algorithm, in green (Berlin), blue (London), and red (New York).

able) runners who are likely to need them most. For instance, 270-minute, male finishers in London can expect an error rate of about 12%; that’s a potential margin of more than 30 minutes for their PB.

The results for the best representation (MR LR PPB) offer a significant improvement across all levels of ability for men and women. Now, our 270-minute, male, London finisher can expect a PB prediction with an error rate of less than 3%, a 4x improvement on the baseline. Moreover, the pacing plan similarity results in Figure 6(b, e, h) indicate that this improved prediction accuracy is available without any material loss of pacing plan quality, compared to the baseline.

On the Degree of PB Improvement
The third column of Figure 6 examines the relationship between prediction error and PB improvement. Some runners achieve a PB that is a big improvement on their most recent race, while other PBs might be more modest. In very small and very large PB improvements were associated with less accurate predictions. This is also evident in Figure 6(c, f, i) for the baseline, but, once again, our best representation reduces this error across all levels of PB improvement.

5 Conclusions
The main contribution of this paper is a novel application of AI to support marathoners with targeted PB advice, by suggesting a suitable goal-time and a pacing plan for an upcoming race. We extend the work of by using richer race representations to demonstrate significant improvements in prediction performance, across all levels of ability, and without compromising race-plan quality. This makes the approach especially useful to recreational runners, who represent a majority of participants. Moreover, we evaluated the relative merits of different types of landmark races to serve as predictive features, confirming that richer representations do not always deliver better prediction performance.

As always, this work has its limitations. A focus on more experienced runners (>3 races) technically excludes novice marathoners, for instance. However, experienced runners are likely to be more motivated to run a PB, and so more likely to use the system. In the future we will examine whether it may be feasible to accommodate novice runners by substituting in more common races (half marathons, 10k’s etc.) in place of missing marathons.

Another limitation is that, although the evaluation is tested with real race data, we did not test ‘live’ predictions, to determine whether runners actually run better races based on the app’s advice. We have made arrangements to address this by testing the approach with runners of the upcoming London Marathon in 2018.

Finally, our future plans include working with more fine-grained race data (<1km segments), rather than 5km segments
and different types of activities (triathalons, cycling, swimming). We will also apply these ideas to other tasks, such as injury prevention, personalised training, recovery advice etc.

References