15TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
APRIL 8-9, 2021 | VIRTUAL
presented by ESPN 42 ANALYTICS

I Think We'll Go to Boston - Marathon Performance Prediction

Varun Pemmaraju, Strava, varun.pemmaraju@gmail.com
Dave Hoch, Strava, davehoch12@gmail.com

1. Introduction

Establishing a link between performance improvements and training is the holy grail of sports science. This question captivated the Cold War era physiologists who developed the concept of periodization to aid their Olympic athletes; it was studied and practically applied by a British doctor who redefined what was humanly possible by breaking the four-minute mile barrier[1]. More recently, it was studied by Nike as they searched for every last marginal gain in the pursuit of breaking 2 hours in a marathon[2]. It is this endeavor - the marathon - that is of special interest.

The marathon has a humble origin, albeit one steeped in folklore, as it commemorates the supposed distance a messenger ran during the Battle of Marathon in 490 BC between the ancient Greek and Persian kingdoms. The unforgiving nature of marathons was evident even then, as Pheidippides the Messenger apparently collapsed and died upon reaching his finish line. Two millennia later, marathons are a universal and addicting endeavor, having attracted 20 million people worldwide[3] to a start line in the last decade. Endurance running is in many ways an innately simple sport - from an evolutionary perspective, humans as a species are well-adapted to it compared to almost all other animals[4]. It is also a relatively egalitarian sport as it is the only one in which professionals, amateurs, celebrities, moms and dads, and challenged individuals all share the same "playing field" at the same time. A particularly fascinating aspect is that there are numerous world-class athletes - Jimmie Johnson, Apolo Anton Ohno, Caroline Wozniacki, Lance Armstrong, and Pat Tillman to name a few - who crossed over from their primary sport and allow us to observe how their abilities translate to running.

This democratic nature has caused marathons to captivate an audience far beyond just medal-seekers, including physiologists, sports psychologists, nutritionists, and even historians and philosophers. Many have tried to understand what factors contribute to a good marathoner's performance. However, the insights they have drawn mostly come together as a result of smaller-scale studies in the laboratory on individuals. No one has yet studied this question in a way that takes advantage of both the granularity of available data and the enormous addressable population (not even FiveThirtyEight![5]).

We thus set out to do the following:

1. Predict marathon performance
    - Across a broader population than ever before studied
    - Using a volume and variety of data that has not previously existed
    - By deriving novel features from an extremely granular dataset, which allows us to more holistically describe an athlete's training - especially around specific terrain, intensity, and normalized aerobic efficiency

    - By utilizing those features to train a gradient-boosted regression model that makes predictions

2. Uncover trends in marathon performance amongst various athlete populations based on age, gender, and finish time range
3. Determine the most important features to preparing for a marathon, and see if the training principles corresponding to those features agree or challenge conventional wisdom
4. Demonstrate that these findings enable us to prescribe training that is personalized to an athlete's background, previous performance, and goals

2. Methodology

2.1. Dataset

We assembled the following dataset from Strava, a fitness-tracking app used by 70 million athletes worldwide:

- 356,618 marathons from 2016-2020 (including almost 40% of the entire 2019 Boston Marathon field![6])
- The preceding 6 months of an athlete's training before their race, totaling 47 million activities
- For each activity, the raw "streams", or arrays of metrics for GPS coordinates, speed, moving and resting times, elevation, heart rate, and step rate. These metrics are generally sampled by recording devices at a rate of 1 reading, or "point", per second, and results in 72 billion total points.

2.2. Feature Determination and Derivation

Using our combined knowledge of marathoning and a perusing of some training literature, we postulate that the following broad training principles are descriptive of an athlete's training and should be encompassed by our choice of features:

A. *Training More*: Up until some point of diminishing returns and within some bounds of reason, those who train more are going to perform better

B. *Consistency:* Runners who are more consistent day-to-day and week-over-week, either because they do not get injured or better adhere to their training plan or reach the start line closer to their full potential

C. *Training Cycle Length:* Physiological gains are accumulated over an entire lifetime of miles run, but common wisdom (and some intuition and experience) supports that a marathon-focused training cycle should be around three to four months. Shorter periods may not allow full adaptation to training, and longer periods may cause burnout or injury.

D. *Long Runs:* Long runs, designed specifically to increase endurance, are especially critical for the marathon. That said, it is rare to do the entire 26.2 miles distance in practice because of the physiological stress it causes. Regardless of ability, the human body can generally sustain itself without nourishment on glycogen stores for 90 to 120 minutes[7]

E. *Workouts:* Structured interval sessions that alternate between higher and lower intensity periods induce fitness responses to training more than just steady running

F. *Threshold:* The energy systems required to run produce lactate in the blood, which is what causes the burning sensation in muscles. Runners who are more efficiently able to process and remove this lactate can hold the same speed for longer. The speed or heart rate that an athlete can hold for 1 hour all-out is called a *lactate threshold (LT)* and closely approximates that tipping point.

G. *Running Economy:* The exertion, or energy it takes to run at a given pace, which is one of the reasons that two athletes with otherwise similar biometrics for lactate threshold or VO$_2$Max (the maximal oxygen an athlete can uptake to fuel exercise) may perform vastly differently

H. *Terrain:* Running at a fixed pace has different energy costs depending on the elevation gradient, altitude, and terrain. Athletes may seek out hilly courses to build strength or practice for a specific race course.

I. *Demographics:* As with many sports, athlete demographics such as age, gender, and weight come into play in ways that are often unclear, but can be significant

Keeping those previously discussed facets and these principles in mind, we then needed to transcribe these high-level principles into more specific features for our machine learning model. The following sections describe the key features, and Appendix B contains the full list of 41 features used.

2.2.1. Training Volume

Training Cycle Length caused us to attempt to train the model using data from the previous 3 months, 6 months, and 12 months of an athlete's data from the date of their marathon. The two longer time frames provided better results but were not themselves that distinguishable. We thus went with a 6-month window as this provided a larger sample size of athletes who had 6 months of training data.

We then determined the following features:

- Average weekly distance, moving time, and elapsed time (*Consistency*, *Training More*)
- Total active days and weeks (*Consistency*)
- Average longest run and 2nd longest run (often called a "middle-long" run) each week in both miles and minutes (*Long Run*)
- Number of runs over 90 minutes and 120 minutes (*Long Run*)
- Number of interval sessions (*Workout*s)

Appendix A lists some nuances about how we aggregated training volume.

2.2.2. Gradient-Adjusted Distance (GAD) and Pace

We needed a way to account for the absolute elevation and changes within a run (*Terrain*). One of the benefits of having the full-fidelity streams of data for distance and elevation is that we can determine gradient-adjusted distances[8] for runs and subsections of those runs. Gradient-adjusted distance (GAD) can be thought of as the effective distance of a course taking into account the elevation changes, where a course that is harder than the baseline flat distance will have a longer distance (and consequently, for the same effort corresponding to a fixed pace on flat ground, would take longer to cover).

For example, see the following examples of runs of the same length with varying elevation gain and loss, as well as varying gradients to accumulate those gains and losses. One could speculate that:

- The straight uphill course is the slowest
- The straight downhill is fastest, as long as the gradient is not too steep
- Of the bottom 3 (which have the same total loss and gain), the one with a series of shorter climbs is probably covered fastest

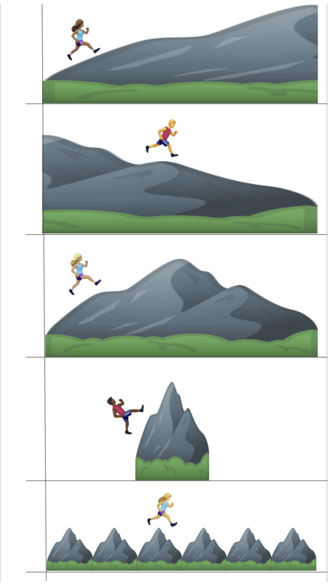


Figure 1 - Depiction of various courses of the same length but different gradients

Having GAD at our disposal is extremely powerful as it allows us to quantify those speculations and normalize both training activities and races themselves with respect to the elevation profiles, which would not have otherwise been possible.

2.2.3. Best Efforts over Fixed Time/Distance

An input into most other marathon prediction models is prior race times. Due to the physiological toll of marathons, often athletes will run a half marathon in their build-up to measure progress. For a prediction model, this best half marathon time is inputted, and often the athlete will be asked to qualitatively self-rate the conditions between "flat/easy and hard/hilly"[5]. This seems to add a constant scaling factor to their half marathon time, which is inputted into the prediction model. Our model, however, can use the exact ratio between the GAD and real-world distance to normalize an athlete's effort to what it would have been on a flat course.

Furthermore, we do not have to rely solely on discrete race efforts to get a sense of an athlete's "best-effort" fitness leading into a race. With the full activity distance and time streams, we can find the best sub-stream (or section of the run) of various fixed distances (gradient-adjusted or normal) and times, and then determine what an athlete's single best time was over the training cycle. Thus an athlete does not have to have done a half marathon or even a prior race at all for the model to be applicable. We can instead generate a "best-effort curve" for every athlete and use these as inputs into the model.

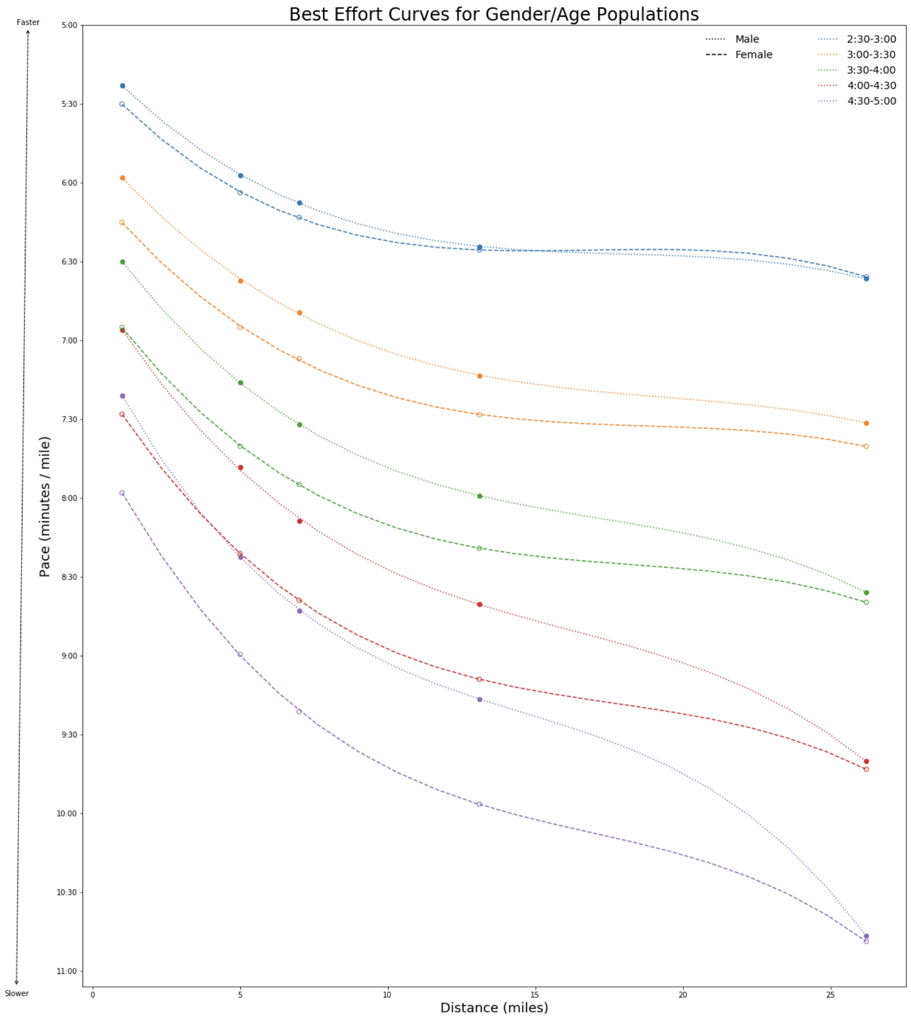


Figure 2 - Best Effort Curves of pace vs. distance for various gender/age populations

The average curves for various gender and finish time populations alone have interesting insights. For example, as distances get longer, the gap between men and women closes. There is research[9] that supports this hypothesis, theorizing that it has to do with testosterone in men being utilized for greater bursts of strength. In contrast, increased body fat percentages or even pain tolerance serves women better in longer endurance events.

These curves themselves tell us a great deal about the athlete, and thus people have attempted to tabulate with far less data in the past[10, 11]. Comparing individuals against the curves can tell us who has more natural speed, and who may be more blessed with or focuses more on endurance, which may benefit them especially in the marathon.

2.2.4. Specific Race Course

Not only do we utilize GAD for the training leading into an athlete's marathon, but also for the race itself. This allows us to predict how athletes would perform on a standard, completely flat course. Even more interestingly, at least to athletes themselves, it allows us to predict their time on any course *(Terrain)*. We no longer have to hypothesize or rely on forums to determine whether New York is faster or slower than Boston, or if you can beat your friend's time in Berlin, should you choose to run there.

We show the GAD of the 6 World Marathon Majors as well as 2 additional ones from our case studies below. For interpretability, we assign each a normalized course difficulty, as well as show what a 3:00 marathoner on a flat course would expect to run on that course if scaled by the difficulty. Note that the difficulty here does not encompass where in the course the difficulty may arise - for example, the hills are backloaded (from miles 16-21) at Boston which comes on tired legs and is much worse than coming at miles 1-6. It also does not encompass the weather, which is notoriously fickle in Boston in April compared to Berlin in September.

$$Course\ Difficulty = Course\ GAD/Marathon\ Distance\ (26.22\ miles)(1)$$

Course	Cumulative Elevation Gain (meters)	GAD (miles)	Difficulty	Equivalent 3:00 Marathon
New York	290	26.471	1.0096	3:01:44
Boston	247	26.446	1.0087	3:01:33
London	65	26.295	1.0029	3:00:31
Berlin	84	26.285	1.0025	3:00:27
Tokyo	41	26.255	1.0014	3:00:15
Chicago	35	26.236	1.0007	3:00:07
Twin Cities	167	26.38	1.0061	3:01:06
Santa Rosa	118	26.33	1.0042	3:00:46

Figure 3 - Gradient-Adjusted Distances and Course Difficulties for World Marathon Majors and select other courses

This course-comparison is perhaps even more salient this year with COVID causing the entire global race community to virtual racing on self-designed courses.

2.2.5. Practicing Race Pace

In conjunction with *Running Economy*, it is widely accepted that practicing at marathon pace, or MP, helps the body adapt to running at that pace. We thus calculated a few related but distinct features:

- Longest distance/GAD traversed in a single continuous stretch where your pace always remained within a threshold (only slightly faster or slower) of MP - i.e. doing a half marathon at MP would be 13.1 miles
- Cumulative distance/GAD traversed where your pace remained within a threshold of MP - i.e. doing a 9 mile run as 3 intervals of 2 miles at MP with a recovery mile in between each interval would be 6 miles
- Longest distance/GAD traversed in a single continuous stretch where your average pace over the entire stretch was strictly faster than MP - i.e. doing a 13-mile run where you had 3 miles of warmup and cooldown at slower than MP, and the middle 7 miles were significantly faster than MP. If the net average for miles 1-10 was faster than MP but the 11th mile made it slower, this would be 10 miles
- We also calculated the above features for 90% and 95% of MP, as training programs often prescribe runs of those paces as ways to get used to running closer to race pace but being able to go for longer durations without as much physiological toll

For prediction, this class of features presents a bit of a chicken-and-egg issue. While training the model, since we have the outcome MP in the dataset it is possible to determine these features. However, when making predictions the outcome MP is not known, as that is exactly what we are trying to predict. Our solution at test time is to calculate the features using an athlete's previous best or a ballpark estimate.

2.2.6. Workouts

Workouts are intended to practice at paces close to or even faster than race pace without actually inducing the toll of a full-blown race. They broadly will have a structure of faster sections broken up by slower recovery sections. Most recording devices have "lap buttons" for athletes to indicate and record these sections separately, though some devices are set to "auto-lap" every mile or kilometer. Strava also allows athletes to "tag" a workout, though not all athletes utilize this feature. We thus augmented the number of workout "tags" by clustering the distances and paces of the sections indicated by the lap button and then classifying runs as workouts. For example, if a run had laps of varying distance - say 1 mile at faster paces followed by 3 minutes of slower running - we hypothesize that the workout was mile repeats and not just a normal run.

2.2.7. Aerobic Efficiency - Heartbeats per Mile (HBPM)

Threshold and *Aerobic Efficiency* both are intrinsically tied to heart rate and specifically lactate threshold (LT). In a lab, this can be measured as the distance an athlete can cover in 1-hour all out. However, not every athlete will have such an effort and so we need a different reference.

We observed that for a single athlete, their fitness improved as they required fewer heartbeats to traverse a mile (termed Heartbeats Per Mile or HBPM). Between athletes of even similar fitness, however, this number will vary due to some athletes having higher heart rates than others[12], which we must account for.

We first looked through an athlete's corpus of heart rate data and determined what their LT might be, using both percentages of heart rates sustained for shorter windows (such as 30 minutes) and highest heart rates sustained over a full hour. We then determined the distance an athlete was able to cover in a fixed number of heartbeats; this fixed number is the total number of heartbeats if they were at LT for one hour.

As an example, let us take two athletes with identified LTs of 180 and 150 beats/minute respectively, who are equally fit and can hold 6:00/mile at their threshold heart rate.

$$LT1 = 180\ beats/minute$$
$$At\ 6{:}00/mile, HBPM = 1080\ beats/mile$$
$$1\ hour\ @\ LT = 180\ beats/minute * 60\ minutes = 10800\ heartbeats$$

$$LT2 = 150\ beats/minute$$
$$At\ 6{:}00/mile, HBPM = 900\ beats/mile$$
$$1\ hour\ @\ LT = 150\ beats/minute * 60\ minutes = 9000\ heartbeats$$

Each athlete is allotted this budget of heartbeats. We now find every subsection of an athlete's runs that encompass this number of heartbeats, regardless of how much time it takes, and calculate the distance covered by that subsection. That subsection could be an hour all-out at 6:00/mile, or perhaps a longer period at a lower heart rate and slower pace that results in a lower HBPM and longer distance covered. For athlete 1:

$$At\ 6{:}00/mile, 180bpm => 10800\ beats\ takes\ 60\ mins => 10\ miles\ covered$$
$$At\ 7{:}00/mile, 180bpm => 10800\ beats\ takes\ 60\ mins => 8.6\ miles\ covered$$
$$At\ 7{:}00/mile, 150bpm => 10800\ beats\ takes\ 72\ mins => 10.3\ miles\ covered$$

We look for the athlete's longest distance covered during all activities as this indicates their peak aerobic efficiency. Athletes who cover greater distances are more aerobically efficient.

Due to privacy regulations, we are unable to process all athletes' heart rate data for research purposes without first obtaining consent. For the small sample of athletes we were able to use it on, the HBPM metric showed a lot of promise as a feature in improving the model. Thus, we intend to follow the proper processes that would allow us to use this valuable feature as an input for all athletes with heart rate data.

2.3. Machine Learning Model

Having determined the above features as important, we first needed to materialize them from the raw data. Extracting the best efforts over time and distance would require multiple passes at the 72 billion point dataset - we thus used a cluster-computing framework called Spark[13] which allowed us to parallelize the feature extraction. We also recognized that predicting a marathon time had been distilled into a regression problem from a large number of both continuous (i.e., average mileage) and categorical (i.e gender) features and Spark also comes with fairly robust machine learning support. We chose to use a gradient-boosted framework called XGBoost[14], as it is easily run in a distributed framework and has gained increasing popularity in machine learning circles. Moreover, the math behind XGBoost allows for a somewhat quantitative understanding of feature importance to the model, which we knew would be useful for the prescribed training suggestions we desired to produce.

We also split our data into a training, validation, and test set and used GridSearchCV[15] to tune the model to avoid overfitting. While training the model, we optimized for the Root Mean Squared Error (RMSE), as this conveys confidence that is expressed in the same intuitive units as the prediction variable - in this case, estimated finish time in minutes.

Appendix C lists the hardware used and XGBoost parameters chosen.

3. Results

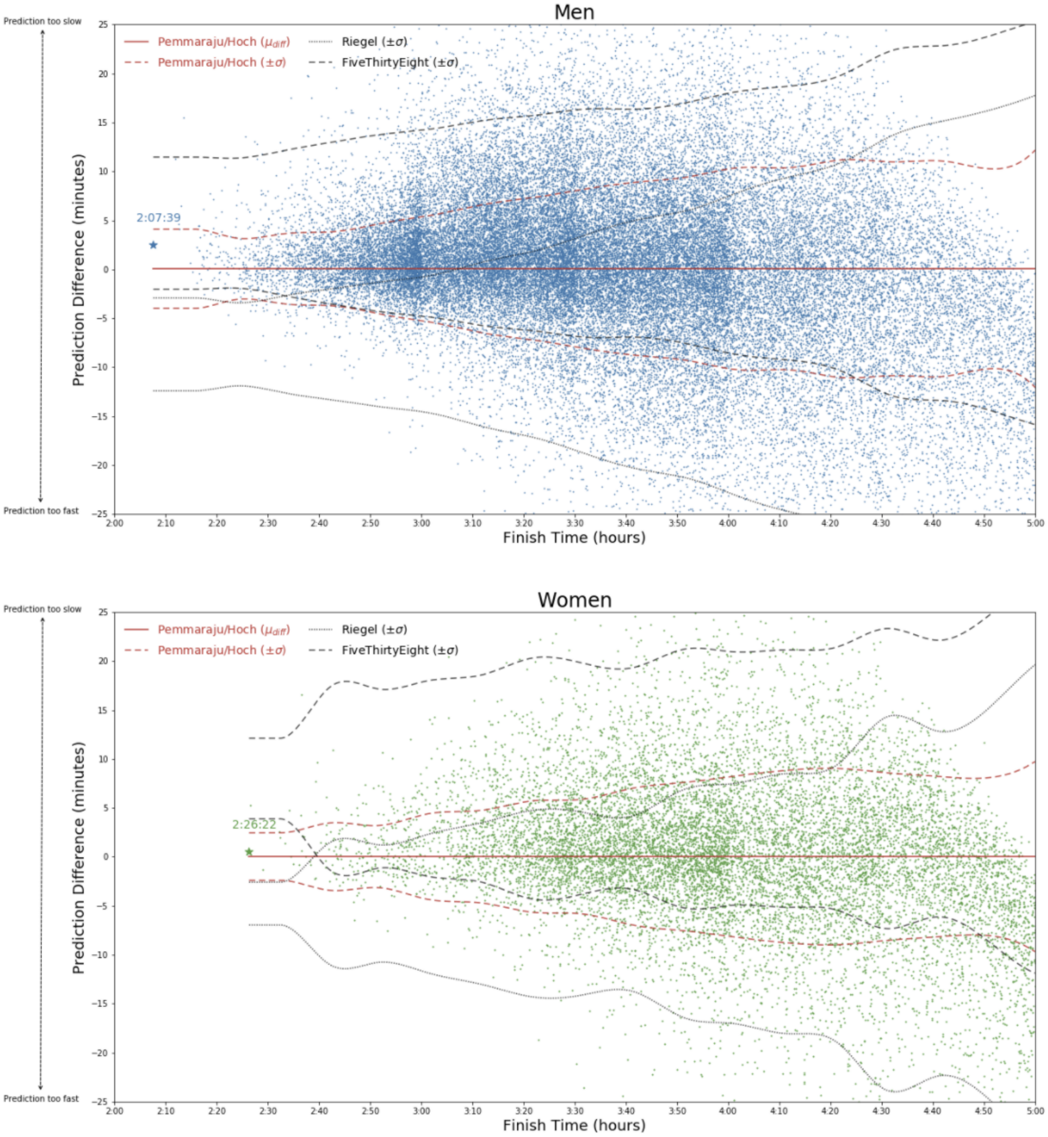


Figure 4 - Actual finish times vs. Differences from Prediction on Strava dataset

Population **(from test set unless noted)**	**Population Size**	**Average Finish**	**Mean Error**	**RMSE**	**MAPE**	**R^2**
Training Set (70% of total data)	293,685	3:44:01	0:01	7:13	2.2%	0.95
Test Set (15% of total data)	62,933	3:43:56	0:03	8:46	2.7%	0.93
Women	12,443	4:01:18	0:02	8:03	2.3%	0.93
Men	49,105	3:39:27	0:03	8:57	2.8%	0.93
Boston Qualifiers	11,176	3:04:18	1:59	5:42	2.0%	0.91
Riegel + Runner's World[16]	-	-	-	19:52	-	-
Vickers/Vertosick[17] + FiveThirtyEight	721	3:38:14	-	14:25	-	-
Altini/Amft[18] (results only reported for 10km and extrapolated)	2,113	-	-	-	4.0%	0.87
Smyth/Muniz-Pumares[10]	25,000	3:39:05	-	-	7.7%	0.69
Riegel + Runner's World[16]	Using Strava Test Set		-7:03	16:55	4.6%	0.75
Vickers/Vertosick[17] + FiveThirtyEight			5:24	15:02	4.6%	0.80

Figure 5 - Comparison of distributions of various populations. Reported results from other studies are shaded red. Studies that have published models that we ran against our own test set are in blue.

3.1 Predicting Marathon Times

With our output model in hand, we then made predictions for the entire test set. We were also able to make predictions against the same test set using models we found in literature, including one from Runner's World[16] developed by Peter Riegel and another done more recently by a collaboration between Slate and FiveThirtyEight[5]. We provide both a graphical and tabular summary of our results measured against prior work and show various measures of accuracy, including Root Mean Squared Error (RMSE), Mean Absolute Percent Error (MAPE), and Regression Value (R^2).

Previous models have generally been linear regressions with simple features like total volume and prior races. Processing a dataset composed of raw points enables us to derive novel features that more holistically characterize an athlete's training. Our model makes predictions with far more accuracy - with a mean error on the order of a couple of minutes - and precision - with a much tighter confidence band.

Means and standard deviations, however, are not the language that athletes speak. As Outside Magazine notes in an analysis of the error of another "big-data" solution to this problem - "7.7% for a three-hour marathoner is almost 14 minutes, which is a pretty big deal… this looks a bit like BMI: very useful for population-level trends, not so good for making individual decisions"[19]. In a sport where minutes, rather than means, are the difference between success and failure, our model bridges the utility gap and tells an athlete what their actual fitness and an appropriate race goal are.

We make a few further interesting observations from Figures (4) and (5) as well:

- Our predictions tend to be better for women and for faster athletes, which is potentially an indication that both of these populations are more judicious at pacing and less likely to have large variations from what their training would suggest
- A series of distinctly clustered bands emerge from the scatter plot at round boundaries - 3:00, 3:30, 4:00, etc. This is an interesting insight into human psychology and the desire for defined barriers, and likewise indicates how important each minute is even in a race spanning hundreds of them
- Our initial hypothesis was that the Strava athlete population might not be representative of the entire sample - historically Strava has had the impression of being for more "hardcore" athletes and skewing more heavily towards men than the average demographics of marathon runners. The speed skew based on average finish time does not appear when compared to other studies; however, both the speed and gender skew are apparent when looking at global averages[3]. This represents a potential bias and area for improvement.

3.2 Feature Significance

Decision tree models also provide feature importance through various mathematical measurements. One such measurement is *gain*, which can be understood as the contribution to accuracy by a feature in the branches of the tree that it is part of. Gain also fairly represents both categorical and continuous features and is even impartial between categorical features with fewer buckets, like gender, and those with far more, like the number of interval sessions.
Appendix D has a quantitative depiction of gain, but we focus here on the relative positions of features rather than the magnitude of differences. In particular, we observe:

- The best predictors are prior best efforts, notably those closest to most athlete's lactate thresholds (7 miles, 1 hour) rather than strictly the longest ones (half marathon)
- Specificity (time spent around, but not faster, than goal pace) beats volume (long runs and weekly mileage)
- Age is less significant than gender or weight, which promotes endurance sport remaining accessible even when athletes are past their traditional "prime"

4. Case Studies: Prediction and Prescribed Training

The most consequential outcome of a population-study such as this one is if the learnings can be applied to individuals - in this case, their preparation and planning for race day. We thus look more discriminately at three rather relevant case studies:

- Varun Pemmaraju (Author 1) - Boston Marathon (2018) and Personal Record, or PR, from Twin Cities (2019)
- Dave Hoch (Author 2) - Boston Marathon (2019) and prior PR from Santa Rosa (2018)
- Nikhil Byanna[1] (a current Sloan student involved with the conference) - Boston Marathon (2019) and prior PR from Berlin (2017)

Feature	Varun Boston 18	Varun TC 19	2:25-2:30 Average	Nikhil Berlin 17	Nikhil Boston 19	Dave SR 18	Dave Boston 19	2:50-2:55 Average
Predicted	2:47:18	2:34:06	2:28:24	2:59:07	3:01:06	3:00:38	2:58:08	2:53:33
Actual	2:48:54	2:35:42	2:28:05	2:58:57	2:58:41	2:58:14	2:58:39	2:52:52
Course GAD	26.446	26.38	-	26.285	26.446	26.33	26.446	-
Avg Mileage / Week	36.3	40.9	63.3	18.8	23.6	41.0	36.0	39.5
Avg Longest Run / Week	13.3	14.3	16.4	10.4	11.0	14.4	14.0	13.8
Avg 2nd Longest Run / Week	8.6	8.2	12.1	5.8	6.9	8.7	8.3	8.8
Total Runs > 90 Mins	16	15	43	7	8	19	20	20
Total Workouts	19	33	3	-1	-1	15	17	-1
Total Active Days	99	138	153	64	73	133	100	115
Best 7 Mile Time	40:19	37:27	37:23	43:56	44:01	42:16	43:34	43:04
Best Half Marathon Time	1:15:54	1:11:08	1:12:19	1:34:47	1:23:07	1:28:17	1:22:25	1:24:20
Longest Distance - Avg Faster Than MP	15.1	13.3	14.1	12.8	10.6	21.6	13.0	14.9
Longest Distance Around MP (1 Activity)	11.4	9.0	8.1	11.5	11.1	14.2	6.9	8.0
Cumulative Distance Around MP (Whole Cycle)	154	150	119	179	270	224	119	133
Age	26	27	31	22	24	28	29	34
Weight	60	58	65	70	70	81	81	67

Figure 6 - Selected features and predictions for case studies and their respective goal's population averages. Darker shaded boxes indicate identified areas of improvement to achieve that goal.

[1] Nikhil reached out to us after the Abstract submission as he uses Strava, and we in turn asked for his consent to show his data as a case study for the Research Paper, which he kindly gave us

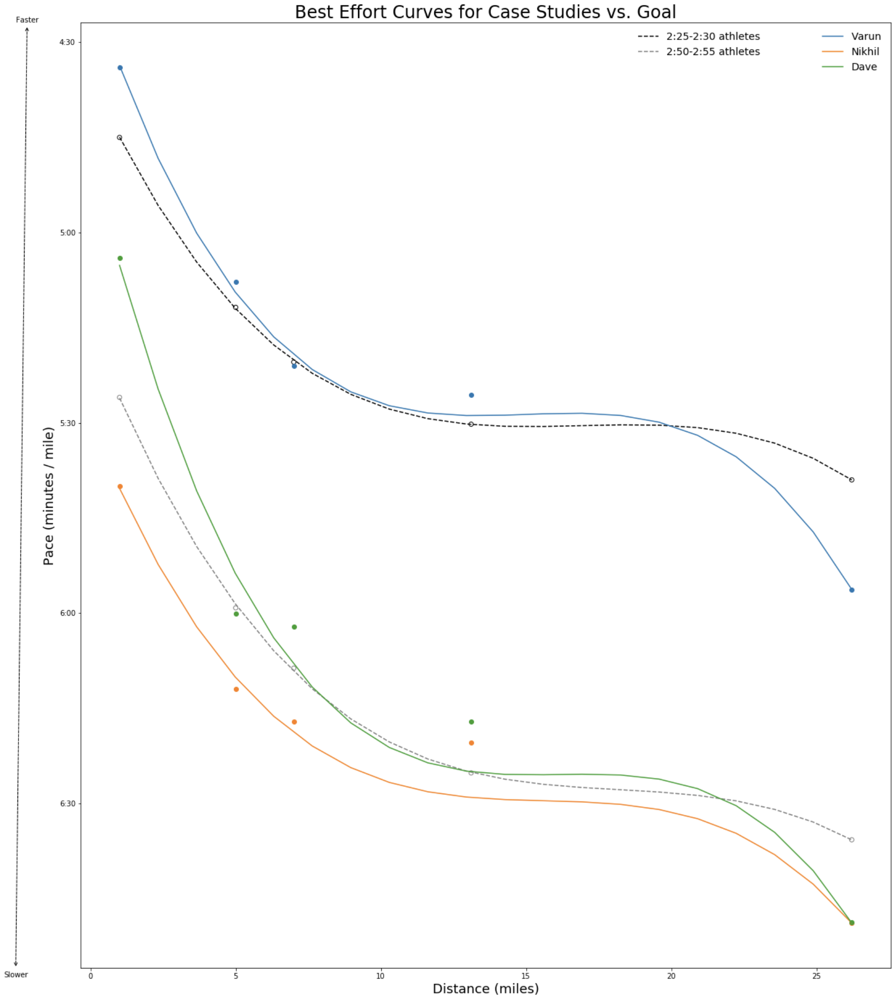


Figure 7 - Best Effort curves for case studies overlaid on their respective goals.

Besides the obvious vested interest, using these three case studies presents the advantage of examining two athletes - Dave and Nikhil - who have remarkably similar times (a mere two seconds separated them!) in the same race and identical future goals - to break 2:55. In selecting two races from each athlete, we can also theorize how changes to their training approaches may (or may not have) changed the outcome.

4.1. Varun: 2018 to 2019

Between April 2018 and October 2019, Varun dropped 13 minutes in his marathon over fairly similar course profiles. We observe the following:

- Measures of total training volume (*Training More*, *Long Runs*) do not show marked differences

- Even though total volume did not increase by that much, the *Consistency* of the number of active days increased by quite a bit, bringing Varun more in line with what other people around his pace were doing
- The most drastic difference is in the number of *Workouts*, which Varun perhaps even over-indexed on compared to the average marathoner of his speed. This resulted in a developed speed that carried into success over the longer marathon distance
- On *Demographics*, Varun dropped a few pounds, likely as a result of giving up his former life as a triathlete (but notably, not giving up the Dunkin Donuts) and no longer swimming. The authors would like to call out that "race weight" is a complicated concept - generally speaking, lighter weight means lowered energy expenditure required, but it is more important to stay biomechanically, nutritionally, and psychologically healthy and not let weight jeopardize those far more important characteristics

4.2. Varun: Top 100 at Boston

Since the day he first saw the Boston Marathon in person in 2016, Varun has had the ambition of finishing in the top 100, which tends to require about 2:30. In looking at trends for athletes in this range and the Best Effort curve, we see he already has the speed, hitting or exceeding the typical best efforts but then falling off spectacularly at the marathon distance.

However, his lack of high mileage and long run distance and frequency is glaring. We would suggest he build that sustained mileage up (*Training More, Long Runs)* if he wishes to hit his goals. As quantitative evidence, keeping all other features the same and setting his mileage to the average for 2:25-2:30 marathoners changes his prediction to 2:32:12.

4.3. Dave vs. Nikhil at Boston

For finishing a scant 2 seconds apart, Dave and Nikhil took very different journeys to the finish line in Copley Square. Dave focused on the mileage and long runs (*Training More, Long Runs)* and on paper, was much closer to the average 2:55 marathoner. Nikhil, on the other hand, was in the lower 10-15% for those same measures but spent an impressive and significant amount of the miles he did run accumulating time around his eventual marathon pace.

4.4. Dave and Nikhil: Breaking 2:55

Dave went to college in Boston and the Boston Marathon has had a special meaning for him since then. Boston has the distinction of being one of the only races in the world that you can only qualify on time (or via charity), and not enter a lottery. There is also a nuance where if a qualifying time exists and if more than the allotted number hit that time, they select racers in order of finish time and you may still miss the eventual cutoff. This happened to Dave when he appeared to secure his second chance to run up Heartbreak Hill, hitting the qualifying time of 3:00 with a 2:58:39, but ended up being heartbreakingly done in by the eventual cutoff being set at 2:58:21. He has thus stated a goal of running 2:55 to effectively guarantee his entry. By coincidence, when asked, Nikhil stated that his next ambition is to break 2:55 too, which makes for both an interesting story and rematch when that day arrives, and also interesting for our analysis.

Just as their previous training differed, so too are the recommendations for achieving their respective yet shared goals. We advise, quite simply, that Nikhil increases his mileage across the board - long and medium-long runs and number of runs per week (*Training More, Long Runs)*. In contrast, we would recommend for Dave to focus perhaps less on accumulating time on feet and instead, taking a page out of Nikhil's training log, to more efficiently spend those miles at faster speeds (*Workouts, Threshold, Running Economy)*. Another peculiar observation about Dave's training is that his preparation for Santa Rosa in 2018 appears better across the board than Boston 2019 even though the times in his two races are similar, which probably highlights how people learn to race the marathon and their bodies adapt to it over time (*Consistency)*. The encouraging commonality between both athletes is that while they each have left gaps in fulfilling certain training principles (that they even perhaps were better at filling in past cycles!), they also both have the characteristics, notably speed, to accomplish their goals.

Lastly, in all three cases - Varun, Nikhil, and Dave - we see that none of them have yet reached the age for marathon peak (*Demographics)*, which seems to be in the early to mid-30s. This is much later than most other running races, let alone other sports, indicating the importance of consistent and sustained mileage over time.

5. Conclusion

Utilizing big data techniques on a rich dataset yielded marathon prediction capabilities that are *miles ahead* of what was previously possible in both accuracy and precision. In particular, confidence in the predictions is high enough that athletes can utilize the model to plan training and race approach. Furthermore, we have a framework with which we can provide personalized insights and training recommendations to athletes for achieving their future goals.

All that said, there are numerous areas to pursue that would potentially improve this work further. For example, we could incorporate data about gear, which has received attention as part of an innovation arms race in the massive shoe industry[20, 21], as well as weather - both of these are already available in the Strava dataset.

We also look to apply some of these approaches to other sports such as cycling that have metrics, training principles, and desired outcomes that resemble running, and perhaps even sports with heavy endurance components like soccer that are starting to generate large quantities of biometric data for analysis. In comparison to sports like soccer, running certainly lacks mainstream publicity. However, it is a uniquely accessible sport that inspires a diverse audience to ask, "how do I become a better athlete?" The insights derived from this research have the potential to significantly change how people answer that question.

6. References

1. "Roger Bannister: First sub-four-minute mile | Guinness World ...." 17 Jun. 2015, https://www.guinnessworldrecords.com/records/hall-of-fame/first-sub-four-minute-mile

2. "Breaking2. Nike.com." https://www.nike.com/running/breaking2.

3. "Marathon Statistics 2019 Worldwide (Research) | RunRepeat." 30 Mar. 2020, https://runrepeat.com/research-marathon-performance-across-nations.

4. Bramble, D., Lieberman, D. Endurance running and the evolution of Homo. Nature 432, 345–352 (2004). https://doi.org/10.1038/nature03052

5. "How Fast Would You Run A Marathon? | FiveThirtyEight." 20 Nov. 2018, https://projects.fivethirtyeight.com/marathon-calculator/.

6. "Boston Marathon | Marathon road race in Boston ... - Strava." 15 Apr. 2019, https://www.strava.com/running_races/2019-boston-marathon.

7. Javier T. Gonzalez, Cas J. Fuchs, James A. Betts, and Luc J. C. van Loon. Liver glycogen metabolism during and after prolonged endurance-type exercise American Journal of Physiology-Endocrinology and Metabolism 2016 311:3, E543-E553. https://doi.org/10.1152/ajpendo.00232.2016

8. "An Improved GAP Model... - Medium." 3 Oct. 2017, https://medium.com/strava-engineering/an-improved-gap-model-8b07ae8886c3.

9. Amelia C. Lanning, Geoffrey A. Power, Anita D. Christie, and Brian H. Dalton. Influence of sex on performance fatigability of the plantar flexors following repeated maximal dynamic shortening contractions. Applied Physiology, Nutrition, and Metabolism. 42(10): 1118-1121. https://doi.org/10.1139/apnm-2017-0013

10. Smyth, B., & Muniz-Pumares, D. (2020). Calculation of Critical Speed from Raw Training Data in Recreational Marathon Runners. Medicine & Science in Sports & Exercise, Publish Ahead of Print. https://doi.org/10.1249/mss.0000000000002412

11. "VDOT Running Calculator - Apps on Google Play." https://runsmartproject.com/calculator/.

12. van de Vegte, Y. J., Tegegne, B. S., Verweij, N., Snieder, H., & van der Harst, P. (2019). Genetics and the heart rate response to exercise. Cellular and molecular life sciences : CMLS, 76(12), 2391–2409. https://doi.org/10.1007/s00018-019-03079-4

13. "Apache Spark™ - Unified Analytics Engine ...." https://spark.apache.org/.

14. "XGBoost Documentation - Read the Docs." https://xgboost.readthedocs.io/.

15. "sklearn.model_selection.GridSearchCV — scikit-learn 0.23.2 ...." http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

16. "RW's Race Time Predictor - Runner's World." 15 Jul. 2013, https://www.runnersworld.com/uk/training/a761681/rws-race-time-predictor/.

17. Vickers, A. J., & Vertosick, E. A. (2016). An empirical study of race times in recreational endurance runners. BMC Sports Science, Medicine and Rehabilitation, 8(1). https://doi.org/10.1186/s13102-016-0052-y

18. Altini, M., & Amft, O. (2018, July). Estimating Running Performance Combining Non-invasive Physiological Measurements and Training Patterns in Free-Living. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). https://doi.org/10.1109/embc.2018.8512924

19. "A Big Data Approach to Predicting Your Marathon Pace ...." 22 Jun. 2020, https://www.outsideonline.com/2414931/critical-speed-marathon-prediction-study.

20. "Nike Says Its $250 Running Shoes Will Make You Run Much ...." 18 Jul. 2018, https://www.nytimes.com/interactive/2018/07/18/upshot/nike-vaporfly-shoe-strava.html

21. "Nike's Fastest Shoes May Give Runners an Even Bigger ...." 13 Dec. 2019, https://www.nytimes.com/interactive/2019/12/13/upshot/nike-vaporfly-next-percent-shoe-estimates.html.

7. Appendices

A - Training Volume Nuances

- The duration of a run can be measured either by moving time (excluding stops for water, bathroom breaks, photos, petting dogs, etc.) or elapsed time (total door-to-door time on feet). We hypothesize that moving time is the more important one for training, and elapsed time is of course the one relevant to races, but used both in the model to confirm
- Runners may upload multiple activities on a single day, either to break up the warmup and cooldown portions of an interval session or even for more advanced athletes, to do "double-days" where they have multiple runs in a day. For simplicity of calculation of measures of training volume, we have combined metrics for all runs occurring on the same day
- Runners vary in how they measure volume - some use time and some use distance (and within those some use kilometers and others miles). We often chose many very similar features in the initial model training to represent this and dropped down to a smaller number once the most significant features from the group were determined

B - Full Feature List

- Avg Miles Per Week
- Avg Time Per Week (moving)
- Avg Time Per Week (elapsed)
- Avg Longest Run Per Week
- Avg Longest Run Per Week (moving)
- Avg 2nd Longest Run Per Week
- Avg 2nd Longest Run Per Week (moving)
- Total Runs Over 2 Hours
- Total Runs Over 90 Minutes
- Total Workouts
- Total Long Runs
- Total Active Days
- Total Active Weeks
- Longest Distance Covered 20 Min
- Longest Distance Covered 60 Min
- Longest Distance Covered 90 Min
- Longest Grade Distance Covered 20 Min
- Longest Grade Distance Covered 60 Min
- Longest Grade Distance Covered 90 Min
- Fastest Time 1 Mile
- Fastest Time 5 Km

- Fastest Time 5 Miles
- Fastest Time 10 Km
- Fastest Time 7 Miles
- Fastest Time 10 Miles
- Fastest Time Half Marathon
- Fastest Time 30 Km
- Fastest Time 20 Miles
- Longest Continuous Distance - Avg Faster Than Goal Pace
- Longest Continuous Distance - Avg Faster Than 95% Goal Pace
- Longest Continuous Distance - Avg Faster Than 90% Goal Pace
- Longest Continuous Grade Distance - Avg Faster Than Goal Pace
- longest Continuous Grade Distance - Avg Faster Than 95% Goal Pace
- Longest Continuous Grade Distance - Avg Faster Than 90% Goal Pace
- Longest Distance Around Goal Pace - Single Activity Total
- Longest Grade Distance Around Goal Pace - Single Activity Total
- Total Distance Around Goal Pace - Whole Cycle
- Total Grade Distance Around Goal Pace - Whole Cycle
- Gender
- Age
- Weight

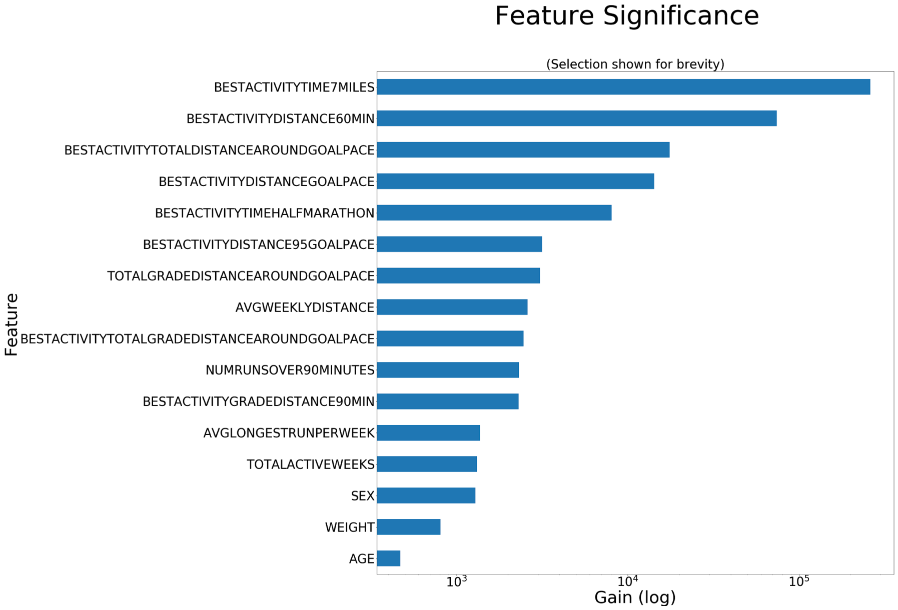
C - Model Training Hardware and Parameters

For hardware, we used 60 AWS i3.2xlarge machines with 8 CPUs and 40 GB of memory each.

The XGBoost hyperparameters chosen via tuning are listed below, and further explanation of each can be found in the documentation[8]

- Max tree depth: 8
- Minimum child weight: 4
- Learning rate (η): 0.03
- Gamma (γ): 0
- Number of estimators: 2000

D - Feature Significance from XGBoost Gain


Feature Significance
(Selection shown for brevity)
BESTACTIVITYTIME7MILES
BESTACTIVITYDISTANCE60MIN
BESTACTIVITYTOTALDISTANCEAROUNDGOALPACE
BESTACTIVITYDISTANCEGOALPACE
BESTACTIVITYTIMEHALFMARATHON
BESTACTIVITYDISTANCE95GOALPACE
TOTALGRADEDISTANCEAROUNDGOALPACE
AVGWEEKLYDISTANCE
BESTACTIVITYTOTALGRADEDISTANCEAROUNDGOALPACE
NUMRUNSOVER90MINUTES
BESTACTIVITYGRADEDISTANCE90MIN
AVGLONGESTRUNPERWEEK
TOTALACTIVEWEEKS
SEX
WEIGHT
AGE
Feature
10^3
10^4
10^5
Gain (log)