Predicting Human Performance with Machine Learning

Fix citations – use numbers after fixed works cited ordering [#]

**Abstract**

1. **Introduction**

This study explores the possibility of leveraging machine learning and personal training data to predict race performance more accurately, specifically focusing on the sport of running. Most of the tools that exist around race performance today generate a prediction either based on a previous race result or by considering another single data point, such as an individual’s VO2 Max. These tools also tend to rely on vague generalizations, such as basing the prediction result on the assumption that the subject is doing the “proper training” (CITE). Section 2 of this paper explores some of these techniques in greater detail, as well as their shortcomings.

Running is one of the most popular sports worldwide, with an estimated 50 million participants in the United States alone in 2021 (“Running & jogging”, 2023). The popularity of running and jogging is often traced back to the 1970s, a phenomenon often coined the “long-distance-running boom” (Haberman, 2017) which coincided with a larger “exercise boom” that was observed across America throughout the 1960s-70s (PODCAST). The Boston Athletic Association Marathon, America’s most esteemed race, had just 197 participants in 1960. This number reached 1,342 by 1969 and saw nearly a sixfold increase to 7,927 runners by 1979 (Plymire, 2004). In 2023, the Boston Marathon had nearly 30,000 participants (baa.org).

As the number of runners increased, so did the technology surrounding the sport. From the emergence of fitness trackers in 1965 with the Manpo-kei (the ’10,000 steps meter’, invented by Japanese professor Dr. Yoshiro Hatano), to the introduction of wireless heart rate monitors in the 1980s, to the launch of modern wearables with the Fitbit in 2009, and eventually feature-rich smartwatches, beginning with the 1st generation Apple Watch in 2015 (Douglas-Walton, 2020), personal fitness and performance data is now more accessible than ever.

With the abundance of data today, we’ve seen powerful changes in the consumer experience. Considering the healthcare industry momentarily, wearable and sensory devices have enabled 24/7 monitoring of patient metrics, such as heart rate, heart rate variability, and rate of respiration, opposed to relying on collecting this data from a single in-office visit. Patients’ health can now be analyzed more holistically, and this data can be leveraged with machine learning techniques to recognize patterns and anomalies, hopefully enabling more accurate and proactive care. This has allowed for personalization in an industry that traditionally relied on a “one-size-fits-all” approach for patient treatment (Sebastian, 2023).

Much like healthcare needs varying amongst patients, so do training needs amongst athletes (Haugen et al., 2022). The ability to better understand how your personal training effects your performance potential could be extremely valuable in a sport where world records are determined by 34 seconds (Puleo, 2023). \*

In the remainder of this report, we will examine popular race prediction techniques and the various factors that have been proven to correlate with running performance (section 2), the steps taken to prepare our training data and build our prediction model (sections 3 and 4), the results of our machine learning model (section 5), and final remarks, including possibilities of future improvements to the model (section 6).

1. **Research Background**
   1. **Factors Effecting Running Performance**
      1. **Intrinsic Factors**

When attempting to estimate running performance, the basis for the prediction tends to be the athlete’s intrinsic qualities. This includes both quantitative measures of fitness, as well as genetic characteristics. Three of the primary metrics used to quantify an athlete’s aerobic ability and determine their performance potential are maximal oxygen uptake (VO2 max), running economy (RE), and anaerobic threshold (AT) (Venturini & Giallauria, 2022).

VO2 max is a measure of the maximum rate of oxygen delivery and utilization to your muscles during cardiovascular exercise (Heins, 2023). In other words, it is a measure of fitness, with a higher VO2 max indicating a higher degree of fitness. Many modern fitness trackers, such as Garmin watches, provide VO2 max estimations based on the relationship between the user’s heartrate and pace (Heins, 2023). VO2 max can be particularly useful to pinpoint key metrics such as heart rate zones and anaerobic threshold, which can facilitate more effective training regimes (Ritterbeck & Williams, 2020). Although VO2 max has a clear correlation with endurance race results, predicting an estimated 59% of the variance in marathon times (Venturini & Giallauria, 2022), the absence of significant differences in VO2 max amongst top runners (Lucia et al., 2006) and the observed variance in maximum oxygen uptake (O'Toole et al., 1987) suggests that VO2 max may not be the best indicator of one’s relative race performance.

Running Economy (RE) refers to the amount of energy required to run a specific speed, and has been shown to be a more useful performance predictor amongst a homogeneous group of runners with similar VO2 maxes (Diebel et al., 2017). Improvements in RE will allows athletes to run at higher speeds with the same oxygen uptake (O'Toole et al., 1987). RE is also highly trainable, and therefore improvements in RE is a key focus amongst coaches. Typical training interventions aimed at RE advancement include strength work, training in hot environments, high-carbohydrate diets, and training at altitude (Diebel et al., 2017). Altitude training in particular has been consistently effective in improving RE, as demonstrated by a 20-day study where 22 elite distance runners were divided into three groups, each group training at a different altitude over the course of the study (2,000-3,100m, 1,500-2,000m, and 600m, respectfully). Each of the athletes’ performance was evaluated at low altitude (600m) before and after the training interval, and it was determined the athletes training at higher altitude experienced improvements to their RE (Saunders et al., 1985).

A separate study attributes the improvements in RE observed during altitude training specifically to the changes in net muscle lactate release (Brooks et al., 1992). Lactate acid is naturally produced by the body during normal metabolism and exercise, and is essential to maintaining proper bodily function, acting as the body’s primary energy source, a precursor for producing glucose, and a signaling molecule (Rush & Barrell, 2022). However, if the body produces too much lactic acid that it cannot remove quickly enough, sever conditions such as lactic acidosis or hyperlactatemia may be a risk. Lactate levels in the body can increase nearly twenty-fold during intense exercise, and proper training can help regulate this by allowing you to perform at a higher rate of work without raising your blood lactate levels (UC Davis Health). More specifically, training results in reduced lactate production and increased lactate re-uptake within the body. Active recovery practices, or light intensity exercise following strenuous exertions, has also been shown to improve lactate removal (Hatta, 2023).

The final metric often used to predict running performance is one’s Anaerobic Threshold, or AT. AT is a sustainable level of effort that an athlete could maintain for hours (Walsh, 2023). More specifically, the aerobic threshold is the point where the lactate levels start to rise in the bloodstream and anaerobic energy pathways start to assist in energy production. A higher AT means you can train at higher intensities without lactate building up (Walsh, 2023). Training at one’s anaerobic threshold level generally results in minimal increases to VO2 max and AT speed, but significantly increases in the time to exhaustion at AT speed, once again likely do to improving lactate level regulation (Billat et al., 2004). In summary, a higher anaerobic threshold indicates a higher level of lactic acid tolerance, which corresponds to a higher VO2 max (Diebel et al., 2017).

The intrinsic factors discussed so far (VO2 max, RE, and AT), are all trainable to some degree. The final intrinsic factors we will consider, however, are those that are unchangeable – one’s genetics.

Gender difference for one, will always have an impact on performance capabilities. Males on average have greater muscle mass, heart size, and hemoglobin concentration, meaning the VO2 max potential is higher in men than women (Venturini & Giallauria, 2022), explaining the 10-12% slower race times seen by women compared to men at an elite level (Joyner, 2017).

Age is another clear factor in performance. An athlete’s peak maximum oxygen consumption (VO2 max ) is achieved around 27 years for males, and 29 years for females (Lara, 2014). Peak endurance capability is generally thought to be maintained until around 35 years of age, followed by a moderated decrease until 50-60 years, and a more dramatic decrease thereafter (Parker, 2011). Decline in endurance capability with age is attributed to reductions in VO2 max and lactate threshold.

Nationality also has been proven to play a role in performance ability. African runners, particularly East African, have historically had faster marathon times than non-Africans (Venturini & Giallauria, 2022).

Additional genetic characteristics, such as longer legs, lower body fat, and higher flexibility, are of course going to make an individual better equip for running from the beginning (Venturini & Giallauria, 2022).

* + 1. **Extrinsic Factors**

Likely carrying less weight on performance prediction than the intrinsic factors discussed previously are extrinsic considerations, such as climate, topography, and running equipment.

Topography, specifically elevation gain, can be a significant indicator of performance potential on a particular course. The Berlin Marathon, for example, is known to be a relatively flat course, and holds the fastest winning times by comparison (Díaz, 2019).

Other environmental factors also appear to have a correlation with race times. The best marathon records generally occur in the autumn or springtime, with moderate temperatures ranging between 10 and 15°C (Díaz, 2019). Additionally, races that occur in cities with higher degrees of air pollution have shown reductions in average performance (Marr, 2010).

Finally, new advanced shoe technology, or NAST became popular around 2017 when NIKE introduced their Vaporfly shoe as part of a larger project aimed at running the first sub-2-hour marathon. NAST includes shoes with carbon fiber plates, greater stake height, and lower weight. After the introduction of this technology, runners wearing NAST ran ~1% faster in the marathon compared to those who did not use it (Rodrigo-Carranza et al., 2021)

* 1. **Race Prediction Tools**

Prior to building out the personalized race predication model detailed in this report, extensive research was done on the current tools and techniques that are used to predict race performance. The remainder of this section outlines how these tools work, and Section 5 compares the predictions generated by these tools with the predictions generated by the custom machine learning model.

* + 1. **Online Race Predictor Calculators**

Many of the online race prediction tools prompt the user to input a previous race distance and finish time and input these measures into a formula to calculate an estimated finish time for a different distance. Some of the tools also ask for additional data points, such as Sex, ..

A few of the online tools offer a more detailed breakdown of how the race time prediction is derived. Marathon Handbook’s online race time calculator, for example, provides time estimations based on two formulas: the Pete Riegel and David Cameron formulas. The Pete Riegel formula was published in Runner’s World magazine back in 1977, and proposes that a linear relationship between distance and speed, where a person’s race speed declines about 6% as the distance doubles (Radford, 2023). The Riegel formula is one of the most popular calculation methods for many of these online tools.

Pete Riegel formula. D1 and T1 correspond to the previous race distance and time, D2 and T2 correspond to the upcoming race distance and time.

The David Cameron formula is more complex and leverages a non-linear regression model that incorporates race times of world-level athletes for distances between 400 meters to 50 miles. (“Cameron Race Time Prediction Formula and Calculator”, n.d.). One limitation with either of these formulas is that both initial studies were based on racing patterns of professional athletes, rather than the average runner. This can lead to underpredicted race times, especially for longer distances.

To compensate for the shortcomings of Riegel and Cameron formulas, the Marathon Handbook also offers a specific Marathon Race Time Calculator, as well as an Age-Grade Calculator. The Marathon Calculator aims at predicting race times for non-professional runners and is based on formulas derived by Andrew Vickers and Emily Vertosick. It also takes into consideration two previous races, as opposed to one, as well as an estimate of you weekly mileage during training. Potential problems here include the fact that weekly mileage often varies throughout the course of a training period, and there is no indication of how many weeks you have been averaging this mileage.

The Age-Grade calculator considers the runner’s age and gender and applies an “age factor” between 0 and 1 to the prediction. The calculator assumes peak running potential occurs between 20 and 30 years of age, which corresponding to an age factor of 1. The prediction is based on world record race time data and attempts to estimate the best possible time an athlete could run a certain distance based on their age. The below formulas are used to calculate the prediction with this tool.

A white background with black text

Description automatically generated

Each calculator makes certain assumptions and generalizations about a runner’s performance, and all the online race predictor tools and tend to emphasize that the predicted times assume “you have trained appropriately for said distance” (“Race Time Calculator”, n.d.). Necessary training can vary amongst runners, and it is unclear what these tools consider to be “proper training”.

* + 1. **VDOT – an adjusted VO2 Max**

Where the online calculators explored above primarily focus on using past performance to predict race times, the intrinsic factors mentioned previously are also commonly used for predictions. One factor, known as VDOT, is particularly popular.

Famous running coach Jack Daniels developed VDOT as a combination of VO2 Max and Running Economy to serve as a more holistic measurement of aerobic fitness, as opposed to relying solely on VO2 Max (Tri2Max Coaching LLC, n.d.). Online tools allow you to calculate your VDOT based on a previous race result (“VDOT Running Calculator”), and then use this value to look-up the corresponding race time predictions at various distances in a VDOT table (Jordan, n.d.). VDOT can also be a helpful metric in determining target training paces bases (i.e. threshold pace) (Daniels, 2022).

* + 1. **Garmin**

The final race prediction tool explored in this paper is a slightly more personalized tool, Garmin’s Race Predictor. Garmin watches have a built-in race prediction feature that primarily uses your VO2 Max, as well as training history from the past several weeks, to give you an estimate of predicted paces for several different race distances (“Forerunner 245/245 Music Owners Manual”). The VO2 Max that Garmin watches use for this estimation are calculated based on the relationship between the user’s heartrate and pace over recent activities (Heins, 2023). This VO2 Max estimation does not consider other factors that may affect the relationship between heart rate and pace, such as illness, dehydration, lack of sleep, travelling, or running in heat or at altitude.

Limitations with Garmin’s method of race prediction include that fact that, as stated earlier, VO2 Max alone is not always the most accurate race predictor across individuals. Additionally, the VO2 Max calculation for Garmin watches rely on the accuracy of one’s heart rate measurement. Studies have found wrist-based heart rate monitors to have a margin of error ranging from 1 to 13.5% (Thomson et al., 2018), which often leads to underestimations of heart rate, and thus overestimations of VO2 Max and race pace. Also, as with the other race prediction methods mentioned, Garmin’s calculation makes the assumption that you have completed proper training for that race distance (Garmin, 2020) and does not take into account any additional factors such as weather, course difficulty, terrain, or training regime.

In summary,

**NONE of the above take into account extrinsic factors**

**All rely on a previous race time, and race course**

1. **Data Pre-Processing**

Common limitations between all the race prediction methods explored in Section 2 is the prediction is not personalized to the individual, nor does it consider any extrinsic factors. The individual’s specific training patterns (mileage; running pace - intervals, threshold, easy miles, etc.; training at elevation; training at heat; cross-training), level of effort they may put into a race activity vs their potential (do they run to exhaustion during a race, or do they maintain a comfortable level of exertion), anomalies (such as injuries, illnesses, amount of sleep) all may effect an athlete’s performance differently, and to varying degrees. Additionally, the extrinsic factors unique to the environment where the race is occurring (temperature, altitude, elevation) will all impact the level of difficulty of the course, and thus the individual’s performance. The data used to train the personalized, machine learning-based model explored in this project was specifically selected and curated to compensate for these shortcomings.

* 1. **Data Overview**

Strava is a comprehensive, sport-centric social media platform with features that enable athletes to track their own training progress, connect with other athletes, explore new places, and compete for leaderboard positions and in virtual competitions (Strava, 2023). Most of Strava’s features are enabled by its activity tracking capabilities. Strava allows you to track a range of athletic activities either from the app itself, or from a third-party fitness tracker or smartwatch (i.e. Garmin watch, Apple watch) and upload them to your Strava profile, memorializing this activity as part of your athletic journey. Once the activity is uploaded to your profile, you can holistically review this activity by analyzing key metrics including speed/pace, elevation, distance, heart rate, and cadence, as well as share this activity with your followers (Strava, 2023).

Strava offers a publicly available API that allows developers to access Strava data (“Strava API V3”). The data set we will be using is my personal Strava activity data over the last 1-2 years. Additionally, we enriched the activity dataset to include weather data for each specific location/activity by leveraging Meteostat’s Python library, which provides access to open weather and climate data via Pandas DataFrame objects (Meteostat). By using one’s personal data to build out a race predication model, we can consider the majority of the intrinsic and extrinsic factors effecting performance that were introduced in Section 2, and hopefully build a more accurate, individualized race predication tool.

Prior to any of the data preparation steps detailed in this section, Strava’s activities API was called to retrieve my recent activity data starting from START DATE. This data was then loaded into a DataFrame using Python’s Pandas library. The code used to retrieve and prepare the data, as well as build out the models discussed in Section 4, can be found in the GitHub repository referenced in the footnotes.

* 1. **Feature Selection and Data Enrichment**

Strava’s activities API returns 55 attributes corresponding to a given activity. Initial feature selection was determined by the presence of null values in each column, removing one of two closely coupled columns (i.e. average pace and moving time), and focusing on the key determinants of run performance that were introduced in Section 2. Selected features included local date-time, distance, moving time, total elevation gain, activity type (i.e. Run, Bike, Swim, etc.), starting latitude/longitude, maximum speed, average cadence, average heart rate, maximum heart rate, and elevation high.

After initial feature selection, the local start date-time and the coordinates of the activity were leveraged with the Meteostat Python library to retrieve weather station data for each activity and enrich the data with an average temperature feature. The decision to include this feature was based on the proven correlation between temperature, heart rate, and running pace (Venturini).

Another key step of our data enrichment involved adding an additional column “race”. This column would hold Boolean values indicating whether an activity was a race activity (True) or not (False). The dataset had # race activities, and thus # True values in the “race” column.

Chart showing distribution of values in race col

* 1. **Handling Null Values**

Handling missing values is a key step in data pre-processing, as most machine learning models and Python libraries are unable to handle these Null or NaN values on their own. Missing values can also lead to biases, and thus inaccuracies, in the final model (simplilearn). The null values in our data set were limited, but there were still a few attributes that needed to be addressed. Below is a summary of the initial null values by column in our data set.

UPDATE

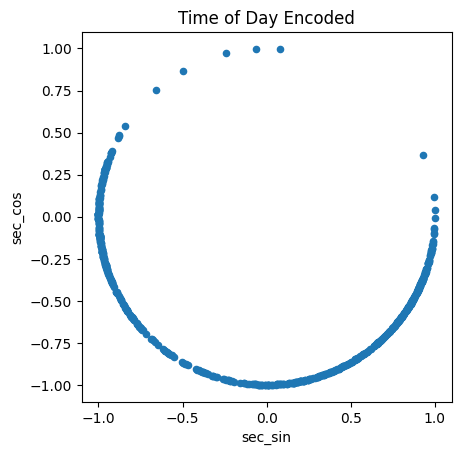
|  |  |
| --- | --- |
| Column Name | Number of Null Values |
| start\_date\_local | 0 |
| start\_latlng | 7 |
| distance | 0 |
| moving\_time | 0 |
| total\_elevation\_gain | 0 |
| type | 0 |
| max\_speed | 0 |
| average\_cadence | 16 |
| average\_heartrate | 15 |
| max\_heartrate | 15 |
| elev\_high | 7 |
| avg\_temp\*\* | 20 |

Many strategies exist for handling null values, but due to the ordered nature of the dataset (activities are sequenced by timestamp), it was determined that interpolation would be a sufficient method. Interpolation estimates the missing values based on the values of surrounding data points (Saha, 2023), and thus we are assuming that nearby activities occurred at a similar location and at a similar level of effort. The built-in interpolate method in the Pandas library was used to apply interpolation to each column with missing values (pandas doc).

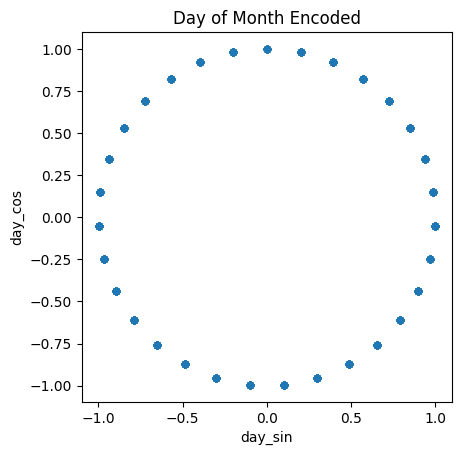
* 1. **Data Encoding: Handling Categorical Features**

Categorical values are those that are discrete and non-continuous. There are both ordinal (ordered) and nominal (unordered) categorical values, each of which need to be preprocessed differently (Kumar). In the case of our data set, the only categorical feature that would ultimately be included in our final model was the datetime column start\_date\_local.

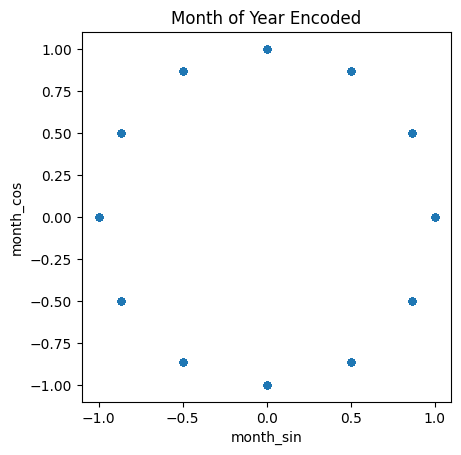
Date and time were important factors to consider in our model because one’s training and performance my vary based on time of day and throughout the year in response to factors such as climate (i.e. temperatures, rain/snow, air quality). Date and time are unique categorical fields, as they are cyclical in nature. To preserve the cyclicality, Eryk Lewinson’s method of feature transformation via sine and cosine functions was used to encode the activity’s time of day, day of the month, and month of the year [LEWINSON]. The following images show the resultant sine/cosine values of each feature respectfully following the transformation.



Time of Day encoded



Day of Month encoded



Month of Year encoded

Lastly, we handled year via one-hot encoding, as there were only two different year values in our dataset. Columns “2022” and “2023” were added to the dataset, and their value was set as True or False accordingly.

* 1. **Data Encoding: Handling Numerical Features**

Most of the attributes in our dataset are numeric. However, each column’s numerical values are at a different scale, meaning each column needs to be standardized to prevent columns with bigger values from dominating decisions made by our machine learning models.

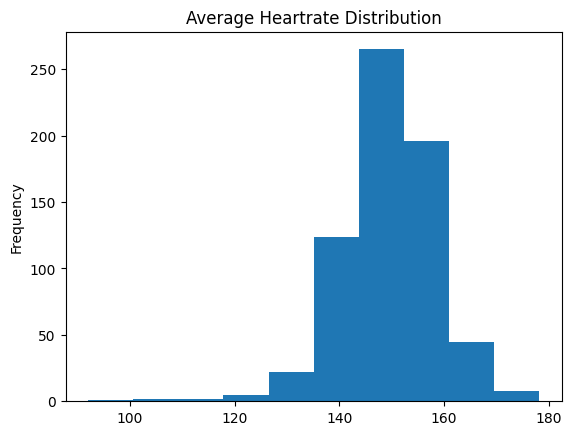
There are two common methods of transforming numerical values – standardization and normalization. Normalization involves scaling the column values on a set range, typically between 0 and 1 or -1 and 1. Normalization is useful when the distribution of a column is unknown or not normal, but normalized columns tend to be more effected by the presence of outliers (Jaadi, 2023). The formula below demonstrates how some of the columns in this study were normalized on a scale of 0 to 1.

Normalized value calculation. X represents the raw score, and min(x) and max(x) represent the minimum and maximum values of the column, respectfully.

Standardization, on the other hand, involves scaling a numeric column so that the data has a mean of 0 and standard deviation of 1. Standardization is typically the preferred method for normally distributed values and can be accomplished by calculating the z-score of each value using the formula below.

Z-score calculation. X represents the raw score, represents the mean value of the column, and represents the column’s standard deviation.

In the case of our data set, we first looked at the distribution of values across each numeric column and then made the decision on whether to use standardization or normalization to transform the values. The distribution of values across each column are illustrated in the figures below.



Mean: 149.26

Standard Deviation: 9.10

A graph of a number of heart rate

Description automatically generated

Mean: 171.22

Standard Deviation: 10.50

A graph of a speed distribution

Description automatically generated

Mean: 5.36

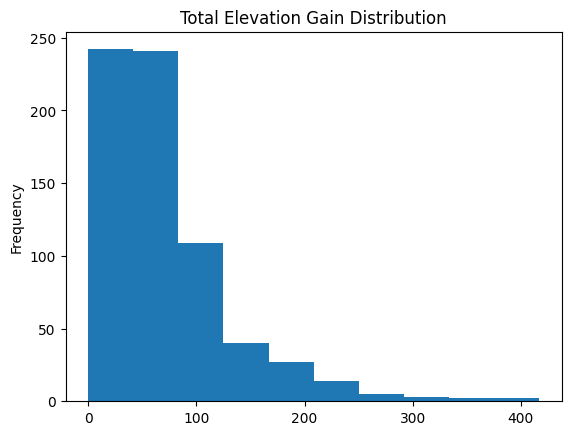
Standard Deviation: 1.71

A graph of elevation high distribution

Description automatically generated

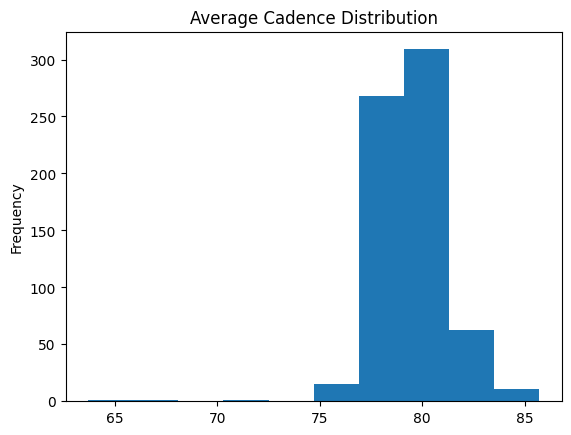
Mean: 290.65

Standard Deviation: 417.36



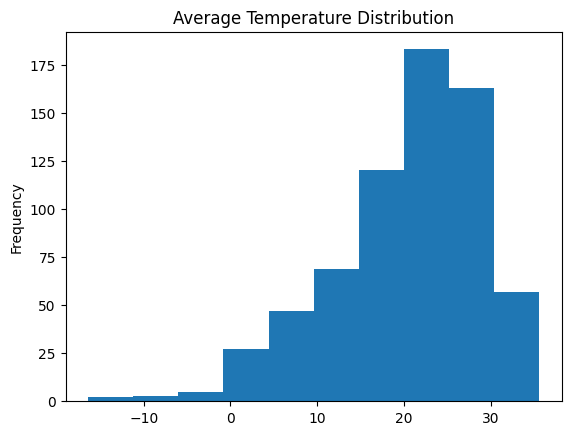
Mean: 71.34

Standard Deviation: 59.91



Mean: 49.41

Standard Deviation: 1.80



Mean:20.23

Standard Deviation: 8.54

Standard deviation is an indication of how close the values of a dataset are to the mean value. A lower standard deviation indicates a smaller distribution of values, i.e. values are closer to the mean. A normal “bell curve” distribution is recognized by a symmetrical distribution of values, where the mean is 0 and standard deviation is 1. Columns average\_cadence, max\_speed, max\_heartrate, average\_heartrate, and avg\_temp exhibited relatively symmetric, normal distributions with lower standard deviations, and thus were standardized via z-score. Columns total\_elevation\_gain and elev\_high, had more of a skewed distribution, and were thus transformed via normalization.

1. **Method**

After pre-processing, the encoded data was used to build a series of machine learning models, the results of which are summarized and compared in Section 5. This section will walk through the model selection process and how the accuracy of the models will be assessed.

* 1. System Overview

Two machine learning models were constructed based on the data set described above – a Linear Regression model and a Neural Network. These models were both trained on the same data; a subset of my personal Strava activity over the last approximately 2 years cleansed and encoded following the methods described in Section 3. The output of each model was a prediction of the total moving time of a running activity. The features used to train the models are summarized in the figure below.

Encode distance?

|  |  |  |
| --- | --- | --- |
| Feature | Description | Encoded |
| distance | Total distance (meters) of activity | N |
| total\_elevation\_gain | Total elevation gain over the course of activity | Y |
| max\_speed | Maximum speed run during activity | Y |
| average\_cadence | Average cadence during activity | Y |
| average\_heartrate | Average heart rate during activity | Y |
| max\_heartrate | Maximum heart rate during activity | Y |
| elev\_high | The highest elevation recorded during activity | Y |
| avg\_temp | Average temperature of the location/day of activity | Y |
| race | Boolean value indicating if activity is a race activity (T) or not (F) | N |
| month\_sin, month\_cos, day\_sin, day\_cos | Cyclically-encoded columns to represent month and day of activity | Y |
| sec\_sin, sec\_cos | Cyclically-encoded columns to represent time of day of activity | Y |
| 2021, 2022, 2023 | One-hot encoded columns indicating the year of the activity | Y |

Model Training Features

To test each of these models, I used two of the most recent races I ran as inputs, and compared the models’ predictions to my actual results. The first race, the Richmond Marathon, occurred on November 11th, 2023 in Richmond, Virginia. The second race was a 15K in Seattle, Washington on November 23rd, 2023. For each race, input features regarding distance, elevation, and date/time were all known. Features regarding speed, cadence, and heartrate were populated in the code using the average values of those respective fields from other race activities in the training data set (i.e. where “race” == True).

A special case was used to handle the average temperature. The average temperature value corresponds to the average temperature at a given location on that day the previous year. This logic was built into the code with the assumption that this race predication model would be used the predict a race result in the *future*, and thus the exact temperature on the day of the race was unknown.

For each race result prediction, the training data used to build the model spans from August 2021 to the day before the race being predicted.

* 1. Linear Regression Modelling

The first series of models constructed were multiple Linear Regression models. Multiple linear regression is a very popular supervised machine learning technique where multiple independent variables (the input features described above) are used to predict the outcome of a single dependent variable (total moving time) by establishing a linear relationship between the two. A linear regression algorithm finds a best-fit line between the independent and dependent variables that minimizes the sum of squared errors between the actual and predicted values [Biswal].

Regression line equation where Y is the dependent variable, x1, x2, … xn are the independent variables, and , , … are coefficents.

The models were constructed using Python’s scikit-learn LinearRegression (sci-kit) module. This module allows you to train a linear model on a dataset, predict new values with this model, and evaluate the model’s performance.

* 1. Neural Network Modelling

The second series of models were Neural Networks. This is a type of deep learning that is loosely modelled after the human brain and consists of upwards of thousands densely interconnected nodes [Hardesty]. Typically, these nodes are organized into layers, and the layers pass data between them. Oftentimes a single node is receiving information from multiple nodes in the layer proceeding it and is also sending data to multiple nodes in the layer following it.

The neural network models were built using Python’s Keras library. This is part of the larger TensorFlow Python library, and allows you to develop and build deep learning models with minimal code [keras].

1. **Results**

This section compares the results and accuracy of the machine learning models created during this study.

* 1. **Actual Race Results**

The below figure summarizes my actual race results from the marathon and 15K. These times will be compared to the models’ predicted race time to assess accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Race Name | Race Date | Distance (miles) | Total Moving Time (as recorded by Garmin) | Pace |
| Richmond Marathon | 11/11/2023 |  |  |  |
| Seattle Turkey Trot 15K | 11/23/2023 |  |  |  |

* 1. **Prediction Results from Existing Tools**

Prior to evaluating the race predictions generated by the custom machine learning models, race estimations for the marathon and 15K were generated using the existing tools described in Section 2. The prediction results from these tools, as well as my past race results that were supplied to the tools when applicable, are summarized in the figures below.

\*\* feature importance

**Linear Regression**

All encoded features

Marathon 🡪 3:10:49

15K -> 59:27

Not datetime encoded features

To evaluate the accuracy of the predication for each model

**Notes about each of the previous races I used (i.e. hilly, crowd support, injuries)**

**Compare results of all the prediction tools mentioned above**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Input Race** | **Date** | **Distance (miles)** | **Chip Time** | **Pace (min/mile)** |
| **1** | **3M Half Marathon** | **1/22/2023** | **13.1** | **1:26:28** | **6:36** |
| **2** | **Austin Half Marathon** | **2/19/2023** | **13.1** | **1:29:42** | **6:51** |
| **3** | **Dallas Turkey Trot 5K** | **11/24/2022** | **3.1** | **00:19:17** |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Finish Time Prediction Tool** | **Input Race(s)** | **Extra Informaion** | **Predicted Marathon Time** | **Predicted Marathon Pace** | **Predicted 15K Time** | **Predicted 15K Pace** |
| [**https://sporttracks.mobi/labs/race-finish-time-predictor**](https://sporttracks.mobi/labs/race-finish-time-predictor) | **1** | **Sex (F)** | **3:07:05** | **7:08** | **1:00:14** | **6:28** |
| [**https://marathonhandbook.com/race-time-calculator/**](https://marathonhandbook.com/race-time-calculator/)  **(Pete Riegel formula)** | **1** |  | **3:00:24** |  | **N/A** |  |
| [**https://marathonhandbook.com/race-time-calculator/**](https://marathonhandbook.com/race-time-calculator/)  **(David Cameron formula)** | **1** |  | **3:03:53** |  | **N/A** |  |
| [**https://marathonhandbook.com/marathon-race-time-predictor/**](https://marathonhandbook.com/marathon-race-time-predictor/)  **Marathon Race Time Prediction** | **1,3** | **Weekly Mileage: 55 miles** | **3:08:39** |  | **N/A** |  |
| [**https://marathonhandbook.com/age-grade-calculator/**](https://marathonhandbook.com/age-grade-calculator/)  **Age-Grade Calculator** | **1** | **Age: 24**  **Gender: F** | **2:58:27** |  | **1:00:41** |  |
| <https://vdoto2.com/calculator>  **VDOT** | **1** | **VDOT: 53.4** | **3:00:28** | **6:53** | **1:00:05** | **6:27** |
| Garmin Race Predictor | **N/A** | **Personal VO2 Max: 64 (as measured by Garmin watch)** |  |  |  |  |

1. **Conclusion**

Fine-tuning these race predication models is an ongoing project. The accuracy of these models may benefit from incorporating more of the variable, intrinsic features that can have an impact on performance. For example, sleep, sickness, and injuries can all have a negative impact on performance and could be additional metrics to consider. Factors that typically lead to performance improvements, such as cross-training and nutrition, could also be considered.

**How to improve model**

**More factors to bring in**

1. **Sources**

FOOTNOTES

\* Kelvin Kiptum broke the marathon world record during the Chicago Marathon in 2023, completing the race with a time of 2:00:35, which was 34 seconds faster than Eliud Kipchoge’s previous record (Puelo, 2023). Kipchoge still holds the fastest marathon time recorded, running a 1:59:40 on a special course in Vienna, Austria in 2019. This race, however, did not meet the qualifications to be considered for the world record title (Angus, 2023).

\*\* The data in the avg\_temp column was added to the activity dataframe from an external weather data source and relied on the activity’s coordinates and date-time columns to populate. Thus, any null values in the start\_date\_local and start\_latlng columns had to be addressed first, then the avg\_temp column was populated, and finally the null values in the avg\_temp column could be addressed using the mean avg\_temp value to populate.

WORKS CITED – needs to be formatted / alphabetized

Angus, S. D. (2023, November 16). *Eliud Kipchoge broke the men’s marathon record by 30 seconds. how close is the official sub-2 hour barrier now?* The Conversation. https://theconversation.com/eliud-kipchoge-broke-the-mens-marathon-record-by-30-seconds-how-close-is-the-official-sub-2-hour-barrier-now-191421

Billat, V., Sirvent, P., Lepretre, P. M., & Koralsztein, J. P. (2004). Training effect on performance, substrate balance and blood lactate concentration at maximal lactate steady state in master endurance-runners. *Pflugers Archiv : European journal of physiology*, *447*(6), 875–883. <https://doi.org/10.1007/s00424-003-1215-8>

Biswal, A. (2023, April 3). *Sklearn linear regression (step-by-step explanation): Sklearn Tutorial*. Simplilearn.com. https://www.simplilearn.com/tutorials/scikit-learn-tutorial/sklearn-linear-regression-with-examples#:~:text=The%20scikit%2Dlearn%20library%20in,linear\_model%20module.

Brooks, G. A., Wolfel, E. E., Groves, B. M., Bender, P. R., Butterfield, G. E., Cymerman, A., Mazzeo, R. S., Sutton, J. R., Wolfe, R. R., & Reeves, J. T. (1992). Muscle accounts for glucose disposal but not blood lactate appearance during exercise after acclimatization to 4,300 m. *Journal of applied physiology (Bethesda, Md. : 1985)*, *72*(6), 2435–2445. <https://doi.org/10.1152/jappl.1992.72.6.2435>

*Cameron Race Time Prediction Formula and Calculator*. Running Time Prediction Calculator | Cameron Formula. (n.d.). <https://www.had2know.org/sports/race-performance-prediction-calculator-cameron.html>

Daniels, J. (2022, December 31). *Threshold training | how to use T-pace running to PR - runner’s world*. Runner’s World. https://www.runnersworld.com/advanced/a20807282/threshold-training/

*Demystifying running Vdot*. Tri2Max Coaching LLC. (n.d.). https://www.tri2max.com/demystifying-vdot#:~:text=Efficient%20runners%20could%20run%20at,one%20factor%20he%20called%20VDOT

Díaz, J. J., Renfree, A., Fernández-Ozcorta, E. J., Torres, M., & Santos-Concejero, J. (2019). Pacing and Performance in the 6 World Marathon Majors. *Frontiers in sports and active living*, *1*, 54. <https://doi.org/10.3389/fspor.2019.00054>

Diebel, S. R., Newhouse, I., Thompson, D. S., & Johnson, V. B. K. (2017, July 1). *Changes in running economy, respiratory exchange ratio and VO2max in runners following a 10-day altitude training camp*. International journal of exercise science. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5466410/>

Douglas-Walton, J. (2020, May 20). *A study of fitness trackers and Wearables*. HFE. <https://www.hfe.co.uk/blog/a-study-of-fitness-trackers-and-wearables/>

*Forerunner® 245/245 music*. Forerunner 245/245 Music Owners Manual - Viewing Your Predicted Race Times. (n.d.). https://www8.garmin.com/manuals/webhelp/forerunner245/EN-US/GUID-31B2458A-859A-4A34-AB83-224E4A29387A.html#:~:text=Your%20device%20uses%20the%20VO2,refine%20the%20race%20time%20estimates

Garmin. (2020, November). *Predicted race times on app or Garmin Connect website.* Predicted race times on app or Garmin Connect website. - Forerunner 645/645 M - Running/Multisport - Garmin Forums. <https://forums.garmin.com/sports-fitness/running-multisport/f/forerunner-645-645-m/244367/predicted-race-times-on-app-or-garmin-connect-website>

Haberman, Aaron L. (2017). Thousands of Solitary Runners Come Together: Individualism and Communitarianism in the 1970s Running Boom. *Journal of Sport History*, *44*(1), 35–49. <https://doi.org/10.5406/jsporthistory.44.1.0035>

Hardesty, L. (2017, April 14). *Explained: Neural networks*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Hatta H. (1990). Oxidative removal of lactate after strenuous exercise. *The Annals of physiological anthropology = Seiri Jinruigaku Kenkyukai kaishi*, *9*(2), 213–218. <https://doi.org/10.2114/ahs1983.9.213>

Haugen, T., Sandbakk, O., Seiler, S., & Tonnessen, E. (2022, April 1). *The training characteristics of world-class distance runners: An integration of scientific literature and results-proven practice - sports medicine - open*. SpringerOpen. https://sportsmedicine-open.springeropen.com/articles/10.1186/s40798-022-00438-7

Heins, W. (2023, January 26). *Garmin VO2 Max explained & metrics reviewed*. The Mother Runners. https://www.themotherrunners.com/garmin-vo2-max-explained-metrics-reviewed/

Jaadi, Z. (2023, August 4). *When and why to standardize your data*. Built In. https://builtin.com/data-science/when-and-why-standardize-your-data

Jordan, K. (n.d.). Coaches education - running training: Determining your current level of fitness. https://www.coacheseducation.com/endur/jack-daniels-nov-00.php

Joyner M. J. (2017). Physiological limits to endurance exercise performance: influence of sex. *The Journal of physiology*, *595*(9), 2949–2954. <https://doi.org/10.1113/JP272268>

Keras. (n.d.). *Keras 3 API documentation*. https://keras.io/

Kumar, D. (2021, June 20). *Introduction to data preprocessing in Machine Learning*. Medium. https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

Lara, B., Salinero, J. J., & Del Coso, J. (2014). The relationship between age and running time in elite marathoners is U-shaped. *Age (Dordrecht, Netherlands)*, *36*(2), 1003–1008. <https://doi.org/10.1007/s11357-013-9614-z>

Lewinson, E. (2022, August 21). *Three approaches to encoding time information as features for ML Models*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/three-approaches-to-encoding-time-information-as-features-for-ml-models/#:~:text=We%20can%20use%20the%20following,time%20feature%20into%20two%20features.&text=In%20the%20snippet%20below%2C%20we,using%20the%20sine%2Fcosine%20transformations

Lucia, A., Esteve-Lanao, J., Oliván, J., Gómez-Gallego, F., San Juan, A. F., Santiago, C., Pérez, M., Chamorro-Viña, C., & Foster, C. (2006). Physiological characteristics of the best Eritrean runners-exceptional running economy. *Applied physiology, nutrition, and metabolism = Physiologie appliquee, nutrition et metabolisme*, *31*(5), 530–540. <https://doi.org/10.1139/h06-029>

Marr, L. C., & Ely, M. R. (2010). Effect of air pollution on marathon running performance. *Medicine and science in sports and exercise*, *42*(3), 585–591. <https://doi.org/10.1249/MSS.0b013e3181b84a85>

Meteostat. (n.d.). *Python Library*. Meteostat Developers. https://dev.meteostat.net/python/

O'Toole ML, Hiller DB, Crosby LO, Douglas PS. The ultraendurance triathlete: a physiological profile. Med Sci Sports Exerc. 1987 Feb;19(1):45-50. PMID: 3821454.

*Pandas documentation*. pandas documentation - pandas 2.1.3 documentation. (n.d.). https://pandas.pydata.org/docs/

Parker B. A. (2011). The 2-hour marathon--more than physiology. *Journal of applied physiology (Bethesda, Md. : 1985)*, *110*(1), 281–294. <https://doi.org/10.1152/japplphysiol.01259.2010>

Plymire, D. C. (2004). Positive Addiction: Running and Human Potential in the 1970s. *Journal of Sport History*, *31*(3), 297–315. <http://www.jstor.org/stable/43611673>

Puleo, M. (2023, October 11). *When could the marathon world record be broken next?*. The Athletic. https://theathletic.com/4946674/2023/10/11/marathon-world-record-kelvin-kiptum-next-race/

*Race Finish Time Predictor*. SportTracks. (n.d.). https://sporttracks.mobi/labs/race-finish-time-predictor

*Race Time Calculator: Predict your race finish times*. Marathon Handbook. (2023, November 13). https://marathonhandbook.com/race-time-calculator/

Radford, S. (2023, February 4). *Riegel’s formula*. TrainAsONE. https://trainasone.com/ufaq/riegels-formula/#:~:text=T2%20%3D%20T1%20x%20(D2%20%2F%20D1)%20%5E%201.06&text=D1%20%3D%20Previous%20race%20distance,race%20time%20for%20upcoming%20race

Ritterbeck, M., & Williams, L. (2020, November 4). *How to Boost Your VO2 Max So Running Faster Feels Easier*. Runner’s World. <https://www.runnersworld.com/training/a26066029/increase-vo2-max/>

Rodrigo-Carranza, V., González-Mohíno, F., Santos Del Cerro, J., Santos-Concejero, J., & González-Ravé, J. M. (2021). Influence of advanced shoe technology on the top 100 annual performances in men's marathon from 2015 to 2019. *Scientific reports*, *11*(1), 22458. <https://doi.org/10.1038/s41598-021-01807-0>

Rush, T., & Barrell, A. (2022, January). *The role of lactic acid in the body*. Medical News Today. <https://www.medicalnewstoday.com/articles/326521#:~:text=It%20assists%20in%20cell%20respiration,to%20hyperlactatemia%20and%20lactic%20acidosis>.

Saha, D. (2023, January 25). *Top 10+ missing data imputation strategies in Pandas*. Medium. https://dsdojo.medium.com/top-6-missing-data-imputation-strategies-in-pandas-3a7090518e8a

Saunders PU, Telford RD, Pyne DB, Cunningham RB, Gore CJ, Hahn AG, Hawley JA; (n.d.). *Improved running economy in elite runners after 20 days of simulated moderate-altitude exposure*. Journal of applied physiology (Bethesda, Md. : 1985). <https://pubmed.ncbi.nlm.nih.gov/14607850/>

sci-kit learn. (n.d.). *Sklearn.linear\_model.linearregression*. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html

Sebastian, A. (2023, October 17). *Consumers are ready for personalized health care. data can make it happen.* MedicalEconomics. <https://www.medicaleconomics.com/view/consumers-are-ready-for-personalized-health-care-data-can-make-it-happen->

Simplilearn. (2023, August 16). *Introduction to data imputation: Simplilearn*. Simplilearn.com. https://www.simplilearn.com/data-imputation-article

Statista Research Department. (2023, September 14). *Running & jogging - statistics and Facts*. Statista. <https://www.statista.com/topics/1743/running-and-jogging/#topicOverview>

Strava. (2023, January 13). *Strava features complete overview*. Complete Overview of Strava Features. https://communityhub.strava.com/t5/athlete-knowledge-base/strava-features-complete-overview/ta-p/275

*Strava API V3*. Strava Developers. (n.d.). https://developers.strava.com/docs/reference/

Stojiljković, M. (2023, June 26). *Linear regression in python*. Real Python. https://realpython.com/linear-regression-in-python/#simple-linear-regression-with-scikit-learn

Thomson, Nuss, K., & Comstock, A. (2018, September 6). *Heart rate measures from the Apple Watch, Fitbit charge HR 2, and ...* Taylor and Francis Online. https://www.tandfonline.com/doi/full/10.1080/02640414.2018.1560644

UC Davis Health. (n.d.). *Lactate profile*. UC Davis Sports Medicine. <https://health.ucdavis.edu/sports-medicine/resources/lactate>

*VDOT Running Calculator*. V.O2 Running Calculator. (n.d.). https://vdoto2.com/calculator

Venturini, E., & Giallauria, F. (2022, March 2). *Factors influencing running performance during a marathon: Breaking the 2-H barrier*. Frontiers in cardiovascular medicine. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8924290/>

Vickers, A. J., & Vertosick, E. A. (2016, August 26). *An empirical study of race times in recreational endurance runners - BMC sports science, medicine and Rehabilitation*. BioMed Central. https://bmcsportsscimedrehabil.biomedcentral.com/articles/10.1186/s13102-016-0052-y

Walsh, E. (2023, February 7). *What is aerobic threshold: An Athlete’s Guide: Polar Blog*. Polar Journal. https://www.polar.com/blog/understand-aerobic-threshold/