Final Year Project

Human running performance from real-world big data

Thomas Thornton

Student ID: 18466574

A thesis submitted in part fulfilment of the degree of

BSc. (Hons.) in Computer Science

Supervisor: Barry Smyth

UCD School of Computer Science

University College Dublin

April 28, 2022

Table of Contents

Chapter 1: **Introduction**

This research paper focuses on predicting marathon performance from real world training data using machine learning models. The most important training features to improve race performance will also be looked for.

With the increase in popularity of running as a pass time there are more runners using exercise tracking equipment to track their training data and their running improvement. This provides an exciting amount of data, ready to be studied. The data for this project will come from these exercise tracking devices. I obtained this data from Strava Inc [1] as part of a data sharing agreement between University College Dublin and Strava Inc. This data opens up the opportunity to study the training and race performance of average people. In the past the only way to study running performance was in a lab with a lot of high tech equipment, which is often used to test extremely high level competitive runners.

This project focuses on average runners and the aim is to create a prediction model based on the data of average runners that can be used to predict the performance of other average runners. This paper will cover the process of extracting important training features from our data and inspecting how they correlate to race performance. At the end of this project these features will be combined and used to train various machine learning models. These models will then be tested and their performance for male and female runners will be compared.

The results will show such relationships as the relationship between long training runs and improved race times, and how a training schedule can effect a runner's race performance. At the end of the paper we will look at how an effective race performance prediction model can be created from the training data of average runners.

All code used for this paper can be found at this Gitlab repository [2].

https://csgitlab.ucd.ie/ThomasThornton/marathon-performance-real-world-data

Chapter 2: Project Specification

2.1 Problem Statement

Running has grown in popularity in recent years and has had a significant boom since the beginning of the COVID-19 pandemic. Naturally with the increase in recreational runners there has been an increase in the popularity of marathon running. Marathon running requires a lot of intensive training. It would be extremely helpful for an athlete to have a general idea of how they will perform in a race given the type of training they have been doing and the intensity of that training. A common sentiment in marathon training is that the number of long runs (>30km) in a training cycle is a major determinant of performance. Classic pieces of running advice like this will be studied in this paper.

RQ1. All else being equal, does the number of long runs during a training cycle affect marathon performance?

1. How many long runs are most effective in a training cycle?
2. Does the intensity of the long runs affect race performance?
3. Does the length of the longest run in training affect when a runner hits the wall in a race?

RQ2. All else being equal, what affect does training pace and training frequency have on marathon performance?

1. Does training at race pace improve a runner's halfway split?
2. What affect does training frequency have on a runner's race performance?
3. Is a consistent training schedule more effective than a disjointed schedule?

RQ3. Can we predict the performance of an athlete in a marathon given their training data and previous race performances?

1. Can we create models to predict runner performance from training data?
2. How do these models vary depending on the athlete's gender?

2.2 Background

Exercise tracking technology has become more and more prevalent in recent years [3], such as watches that track your speed and heart rate while running and other such data. The increase in the popularity of these devices has given data analysts and running enthusiasts a large volume of data to study.

This has allowed runners and coaches to get a more clear and concise view of their training and racing performances, however a more advanced study of this data could provide great benefit to an athlete. This project proposes to use this data to formulate a method that could help predict the performance of an athlete based on the quality and type of training they are doing. This project will look at which kinds of training variables and patterns have the biggest impact on race performance.

Figure 1.1 Exercise tracking smart watch

This project will look at classic pieces of training advice and put them to the test. By looking at the training data of 5 thousand runners of a variety of ages, experience levels and genders, this paper will evaluate if the classic training advice is effective.

The ultimate goal of this project is to assess which training features correlate to an improvement in race time and to create a model that can predict a runner's race time based on their training data.

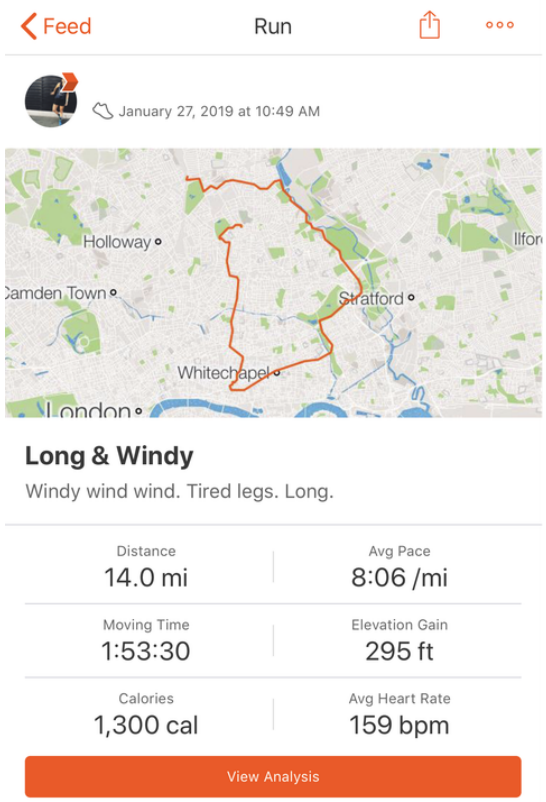

Figure 1.2 Strava tracking app

2.3 Datasets

This project will be using detailed training data from Strava [1]. Strava is a service used on many exercise tracking devices to track running and cycling data from GPS. I have obtained this information through a data sharing agreement between the university and Strava. The dataset contains the full training data of roughly 5,000 anonymous runners across roughly 1 million training sessions.

Chapter 3: Related Work and Ideas

3.1 Predicting performance from training and race data

Many papers have been written on the subject of predicting race performance based on running data [4–10]. Predicting marathon performance from training data and/or from previous race performances are both areas that have been broadly researched. The methods and results from these papers may be applicable to this study.

3.1.1 Predicting marathon time from training data

It is very common for runners to try and predict their marathon performance based on previous race performances. This is an important part of marathon preparation because it allows the runner to calculate a reasonable target finishing time. There are plenty of online calculators [11] which runners use to get an idea of how they are likely to perform in the marathon. It is much rarer that runners try to predict their marathon time based on training performance. This is because it is much more complex to try and predict running performance from a runners training schedule and, until relatively recently, it was very difficult to track the exact amount an athlete was training without having a personal coach and lots of expensive equipment. With the growing popularity of exercise tracking technology [3] it has become more viable to use training data of average runners to predict marathon performance.

The inspiration for this research project comes from the study "Human running performance from real-world big data" by Thorsten Emig and Jussi Peltonen [4]. In this paper Emig and Peltonen used running data from 19k runners to create models which would predict the race performance of a runner. Emig and Peltonen used calculations from a study by Emig et al. [5] along with some new calculations to find the "performance indices" [4] of a runner. By calculating endurance, power and training intensity they could create a model that was reasonably accurate at predicting race performance.

They also discuss some of their key findings from their model. There is a "strong sensitivity of performance to endurance" [4]. They found that increasing training intensity and training frequency was only beneficial to a runner up to a certain point. Too much training caused the runner to burn out over multiple weeks or get injured or other such problems.

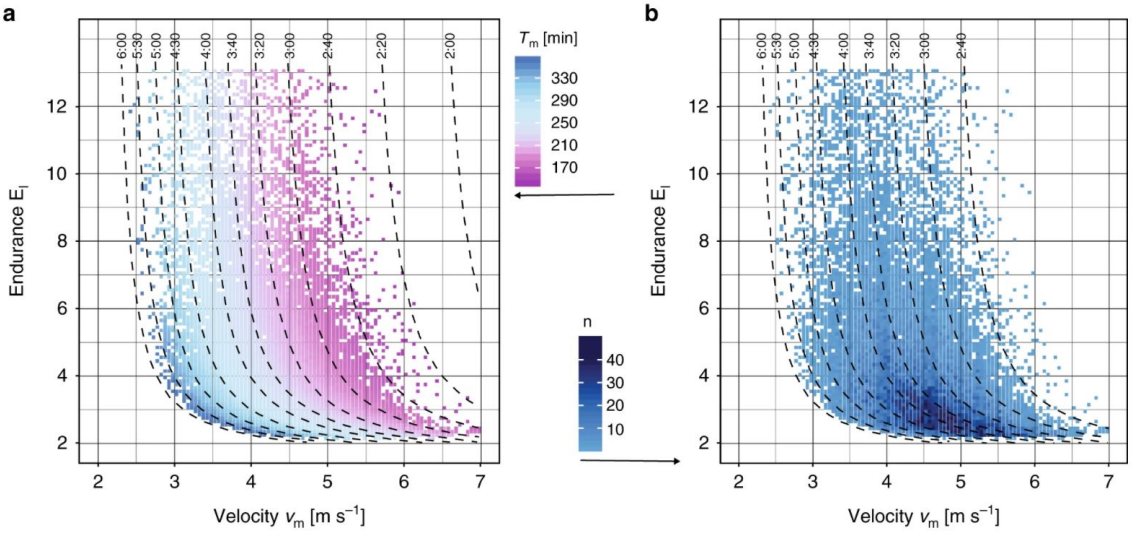

Figure 2.1: Correlation between performance indices and marathon race time. [4]

A problem with this study, which Emig and Peltonen mention themselves, is that the data they used only contained the final time and distance of an individual's running session. This could lead to incorrect velocity if the runner stopped for a rest. The data that will be used in this project contains the average velocity for every 100m segment of the run, which will provide more helpful insight than final total time and distance.

Hoch and Pemmaraju carried out a similar study [6] to that of Emig and Peltonen. Hoch and Pemmaraju gathered their data from Strava, which is how we have gathered our data. They started by selecting features in training data such as "average weekly distance", "moving time", "total active days", "number of long runs" [6]. The features selected by Hoch and Pemmaraju could be help us decide the features to look at in this paper. They gathered these features by using Spark [12]. They then used XGBoost [13] to create a model to predict marathon performance. Hoch and Pemmaraju tested their model on a few specific runners. They found that their model was significantly better than previous race time calculation techniques in "both accuracy and precision" [6].

3.1.2 Predicting marathon time from shorter races

The online calculators [11] to predict marathon time that were mentioned earlier have been created from their own studies. In the study "An empirical study of race times in recreational endurance runners" [7], by Andrew J. Vickers and Emily A. Vertosick they go in to depth on creating a model to predict marathon performance from shorter race times. A key aspect of this study is based upon the fact that many studies in this area "have typically involved elite athletes [and] small sample sizes" [7]. The data for this study was acquired from a questionnaire of average runners which asked for information such as the runner's age, gender, injury history, BMI and so on. Using this data they created two models, one model which used the performance from one previous race and one model which used the performance from multiple previous races. They found that "the commonly used Riegel formula" [7] predicted times that were 10 minutes too quick for the majority of runners. The mean squared error of the two models they created were smaller than the error for the Riegel formula. This study focused on providing a model for average runners without the need of expensive equipment to measure oxygen intake and other such factors, our study has a very similar goal but we aim to create a more comprehensive model using more accurate training data from thousands of runners. The data used by Vickers and Vertosick is also likely to be somewhat biased as it was self-reported by the runners in a questionnaire.

It should be mentioned, Emig and Peltonen used a model in their study in combination with training data to create a new model which predicts marathon performance. The model they used was created by Emig et al. (2018) [5]. This model was created by using race performance of a large data set of runners.

3.2 Race pacing and training intensity

A crucial part of this project is understanding performance from race pace. We can't simply look at speed of a runner, this could vary depending on experience, age, gender. It would be helpful for us to understand how well a runner is performing based on their race splits and when/if they hit the wall.

3.2.1 Race pacing

Race pacing has a big impact on performance in a race. Starting too fast can cause a runner to "hit the wall" but starting too slow could mean the runner performed below their maximum potential.

The supervisor of this project, Barry Smyth, carried out a study on runners hitting the wall in marathons [9]. The data used in the study contained over 4 million individual race performances from over 2 million unique runners. The data also contained the gender and age range of each runner. Hitting the wall "refer[s] to the sudden onset of debilitating fatigue that can occur late in the race" [9]. A runner is classified as having hit the wall when they slow down by a certain amount for a long enough distance after the 20km mark. This method to define when a runner has hit the wall will be helpful for research question two in this paper. Smyth found that a higher proportion of men than women hit the wall in a race, he also found "evidence that younger runners are more likely to hit the wall" [9].

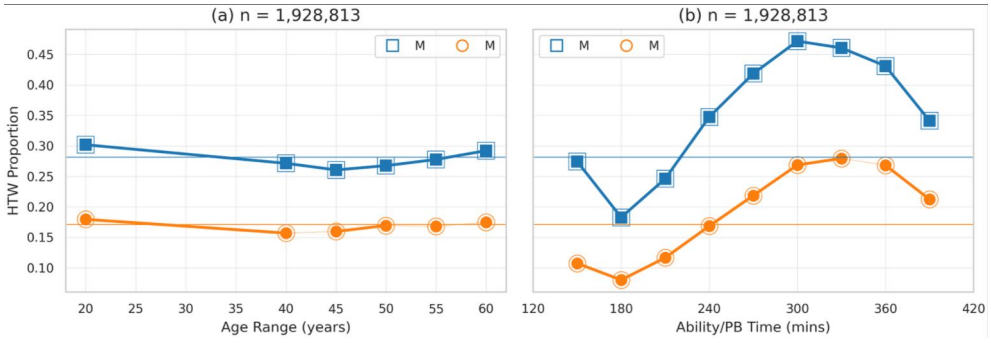

Figure 2.2: The proportion of male and female runners hitting the wall by (a) age range and (b) ability level. [9]

Our paper aims to recreate results similar to this. Smyth also calculated the estimated cost of hitting the wall on a runners race time. This is beyond how much we will look at the effect of hitting the wall in a race, our aim in this paper is to see the effect of training on when, or if, a runner will hit the wall.

"Fast starters and slow finishers" [8] is another paper by Barry Smyth. This paper looks at the impact of pacing on a runner's race time. Does starting slow and finishing fast work better than

starting fast and finishing slow, or is an even split more effective? This subject has been studied many times by sports scientist, but generally on small sample sizes of elite level athletes. Smyth's paper uses data from over "1.7 million recreational runners" [8]. The paper found that, for men and women, fast starts resulted in slower finishing times. It also found that slow starts resulted in slower finishing times than running the race at an even split. These results were consistent across a "variety of conditions" [8]. The paper also demonstrated that runners who started fast were far more likely to "hit the wall". I intend to expand on these results by looking at how training schedule can affect a runner's split and when they hit the wall.

3.2.2 Training intensity

Training with intensity obviously improves a runner's running ability. the problem for many athletes is finding the right balance of intensity. Training too lightly could make the runner perform a lot lower than their potential but training too intensively could cause a runner to get burned out or injured.

Esteve-Lanao, J. et al. [10] carried out a study to compare the impact of two different training regiments. One training regiment with a low intensity and one with a high intensity. This study was carried out on "Twenty competitive subelite (regional to national level, competition experience 5 years) male Spanish runners" [10]. They split the subjects in to two groups and gave both groups identical training loads with differing intensities. Both groups ran a 10.4km race before and after their 5 month training regiment. The results of the studied showed that, while there was an improvement in the times of both groups, the group that trained at a lower intensity had a much more significant improvement in time.

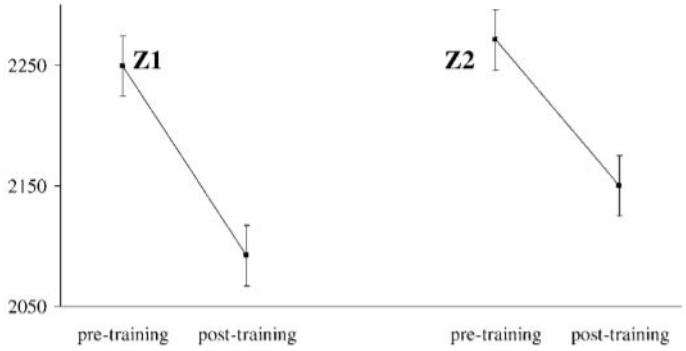

Figure 2.2: Change in performance after the training period during the simulated 10.4-km cross-country race in both groups. [10]

The problem with this study is that the only subjects are experienced, male, Spanish athletes. One aim of my paper is to find the impact of training intensity on runners of varying age, gender or experience.

Emig and Peltonen also covered training intensity in their study of "Human running performance from real-world big data" [4]. They found that a higher intensity in training correlated to an improvement in race performance, but only up to a certain point. After a certain level of training intensity most runners performance started to decline.

3.3 Traditional marathon training advice

In marathon training there are some training methods that are widely accepted as the best technique for training for a marathon. This paper will look at some of these traditional pieces of training advice to find if there is truth to them. In this [14] article by Boston Athletic Association mentions that "The long training runs of over 18 miles are the most important workouts in any training program" [14]. Most training programs from HalHigdon echo this sentiment "The key to the program is the long run on weekends" [15]. The importance of long runs is something that will be tested in this paper. Both articles also mention the importance of training at race pace "One of the most important factors in marathon training is tempo running" [14]. This paper will look at the importance of training at race pace and how that effects the runner's performance. This paper will also discuss the effects of training at race pace too frequently.

Chapter 4: Project Workplan

The Gantt chart in Figure 3.1 below outlines the planned order of operations for this project. I plan to start work on this project on the 10th of January 2022. My supervisor will have already gathered the data necessary for this project courtesy of Strava. I have planned 3 weeks for data pre-processing. I have already familiarised myself with a sample of the data and have done some basic cleaning and I feel that 3 weeks is more than enough time for pre-processing. Research question 1 and research question 2 have both been given 15 days. I have allotted 3 weeks of time for research question 3 as I believe this will be the most difficult question to answer. The report will start after research question 1 is finished. The majority of the work for the report won't come until all research questions are answered but I plan on updating parts of the report while I am still answering RQ2 and RQ3.

- Data gathering: My supervisor will send me the data from Strava when I am ready to start working.
- Data pre-processing: Cleaning and processing the data to have it ready for each research question.
- Research question 1: Using the processed data to answer each part of research question 1.
- Research question 2: Using the processed data to answer each part of research question 2.
- Research question 3: Using the processed data to answer each part of research question 3. I expect this question to be the most difficult to answer.
- Report: Will start writing the final report when I finish research question 1.

Figure 3.1: Plan

Chapter 5: Description of Approach

5.1 Overview

The data for this project was given to us from Strava Inc[1] as part of a data sharing agreement between Strava Inc and UCD. The original dataset was comprised of 922,362 running sessions from 5,000 unique runners. Before any cleaning or filtering the data contained 930 female runners and 4070 male runners. The data contained the gender of the runner and the runner's age. It also contained each running session the runner carried out while using their tracking equipment. The data tracked from these sessions included the columns 'startdatelocal', 'totaldistance', 'heartrate_100', and 'pace_diff_100'. 'startdatelocal' stored the date and time that the this session started, 'totaldistance' stored the distance that was run in this session in meters, 'heartrate_100' contained an array of the runner's heart-rate broken in to 100m intervals and 'pace_diff_100' contained an array of the pace that the runner ran at broken down in to 100m segments. The structure of the original dataset can be seen in Figure 5.1 below.

	athlete_id	activity_id	sex	age	startdatelocal	totaldistance	heartrate_100	pace_diff_100	race_id
0	3490	29909151	M	35.0	2015-11-08 07:37:55	17611.0	[93. 121.82321 124.52922 126.049995 1...	[9.9999997e-06 5.2627940e+00 5.1457191e+00 5.9...	-1
1	3490	20922315	M	36.0	2016-04-20 14:20:47	9385.7	[65. 95.62404 110.85987 121.14349 1...	[9.9999997e-06 2.8561039e+00 9.9999997e-06 5.5...	-1
2	3490	21885198	M	37.0	2017-09-25 17:17:44	6018.9	[86. 118.68496 130.27728 134.70168 136.0...	[1.000000e-05 4.052059e+00 4.912369e+00 4.3280...	-1

Figure 5.1: Structure of data

The data was static for the entirety of the project and was stored locally. The identity of each runner was anonymised before I received it by demand of Strava. A full description of the contents of the data can be found in section 6.1, which details the data left after cleaning and filtering.

5.2 All else being equal, does the number of long runs during a training cycle affect marathon performance?

This research question looks at how the number of long runs and type of long runs affects how a runner performs in a marathon. To answer this research question we first need to assign a target race for each runner. The most recent marathon race by each runner was selected as the target race for each runner. All training data more than 12 weeks before the target race was removed from the dataset as this could be considered outside the training window for the marathon.

5.2.1 RQ1.1 Number of long runs

This sub question of the research question focuses on how the number of long runs affects a runner's marathon performance. The difference in effect on males vs females and the difference in effect on 3 different age ranges is also studied in this section. 'Gender' and 'Age' of the runner are already in the dataset. 'totaldistance' also allows us to easily note how many long sessions a runner has run. There are two definitions of long sessions that will be used. One long session is 30km and the other is 20km.

5.2.2 RQ1.2 Intensity of long runs

This sub question looks at how the intensity of long runs affects race performance. Intensity was measured using the 'heartrate_100' column, the 'age' column and a basic formula for the maximum heart rate of a person [16]. The mean heart-rate of a runner during their training session divided by the theoretical maximum heart-rate of a person that runner's age gave us the intensity of the training session, this is a common way of measuring training intensity [17]. The affect of intensity of long runs on race performance is also compared between men and women in this section.

5.2.3 RQ1.3 Length of longest run

In this sub question we study the affect of the length of the longest run on how a runner performs. The longest run in the 12-week training window before the marathon was used as the runner's longest run. The length of this run is then compared to race time and if this specific runner hit the wall or not. Two separate methods for hitting the wall were used and compared.

5.3 All else being equal, what affect does training pace and training frequency have on marathon performance?

"One of the most important factors in marathon training is tempo running" [14], in this question the validity of this statement is tested. It is also feels intuitive that training more frequently and on a consistent basis would lead to better race performances. These intuitions are also tested in this question. They are tested by checking for runners hitting the wall, checking how their pace varies throughout the race and how their heart-rate varies throughout the race.

5.3.1 RQ2.1 Training at race pace effect on halfway split

The halfway split is the split in time it takes to run the first half of the marathon compared to the second half of the marathon. A positive split means the runner ran slower in the second half of the race than the first half. A neutral split means they ran the same speed in both halves of the race. This section also largely looks at how training at race pace affects a runner's probability of hitting the wall during the marathon race. To answer this question a 'split_diff' column was created which compared the speed between the first and second half of the race. A column called

'hit_wall' was also created which tracked whether a runner hit the wall or not. The formula for hitting the wall can be found in this paper, 'How recreational marathon runners hit the wall', by B. Smyth [9]. These two new columns were compared against the number of long runs a runner completed in the weeks leading up to the race.

5.3.2 RQ2.2 Effect of training frequency

In this section the effect of training frequency on race performance is assessed. To do this a simple calculation of sessions per week was carried out and compared to various race performance indicators such as race time, coefficient of variation of pace and coefficient of variation of heart-rate.

5.3.3 RQ2.3 Importance of a consistent schedule

A common intuition is that a regular training schedule is very important for running performance. This intuition is tested in this section by calculating the largest gap between training days per runner, the minimum gap between training days per runner and the mean gap between training days per runner. The column 'day_std_var' was created which is the standard variation in the gap between training days. These columns are then compared to the same race performance indicators as before.

5.4 Can we predict the performance of an athlete in a marathon given their training data?

For this research question multiple various machine learning models will use the features that have been created while answering the previous questions to create a regression model that can predict marathon performance.

5.4.1 RQ3.1 Marathon performance model

A variety of machine learning models are used to predict the performance of a runner given their training data. Linear regression, decision tree, random forest and gradient boost[13] are all tested. Each model uses all the features that were created in the previous questions such as the number of long runs and other features.

5.4.2 RQ3.2 Marathon performance model for Men vs Women

This section is much the same as the previous section but the data is split in to male and female. A model is created for each gender and compared to see if different features tended to be more important for women than men and vice-versa.

Chapter 6: Methodology

6.1 Initial Data Processing

The data did not come in perfect condition so some initial cleaning was required before new features could be created that could be used to answer each research question. The first thing to sort out was null data cells. In the columns containing integers or string objects there were no missing values, but in the two columns containing lists, 'heartrate_100' and 'pace_diff_100', there turned out to be many empty lists. Both columns combined to have 256,937 rows with at least one empty list. Both lists are crucial to the rest of the project so all these rows had to be dropped. The pace of the runner, stored in 'pace_diff_100' was stored as minutes per kilometer. To make the data easier to work with and more recognisable to the average person the pace was converted to meters per second.

Another issue with the lists was that every list was stored as a string so each list in the dataset had to be converted from a string to a list. This was accomplished using the replace() and eval() methods in python.

Some of these lists contained inaccurate data, I suspect due to the GPS in the exercise tracking devices miss-tracking the runner which is a common issue with GPS [18]. These errors also could've been caused by misclassification of data, such as a cycle being stored as a run. The main problem was with the 'pace_diff_100' column. There were some runners supposedly running at a faster speed than Usain Bolt reached during his 100m world record race [19]. I added a limit of 9.5m/s to the 'pace_diff_100' column, any value greater than 9.5m/s in any of the lists in the 'pace_diff_100' column were changed to 9.5m/s. There was also a lower limit of 1m/s added as this is roughly the speed of a very slow walking pace [20].

The lists in the column 'heartrate_100' had some flaws too. There were some runners that had impossibly high heart-rates so a limit of 220 was added to the heart-rate lists. 220 was picked because this is generally regarded as the maximum human heart-rate [16]. A lower limit of 40 was also added to the heart-rate list. It is not impossible to have a heart-rate below 40 but it is exceptionally unlikely for someone to have a heart-rate below 40 while exercising. Finally the 'heartrate_100' and 'pace_diff_100' lists had to be checked in each column to make sure they were the same length. If they were not the same length it meant there was an error in the data.

By the end of the cleaning there were 119,245 running sessions across 3,308 unique runners. 618 [20.3%] of those runners identified as female and the remaining 2,690 [79.7%] identified as male. A breakdown of some important features can be seen in Figure 6.1 below.

Column1	No. runners	Mean sesh per week	Mean race time	Percent hit the wall	Mean age	Mean long sessions	Mean 20km sessions
Male	4070	3.57	4.02	0.96%	39	1.33	5.77
Female	930	3.82	4.45	0.43%	37	1.37	5.61

Figure 6.1: Breakdown of dataset.

The mean age for male runners was 39 while the mean age for female runners was 37. Runners completed fewer than 2 long runs in a training cycle but close to 6 20km sessions. We can also see that a very low number of runners hit the wall, less than 1% of male runners hit the wall and less than half of 1% of female runners hit the wall.

6.2 Research Question 1

All else being equal, does the number of long runs during a training cycle affect marathon performance?

Long runs are an important part of preparation for a marathon. If a runner has never ran close to a marathon distance before it will be hard for them to run the marathon in the race. There is no one definitive definition of a long run in marathon training. In the training data a long run was defined as a session where the runner ran more than 30km. There was a second definition of long runs created as 20km. This is to compare how different definitions of long run compare to one another.

6.2.1 RQ1.1 Number of long runs

After the data was cleaned new features could be calculated. The first new features were 'heartrate_mean' and 'pace_mean'. These were calculated using the stats trim_mean function from the Scipy library [21]. Trimmed mean was calculated because it is a very noisy dataset with a lot of outliers, even after cleaning. Then the standard variation of the heartrate and pace lists were calculated, 'heartrate_std_dev' and 'pace_std_dev' were the names of the new columns. These two columns were then used to calculate the coefficient of variation of heart-rate and pace. The coefficient of variation of pace and heart-rate were two methods of evaluating how a runner performed in the marathon. Runners that perform well in marathons tend to have even splits throughout the race [22], meaning they do not vary their speed much. This means that a runner with a low coefficient of variation of pace in the race was regarded as someone with a good performance. This is important because it is difficult to judge how runners performed based on race time alone. A runner may have finished quicker than someone else even while performing worse than how they could have performed.

After calculating these features they were compared primarily using a series of boxplots. Multiple boxplot graphs were made that compared the number of long runs and the data was split by male and female or by age ranges. These age ranges were chosen somewhat arbitrarily but also in conjunction with research from a study about the peak age of athletes [23]. The age ranges chosen were <32 years old, 33-45, and 46+

6.2.2 RQ1.2 Intensity of long runs

Intensity was a new feature created for this question. To calculate intensity the mean heart-rate for each session needed to be calculated. The pandas apply function was used in conjunction with the Scipy [21] trim_mean function to calculate the mean heartrate for each running session. The max heart-rate for each runner was then calculated using the method from this [16] article about calculating maximum heartrate. This function was applied to the whole dataframe simply by subtracting each runner's age from 220. Pandas makes this type of calculation extremely quick for large dataframes. Intensity was then calculated for each runner by dividing the 'heartrate_mean' column by the 'max_heartrate' column [17].

$$df[intensity] = \frac{df[heartrate\_mean]}{df[max\_heartrate]} \tag{6.1}$$

Where df is the name of the dataframe

Runners were then grouped in to bands of 5% intervals of intensity and split by gender. This can be plotted as a line chart against race time and coefficient variation of pace. This will also be split by age range to see how different ranges of intensity in training effect race performance. Intensity in long runs was compared to race time and coefficient variation of pace.

6.2.3 RQ1.3 Length of longest run

First feature necessary for this question was the 'long_session' feature. This is a binary column that contains a 1 if the training session was longer than 30km and 0 if the session was shorter than 30km. The second new feature was identical but the limit was set at 20km, '20k_session'. The next thing necessary here was to calculate if the runner hit the wall during the race or not. 2 separate methods were used for this. The first algorithm used was taken from this [9] paper by Barry Smyth and adjusted slightly for the sake of this project.

Calculating hitting the wall in this fashion requires several steps. A new list needed to be created called 'pace_per_km' which was a list of the mean speed of the runner over 1km intervals. This list is then used to calculate the base pace. The base pace is a reference pace used to judge if the runner has hit the wall or not. The calculation for base pace again comes from "How recreational marathon runners hit the wall" [9]. The base pace (BP) is calculated as the average speed of a runner from the 5km to 20km segment of the race, "we exclude the initial 5km segment because pacing during the very early stages of the marathon tends to be more erratic" [9].

$$BP = \frac{\sum_{4<seg<21} pace(r,seg)}{15} \tag{6.2}$$

The base pace and race pace were then converted from kilometers per hour to minutes per kilometer. Finally the degree of slowdown (DoS) of the runner for each segment of the second half of the race is calculated. Equation 6.3 below shows the Degree of Slowdown formula, which has been again been taken from B.Smyth [9].

$$DoS(r, seg) = \frac{pace(r,seg)}{BP(seg)} - 1 \tag{6.3}$$

Now that degree of slowdown has been calculated it is possible to calculate if a runner hit the wall. If the runner had a degree of slowdown of 1.25 or greater that was defined as hitting the wall.

The second definition for hitting the wall was based on the halfway split. A new feature called 'split_percent' was calculated by dividing the runners second half split by their first half split. If a runner had a 'split_percent' greater than 1 that meant they ran faster in the second half of the race than the first. If they had a 'split_percent' less than 1 that meant the second split was slower than the first. A runner that had a 'split_percent' less than 0.95 in the race was considered as hitting the wall. The length of the longest run was then plotted against whether a runner hit the wall or not. This was done using a boxplot and a scatter plot. Separate graphs were used for each definition of hitting the wall.

6.3 Research Question 2

All else being equal, what affect does training pace and training frequency have on marathon performance?

In this question the goal is to analyse the importance of training at race pace and training frequency. Race pace is simply the pace that the runner will be running at during the race. The number of these sessions will be compared to the race performance features established earlier: race time, coefficient of variation of pace and coefficient of variation of heart rate.

6.3.1 RQ2.1 Training at race pace effect on halfway split

First off, the race pace for each runner needs to be created. This is done by getting the base pace of each runner in the marathon. The runner's base pace is taken to be their race pace as this is the pace that the runner tends towards for the race after the clutter of the start of the race has cleared and before they may have hit the wall. After getting base pace a new dataframe is created with each runner's ID number, date of the race and their base pace in the race. This new dataframe is then merged with the original dataframe on 'athlete_id' and 'target_race' to add the race pace to each training row, in Figure 6.1 you can see how 'target_race' is now a feature in each training session.

	athlete_id	weeks_until_race	startdatelocal	activity_id	sex	age	totaldistance	target_race
0	12	12.0	2017-07-01 18:11:09	15021704	M	32.0	12335.5	2017-09-24 08:15:09
1	12	12.0	2017-07-02 06:57:14	22275865	M	32.0	6934.1	2017-09-24 08:15:09
2	12	11.0	2017-07-02 19:55:49	16475019	M	32.0	6988.8	2017-09-24 08:15:09

Figure 6.1: Dataframe structure with 'target_race' column (some columns are filtered out for readability)

Next a new feature called 'train_rp' is created which is a Boolean values that says whether a runner ran at race pace during their training session. A custom function (Algorithm 1 below) was used, which determined if the runner ran at race pace by checking if their mean pace in training was within 0.25m/s in a positive and negative direction. If the runner ran at race pace in that particular training session then 1 was stored in the 'train_rp' column, if they didn't run at race pace then 0 was stored.

Algorithm 1 Training at race pace

if $mean\ pace\ in\ training < base\ pace\ in\ race + 0.25$ AND
$mean\ pace\ in\ training > base\ pace\ in\ race - 0.25$ **then**
$train\_rp \leftarrow 1$
else
$train\_rp \leftarrow 0$
end if

The number of runs at race pace was then plotted against the split percent (the ratio of the time taken between the first half and second half of the race). The number of runs at race pace was also plotted against whether the runner hit the wall or not. We then compared B. Smyth's [9] version of hitting the wall and our own version of hitting the wall, which is based on the halfway split. The comparison of the effect on men vs women is also plotted in this section.

6.3.2 RQ2.2 Effect of training frequency

To test training frequency the feature 'sessions_per_week' was created which is simply the mean of the number of sessions a runner ran each week. This is done by grouping each training session in week intervals going back 12 weeks from the race. The number of sessions in each week are then counted. We then find the mean number of sessions per week by summing these counts and dividing by the number of weeks, this can be seen in Equation 6.4 below.

$$SPW = \frac{\sum_{i=0}^{w} S(i)}{w} \tag{6.4}$$

Where S(i) = number of sessions in week i, w is the total number of weeks a runner trained and SPW is the mean sessions per week

This was then scatter plotted against the race performance indicators that have been used before. The effect on men and women are compared. A second plot where the runners are grouped by the number of training sessions they do per week are grouped. These grouped runners are then evaluated by their race performance indicators.

6.3.3 RQ2.3 Importance of a consistent schedule

In this section three new features were created. Max training gap, mean training gap and standard deviation of training gap. Max training gap is the largest gap in days that was taken between two training sessions over the course of the 12 week training schedule. Mean training gap is the mean gap between training sessions in this time and standard deviation of training gap is the standard deviation of the gap between training days over the 12 weeks.The equation for max training gap can be seen in Equation 6.5 and 6.6. The calculation for min training gap and standard deviation of training gap was calculated similarly.

$$DST = date\_of\_current\_session - date\_of\_previous\_session \tag{6.5}$$

Where DST is Days since training,

$$MaxDST(r) = max(DST(r, s)) \tag{6.6}$$

Where MaxDST(r) is the max days since training for runner r and DST(r,s) is the set of 'days_since_training' for each session s for runner r

These 3 features were then compared for each runner against the race performance indicators. Again, these features were also grouped by number of days and plotted as a line graph against the race performance indicators as well.

6.4 Research Question 3

Can we predict the performance of an athlete in a marathon given their training data and previous race performances?

This section is a culmination of the results from Research Question 1 and Research Question 2. A variety of data models will be tested using the features that were created in the previous two research questions.

6.4.1 RQ3.1 Marathon performance model

In this section various machine learning regression models are created and tested to try and find the most effective ML model. First, certain columns are removed from the dataset that have no correlation to the performance of a runner such as the runner's ID and the race ID. Features from the race itself were also removed such as the race pace. These features needed to be removed because the model would simply look at the pace of the runner and use that to calculate the runner's time. Categorical data is then transformed to make it workable with the ML models. The dataset is then scaled. The technique for this was taken from this article by Angelica Do Luca [24]. K-fold cross validation is also used to thoroughly test the regression models. The sklearn model_selection KFold [25] library is used to perform the k-fold cross validation.

The regression models that were evaluated in this section are the scikit learn linear regression model [26], the scikit learn decision tree regressor model [27], the accelerated gradient boost regressor [13] and the scikit learn random forest regressor [28]. Each model was trained and tested. The predicted results were plotted against the actual results on a scatter plot. The R-squared score and mean squared error for each model was also computed. For the decision tree and random forest model these were run through a loop to test for which tree depth was most accurate. The most important features for each model was also calculated and returned to see which training features are most important to a runner.

6.4.2 RQ3.2 Marathon performance model for Men vs Women

This section is much the same process as the previous section except a separate model is created for men and women. The data is split by the 'sex' column and the rest of the model evaluation is carried out in the same way.

Chapter 7: Results and Discussion

7.1 Research Question 1

All else being equal, does the number of long runs during a training cycle affect marathon performance?

7.1.1 RQ1.1 The effectiveness of long runs

Figure 7.1 shows the number of 30km long runs by a runner compared to their race time in the form of a boxplot. The number of runs longer than 30km is on the x-axis with race time on the y-axis. The data is split by male and female, male runners are represented by blue and female runners are represented by red. This graph shows a clear correlation between running long runs in training and an improved marathon race time.

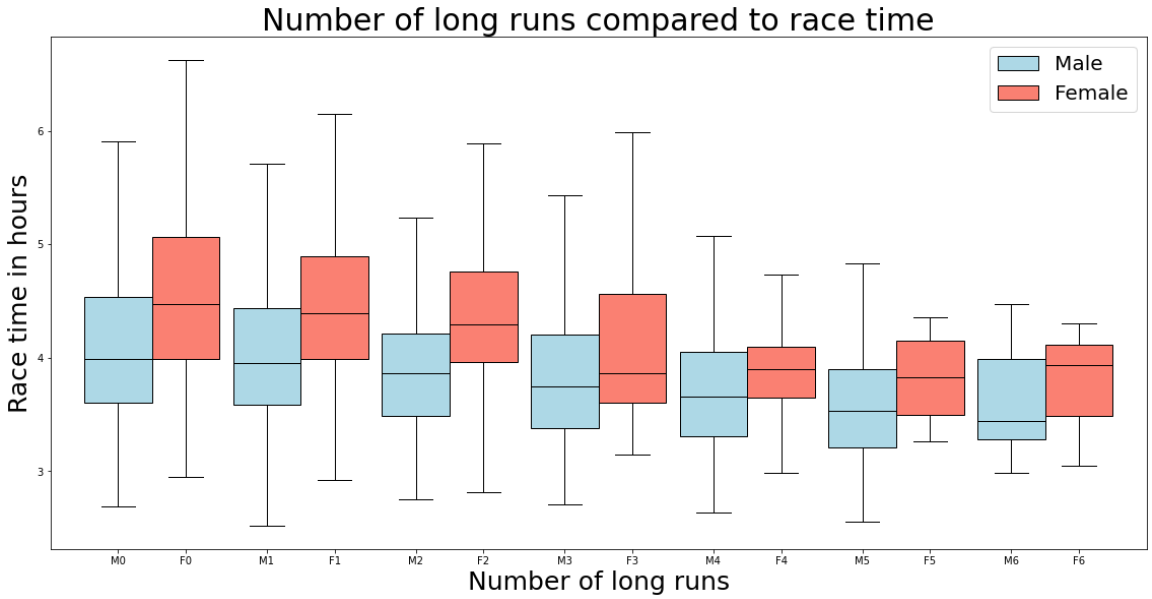

Figure 7.1: Number of long runs compared to race time by gender.

This correlation applies to male and female runners and it holds true for all age ranges too, see Figure A.1 in the Appendix. These results backup the quote from the Boston Athletic Association article [14] that "The long training runs of over 18 miles are the most important workouts in any training program". We cannot yet confirm that they are the most important workout but they do appear to be effective. It should be noted that it is impossible to confirm if this correlation is causation. Do long runs lead to better race times or do better runners just tend to complete more long runs?

Figure 7.2 is structured in a similar fashion to Figure 7.1 but instead of race time on the y-axis there is coefficient of variation of pace in the marathon. This was one of our other race performance measurement methods.

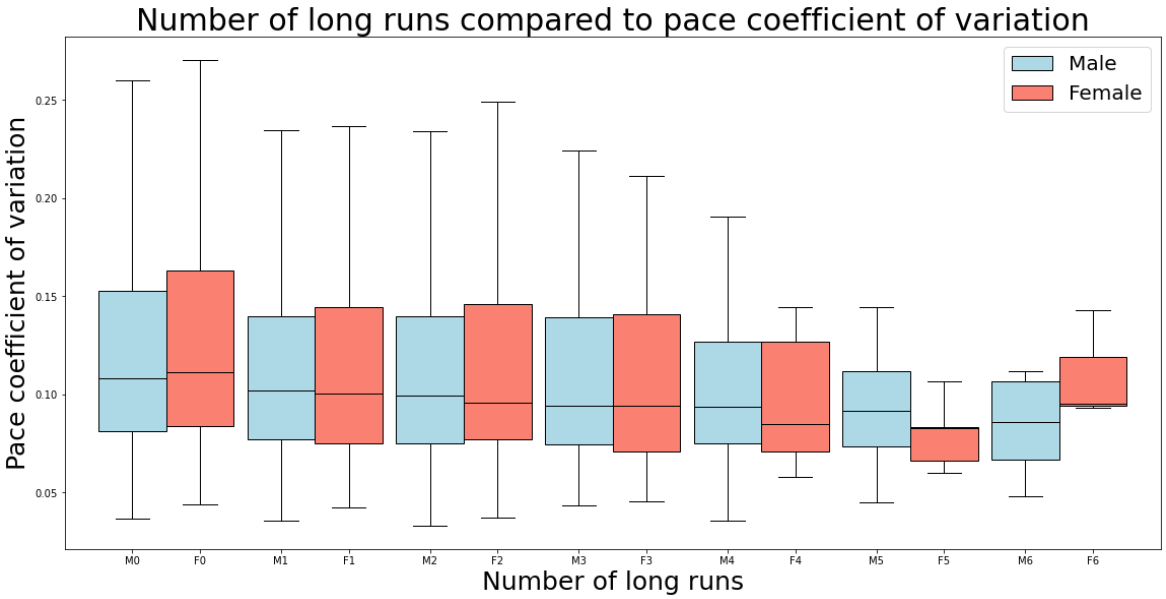

Figure 7.2: Number of long runs compared to coefficient of variation of pace by gender.

We can see that the relationship is not nearly as strong for coefficient of variation of pace. There is a clear improvement in coefficient of variation of pace from 0 long runs to 1 long run but after that there is not much of an improvement. There is some improvement but not much. There appears to be less deviation in the coefficient of variation of pace as the number of long runs increases but this is simply because fewer runners ran 6 long runs rather than 2 long runs. Overall, there may not be a large improvement in the consistency of a runner's pace based on the number of long runs they are running but there is a clear improvement in race time.

Figure 7.3 compares how effective long runs are when they are defined as 20km rather than 30km. This graph is not split by gender, instead it is split by long runs defined as 30km and long runs defined as 20km. Purple is 30km long runs and beige is 20km long runs. Race time is on the y-axis again.

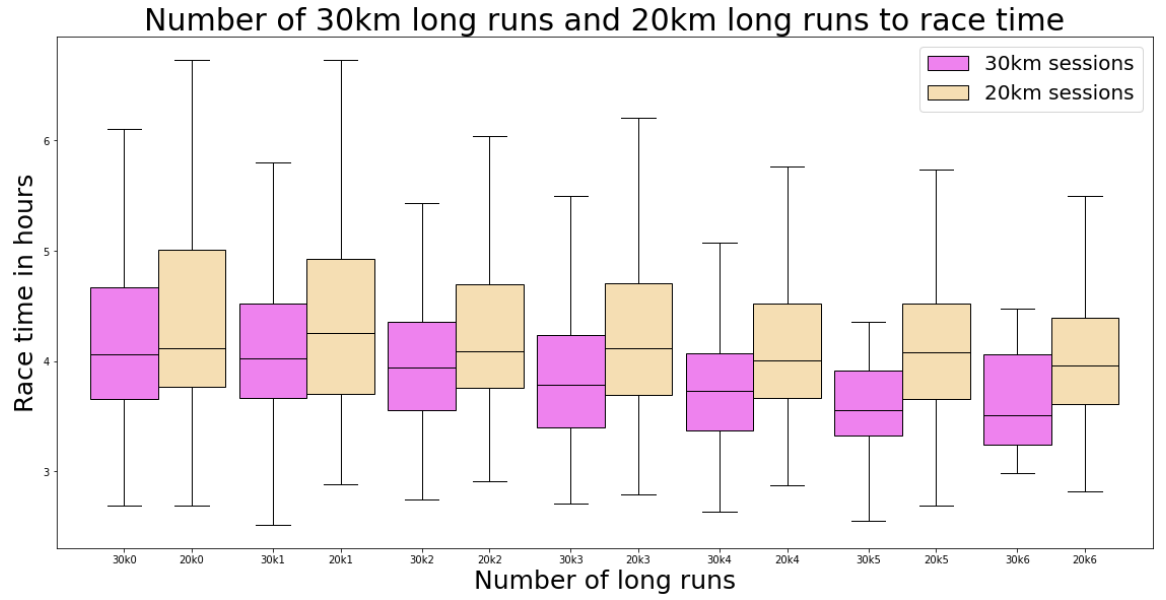

Figure 7.3: Number of long runs compared to race time by 30km and 20km long runs.

There is a significant difference in race time in favour of runners that ran 30km long runs. This

tells us it is very important for a runner to complete some runs near marathon distance for them to be successful in the marathon race. In Figure A.2 we can see that there is still not much of a relationship between the number of long runs ran by a runner and coefficient of variation of pace even when a long run is defined as 20km.

7.1.2 RQ1.2 Does intensity matter?

Naturally, one would expect training intensity to be a large part of improving one's running abilities. It appears that this instinct is only partially correct, in Figure 7.4 we can see that, particularly for men, higher training intensity during long runs (30km) resulted in faster race times, until a point. This graph shows the race time of runners grouped by their training intensity. The x-axis shows the intensity and the y-axis shows the race time, split by male and female.

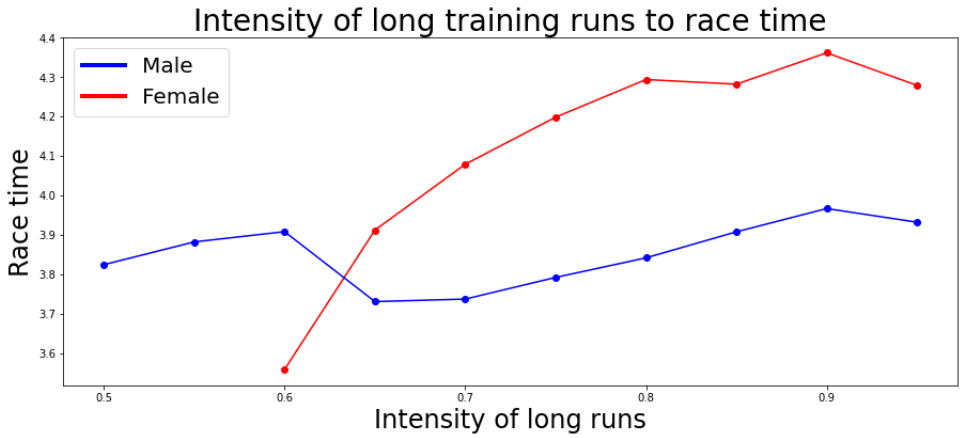

Figure 7.4: Intensity compared to race time.

We can see that an intensity of around 0.65 correlates to the best race time for male and female runners. The race time for women at an intensity of 0.6 can be attributed to outliers. It is possible that an intensity of around 0.65 is the ideal training rate for runners to improve as much as possible without burning out, this is a similar result to what we saw in this paper [10]. However, we cannot rule out that this correlation is not causation of race time, but that advanced runners don't reach a heartrate as high as a novice runner while training simply because running is easier for an advanced runner. Figure A.3 shows that this relationship holds true when a long run is considered to be 20km. Figure 7.5 below compares intensity on the x-axis with coefficient of variation of race pace on the y-axis. This figure shows that, again, 0.65 appears to correlate with the best race performance.

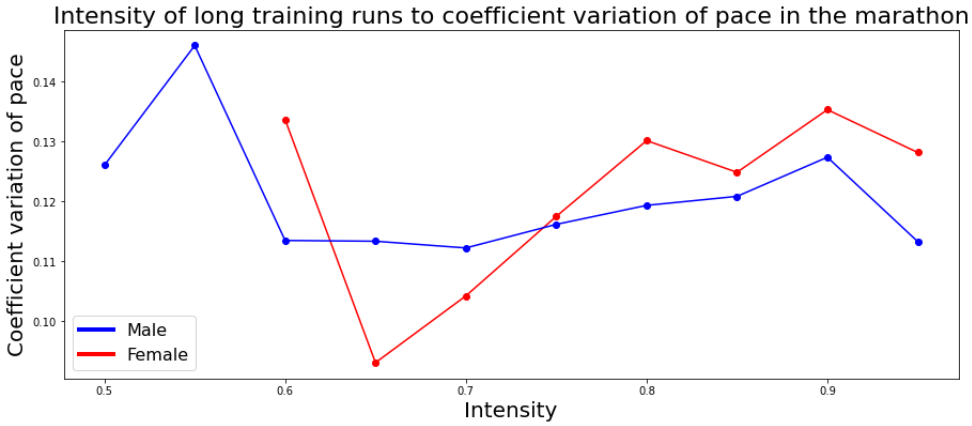

Figure 7.5: Intensity compared to coefficient of variation of pace in the race.

7.1.3 RQ1.3 Length of long run effect on hitting the wall

We have seen that intensity of long runs has a positive relation to race time, now we will see a positive correlation between the length of the longest run and hitting the wall. First we will look at hitting the wall based on degree of slowdown [9]. Figure 7.6 below shows the distance of the runners longest run on the x-axis. Runners that hit the wall are marked in red and separated from runners that did not hit the wall.

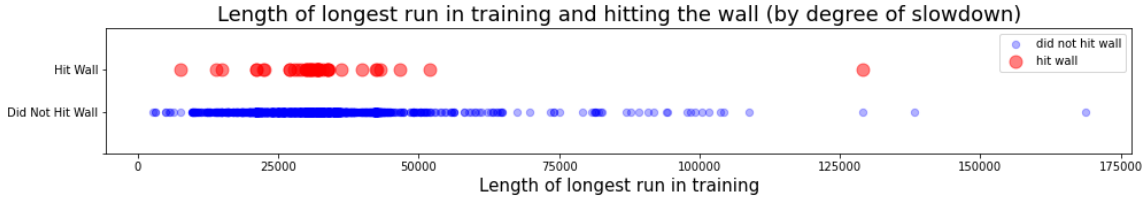

Figure 7.6: Length of longest run and if the runner hit the wall by degree of slowdown.

It's clear from Figure 7.6 that runners who's longest run in training was less than 37km are far more likely to hit the wall. There were only two runners that ran a session of over 50km in training that hit the wall in the marathon. A marathon is 42.2km long. This clearly tells us that a runner should attempt at least 1 run in training near the distance of a marathon to get a better feel for the pace they will need to maintain in the race. Figure 7.7 below shows our own method for determining if a runner hit the wall. It was slightly less lenient when determining a runner hitting the wall.

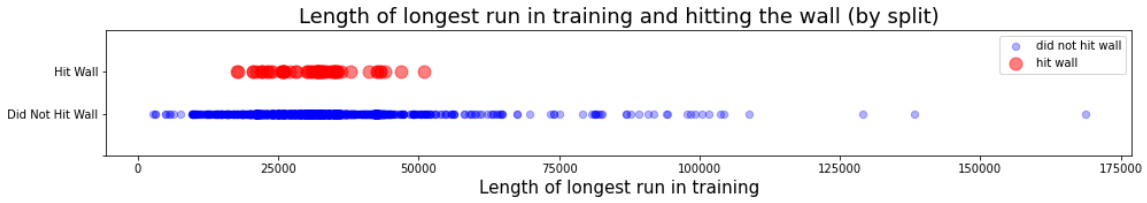

Figure 7.7: Length of longest run and if the runner hit the wall by split.

Figure 7.7 shows the method of labelling a runner as hitting the wall based on the amount they slowed down in the second half of the marathon. There are more runners that hit the wall that ran a training session over 37km in this figure, but again it's clear that a practice run longer than 50km means there is very little chance of a runner hitting the wall in the race. Not because they become immune to hitting the wall, but because they are more comfortable with the pace needed to complete the race.

7.2 Research Question 2

All else being equal, what affect does training pace and training frequency have on marathon performance?

7.2.1 RQ2.1 Training at race pace

Training at race pace is a crucial aspect of marathon training. At first glance from our data the number of sessions at race pace had a small correlation to an improved halfway split of a runner, this can be seen in Figure 7.8 below. This graph shows the number of 15km sessions a runner ran at race pace on the x-axis and the halfway split on the y-axis. The limit of 15km was set for a session at race pace, this was to filter out training sessions that were only a kilometer or so long. Sessions shorter than 15km are too short to have any meaningful relation to the marathon pace.

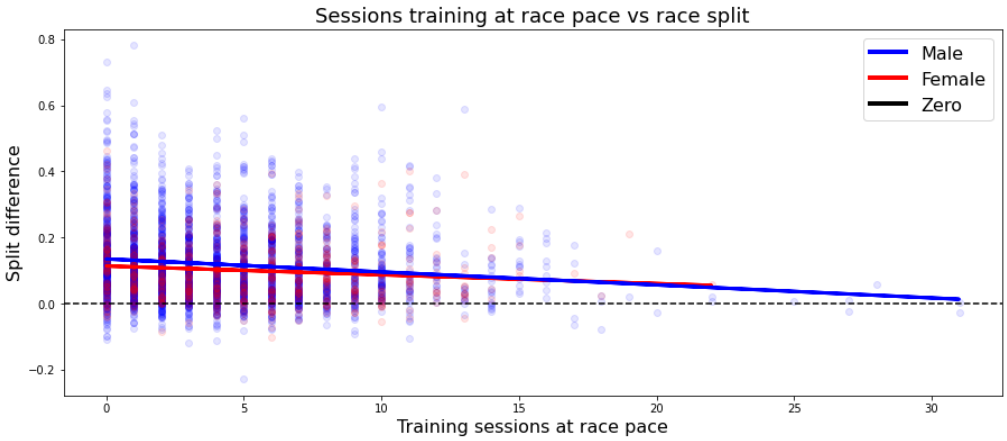

Figure 7.8: Number of runs at race pace compared to halfway split.

A halfway split greater than 0 means the runner ran slower in the 2nd half of the race. A halfway split below 0 means the runner ran faster in the 2nd half of the race and a halfway split of 0 means the runner ran the same speed in both halves of the race. It's clear that runners who completed more training sessions at race pace tended closer towards a halfway split of 0. This is a good indication of a well run race because runners with a halfway split above 0 means they ran too hard in the first half of the race and burned out in the second half of the race. Runners with a halfway split of less than 0 are very rare but this means a runner ran faster in the second half of the race, meaning they went too easy in the first half. However, when we look at training at race pace compared to hitting the wall (by halfway split) we find that training at race pace can be a double edged sword. Figure 7.9 shows the number of sessions longer than 15km a runner ran at race pace on the y-axis and the x-axis shows whether they hit the wall or not. The data is split by male and female represented by blue and red respectively.

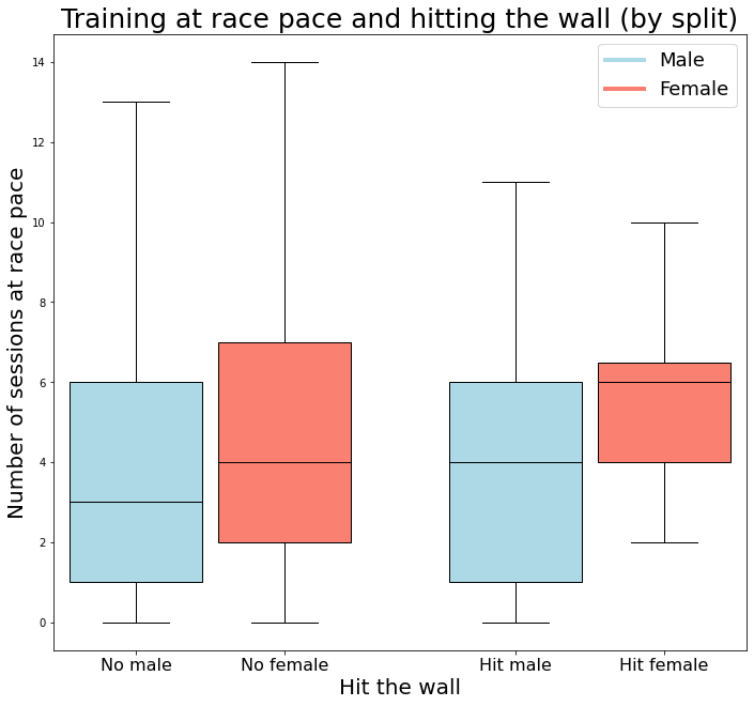

Figure 7.9: Length of longest run and hitting the wall by halfway split.

It appears that training at race pace may make you slightly more likely to hit the wall in a race. This may be because runners choose a pace in training that is too fast for them and they don't realise this until they're halfway through the marathon. It should be noted that it is unlikely that race pace was dependent on a runner hitting the wall because each runner's base pace in the race was taken as their race pace and it is unlikely for a runner to hit the wall during the first half of the marathon.

7.2.2 RQ2.2 Training frequency

The hypothesis that training more frequently leads to faster race times is a simple one but one that should be confirmed nonetheless. Perhaps there is a limit where too much training leads to burnout which could result in a worse race time. Figure 7.10 shows the mean number of training sessions per week for a runner on the x-axis compared to their race time on the y-axis. The linear regression line for male and female runners are plotted over the individual result of each runner.

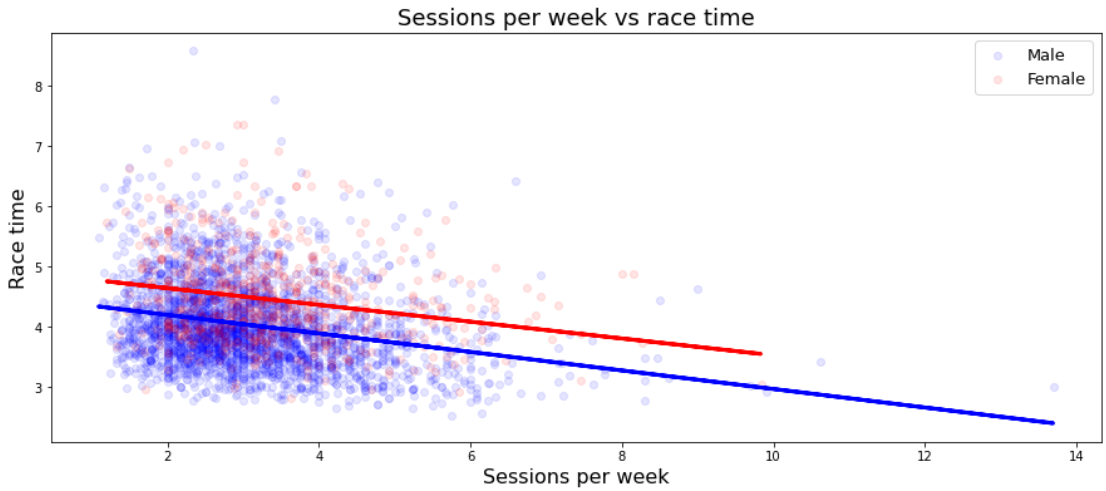

Figure 7.10: Mean sessions per week to race time split by gender.

There is a clear improvement in race time as the number of training sessions per week increases. Again, it can not be stated for certain that sessions per week is causation of a faster race time or if faster runners just train more. The coefficient of variation of pace shows a similar result to race times improvement. Figure 7.11 shows runners grouped by sessions per week on the x-axis and their respective coefficient of variation of pace in the race on the y-axis.

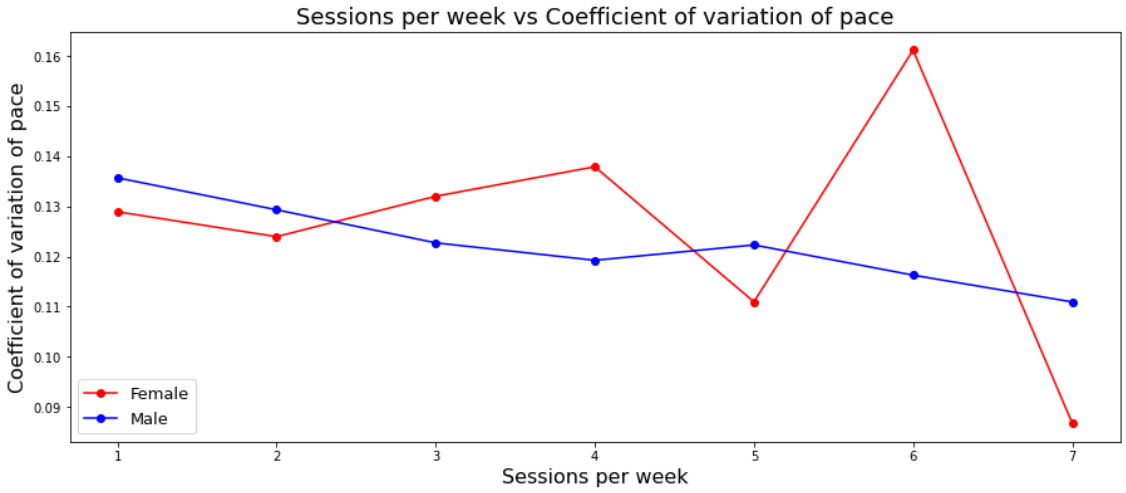

Figure 7.11: Mean sessions per week to coefficient of variation of pace split by gender.

There is a clear improvement in male runners pacing ability during a race with an increase in the number of sessions per week. The results for female runners is very erratic but appears to be trending towards a lower coefficient of variation of pace with an increase in sessions per week. Clearly sessions per week is very important to the performance of a runner in the marathon, both for speed and for running a consistent pace.

7.2.3 RQ2.3 Importance of a consistent schedule

A consistent training schedule is common for many runners but does a consistent schedule improve race performance? Figure 7.12 shows the mean number of days between training on the x-axis and race time on the y-axis. The mean training gap has an obvious effect on the race time of a runner.

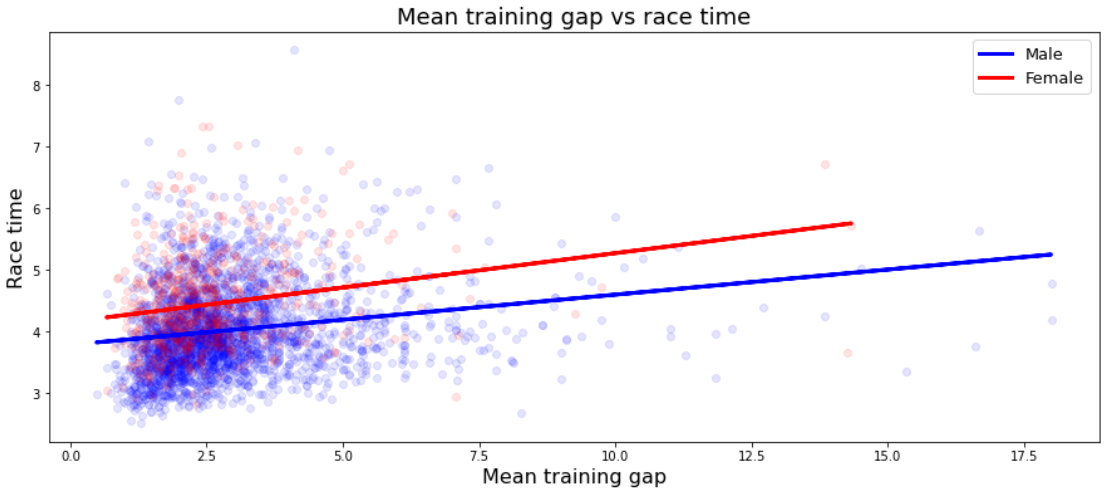

Figure 7.12: Mean days between training sessions to race time.

The standard deviation in the training gap should also be checked. A runner with a large mean training gap may still be training consistently, they just leave a longer gap between sessions than other runners. Figure 7.13 shows the standard deviation of the number of days between training on the x-axis and the runner's race time on the y-axis.

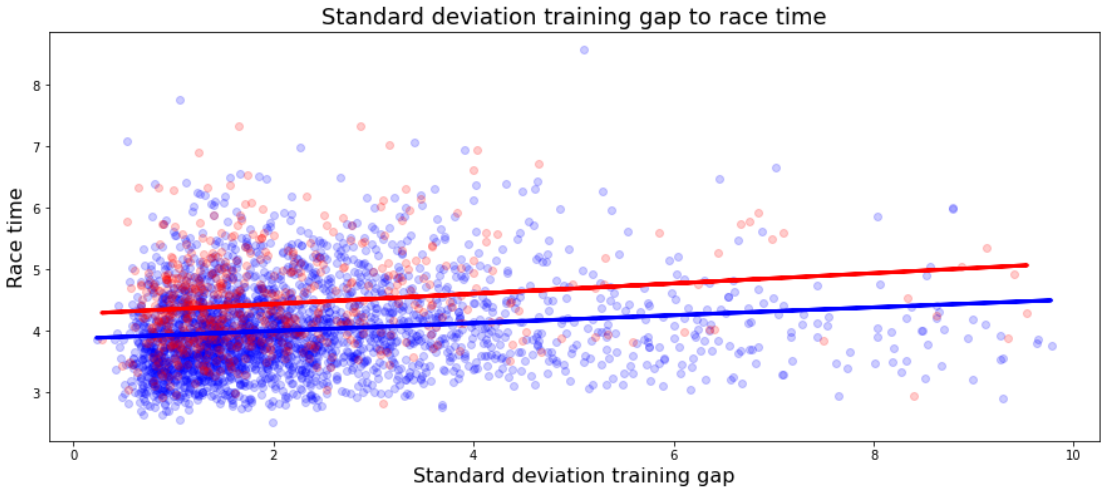

Figure 7.13: Standard deviation of days between training to race time.

The relationship here is weaker than the mean days between training but there is an obvious slowdown in race time with a larger standard deviation. In Figure A.5 you can also see that there is a very slight increase in the coefficient of variation of pace in the race when the standard deviation of days between training increases.

7.3 Research Question 3

Can we predict the performance of an athlete in a marathon given their training data and previous race performances?

7.3.1 RQ3.1 Marathon performance model

All the features that were created in the previous 2 research questions were brought together and run through 4 different machine learning regression models: linear regression [26], decision tree regression [27], gradient boost [13] and random forest regression [28]. Up until this point the dataset has mostly been in the form of 1 row per session for each runner. This means if athlete 12 ran 50 sessions and athlete 13 ran 20 sessions then the dataset would have 70 rows containing both runners training sessions. If we try to train a machine learning model on this data set it would cause overfitting, so each athlete was grouped and we got the mean, min, max and standard deviation values from their training features. Now there is only 1 row for each runner and each row contains features that summarised their training, such as the mean gap between training and max pace. Each model was also tested using K-fold cross validation with 10 splits.

Important features

First, let's look at which features each model found to be the most important in predicting the runner's marathon time. Gradient boost, random forest and decision tree all returned similar features as one another. They each found 'pace_mean', which is the mean pace a runner ran during all of their training, to be by far the most important feature in predicting race time. Those three models also shared the same top 3 features in terms of importance. Figure 7.14 shows the most important features returned from decision tree at max depth 5. The reason max depth 5 was chosen was because this was the best performing level of decision tree which can be seen in Figure A.6.

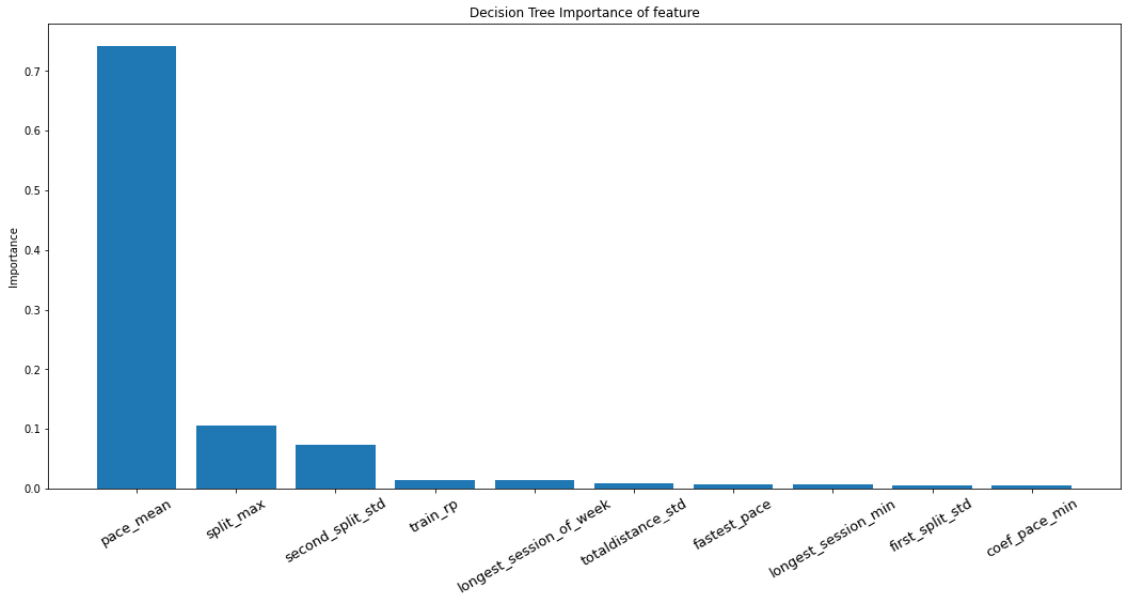

Figure 7.14: Most important features in predicting marathon time for decision tree.

It's clear that 'pace_mean' is by far the most important feature, Figure A.7 and Figure A.8 show the important features for gradient boost and random forest respectively. The other 2 most important features are 'split_max' and 'second_split_std'. The feature 'split' is the length of the runners second split in a training session in seconds minus the first split of the training session in seconds, therefore the 'max_split' is the largest improvement in time in the second half of a training session compared to the first half of that training session. I suspect this feature is helpful to the training models because it shows how well a runner can maintain their pace in the second half of the marathon when they're tired. The feature 'second_split_std' is the standard deviation of all the second splits the runner ran in training. This feature essentially tracked how erratic the

second half of a runner's training session was compared to their other training sessions. The linear regression model returned significantly different features when measuring importance. Figure 7.15 shows these features in order of importance.

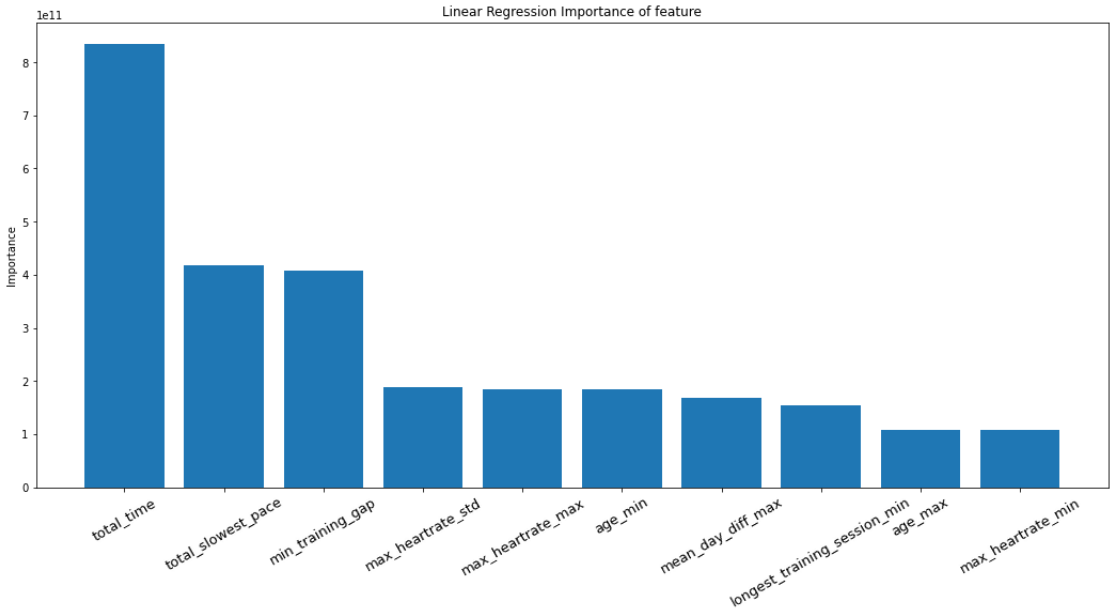

Figure 7.15: Most important features in predicting marathon time for linear regression.

This time 'total_time' was the most important feature. This feature has a slightly misleading name, it is not the total time a runner spent training it's actually the mean length of time of a runner's training sessions in seconds. Runner's averaging longer training sessions would have a higher 'total_time'. This result in conjunction with our findings from research question 1 tells us that performing long runs in training is one of the most important factors for success in the marathon. 'total_slowest_pace' is the slowest pace that a runner ran at in a training session, this pace was likely from a long run which would explain why it would be relevant to predicting marathon time, as their pace in a long run would naturally be similar to their pace in the marathon.

Model performance

Gradient boost performed best of all the models, followed closely by linear regression. Figure 7.16 shows a table containing the R-squared score, the mean absolute error and the mean squared error of each model. The mean absolute error and the mean squared error are the errors between the actual time and the predicted time which have been scaled between 0 and 1. This means an absolute mean error of 0.03 is roughly equivalent to an error of 3%.

	R-Squared	Mean Absolute Error	Mean Squared Error
Linear	0.807930	0.036017	0.002843
Decision Tree	0.628646	0.055515	0.005472
Gradient Boost	0.828678	0.034865	0.002521
Random Forest	0.695454	0.049826	0.004502

Figure 7.16: Accuracy of prediction models

Both tree models struggled to produce a model as accurate as linear and gradient boost. Figure 7.17 below shows the predicted time vs actual time for each runner. Figure A.9 shows this same

graph for the linear regression model. The x-axis shows the actual time and the y-axis shows the predicted time with the regression line plotted over it. Figure A.10 shows the scatter plot for random forest.

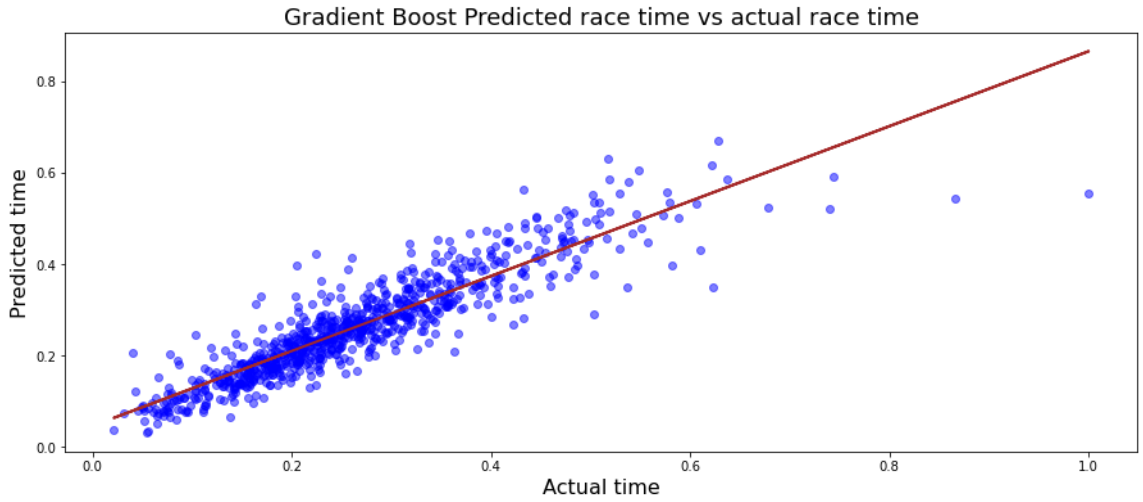

Figure 7.17: Gradient boost predicted time vs actual time.

It's clear that gradient boost did an excellent job in predicting race time based on the training data. The gradient boost model received an R-squared score of 0.829. This means that 82% of the target data can be explained from the training features. It's clear that the model struggled with runners that took a very long time to finish. I suspect these runners were rather inexperienced and either hit the wall or had to stop at some points during the marathon to rest, causing them to under perform based on their predicted time.

7.3.2 RQ3.2 Marathon performance model for Men vs Women

This section is carried out in the same way as the previous section but before any models were tested the data was split in to male and female data. After the data was split the min, max, mean and standard deviation features were all merged to create a male dataframe and a female dataframe. New models were then trained and tested in both datasets.

Important features

The important features for male and female athletes often varied but the most important features remained the same. For decision tree, gradient boost and random forest 'pace_mean' remained the most important feature. However for linear regression 'total_slowest_pace' became the most important feature for male and females, this was previously the second most important feature before the split. Figure 7.18 and Figure 7.19 shows the most important features for males and females, respectively.

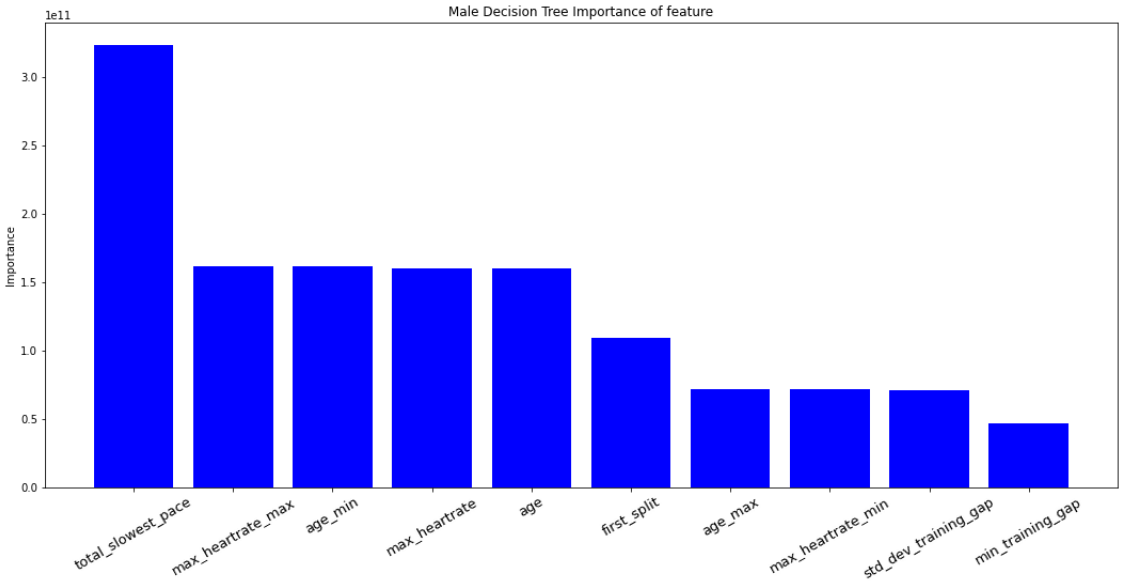

Figure 7.18: Linear regression important features for male runners.

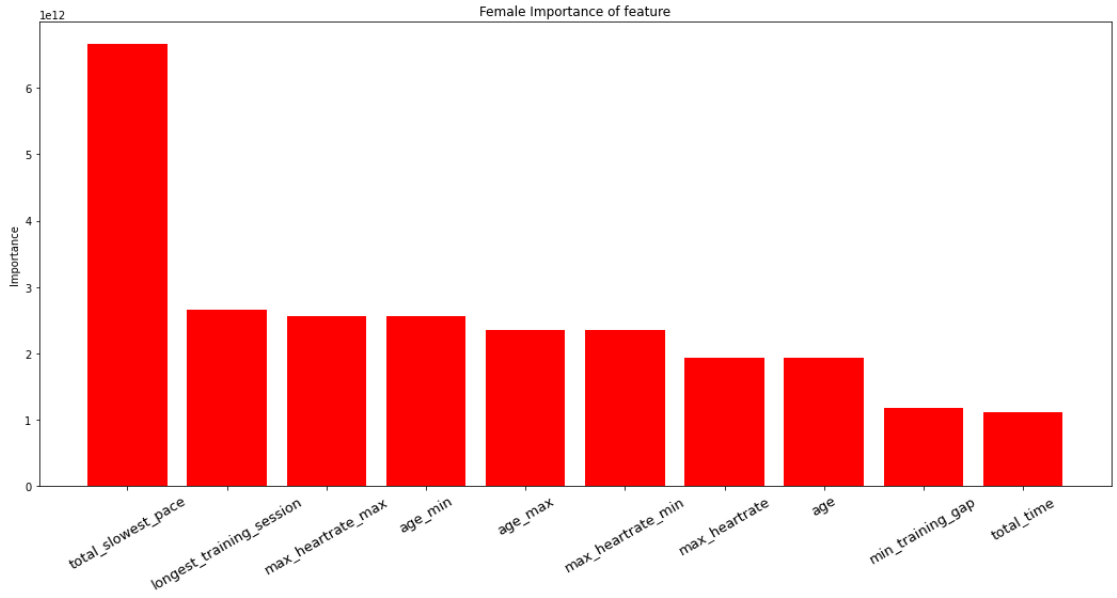

Figure 7.19: Linear regression important features for female runners.

After 'total_slowest_pace' there is an assortment of features based on the runner's age and max heartrate reached in training. The features for males and females for the other 3 models were very similar to the features chosen by those models in the previous section. The most important features in the gradient boost model for male and female runners can be seen in Figures A.11 and A.12.

Model performance

Gradient boost performed best of all the models for male runners and for female runners. Figure 7.20 shows a table containing the R-squared score, the mean absolute error and the mean squared error of each model. I believe an error occurred somewhere while calculating the R-Squared score for the linear regression models for female runners. When testing a random split generally an R-squared score of 0.77 was found. One of these random splits with an R-squared of 0.77 is

what returned the important features in the previous subsection. There is also a very low mean absolute error for linear regression for female runners.

	R-Squared M	R-Squared F	Mean Absolute Error M	Mean Absolute Error F	Mean Squared Error M	Mean Squared Error F
Linear	0.784985	-1.331425	0.036098	0.065395	0.002911	0.052881
Decision Tree	0.618188	0.624954	0.053656	0.074560	0.005266	0.009748
Gradient Boost	0.830671	0.748179	0.033823	0.057931	0.002324	0.006677
Random Forest	0.678435	0.687084	0.048926	0.069097	0.004413	0.008175

Figure 7.20: Accuracy of prediction models by male and female

Gradient boost was again the best prediction model. It was significantly better at predicting male runners than female runners. The sample size of male runners is much larger than the sample size of female runner, this is likely the cause of the discrepancy. Figure 7.21 below shows the scatter plot of male runner and female runner predictions from the gradient boost model.

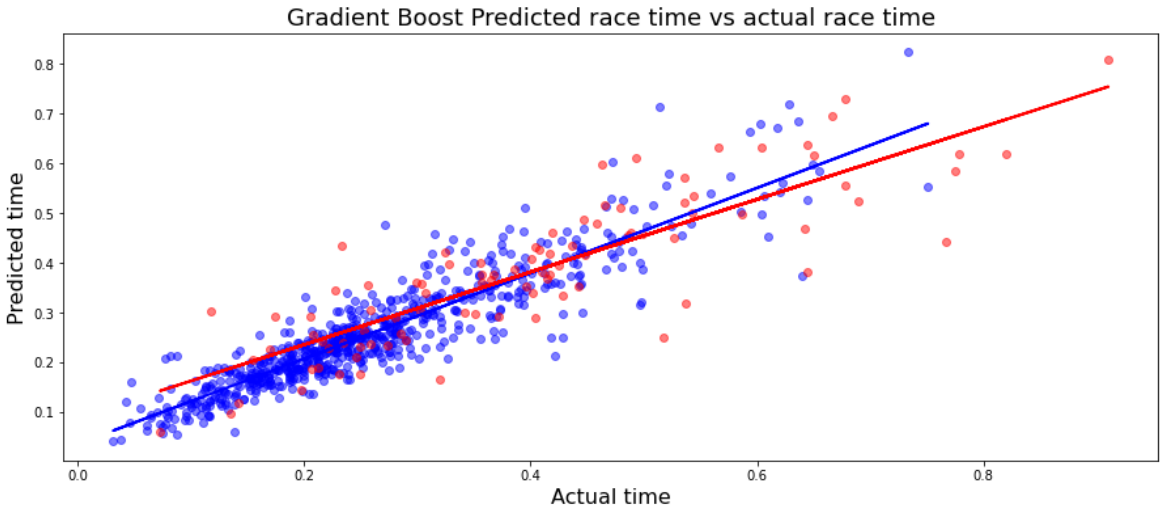

Figure 7.21: Gradient boost actual time vs predicted time for male and female runners.

There is a very clear and strong prediction accuracy for male runners. The prediction for female runners is noticeably more scattered.

Chapter 8: **Future Work**

Improve prediction model

I think it would be interesting to continue working on this model and discover other training features which could be helpful in predicting marathon performance. I would like to develop this data so that it could predict races of different lengths rather than just marathons. How well would some of the features designed for predicting marathon performance, like number of long runs, transfer to a race of shorter length?

Consumer-facing race time prediction app

Another idea that came to mind during the course of this project was to potentially create an app that has the linear regression model from this project built in to it. Runners could pair it to their exercise tracking equipment and the model would predict their marathon time so they could use it as a tool to estimate what they should aim for their race pace to be. I would also like if the data from that runner was added to the training data for the next training cycle and next race. Weights could be added to each row of data and any data from the owner of the app would have a higher importance than the rest of the data in the training set. This could hopefully improve prediction results for a runner as they continue training for more marathons.

Geographic training data

It would also be interesting to include more geographic data in the dataset. Altitude training is very popular among competitive athletes so it would be interesting to include altitude data in the prediction model and check how sessions at altitude compare to the race performance indicators. I have seen suggestions that training beside the ocean helps runners improve their performance while training but there were no scientific papers or evidence of any kind cited so I would like to test the validity of this statement.

Chapter 9: **Summary and Conclusions**

In this paper we found features that correlate to an improvement in race time. Performing long runs in training and the number of sessions per week appeared to have a particularly strong correlation to improved race performance. It cannot be stated for certain whether these features were the actual causation of the improved race performance or not but the relationship was apparent.

The prediction models performed better than I anticipated before starting the project. The gradient boost model received an R-squared score of 0.829 for the model that grouped male and female runners. It would be interesting to find out if the male-only gradient boost model performed better than the female-only gradient boost model because there was a smaller dataset of female runners or if female runners just have a higher variance of running performance. Regardless, the aim of creating a running performance model that could accurately predict marathon times for the average person was successful.

There are some ethical concerns about future iterations of this paper relating to geographic data. Data that tracks the exact location of a runner may be difficult to find and would be ethically questionable, particularly if the runners did not know that their data was being used for a study like this, but that concern was not an issue in this study.

All code used for this paper can be found at this Gitlab repository [2].

https://csgitlab.ucd.ie/ThomasThornton/marathon-performance-real-world-data

Acknowledgements

I would like to thank Barry Smyth for all his help during this project. He was extremely helpful on all technical questions and provided excellent guidance throughout the year.

I would also like to thank Strava Inc [1] for supplying the data for this project.

Appendix

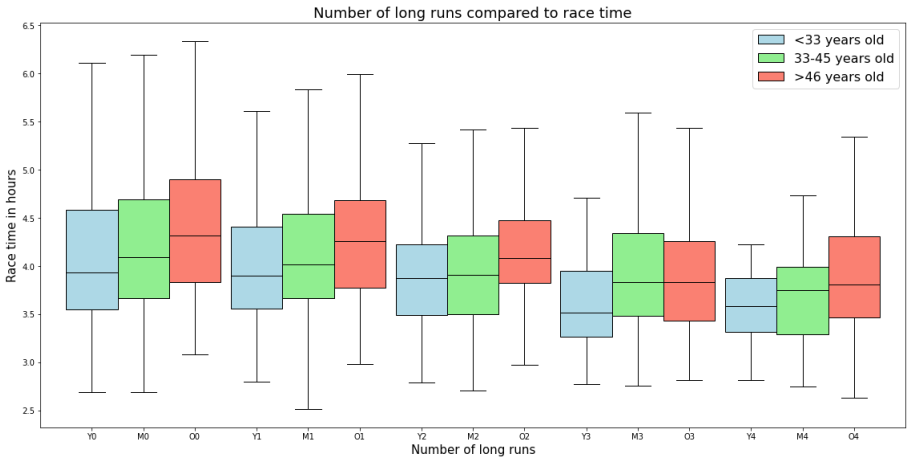

Figure A.1: The number of long runs completed by runners in each age range compared to their race time.

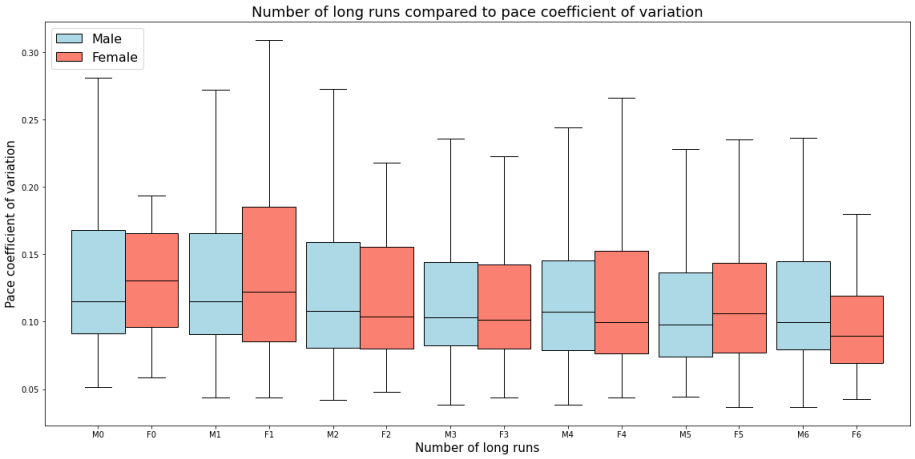

Figure A.2: The number of 20km long runs completed by male and female runners to coefficient of variation of pace.

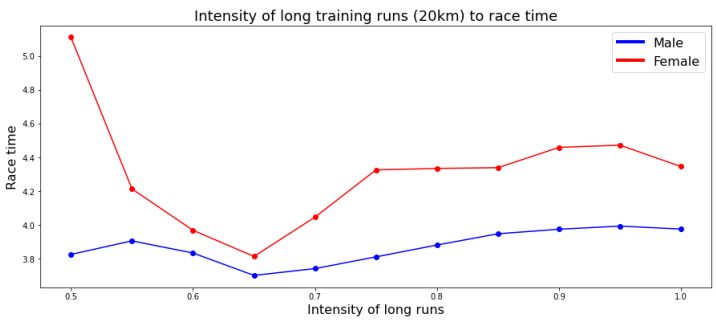

Figure A.3: Intensity of long runs to race time.

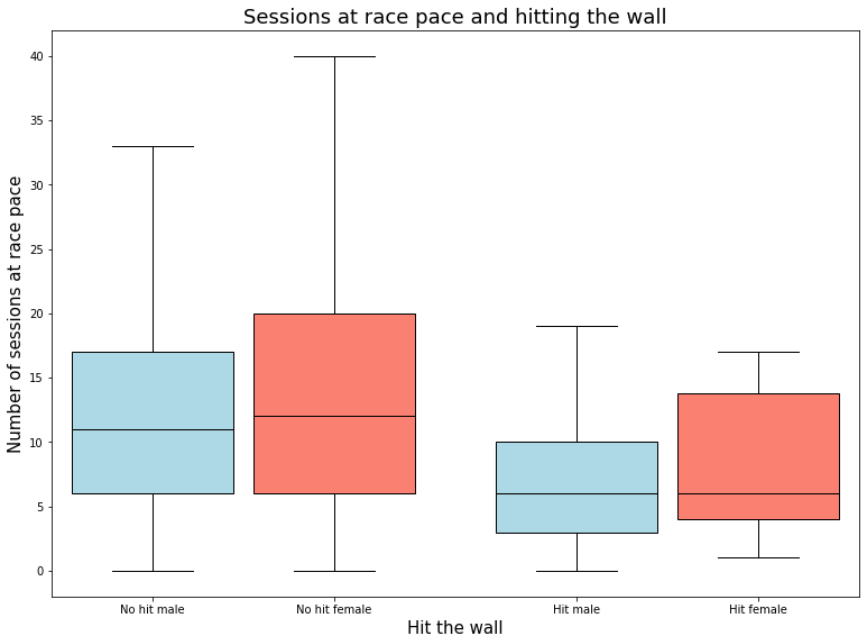

Figure A.4: Number of sessions training at race pace and hitting the wall.

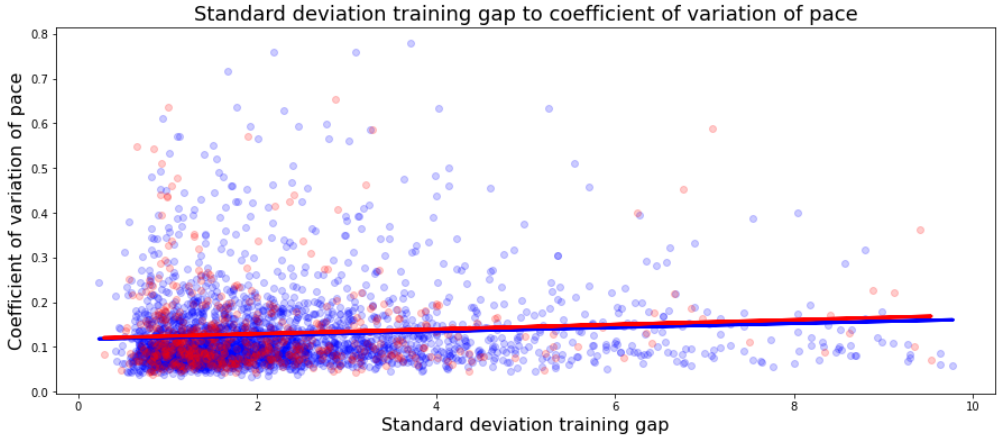

Figure A.5: Standard deviation of days between training to coefficient of variation of pace.

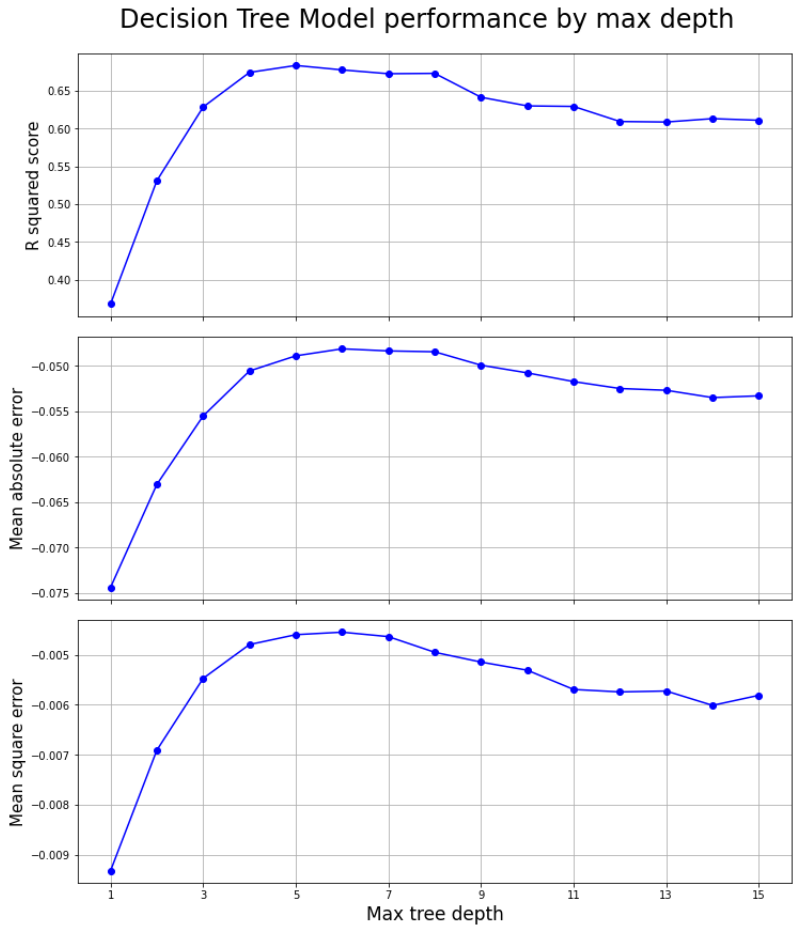

Figure A.6: Best performing level for decision tree.

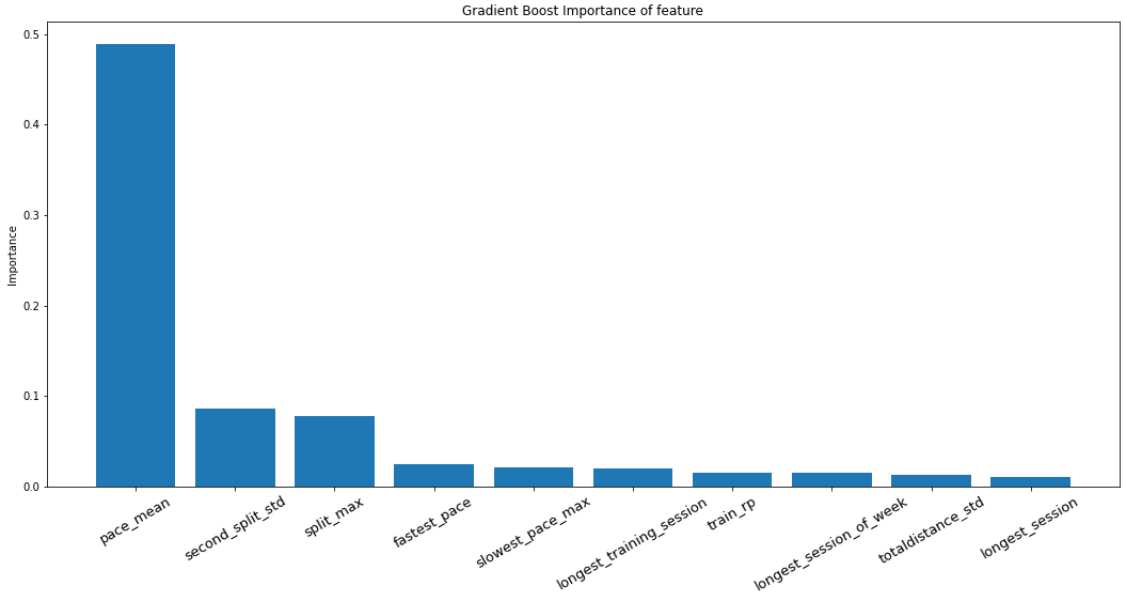

Figure A.7: Most important features gradient boost.

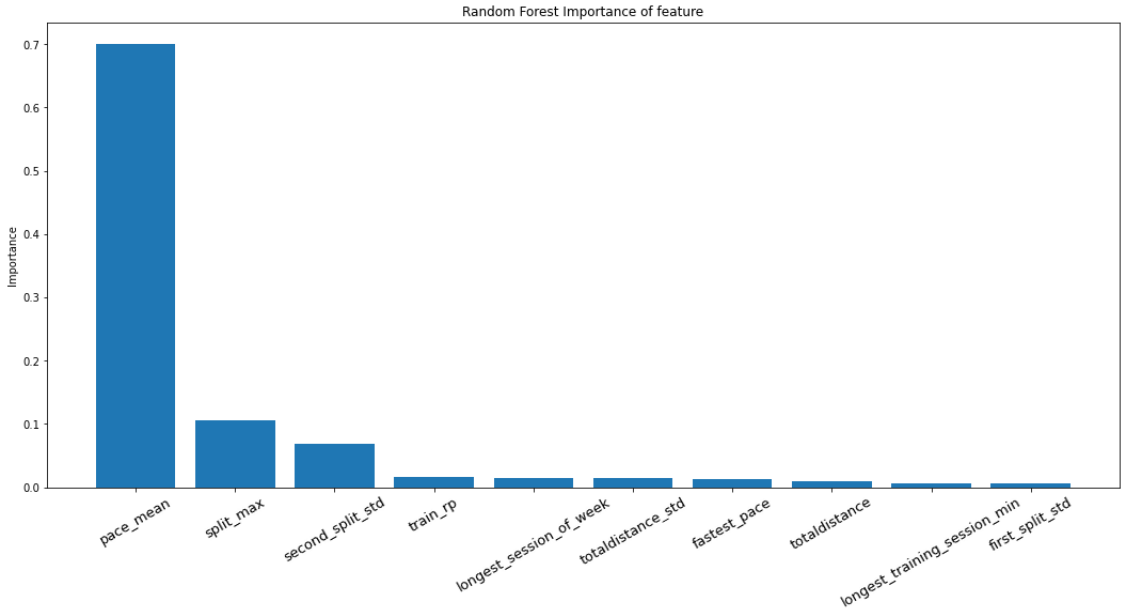

Figure A.8: Most important features random forest.

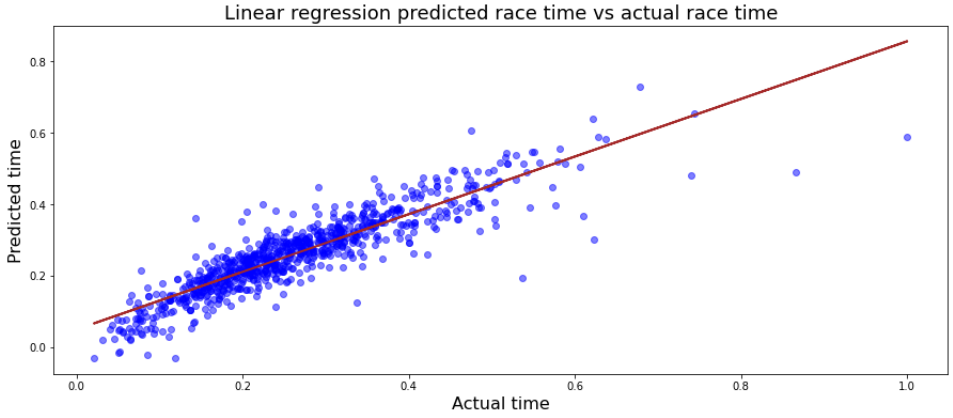

Figure A.9: Linear regression actual time vs predicted time.

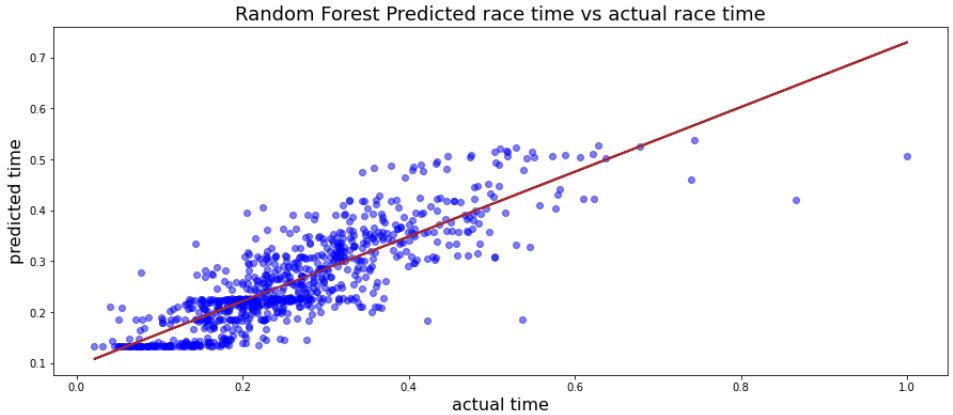

Figure A.10: Random forest max depth 5 predicted time vs actual time.

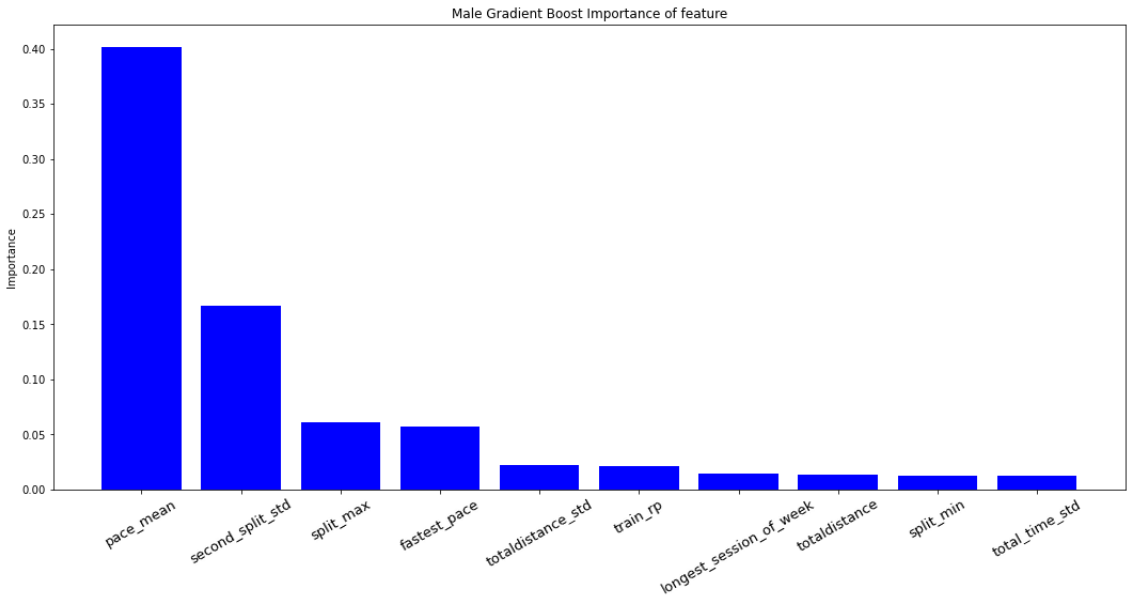

Figure A.11: Most important features gradient boost male runners.

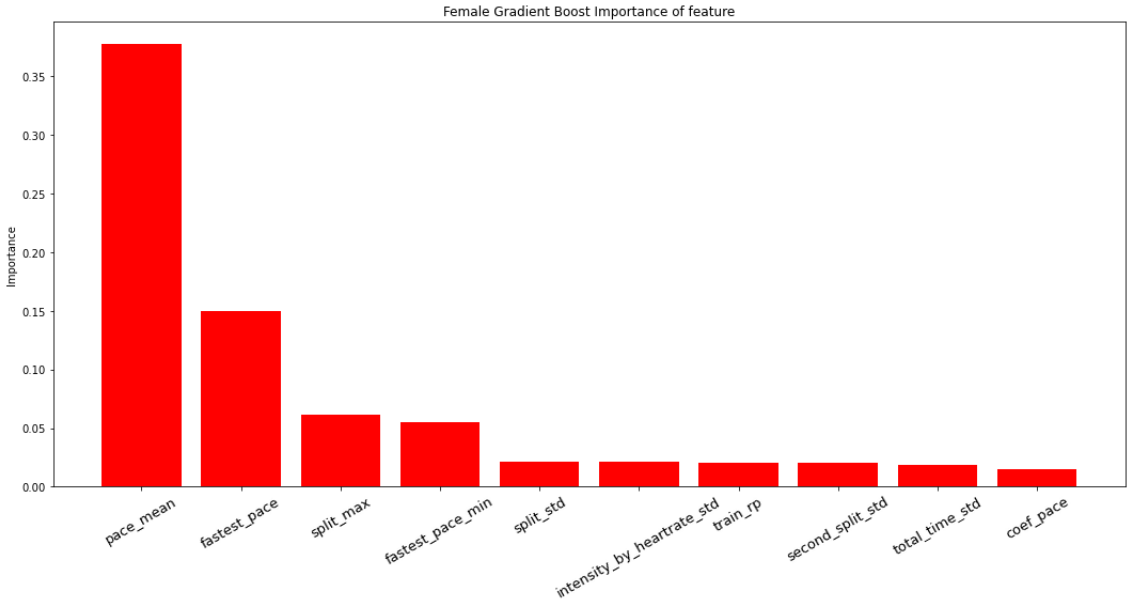

Figure A.12: Most important features gradient boost female runners.

Bibliography

1. *Strava website* https://www.strava.com/.
2. *Gitlab repository for this project* https://csgitlab.ucd.ie/ThomasThornton/marathon-performance-real-world-data.
3. Vailshery, L. S. *Fitbit - statistics & facts* https://www.statista.com/topics/2595/fitbit/.
4. Emig, T. & Peltonen, J. Human running performance from real-world big data. *Nature Communications* **11.** https://www.nature.com/articles/s41467-020-18737-6 (2020).
5. Mulligan, M., Adam, G. & Emig, T. A minimal power model for human running performance. *PloS one* **13,** e0206645 (2018).
6. Pemmaraju, V. & Hoch, D. *I Think We'll Go to Boston - Marathon Performance Prediction* https://www.sloansportsconference.com/research-papers/i-think-well-go-to-boston-marathon-performance-prediction.
7. Vickers, A. J. & Vertosick, E. A. An empirical study of race times in recreational endurance runners. *BMC Sports Science, Medicine and Rehabilitation* **8,** 1–9 (2016).
8. Smyth, B. Fast starters and slow finishers: a large-scale data analysis of pacing at the beginning and end of the marathon for recreational runners. *Journal of Sports Analytics* **4,** 229–242 (2018).
9. Smyth, B. How recreational marathon runners hit the wall: A large-scale data analysis of late-race pacing collapse in the marathon. *PloS one* **16,** e0251513 (2021).
10. Esteve-Lanao, J., Foster, C., Seiler, S. & Lucia, A. Impact of training intensity distribution on performance in endurance athletes. *The Journal of Strength & Conditioning Research* **21,** 943–949 (2007).
11. *RW's Race Time Predictor* https://www.runnersworld.com/uk/training/a761681/rws-race-time-predictor/.
12. *Apache Spark™ - Unified Analytics Engine* https://spark.apache.org/.
13. *XGBoost Documentation - Read the Docs* https://xgboost.readthedocs.io/en/stable/.
14. *10 Marathon Training Tips* https://web.archive.org/web/20120307074320/http://www.baa.org/programs/training-programs/marathon-training.aspx.
15. *Marathon Training : Intermediate 2* https://www.halhigdon.com/training-programs/marathon-training/intermediate-2-marathon/.
16. *HOW TO CALCULATE YOUR MAXIMUM HEART RATE* https://www.polar.com/blog/calculate-maximum-heart-rate-running/.
17. *How To Measure Exercise Intensity* https://wellnessed.com/exercise-intensity/.
18. *Common Problems with GPS* https://hellotracks.com/en/blog/How-to-Improve-your-GPS-Accuracy/.
19. *Usain Bolt 100m 10 meter Splits and Speed Endurance* https://speedendurance.com/2008/08/22/usain-bolt-100m-10-meter-splits-and-speed-endurance/.
20. *What Is the Average Walking Speed of an Adult?* https://www.healthline.com/health/exercise-fitness/average-walking-speed.
21. *Scipy Library* https://scipy.github.io/devdocs/index.html.

22. *Why Marathon Pace Is So Important* https://marathonhandbook.com/marathon-pacing-important/.
23. Schulz, R. & Curnow, C. Peak performance and age among superathletes: track and field, swimming, baseball, tennis, and golf. *Journal of Gerontology* **43,** P113–P120 (1988).
24. *Model Evaluation in Scikit-learn* https://towardsdatascience.com/model-evaluation-in-scikit-learn-abce32ee4a99.
25. *sklearn kfold Library* https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html.
26. *sklearn linear regression* https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.
27. *sklearn decision tree regression* https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html.
28. *random forest regression* https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.