Predicting Marathon Finish Times

Capstone Project

BSTN September 2021 Cohort

Beth McGregor

2021-12-06

Introduction

Marathons, especially larger races such as the World Marathon Majors, are prestigious races both for elite and recreational runners alike. These races attract substantial sponsorship dollars and bring tourism revenue to the host cities. For example, in 2019, the Chicago Marathon generated $378 million dollars for the Chicago economy (Bank of America Chicago Marathon, 2019). Gaining a better understanding of factors impacting the race finish times, could potentially help races to attract both sponsors and runners. I would like to try this model with other marathon race courses and potentially be adapted for other distances.

The goal of the project was to identify factors that influence marathon race finish times and build a predictive model.

Data Acquisition

The data was primarily sourced from the historical marathon results of the London (TCS London Marathon, 2021), Chicago (Bank of America Chicago Marathon, 2021) and Berlin Marathons (BMW Berlin Marathon, 2021) for the years of 2014-2019 via web scraping. To this, weather data (temperature and precipitation; Chicago - US National Weather Service, 2021, Berlin and London - Klein Tank et al., 2002), and course data (Maffetone et al., 2017) was added. Unfortunately, not all information was available consistently between the marathons. A combined dataset without the age, bib number, and half split times was created so that the three races could be explored and analyzed together. However, in order to investigate the predictive contribution of as many features as possible, a subset of the data, the London and Chicago Marathon results, were used for modelling independently.

Web scraping helped to provide some initial control with quality issues, but some processing was required prior to starting analysis, including splitting marathon result details stored together (name and nationality), removing results with missing or erroneous values, and converting finish results to seconds. Additionally, a few non-numerical features (nationality and age class) were converted to numeric features for inclusion in modelling.

Results and Discussion

The exploratory data analysis provided some interesting insights about the Chicago, Berlin and London marathons. Firstly, the mean finish time (seconds) between the marathons are significantly different (Berlin - 15034.3 vs. London - 16311.8 vs. Chicago - 16536.3, p-value = 0.0). The Berlin course is known to be fast - in 2018, elite marathon runner Eliud Kipchoge broke the marathon world record in this race (with a finish time of 2:01:39). However, I had not anticipated that the differences between the mean finish times would be this large. There are differences in how the marathons accept runners (guaranteed entries with qualifying times and non-guaranteed entries via lottery entries) which could partially explain this effect.

It is possible that temperature (primarily minimum temperature) may have an effect on some finish times. This effect was observed for the London Marathon in 2018, in which the mean finish time was significantly higher compared to all other years. The minimum temperature recorded for race day in 2018 was 4-5 degrees warmer compared to the other years. High temperatures can negatively impact finish times by increasing the perspiration (and consequently dehydration) and perceived effort of runners. Minimum temperature was identified as an important feature in the final London Marathon model that will be discussed below.

Modelling was attempted with a few variations of the data: the combined results of all 3 marathons, the London Marathon only (2014-2019 results with half split, age classes and bib numbers added), and the Chicago Marathon only (2014-2019 results with age classes and bib numbers added). Models explored included linear regression (with and without ridge and lasso regularization, and scaled principal component analysis) and XGBoosting (with linear and tree based learners). Overall, an XGBoost model with a tree based learner resulted in a slightly improved performance over the linear regression models (a mean absolute error of 614 vs. 648 seconds; Table 1). The London Marathon results that included half split time resulted in the greatest improvement in model performance. Removing outlying results also provided a modest further improvement in error.

Table 1 - Overview of Models Attempted on the Marathon Results Data.

Model	Dataset	Mean Absolute Error (seconds)	Top Features
Linear Regression	Combined	74406.61	gender, berlin, usa
Linear Regression	London	647.76	Half_split, bib_number, age_class
XGBoost (outliers removed)	London	613.92	half_split, bib_number, age_class
XGBoost (outliers removed)	Chicago	1684.4	bib_number, year, gender

The final model was able to predict finish times within approximately 10 minutes overall, but did show improved performance when looking at elite athletes vs. non elite athletes (158 vs. 611 seconds mean absolute error).

Conclusion

Overall, the best iteration of the model was able to predict finish times with an error of approximately 10 minutes. This is not particularly helpful for an elite or semi-competitive runner, but may be helpful to a new or casual marathon runner to predict their finish time. The data analysis comparing the three marathons provided interesting insight into differences in mean finish times and the potential impact that temperature may play. This is especially interesting to explore in the context of climate change and how this may impact future marathon planning. As next steps, I would like to secure the features missing from the Chicago and Berlin marathons, and look at smaller subsets of the data in order to try and improve the model.

References:

Bank of America Chicago Marathon (2021) Official Race Results 1996-2019. Available from: https://chicago-history.r.mikatiming.com/2019/

“Bank of America Chicago Marathon Generates Record-Breaking $378 Million for Chicago Economy in 2018.” Bank of America Chicago Marathon, 2 Oct. 2019., Available from: https://assets-chicagomarathon-com.s3.amazonaws.com/wp-content/uploads/2019/10/093019_WAGNER_2018-BACCM-Economic-Impact-newsroom_FINAL.pdf. Press release, PDF download.

BMW Berlin Marathon (2021) Results List. Available from: https://www.bmw-berlin-marathon.com/en/impressions/statistics-and-history/results-archive/

Klein Tank AMG and Coauthors (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453. Available from: http://www.ecad.eu

Maffetone PB, Malcata R, Rivera I, Laursen PB (2017) The Boston Marathon versus the World Marathon Majors. PLOS ONE 12(9): e0184024. https://doi.org/10.1371/journal.pone.0184024

TCS London Marathon (2021) Race Results - Past London Marathon results. Available from: https://www.tcslondonmarathon.com/results/race-results

US National Weather Service, 2021. NOWData - NOAA Online Weather Data. Available from: https://www.weather.gov/wrh/Climate?wfo=lot