Quantifying Uncertainty in Marathon Finish Time Predictions

Brandon Onyejekwe, Eric Gerber

Khoury College of Computer Sciences, Northeastern University

Introduction

In the middle of a marathon, expected finish times are traditionally estimated by naively extrapolating the average pace covered so far, assuming it will be held constant for the rest of the race. These predictions have two issues: (1) the estimates do not consider in-race context that can determine if a runner is likely to finish much slower or faster than expected, and (2) the prediction is a single point estimate with no information about uncertainty. A Bayesian inference model addresses both concerns by using the runner's previous splits in the race to generate a probability distribution of possible finish times, using conditional probability and empirical likelihood estimates. In this project, a Bayesian model is evaluated in comparison to the traditional estimate method.

Methods

Data: We scraped the data for this project from the Boston Athletic Association website. It contains a runner's name, age, gender, the intermediate splits of their race (5K, 10K, 15K, 20K, HALF, 25K, 30K, 35K, 40K) as well as their finish time. The splits and finish times are recorded in seconds. The data includes every finishing runner from each Boston Marathon from 2009-2023, resulting in 312805 rows of data. The data was partitioned into two groups: a training set (286777 runners from 2009-2022) and a test set (26028 runners from 2023).

Model: The model utilizes Bayes theorem to iteratively update the posterior finish time distribution: Using a runner's splits so far (s_1, …, s_t), the model predicts the finish split (s_f) according to the following equation:

$$P(s_f|s_{1:t}) \propto P(s_t|s_f, s_{1:t-1}) * P(s_f|s_{1:t-1})$$

In this equation, the posterior distribution is the normalized product of the likelihood distribution and the prior distribution.

Results

We found that the model predictions have similar absolute errors to the default predictor, and outperform the default predictor as more splits are used.

From these tables, we are able to plot the posterior distributions at each iteration on a single plot, an example of which is shown below.

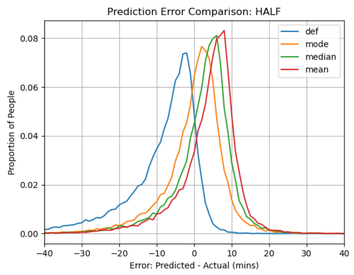

Fig.1 - The posterior modes of the model's predictions have comparable errors to the naive default predictor. Posterior medians and means are also included.

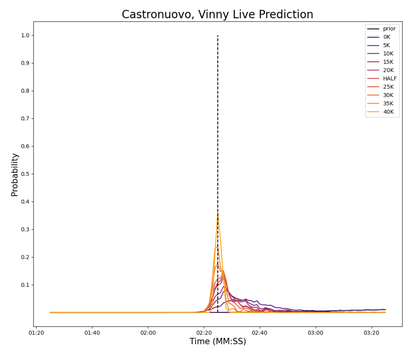

Fig.2 - Finish time estimate plot containing all splits for Vinny Castronuovo (fastest male Bostonain, former Northeastern Club Running President) The actual finish time is denoted with the vertical dotted black line.

Conclusion

Scan the QR code to access the website for this project. It contains functionality to automatically generate a finish time distribution plot given your own splits, and also allows you to view the plots of select past runners (namely, Northeastern Club Running runners that ran in 2023). This app can be used to get a better sense of uncertainty than the traditional estimates given by the BAA during the marathon.

References