ECE 6254 Spring 2024: Data Challenge

In the “data challenge” for this course I have prepared a small dataset. This dataset consists of my times in various races over the last few years, together with a set of (possibly) relevant features. Your task will be to learn a model that, when given this set of features, predicts the associated finish time in the race. You will use this to predict my race time in this year’s Pi Mile Road Race.

Dataset Details

The dataset can be downloaded from the course website as a csv file. The desired output of your model is in the “Finish Time” column. This indicates the number of minutes I took to complete each race. Possible features that you might want to consider in training your model include:

- The date and time of the race.
- The race distance (this feels like an important one to include!)
- The elevation gain and loss of the course (this was estimated by me either using my watch or looking up online estimates if I had doubts about the accuracy of my watch).
- Various facts about the weather at the start of the race.
- The number of miles ran in the week leading up to the race, as well as the average number of miles per week ran in the preceding 4, 13, and 52 weeks.
- The number of hours ran in the week leading up to the race, as well as the average number of hours per week ran in the preceding 4, 13, and 52 weeks.
- The number of “easy runs” completed in the last 1, 4, 13, and 52 weeks. An “easy run” is anything over 3 miles where I’m not targeting a specific speed.
- The number of “interval workouts” completed in the last 1, 4, 13, and 52 weeks. An “interval workout” consists of intervals at my 5K pace or higher (typically lasting from 1-6 minutes, depending on the speed) with short breaks between intervals.
- The number of “tempo workouts” completed in the last 1, 4, 13, and 52 weeks. A “tempo workout” is a run which typically includes a segment of 20-40 minutes ran at a speed near my 10K or half-marathon pace.
- The number of “long runs” completed in the last 1, 4, 13, and 52 weeks. I defined a “long run” as any run the length of a half-marathon or longer (13.1 miles).
- The longest run completed in the last 1, 4, 13, and 52 weeks.

I do not claim that all of the above features are necessarily the best features. Indeed, this probably leaves out many of the most useful features, but I don't necessarily have all the data that might be most predictive. If you would like to request any additional features (that you cannot simply compute yourself), let me know. I will consider generating additional features if they are easy to compute based on the data that I have available, and would release these to the entire class.

Deliverables

On April 18 I will release an additional feature vector. By **7:30am, April 20** you must submit your prediction via **Gradescope**. This will need to be in the form of the number of minutes (decimal form) for your prediction, as well as a "confidence interval" with lower and upper bounds defining a range where you are reasonably confident that my finish time will land within. You will receive a bonus on the assignment as follows:

- If your prediction ranks among the top 5% in the class, you will receive a 3 point bonus.
- If my race time lands within your confidence interval, you will receive a bonus of 30/(length of your interval in seconds) points.

The full assignment will not be due until **11:59pm, April 23**, and will need to include answers to the following additional questions. The assignment will be graded out of 10 (although your score can be higher factoring in bonus points). Each of the following questions will be answered using only a short paragraph.

1. Describe your overall method for making your prediction. What algorithm/algorithms did you use, and why? If you explored multiple algorithms, describe your methodology for selecting the one you chose. Include details such as the form of any loss or regularization functions.
2. Describe which features were used and how they were selected/computed.
3. Describe any pre-processing performed such as normalization or outlier rejection.
4. Comment on the sensitivity of your algorithm to the features used. If possible, comment on the following question: Based on your algorithm, if there is one feature over which I have control (details of my workouts), what would have the biggest impact on my race time? If answering this is not possible using your approach, explain why not.
5. Describe your procedure for determining your confidence interval and comment on how accurate your prediction was. Given the outcome, how would you change your approach if you had to make another prediction in the future?