Data Report

Will Cho
Computer Science Topics II
Hamilton College
Clinton, New York
wcho@hamilton.edu

I. Data Source

For my final project, I primarily sourced data from Strava, a platform that allows runners to post or upload their runs from other devices. In cases where Strava was not available, I utilized Garmin Connect, a platform directly affiliated with Garmin. Both platforms automatically upload data from watches or allow for manual entry. I chose these platforms because they track the metrics I consider essential for each run, such as pace, distance, heart rate, and time, and they allow for easy data export.

II. Data Collection

To collect the data, I employed two methods. First, I reached out to people I knew had previously run marathons, including friends, family members, and professors. This approach yielded only a few data points, so I needed an alternative strategy to collect more data.

Next, I created a Google form that allowed runners to upload their training data and provide their marathon times to serve as labels. To incentivize participation, I offered a raffle prize—specifically a pair of running shoes. Each submitted marathon build-up was counted as one entry in the raffle. Initially, I planned to post this form on Reddit running threads, but most threads prohibited self-promotion. Consequently, I only received one response from Reddit.

I then attempted to use other running forums and websites but was similarly unsuccessful. Finally, I turned to Strava's club feature, which allows users to post in various marathon clubs. By joining marathon clubs worldwide, I was able to broadcast my form and collect more responses. I am currently waiting for additional responses to complete my dataset.

III. Preprocessing

Since I do not have much data yet, I have not begun extensive preprocessing. However, my plan for preprocessing is as follows:

I will first separate each CSV file into its respective marathon build-ups. For example, if a runner completed two marathons—one in 2022 and another in 2024—I will split the data so that the first build-up includes all runs leading up to the 2022 marathon and the second includes runs leading up to the 2024 marathon. Any runs recorded after the 2024 marathon will be discarded.

Next, I will discard unnecessary features, such as the activity title, which has no direct impact on the run itself. I will also remove non-running activities, such as cycling, swimming, or weightlifting sessions, to ensure the dataset is focused solely on running performance.

In addition to these steps, I will also handle missing data carefully to ensure consistency across the dataset. For instance, if any runs are missing key metrics like heart rate or pace, I will impute these values using reasonable estimates, such as the average from nearby runs in the same build-up. If critical data for a run is missing entirely, I may choose to discard that specific run to avoid introducing noise into the model.

Another important aspect of preprocessing will be normalizing the numeric features, such as pace, distance, and heart rate, using Min-Max scaling. This will ensure that all features are on a similar scale and prevent larger numerical ranges from dominating smaller ones. Once the numeric data is scaled, I will address any categorical features, such as the type of run (e.g., long run, interval training), by using one-hot encoding. This will allow me to represent categorical variables as binary features, making them compatible my neural network.

Feature engineering will also play a key role in improving the dataset. I plan to create additional metrics, such as total weekly mileage, changes in average pace over time, and heart rate variability, which can offer more insights into a runner's performance. By breaking the training data into smaller windows, I can track how training progresses across different phases of the marathon build-up.