Edward

DSC412 Milestone 2 - Project Proposal and Plan

# Proposal

**Problem**

Many recreational runners, including myself, may want to step up to the next level and sign up for a race to challenge themselves. Race participants often set goals for themselves, such as a sub-two-hour half marathon or a sub-four-hour full marathon. Such a feat requires strategic pacing, efficient running and breathing techniques, and consistent training. Thankfully, training apps such as Strava are available to record training data and turn them into actionable insights to help runners achieve their goals. However, a premium subscription is required to access many of these features. As a result, there is a need for a tool to analyze running data and provide realistic race time predictions.

**Proposed solution**

I propose using ML algorithms to analyze historical running data to develop a predictive model for race finish times. The model will analyze data from Strava (distance, pace, heart rate, elevation, etc) and utilize regression techniques (linear regression or time series) to estimate finish times based on the user's training patterns. In addition, a classification algorithm (decision tree) can be used to separate actual training data from recovery runs (slow cooldown runs after training) or commute runs to improve prediction accuracy.

**Potential Stakeholders**

Potential stakeholders include runners and trainers seeking predictions for their marathon times to set realistic goals and fitness apps to enhance the user experience by providing this extra feature.

**Potential Obstacles**

Potential obstacles include:

- Inconsistent data entries or missing values (heart rate may be unavailable if a monitoring device is not worn) can affect model performance and accuracy.
- I am currently injured and cannot produce more data. Insufficient data may lead to poor generalization and inaccuracy.
- Many more factors that can significantly affect running (such as weather and temperature) may not be available.

**Novelty**

This project is novel in its simple and economical approach to marathon time prediction using individual training data from Strava. While many tools and resources are available online, a lot of them require some form of subscription. In comparison, my model provides simple estimates for runners like me who do not want to spend money and need all the professional tools.

# Plan

### Data Source

I plan to extract my own running data from Strava and Garmin apps, which includes:

- Distance
- Pace
- Elevation
- Heart rate
- Time
- Temperature (if available)
- Speed
- Cadence

And more if required later on.

### Data Creation

I will be creating my own data through my Garmin watch. My running data will be stored in the Garmin Connect app and transferred to Strava. The classification model will be completed first as a part of data cleaning (for clearing out non-training runs before creating the regression model).

### Data Organization

The Garmin Connect and Strava app will preorganize my running data. I can use the Strava API directly or transcribe data into my own dataset.

### Data Analysis

I plan on applying correlation analysis to determine which factors influence marathon time the most, data analysis to visualize trends, and finally K-Fold Cross Validation (splitting the dataset into training and testing sets) for model validation and accuracy.

**Model Selection**

I plan to use linear regression (or time series regression) as a baseline to predict finish times based on average pace, distance, and other factors. In addition, I will also implement a decision tree for the classification function. If possible, I will also apply any boosting algorithm to improve prediction accuracy.

**Model Accuracy**

I will evaluate the models' accuracy by using metrics such as R-squared and K-fold CV accuracy. R-squared tells us how well the regression line fits the data, and the CV accuracy will tell us how accurate the predictions were (for the decision tree classification part).

**Baseline Comparison**

There are a couple of Kaggle datasets about Strava running data that I can use for comparison for my data collection part. For my models, there are many time prediction and decision tree models online that I can sample and compare.