Milestone 2 Project Proposal
Caden Phan

https://www.kaggle.com/datasets/arnavsmayan/fitness-tracker-dataset/data

---

The Proposal

A problem (this can be related to your interested you completed in milestone 1)

- How do the number of calories burned by someone in a day impact the hours of sleep that they get? The aim is to understand the relationship between increased calorie expenditure and better sleep.

A proposed method of solution involving a machine learning algorithm

- Apply a multivariate linear regression model to predict daily step count based on calories burned and hours of sleep. This can show how these two variables influence the physical activeness of an individual.

Potential Stakeholder - who could benefit from using your model? Is it more than just yourself?

- Fitness tracker app developers, health conscious individuals, and researchers in fitness could all benefits from this type of data analysis

Potential Obstacles - what are some obstacles you think you might encounter, these can be technical problems (ei I may not find the data set that describes my problems statement), or general problems.

- Since I am basing this off of one other person and my own data sets, the sample size is pretty small. I may have to look for other larger samples of data, as two people might have a lot of other complex variables going on to look at what is going on in this context.

What is novel about your problem and your potential approach?

- While many fitness models focus on the number of steps or calories burned independently, this project takes a novel approach by focusing on the combined effect of sleep quality and calories burned on predicting daily activeness and step count. This could provide a more holistic view of how these two lifestyle factors together influence activity levels, offering a more tailored approach for improving fitness.

The Plan

Where do you plan to get your data sets?

- I have found a few good datasets relating to the problem I have planned on Kaggle. Additionally, because my problem is related to fitness, I want to use some of my own data collected on my Apple Watch.

If you plan to create your own dataset, please include a detailed plan of how you are going to collect, where you are going to collect it and where you plan to store the data. - This data will be accessible publicly. It is necessary that you have the correct permissions/rights to use the data you collect.

- The datasets I have found are already on Kaggle and thus on public domain, additionally my own data is my own to publish.

Take some time to sift through your datasets, how do you plan to organize your data?
How do you plan to analyze your data? What are you looking for? What might you find?

- Though the data sets have various other categories of information on them, I only plan on looking at calories burnt and hours slept. I am looking for the correlation between those two variables. I am thinking that I will find that burning more calories leads to better sleep at night.

Based on your problem statement, what type of model do you plan to use? Is it one of the ones we have discussed/will discuss in class?

- Due to the fact that I am focusing on two variables, I feel that linear regression may be the best type of model to focus on.

If we have not done a model overview in class yet it may be necessary to do a quick google of what models we will talk about in class and how they may be used to solve your problem statement. - You may also use models not discussed in class.

How will you determine if your model is accurate?

- I will likely compare it with science already published on the internet and then double check with the raw data sets to see if my model is accurate.

Are there similar models/paper/code online that are trying to do what you are doing? If there are, use them as a baseline for improvement

- At this point, I haven't scoured the internet for what I am planning on doing, but due to the relatively simplistic nature of my model, I feel as if it won't be too hard to find and replicate.