Final Project Proposal

For the final project, we propose analyzing Olympic marathon data from all modern Olympic games after 1904, when the marathon was standardized to 26.2 miles. We plan on using either the top 100 runners from each marathon, or just the top 1000 overall.

We will try to answer several research questions. First, can we accurately predict the outcome of the 2021 Olympics using the data from the previous Olympics? Second, what relationship does the GDP of a country have with the results of its runners? Finally, what variables are the most important in predicting Olympic marathon results?

We are going to get data on the runners for the Olympics until 2016 from a Kaggle dataset. We are going to get the GDP data from publicly available tables on google. We are going to scrape the 2021 Olympics data from the same places as the original Kaggle dataset.

There will be a fair amount of data wrangling required. We will need to join several tables (both from Kaggle and GDP). Because we will be scraping uncleaned data, we might need to clean the 2021 data. We are planning on adding features, with the most likely additions being the number of previous Olympic marathon appearances as well as the time from the last Olympic marathon appearance. We will also need to one-hot encode many of the features (like country of origin).

For EDA, we plan on plotting inter-variable correlation, finding and plotting interesting relationships between variables, and several other things to be determined upon having the data.

We are going to use several machine learning models. Because the observations will be correlated both across time and within marathons, the models we use will need to have minimal assumptions about independence. We will therefore be using a random forest model, a gradient boosting tree ensemble, and an extreme gradient boosting model since those have some level of interpretability with minimal assumptions. We will use partial effect plots to understand the marginal relationships of variables with the marathon outcome.

We anticipate that the model will be fairly accurate in its predictions. We also anticipate that GDP will have a bathtub shaped effect on marathon results (weighted on one end by Kenya and Ethiopia and weighted on the other end by the United Kingdom and the United States). We don’t have any hypothesis about the relative importance of variables.

We plan on having Jace scrape the missing data and perform analysis on the results. Gavin will perform EDA and wrangle the data. Colin will cross validate the models and select hyperparameters, then perform the final predictions. We will use Jace’s GitHub account to host the project.