Braden Wellman

Jessica Hamblin

Spencer Kimball

Stat 426 Group Project

1. What research question are you trying to answer?

Can we predict an athlete’s rank based on their profile (which consists of their basic demographics, performance in exercises, and diet).

1. What data are you planning on using and where are you going to get the data?

We are planning on using a Crossfit data set from data world.

<https://data.world/bgadoci/crossfit-data>

1. How much data wrangling or cleaning will be required? Are you planning on making new features?

We are going to try and extract data (converted to factors) from the “eat” column since it was entered by the athletes manually. The “eat” column will need to be split based on the “|” which separates the different information that is found in that column. The training column also will need features created from the string that includes where they workout, which program they follow, and whether or not they record their workouts.

We will also need to merge the two datasets based on the athlete id column. In many of the columns there are “na” values as well as spots where it just says “No data”.

1. What types of EDA are you planning on doing?

We will be doing the Multivariate graphical type of EDA. We will be doing pairwise plots to show comparison between the different explanatory variables. We would want to compare rank to score to try and decide which of the two we would want to use as the response variable.

There are four different years within the two datasets with the athletes, and we would be able to split the data based on the years, and create a model for each year.

1. What kind of machine learning model will you most likely use?

We think we would want to try do decision trees but combine it with bagging to try and avoid overfitting. We would also want to use knn because in our previous group project it seemed to be a fairly reliable model.

1. What are your anticipated results? Do you have a hypothesis?

We hypothesize that if someone is on a good team, have been lifting for a longer period of time (“howlong”), and have a good eating regiment then they would have a better score.

1. Group Member’s Outline for Project:

Spencer:

Separate the training column into its respective features.

Knn machine learning model.

Braden:

Fix the na and “No Data” values in the columns

Add the bagging model to the decision tree learning model.

Jessica:

Separate the eat column into the distinctive columns based on responses.

Decision Tree learning model.

1. Who’s github account are you planning on using to host this project?

Jessica created the github repository on her account.