Executive Summary: Problematic Internet Usage

Team Members: Dan Visscher, Aaron Weinberg, Emilie Wiesner

Note: The organization of the GitHub repo is:

* CSV files:
  + train\_original is the data file downloaded from Kaggle
  + train and test are an 80/20% split of train\_original
  + train\_cleaned is train after being processed by Data\_Cleaning
  + train\_cleaned\_outcome\_imputed is train\_cleaned after being processed by Outcome\_Imputing
  + train\_cleaned\_outcom\_imputed\_feature\_selected is the result of processing by Feature\_Reduction
  + Accelerometer\_enmo\_anglez\_daily\_averages is the predictors generated from the actigraphy data
* Notebooks:
  + Accelerometer\_Computations computes predictors from the actigraphy data
  + Data\_Cleaning cleans the data
  + Feature\_Importance identifies a list of “key features” to use in multiple linear regression
  + Feature\_Reduction removes highly-correlated and problematic predictors
  + Modeling tests out-of-the-box performance for a collection of models, tunes, and re-tests the models
  + Outcome\_Imputing imputes missing values of PCIAT scores
* Other files:
  + CustomImputers includes classes for doing our iterative imputing and computation of Zone predictors
  + OrdinalClassifier includes a class that “wraps” a classifier in an algorithm that performs ordinal classification based on a method proposed by Frank and Hal (2001)
  + series\_test and series\_train are folders that contain the actigraphy data

The research question investigated by our team is: “Can you predict the level of problematic internet usage exhibited by children and adolescents, based on their physical activity?”

Researchers have identified internet use as having the potential to rise to the level of addiction, and problematic internet is associated with increased rates of anxiety and depression. For these reasons, being able to identify problematic internet use is a worthwhile goal. However, under current circumstances, identifying problematic internet usage requires evaluation by an expert, for example, a psychologist. This significantly limits the ability to evaluate all children and adolescents. If it could be effectively predicted based on physical activity, prediction could potentially be done by, say, a family physician.

This project came to us by way of a current competition hosted by Kaggle (a data science competition platform) competition. This particular competition was sponsored by the Child-Mind Institute. The Child-Mind Institute is focused on mental health and learning disabilities among children and adolescents, supporting research, clinical work, and educational outreach.

The data for the competition comes from a research study conducted by the Child-Mind Institute. The target variable is the severity impairment index, or SII. This is an ordinal variable, values ranging from 0 (no impairment) to 3 (severe impairment). The SII score is derived from scores on a 20-question survey completed by participants and their families. We had access to individual survey question responses as well.

We were given a total of 33 predictor variable, spanning a variety of physical measurements. The variables naturally fall into several groupings. This includes demographics, hours of internet use, a clinical assessment of general functioning, physical measurements that one might take as part of a check-up, results from a fitness test similar to the Presidential Fitness Test that some of you may have completed in high school, measures from a bio-electric impedance analysis, a score from a physical activity questionnaire, and an evaluation of sleep disturbance. We also had accelerometer data: some of the participants wore an accelerometer wristwatch for roughly one month; and we had accelerometer data on 5-second intervals for those participants. Data was provided for 3960 participants.

The data included time series information from accelerometers worn by approximately 1000 participants over roughly one month. From this data, we developed one variable to summarize participant activity: we computed Euclidean Norm Minus One values,

grouped the data into 5-minute bouts to compute means, and identified whether each mean was above thresholds we identified from actigraphy research

There was considerable missing data. Almost 1000 participants did not complete any of the Parent-Child Internet Addiction Survey, so there was no information on the target variable; these cases were dropped. Other participants did not answer some of the survey questions, leading to potential inaccuracies in the SII score. Among the 3000 cases with SII scores, every case was missing data for at least one variable.

There were some other issues with the data itself. (SAY SOMETHING here about grip strength? The way they computed SII score?)

In our exploratory data analysis, we saw no strong relationships between any of the predictor variables and SII, suggesting that it may simply be difficult to predict problematic internet usage based on physical activity. Additionally, the distribution of participants across SII scores was very uneven. In particular, there were only 30 participants who were measured with an SII score of 3, or severe, in the entire data set; and a very high rate of 0’s.

Some of these issues we attempted to address prior to modeling and some as part of the modeling process. Here we wanted to share some of the interesting aspects of our modeling work.

The competition evaluates submissions using Cohen’s Kappa, which is a measure of accuracy for ordinal variables. We couldn’t directly incorporate this into our model training as the loss function, but we did use this to compare models.

Because there was so much missing data, we needed to use imputation, rather than just remove cases with missing values.

Recall that the PCIAT questionnaire is the basis for the SII score, our ultimate target variable. We eliminated all cases that had no PCIAT data, and then imputed any remaining missing data for individual PCIAT questions. All questions were scored on the same scale, so we used KNN imputation here.

For the predictor variables, we tested both KNN and a MICE imputation. Some of the variables, various fitness zones, were derived from the fitness test quantitative measurements. Rather than imputing the zone values, we computed these separately after running the imputation.