Vicki Moore
Springboard Project Proposal – Capstone 1

Title: Predicting heart disease risk

Problem: Heart disease is one of the greatest causes of death, though it is not necessarily clear which patient health factors contribute most to risk of heart disease.

Client interest: Understanding of the relative risks leading to heart disease would be useful to multiple stakeholders. The patient care community, as well as patients themselves, could benefit from awareness of specific risk factors. Insurers, whether private or publicly administered, would have an interest in knowing which factors make people vulnerable to heart disease and its outcomes, both to administer resources and to incentivize strategies to lead to optimal health. Heart disease is not simply a problem of industrialized nations, either, and NGOs could make use of knowledge of risk factors for treatment and prevention, including education, in areas where they are able to make an impact on public health. Publicly funded medical researchers and pharmaceutical companies could use information about the etiology of heart disease to develop treatments or preventive medications, since it is quite possible that heart disease is mediated not by lifestyle choices alone.

Data source(s) and how acquired: Data used for this machine learning project will come from the University of California, Irvine Machine Learning Repository. There are datasets regarding heart disease risk factors available from 4 distinct patient populations, including Cleveland, OH (N=282), Long Beach, CA (N=200), Switzerland (N=123), and Hungary (N=294). There are processed datasets available (examining 13 possible contributing factors, in addition to the target variable), but this project will explore the unprocessed datasets for any factors (among a total of 75 possible contributors) that may be missed in the processed datasets. Many of the factors from the unprocessed datasets will likely be disregarded after further analysis, because preliminary examination suggests that they comprise mostly missing data. However, hypertension, for instance, is included in the unprocessed datasets as a boolean factor, but it is not present in the processed datasets, and it will be examined in this study.

Approach to solving problem: The target variable to be predicted in our machine learning model is presence of heart disease, which is organized within the data as a classification system with 4 levels (including a category for absence of heart disease). We can choose a binomial, categorical classification system (heart disease diagnosis likely versus not) based on labeled data, and a supervised classification model would be a prudent approach for this. The model would include probability of categorization into either class. Alternatively, we can choose to categorical predictions into any of the 4 possible classes, using a multiclass classification approach with logistic regression, a neural network, or a decision forest.

While it is useful to have 4 separate populations from 3 countries included in this data analysis, a caveat is that all will have come from western, industrialized nations, possibly limiting the power of any model arising from this analysis to extrapolate into other regions of the world.

Deliverables: This project will include development of the following items: codes for importation of data, data cleanup, exploratory data analysis, and model development. A project report and a slide deck summarizing the project and its findings will also be provided.