Machine Learning Project Proposal

Matt Lombardo, John Michals, Kenneth Reed, Yamini Sasidhar, Khrystyne Vaughan, Fardi Yeung

Project Summary. For this machine learning project, we will create a series of robust models with the goal of predicting user likelihood of being diagnosed with different medical conditions based on patient blood profile, behavioral practices, individual attributes, and medical history. Current models of interest include but are not limited to: (1) logistic regression, (2) deep neural networks, and (3) random forests.

The dataset examined was requested from and provided by The National Heart, Lung and Blood Institute[1]. The architecture of the dataset is as follows:

- Began in 1948 with 5,209 subjects, all living in the community of Framingham, MA
- Focused on a subset of 4,434 subjects, complete with participant clinical data
- Data collected approximately 6 years apart, from 1956 to 1968
- Longitudinal form data given each subject has 1 to 3 observations depending on the number of exams the subject attended, resulting in 11,627 unique observations

Users will be able to enter their medical history and associated details into a web application to view their predicted odds for the following events: (1) angina pectoris, (2) myocardial infarction, (3) atherothrombotic infarction, and (4) cerebral hemorrhage (stroke).

Data Limitations. Current limitations of our dataset include: (1) missing HDLC/LDLC profiles prior to the subject's third observation, (2) general nulls in the data, and (3) inconsistency of the subject's attendance throughout the study.

We will be documenting our attempts at cleaning the dataset, training different models, and any visualizations in this GitHub repo. The tools we plan on using include: R, Pandas, Matplotlib, Scikit-learn, Flask, Heroku, HTML/CSS/Bootstrap, and Tableau.

[1] https://biolincc.nhlbi.nih.gov/