ORIE 4741 Project Proposal

Team Members: Krishna Raju, Karan Maheshwari, Adam Guo

Proposal Topic

Determine Factors Leading to Cardiovascular Disease and Predict Risk of Having CD

Problem

A news report came out in 2019 showing that chronic disease made up a big portion of an average household's expenditure. They are usually harder to diagnose, harder to prevent, and take a lot of effort and resources to treat. Among the all the chronic diseases, cardiovascular disease can be the mostly deadly. In the United States alone, cardiovascular disease is the leading cause of death, attributing to almost 1 million death annually, and the cost of treating cardiovascular disease is estimated to be around 400 billion dollars every year. Many research have been done on the causes that lead to this disease (blood pressure, weight, age, etc.), but not many have been done on patient data. We believe that preventing cardiovascular disease is an important enough field to investigate in. If possible we want to find the factors that could lead to cardiovascular disease the most, and use the model to predict risk of CD given new data.

Dataset

The dataset that we will be using for this project is the 'Cardiovascular Disease dataset' available on Kaggle. The data can be found at: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset/data. The dataset consists of patient information, with 3 types of input features: objective, examination, and subjective. The objective information are factual, the examination data are collected from medical examination, and the subjective information are given by the patients themselves. All of the dataset values are collected at the time of the medical examination, and overall there are 12 features, including gender, weight, height, and 70,000 data samples. The outcome variable in the dataset

Solution

Our goal for the research are two-fold. First, we want to come up with a machine learning model to determine which factors (features) in the dataset contribute the most to the patient having cardiovascular disease. Second, we want to be able to predict the risk of a new patient having cardiovascular disease given his/her data, and maybe come up with a confidence interval for the prediction. To evaluate our model performance, we want to look at the test error for the model, and also by comparing the important factors we got from the model to some clinical proven results.