Machine Learning Engineer Nanodegree final project proposal

Capstone Proposal

Heba Ghonaemy
March 5th, 2019

Proposal

Domain Background

Heart disease is the number one killer of men and women in the United States today. The Centers for Disease Control and Prevention (CDC) estimate that heart disease causes about 1 in 4 deaths in the United States each year. That's 610,000 people per year. About 735,000 people in the United States have a heart attack each year.

Problem Statement

Heart disease is considered one of the top preventable causes of death in the United States. It is believed some genetic factors can contribute, however the disease is largely attributed to poor lifestyle habits. *By discovering patterns in data like* (Blood Glucose Level / Blood Pressure / Cholesterol / etc.). We can detect the presence of heart disease. The goal of this project is early detection of cardiovascular disease which can be the difference between life and death.

Datasets and Inputs

The dataset used is the Heart disease UCI from kaggle. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

Details like the names and social security numbers of the patients were removed from the database. Only 14 attributes used:

- Age : age in years
- Sex : (1 = male; 0 = female)
- CP : chest pain type
- Trestbps : resting blood pressure (in mm Hg on admission to the hospital)
- Chol : serum cholesterol in mg/dl
- Fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- Restecg : resting electrocardiographic results
- Thalach : maximum heart rate achieved
- Exang : exercise induced angina (1 = yes; 0 = no)
- Oldpeak : ST depression induced by exercise relative to rest
- Slope : the slope of the peak exercise ST segment
- Ca : number of major vessels (0-3) colored by fluoroscopy
- Thal : 3 = normal; 6 = fixed defect; 7 = reversible defect
- Target : 1 or 0

Solution Statement

For this project I will be using supervised learning to detect the presence of heart disease. Will be comparing various classifiers results to determine the best model for this case.

Benchmark Model

Some of the tests used to determine cardiovascular disease are known to produce many false positive results. For example exercise stress test, despite its low sensitivity and specificity (67% and 72%, respectively), exercise testing has remained one of the most widely used noninvasive tests to determine the prognosis in patients with suspected or established coronary disease.

Evaluation Metrics

For this project our main focus is on Classifiers being sensitive to false negatives (FN). For this dataset, false negative is a person that has heart disease but classifier decided the person does not have any heart diseases. In other words, classifier said that the ill person is healthy. Because of this, I believe the recall score should be used as the most appropriate measure for classifier validation.

Project Design

- Start by importing concerned data from kaggle.
- Explore the data we have. Describe data and show some statistics.
- Chose the appropriate metric. For this case it will be recall.
- Split data into (train, test, validation) sets.
- Define the model used. It will be more than one classifier. This will give the chance to compare results deciding on the better model.
- Train the model(s).
- Analyze test results (high bias, high variance).
- Deploy the model.
- Inspect the metric.
- Compare results between models used and decide on optimal model.