CardioPredict: Machine Learning for Heart Disease Prediction

Certification Project

Domain – Healthcare

Context

Heart Disease is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. In the United States alone, heart disease claims roughly 647,000 lives each year — making it the leading cause of death. The buildup of plaques inside larger coronary arteries, molecular changes associated with aging, chronic inflammation, high blood pressure, and diabetes are all causes of and risk factors for heart disease.

The Centers for Disease Control and Prevention has identified high blood pressure, high blood cholesterol, and smoking as three key risk factors for heart disease. Roughly half of Americans have at least one of these three risk factors. The National Heart, Lung, and Blood Institute highlights a wider array of factors such as Age, Environment and Occupation, Family History and Genetics, Lifestyle Habits, Other Medical Conditions,

Race or Ethnicity, and Sex for clinicians to use in diagnosing coronary heart disease.

Diagnosis tends to be driven by an initial survey of these common risk factors followed by bloodwork and other tests.

"AIHealth" is a new age startup laying foundations in the healthcare domain by solving some of the most prominent problems by using Data Science and Machine Learning.

They are using a lot of open source data to do a lot of experimentation. You were recently hired as a Data Scientist in their research team and your role is to create a model to determine probability for a patient having heart disease or attack.

Objective

- Provide the best performing model to determine probability for a patient having a heart disease or attack.
- Providing the most important drivers for a heart disease or attack.

Data Description

The data provided consists of the following Data Dictionary

- HeartDiseaseorAttack: Target variable determining whether patient had prior heart disease or heart attack.
- HighBP: Binary flag determining whether a patient has high blood pressure.
- HighChol: Binary flag determining whether a patient has high cholesterol levels.
- BMI: Numeric value representing the Body Mass Index.
- Smoker: Binary flag determining whether a patient smokes or not.
- Diabetes: Binary flag determining whether a patient has diabetes or not.
- Fruits: Binary flag determining whether a patient consumes fruits in daily diet or not.
- Veggies : Binary flag determining whether a patient consumes vegetables in daily diet or not.
- HvyAlcoholConsump: Binary flag determining whether a patient is a heavy consumer of alcohol.
- MentHlth: Numeric value representing mental fitness, ranging from 0 to 30.
- PhysHlth: Numeric value representing physical fitness, ranging from 0 to 30
- Sex: Determining gender of the patient
- Age: The age of the patient binned into buckets between 1-13
- Education: The education level of the patient binned into buckets between 1-6.
- Income: The income of the patient binned into buckets between 1-8

Steps and Tasks

- Import libraries and load dataset
- Exploratory Data Analysis :
  - Including univariate analysis to understand the distribution of features.
  - Including multivariate analysis to determine the correlations and analysisof target variables.
  - To determine if new features can be created, based on the given data.
- Layout binary classification experimentation space (i.e. determine the list of models you would like to experiment with)
- Using precision-recall curves to determine best threshold
- Publish the performance of all the models on the same hold-out/ test dataset.
- List out the most important drivers of heart disease or attack.
- Using techniques such as oversampling, undersampling to handleclass-imbalance.

- Additional: Using model pipeline to create end to end training and inferencepipelines.