Heart Disease Prediction

Domain Background

As is mentioned in MATH 2319 Machine Learning Applied Project Phase I[1], heart disease is currently the leading cause of death across the globe. It is anticipated that the development of computation methods that can predict the presence of heart disease will significantly reduce heart disease caused mortalities while early detection could lead to substantial reduction in health care costs.

Machine Learning methods can use a large number of often complex variables obtained from a variety of medical data banks to predict whether a patient has heart disease. Recent efforts to develop computational models capable of analyzing and predicting whether a person has heart disease have shown great promise.

Problem Statement

The objective of this project is to build a binary classifier, that will take as input the attributes described in the section below, and will be able to predict whether a person has cardiovascular disease (value of 1) or not (value of 0).

Datasets and Inputs

Datasets (processed.cleveland.data.csv, processed.hungarian.data.csv, processed.switzerland. data.csv, long beach processed.va.data.csv and heart-disease. names.csv) are obtained from the UCI Machine Learning Repository[2] and will be concatenated to one dataset.

The heart-disease.names.csv file contains the details of attributes and variables.

Each dataset contains 14 attributes:

1. age in years
2. sex (1 = male; 0 = female)
3. chest pain type
   - 1 = typical angina
   - 2 = atypical angina
   - 3 = non-anginal pain
   - 4 = asymptomatic

4. resting blood pressure (in mm Hg on admission to the hospital)
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
    - 1 = true
    - 0 = false
7. resting electrocardiographic results
    - 0 = normal
    - 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
8. maximum heart rate achieved
9. exercise induced angina
    - 1 = yes
    - 0 = no
10. ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
    - 1 = upsloping
    - 2 = flat
    - 3 = downsloping
12. number of major vessels (0-3) colored by flourosopy
13. thallium test
    - 3 = normal
    - 6 = fixed defect
    - 7 = reversable defect
14. target feature: number of major vessels with >50% narrowing (0,1,2,3, or 4)

The 'target feature' refers to the presence of heart disease in patients and is comprised of an integer value for 0 (no presence) to 4 (cardiac disease present).

Target feature Distribution:

	Target feature:	0	1	2	3	4	Total
Dataset:							
Cleveland		164	55	36	35	13	303
Hungarian		188	37	26	28	15	294
Switzerland		8	48	32	30	5	123
Long Beach VA		51	56	41	42	10	200

Solution Statement

SageMaker's LinearLearner will be used to create the Heart Disease classifier, because this model is well-suited for a binary classification task that involves managing class imbalance in the training set.

We want this classifier to be able to detect cases of cardiovascular disease with high accuracy. This corresponds to a model with as many true positives and as few false negatives as possible, that is a model with a **high recall.**

The achieved recall of the classifier will have to be close or higher to the recall defined in model's benchmark, described below**.**

Benchmark Model

Heart Diseases Detection Using Naive Bayes Algorithm[3] article presented a model, which used the same dataset and achieved average recall of **74%** for the test set.

This recall will be used to benchmark our model.

Evaluation Metrics

To evaluate our classifier we are going to test that it at least has **recall** about **74%**

Project Design

The workflow for approaching the solution will include the following steps:

1. Reading our concatenated dataset and cleaning rows with missing attributes
2. Classifying rows with target feature of 2, 3 or 4 as positive (1) and rows with target feature of 0 or 1 as negative (0)
3. Visualizing attributes value distribution and class distribution
4. Splitting data to train and test sets
5. Loading the above sets to S3
6. Defining and training a classifier
7. Creating and deploying an estimator
8. Evaluating the accuracy, precision and recall of our model on the test set
9. Improving our classifier (as many times as needed), if the achieved recall is quite lower than the benchmark value. The improvement steps will include tuning out model

hyperparameters during training so that the training is optimized for the best possible recall and for taking into account the class imbalance.

References

1. *Charles Galea.* "MATH 2319 Machine Learning Applied Project Phase I." *8 Apr. 2018. https://rstudio-pubs-static.s3.amazonaws.com/396380_639e2f68b09e41a0b05f97b5dc8eb3f2.html.*

2. David W. Aha. "Machine Learning Repository." UCI Machine Learning Repository. *14 Aug. 1991. http://archive.ics.uci.edu/ml/datasets/heart+disease.*

3. K.Vembandasamy. "Heart Diseases Detection Using Naive Bayes Algorithm." IJISET. 9 Sep. 2015. *http://ijiset.com/vol2/v2s9/IJISET_V2_I9_54.pdf.*