HEART DISEASE RISK PREDICTION SYSTEM

ISM6136.001S21.23180
Data Mining

Group: Sigmoid

Kalidindi, Satya Mounica
Matta, Geeta
Surampudi, Vinuthna
Vangala, Uday Reddy
Yarra, Leela Sai

USF UNIVERSITY OF SOUTH FLORIDA

HEART DISEASE RISK PREDICTION SYSTEM

ABSTRACT:

Heart disease is a leading cause of morbidity and mortality worldwide, accounting for approximately one third of all deaths. Prevention of Heart disease requires timely identification of people at increased risk to target effective dietary, lifestyle or drug interventions. Our organization addresses this problem and provides services to the health care industry by predicting the risk of heart disease in a patient. Health care organizations collect huge amount of patient data that might contain some hidden information which could be useful for making effective decisions on whether a person is prone to heart disease or not. We develop different Machine Learning algorithms using this data which can be used to implement a Heart Disease Risk Prediction System.

BACKGROUND:

Heart disease is the biggest killer of humans. The World Health Organization (WHO) lists cardiovascular diseases as the leading cause of death globally with 17.9 million people dying every year [1]. The risk of heart disease increases due to harmful behavior that leads to overweight and obesity, hypertension, hyperglycemia, and high cholesterol [1]. Furthermore, the American Heart Association [2] complements symptoms with weight gain (1–2 kg per day), sleep problems, leg swelling, chronic cough and high heart rate [3]. Diagnosis is a problem for practitioners due the symptoms' nature of being common to other conditions or confused with signs of aging.

The Heart Disease and Stroke Statistics 2019 states that 116.4 million, or 46% of US adults are estimated to have hypertension. By 2035, more than 130 million adults, or 45.1% of the US population, are projected to have some form of CVD. Total costs of CVD are expected to reach $1.1 trillion in 2035, with direct medical costs projected to reach $748.7 billion and indirect costs estimated to reach $368 billion.

According to the World Health Organization, people with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management using counselling and medicines, as appropriate [4].

Sometimes heart disease may be "silent" and not diagnosed until a person experiences signs or symptoms of a heart attack, heart failure, or an arrhythmia. When these events happen, symptoms may include:

- **Heart attack:** Chest pain or discomfort, upper back or neck pain, indigestion, heartburn, nausea or vomiting, extreme fatigue, upper body discomfort, dizziness, and shortness of breath.
- **Arrhythmia:** Fluttering feelings in the chest (palpitations).
- **Heart failure:** Shortness of breath, fatigue, or swelling of the feet, ankles, legs, abdomen, or neck veins.

Therefore, it is important to identify whether the person is prone to a heart disease so that it can help everyone to take preventive measures.

MOTIVATION:

Our organization wants to predict whether a person is prone to heart disease based on different factors. This could be an important function to the medical field. If such a prediction is accurate enough, we can not only avoid wrong diagnosis but also save human resources. When a patient without a heart disease is diagnosed with heart disease, he will fall into unnecessary panic. Conversely when a patient with heart disease is not diagnosed with heart disease, he might miss his/her best chance for survival. Such wrong diagnosis is painful to both patients and hospitals. With accurate predictions, we can help both patients and doctors.

Data mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile computing. Of late, data mining has been applied successfully in healthcare fraud and detecting abuse cases. It has also become extremely important for heart disease prediction and treatment. It uses relevant health exam indicators and analyzes their influences on heart disease. The prediction of cardiac disease helps practitioners make more accurate decisions regarding patients' health. Therefore, the use of machine learning (ML) is a solution to help reduce and understand the symptoms related to heart disease. Besides, if we can apply our machine learning tool for individual health related predictions, we will save human resources time and money with pre-emptive care.

Clinical decisions are often made based on doctor's insights and experience rather than the Machine Learning Algorithms. This practice may lead to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

Heart disease can be managed effectively if it is predicted at an earlier stage with a combination of lifestyle changes, medicine and in some cases surgery. With the right treatment, the symptoms of heart disease can be reduced, and the functioning of the heart improved. The predicted results of our system can be used to take preventive measures and thus reduce cost for surgical treatment and other expenses.

For providing appropriate results and making effective decisions on data, some advanced data mining techniques must be used. Our organization has developed a Heart Disease Risk Prediction System using algorithms such as Two-Class Boosted Decision Tree Classifier and Two-Class Neural Network. It will find new and hidden patterns in the data and predicts the level of risk for heart disease. This can be used by the healthcare experts to deliver better quality of service.

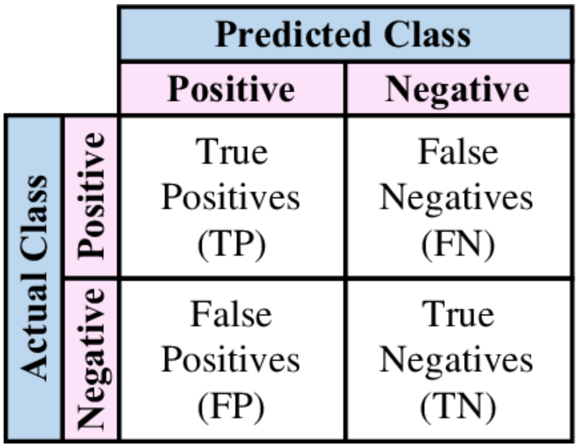
SOLUTION METHODOLOGY AND EVALUATION METRICS:

As our problem deals with predicting whether a person will be prone to Heart Disease or not, we use classification models on our data. We divide our dataset into Training and Testing datasets. We train our Training data using classifier models to find the inner patterns in our data for determining the heart disease prediction and Test our model on Test dataset and get required Evaluation Metrics.

Evaluation Metrics we use in this Scenario/Case are:

Confusion Matrix:

Confusion Matrix shows us the number of data instances which are classified into Positive class and Negative class from the Actual Positive and Negative classes. The instance which is from positive class classified as positive class it is termed as True Positive. The instance which is from negative class classified as negative class it is termed as True Negative. The instance which is from positive class classified as negative class it is termed as False Negative. The instance which is from negative class classified as positive class it is termed as False Positive.



		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Figure 1.1 - Confusion Matrix

In our case we should not consider taking the risk of False Negatives, which means, our model is predicting the Actual Heart Disease prone person into a Negative class (not prone to Heart Disease).

Precision and Recall:

Precision: It is the ratio of True Positives (TP) to the Predicted Positive Classes (TP + FP).
i.e., Out of all the Predicted Positive Classes how many are Actually Positive.

Recall: It is the ratio of True Positives (TP) to the Actual Positive Classes (TP+FN)
i.e., Out of all the Actual Positive Classes how many are Predicted Positive.

During all our tasks in this project we try to decrease the False Negatives cases, which means we are trying to increase the Recall Score.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad \text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

DATASET:

Predicting a person is prone to Heart Disease requires extensive examination of that person in many ways. Person’s living style, eating habits, his past medical history and many more ways. The dataset we have considered for this project is created by - *Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.* and *University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.* The dataset has as many as 76 attributes for a person. Out of which we only use 14 important attributes in our model.

Explanation of Attributes:

`Age`: An important attribute as the person getting old the probability of prone to heart disease is more either in the males or females.
`Sex:` This attribute finds the gender specific vulnerability patterns to the heart disease in our dataset.

cp: Chest Pain type

-- Value 1: typical angina

-- Value 2: atypical angina

-- Value 3: non-anginal pain

-- Value 4: asymptomatic

trestbps: resting blood pressure (in mm Hg on admission to the hospital)

chol: serum cholestoral in mg/dl

fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg: resting electrocardiographic results

-- Value 0: normal

-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05mV)

-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

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

-- Value 1: upsloping

-- Value 2: flat

-- Value 3: downsloping

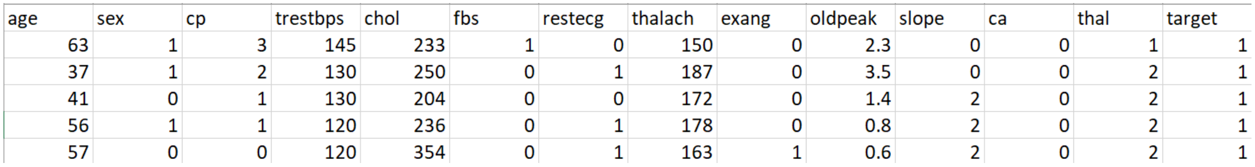
ca: number of major vessels (0-3) colored by fluoroscopy.

thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

target: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 1.2 - Snippet of dataset

The Dataset we have here is cleaned and it has no missing values. So, we can directly go ahead and fit the models to the dataset without any preprocessing steps.

BUILDING MODELS:

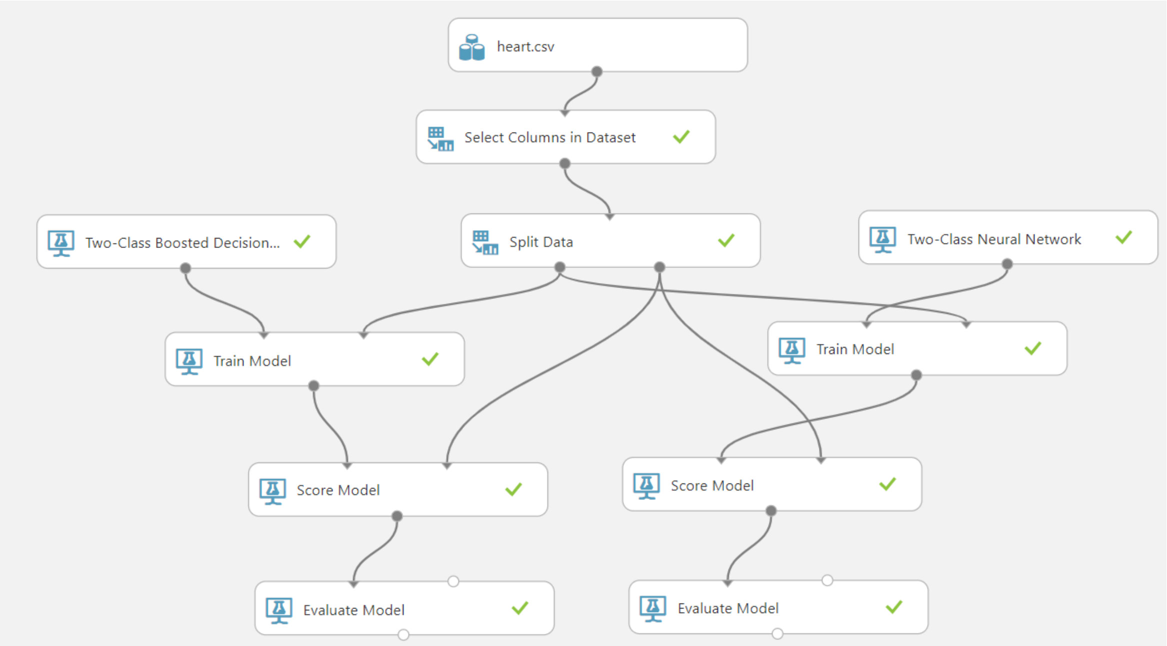


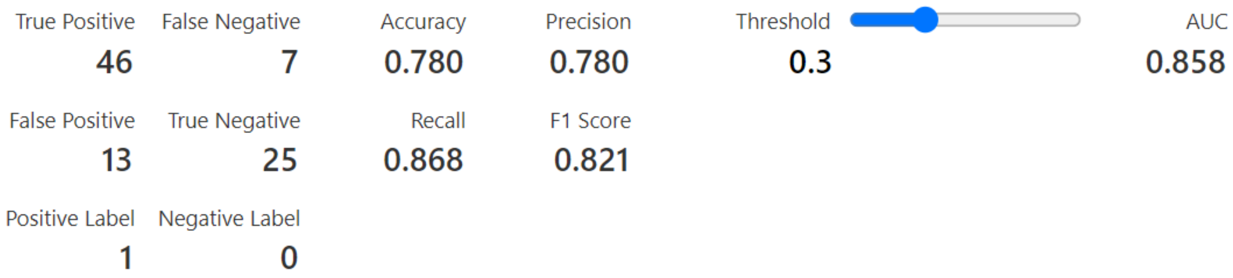
Figure 1.3 - Snippet of model creation on Azure ML

We uploaded our heart dataset into Azure ML database and used that to create and run the model. The step by step of model creation is shown in the Figure 1.3. Whiling splitting data into training and testing datasets we used 70 percent of the data for training the model and rest 30 percent of the data for testing the built model. To classify a person whether he/she is prone to Heart Disease we use the most common and most advanced models namely, Two-Class Boosted Decision Tree Classifier and Two-Class Neural Network algorithms for training our model. Later we scored the model and evaluated it on the test dataset and found the appropriate evaluation metrics to choose the best algorithm that gives better results in this particular scenario.

MODEL COMPARISION:

1. Two-Class Boosted Decision Tree:

We used the Two-Class Boosted Decision Tree Classifier on our dataset as we have only two output classes i.e., whether a person is prone to heart disease (1) or not (0). Initially we used the default parameters that is set by the Azure ML for the classifier. Default Parameters – Minimum Number of leaves per tree = 20, Minimum number of training instances required to from a leaf = 10, Learning rate = 0.2, Total number of tress constructed = 100. Metrics for the model with default parameters are as follows:



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
46	7	0.780	0.780	0.3	0.858
False Positive	True Negative	Recall	F1 Score		
13	25	0.868	0.821		
Positive Label	Negative Label				
1	0				

Figure 1.4 - Confusion Matrix and Evaluation Metrics for Two-Class Boosted Decision Tree with default parameters.

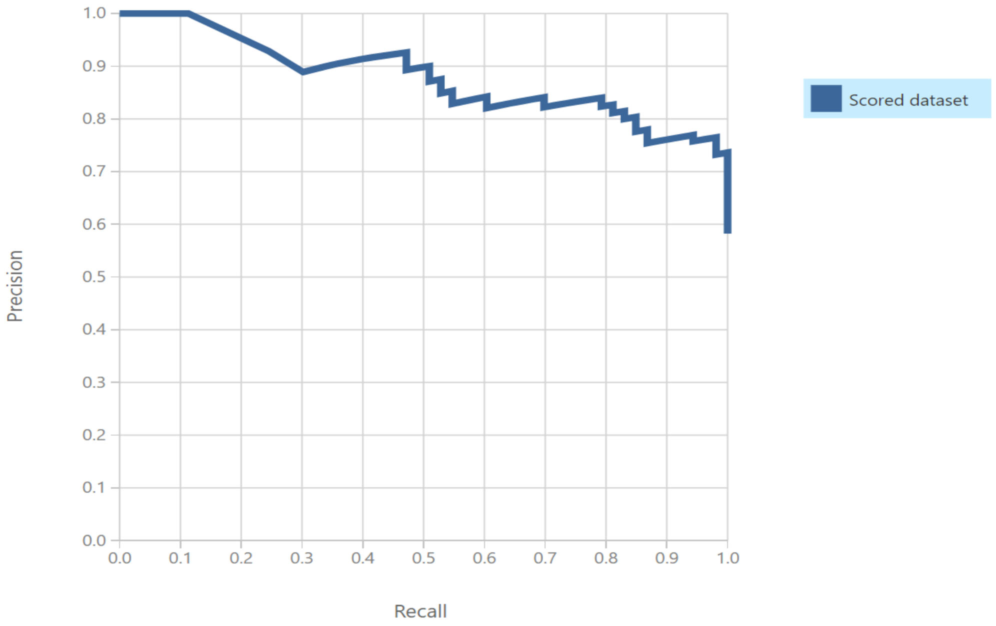
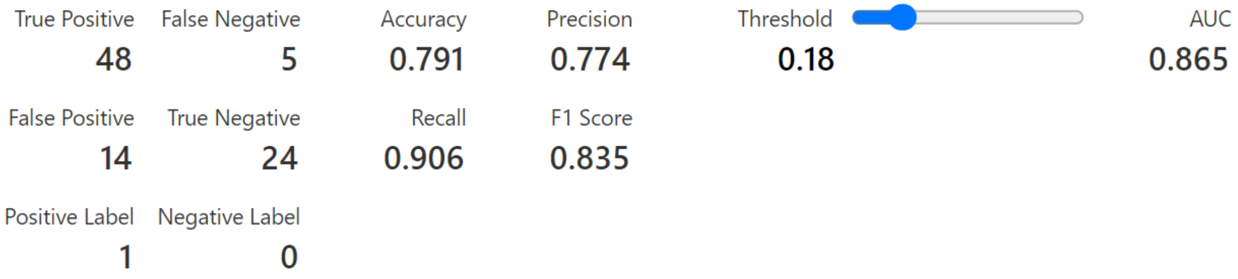


Figure 1.5 - Precision Vs Recall Curve for Two-Class Boosted Decision Tree model with default parameters

As our goal is to predict the persons prone to Heart Disease, we should concentrate more on Recall than Precision. As we should not let the person who is probably prone to heart disease to classify as not prone to heart disease. For a default threshold value (= 0.5) the Recall is 83% and Precision is 81.5%. Those are really good metrics, but can we change the threshold value to improve the Recall by maintaining the reasonable Precision? We changed the threshold value to 0.3 the result is provided in the above figure 1.4.

Now the new Recall is around 87% with Precision 78% which is a good improvement in terms of Recall improvement as well as Recall vs Precision trade off. The Precision vs Recall curve also provides evidence that the model is a good one and there is some chance of improving the model. Area Under the Curve is 0.858.

Now we try to tweak the default parameters set by Azure ML for our classifier and see if we have any improvement in the evaluation metrics. We used trial and error method for several combinations of parameters and finalized to one new set of parameters. New Parameters – Minimum Number of leaves per tree = 100, Minimum number of training instances required to from a leaf = 2, Learning rate = 0.2, Total number of tress constructed = 100. Metrics for our updated model with new parameters are as follows:



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
48	5	0.791	0.774	0.18	0.865
False Positive	True Negative	Recall	F1 Score		
14	24	0.906	0.835		
Positive Label	Negative Label				
1	0				

Figure 1.6 – Confusion Matrix and Evaluation Metrics for updated model of Two-Class Boosted Decision Tree with new parameters.

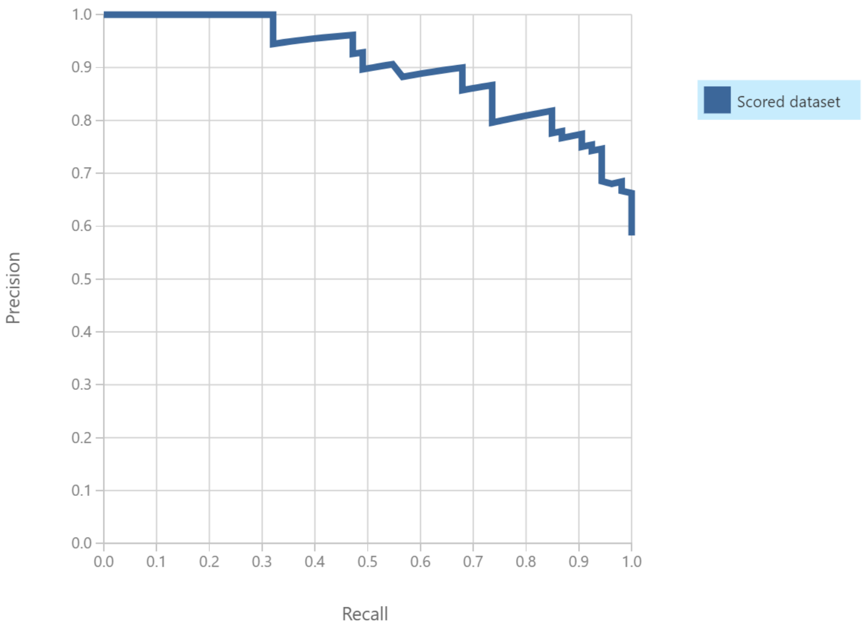


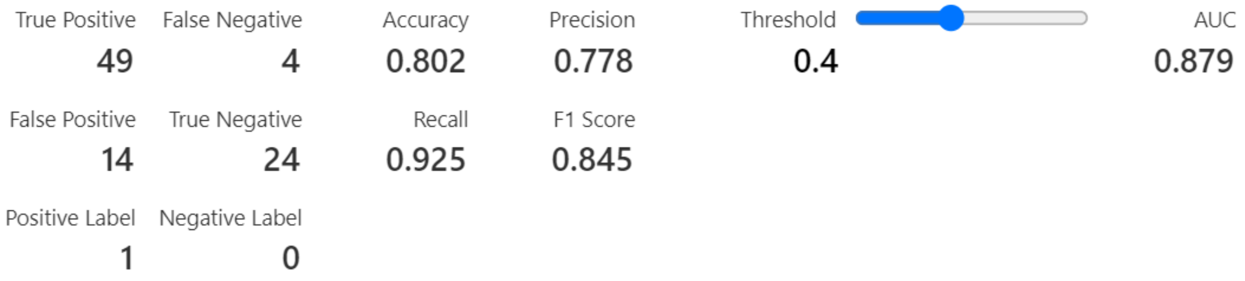
Figure 1.7 – Precision Vs Recall curve for the Two-Class Boosted Decision Tree model with new parameters.

With the new parameters for Two-Class Decision Tree model for a default threshold value (= 0.5), Recall is at 83% with the precision of 81.5%. Which is a really good improvement. But can we still adjust the model threshold value and gain significant amount of recall with reasonable precision? We tried different levels of threshold values and came to conclusion at threshold value equal to 0.18. At this threshold we have the Recall of 90.6% with Precision of 77.4% which are very significant metrics than the default model with the adjusted threshold value.

As we see the Two-Class Decision Tree model has given significant metrics, now we try Two-Class Neural Network model on our same Heart dataset and evaluate its metrics. We go with the same procedure as we went through Two-Class Decision Tree model. First, we evaluate the model with default parameters and later we tweak the parameters to get the new parameters which improves our Two-Class classification model.

2. Two-Class Neural Network:

Initially we build the Two-Class Neural Network classification model using the default parameters set by the Azure ML. Default Parameters - Number of hidden nodes = 100, Learning rate = 0.1, Number of learning iterations = 100, The initial learning weights diameter = 0.1. Metrics for the model with default parameters are as follows:



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
49	4	0.802	0.778	0.4	0.879
False Positive	True Negative	Recall	F1 Score		
14	24	0.925	0.845		
Positive Label	Negative Label				
1	0				

Figure 1.8 - Confusion Matrix and Evaluation Metrics for Two-Class Neural Network with default parameters.

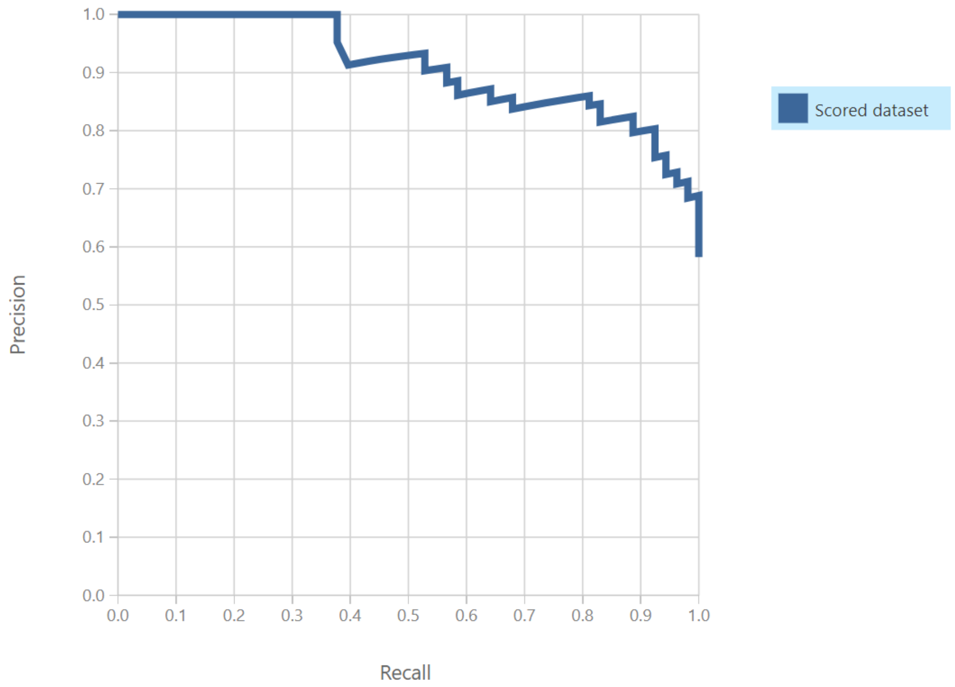
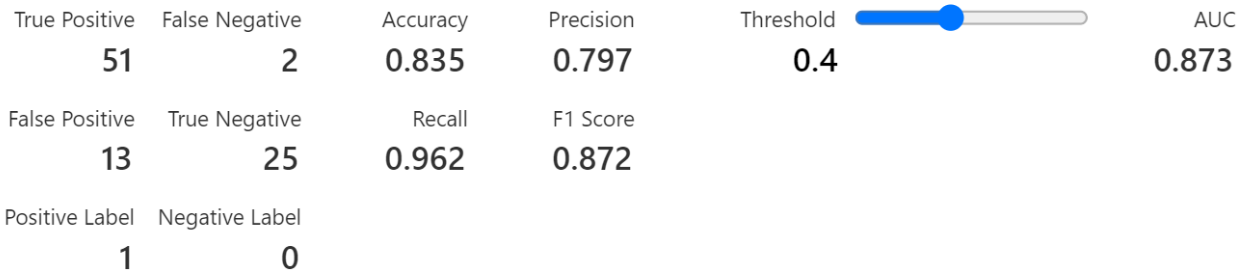


Figure 1.9 - Precision Vs Recall Curve for Two-Class Neural Network model with default parameters

For a baseline model with default threshold value (=0.5) we have Recall at 88.7% with the precision 79.7% which is on par with the tweaked Two-Class Decision Tree model. Out of curiosity, can we still try to adjust the threshold value to make our model more powerful by increasing the Recall with reasonable adjustment in Precision? We did try that and found that for a threshold value of 0.4, we have a good Recall and Precision tradeoff. Recall is at 92.5% with the Precision 77.8%. We can also observe this in the Precision vs Recall curve, as the Recall pass 0.9 the curve starts to fall more drastically. The Precision in this case is almost equal to the Precision of Two-Class Decision Tree with new parameters but Recall we have in this case is 2.5% more than that. With the same Precision, improvement in the Recall is shows that Neural Networks are doing good job here.

Now, as we did in the previous model, we try to tweak our Two-Class Neural Network model and check on the evaluation metrics. By doing several trail and errors, we set to a particular set of parameters. New Parameters - Number of hidden nodes = 150, Learning rate = 0.1, Number of learning iterations = 200, The initial learning weights diameter = 0.1. Metrics for the model with new parameters are as follows:



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
51	2	0.835	0.797	0.4	0.873
False Positive	True Negative	Recall	F1 Score		
13	25	0.962	0.872		
Positive Label	Negative Label				
1	0				

Figure 1.10 - Confusion Matrix and Evaluation Metrics for updated model of Two-Class Neural Networks with new parameters.

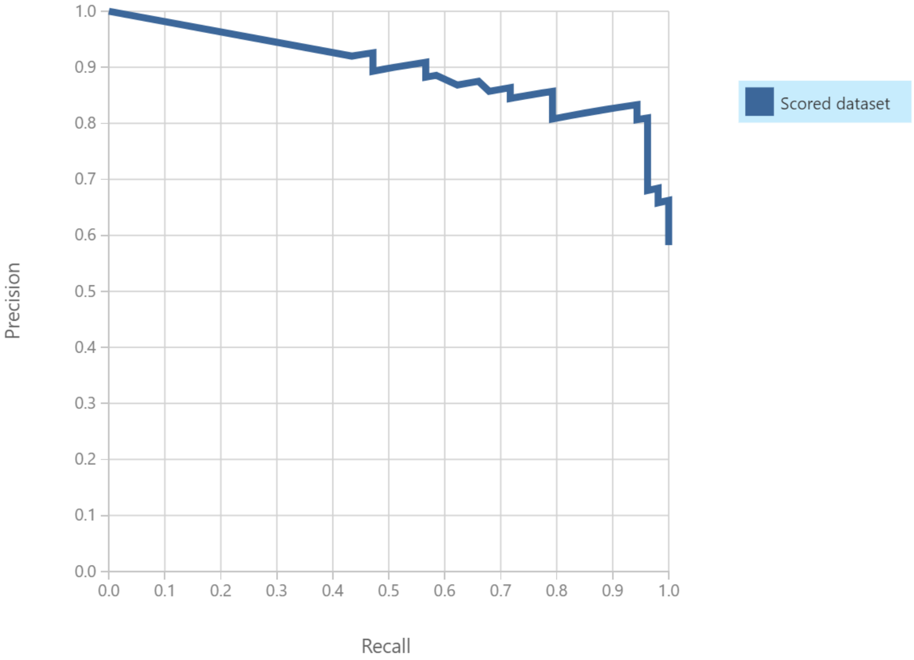


Figure 1.11 - Precision Vs Recall curve for the Two-Class Neural Network model with new parameters.

With the new parameters for our Two-Class Neural Network model at default threshold value (=0.5) we have the Recall of 94.3% with Precision 80.6%. Which are very high in both Recall and Precision for any model till now we have seen here. But still can we try to increase our Recall if we have any chance without much disturbing the precision? Let us adjust the threshold value and check this. We found at 0.4 threshold we have our model at best with new parameters. We have the Recall of 96.2% which is very huge and Precision of almost 80%, which is a great improvement from the previous model. The Two-Class Neural Networks with little tweak is doing a better job in this scenario.

We have seen 2 different classifier models with 2 variants each. Two-Class Decision Boosted Trees and Two-Class Neural Network with 2 variants each with different parameters. Of all the cases we have seen in this report, we observe that Neural Networks work well in our case and if we tweak our model a bit with cross validation and get a good set of parameters, we can get best from the Neural Network model.

CONCLUSION AND RECOMMENDATIONS:

This model is Successful in predicting the probable Heart Disease persons based on the attributes given. To make this model more reliable we can train this on large amounts of real-world data, add more attributes which may contribute towards our model if necessary and tweak a bit to get better the Evaluation Metrics.

The proposed system is scalable, reliable, and expandable. The model can also serve as a training tool for medical students and a soft diagnostic tool for physicians and cardiologists. General physicians can utilize this tool for initial diagnosis of cardio-patients. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. Generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better prediction of heart disease risk is very challenging task for future research.

References:

[1] http://www.who.int/cardiovascular_diseases/en/

[2] https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure

[3] http://www.heart.org/HEARTORG/Conditions/HeartFailure/Heart-Failure_UCM_002019_SubHomePage.jsp

[4] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)