Matthew Babik

Analysis of a 1988 Heart Disease Dataset

This dataset comes from a 1988 study from the locations of Cleveland, Hungary, Switzerland, and Long Beach. The target field simply consists of 0 meaning no disease present in the patient, and 1 meaning disease is present. All factors present in the dataset are listed below

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise-induced angina
10. Old peak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by fluoroscopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversible defect

    The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Since we are accounting for if a patient has heart disease or not the model in this overall dataset was graphed logistically.

StatsModels:

First, a logistic model was built using the StatsModels Python module. After which the data was processed and put into a logistic model. The following values were obtained:

Pseudo R squared	0.4980
Log-Likelihood	-285.28
LL-Null	-568.29

LLR p-value	2.566 e-114

P-values for variables that were observed to be non-significant in relation to the overall model were age and Fasting Blood Sugar. The pseudo R squared increased from the removal of these two variables from the inputs. The final model, when compared to test data, yielded an 85% accuracy. Due to the low LLR p-value, we can reject the null “restricted” model and prefer the more full model presented. The Log-likelihood for this model is quite low and may say for how well the overall fit of the coefficients are but can be somewhat rectified with the fair test accuracy

Sklearn:

The data was standardized and then fit the Logistic Regression function from the Sklearn Python module. Test data was separated using the train_test_split Sklearn function. The model was fit to a logistic function then tested with the created test data. The accuracy of this model was found to lie around 87%. This test accuracy was further analyzed when placed into a confusion matrix to show the values of true and false 0s and 1s.

Tensorflow:

Tensorflow was also applied to the overall dataset. The model consisted of 3 neural layers with the first two hidden layers using ReLu activation functions to firmly establish the importance of each variable from the input data. The 70 nodes per layer hyperparameter were established based on different attempts to maximize the training accuracy. A patience of 2 for the premature stopping function was determined to be optimal based on the fact that if the patient was set to 1 the program would tend to end with a training accuracy of around 80% to 85%. A patience of 3 tended to slightly overfit the overall model. Thus a training accuracy of 2 showed a high accuracy of the training data as well as the test data.