Predicting heart disease from direct cardiac conditions with a multivariable logistic regression

Kirby Leo

Abstract — **The following work describes the use of a multivariable logistic regression model for predicting heart disease from direct cardiac conditions. Five cardiac conditions were screened from the 14 attributes identified in a UCI heart disease experiment. A training set of 80% of 303 individuals with independent variables of resting ECG, maximum heart rate, ST depression, ST segment peak slope, and nuclear stress testing were used to construct a regression to predict the presence of heart disease. The quality of the regression model was analyzed via comparison of area under the curve (AUC) values in ROC (receiver operating characteristic) curves from the testing and training data. Coefficients from the model were interpreted and the overall performance was determined to be acceptable for a preliminary analysis of heart disease.**

I. INTRODUCTION

Heart disease is a broad term to describe heart conditions that involve diseased vessels, structural problems, or clotting. According to the Center for Disease Control (CDC), 1 in every 4 deaths is due to heart disease and it is the leading cause of death for men and women. There are many warning signs for heart disease and specifically the onset of a heart attack, such as chest pain or shortness of breath. However, these symptoms are not exclusive to heart disease. Thus, in this work, we examine the direct electrophysiology and cellular function of the heart to predict heart disease outcome. We hypothesize that creating a logistic regression model from direct cardiac conditions will accurately predict whether or not someone has heart disease.

II. METHODS

A. Data Processing

The dataset of interest was download from Kaggle (https://www.kaggle.com/ronitf/heart-disease-uci). This is a dataset originally with 76 attributes but narrowed down to 14 for experimental use. Data was downloaded as a .csv file and read into Colab using the Python 'pandas' module. Attributes not of interest were dropped from the set. Data were separated into training and testing sets using random assignment. The training set was derived from 80% of the data.

B. Model Fitting

A linear regression was found to be insufficient for analyzing this set of data. Due to the limited value domain for certain variables, such as the classifier itself, a linear prediction would result in arbitrarily small or large values. A multivariable neural net would be inappropriate due to the relatively small sample size of the data. Therefore, a logistic regression in the form of a Logit model using maximum likelihood estimation was determined to be the most appropriate of the three. Assuming that the impact of the multiple predictors is linear and separable:

(1) $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}$

Where each β value is a coefficient and each x value is one of the predictors.

B. Evaluating Model Performance

To evaluate model performance, the 'sklearn' module was utilized to present ROC (receiver operating characteristic) curve data for both testing and training sets. The **sensitivity**, which is defined as the true positive rate (TPR), **specificity**, which is defined as the false positive rate (FPR) subtracted from one, and the **accuracy** of the model predictions were evaluated for probability thresholds from 0 to 1. ROC curves (TPR vs FPR) and their AUC were compared.

III. RESULTS

A. Data Summary

Independent variables included from this dataset were determined by whether or not they were directly related to cardiac behavior. Resting electrocardiac results (*restecg*) were described by the values 0, 1, or 2, in which 0 corresponded to left ventricular hypertrophy, 1 corresponded to normal results, and 2 corresponded to have ST-T wave abnormalities. (*thalach*) corresponded to the maxiumum heart rate achieved. (*oldpeak*) corresponded to ST depression induced by exercise relative to rest. (*slope*) corresponded to the slope of the peak exercise ST segment. (*thal*) corresponds to nuclear stress testing in which a radioactive tracer is taken up by healthy cardiomyocytes in the heart. The values of 3, 6, and 7 correspond to different types of uptake in which 3 corresponds to normal phenotype, 6 corresponds to a fixed defect from scarred myocardium with no tracer uptake, and 7 corresponds to a reversible defect in which tracer fails to enter the myocardium under cardiac stress only. (*target*) is the classifier of this regression and is values at 0 and 1, indicating diseased or no disease respectively. These variables were individually plotted against target values (**Figure 1**).

B. Model Summary

The Logit Regression Results from the training set revealed the coefficients of our equation (1) (**Figure 2**). With respect to all other regressors held constant: it was observed that there would be a 0.2802 change in the outcome of heart disease for every 1 unit change in *restecg*, a 0.0338 change in the outcome for every 1 unit

change in *thalach*, a -0.5231 change in the outcome for every 1 unit change in *oldpeak*, a 0.2158 change in the outcome for every 1 unit change in *slope*, and a -1.3299 change for every 1 unit change in *thal*. The constant intercept is the starting point of the model, but has no physical relevance.

C. *Model Performance*

The sensitivity, specificity, and accuracy over the probability thresholds were visually similar between the testing and training sets (**Figure 3**). Decreases and increases in sensitivity and specificity respectively were observed with increasing threshold. Interestingly, a decrease in accuracy at the extreme low and high ends of the probability thresholds was observed. The difference in AUC of the ROC curves constructed from the sets was 0.0254 (**Figure 4**).

IV. DISCUSSION AND CONCLUSION

From our interpretation, we can see that decreases, due to the negative coefficients, in ST depression (*oldpeaks*) and nuclear stress testing score (*thal*) relate to a prediction less weighted towards heart disease. Considering that each variable is on a different range scale, it is not appropriate to assign importance to specific variables from this model.

The performance of the model itself was evaluated via ROC curves and associated test performance metrics. The trends in sensitivity and specificity align with typical interpretation: that thresholds set higher will result in decreased sensitivity and increased specificity (vice versa for thresholds set lower). Setting a threshold at ~0.6 results in a convergence of equal sensitivity and specificity along with relatively high accuracy. For the ROC curve, the closer the ROC curve approaches the red diagonal, the less accurate the test. A lower AUC value indicates a lower probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.

Overfitting of the model was controlled by checking the differences between testing and training sets. Due to the similarity between sets in terms of sensitivity, specificity, and accuracy in addition to the low difference in ROC curve AUC values, we could conclude that overfitting was not causing major issues with prediction. Directional differences in AUC between the testing and training sets are not strong enough for any indicative argument of fitting behavior.

The function minimized in logistic regression, called the cross-entropy, is under the assumption that the gold standard values (targets) are independent. However, heart disease is incredibly complex with cross-tendencies. One way to optimize the model might be to create pseudo testing sets within the training set itself. Iterating through the most accurate models would result in a higher accuracy model to use with the real test data.

Future work may be to fit this similar type of regression model for the non-direct cardiac attributes. For example, cholesterol levels and blood sugar may give insight on what symptoms are better predictors for heart disease if the accuracy of the models are compared.

V. FIGURES

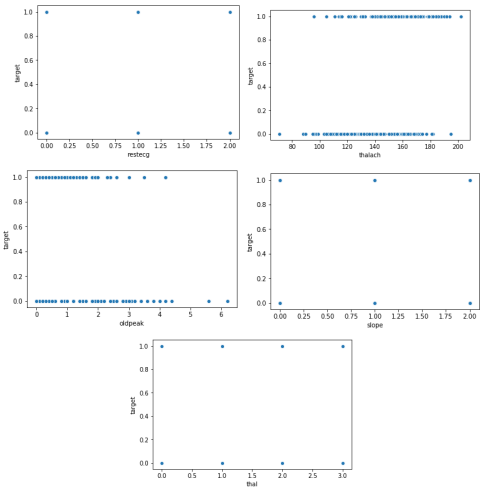

Fig. 1. Summary of initial data

Logit Regression Results

Dep. Variable:	target	**No. Observations:**	237
Model:	Logit	**Df Residuals:**	231
Method:	MLE	**Df Model:**	5
Date:	Sun, 06 Oct 2019	**Pseudo R-squ.:**	0.2683
Time:	18:00:32	**Log-Likelihood:**	-118.90
converged:	True	**LL-Null:**	-162.50
Covariance Type:	nonrobust	**LLR p-value:**	2.605e-17

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6865	1.390	-1.213	0.225	-4.411	1.038
restecg	0.2802	0.305	0.918	0.359	-0.318	0.879
thalach	0.0338	0.009	3.945	0.000	0.017	0.051
oldpeak	-0.5231	0.186	-2.807	0.005	-0.888	-0.158
slope	0.2158	0.309	0.698	0.485	-0.390	0.822
thal	-1.3299	0.282	-4.720	0.000	-1.882	-0.778

Fig. 2. Logit regression results

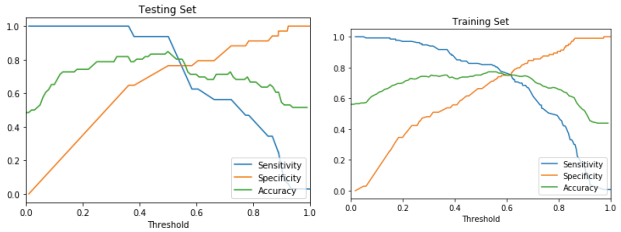

Fig. 3. Sensitivity, specificity, and accuracy for testing (left) and training (right) data

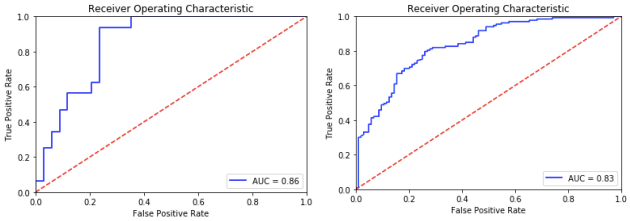

Fig. 4. ROC curves for testing (left) and training (right) data

APPENDIX

Github link to code:
https://github.com/ds4bmeIntroFall2019/project-krbyktl/blob/master/FinalProject.ipynb