JSC370 Final Project

John Chen

Introduction

The purpose of this analysis is to attempt to create a model that best predicts the existence of a heart disease given the body condition of an individual. The model will also be used to examine the relationship and significance between selected features and the response variable. The dataset we will be using is the Heart Failure Prediction Dataset. It contains 918 observations 11 features: Age, Sex, Chest Pain Type, Resting blood pressure, Cholesterol, Fasting blood sugar, Resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, oldpeak, and the slope of the peak exercise ST segment. Duplicates were removed.

Method

Our dataset is cross-sectional and combined from 5 datasets used for heart disease research: Cleveland, Hungarian, Switzerland, Long Beach, Stalog Data set. They are all from the UCI Machine Learning Repository. The dataset will be split into 80-20 portions for training and testing respectively. We will compare the accuracy between logistic regression, random forest, and Extreme Gradient Boosting and interpret the best model. The assumptions of logistic regression is met: observations are independent from each other(each patient only has one observation), the response variable is binary, and we will be using the logit function as the link function. For testing, the predicted probability will be evaluated to 1 if its above 0.5, otherwise it will be evaluated to 0. For random forest and Extreme Gradient Boosting, we will fine tune the parameters for best accuracy.

EDA

Through basic data wrangling, this data contains no missing values and all variables types are in their expected type, categorical predictors will be transformed into factors for decision tree models training, and all unique values of categorical variables make sense, and the response variable only has two values for heart disease indication.

The distribution of heart disease is about uniform. The clustering of the scatterplots between predictors does not show non-linear pattern. This suggest that the regression should be a good fit.

There are 172 observations with 0 serum cholesterol which doesn't make any sense. Based on this study, there is no evidence that there is a correlation between cholesterol and heart disease. Therefore, it is expected that replacing 0 cholesterol with the median will not have significant impact on training results.

Results

AUC of the logistic regression model is 0.89 which indicates that the model is good at ranking the probabilities of the presence of heart disease in the training set. Accuracy of the model is around 16% which indicates the it is not good at predicting new dataset. This could mean that the model is overfit. The coefficients of the model is presented below.

term	estimate	std.error	statistic	p.value
(Intercept)	0.2800439	0.9096977	0.3078428	0.7582020
SexM	1.5169741	0.3116394	4.8677225	0.0000011
ChestPainTypeATA	-2.1393969	0.3907402	-5.4752413	0.0000000
ChestPainTypeNAP	-1.5470520	0.2920778	-5.2967126	0.0000001
ChestPainTypeTA	-1.7367296	0.5011811	-3.4652737	0.0005297
FastingBS	1.4607001	0.3012573	4.8486799	0.0000012
MaxHR	-0.0122388	0.0051253	-2.3879102	0.0169445
ExerciseAnginaY	0.9220335	0.2751079	3.3515347	0.0008036
Oldpeak	0.3480843	0.1289783	2.6987824	0.0069594
ST_SlopeFlat	1.4988307	0.4682215	3.2011145	0.0013690
ST_SlopeUp	-1.0444413	0.4871879	-2.1438165	0.0320476

Logistic Regression	Random Forest Model	Extreme Gradient Boosting
16.30435	16.30435	12.5

We can interpret the coefficient as odds ratio. For example, the odds of male having heart disease is 0.212248 times higher than the odds of female having heart disease.

We have the following accuracy percentage after training random forest and extreme gradient boosting models.

The accuracy for all model is low. This could a sign of overfitting for all models. This result might also indicate that the features are not good predictors of heart disease. Refering to the ranking table of on home webpage, slope of the peak exercise ST segment and max heart rate seems to be extremely relevant because it was showed as the most relevant variables in both models. It is unexpected that Cholesterol is relatively high on the ranking since literature concluded that there is no evidence between high cholesterol and higher risk of heart disease. This could be because that the relationship between cholesterol and heart disease depends on cofounders.

Conclusion

We have failed to train a model that accurately predict the presence of heart disease. There is an indication of overfitting. Models like deep learning with heavy regularization might be more fitting for this analysis. However, we did concluded that slope of the peak exercise ST segment and max heart rate are extremely relevant in prediction, more specifically, Upsloping ST segment and low max max heart rate.