Predicting heart disease status by patients' information and medical test results

Yun He (yh3094), Jun Lu (jl5297), Haoran Hu (hh2767)

Introduction

Heart diseases are the leading cause of death in the US[1]. According to the American Heart Association (AHA), the number of deaths due to cardiovascular diseases (CVD) is 840,678 in the US in 2016, and that is one third the number of all deaths[2]. In many cases, cardiovascular diseases do not show any visible symptoms until it is too late[3]. Early accurate diagnosis of CVD can help patients take timely appropriate treatments and improve survival.

The gold standard for establishing the presence of coronary heart disease is invasive coronary angiography. However, this technique is invasive and costly. Potential cardiovascular patients are often sent for other multiple noninvasive tests, and these test results help doctors with the diagnosis. However, the accuracy of the diagnosis using multiple medical tests often relies on the individual doctor's knowledge and experiences, which varies from person to person.

In this project, the Cleveland CAD dataset from the University of California, Irvine (UCI) was used. The aim of this project was to build a model using machine learning methods to predict the presence of heart disease based on the results of several medical tests and patients' gender and age, and provide some insights on day-to-day medical practice.

The presence of heart disease (determined by the gold standard) is treated as a response, while 13 other variables (including 11 different medical test results and patients' gender and age) in the dataset are all potential predictors. The detailed description of these variables can be found in Table.1. The original dataset contains 303 observations in total. Two observations with missing values were removed from the dataset.

Exploratory Analysis

In the exploratory data analysis for the continuous variables, it can be seen that people with the narrowing vessels tend to be older and have higher ST depression value induced by exercise relative to rest and lower maximum heart rate achieved during the thallium stress test (Table 1, Appendix Fig.1 and Fig. 2). For categorical variables, the presence of heart disease is likely to be associated with sex, chest pain type, exercise-induced angina, the slope of peak exercise ST segment, thallium stress test result and the number of major vessels colored by fluoroscopy (Table 1 and Appendix Fig. 3).

By projecting data into two-dimensional subspace composed of the first principle component and the second principal component of 13 centered and scaled variables, we can see that the response can be separated to some extent based on these variables (Fig. 1). However, there is not a sharp border between the presence and absence of heart disease in this subspace.

Model

Nine classifiers were built in this project, including regularized logistic regression, linear discriminant analysis (LDA), Naive Bayes, decision tree, bagging, random forests, Adaboost, support vector machine (SVM) with different kernels and single-layer neural network[4].

10-fold cross-validation method was utilized for model selection. For those models which can produce class probabilities, we tuned the model and determined the best sets of tuning parameters by comparing the area under the curve (AUC) in the cross-validation. For SVM models, we also chose the ROC as a metric using the train function in the caret package, although the SVM model itself does not generate class probabilities. Therefore, cross-validated AUC can be used to compare models built by different algorithms.

A regularized logistic regression model was first fitted for simplicity and easier interpretation purpose. It kept all variables and performs relatively well on this dataset (median cross-validated AUC 0.900). In the fitted regularized logistic regression, the coefficients of variables are consistent with the patterns we found in the exploratory analysis. For example, older people and people have higher ST depression value tend to have a higher risk of having heart disease. Also, people with angina induced by exercise are at higher risk of having heart disease. Also, people who have fixed or reversible defects in a thallium stress scintigraphy test are at higher risk. Those findings are clinically acceptable[5].

$$
\begin{aligned}
\log\frac{\mathrm{Y}}{1-\mathrm{Y}} = {} & -0.624 + 0.009\,\mathbf{X}_{\mathbf{age}} + 0.001\,\mathbf{X}_{\mathbf{chol}} + 0.199\,\mathbf{X}_{\mathbf{oldpeak}} + 0.306\,\mathbf{X}_{\mathbf{ca}} + 0.005\,\mathbf{X}_{\mathbf{trestbps}} \\
& + 0.462\mathbf{I}(\mathbf{Sex} = \mathbf{Male}) - 0.385\,\mathbf{I}(\mathbf{cp} = \mathbf{Asymptomatic}) \\
& - 0.586\,\mathbf{I}(\mathbf{cp} = \mathbf{Atypical\ angina}) - 0.524\,\mathbf{I}(\mathbf{cp} = \mathbf{Non} - \mathbf{anginal\ pain}) \\
& + 0.522\,\mathbf{I}(\mathbf{exang} = \mathbf{Yes}) + 0.293\,\mathbf{I}(\mathbf{slope} = \mathbf{flat}) \\
& - 0.290\mathbf{I}(\mathbf{slope} = \ \mathbf{downsloping}) - 0.527\,\mathbf{I}(\mathbf{thal} = \mathbf{normal}) \\
& + 0.526\,\mathbf{I}(\mathbf{thal} = \mathbf{Reversible\ defect}) - 0.246\,\mathbf{I}(\mathbf{restecg} = \ \mathbf{abnormal\ ST} - \mathbf{T}\ ) \\
& + 0.198\,\mathbf{I}(\mathbf{restecg} = \mathbf{probable}\ ) - 0.069\,\mathbf{I}\ (\mathbf{fbs} = > \mathbf{120})
\end{aligned}
$$

Y: the predicted probability of the presence of heart disease.
The reference for categorical variables: $\mathrm{Sex} = \mathrm{Female}$, $\mathrm{cp} = \mathrm{Typical\ angina}$, $\mathrm{exang} = \mathrm{No}$, $\mathrm{slope} = \mathrm{Upslope}$, $\mathrm{thal} = \mathrm{fixed\ defect}$, $\mathrm{restecg} = \mathrm{normal}$, $\mathrm{fbs} = < 120$.

Naive Bayes (median cross-validated AUC 0.899) and LDA (median cross-validated AUC 0.900) were also tried. Naive Bayes is a powerful algorithm for predictive modeling which assumes no dependency between variables attempting to maximize the posterior probability in determining the class while LDA assumes the dependency.

A classification tree model (median cross-validated AUC 0.847) was fitted to capture the nonlinearity and interaction effects. The cross-validation resulted in a tree with 8 terminal nodes (Fig. 2). The internal nodes correspond to splitting the values of 6 variables including thallium stress test result, chest pain type, the number of major vessels colored by fluoroscopy, exercise-induced angina, the slope of peak exercise ST segment, and age. The model is easily interpretable, but the prediction performance is not satisfactory.

To improve the predictive performance, ensemble methods such as bagging, random forests and boosting were used. After picking the best tuning parameters, the median cross-validated AUC of the random forest, boosting and bagging models are respectively 0.911, 0.907, and 0.894. The random forests model performs best among these tree-based models in terms of cross-validated AUC, though its interpretability is not as good as that of the classification tree. To interpret the random forest, a variable importance plot (Fig. 3) and partial dependence plots (Fig. 4) were generated. The top three most important predictors in the random forests are thallium stress test result, the number of major vessels colored by fluoroscopy, ST depression induced by exercise relative to rest. In the partial dependence plots, the probability of having CAD tends to increase with the increase of the number of major vessels colored by fluoroscopy, ST depression induced by exercise relative to rest, and age. On the other hand, the probability of having CAD tends to decrease with the increase of thallium stress test result.

We also tried two more complex algorithms: SVM with Gaussian kernel (median cross-validated AUC 0.901) and single-layer neural network (median cross-validated AUC 0.901).

Using the varImp function in the caret package, we computed the variable importance for each model. We found that the thallium stress test result and the number of major vessels colored by fluoroscopy contributed a lot in the prediction of each model. Generally, in the real-world diagnosis of coronary heart disease, fluoroscopy and stress thallium scintigraphy are indeed the two most popular tests among those medical tests[5, 6, 7]

Conclusion

Random forests performs best among the models we tried by comparing the median of the cross-validated AUC, while the simpler model fitted by using the logistic regression also performed well (Fig. 5 and Table 3).

Regarding all the models we tried, their cross-validated AUC did not differ a lot. One possible reason is that our sample size is not large enough for complex algorithms to show their advantages.

In medicine, the clinical interpretation of the decision-making procedure is critical. Hence, complex algorithms (including SVM, ensemble methods and artificial neural network) that use the black-box mathematical methodology are not ideal[7]. Thus, we would choose the regularized logistic regression and Naive Bayes as our final models which have relatively good performance in prediction and can be interpreted easily.

According to the models we fitted, we suggest clinicians pay more attention to fluoroscopy and stress thallium scintigraphy test which play important roles in predicting the response.

One limitation of our project is that we didn’t partition our dataset into the training dataset and test dataset because of the limited sample size. Otherwise, we could use the test dataset to further validate our results. In addition, the patient data were collected from one hospital, which results in the limited generalizability of our findings.

Reference

[1]: Heron, M.P., 2018. Deaths: Leading causes for 2016.

[2]: Benjamin, E.J., Muntner, P. and Bittencourt, M.S., 2019. Heart disease and stroke statistics-2019 update: A report from the American Heart Association. *Circulation*, 139(10), pp.e56-e528.

[3]: National Institute on Aging. (2019). *Heart Health and Aging*. [online] Available at: https://www.nia.nih.gov/health/heart-health-and-aging [Accessed 9 May 2019].

[4]: Jamse G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York, NY, USA: Springer; 2017.

[5]: Kohsaka, S. and Makaryus, A.N., 2008. Coronary angiography using noninvasive imaging techniques of cardiac CT and MRI. *Current cardiology reviews*, 4(4), pp.323-330.

[6]: Detrano, R., Simpfendorfer, C., Day, K., Salcedo, E.E., Rincon, G., Kramer, J.R., Hobbs, R.E., Shirey, E.K., Rollins, M. and Sheldon, W.C., 1985. Comparison of stress digital ventriculography, stress thallium scintigraphy, and digital fluoroscopy in the diagnosis of coronary artery disease in subjects without prior myocardial infarction. *The American journal of cardiology*, 56(7), pp.434-440.

[7]: Vliegenthart, R., 2004. Detection and quantification of coronary calcification. In *Coronary Radiology* (pp. 175-184). Springer, Berlin, Heidelberg.

[8]: W. N. Price II., Black-box medicine. Harv. J. L. Tech. 28, 419–467 (2015).

Table 1. Description of the variables.

Variable	Description	Type	Categories
age	Age in years	int	-
ca	Number of major vessels colored by fluoroscopy	int	-
chol	Serum cholesterol (mg/dl)	con	-
trestbps	Resting blood pressure (mm Hg)	con	-
thalach	Max. heart rate achieved during thallium stress test	con	-
oldpeak	ST depression induced by exercise relative to rest	con	-
sex	Female or male	bin	0-female, 1-male
fbs	Fasting blood sugar (< 120 mg/dl or > 120 mg/dl)	bin	0-<120, 1->120
exang	Exercise induced angina	bin	0-no, 1-yes
cp	Chest pain type	cat	0-typical angina 1-atypical angina 2-non-anginal pain 3-asymptomatic
restecg	Resting electrocardiographic results	cat	0-normal 1-having ST-T wave abnormality 2-showing probable or definite left ventricular hypertrophy
slope	Slope of peak exercise ST segment	cat	0-unsloping 1-flat 2-downsloping
thal	Thallium stress test result	cat	1-fixed defect 2-normal 3-reversible defect
target	The presence of heart disease	bin	0-presence, 1-absence

Variables of the heart disease patients in the dataset; the last row (target) is the response variable; the type column indicates whether a variable is binary (bin), integer (int), categorical (cat), or continuous (con); categories were shown for binary and categorical variables.

Table 2. Descriptive statistics of variables stratified by the response.

		Absence	Presence	p-value
N		164	137	
age (mean (SD))		52.49 (9.58)	56.64 (7.98)	<0.001
trestbps (mean (SD))		129.31 (16.22)	134.45 (18.79)	0.011
chol (mean (SD))		242.39 (53.68)	251.43 (49.47)	0.133
thalach (mean (SD))		158.73 (18.93)	138.98 (22.64)	<0.001
oldpeak (mean (SD))		0.59 (0.78)	1.59 (1.30)	<0.001
sex = 0/1 (%)		71/93 (43.3/56.7)	24/113 (17.5/82.5)	<0.001
exang = 0/1 (%)		141/23 (86.0/14.0)	62/75 (45.3/54.7)	<0.001
fbs = 0/1 (%)		141/23 (86.0/14.0)	116/21 (84.7/15.3)	0.877
slope (%)				<0.001
	0	9 (5.5)	12 (8.8)	
	1	49 (29.9)	90 (65.7)	
	2	106 (64.6)	35 (25.5)	
thal (%)				<0.001
	1	6 (3.7)	12 (8.8)	
	2	130 (79.3)	36 (26.3)	
	3	28 (17.1)	89 (65.0)	
restecg (%)				0.005
	0	67 (40.9)	79 (57.7)	
	1	96 (58.5)	55 (40.1)	
	2	1 (0.6)	3 (2.2)	
cp (%)				<0.001
	0	39 (23.8)	103 (75.2)	
	1	41 (25.0)	9 (6.6)	
	2	68 (41.5)	18 (13.1)	
	3	16 (9.8)	7 (5.1)	
ca (%)				<0.001
	0	129 (78.7)	44 (32.1)	
	1	21 (12.8)	44 (32.1)	
	2	7 (4.3)	31 (22.6)	
	3	3 (1.8)	17 (12.4)	
	4	4 (2.4)	1 (0.7)	

P-values of continuous variables were calculated by the t-test; p-values of categorical variables were calculated by the chi-square test.

Table 3. Median and Mean Cross-validated training AUC.

Model	Median cross-validated AUC	Mean cross-validated AUC
Random Forest	0.911	0.909
AdaBoost	0.907	0.906
SVM (linear kernal)	0.906	0.904
SVM (radial kernal)	0.901	0.903
Neural network (single layer)	0.901	0.903
LDA	0.900	0.898
Regularized logistic regression	0.900	0.902
Naive Bayes	0.900	0.910
Bagging	0.894	0.893
Classification Tree (CART)	0.847	0.845

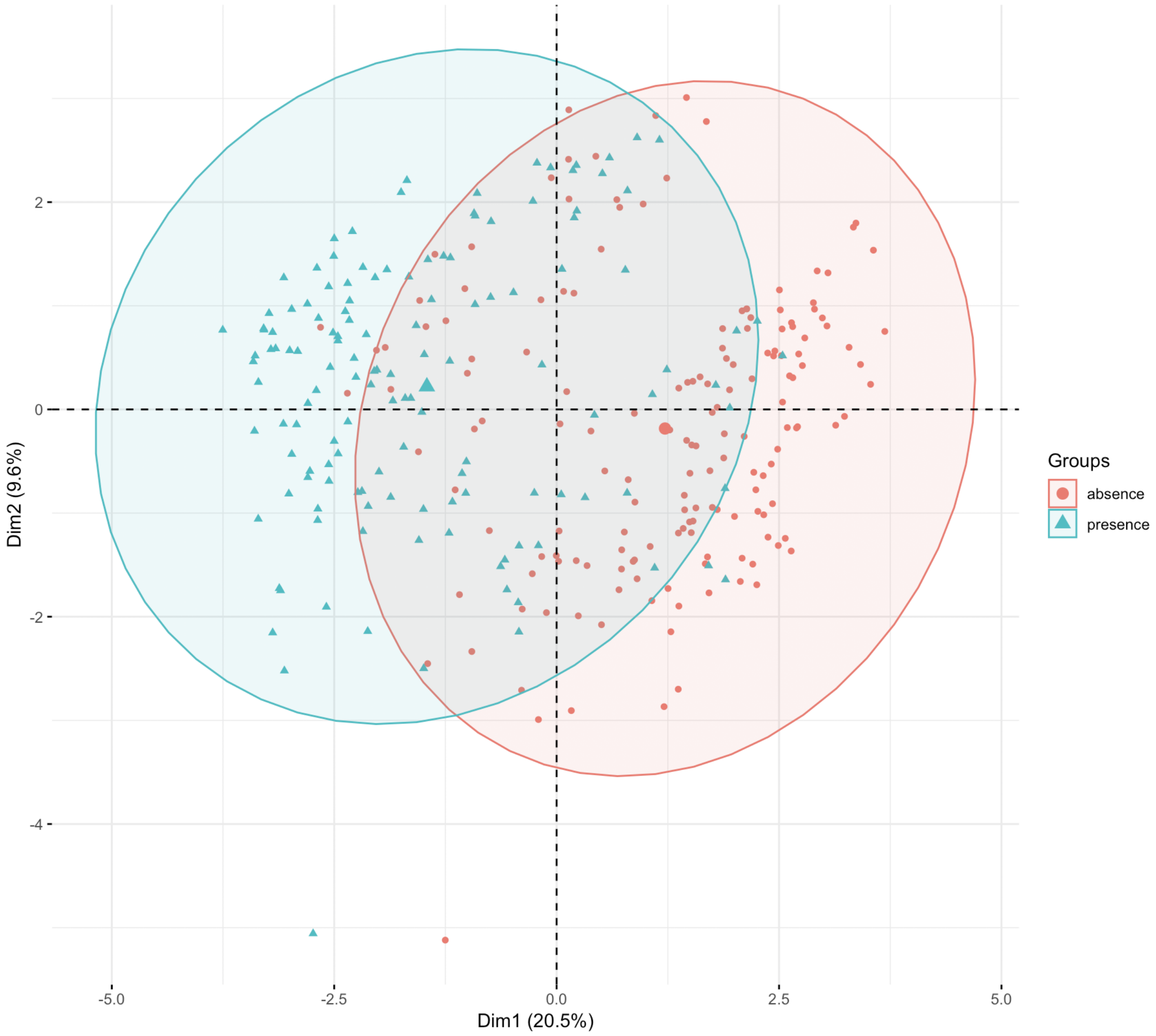

Figure 1. PCA Plot. The x-axis is the first principle component (explain 20.5% variance) score and the y-axis is the second principal component (explain 9.6%) score. Two categories of the response were emphasized in two different colors.

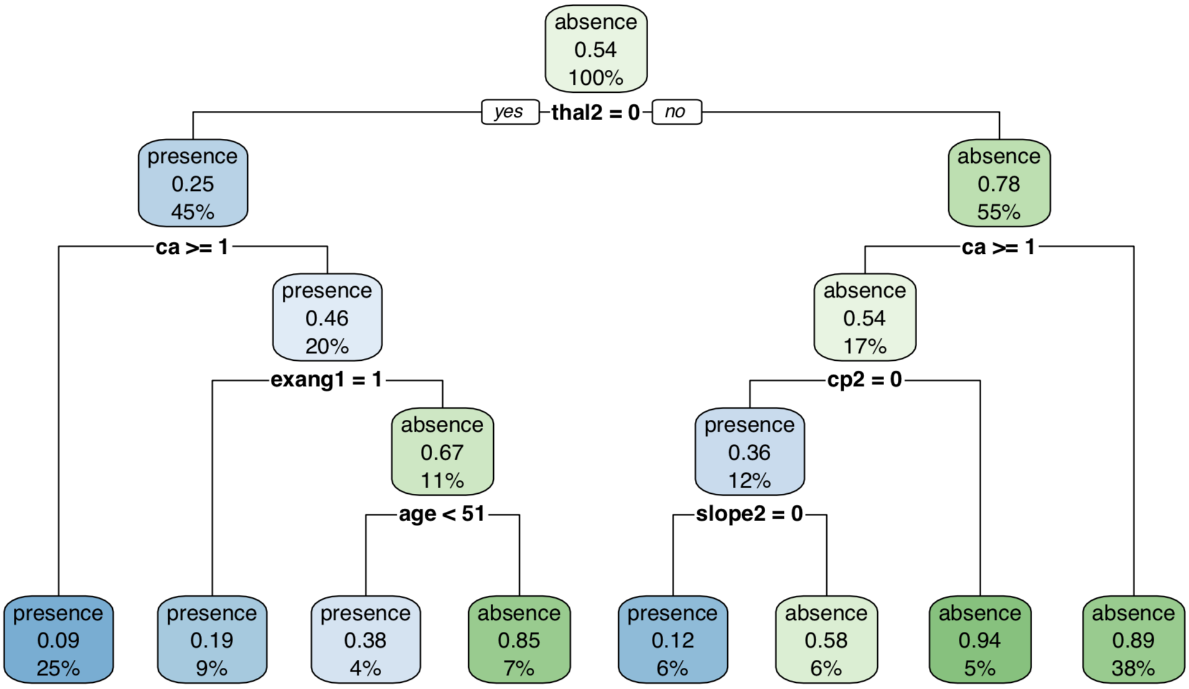

Figure 2. Decision Tree Plot. Each node shows the predicted class (presence or absence), the predicted probability of absence, and the percentage of observations.

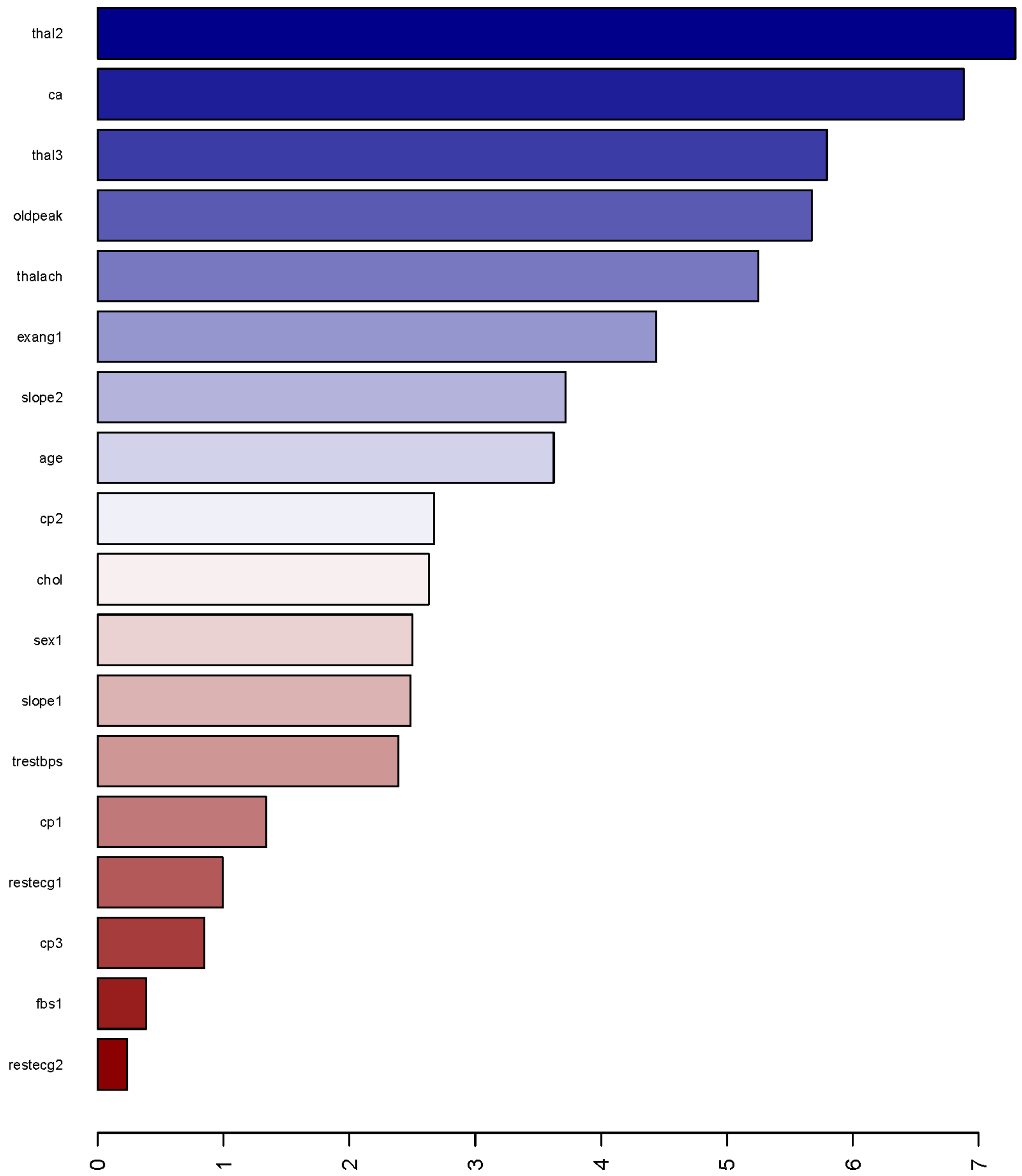

Figure 3. Variable Importance Plot of the Random Forests model.

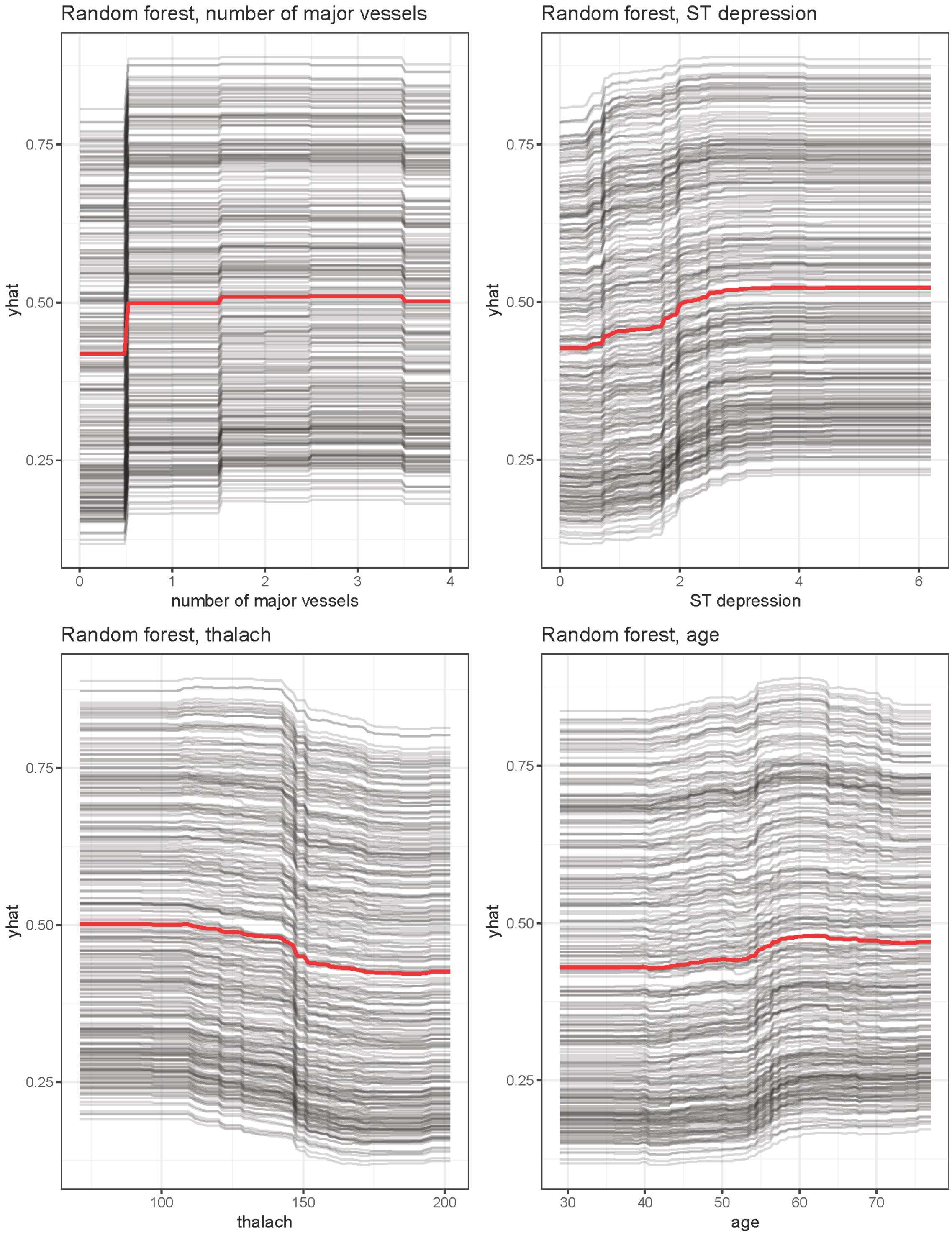

Figure 4. Partial Dependence Plot of 4 Continuous Variables. The y-axis is the predicted probability of the presence of heart disease.

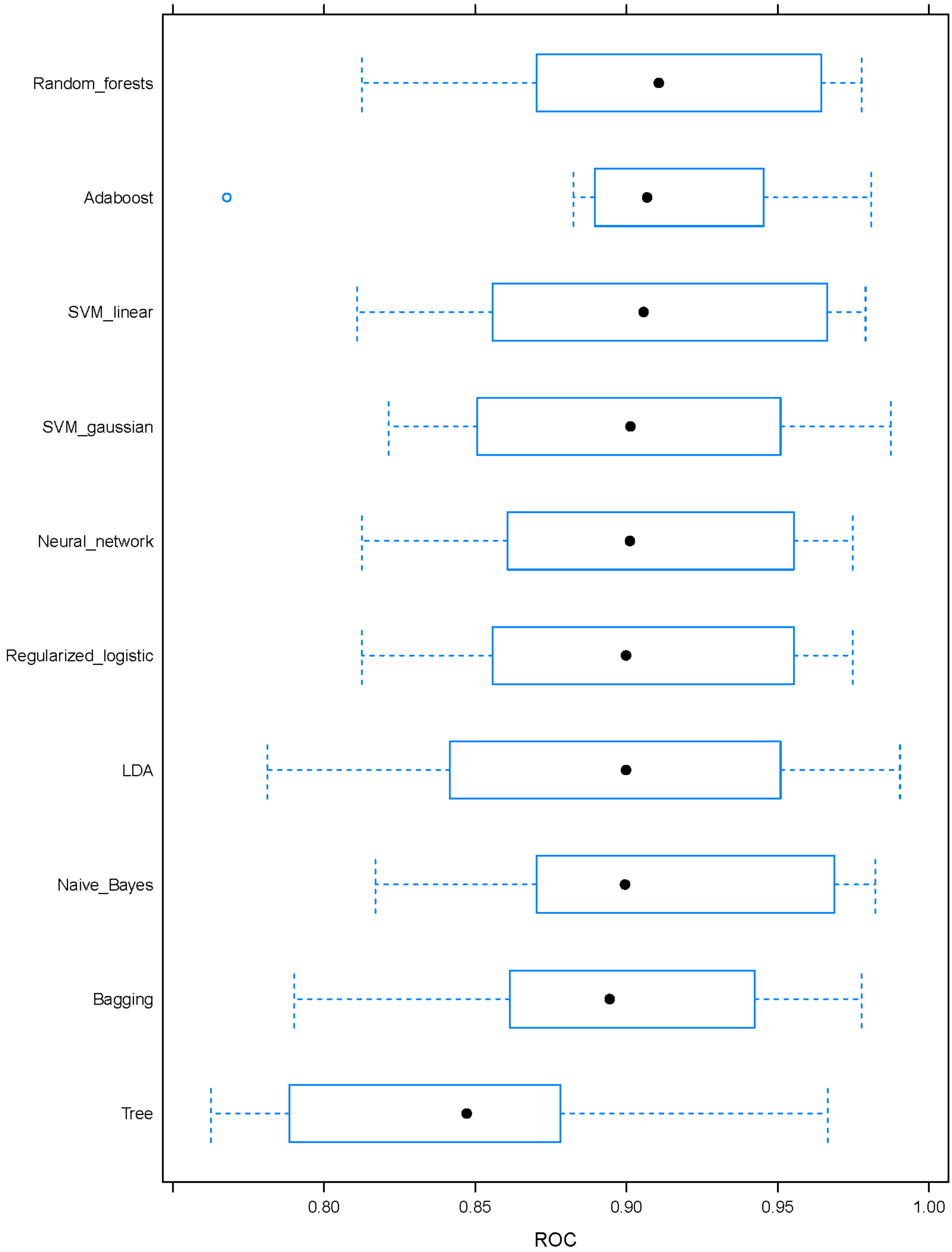

Figure 5. Boxplots of the AUC values for different models by 10-fold cross-validation.