P8106 Midterm Project
Weixi Chen
03/29/2021

Introduction

Heart disease causes 1/4 deaths in the United States and coronary heart disease (CHD) is the most common type. Analyzing the determinants of CHD can help patients to be screened, diagnosed, and treated by medications earlier. This project intends to find the significant determinants in predicting whether the patient has ten-year risk of CHD and to gain a best model in prediction performance by comparing five supervised classification models.

The data used to fit these models is obtained from Kaggle website and is consisted of 15 predictors including demographic, behavioral, history medical, current medical information and one binary response variable of the ten-year risk of CHD from 4,228 patients living in Framingham, Massachusetts. The data is preprocessed by remedying the severe missing data. There are 9.16% missing in the glucose variable which is highly correlated to the diabetes variable (***Figure 1***). Hence, the method of "bagImpute" is implemented to recover the glucose data based on the diabetes data. And since the other variables have few missing data, those observations are dropped.

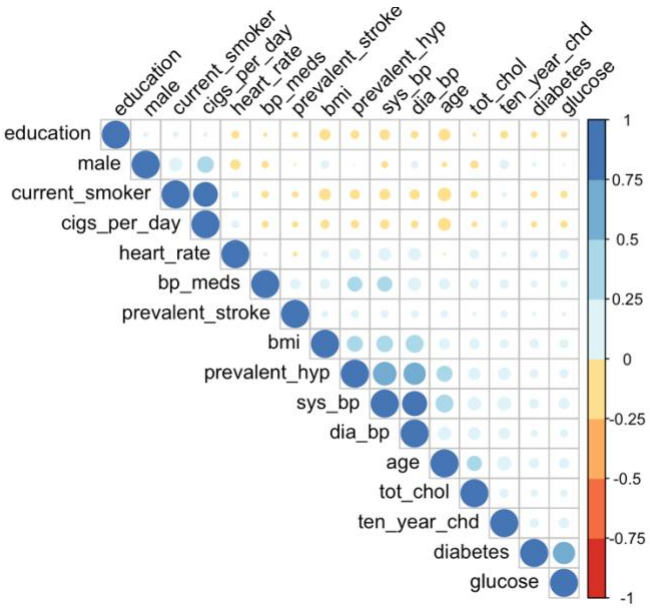

Figure 1. Correlation map of all variables.

Exploratory Analysis/Visualization

The continuous predictors are visualized through density plots while the other categorical predictors are visualized through bar graphs. Each variable is classified and colored by the binary response of ten-year risk of CHD in order to visualize the distribution of a variable in the two

levels of the response variable. The age of patients who have the ten-year risk of CHD is distinctly higher than those without the risk (***Figure 2***). The systolic and diastolic blood pressures of patients who have the risk are slightly higher than those without the risk (***Figure 2***). Patients with male sex, taking blood pressure medication, having a previous stroke, having a hypertensive status, and having diabetes have higher proportions of having ten-year risk of CHD compared to those without these conditions (***Figure 3***). Hence, based on the exploratory analysis, the highly possible determinants of CHD risk include the male sex, higher age, and being with history illnesses such as blood pressure diseases, stroke, and diabetes.

In addition, the correlation map indicates highly correlated relationships (over 50%) between binary variable – being a current smoker and continuous variable – cigarettes per day, between binary variable – having diabetes and continuous variable – glucose level, and among binary variable – having prevalent condition of hypertension and continuous variables – systolic and diastolic blood pressures (***Figure 1***). The reason to this high correlation is that the binary variables can be greatly explained by these continuous variables. Hence, in order to avoid multicollinearity in the later model analysis, these highly correlated binary variables are removed as the continuous ones have higher statistical power.

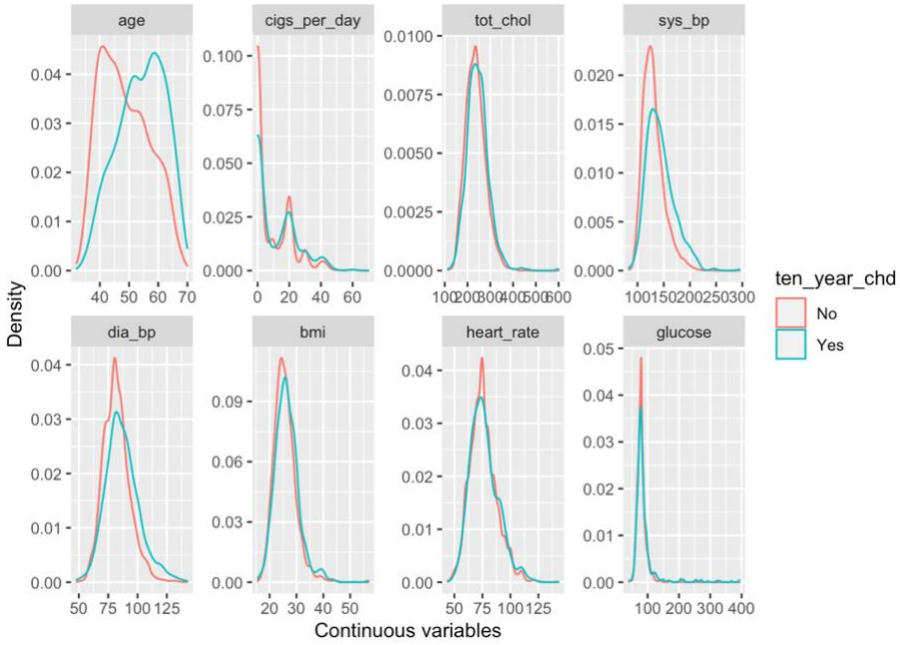

Figure 2. Density plots of continuous predictors by the response variable – ten-year risk of CHD.

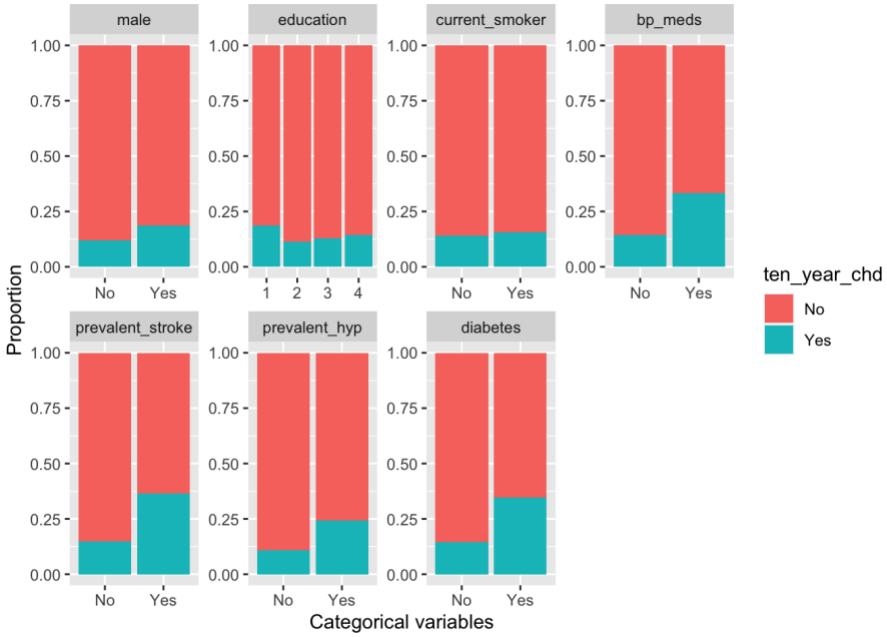

Figure 3. Bar graphs of categorical predictors by the response variable – ten-year risk of CHD.

Models

A generalized linear model with logistic regression is firstly constructed with all the available data and the 15 predictors to determine the most significant predictors. The result shows that there are five variables with p-value smaller than 0.01, including male sex, age, cigarettes per day, diastolic blood pressure, and glucose level. In other words, these five variables are statistically significant determinants in predicting the ten-year risk of CHD from the logistic GLM. Hence, the following model constructions only focus on these five variables to ensure parsimony.

Five classification models, including generalized linear model, generalized addictive model, multivariate adaptive regression splines model, k-nearest neighbors model, and latent Dirichlet allocation model, are then built with cross-validation and compared to find out the best one in predicting the ten-year risk of CHD. The whole data set is randomly divided into two parts, 75% as training data and the remaining 25% as testing data. The models are fitted based on the training data and then their prediction accuracies are tested based on the testing data.

The GLM is built by assuming independence between predictors and linear relationship between the predictors and the response variables. The independence assumption is greatly achieved by removing those highly correlated predictors in advance. For the other four models, since they use non-parametric algorithm, no strong assumptions are needed.

The tuning parameter of the fitted MARS model is automatically selected by the “caret” training method as 1 product degree and 5 terms to minimize prediction error. In addition, the MARS model gives a rank of the important roles in predicting the response: age, diastolic blood pressure, glucose, cigarettes per day, and male sex.

The tuning parameter of the fitted KNN model is also automatically selected to be 185. The limitation to the KNN model is a trade-off between bias and variance in the estimators. Due to the big tuning parameter, although the variance is small in the resulting estimators, there is a great bias existed.

Based on the ROC curve plot, GAM has the best prediction performance when applying on the testing data. All five models have very close ROC values around 70% which are generally good, and GAM has the highest ROC – 70.8%, indicating a relatively higher accuracy (***Figure 4***). The result of GAM indicates that the coefficient of the parametric term – male sex is 0.32 and this term is significant since the p-value is less than 0.05 (***Table 1***). Based on the plots of the smooth terms, cigarettes per day and age are linear and significant, while glucose level and diastolic blood pressure are quadratic and significant (***Figure 5***). The limitation to the GAM is a propensity of overfitting.

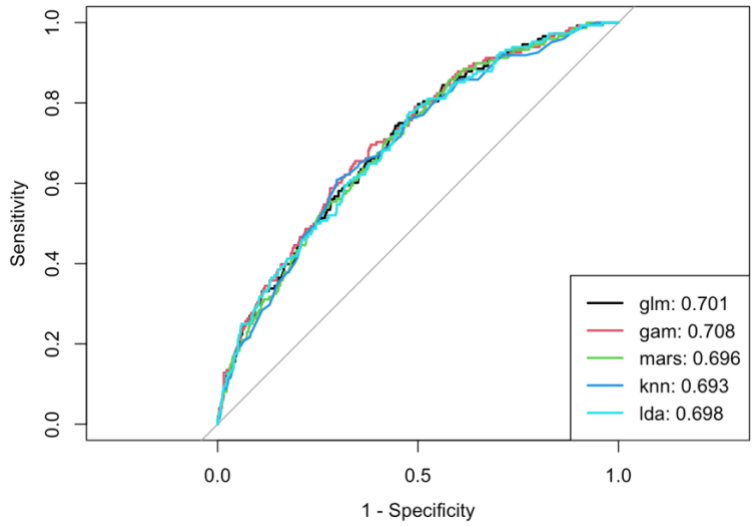

Figure 4. ROC curve for the five models.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.131	0.084	-25.385	0.000
male1	0.321	0.115	2.802	0.005

Table 1. The results of parametric term in generalized addictive model.

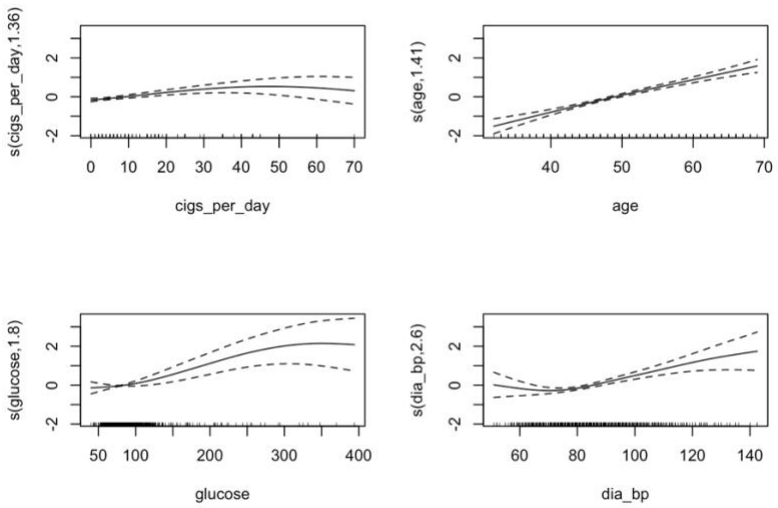

Figure 5. Plots of smooth terms in GAM.

Conclusion

Based on the model analysis, the significant determinants in predicting whether the patient has ten-year risk of CHD are age, diastolic blood pressure, glucose, cigarettes per day, and male sex with descending importance. The result is similar to the expected result based on the exploratory analysis except that the model analysis points out cigarettes per day as another significant determinants. The public health policy for intervening coronary heart disease can be designed by concentrating on men, elder people, smokers, and people with history illnesses such as blood pressure diseases, stroke, and diabetes.

All five models indeed have similar prediction performance, though the generalized addictive model seems to be the best one with the highest prediction accuracy. Hence, GAM is recommended for statisticians to predict the ten-year risk of CHD, but the other classification methods can also be used for more information, such as the rank of important variables from MARS model.