P8106 Final Project Report

Jinghan Liu (jl6048), Yunlin Zhou (yz4184), Jiayao Sun(js5962)

Introduction

The project researches **how to predict Cardiovascular disease through various body function indexes**. By analyzing the distribution of attribute information and building the model, the research aims at predicting the effect of each risk factor on the probability of having heart disease.

Motivation

According to the Heart disease facts on the WHO website, Cardiovascular disease (CVD) is the number one cause of death globally, with an estimated **17.9 million** deaths each year, accounting for **31%** of global mortality. What's more, heart failure is a **leading cause** of CVD. Therefore, predicting the risk factors for developing heart failure is essential. The purpose of the study is to analyze the patient dataset and build an optimal model to predict the likelihood of heart failure in a patient based on predictors such as age, gender, ChestPainType, resting blood pressure, serum cholesterol, etc.

Data Cleaning

The dataset contains 918 observations and 12 variables that can be used to predict the characteristics of possible heart disease. The response variable is Heart Disease. The predictors are:

- Age: age of the patient [years]
- Sex: sex of the patient [M: Male, F: Female]
- ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: resting blood pressure [mm Hg]
- Cholesterol: serum cholesterol [mm/dl]
- FastingBS: fasting blood sugar [high: if FastingBS > 120 mg/dl, other: otherwise]
- RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- Oldpeak: oldpeak = ST [Numeric value measured in depression]
- ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- HeartDisease: disease or normal

Table 1: Data summary

Name	Piped data
Number of rows	918
Number of columns	12
Column type frequency:	
character	7
numeric	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
heart_disease	0	1	6	7	0	2	0
sex	0	1	1	1	0	2	0
chest_pain_type	0	1	2	3	0	4	0
fasting_bs	0	1	4	5	0	2	0
resting_ecg	0	1	2	6	0	3	0
exercise_angina	0	1	1	1	0	2	0
st_slope	0	1	2	4	0	3	0

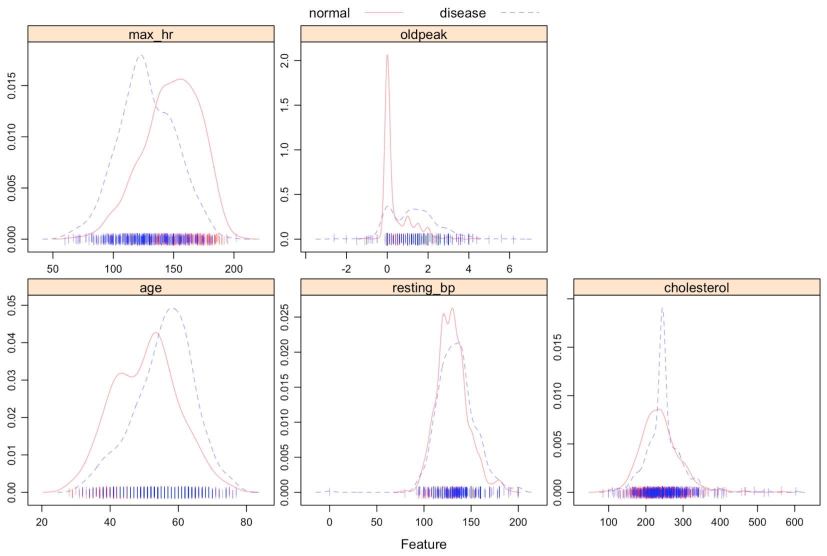
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1.00	53.51	9.43	28.0	47.00	54.0	60.0	77.0	
resting_bp	0	1.00	132.40	18.51	0.0	120.00	130.0	140.0	200.0	
cholesterol	172	0.81	244.64	59.15	85.0	207.25	237.0	275.0	603.0	
max_hr	0	1.00	136.81	25.46	60.0	120.00	138.0	156.0	202.0	
oldpeak	0	1.00	0.89	1.07	-2.6	0.00	0.6	1.5	6.2	

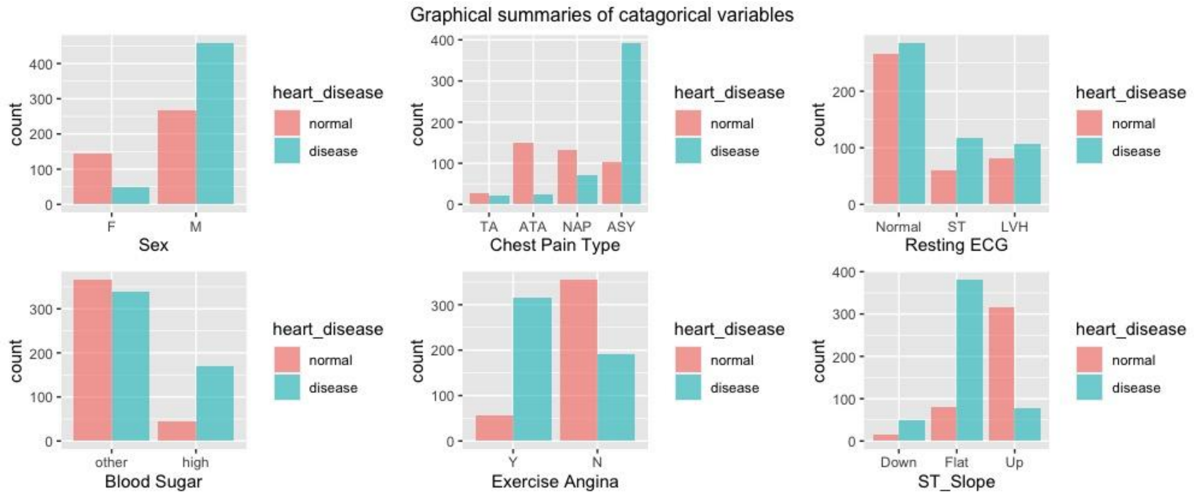
As the table shows above, the data set has 7 categorical variables, and 5 numeric variables, with 918 observations. In the original data set, there were no null observations, but some data on Cholesterol was 0, which is not possible in real life. So the report assumes that those Cholesterol = 0 rows were actually null values when collecting the data. In that case, the report uses the mean value to replace the null observations. For the character variables, function `factor()` is used to change the data type in order to apply the data set to the models. To better use this data set to train the models, the study splits the data set into two parts: training data (70%) and test data (30%).

Exploratory analysis/visualization

From the density plot of continuous variables below, it is obvious that most features have significant differences between the normal and heart-diseased people. The normal people are tending to have higher maximum heart rate; younger people are less likely to have heart disease; normal people have larger chances to have 0 oldpeak; the diseased people's cholesterol are more concentrated between 200 - 300. But for the feature resting_bp, the difference is not significant.



As the plot shows below, male tend to have heart disease. If the patients have Exercise Angina or flat ST slope, they are more likely to have heart disease. However, even if the patient has normal features like no chest pain, normal resting ECG, and blood sugar, they could still have heart disease.



Models

Model Determination

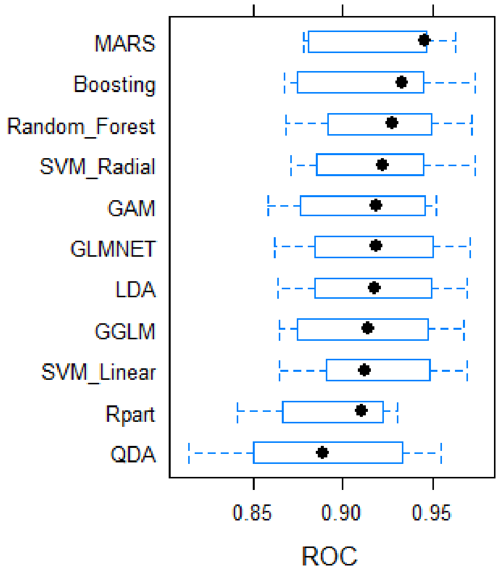
Since the outcome of the study is either having heart disease or not, classification models including logistic regression, penalized logistic regression, GAM, MARS, LDA, QDA, classification tree, random forest, boosting, SVM with linear kernel and radial kernel are built to train the data set, with heart_disease as the response and all other variables in the data set as predictors to fit the model.

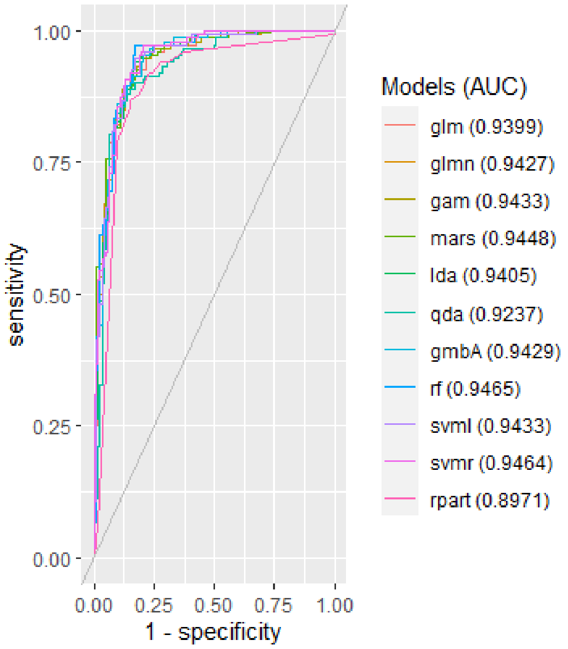
Model Tuning

(1) Penalized Logistic Regression: For penalized logistic regression, the best tuning parameters are alpha = 0.1 and lambda = 0.06105877. (see Figure 1 in the Appendix)
(2) Generalized Additive Model (GAM): The function automatically returned the most optimal model: heart_disease = sexM + chest_pain_typeATA + chest_pain_typeNAP + chest_pain_typeASY + fasting_bshigh + resting_ecgST + resting_ecgLVH + exercise_anginaN + st_slopeFlat + st_slopeUp + s(oldpeak) + s(age) + s(resting_bp) + s(max_hr) + s(cholesterol).
(3) Multivariate Adaptive Regression Splines (MARS): The report tuned degree from 1 to 4 and nprune from 5 to 20. The best model is with nprune = 11 and degree = 1. The best model selected 11 of 19 terms, and 8 of 15 predictors. (see Figure 2 in the Appendix)
(4) Classification Tree Model: the report tuned 50 values of Complexity Parameters from e −8 to e −6. The best model with the smallest cross-validation RMSE is with cp = 0.001862726. (see Figure 7 in the Appendix)
(5) Random Forest Model: All 11 predictor variables are randomly sampled as candidates for the model. The minimum node sizes are 2, 4, 6, 8, 10. The best model with highest ROC in repeated cross-validation is one with minimal node size equaling 6 and randomly selected predictors equaling 1. (see Figure 5 in the Appendix)
(6) Boosting Model: The report uses the Adaboost algorithm for binary classification. The numbers of gradient boosting iterations chosen to be tuned are 2000 and 3000, 4000, 5000, with interaction.depth from 1 to 6 and shrinkage parameter equaling 0.0005, 0.001, 0.002, and n.minobsinnode equaling 1. The model with the highest ROC is the one with boosting iteration = 2000, Max Tree Depth = 3, Shrinkage = 0.002, and the minimum number of observations in trees' terminal nodes = 1. (see Figure 6 in the Appendix)
(8) Support Vector Machine (SVM): The optimal cost parameter for SVM Linear Kernel model was 0.0183 and the optimal SVM Radial Kernel model had a sigma value of 0.013356 and a cost parameter of 3.7937 . (see Figure 3&4 in the Appendix)

Models Comparison

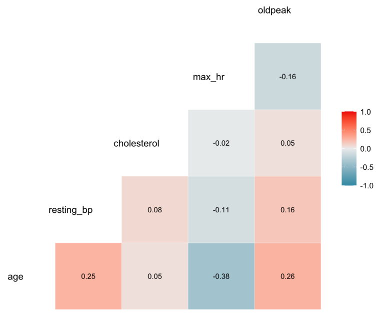
The plots below display the receiver operating characteristic curve (ROC) and area under the ROC curve (AUC). As it shows in the ROC plot that the MARS, Boosting, and Random Forest models all performed well compared to the other models. Also, as shown in the AUC plot, the random forest model has the largest AUC, which suggests that the random forest model is the best fit for our test data set.





For building logistic regression which needs some assumptions: The result is a binary variable; There is a linear relationship between the logit of the outcome and each predictor; No influential value (extreme or outlier) in the continuous predictor; There is no high correlation (i.e. multicollinearity) between the predictors. LDA works when gaussian assumption is violated. The assumption of Random Forest is that there are no formal distributions. An SVM can be defined as a linear classifier under the following two assumptions: The margins should be as large as possible. Support vectors are the most useful data points because they are most likely to be misclassified. The Classification tree also has some of the assumptions: the entire training set is considered the root and the feature values are preferably categorical. If the values are continuous, discretize them before building the model.

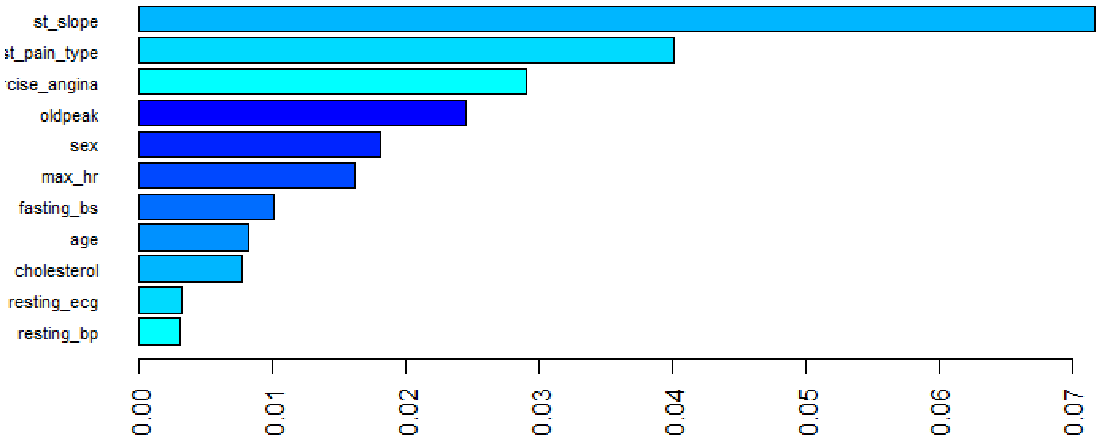
As shown in the Correlation plot below, the report concludes that age and max_hr, as well as age and oldpeak, are relatively highly correlated. To fit a logistic regression model, the predictors must not be correlated. Since age and oldpeak or max_hr are correlated, the result of logistic might be affected. Also, the main limitation of logistic regression is the assumption of linearity between the dependent and independent variables. Overfitting may result if the number of observations is less than the number of features. The tendency to overfit is a limitation of GAM and also the model loses predictability when the value of the smoothing variable is outside the range of the training dataset. Essentially, it sacrifices predictability outside the data range for accuracy within the data range. While MARS has the weaknesses of requiring strict assumptions and dealing with outliers, MARS is not only highly adaptive, but also more accurate in model predictions than some other methods. For the Random Forest model its limitations are that it will overfit particularly noisy datasets and for data containing categorical predictors with different numbers of levels, random forests are biased towards predictors with more levels. Therefore, variable importance scores of random forests are not always reliable for such data. SVM also has some limitations such as its algorithm is not suitable for large datasets; when the dataset has more noise, there will be overlap of target classes; in the case where the number of features per data point exceeds the number of training data samples, SVM will perform poorly. The limitations of the Classification tree are largely unstable and Less effective in predicting the outcome of a continuous variable.



Final Model Interpretation

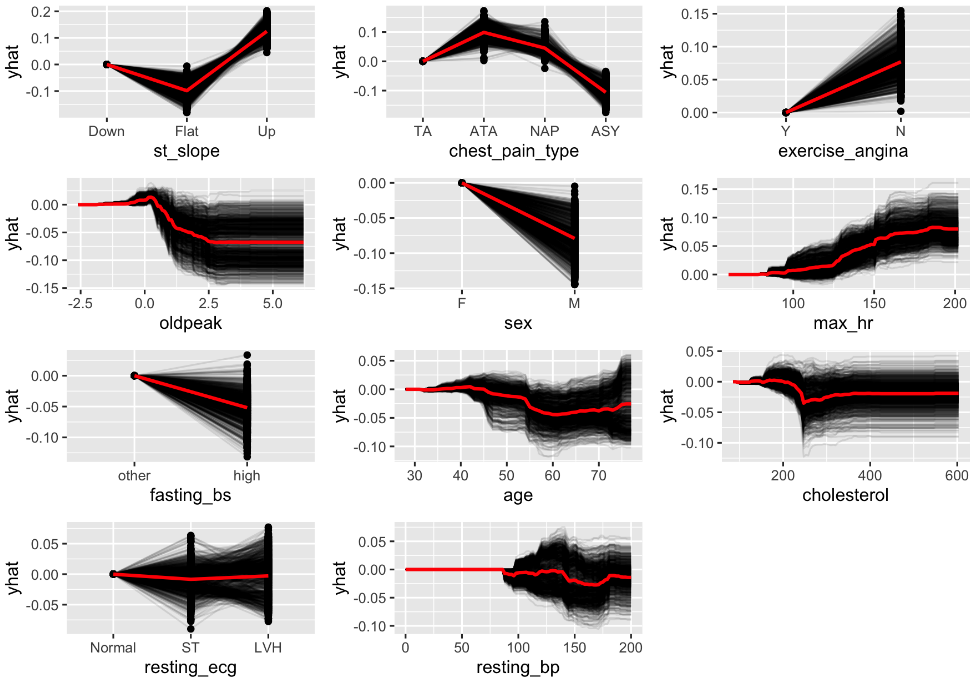
Since random forest is a black box model, a clear function for the response and predictors cannot be generated. However, analyzing the performance, overall importance and partial predictive power could well interpret the random forest model.

For the performance of the classification algorithm, the model confusion matrix based on the test data shows that the overall prediction accuracy is 0.9018. The 95% CI is [0.8604, 0.9343]. The p value is small and close to 0, showing that the fitted model accuracy is better than no information rate. And the no information rate is 0.5527 which means that if there is not any information to predict the heart_disease, the prediction accuracy will be 0.5527 . Kappa equals 0.7992, which means that it is highly related to observation and predictions. The test error rate is 9.82%.



After contrasting the impurity-based importance plot and the permutation importance plot, the study chooses the latter because it could avoid the issue caused by unseen data if the model is overfitted. From the permutation-based importance plot, the importance of 11 predictors is in the sequence: st_slope, chest_pain_type, exercise_angina, oldpeak, sex, max_hr, fasting_bs, age, cholesterol, resting_ecg, resting_bp.

From the centered ICE curves, if the oldpeak, the most relevant numeric variable, is larger than 2.5, the effect of oldpeak on getting heart disease will drop and become steady. Patients with upsloping the peak exercise ST segment are more likely to have heart diseases compared to those with flat slope. ATA chest pain is highly related to heart disease rate. Meanwhile, the heart disease rate increases rapidly when the maximum heart rate exceeds 125.



Conclusion

The report found that in the heart failure prediction dataset, the Random Forest model performed best in the test data set. Given the extreme violation of the normality assumption in the predictors, logistic regression is somewhat expected to outperform LDA. As expected in the report, in this model, the st_slope, st_pain_type, exercise_angina, oldpeak, sex, max_hr, fasting_bs, age, cholesterol, resting_ecg, resting_bp are important to predict the probability of heart failure in this model. It is worth noting that when the type of chest pain is asymptomatic, it has a higher probability of causing heart failure. This is because people do not go to the hospital for examinations when they have no symptoms. This finding has important implications for how to reduce the risk of heart failure. People can reduce risk through regular physical examinations, and the earlier the heart disease is detected, the risk of heart failure can be reduced. In addition, there are also some problems with this model. For example, the gender ratio in this data is very inconsistent due to the fact that the ratio of males and females is very large when the data is collected. There are 79% of male and only

21% of females. Therefore, it is concluded that sex is a significant predictor of heart failure, which requires further expansion of the database and analysis. Second, from the previous literature, it was found that diabetes, ethnicity, obesity, lifestyle, etc. are all important risk factors for heart failure. Therefore, the models used in this report may still be inaccurate in predicting heart failure. Including more patient information with heart failure and more likely risk factors may help improve the predictive performance of the model to more accurately predict risk factors for heart failure.