Predictive Model For Heart Disease Screening Program

Iris Ziyi Wang, Yiming Zhao, Jeffrey Zhuohui Liang

4/7/2021

Introduction

Cardiovascular disease is the leading disease burden in U.S, according to *www.cdc.com* [1], on average one person die from heart disease every 36 seconds. And 1 in 4 death is caused by cardiovascular disease. Heavy disease burden of cardiovascular disease should be manage to improve population health.

One of many important manners is screening, *American Heart Association*[2] lists important screening to help monitor heart condition, which are shown below:

- Blood Pressure
- Fasting Lipoprotein Profile
- Body Weight
- Blood Glucose
- Smoking, physical activity, diet

In light of aiding the screening process, we will use `Heart Disease Data Set` from UCI [3] to build our models and select one for applications.

The `Heart Disease Data` is a dataset with 76 attributes, all data were collected from 4 sites, namely Cleveland, Hungary, Switzerland, and the VA Long Beach. Of all 76 attributes, we selected 14 variables as our training data in this case as there're previously researchs have done similar job and used these 14 pre-selected variables. The predictors used are:

- age: The person's age in years
- sex: The person's sex (1 = male, 0 = female)
- cp: chest pain type — Value 0: asymptomatic — Value 1: atypical angina — Value 2: non-anginal pain — Value 3: typical angina
- trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
- chol: The person's cholesterol measurement in mg/dl
- fbs: The person's fasting blood sugar ($>$ 120 mg/dl, 1 = true; 0 = false)
- restecg: resting electrocardiographic results — Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria — Value 1: normal — Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV)
- thalach: The person's maximum heart rate achieved
- exang: Exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
- slope: the slope of the peak exercise ST segment — 0: downsloping;

  - 1: flat;
  - 2: upsloping

- ca: The number of major vessels (0–3)
- thal: A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously
  - Value 1: fixed defect (no blood flow in some part of the heart)
  - Value 2: normal blood flow
  - Value 3: reversible defect (a blood flow is observed but it is not normal)
- target: Heart disease (1 = no, 0= yes)

Exploratory Analysis

From Fig.1, some features are well distinguish for disease status, eg. `num_major_vessels_flouro`, `chest_pain_typeatypical angina`,`st_depression_exercise`, these variables may be statistical significant for the model.

Model

As shown in Table.1, Table.2, there are missing values in our data. Assuming that these values are missing at random, we impute these values with `knnImpute` method. All data were center and scale before training.

To train classifiers, we choose `Elastic Net logistics`, `MARS`, `KNN`, `LDA`, `QDA` and `TREE`, `Random Forest(Bagging)`, `Boost`, `Support Vector Machine` models to train our data with 5-fold cross validation.

When training, `ROC` is used as loss function for our model, as we intent to build a model with highest classification ability to predict whether a client has heart disease.

Model tunning

Since variables in the data set were selected as important influence factors of heart disease, all of 13 variables were included in the models as predictors. Because the KNN method s applicable in any of Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) assumptions, the missing value would be imputed by knn method.

In order to select best model to predictor heart disease status, this report compares the ROC value of logistic regression, MARS,KNN, LDA, QDA, SVM and tree, random forest, boosting and support vector machine models. These models are flexible enough to capture the underlying truth by adjusting appropriate tuning parameters. The train function was used to select the optimal tuning parameters by 5-fold cross validation for elastic net logistics regression, Mars, KNN, SVM and tree models by choosing the tuning parameter with highest ROC in ROC plot.

For KNN, although the missing values are imputed before applying KNN model to train data, major proportion of lossing the value of peak Exercise ST Segment,Number of visible vessels may still degrade the model performance. Moreover, the heart disease training data has 691 observations. Because KNN is a distance-based algorithm, relies on local neighbor's distance between a new point and each existing point, which might degrade the performance of the algorithm on a global perspective[5].

Elastic net logistics regression is logistics regression which loss function is modified with L1 penalty and L2 penalty, we tune this penalty term λ and the elastic net parameter α for regression model training with cross-validation.

MARS has model predictors' order and prune remaining term as parameters for tuning. Assuming that data can be well-explain with at most cubic model, we tune the order from 1-3 and leaving cross validation to choose for prune term.

Tree model and the ensemble form of tree's family were consider in the model. For single tree model, we tune the complexity of the tree, the model was selected with "1-se" rule to have the simplest working model. Random Forest(and Bagging) ensemble trees with randomly sampling variable in each split into a final tree. Model tunes with number of predictors sampled at each split, splitting rules and minimum node size. Boosting use sequential addictive tree models and ensemble into a tree. Model tuning with number of trees and learning rates.

Support Vectors Machines fit boundary on the support vectors, both linear kernal and radial kernal were consider in the model.

All models are listed in Table.3.

Result

In our trained model have similar ROC performance excepted for `TREE` and `qda`(Table.3). Considering our model is used for improving screening process, we would prefer model with higher sensitivity. Considering both metrics, MARS method which has high ROC and highest mean sensitivity is chosen as our model.

In the MARS model, `chest pain type: atypical angina` has the highest importance, followed by `serum cholesterol` and `st depression excercise`. `fasting blood sugar` has second lowest importance to AUC loss in the MARS model followed by `resting ECG: ST-T wave abnormality`.

The models, included others not selected model is test against the test data. The test performance is similar to the train performance. Which `knn` has the highest ROC, but is similar or not significantly different from other methods except for tree, which has the poorest performance.

Conclusion

MARS model has high predictability and high sensitivity, which is suitable for screening. The MARS model, with it's nature of spline predictor, also provide good reference for critical values for labs/ testing result for diagnosis.

The first implement of the model is to predict/diagnosis heart disease for the screening visitors.

The Second implement of the model is to selected high importance predictors as screening items. For example, `chest pain type` has the most 2 important preditors in the model, as well as `serum cholesterol` and `excercise`, which should included in the screening process. Although the important score plot shows that vessels numbers are very important variables in predicting heart disease, the validity of this result is questionable because of the large missing values of vessel number in the data. Moreover, `fast boold sugar` and `resting ECG: ST-T wave abnormality` has minor importance to the model, which can consider removing for economic package. For example, in the listed recommended screening terms of [2], `Blood Glucose` can be consider remove from screening, as it has little use for our model, but it may has clinical usage which is not our consideration in the projects.

From the important score plot, age(28-77) is not a important variable in predicting the response. This indicates that heart disease may not have age preference. This result is counterintuitive since people usually think that the risk of heart disease would be increased with age. Our analysis result implies that young people also have risk of heart disease as same as old people. It is corresponding with a multi-state study that investigated 28,000 people hospitalized for heart attacks from 1995 to 2014[4]. The research claims that 30 percent of those patients were young, age 35 to 54. Therefore, it is also important for young people to take care of their heart problems.

One of potential limitation for the analysis result is that the raw data was created by combining 4 different location heart disease data. Further more investigation may be necessary to make sure the generalizability of heart disease result in different locations.

References

[1] Centers for Disease Control and Prevention: https://www.cdc.gov/heartdisease/facts.htm#:~:text=Heart%20Disease%20in%20the%20United%20States&text=One%20person%20dies%20every%2036,1%20in%20every%204%20deaths.&text=Heart%20disease%20costs%20the%20United,year%20from%202014%20to%202015.

[2] American Heart Association:Heart-Health Screenings:

[3] USI: http://archive.ics.uci.edu/ml/datasets/heart+Disease

[4] Twenty Year Trends and Sex Differences in Young Adults Hospitalized With Acute Myocardial Infarction

[5] KNN: Failure cases, Limitations, and Strategy to Pick the Right

Appendix

Table 1: Fig.1 Data Discription(Factors)

variable	complete rate	fators count
diagnosis_heart_disease	1.000	pre: 509, abs: 411
location	1.000	cle: 303, hun: 294, va: 200, swi: 123
sex	1.000	mal: 726, fem: 194
chest_pain_type	1.000	asy: 496, non: 204, aty: 174, typ: 46
fasting_blood_sugar	0.902	fas: 692, fas: 138
resting_ecg	0.998	nor: 551, lef: 188, ST-: 179
exercise_induced_angina	0.940	no: 528, yes: 337
peak_exercise_st_segment	0.664	Fla: 345, Up-: 203, Dow: 63

Table 2: Data Discription(Numeric)

variable	complete rate	p25	mean	p75
thalassemia	0.472	NA	NA	NA
age	1.000	47	53.511	60.0
resting_blood_pressure	0.936	120	132.132	140.0
serum_cholesterol	0.967	175	199.130	268.0
max_heart_rate_achieved	0.940	120	137.546	157.0
st_depression_exercise	0.933	0	0.879	1.5

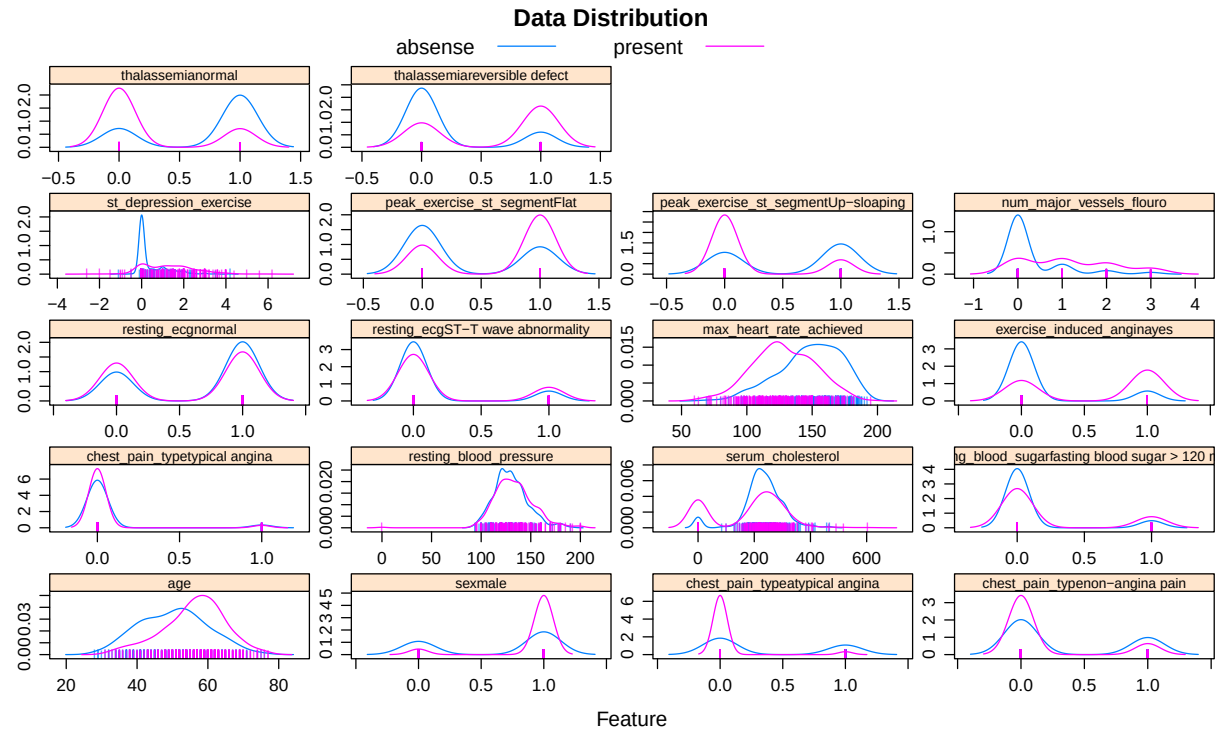


Fig.2 Model Tuning(Part.1)

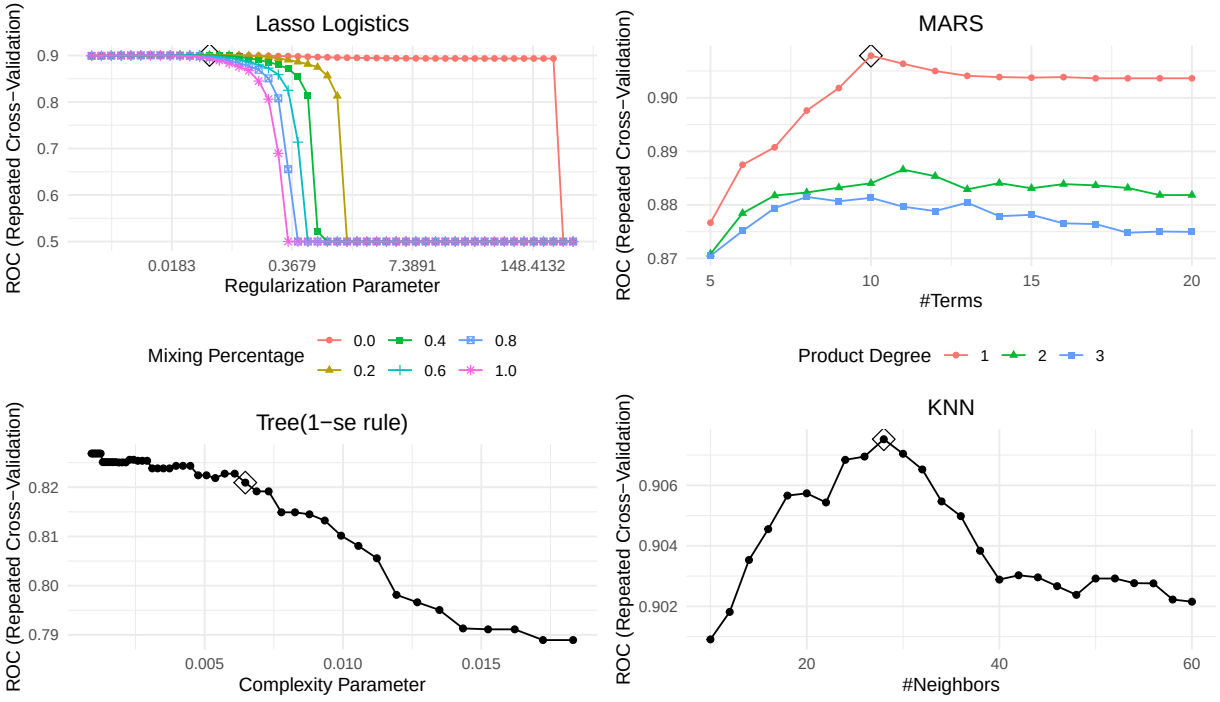


Fig.3 Model Tuning(Part.2)

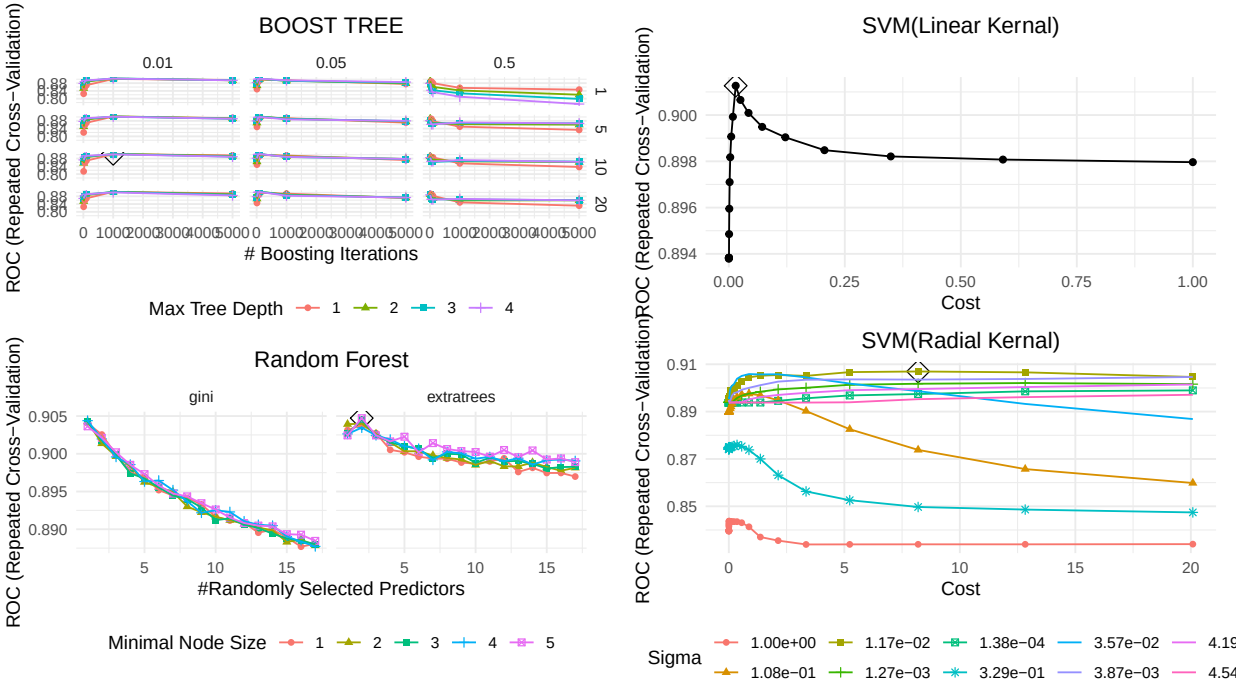
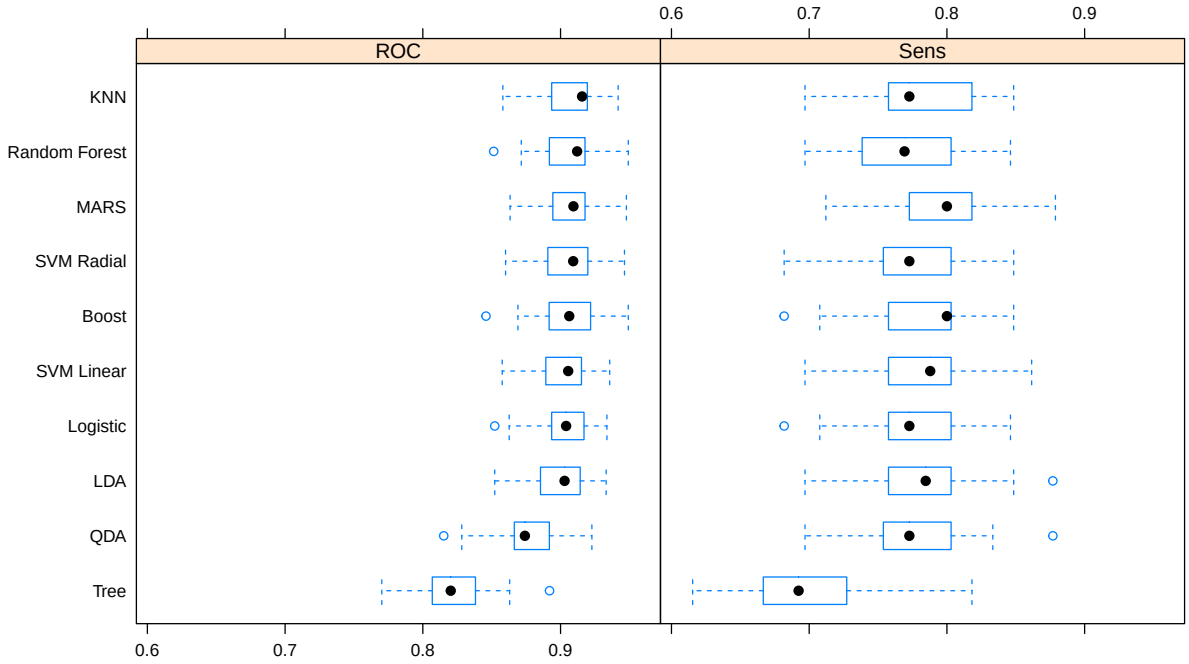


Table 3: Model and train data resampling performance

Model	Metric	p25	Mean	p75
Boost	ROC	0.892	0.904	0.922
Boost	Sens	0.758	0.784	0.803
KNN	ROC	0.894	0.908	0.919
KNN	Sens	0.758	0.781	0.818
LDA	ROC	0.885	0.900	0.914
LDA	Sens	0.758	0.783	0.803
Logistic	ROC	0.894	0.901	0.917
Logistic	Sens	0.758	0.775	0.803
MARS	ROC	0.894	0.908	0.918
MARS	Sens	0.773	0.795	0.818
QDA	ROC	0.866	0.875	0.892
QDA	Sens	0.754	0.779	0.803
Random Forest	ROC	0.892	0.905	0.918
Random Forest	Sens	0.738	0.765	0.803
SVM Linear	ROC	0.889	0.901	0.915
SVM Linear	Sens	0.758	0.781	0.803
SVM Radial	ROC	0.891	0.907	0.920
SVM Radial	Sens	0.754	0.771	0.803
Tree	ROC	0.807	0.821	0.838
Tree	Sens	0.667	0.701	0.727

Fig.4 Resampling Performance



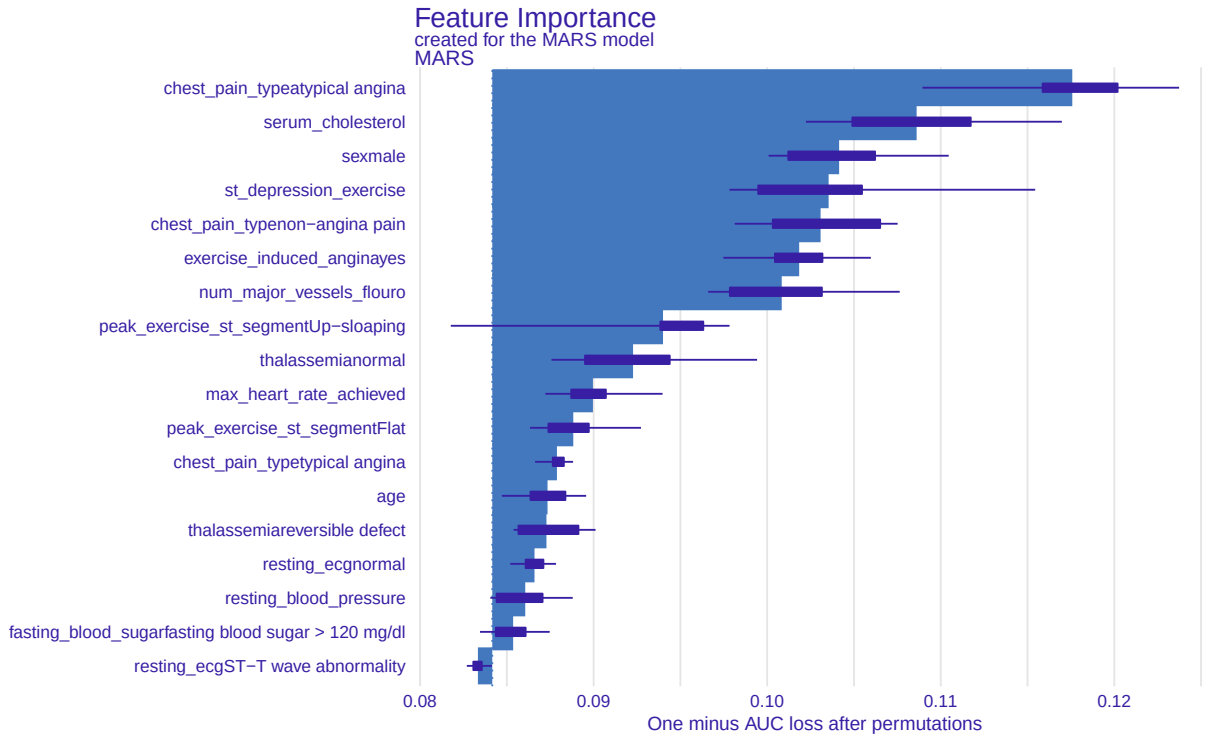


Table 4: Models and Test Dataset's AUC

model	AUC
Elastic Net Logistic	0.872
MARS	0.878

model	AUC
KNN	0.875
LDA	0.874
QDA	0.862
Tree	0.830
Boost	0.886
Random Forest	0.878
SVM Linear	0.872
SVM Radial	0.868

Fig.5 Models Test Dataset Performacne