???????????

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 that

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

are important screening that help monitor heart condition.

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 above plot, 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, 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` 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

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 has tree complexity for tuning.

KNN has the number of closest neighbor as tuning parameter.

LDA and QDA do not have tuning parameters.

All parameter is tune by 5-fold cross validation and choose the one with highest ROC.

Result

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.

Reference

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

Appendix

Table 1: Data summary

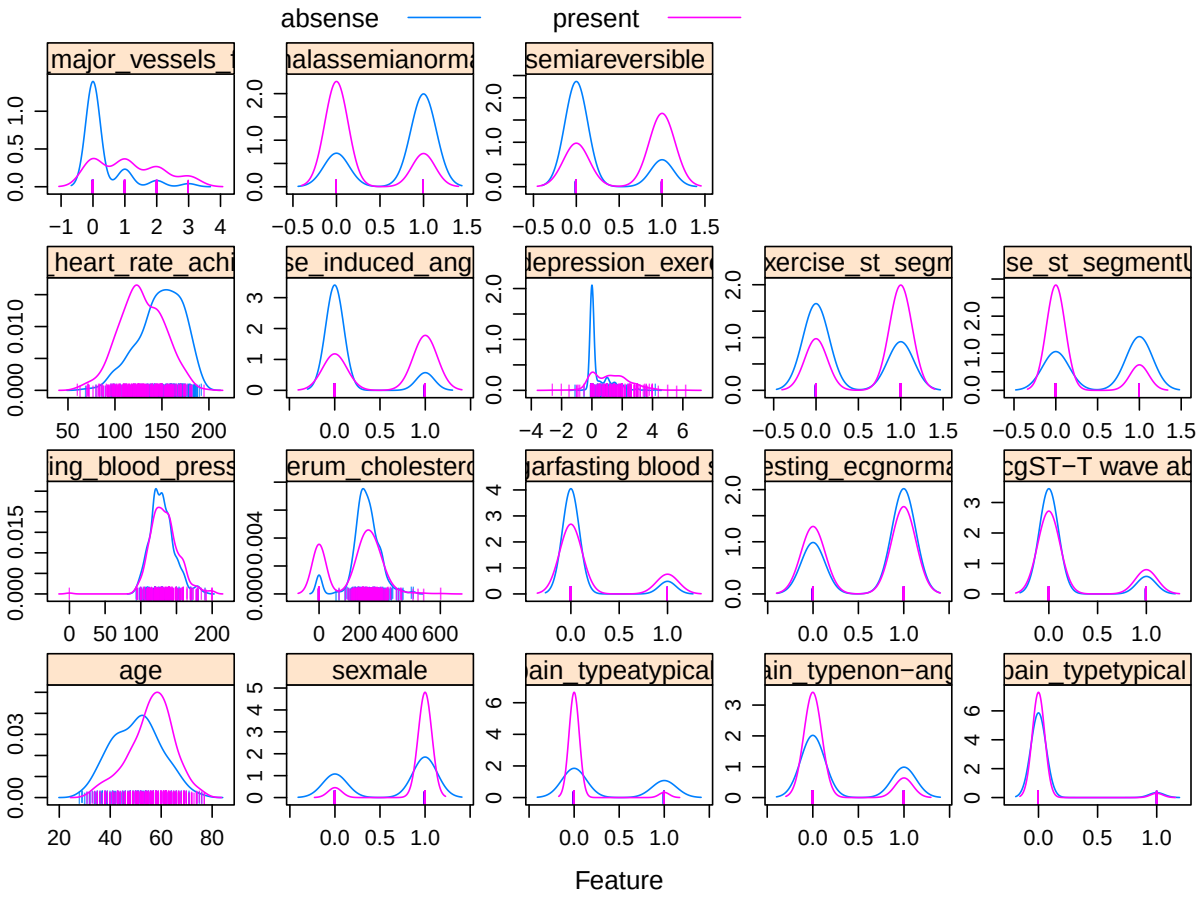
Name	hrt_data
Number of rows	920
Number of columns	15
Column type frequency:	
factor	9
numeric	6
Group variables	None

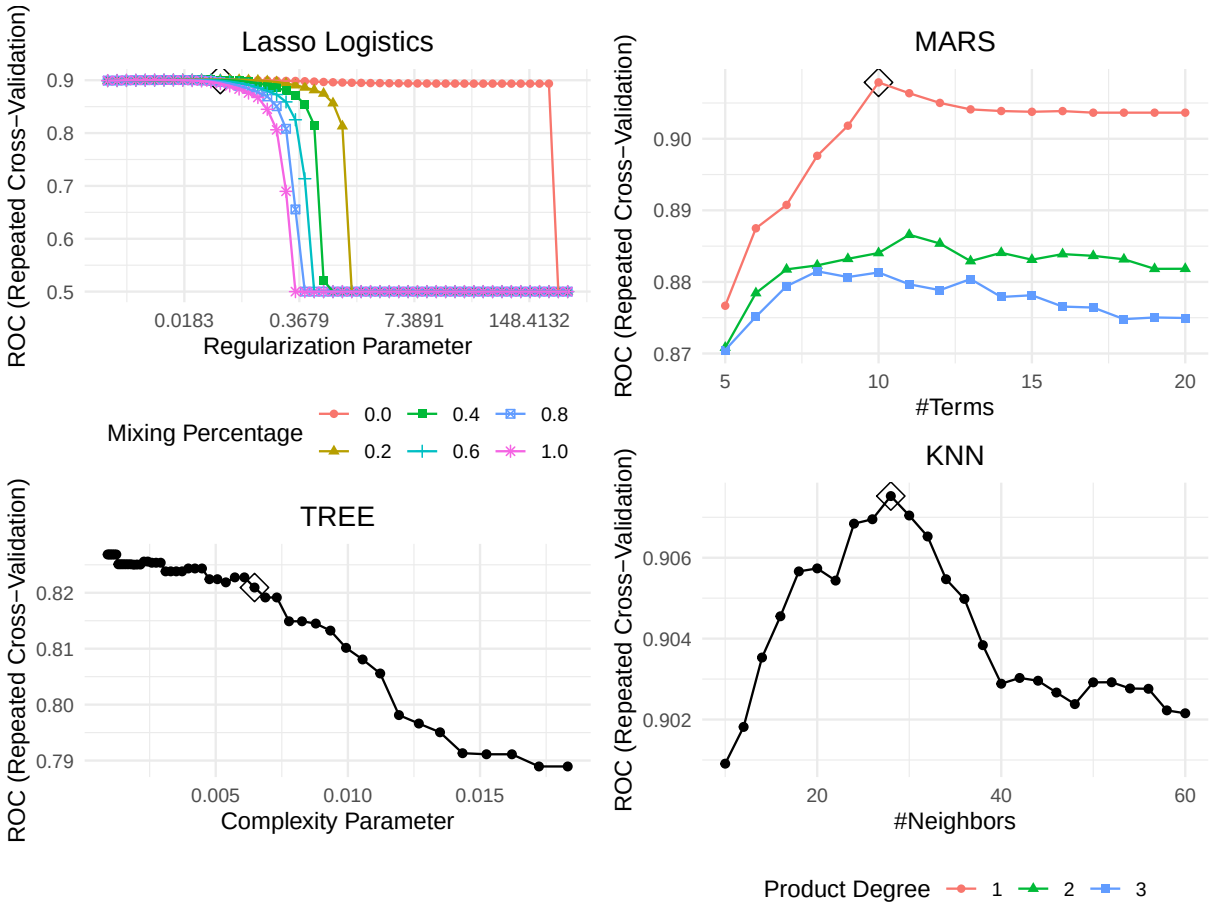
Variable type: factor

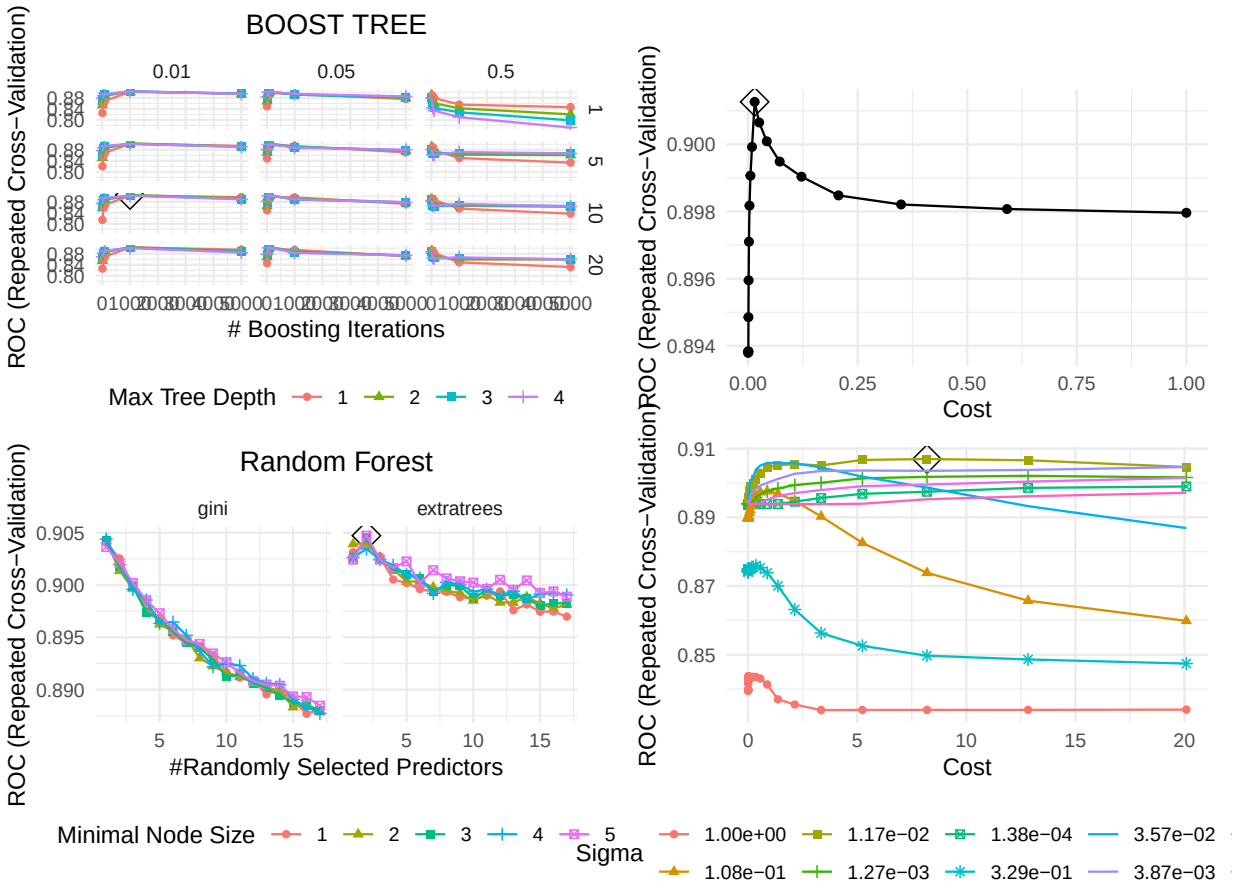
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
diagnosis_heart_disease	0	1.00	FALSE	2	pre: 509, abs: 411
location	0	1.00	FALSE	4	cle: 303, hun: 294, va: 200, swi: 123
sex	0	1.00	FALSE	2	mal: 726, fem: 194
chest_pain_type	0	1.00	FALSE	4	asy: 496, non: 204, aty: 174, typ: 46
fasting_blood_sugar	90	0.90	FALSE	2	fas: 692, fas: 138
resting_ecg	2	1.00	FALSE	3	nor: 551, lef: 188, ST-: 179
exercise_induced_angina	55	0.94	FALSE	2	no: 528, yes: 337
peak_exercise_st_segment	309	0.66	FALSE	3	Fla: 345, Up-: 203, Dow: 63
thalassemia	486	0.47	FALSE	3	nor: 196, rev: 192, fix: 46

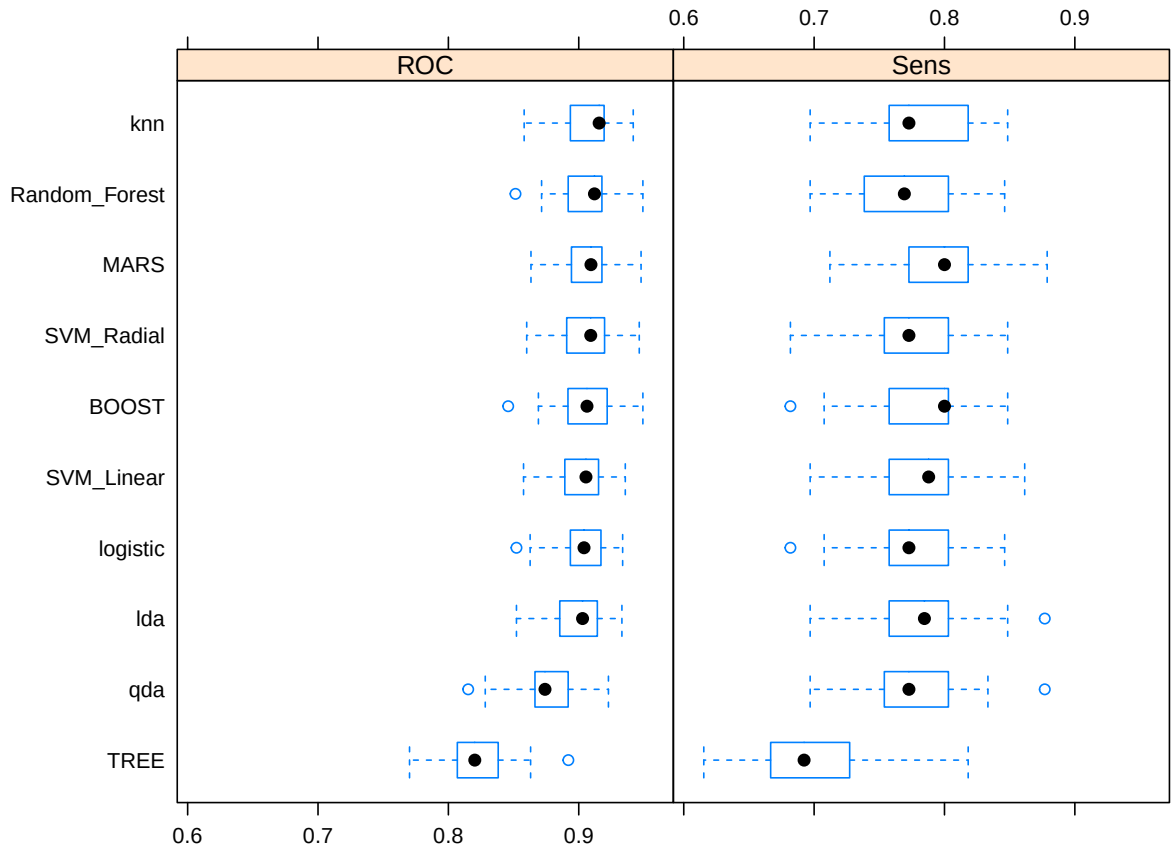
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1.00	53.51	9.42	28.0	47	54.0	60.0	77.0
resting_blood_pressure	59	0.94	132.13	19.07	0.0	120	130.0	140.0	200.0
serum_cholesterol	30	0.97	199.13	110.78	0.0	175	223.0	268.0	603.0
max_heart_rate_achieved	55	0.94	137.55	25.93	60.0	120	140.0	157.0	202.0
st_depression_exercise	62	0.93	0.88	1.09	-2.6	0	0.5	1.5	6.2
num_major_vessels_flouro	611	0.34	0.68	0.94	0.0	0	0.0	1.0	3.0

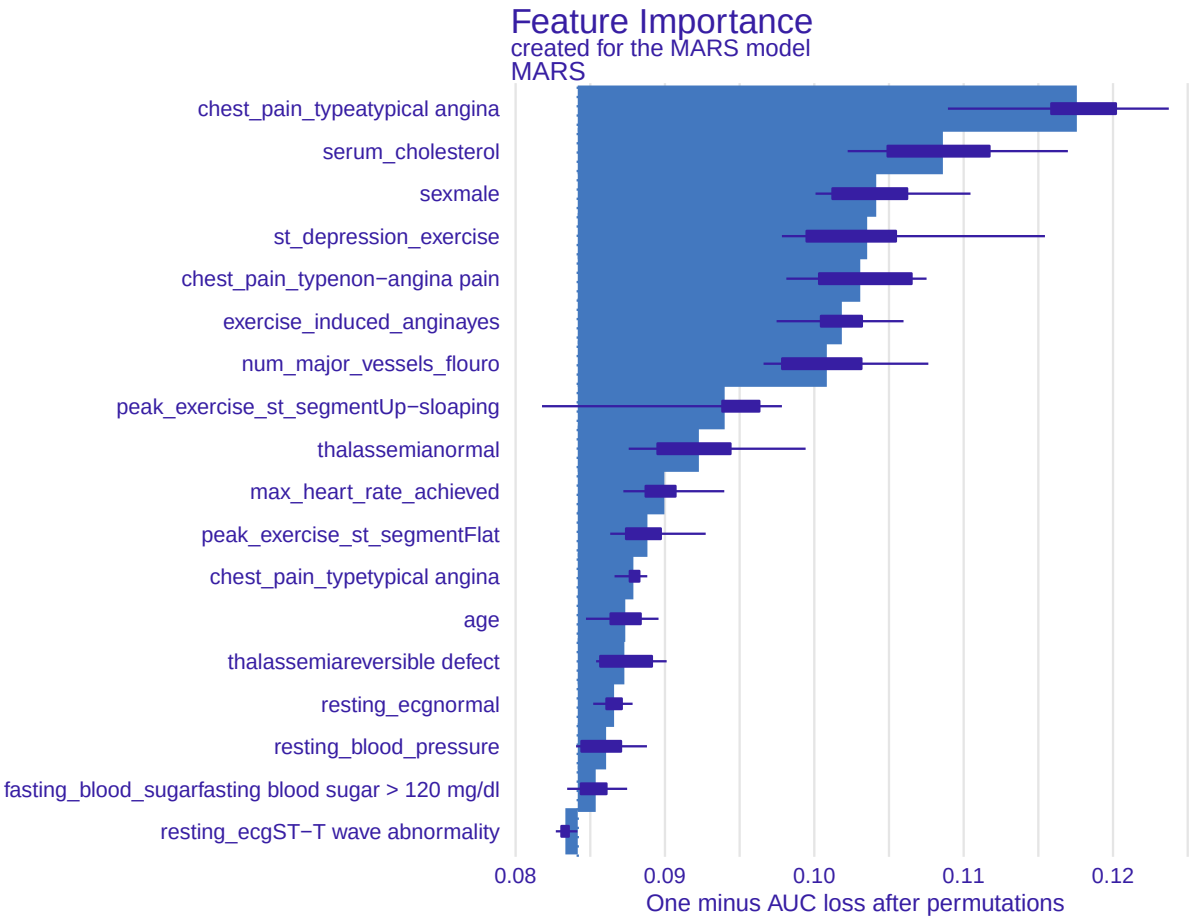

absense
present
major_vessels_
alassemianorma
semiareversible
heart_rate_achi
se_induced_ang
epression_exer
xercise_st_segm
se_st_segmentU
ing_blood_press
erum_cholesterc
arfasting blood s
esting_ecgnorma
cgST−T wave ab
age
sexmale
ain_typeatypical
ain_typenon−ang
pain_typetypical
Feature

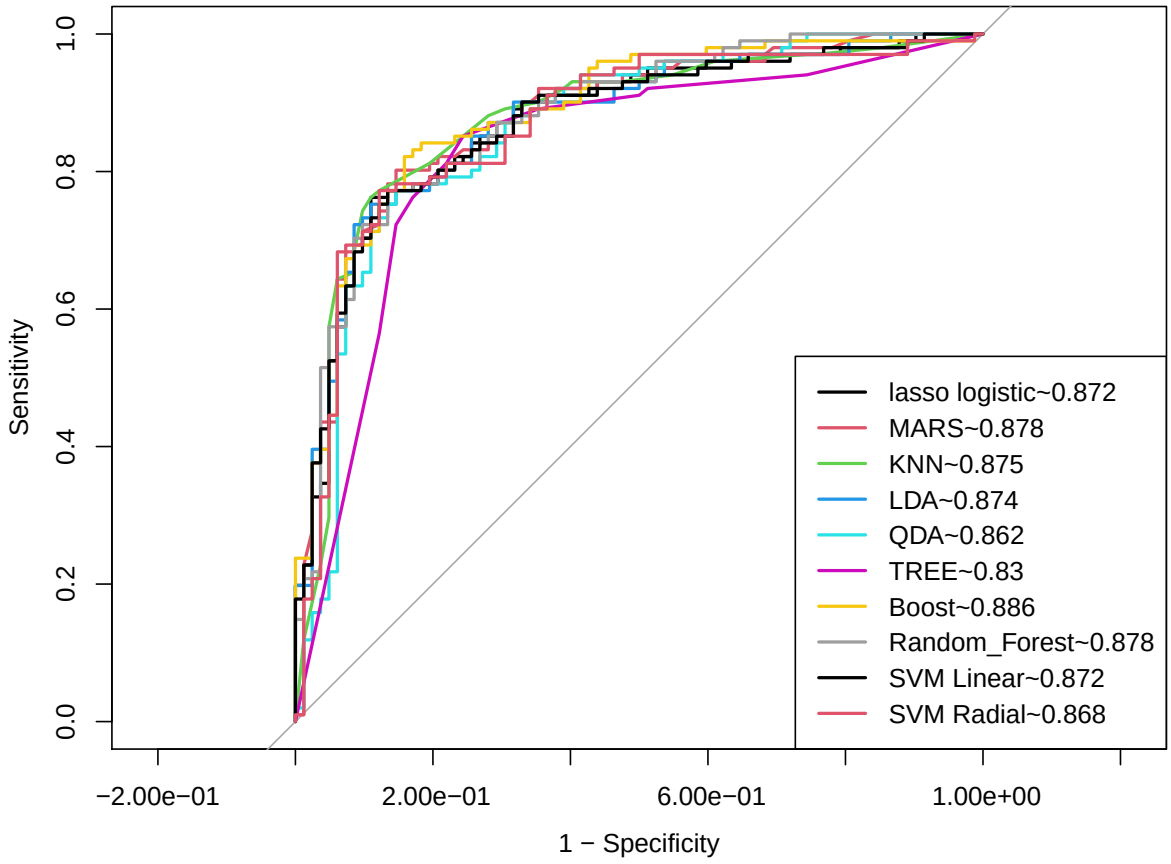

Lasso Logistics
ROC (Repeated Cross−Validation)
0.9
0.8
0.7
0.6
0.5
0.0183
0.3679
7.3891
148.4132
Regularization Parameter
Mixing Percentage
0.0
0.2
0.4
0.6
0.8
1.0
MARS
ROC (Repeated Cross−Validation)
0.90
0.89
0.88
0.87
5
10
15
20
#Terms
TREE
ROC (Repeated Cross−Validation)
0.82
0.81
0.80
0.79
0.005
0.010
0.015
Complexity Parameter
KNN
ROC (Repeated Cross−Validation)
0.906
0.904
0.902
20
40
60
#Neighbors
Product Degree
1
2
3


BOOST TREE
0.01
0.05
0.5
ROC (Repeated Cross−Validation)
Boosting Iterations
Max Tree Depth
1
2
3
4
Random Forest
gini
extratrees
#Randomly Selected Predictors
Minimal Node Size
1
2
3
4
5
ROC (Repeated Cross−Validation)
Cost
Sigma
1.00e+00
1.08e−01
1.17e−02
1.27e−03
1.38e−04
3.29e−01
3.57e−02
3.87e−03



In our trained model have similar ROC peformance excepted for `TREE` and `qda`. 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.


Feature Importance
created for the MARS model
MARS
chest_pain_typeatypical angina
serum_cholesterol
sexmale
st_depression_exercise
chest_pain_typenon−angina pain
exercise_induced_anginayes
num_major_vessels_flouro
peak_exercise_st_segmentUp−sloaping
thalassemianormal
max_heart_rate_achieved
peak_exercise_st_segmentFlat
chest_pain_typetypical angina
age
thalassemiareversible defect
resting_ecgnormal
resting_blood_pressure
fasting_blood_sugarfasting blood sugar > 120 mg/dl
resting_ecgST−T wave abnormality
0.08
0.09
0.10
0.11
0.12
One minus AUC loss after permutations


Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
−2.00e−01
2.00e−01
6.00e−01
1.00e+00
lasso logistic~0.872
MARS~0.878
KNN~0.875
LDA~0.874
QDA~0.862
TREE~0.83
Boost~0.886
Random_Forest~0.878
SVM Linear~0.872
SVM Radial~0.868