???????????

Iris Ziyi Wang, Yiming Zhao, Jeffrey Zhuohui Liang

4/7/2021

Introduction

Exploratory Analysis

Model

Result

Refference

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

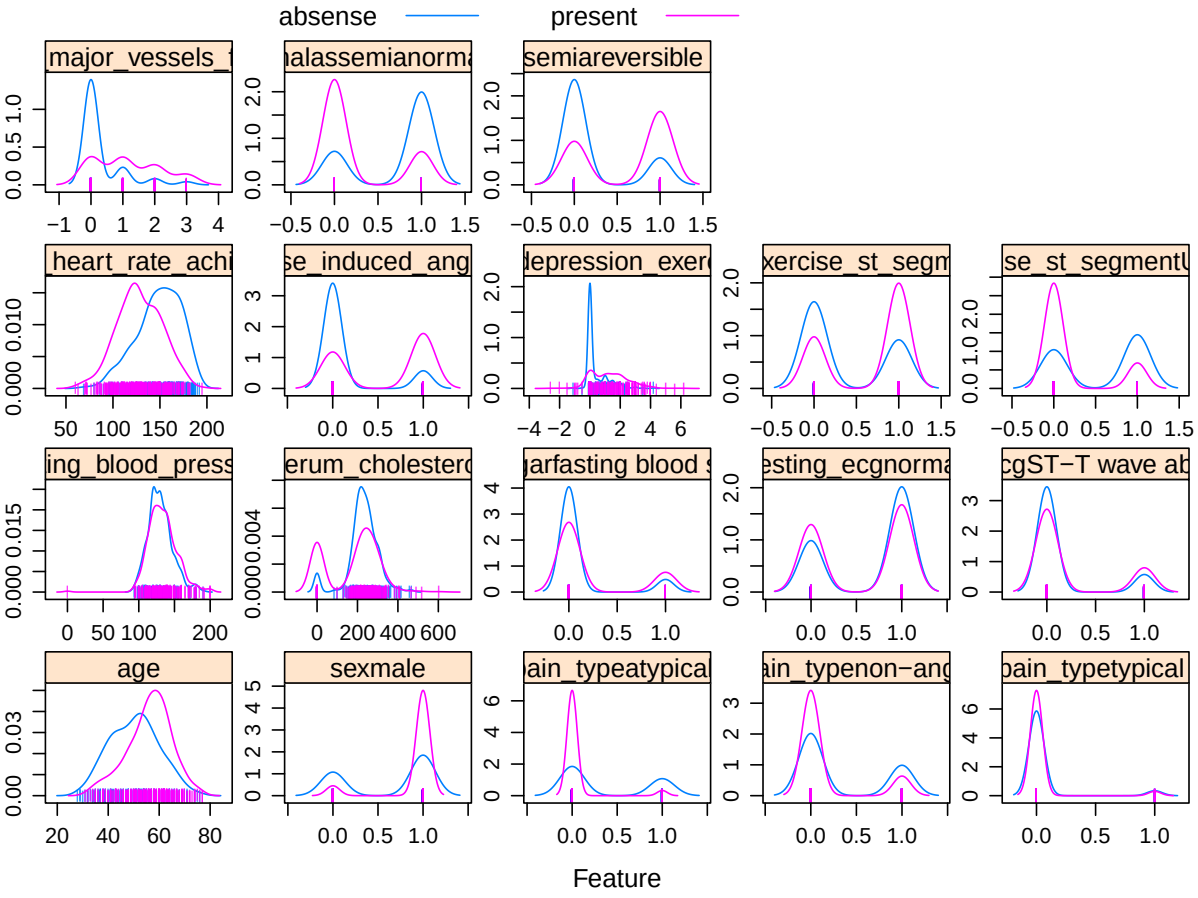
Variable type: factor

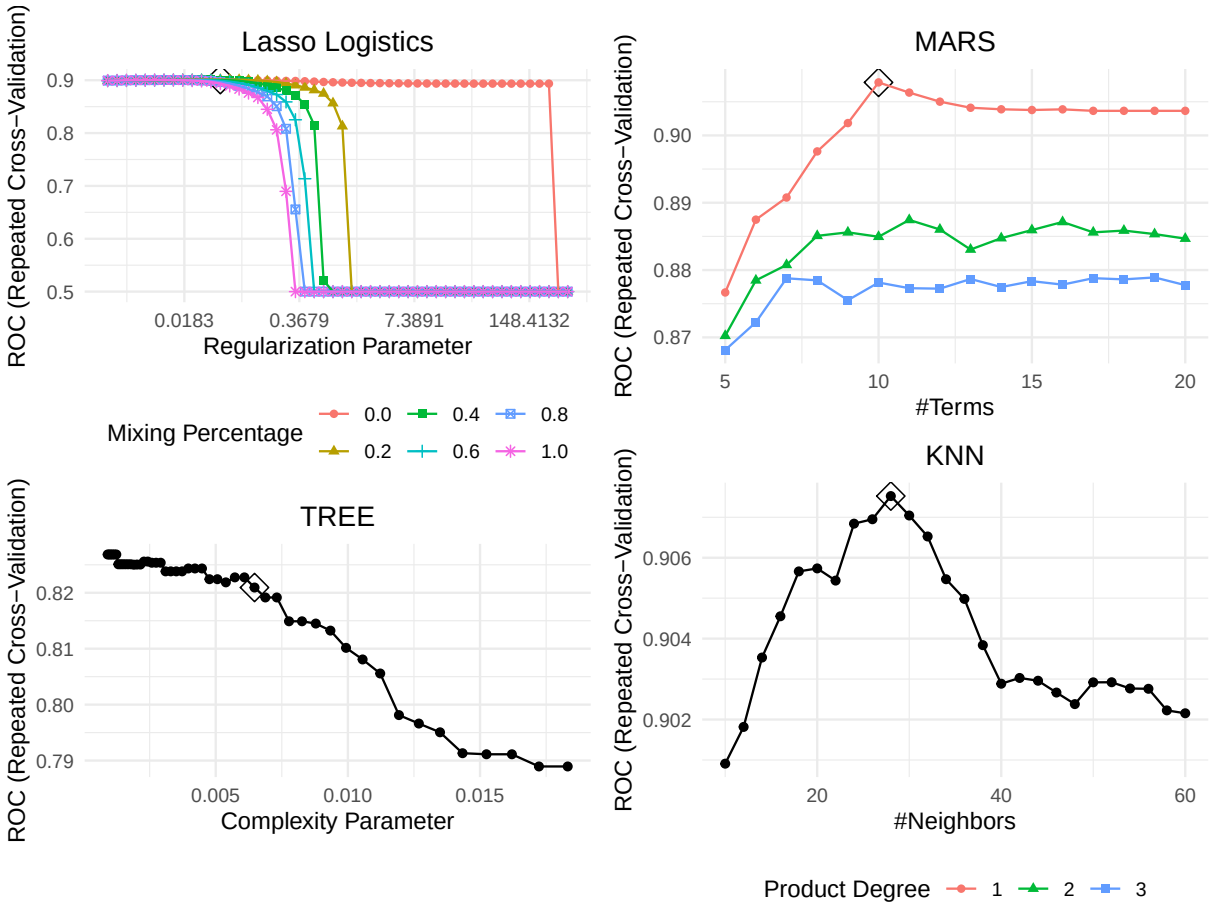
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
diagnosis_heart_disease	0	1.00	FALSE	2	pre: 509, abs: 411
location	0	1.00	FALSE	4	cle: 303, hun: 294, va: 200, swi: 123
sex	0	1.00	FALSE	2	mal: 726, fem: 194
chest_pain_type	0	1.00	FALSE	4	asy: 496, non: 204, aty: 174, typ: 46
fasting_blood_sugar	90	0.90	FALSE	2	fas: 692, fas: 138
resting_ecg	2	1.00	FALSE	3	nor: 551, lef: 188, ST-: 179

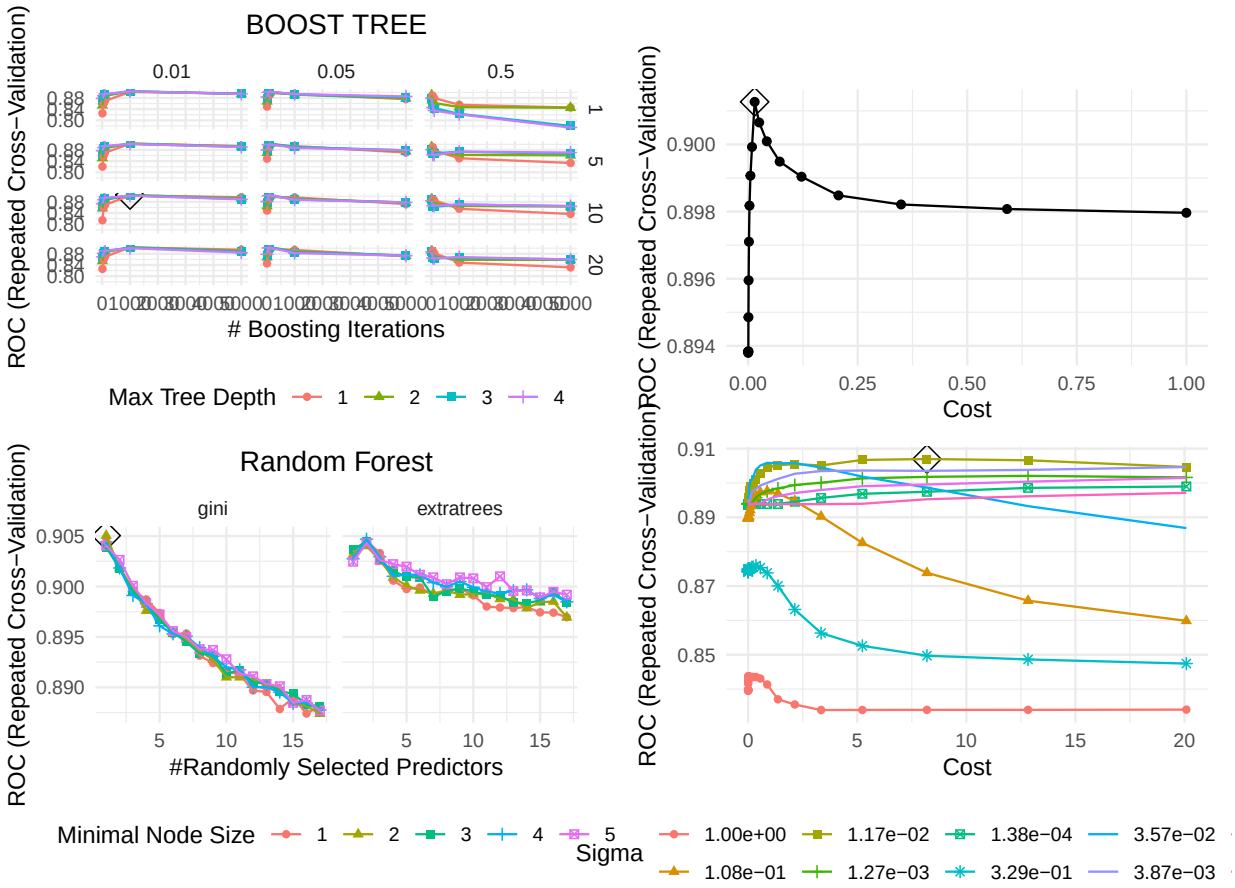
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
exercise_induced_angina	55	0.94	FALSE	2	no: 528, yes: 337
peak_exercise_st_segment	309	0.66	FALSE	3	Fla: 345, Up-: 203, Dow: 63
thalassemia	486	0.47	FALSE	3	nor: 196, rev: 192, fix: 46

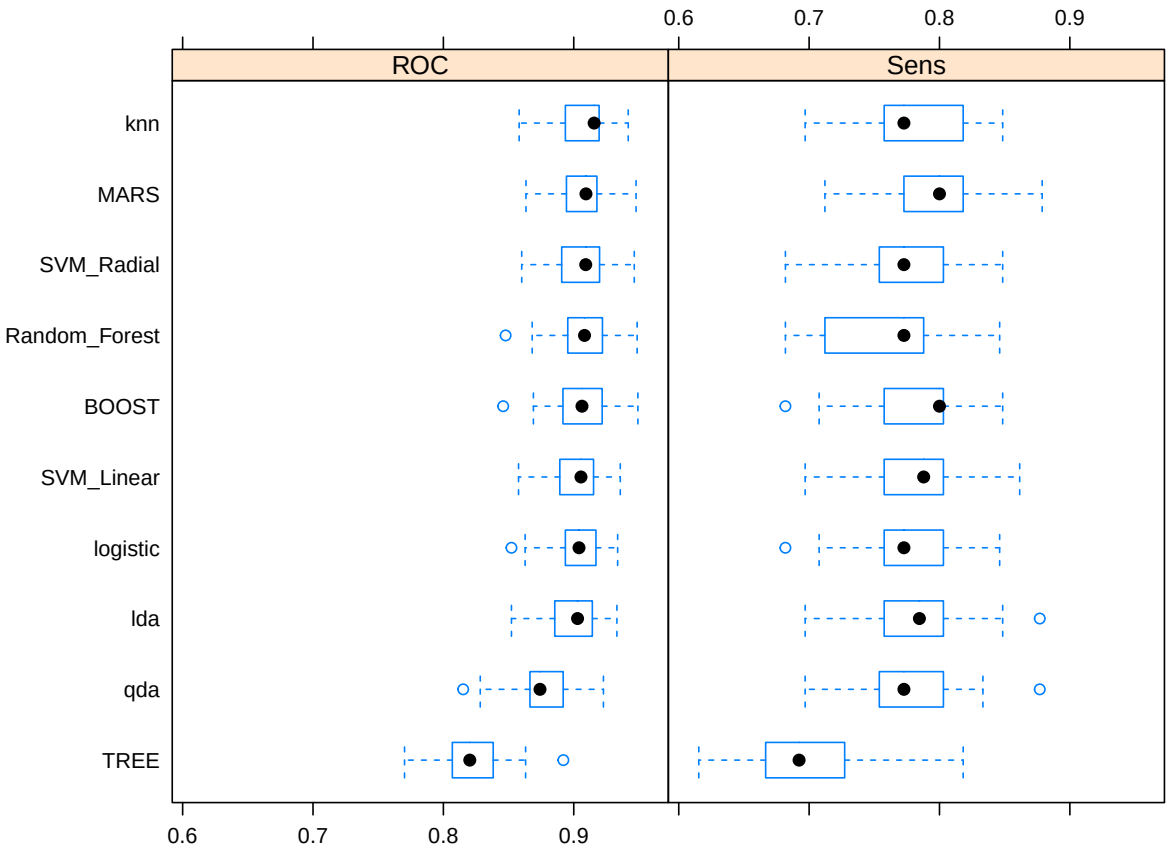
Variable type: numeric

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

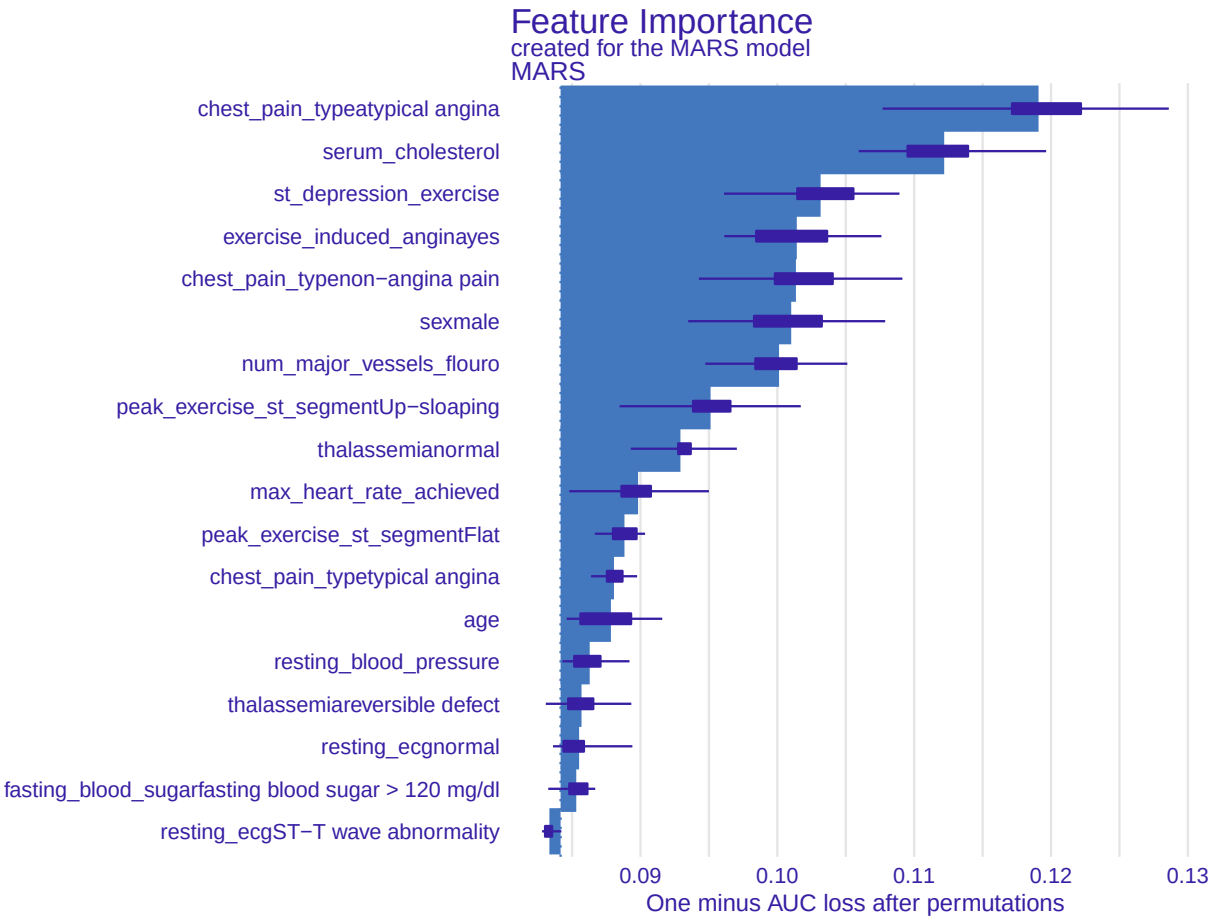


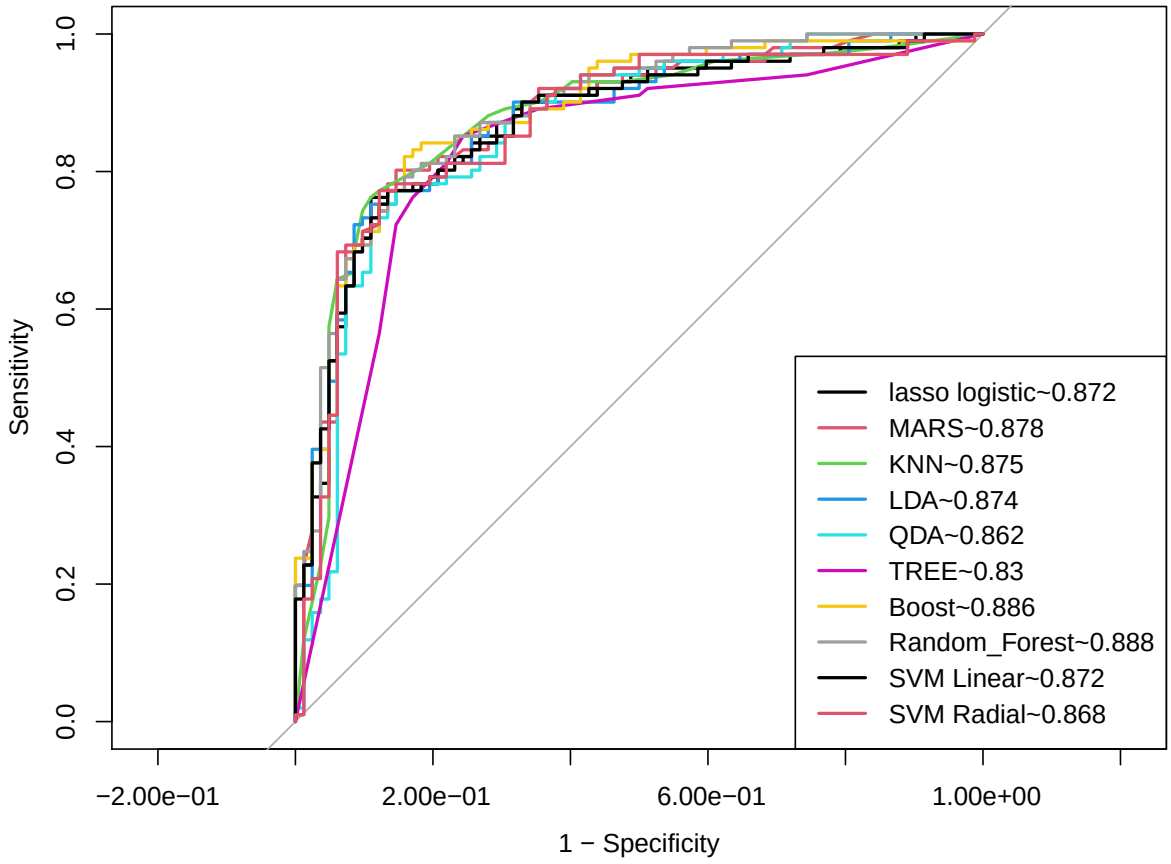

Lasso Logistics
ROC (Repeated Cross−Validation)
0.9
0.8
0.7
0.6
0.5
0.0183
0.3679
7.3891
148.4132
Regularization Parameter
Mixing Percentage
0.0
0.2
0.4
0.6
0.8
1.0
MARS
ROC (Repeated Cross−Validation)
0.90
0.89
0.88
0.87
5
10
15
20
#Terms
TREE
ROC (Repeated Cross−Validation)
0.82
0.81
0.80
0.79
0.005
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
1
5
10
20
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
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
st_depression_exercise
exercise_induced_anginayes
chest_pain_typenon−angina pain
sexmale
num_major_vessels_flouro
peak_exercise_st_segmentUp−sloaping
thalassemianormal
max_heart_rate_achieved
peak_exercise_st_segmentFlat
chest_pain_typetypical angina
age
resting_blood_pressure
thalassemiareversible defect
resting_ecgnormal
fasting_blood_sugarfasting blood sugar > 120 mg/dl
resting_ecgST−T wave abnormality
0.09
0.10
0.11
0.12
0.13
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~0.888
SVM Linear~0.872
SVM Radial~0.868