Midterm

Jeffrey Zhuohui Liang

3/7/2021

Introduction

Cardiovascular disease is the leading disease burden in U.S, according to *www.cdc.com* [?], 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 vardiovascular disease should be manage to improve population health.

```
One of many important manners is screening, \textit{American Heart Association} 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 [?] 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)

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

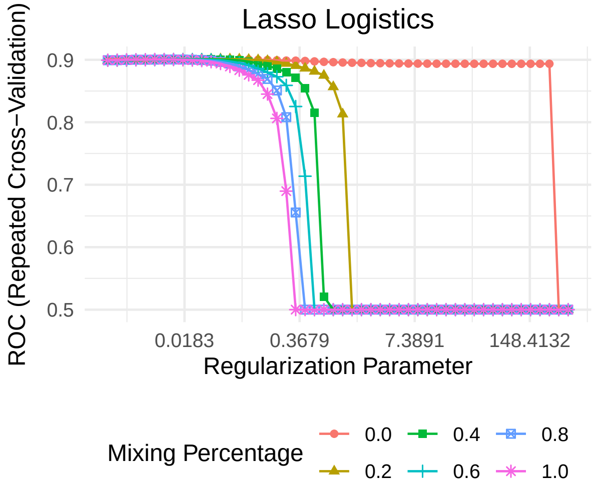
Variable type: numeric

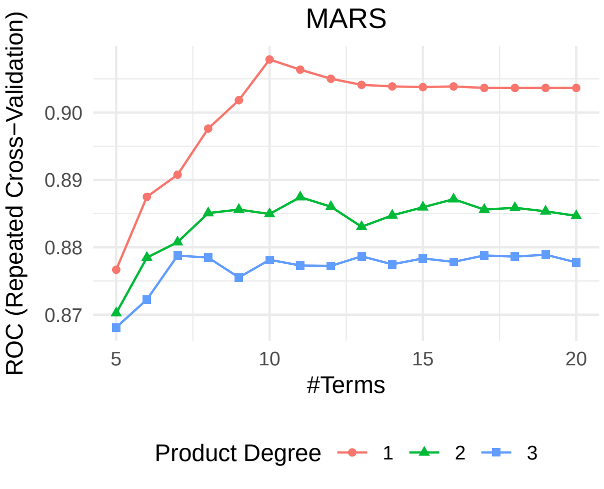
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1.00	53.51	9.42	28.0	47	54.0	60.0	77.0
resting_blood_pressure	59	0.94	132.13	19.07	0.0	120	130.0	140.0	200.0
serum_cholesterol	30	0.97	199.13	110.78	0.0	175	223.0	268.0	603.0
max_heart_rate_achieved	55	0.94	137.55	25.93	60.0	120	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

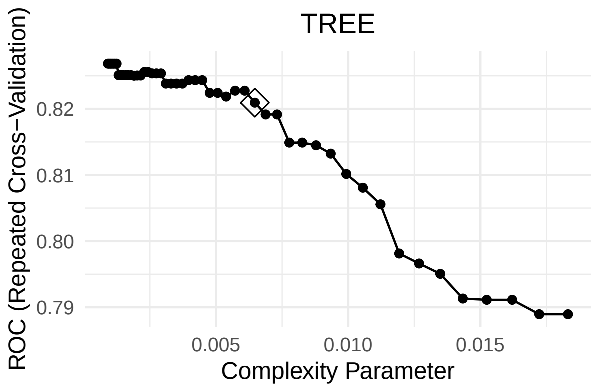
Modeling

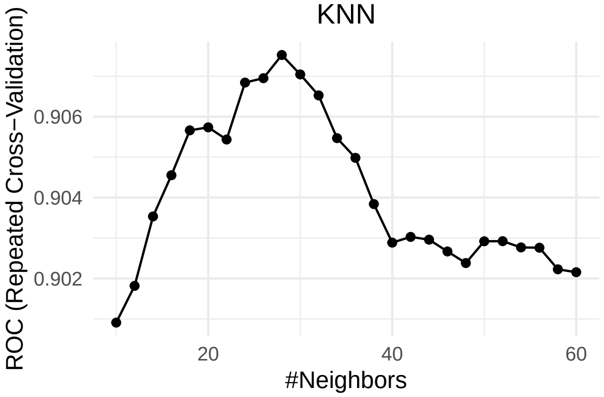
Table 4: Coefficient of Lasso Logistic Regression

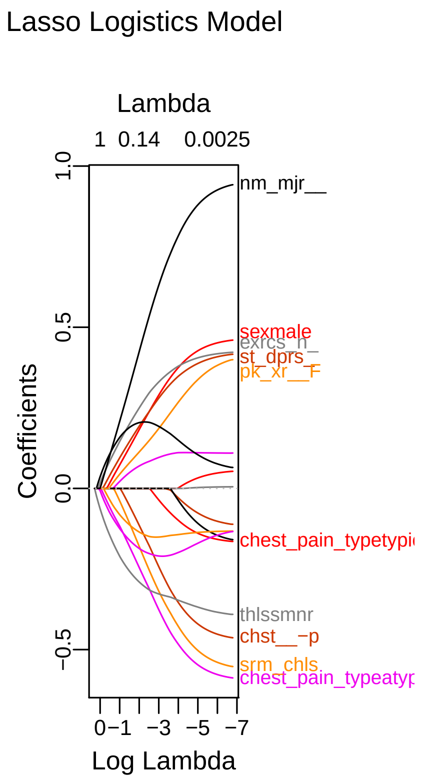
term	coefficient
Intercept	0.333
age	0.097
sexmale	0.296
chest_pain_typeatypical angina	-0.384
chest_pain_typenon-angina pain	-0.251
chest_pain_typetypical angina	-0.040
resting_blood_pressure	0.000
serum_cholesterol	-0.326
fasting_blood_sugarfasting blood sugar > 120 mg/dl	0.000
resting_ecgnormal	0.000
resting_ecgST-T wave abnormality	0.000
max_heart_rate_achieved	-0.151
exercise_induced_anginayes	0.334
st_depression_exercise	0.285
peak_exercise_st_segmentFlat	0.191
peak_exercise_st_segmentUp-sloaping	-0.210
num_major_vessels_flouro	0.642
thalassemianormal	-0.329
thalassemiareversible defect	0.190

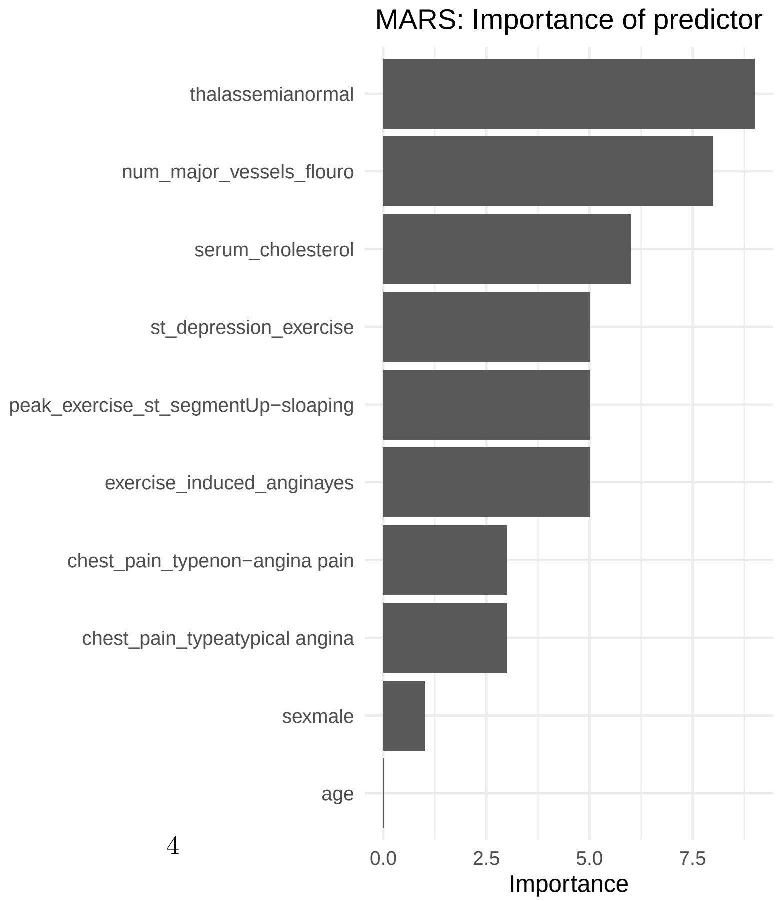

Lasso Logistics
ROC (Repeated Cross−Validation)
0.9
0.8
0.7
0.6
0.5
0.0183
0.3679
7.3891
148.4132
Regularization Parameter
Mixing Percentage
0.0
0.2
0.4
0.6
0.8
1.0

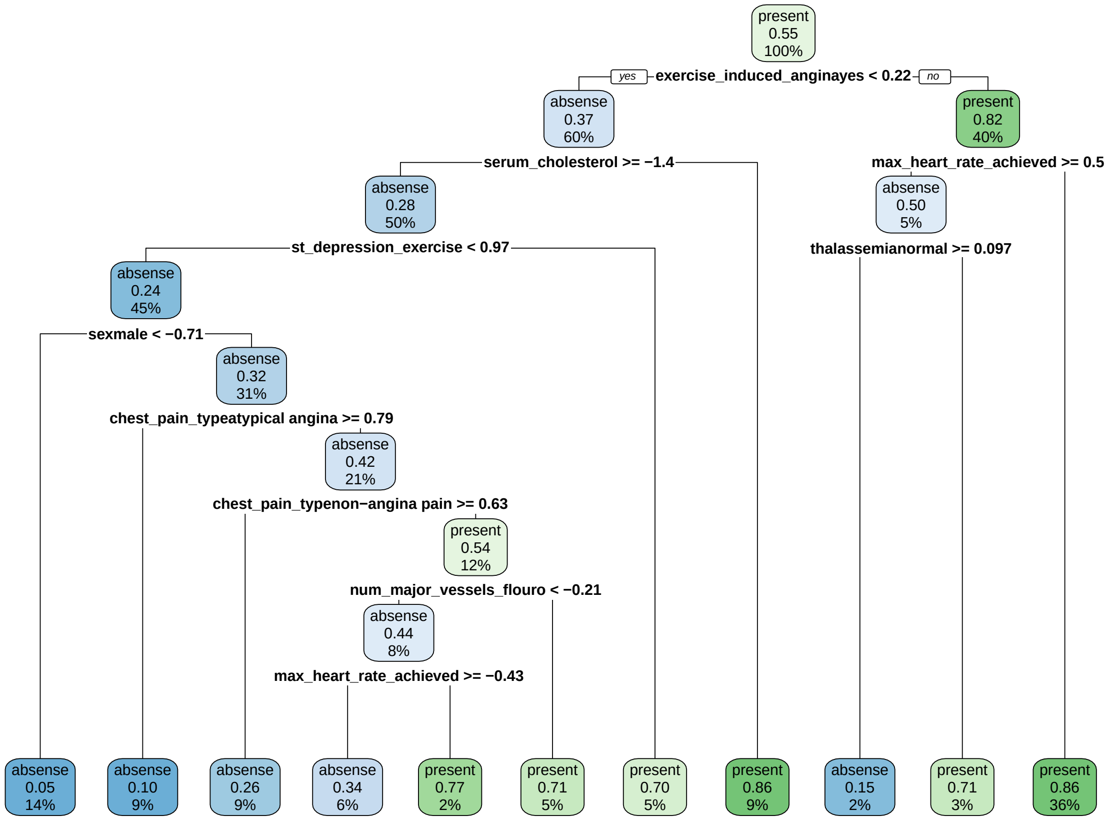

MARS
ROC (Repeated Cross−Validation)
0.90
0.89
0.88
0.87
5
10
15
20
#Terms
Product Degree
1
2
3


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

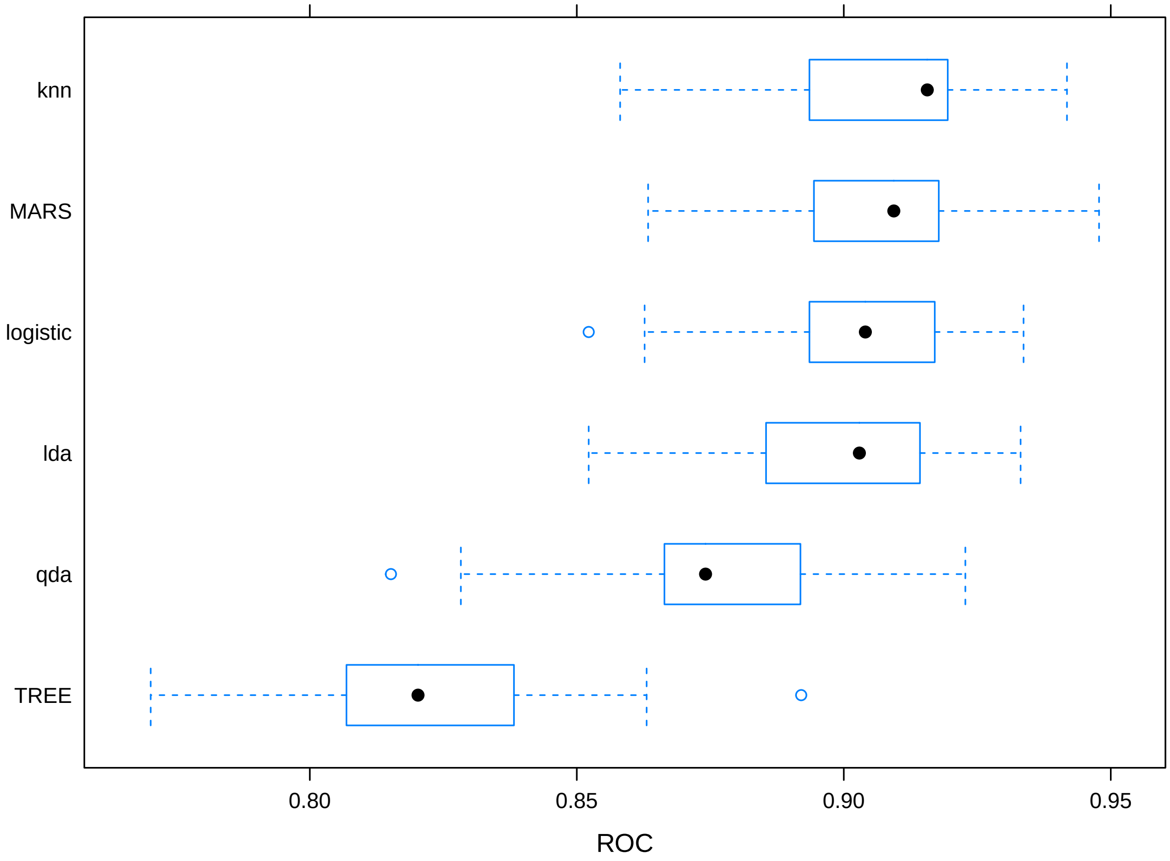

Lasso Logistics Model
Lambda
1
0.14
0.0025
Coefficients
1.0
0.5
0.0
−0.5
nm_mjr__
sexmale
exrcs_n_
st_dprs_
pk_xr__F
chest_pain_typetypi
thlssmnr
chst__−p
srm_chls
chest_pain_typeatyp
0
−1
−3
−5
−7
Log Lambda

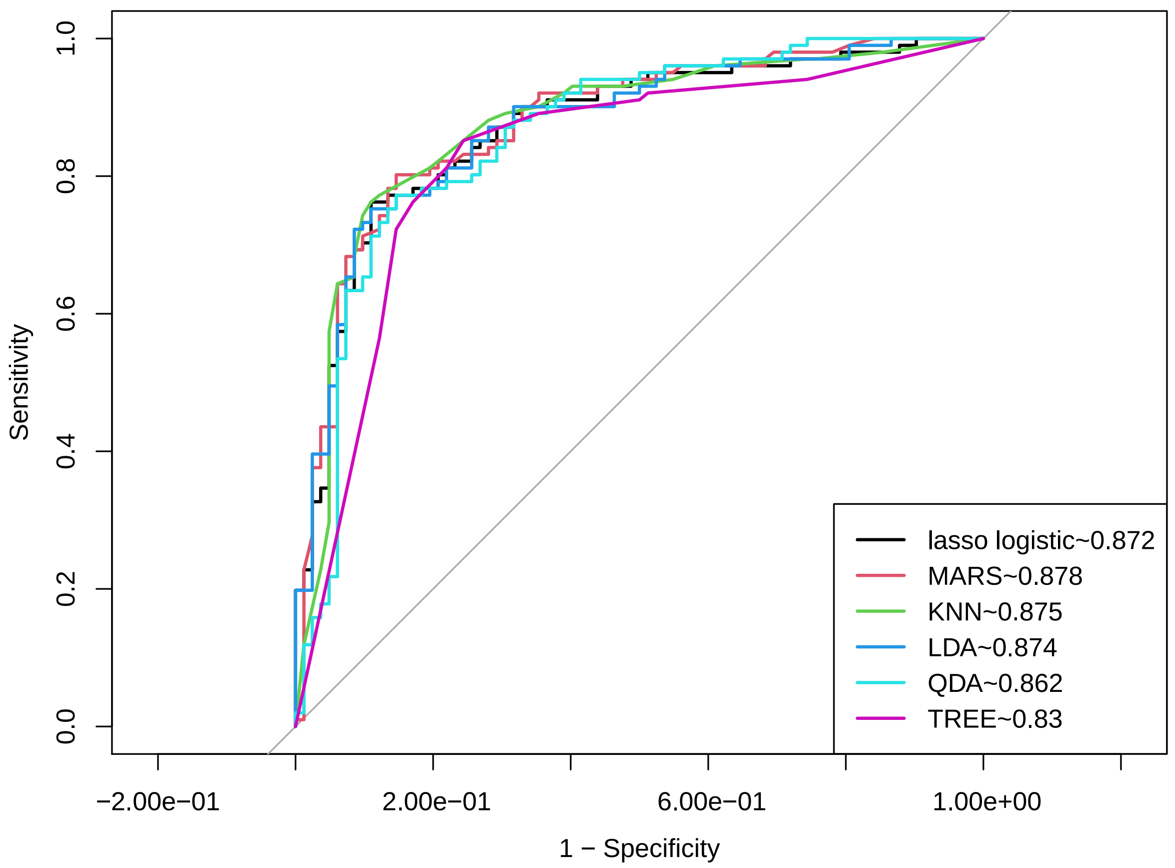

MARS: Importance of predictor
thalassemianormal
num_major_vessels_flouro
serum_cholesterol
st_depression_exercise
peak_exercise_st_segmentUp−sloaping
exercise_induced_anginayes
chest_pain_typenon−angina pain
chest_pain_typeatypical angina
sexmale
age
0.0
2.5
5.0
7.5
Importance



Performance comparison

```
##
## Call:
## summary.resamples(object = rsmp)
##
## Models: logistic, MARS, knn, lda, qda, TREE
## Number of resamples: 25
##
## ROC
##           Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
## logistic 0.852   0.894  0.904 0.901   0.917 0.934    0
## MARS     0.863   0.894  0.909 0.908   0.918 0.948    0
## knn      0.858   0.894  0.916 0.908   0.919 0.942    0
## lda      0.852   0.885  0.903 0.900   0.914 0.933    0
## qda      0.815   0.866  0.874 0.875   0.892 0.923    0
## TREE     0.770   0.807  0.820 0.821   0.838 0.892    0
##
## Sens
##           Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
## logistic 0.682   0.758  0.773 0.775   0.803 0.846    0
## MARS     0.712   0.773  0.800 0.795   0.818 0.879    0
## knn      0.697   0.758  0.773 0.781   0.818 0.848    0
## lda      0.697   0.758  0.785 0.783   0.803 0.877    0
## qda      0.697   0.754  0.773 0.779   0.803 0.877    0
## TREE     0.615   0.667  0.692 0.701   0.727 0.818    0
##
## Spec
##           Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
## logistic 0.815   0.852  0.877 0.868   0.890 0.915    0
## MARS     0.778   0.840  0.854 0.856   0.878 0.915    0
## knn      0.827   0.854  0.878 0.880   0.902 0.915    0
## lda      0.815   0.852  0.864 0.863   0.878 0.915    0
## qda      0.778   0.817  0.852 0.844   0.866 0.915    0
## TREE     0.728   0.815  0.840 0.846   0.878 0.938    0
```




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