P8106 Midterm Project

Yunlin Zhou

Introdution

Motivation

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Through this data set, we would like to explore how those features related to the heart disease, thus we can further use them to predict a possible heart disease.

Data preparation and cleaning

The variables in our data set are below:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [high: if FastingBS $>$ 120 mg/dl, other: otherwise]
7. 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]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. 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

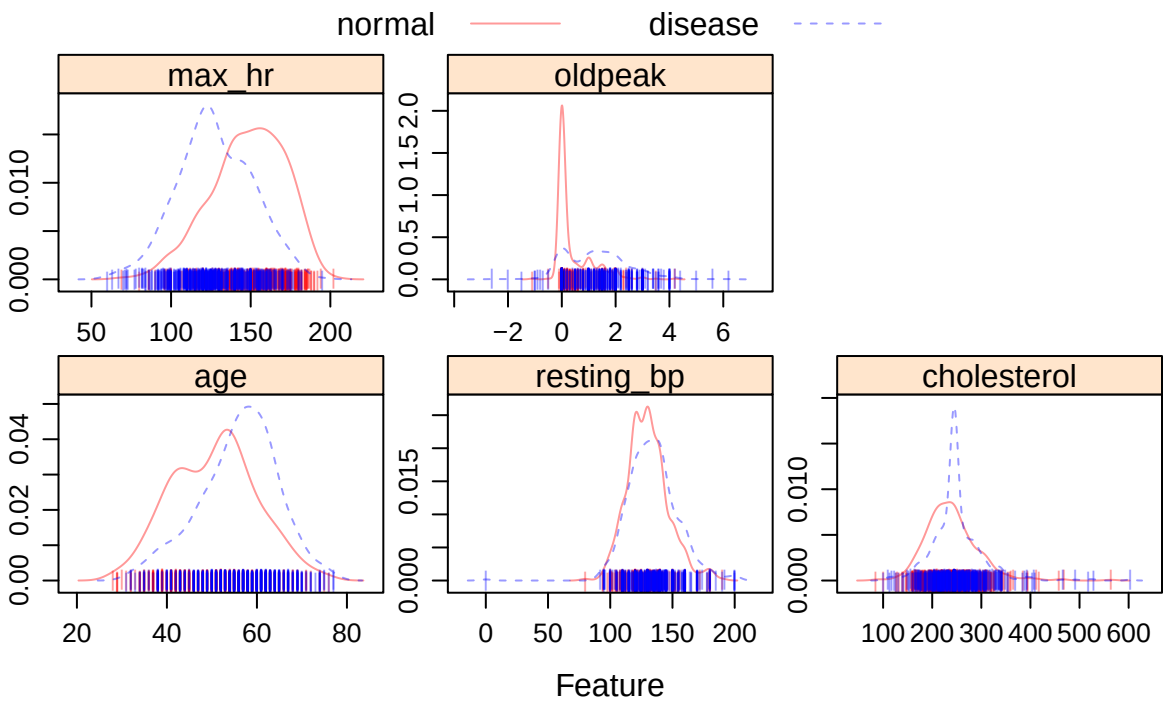
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
heart_disease	0	1	6	7	0	2	0
sex	0	1	1	1	0	2	0
chest_pain_type	0	1	2	3	0	4	0
fasting_bs	0	1	4	5	0	2	0
resting_ecg	0	1	2	6	0	3	0
exercise_angina	0	1	1	1	0	2	0
st_slope	0	1	2	4	0	3	0

Variable type: numeric

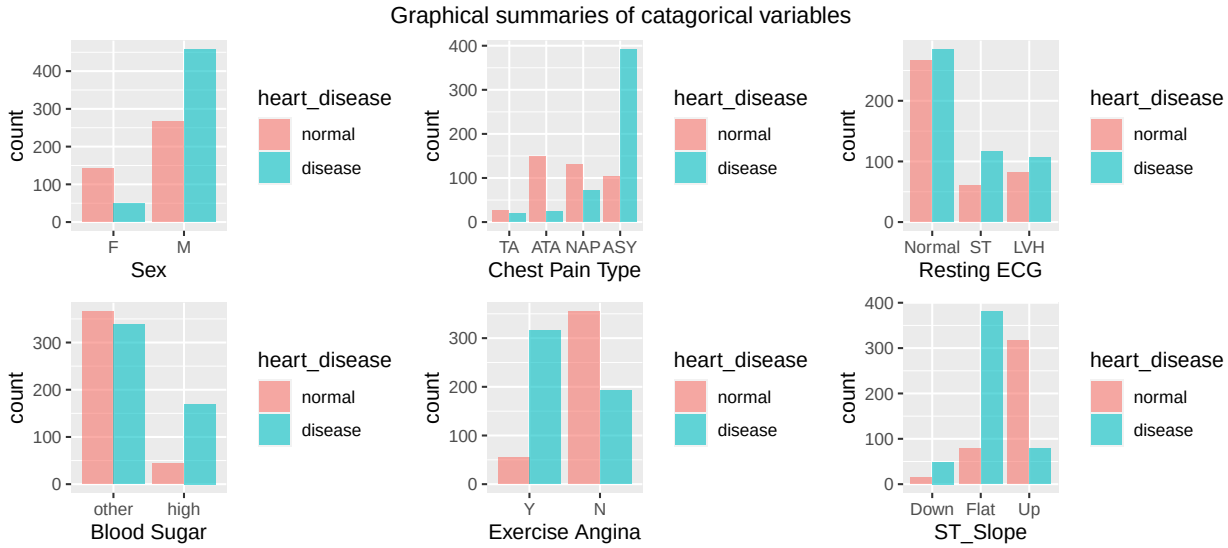
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1.00	53.51	9.43	28.0	47.00	54.0	60.0	77.0	
resting_bp	0	1.00	132.40	18.51	0.0	120.00	130.0	140.0	200.0	
cholesterol	172	0.81	244.64	59.15	85.0	207.25	237.0	275.0	603.0	
max_hr	0	1.00	136.81	25.46	60.0	120.00	138.0	156.0	202.0	
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 character variables, 5 numeric variables, with 918 observations. In the original data set, there was no null observations, but we found out that some data of Cholesterol was 0, which is not possible in real life. So we assume that those Cholesterol = 0 rows were actually null value when collecting the data. In that case, we use the mean value to replace the null observations. For the character variables, we use the function `factor()` to change the data type so that we could apply the data set to the models. For better using this data set to train the models, we split the data set into two parts: training data (70%) and test data (30%).

Exploratory analysis/visualization



From the density plot of continuous variables above, we can see 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 we can see from the plot above: male are tending to have the 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.

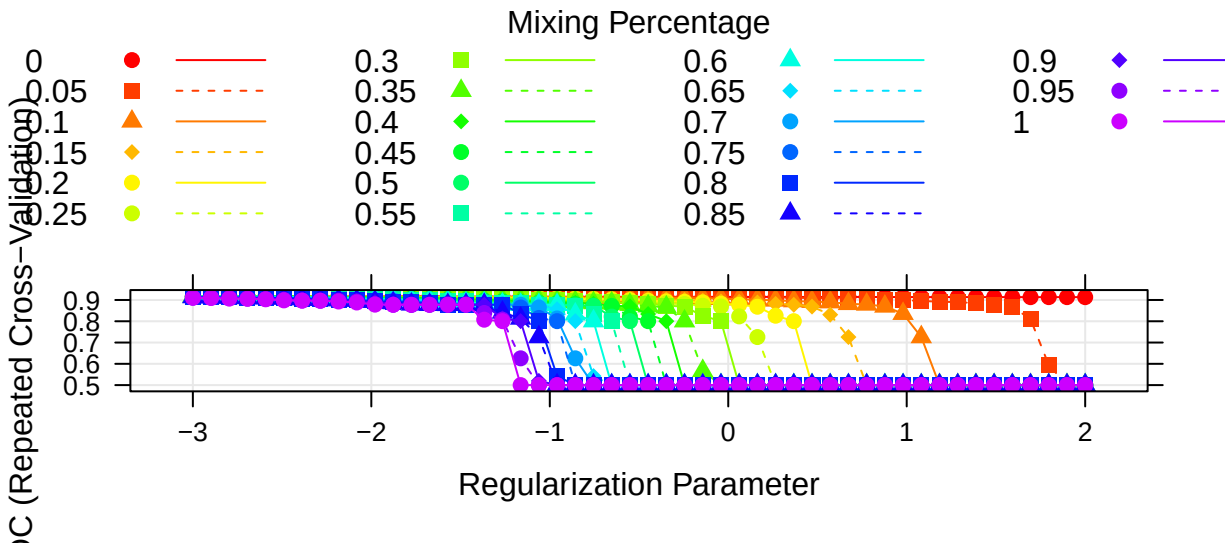
Models

Since our outcome is either having hear disease or not, we would use classification models including logistic regression, penalized logistic regression, GAM, MARS, LDA and QDA to train the data set.

There is no tuning parameter for logistic, GAM, LDA and QDA model.

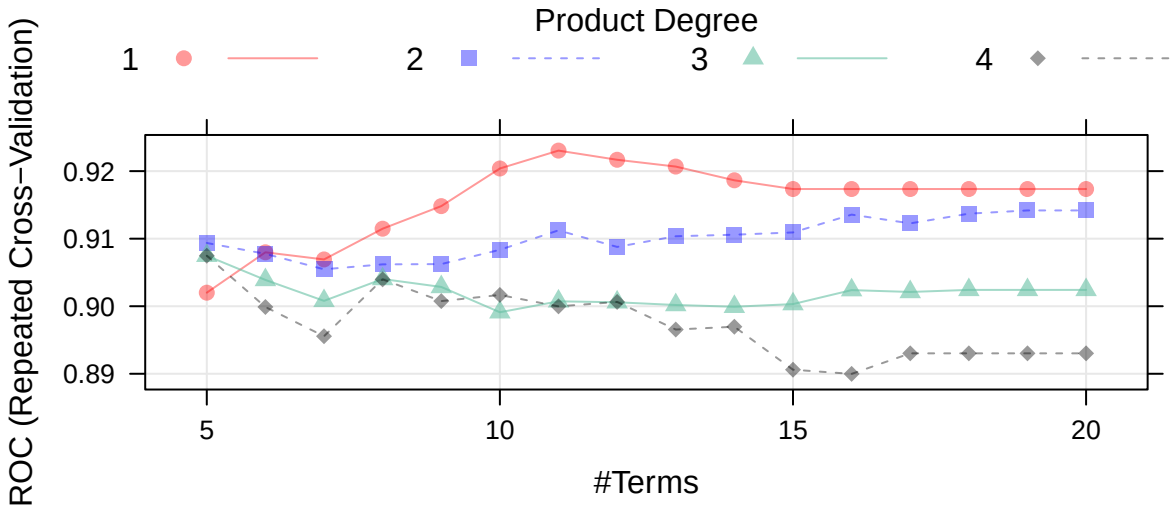
For penalized logistic regression, the best tuning parameters are alpha = 0.1 and lambda = 0.06105877. The plot below shows that the highest point is the best tuning parameter selection.

```
##     alpha     lambda
## 103   0.1 0.06105877
```



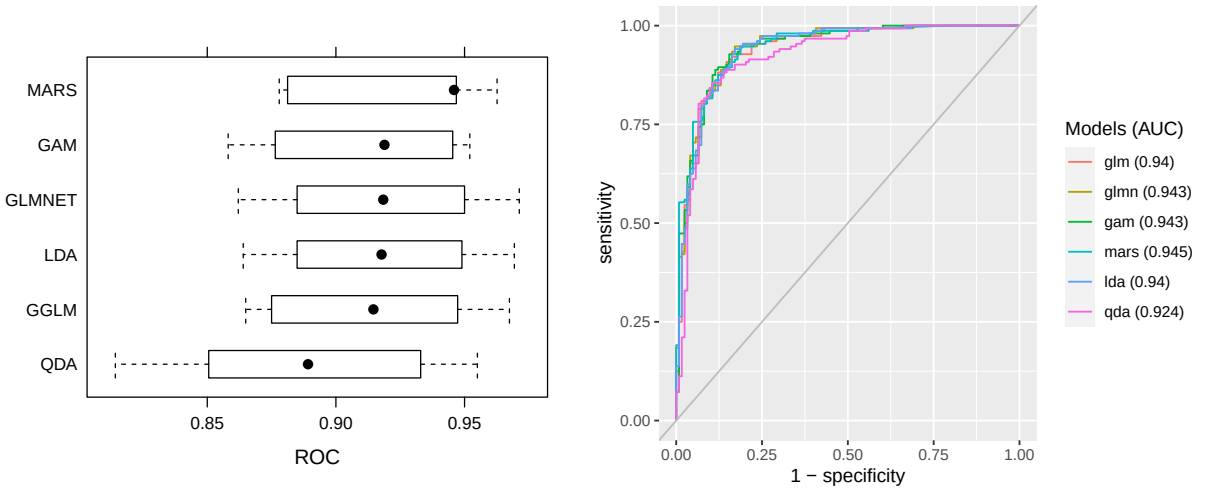
For MARS model, the best tuning parameters are nprune = 11 and degree = 1. The plot below shows that the highest point is the best tuning parameter selection.

```
##   nprune degree
## 7     11      1
```

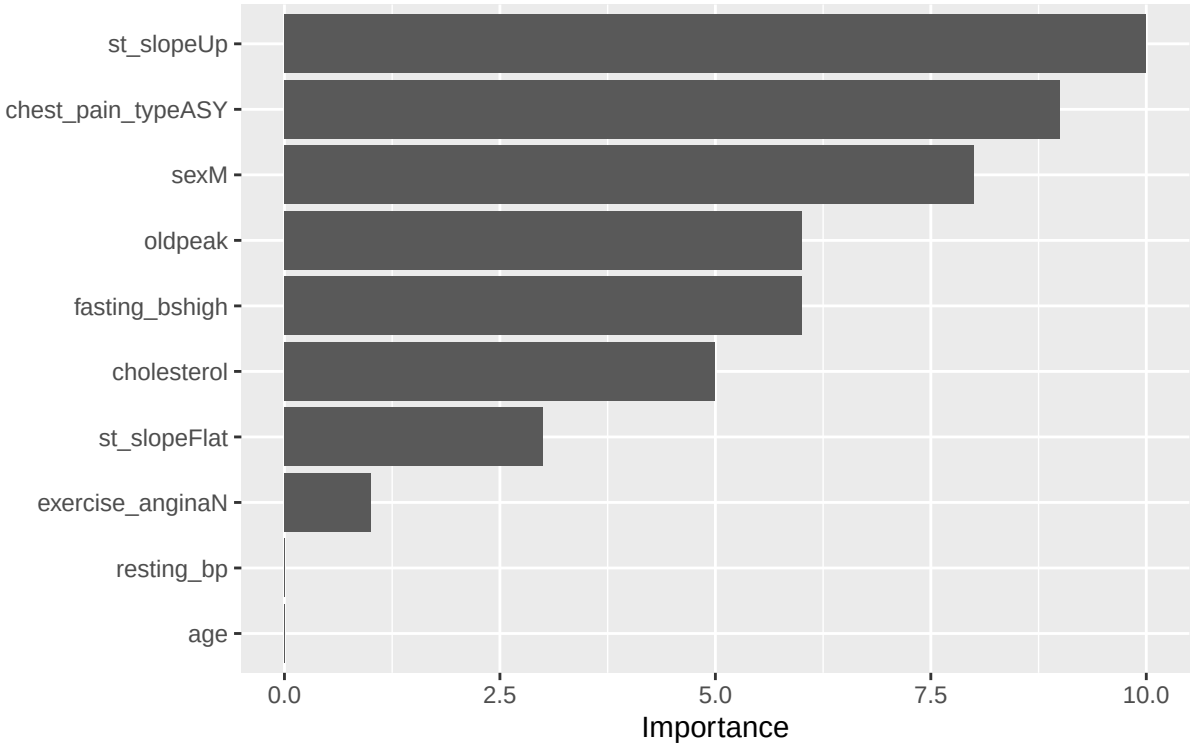


Find the best model

To find the best fitting model, we need to compare the models with their AUC area and ROC. As the plot shows below, the MARS model has the largest AUC area and ROC, so we choose MARS model as the best fitting model.



Feature Importance based on MARS model



According to the vip plot, we can conclude that st_slopeUp, chest_pain_typeASY, sexM, oldpeak, fasting_bshigh, cholesterol, st_slopeFlat, exercise_anginaN are statistically significant.

Conclusions

In the end, we choose the MARS model as our best fitting model because of its high sensitivity and specificity. When a patient has a up slope of the peak exercise ST segment, high old peak, high fasting blood sugar and cholesterol, we need to be more cautious since those features might suggest heart disease. Also, even though some patients have normal features, we might still need further test for accurate diagnoses.