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

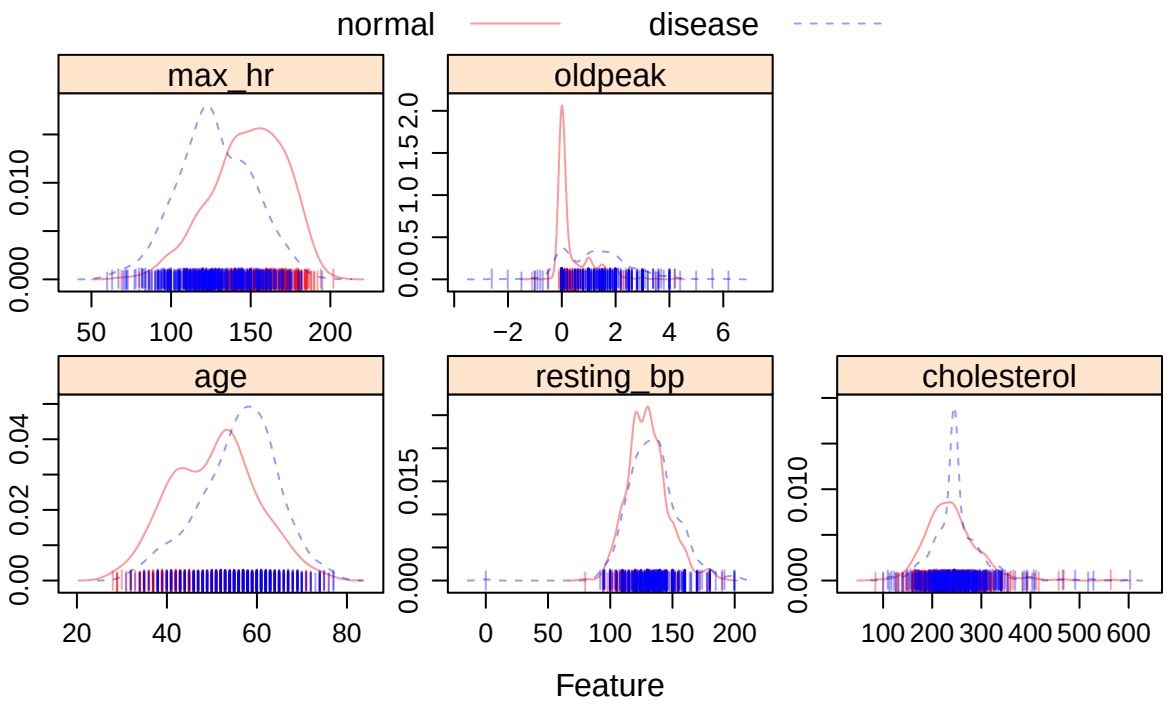
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
heart_disease	0	1	6	7	0	2	0
sex	0	1	1	1	0	2	0
chest_pain_type	0	1	2	3	0	4	0
fasting_bs	0	1	4	5	0	2	0
resting_ecg	0	1	2	6	0	3	0
exercise_angina	0	1	1	1	0	2	0
st_slope	0	1	2	4	0	3	0

Variable type: numeric

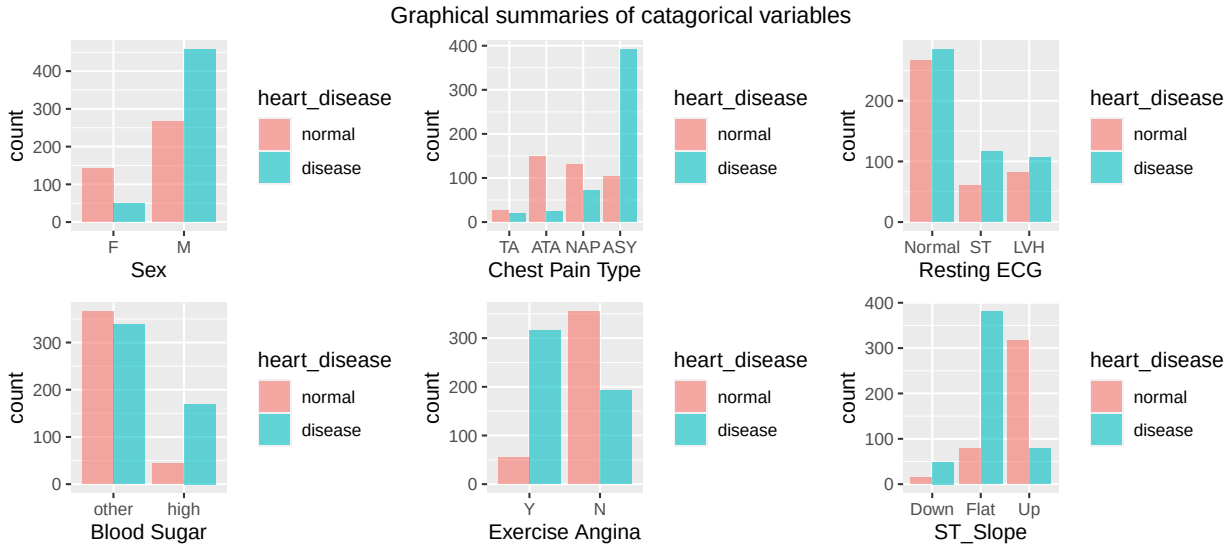
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1.00	53.51	9.43	28.0	47.00	54.0	60.0	77.0	
resting_bp	0	1.00	132.40	18.51	0.0	120.00	130.0	140.0	200.0	
cholesterol	172	0.81	244.64	59.15	85.0	207.25	237.0	275.0	603.0	
max_hr	0	1.00	136.81	25.46	60.0	120.00	138.0	156.0	202.0	
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.

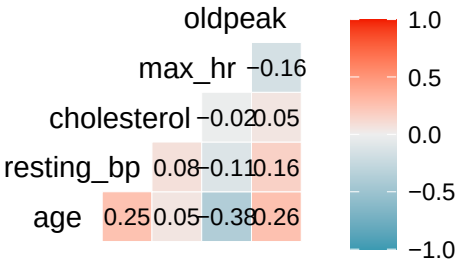


As we can see from the plot above: male are tending to have the heart disease; if the patients have Exercise Angina or flat ST slope, they are more likely to have heart disease. However, even if the patient has normal features like no chest pain, normal resting ECG and blood sugar, they could still have heart disease.

Models

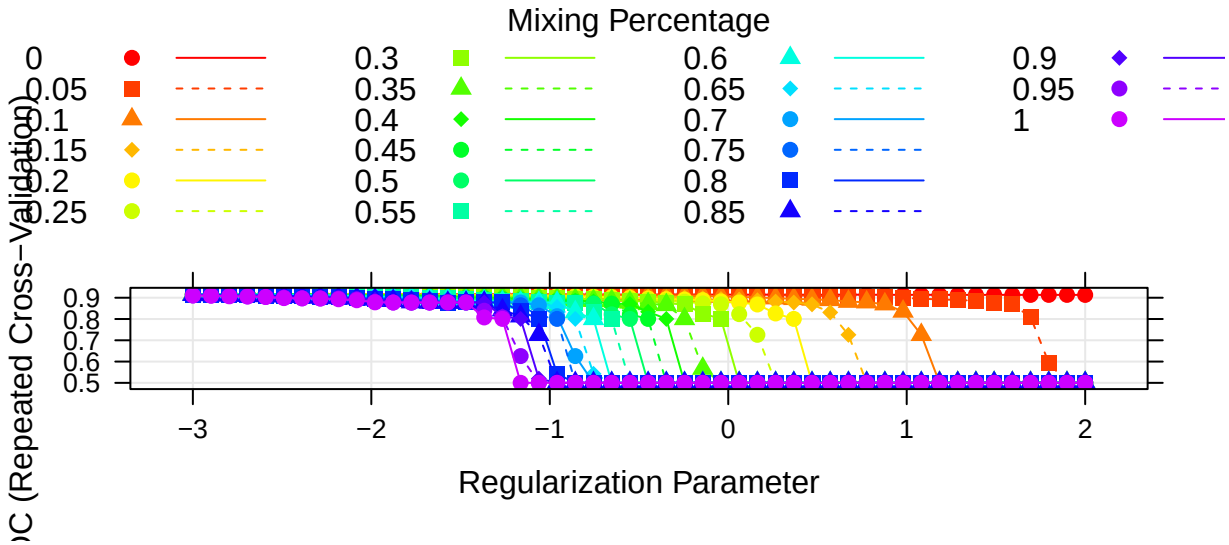
Since our outcome is either having hear disease or not, we would use classification models (including logistic regression, penalized logistic regression, GAM, MARS, LDA and QDA) with 10 fold validation to train the data set. We use all the variables in the data set to fit the model.

As we can see from the Correlation plot below, we can conclude that age and max_hr, as well as age and oldpeak, are relatively highly correlated. To fit logistic regression model, we need to make sure that the predictors are not correlated. Since age and oldpeak or max_hr are correlated, the result might be affected.



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.

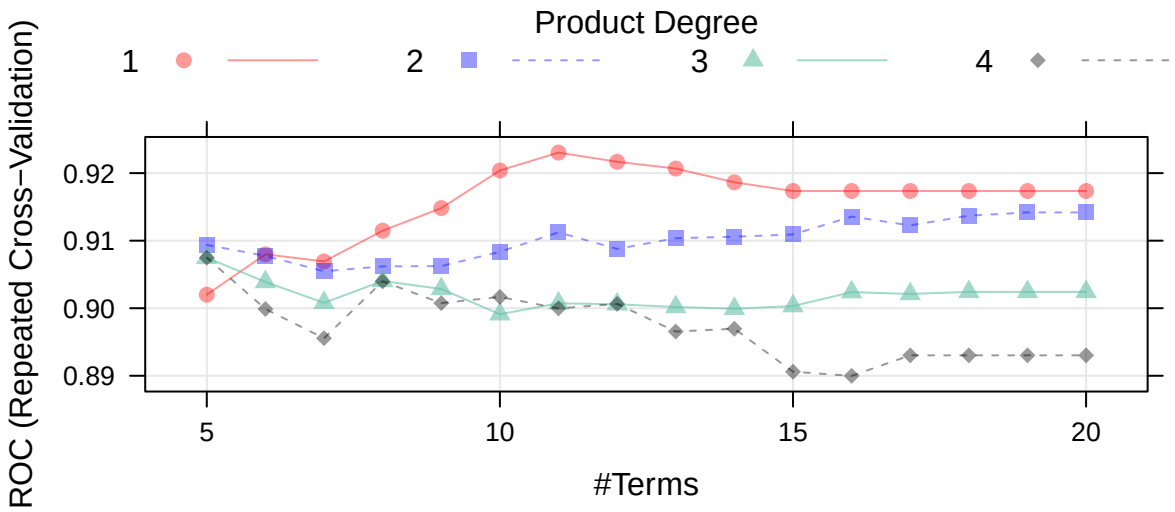
```
##     alpha     lambda
## 103   0.1 0.06105877
```



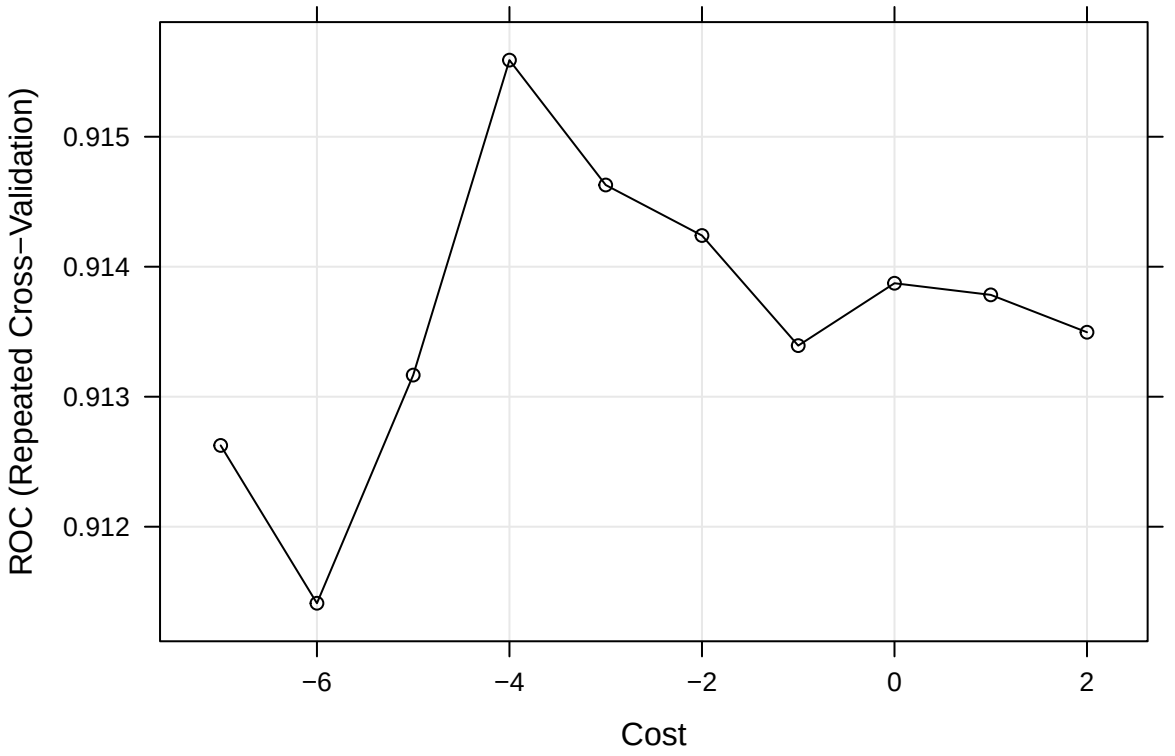
For GAM model, we use GCV to select the degree of freedom. By looking at the formula in the final model, we can conclude that resting_bp is not an important predictor since its df is close to 0. The GAM model could automatically model non-linear relationships that standard linear regression will miss and potentially make more accurate predictions.

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
```



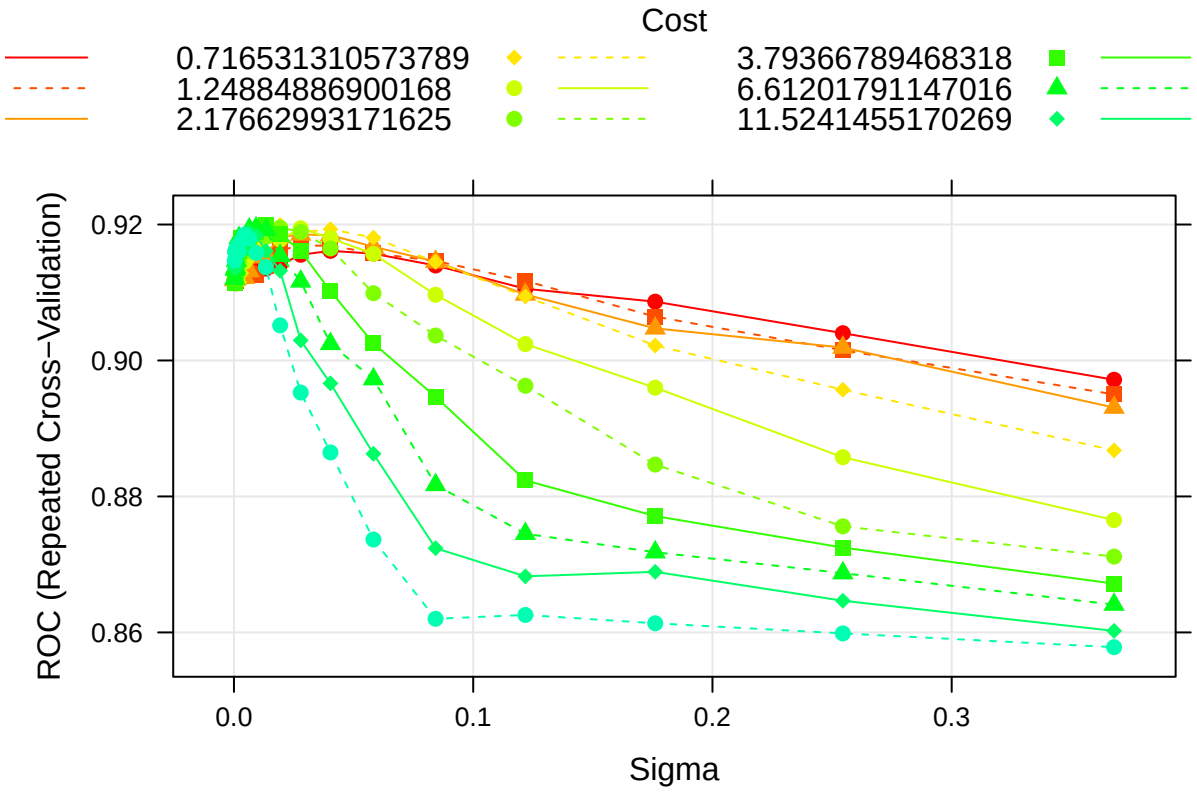
SVM with Linear Kernel




```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 0.0183156388887342
```

```
##
## Linear (vanilla) kernel function.
##
## Number of Support Vectors : 266
##
## Objective Function Value : -4.4482
## Training error : 0.143079
## Probability model included.
```

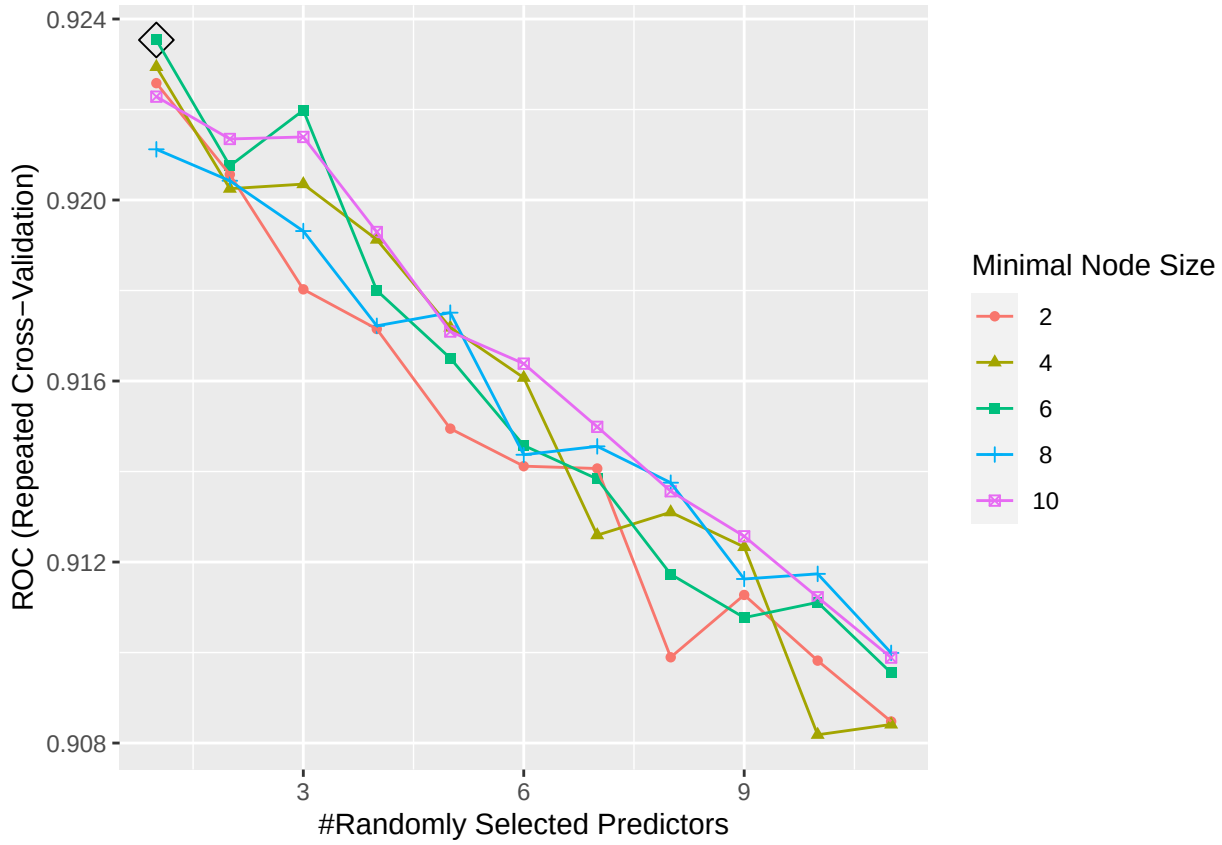
SVM with Radial Kernel




```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 3.79366789468318
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.0133560011114399
##
## Number of Support Vectors : 248
##
## Objective Function Value : -777.7854
## Training error : 0.127527
## Probability model included.
```

Random Forest

Under the random forest model, the Minimal Node Size with highest ROC curve with repeated cross-validation is 6.

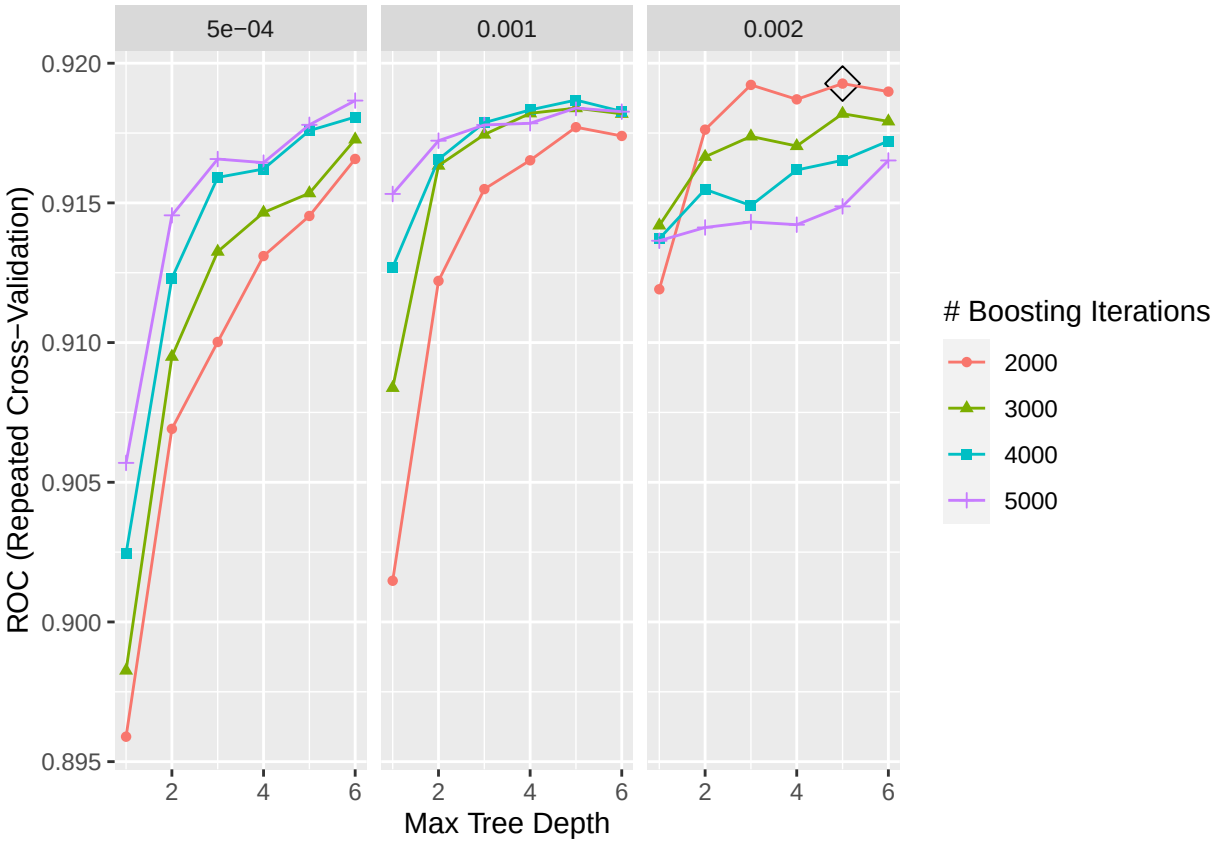


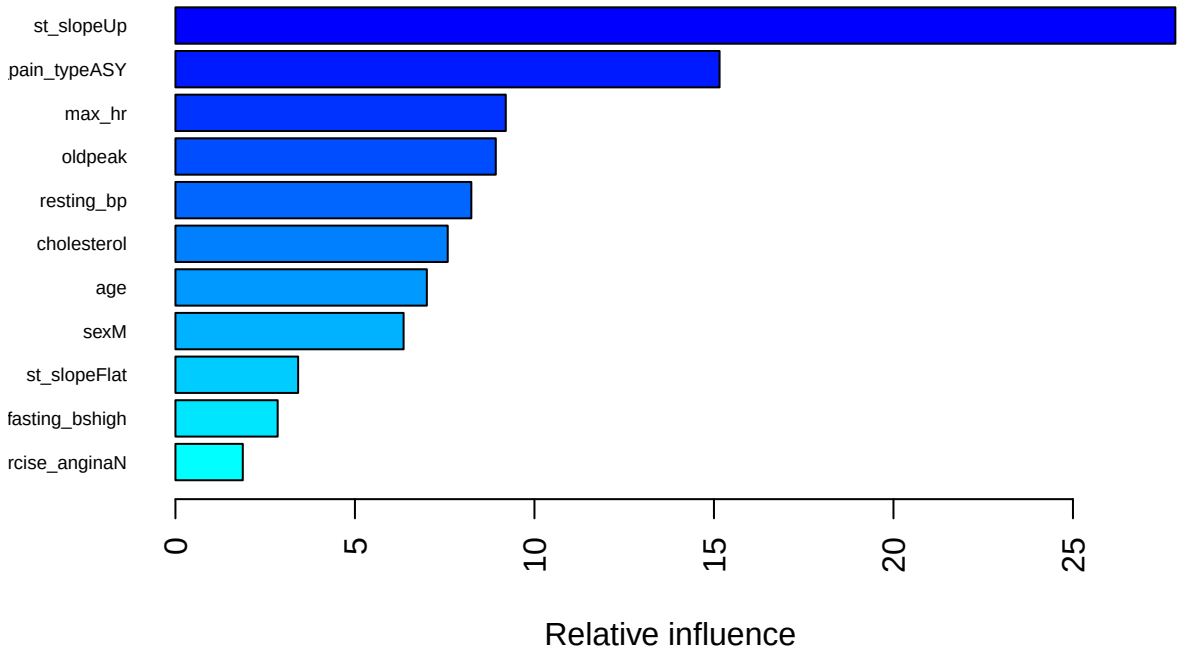

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction normal disease
##    normal     102       6
##    disease     21     146
##
##                Accuracy : 0.9018
##                  95% CI : (0.8604, 0.9343)
##     No Information Rate : 0.5527
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7991
##
##  Mcnemar's Test P-Value : 0.007054
##
##             Sensitivity : 0.8293
##             Specificity : 0.9605
##          Pos Pred Value : 0.9444
##          Neg Pred Value : 0.8743
##              Prevalence : 0.4473
```

```
##          Detection Rate : 0.3709
##    Detection Prevalence : 0.3927
##       Balanced Accuracy : 0.8949
##
##        'Positive' Class : normal
##
```

boosting

For the adaboost model, the number of boosting iterations is 2000 and the maximum depth is 3.



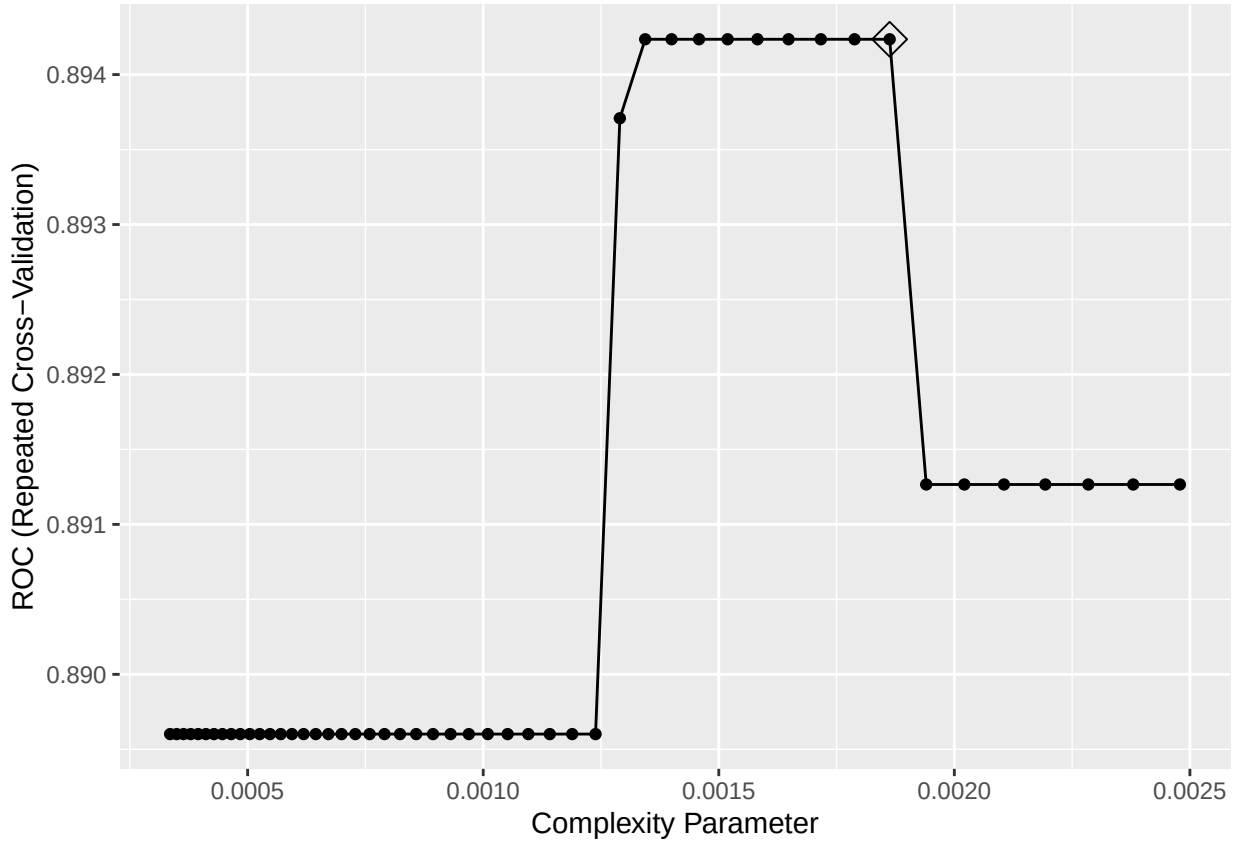



```
##                                   var    rel.inf
## st_slopeUp                 st_slopeUp 27.8520908
## chest_pain_typeASY chest_pain_typeASY 15.1566159
## max_hr                         max_hr  9.2016104
## oldpeak                       oldpeak  8.9233103
## resting_bp                 resting_bp  8.2415726
## cholesterol               cholesterol  7.5843182
## age                               age  7.0029013
## sexM                             sexM  6.3548872
## st_slopeFlat             st_slopeFlat  3.4174186
## fasting_bshigh         fasting_bshigh  2.8478775
## exercise_anginaN     exercise_anginaN  1.8760094
## resting_ecgLVH         resting_ecgLVH  0.6907263
## chest_pain_typeATA chest_pain_typeATA  0.4049399
## chest_pain_typeNAP chest_pain_typeNAP  0.2796364
## resting_ecgST           resting_ecgST  0.1660850
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction normal disease
##    normal      99      13
##    disease     24     139
##
##                Accuracy : 0.8655
```

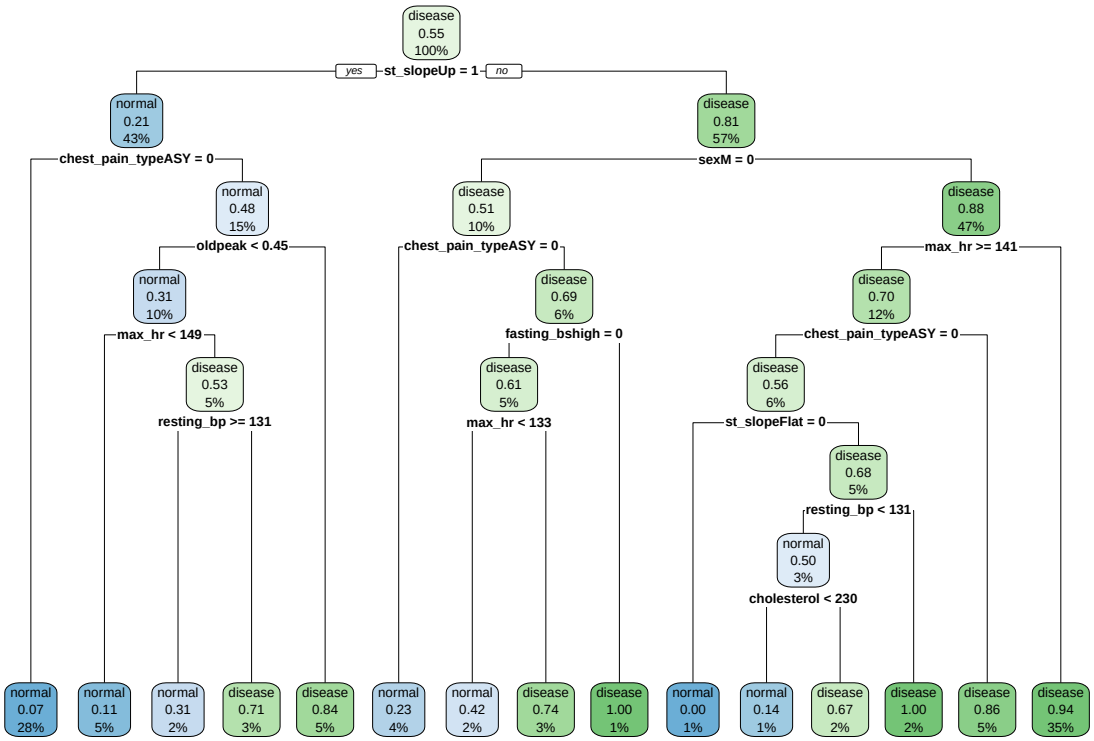
```
##                  95% CI : (0.8193, 0.9035)
##     No Information Rate : 0.5527
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7255
##
##  Mcnemar's Test P-Value : 0.1002
##
##             Sensitivity : 0.8049
##             Specificity : 0.9145
##          Pos Pred Value : 0.8839
##          Neg Pred Value : 0.8528
##              Prevalence : 0.4473
##          Detection Rate : 0.3600
##    Detection Prevalence : 0.4073
##       Balanced Accuracy : 0.8597
##
##        'Positive' Class : normal
##
```

Classification Trees



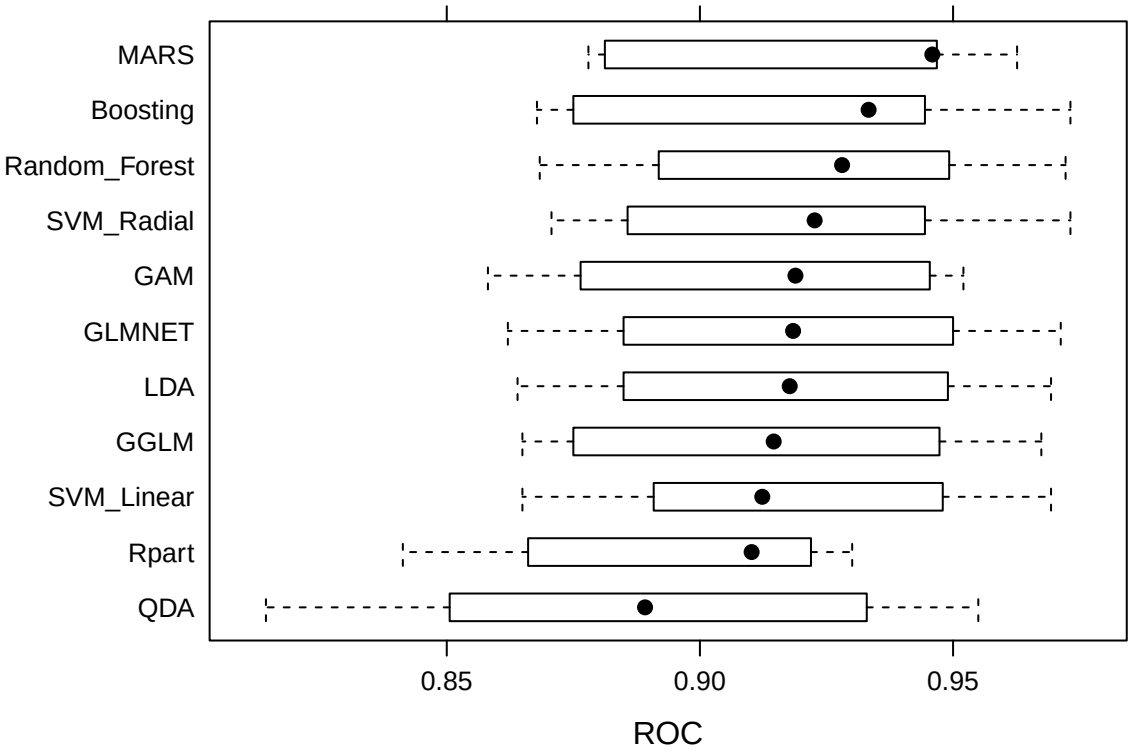

```
##            cp
## 43 0.001862726
```

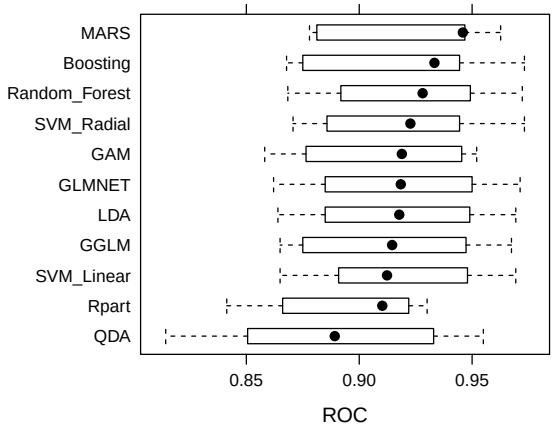
Create a plot of the tree.

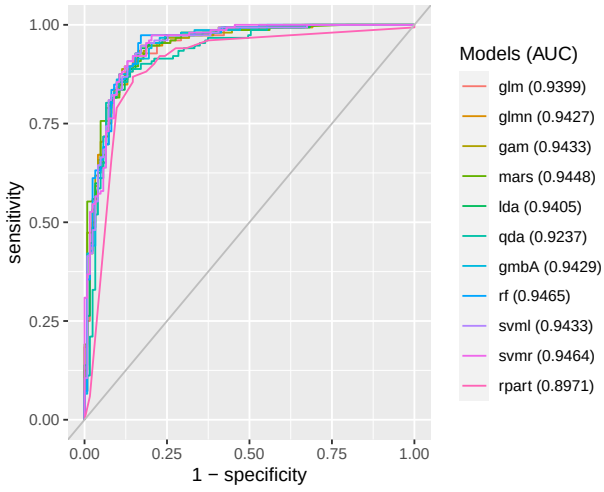


Find the best model

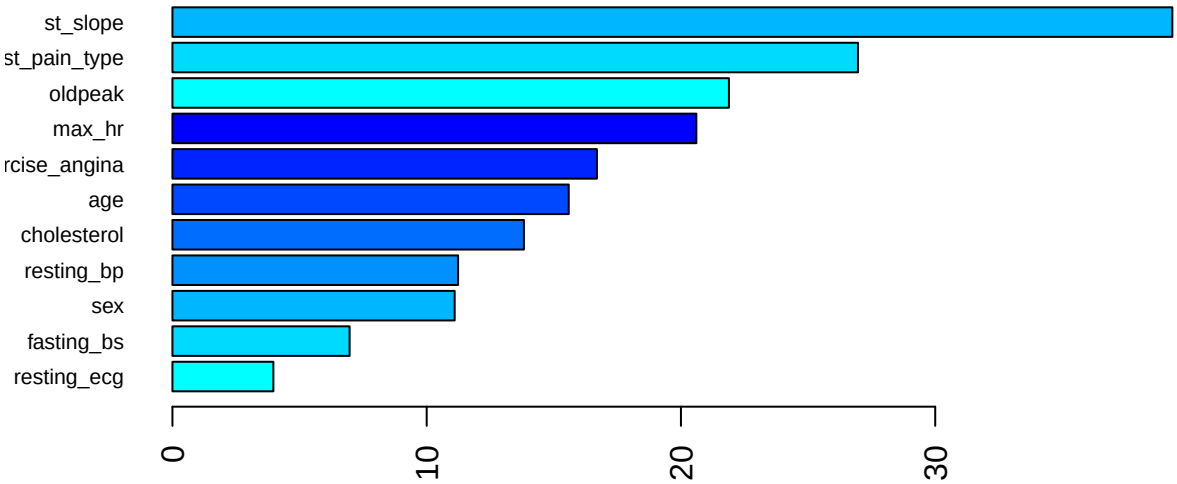
To find the best fitting model, we need to compare the models with their AUC . As the plot shows below, the MARS model has the largest AUC, so we choose MARS model as the best fitting model.

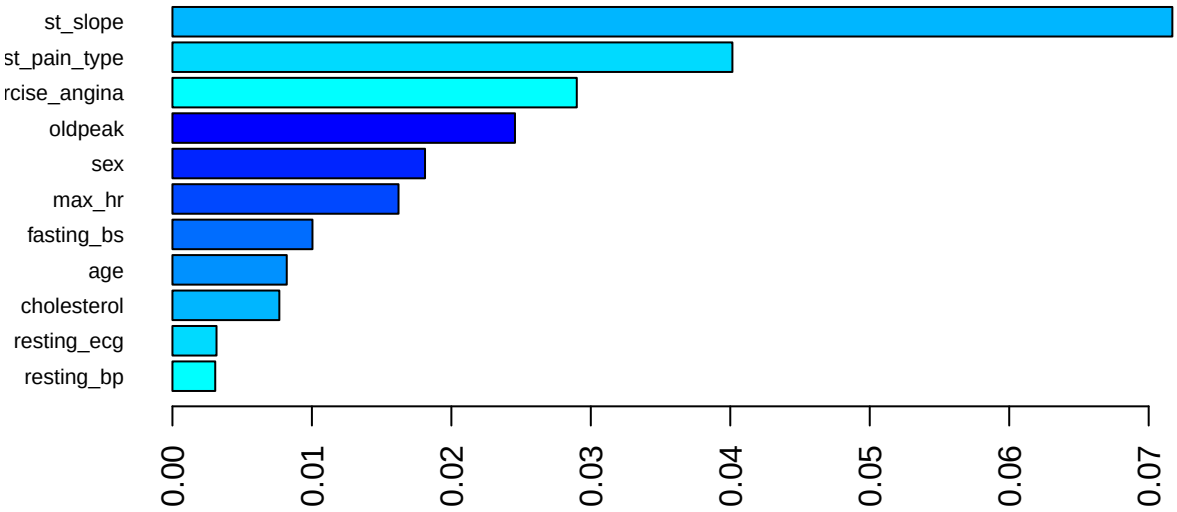

MARS
Boosting
Random_Forest
SVM_Radial
GAM
GLMNET
LDA
GGLM
SVM_Linear
Rpart
QDA
0.85
0.90
0.95
ROC


MARS
Boosting
Random_Forest
SVM_Radial
GAM
GLMNET
LDA
GGLM
SVM_Linear
Rpart
QDA
0.85
0.90
0.95
ROC


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 − specificity
Models (AUC)
glm (0.9399)
glmn (0.9427)
gam (0.9433)
mars (0.9448)
lda (0.9405)
qda (0.9237)
gmbA (0.9429)
rf (0.9465)
svml (0.9433)
svmr (0.9464)
rpart (0.8971)

Feature Importance based on Random Forest model





PDP plot