Midterm

Jeffrey Zhuohui Liang

3/7/2021

Method

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

```r
hrt_data = hrt_data %>% select(-location)
# split into training set
train_index = createDataPartition(hrt_data$diagnosis_heart_disease,p=0.8,list = F)

Y_tr = hrt_data$diagnosis_heart_disease[train_index]

Y_ts = hrt_data$diagnosis_heart_disease[-train_index]

options(na.action = "na.pass")
X_tr = model.matrix(diagnosis_heart_disease ~., hrt_data,na.action = "na.pass")[train_index,-1]

X_ts = model.matrix(diagnosis_heart_disease ~., hrt_data,na.action = "na.pass")[-train_index,-1]

TRC = caret::trainControl(method = "repeatedcv",repeats=5,
                          number = 5,
                          summaryFunction = twoClassSummary,
                          classProbs = T)
```

Modeling

```r
set.seed(123123)
cl = parallel::makePSOCKcluster(5)
doParallel::registerDoParallel(cl)

logistic_model =
  train(
    X_tr,
    Y_tr,
    method = "glmnet",
    tuneGrid = expand.grid(alpha = seq(0,1,length=6),
                           lambda = exp(seq(
                             6, to = -6, length = 50
                           ))),
    family = "binomial",
    preProcess = c("knnImpute", "center", "scale"),
    metric = "ROC",
    trControl = TRC
  )

stopCluster(cl)

p_logistics =
  ggplot(logistic_model) +
  scale_x_continuous(trans = "log")+
  labs(title = "Lasso Logistics")
```

```r
set.seed(123123)
cl = parallel::makePSOCKcluster(5)
doParallel::registerDoParallel(cl)
```

```r
mars_model =
  train(X_tr,
        Y_tr,
        method = "earth",
        tuneGrid = expand.grid(degree = 1:3,
                               nprune = 5:20),
        preProcess = c("knnImpute", "center", "scale"),
        trControl = TRC,
        metric = "ROC")

stopCluster(cl)

p_mars = ggplot(mars_model) +labs(title ="MARS")
```

```r
set.seed(123123)
cl = parallel::makePSOCKcluster(5)
doParallel::registerDoParallel(cl)

knn_model =
  train(X_tr,
        Y_tr,
        method = "knn",
        tuneGrid = expand.grid(k = seq(10,60,2)),
        preProcess = c("knnImpute", "center", "scale"),
        trControl = TRC,
        metric = "ROC")

stopCluster(cl)

p_knn = ggplot(knn_model)+labs(title = "KNN")
```

```r
set.seed(123123)
cl = parallel::makePSOCKcluster(5)
doParallel::registerDoParallel(cl)

lda_model =  train(
  X_tr,
  Y_tr,
  method = "lda",
  preProcess = c("knnImpute", "center", "scale"),
  trControl = TRC,
  metric = "ROC"
)

stopCluster(cl)
```

```r
set.seed(123123)
cl = parallel::makePSOCKcluster(5)
doParallel::registerDoParallel(cl)

qda_model =  train(
  X_tr,
  Y_tr,
  method = "qda",
  preProcess = c("knnImpute", "center", "scale"),
  trControl = TRC,
  metric = "ROC"
)

stopCluster(cl)
```

```
set.seed(123123)
cl = parallel::makePSOCKcluster(5)
doParallel::registerDoParallel(cl)

tree_model =
  train(
    X_tr,
    Y_tr,
    method = "rpart",
    tuneGrid = expand.grid(
      cp = exp(seq(-7,-4,length=50))
    ),
    preProcess = c("center", "scale"),
    trControl = caret::trainControl(method = "repeatedcv",repeats=5,
                          number = 5,
                          summaryFunction = twoClassSummary,
                          classProbs = T,
                          selectionFunction = "oneSE"),
    metric = "ROC"
  )
stopCluster(cl)

p_tree =
  ggplot(tree_model,highlight = T) + labs(title = "TREE")
```

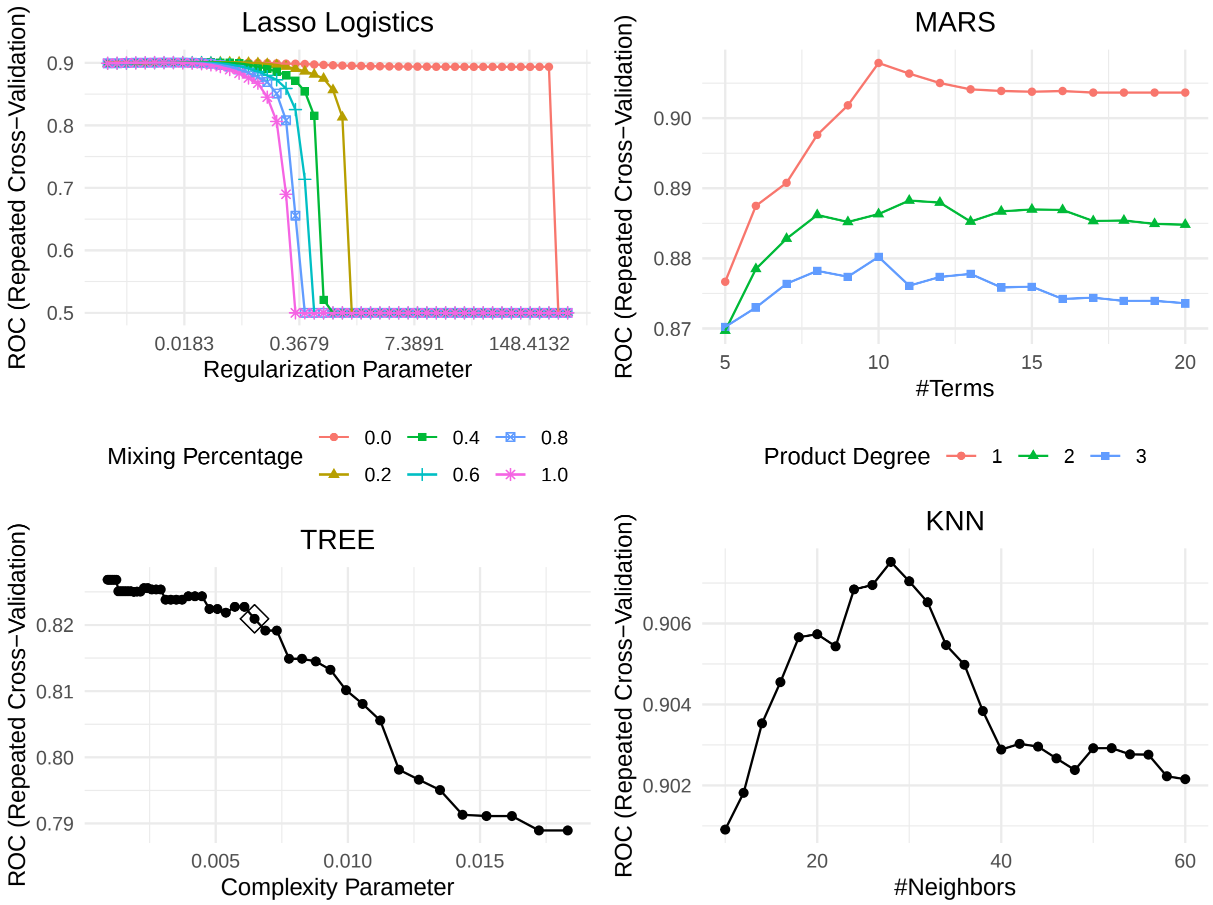
```
coef(logistic_model$finalModel,logistic_model$bestTune$lambda) %>%
  as.vector() %>%
  tibble(term = c("Intercept",colnames(X_tr)),
         coefficient = .) %>%
  knitr::kable(caption = "Coefficient of Lasso Logistic Regression")
```

Table 4: Coefficient of Lasso Logistic Regression

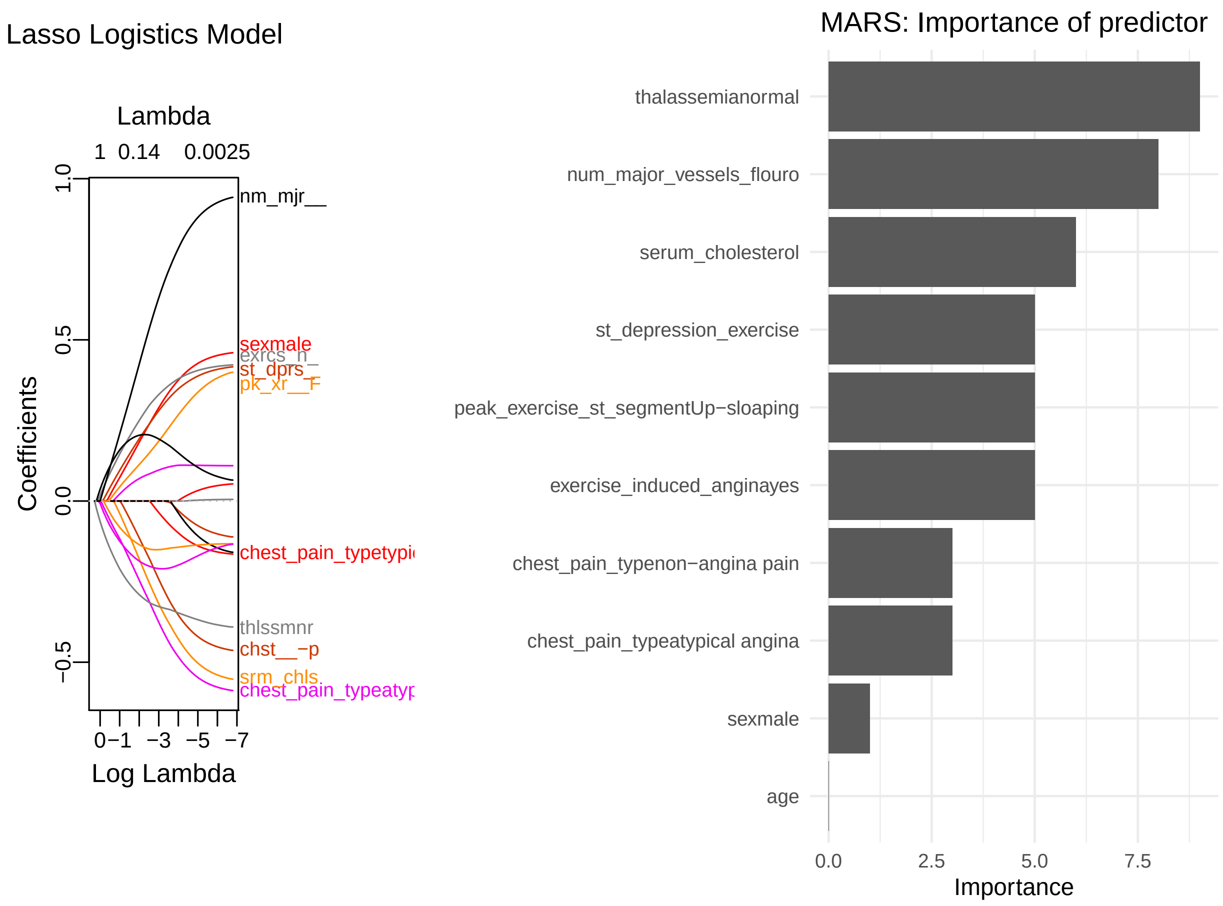
term	coefficient
Intercept	0.333
age	0.097
sexmale	0.296
chest_pain_typeatypical angina	-0.384
chest_pain_typenon-angina pain	-0.251
chest_pain_typetypical angina	-0.040
resting_blood_pressure	0.000
serum_cholesterol	-0.326
fasting_blood_sugarfasting blood sugar > 120 mg/dl	0.000
resting_ecgnormal	0.000

term	coefficient
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

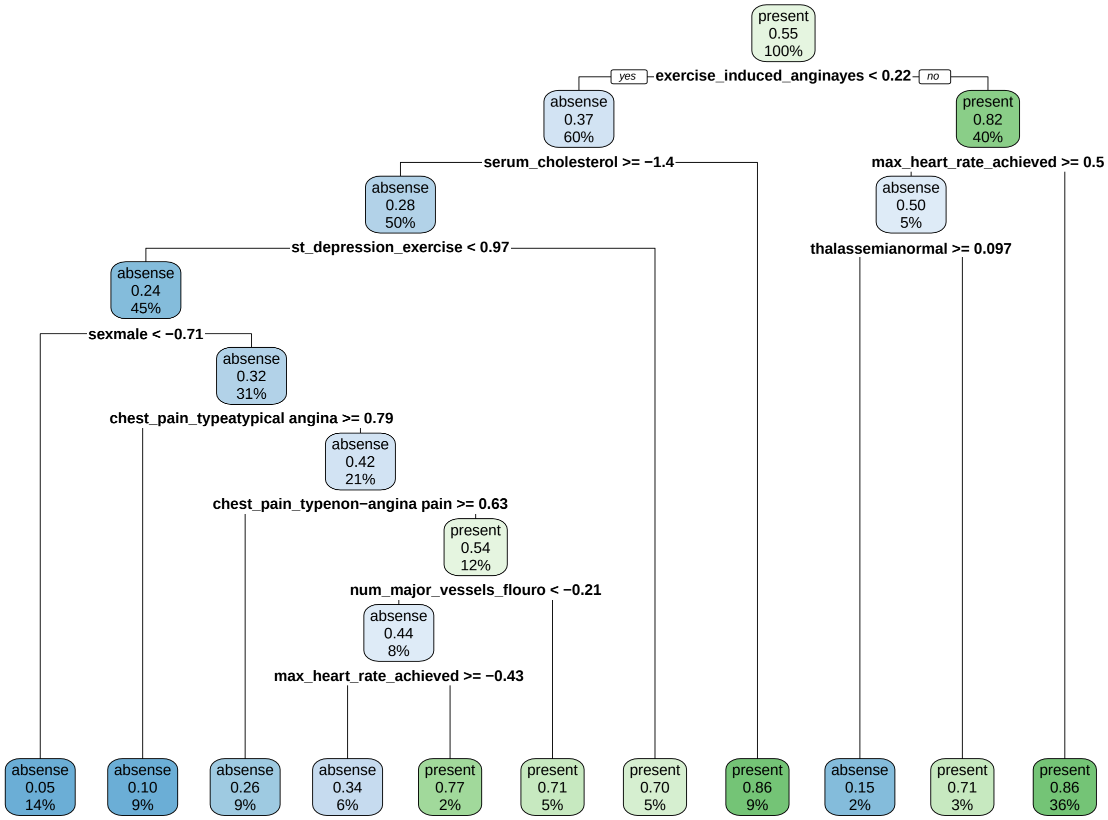
```
p_mv = vip::vip(mars_model$finalModel) + labs(title = "MARS: Importance of predictor")

p_ll = ggplotify::as.ggplot(~plot_glmnet(logistic_model$finalModel))
p_ll = p_ll + labs(title = "Lasso Logistics Model") +
  xlim(c(0.1,1))+
  ylim(c(0.1,1))

(p_logistics / p_tree) | (p_mars/p_knn)
```



```
(p_ll | p_mv)
```



```
rpart.plot::rpart.plot(tree_model$finalModel)
```

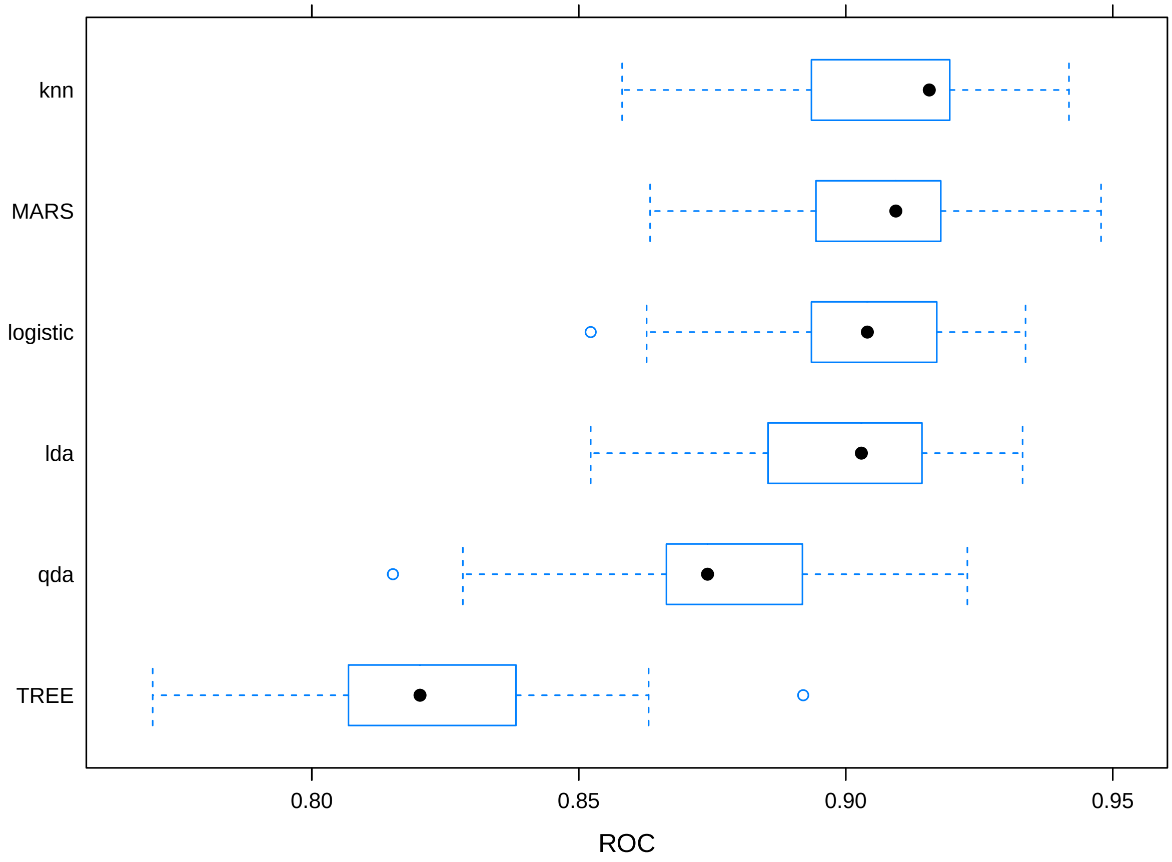


Performance comparison

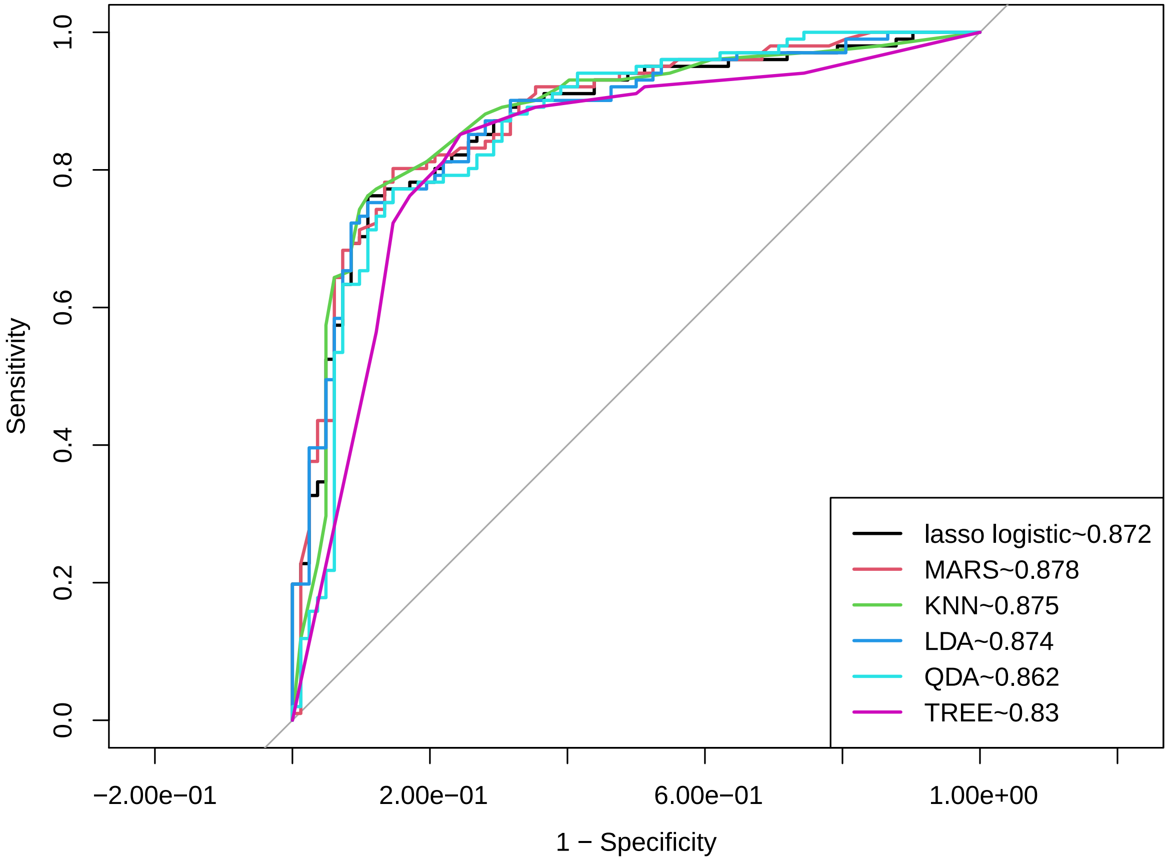
```
rsmp = resamples(
  list(
    logistic = logistic_model,
    MARS = mars_model,
    knn = knn_model,
    lda = lda_model,
    qda = qda_model,
    TREE = tree_model
  ),
  metric = c("ROC", "Kappa")
)

summary(rsmp)
```

```
##
## 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
```

```r
bwplot(rsmp,metric = "ROC")
```



```
ROC =
  expand.grid(
    test_X = list(X_ts),
    test_Y = list(Y_ts),
    model = list(logistic_model, mars_model, knn_model, lda_model, qda_model,tree_model)
  ) %>%
  mutate(
    pred = map2(model, test_X,  ~ predict(.x, newdata = .y, type = "prob")[, 2]),
    roc = map2(test_Y, pred,  ~ pROC::roc(.x, .y))
  ) %>%
  pull(roc)

auc = c()

for (i in 1:6){
  auc = append(auc,ROC[[i]]$auc[1])
  plot(ROC[[i]],col = i, add = T * (i>1), legacy.axes = T * (i==1))
}

model_name =
  c("lasso logistic","MARS","KNN","LDA","QDA","TREE")

legend("bottomright",
       legend = paste0(model_name,"~",round(auc,3)),col=1:6,lwd=2)
```


Sensitivity
−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