

Predicting Diagnosis of Liver Disease

```
# Load libraries
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.3      v purrr  0.3.4
## v tibble  3.0.6      v dplyr  1.0.4
## v tidyr   1.1.2      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(dplyr)
library(ISLR)
library(janitor)

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##   chisq.test, fisher.test

library(AppliedPredictiveModeling)
library(caret)

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##   lift

library(corrplot)

## corrplot 0.84 loaded
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
```

```
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
##      cov, smooth, var
```

```
library(MASS)
```

```
##
```

```
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

```
##      select
```

```
library(readxl)
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
```

```
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
```

```
##
```

```
##      expand, pack, unpack
```

```
## Loaded glmnet 4.1
```

```
library(mlbench)
```

```
library(pdp)
```

```
##
```

```
## Attaching package: 'pdp'
```

```
## The following object is masked from 'package:purrr':
```

```
##
```

```
##      partial
```

```
library(vip)
```

```
##
```

```
## Attaching package: 'vip'
```

```
## The following object is masked from 'package:utils':
```

```
##
```

```
##      vi
```

```
library(klaR)
```

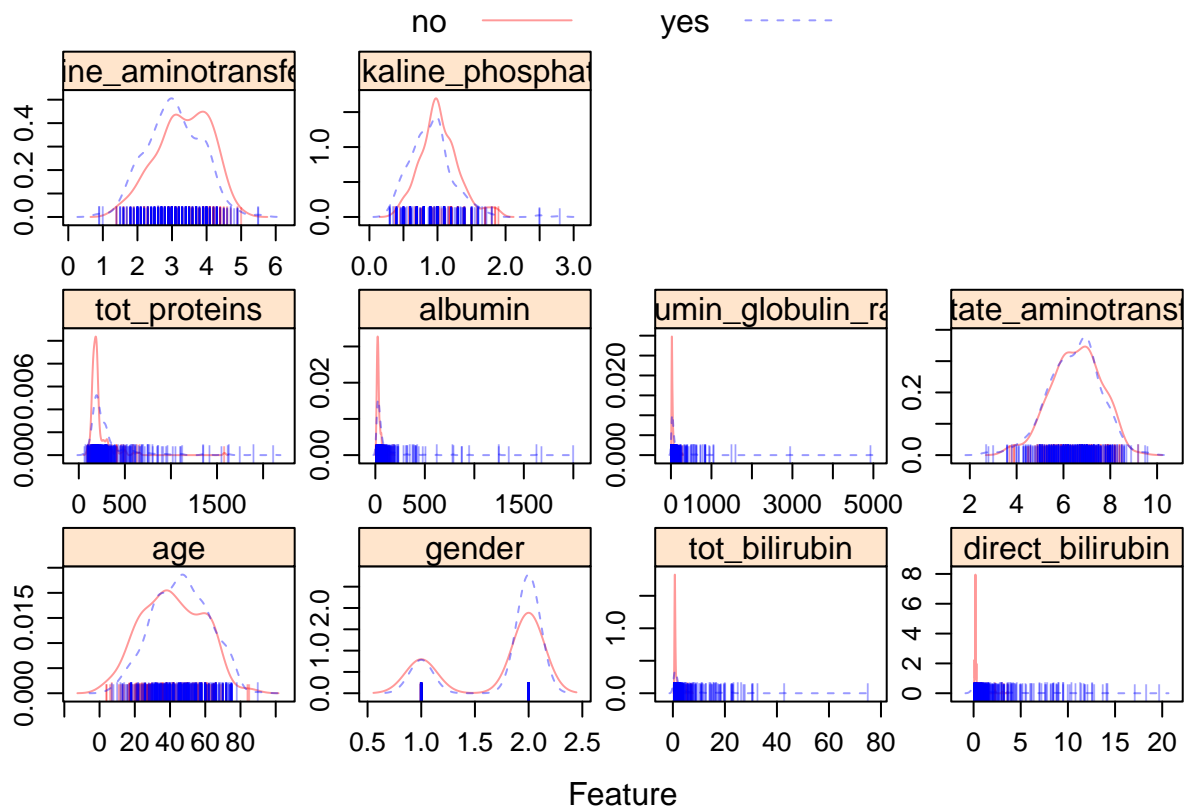
Import data

```
liver_df = read_excel("./data/liver.xlsx") %>%
  mutate(outcome = ifelse(is_patient == 1, "yes", "no"), outcome = as.factor(outcome)) %>%
  dplyr::select(-is_patient) %>%
  clean_names %>%
  rename(
    aspartate_aminotransferase = sgpt,
    alamine_aminotransferase = sgot,
    albumin_globulin_ratio = ag_ratio,
    alkaline_phosphate = alkphos) %>%
  drop_na

liver_df$gender=factor(x=liver_df$gender,levels = c('Female','Male'),labels=c(0, 1))
liver_df$gender = as.double(liver_df$gender)
# female = '1', male = '2'
```

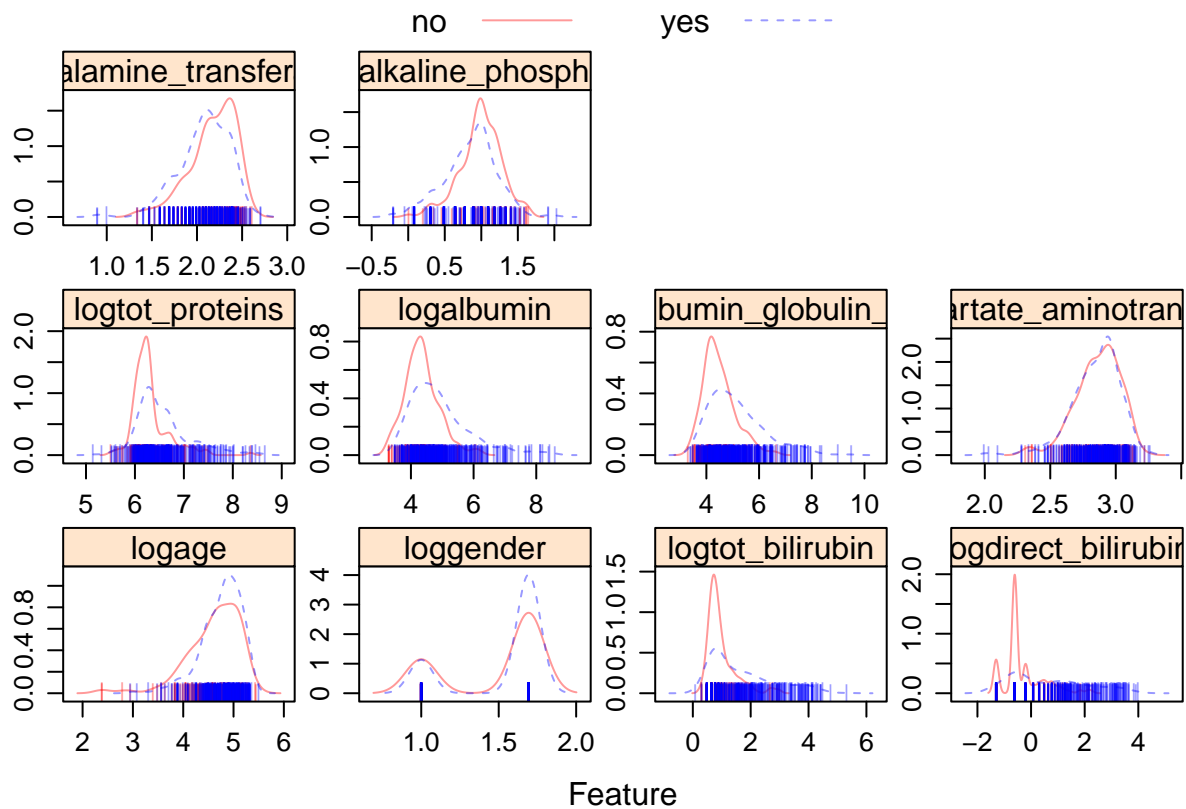
Exploratory Data Analysis

```
# Feature plots
theme1 <- transparentTheme(trans = .4)
trellis.par.set(theme1)
featurePlot(x = liver_df[, 1:10],
  y = liver_df$outcome,
  scales = list(x = list(relation = "free"),
    y = list(relation = "free")),
  plot = "density", pch = "|",
  auto.key = list(columns = 2))
```

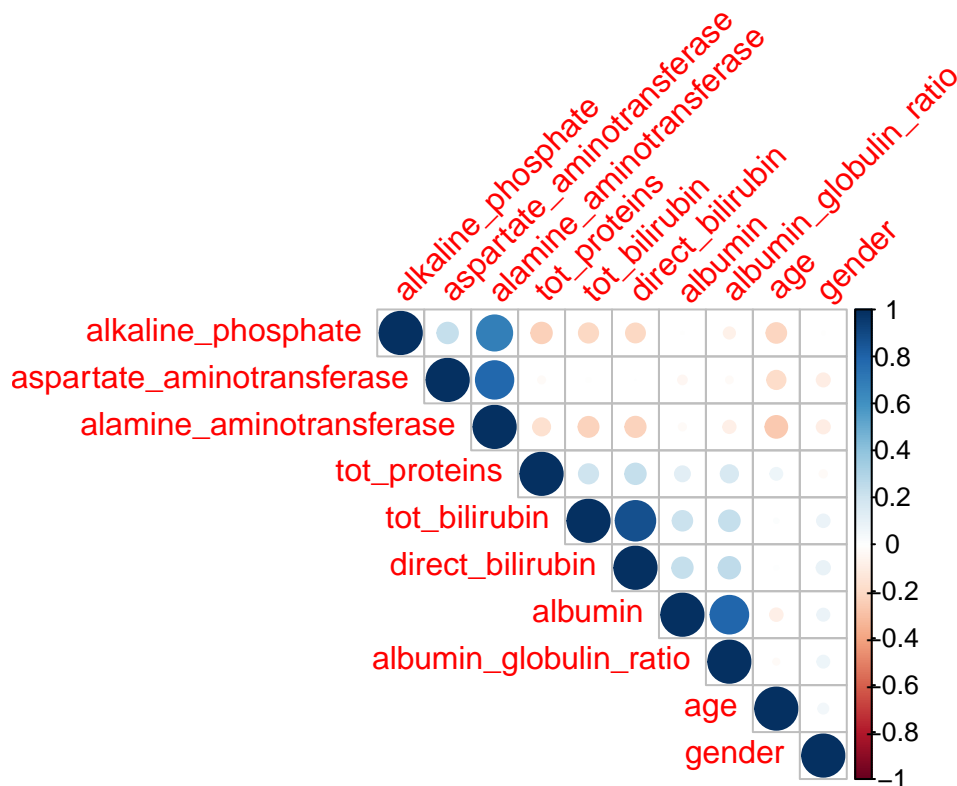


```
# dataset with all the log-transformed predictor variables
liver_df1 =
  liver_df %>%
  mutate(logtot_bilirubin = log(tot_bilirubin) +1,
         logdirect_bilirubin = log(direct_bilirubin) +1,
         logtot_proteins = log(tot_proteins) +1,
         logalbumin = log(albumin) +1,
         loggender = log(gender) +1,
         logalbumin_globulin_ratio = log(albumin_globulin_ratio) +1,
         logage = log(age) +1,
         logaspartate_aminotransferase = log(aspartate_aminotransferase) +1,
         logalamine_transferase = log(alamine_aminotransferase ) +1,
         logalkaline_phosphate = log(alkaline_phosphate) +1) %>%
  dplyr::select(logage, loggender, logtot_bilirubin, logdirect_bilirubin, logtot_proteins, logalbumin,
               logalbumin_globulin_ratio, logaspartate_aminotransferase, logalamine_transferase,
               logalkaline_phosphate, outcome)

theme1 <- transparentTheme(trans = .4)
trellis.par.set(theme1)
featurePlot(x = liver_df1[, 1:10],
            y = liver_df1$outcome,
            scales = list(x = list(relation = "free"),
                          y = list(relation = "free")),
            plot = "density", pch = "|",
            auto.key = list(columns = 2))
```



```
# Correlation plot
corrplot(cor(liver_df[, -11]), tl.srt = 45, order = 'hclust', type = 'upper')
```



```
table(liver_df$outcome)
```

```
##  
## no yes  
## 165 414
```

Data Partition

```
#Create Training and Test Datasets  
set.seed(10)  
dim(liver_df1)
```

```
## [1] 579 11
```

```
rowTrain <- createDataPartition(y = liver_df1$outcome,  
                                p = 0.8,  
                                list = FALSE)
```

Logistic Regression

```
#Fit Logistic Regression Model with all predictors  
glm.fit <- glm(outcome ~ ., data = liver_df1,  
               subset = rowTrain,  
               family = binomial(link = "logit"))  
summary(glm.fit)
```

```
##  
## Call:  
## glm(formula = outcome ~ ., family = binomial(link = "logit"),  
##      data = liver_df1, subset = rowTrain)  
##  
## Deviance Residuals:  
##      Min       1Q   Median       3Q      Max  
## -2.8567  -1.0195   0.4202   0.8414   1.6584  
##  
## Coefficients:  
##  
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept)    -15.11618     4.47123  -3.381 0.000723 ***  
## logage           0.74111     0.26049   2.845 0.004440 **  
## loggender        0.24539     0.37642   0.652 0.514469  
## logtot_bilirubin  0.54443     0.53800   1.012 0.311554  
## logdirect_bilirubin -0.04189     0.35933  -0.117 0.907183  
## logtot_proteins   0.31927     0.30501   1.047 0.295219  
## logalbumin       0.75082     0.28262   2.657 0.007893 **  
## logalbumin_globulin_ratio 0.13377     0.25062   0.534 0.593500  
## logaspartate_aminotransferase 5.03133     2.53732   1.983 0.047375 *
```

```
## logalamine_transferase      -4.88216      2.53278     -1.928 0.053906 .
## logalkaline_phosphate      1.55808      1.26055      1.236 0.216445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 554.14  on 463  degrees of freedom
## Residual deviance: 456.81  on 453  degrees of freedom
## AIC: 478.81
##
## Number of Fisher Scoring iterations: 5
```

Confusion Matrix

```
test.pred.prob <- predict(glm.fit, newdata = liver_df1[-rowTrain,], type = "response")
test.pred <- rep("no", length(test.pred.prob))
test.pred[test.pred.prob > 0.5] <- "yes"
confusionMatrix(data = as.factor(test.pred), reference = liver_df1$outcome[-rowTrain],
                 positive = "yes")
```

Confusion Matrix and Statistics

```
##
##           Reference
## Prediction no yes
##      no    7    5
##      yes  26   77
##
##           Accuracy : 0.7304
##           95% CI : (0.6397, 0.8089)
##      No Information Rate : 0.713
##      P-Value [Acc > NIR] : 0.383747
##
##           Kappa : 0.1866
##
##  Mcnemar's Test P-Value : 0.000328
##
##           Sensitivity : 0.9390
##           Specificity : 0.2121
##           Pos Pred Value : 0.7476
##           Neg Pred Value : 0.5833
##           Prevalence : 0.7130
##           Detection Rate : 0.6696
##      Detection Prevalence : 0.8957
##           Balanced Accuracy : 0.5756
##
##           'Positive' Class : yes
##
```

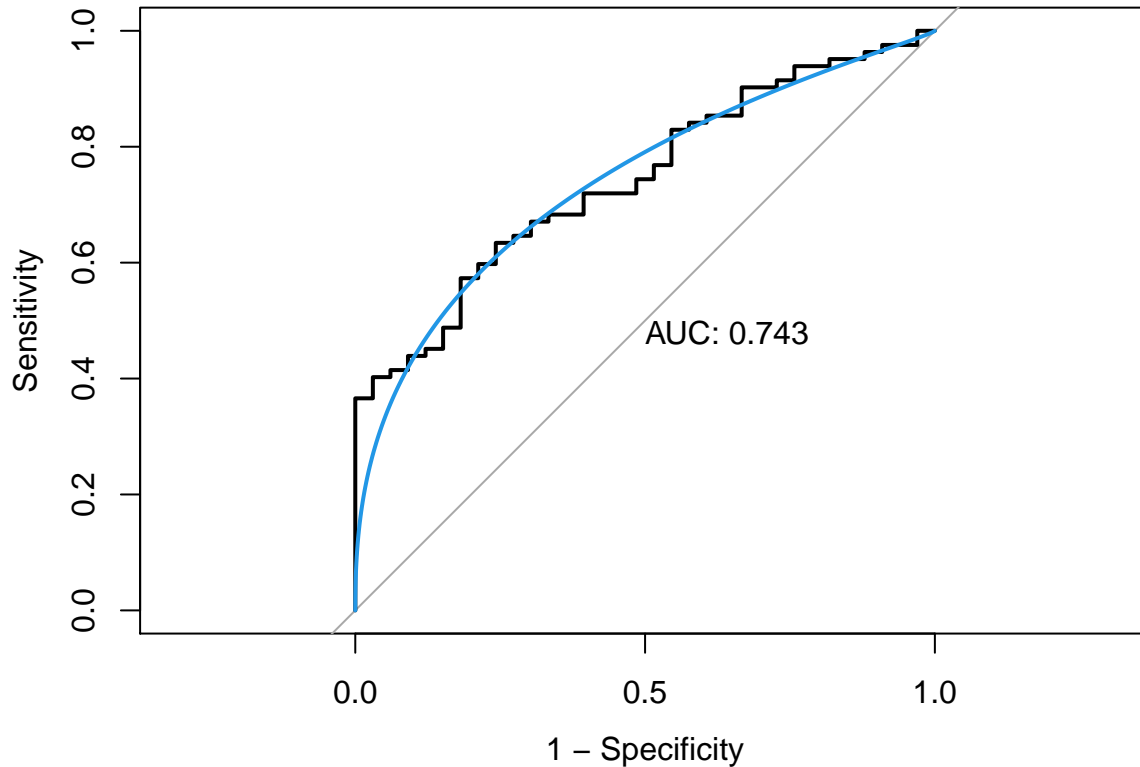
ROC Curve

```
roc.glm <- roc(liver_df1$outcome[-rowTrain], test.pred.prob)
```

```
## Setting levels: control = no, case = yes
```

```
## Setting direction: controls < cases
```

```
plot(roc.glm, legacy.axes = TRUE, print.auc = TRUE)  
plot(smooth(roc.glm), col = 4, add = TRUE)
```



```
# Fit a logistic regression with CARET  
set.seed(10)  
ctrl <- trainControl(method = "repeatedcv", repeats = 10,  
                     summaryFunction = twoClassSummary,  
                     classProbs = TRUE)  
model.glm <- train(x = liver_df1[rowTrain,1:10],  
                  y = liver_df1$outcome[rowTrain],  
                  method = "glm",  
                  preProcess = c("center", "scale"),  
                  metric = "ROC",  
                  trControl = ctrl)
```

```
## Warning: The 'i' argument of '['() can't be a matrix as of tibble 3.0.0.  
## Convert to a vector.  
## This warning is displayed once every 8 hours.  
## Call 'lifecycle::last_warnings()' to see where this warning was generated.
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
## Warning: Setting row names on a tibble is deprecated.
```


[illegible]

[illegible]


```
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
## Warning: Setting row names on a tibble is deprecated.
```

Regularized logistic regression

```
ctrl <- trainControl(method = "repeatedcv",
                     repeats = 5,
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE)

set.seed(10)
glmGrid <- expand.grid(.alpha = seq(0,1,length =6),
                     .lambda = exp(seq(-6,0,length =20)))
model.glmn <- train(x=liver_df1[rowTrain,1:10],
                   y=liver_df1$outcome[rowTrain],
                   method = "glmnet",
                   #preProcess = c("center", "scale"),
                   tuneGrid = glmGrid,
                   metric = "ROC",
                   trControl = ctrl)
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

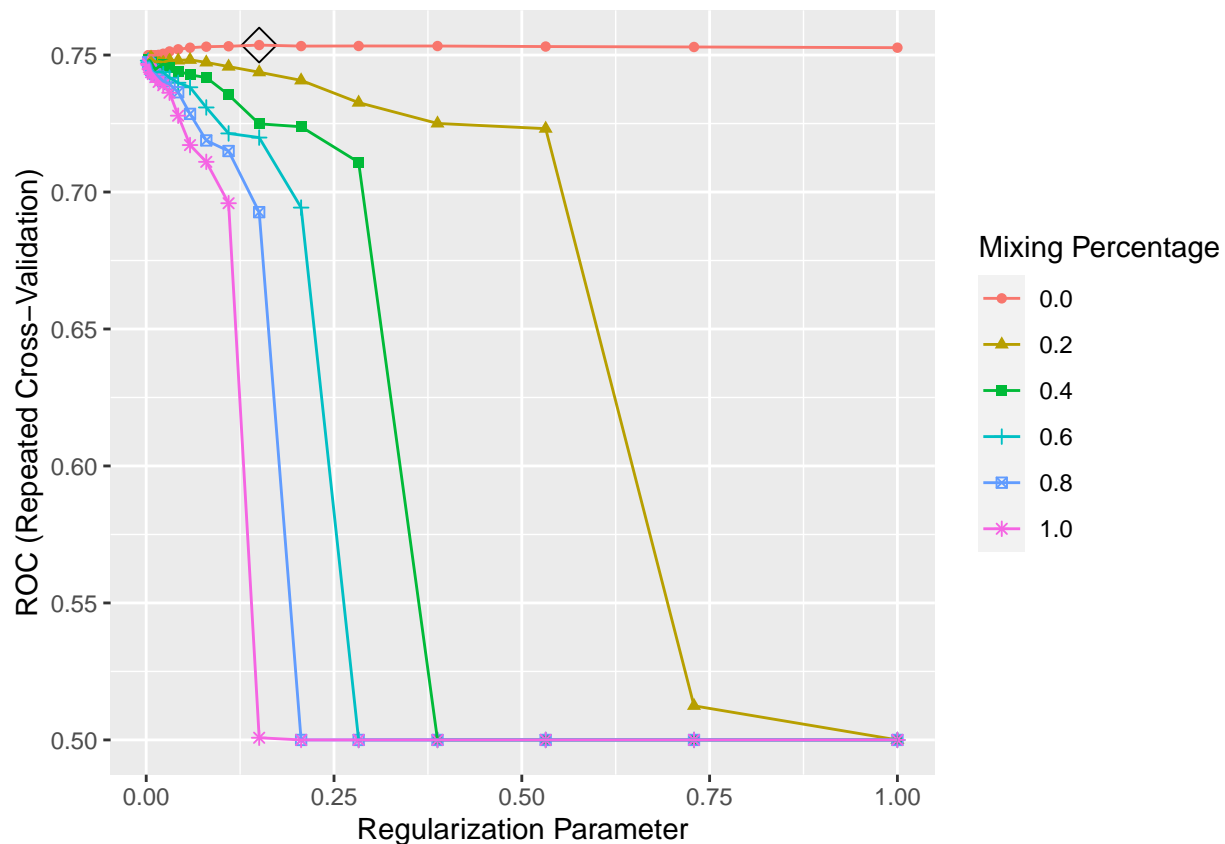
```
## Warning: Setting row names on a tibble is deprecated.
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
ggplot(model.glmn,xTrans = function(x)log(x), highlight = TRUE)
```



```
max(model.glmn$result$ROC)
```

```
## [1] 0.7536379
```

```
model.glmn$bestTune # alpha of 0 indicates a ridge regression.
```

```
##      alpha      lambda  
## 14         0 0.1503579
```

```
coef(model.glmn$finalModel, s = model.glmn$bestTune$lambda)
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"  
##                                     1  
## (Intercept)                      -4.8091934
```


## logage	0.3880113
## loggender	0.2410487
## logtot_bilirubin	0.1845745
## logdirect_bilirubin	0.1342574
## logtot_proteins	0.2520828
## logalbumin	0.2631151
## logalbumin_globulin_ratio	0.1996150
## logaspartate_aminotransferase	0.1483546
## logalamine_transferase	-0.3099717
## logalkaline_phosphate	-0.3247082