BCB 724 Homework 2

Matthew Sutcliffe

2023-12-02

Load libraries and data

```
library(caret)
library(glmnet)
library(pROC)

set.seed(0)

metadata <- read.csv(file = "poore-2020-metadata-subset.csv", stringsAsFactors = TRUE)[, -1]
```

One-hot encode the 6 feature columns

```
binary_features <- model.matrix(object = formula( ~ . - 1), data = metadata[, -1])

dim(binary_features)
```

```
## [1] 18116   217
```

Combine with target “Ovarian Serous Cystadenocarcinoma”

```
x <- cbind(data.frame(
    "OSC" = as.numeric(metadata$disease_type == "Ovarian Serous Cystadenocarcinoma")
    ),binary_features)
```

50/50 train/test split

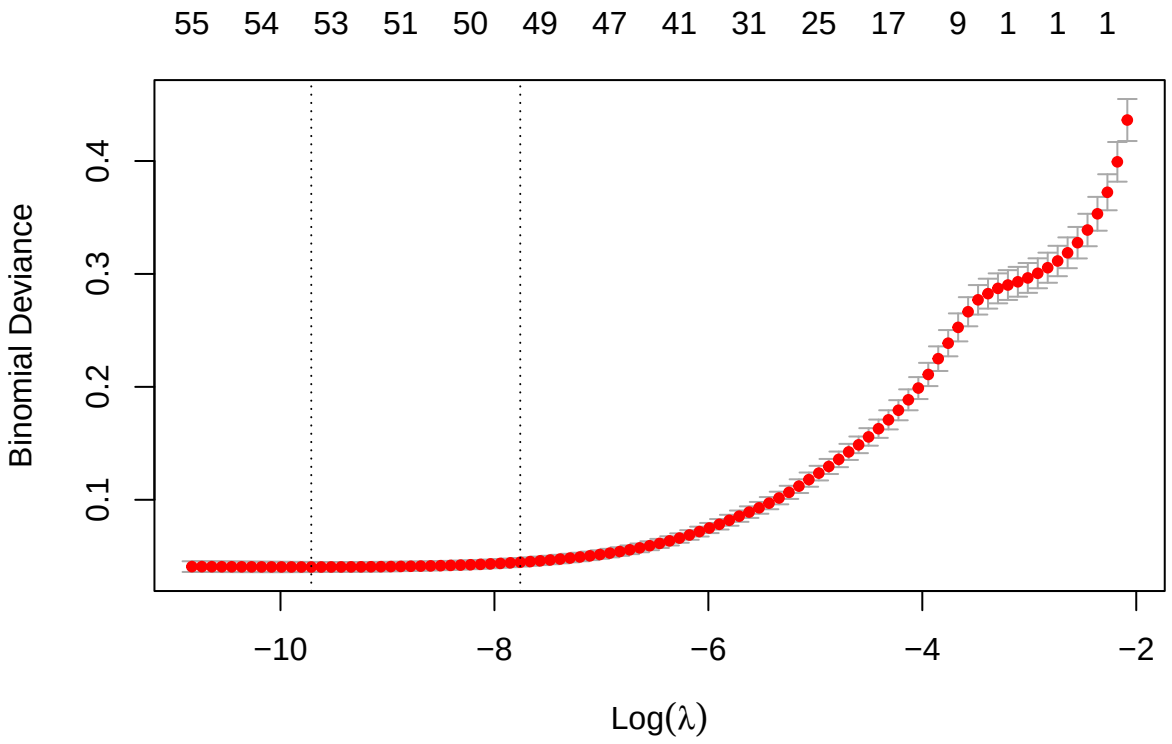
```
trainIndex <- createDataPartition(y = x$OSC, p = 0.5, list = FALSE, times = 1)

x_train <- x[ trainIndex, ]
x_test  <- x[-trainIndex, ]
```

Model Fitting

Fit a multivariate logistic regression model with LASSO penalty.

```
fit <- cv.glmnet(x = as.matrix(x_train[, -1]),
                 y = as.matrix(x_train[,  1, drop = FALSE]),
                 family = "binomial", alpha = 1)
plot(fit)
```



What features does it select?

```
selected_features <- as.matrix(predict(object = fit, s = "lambda.min", type = "coefficients"))[, 1]
selected_features <- selected_features[selected_features != 0] |> sort(decreasing = TRUE)

head(selected_features)
```

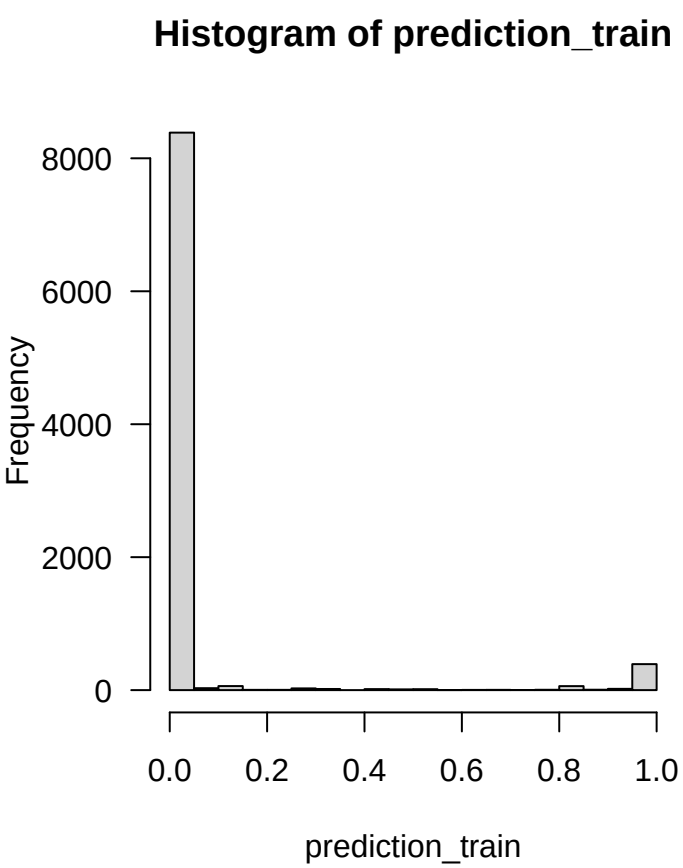
```
##           tissue_source_site_labelBC Cancer Agency
##                                          12.814297
##                      tissue_source_site_labelUCSF
##                                          11.967917
## tissue_source_site_labelGynecologic Oncology Group
##                                          11.108092
##              tissue_source_site_labelCedars Sinai
##                                          10.369788
##           tissue_source_site_labelImperial College
##                                          10.368603
##      tissue_source_site_labelWashington University
##                                           9.967615
```

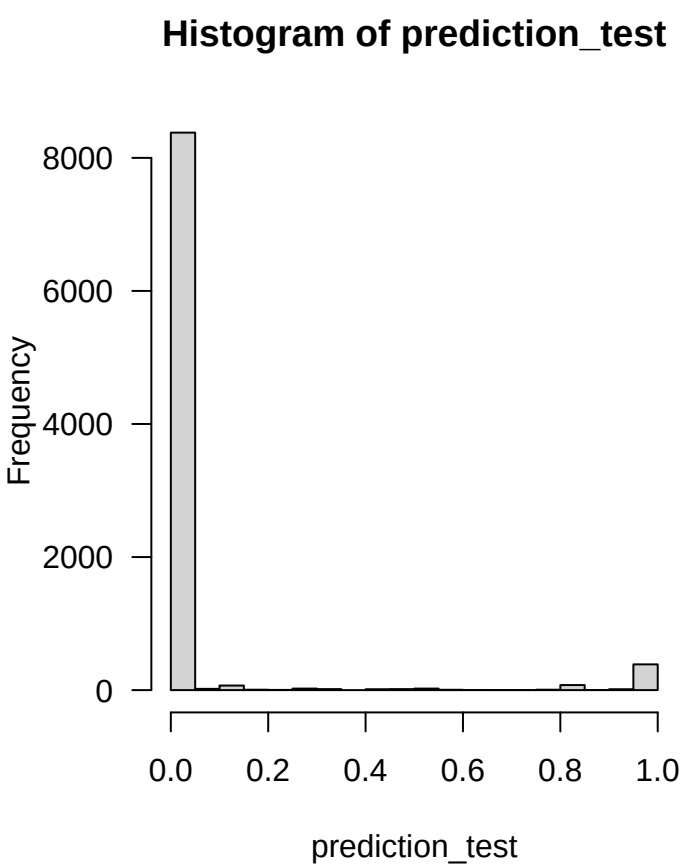
There are 54 features selected. Interestingly, one of the top features include "tissue_source_site_label = Gynecologic Oncology Group", which seems highly relevant for predicting ovarian cancer.

Let's look at the predicted probabilities for the training set and testing set

```
prediction_train <- predict(object = fit,
                            newx = as.matrix(x_train[, -1]),
                            s = "lambda.min",
                            type = "response")[, 1]
prediction_test  <- predict(object = fit,
                            newx = as.matrix(x_test[, -1]),
                            s = "lambda.min",
                            type = "response")[, 1]

par(mfrow = c(1, 2))
hist(prediction_train, las = 1)
hist(prediction_test, las = 1)
```





All the probabilities are very close to 0 or 1, so regardless of whether the model is correct, it seems very confident in the predictions.

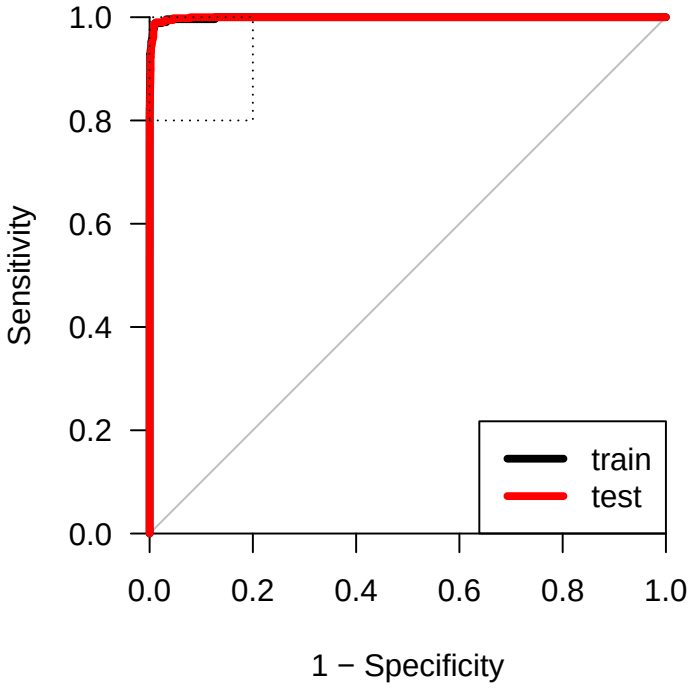
Calculate ROC curves

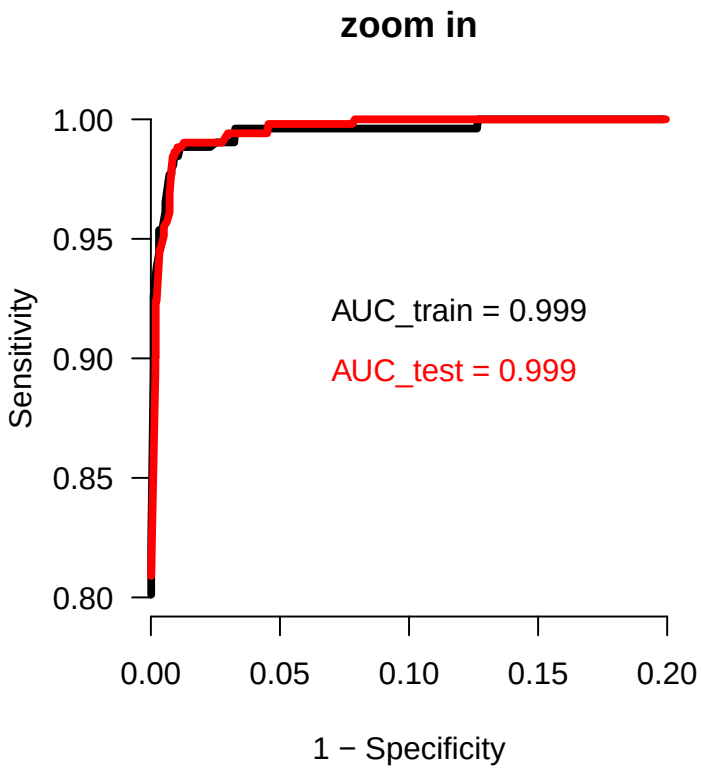
```
x_train$prediction <- prediction_train
x_test$prediction <- prediction_test

roc_train <- roc(OSC ~ prediction, data = x_train)
roc_test <- roc(OSC ~ prediction, data = x_test)

auc_train <- auc(roc_train) |> as.numeric()
auc_test <- auc(roc_test) |> as.numeric()
```

Results


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 − Specificity
train
test


zoom in
1.00
0.95
0.90
0.85
0.80
Sensitivity
0.00
0.05
0.10
0.15
0.20
1 − Specificity
AUC_train = 0.999
AUC_test = 0.999