MEDI 504A: Working with Diabetes Data

Ehsan Karim & Liang Xu, ehsan.karim@ubc.ca

30 November 2021

This document aims to present the general steps for analyzing binary data using machine learning methods. The data source is described in Strack B. et al. [1] link. The dataset can be downloaded from here. Below we present the codes for processing the analytic data following the guideline presented in the paper.

Data Inspection

Coding of predictors

The predictors are coded following the steps outlined in the paper.

Model specifiction

For the purpose of comparison, we select the same set of predictors and interactions as in the paper.

Model Estimation

Outcome and model specification:

```
class(diabetic.data$readmitted)
```

```
## [1] "factor"
```

```
levels(diabetic.data$readmitted)
```

```
## [1] "NO"  "YES"
```

```
model.formula <- as.formula("readmitted ~ discharge + race + source + medical_specialty + time_in_hospi
```

1. Fit a logistic regression model using the above formula and the analytic data `diabetic.data`.

Hint: a) The results should be comparable to the values reported in Table 4 (but may not be exactly the same). b) use `summary()` function to report the fit c) No cross-validation required.

```
lreg_fit <- glm(model.formula, data = diabetic.data,
                family = binomial(link = "logit"))
summary(lreg_fit)
```

```
##
## Call:
## glm(formula = model.formula, family = binomial(link = "logit"),
##     data = diabetic.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9379  -0.4857  -0.3992  -0.3458   2.8224
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       -3.173941   0.176193
## dischargeOther                                     0.292095   0.195068
## raceMissing                                       -0.316646   0.134056
## raceOther                                         -0.295328   0.104848
## raceCaucasian                                      0.015916   0.048438
## sourceOther                                       -0.123723   0.040933
## sourceReferral                                    -0.020914   0.032175
## medical_specialtyFamily/GeneralPractice            0.409992   0.184320
## medical_specialtyInternalMedicine                  0.402032   0.164915
## medical_specialtyMissing or Unknown                0.422134   0.150039
## medical_specialtyOther                             0.288985   0.164017
## medical_specialtySurgery                           0.435112   0.204922
## time_in_hospital                                   0.128699   0.025829
## age< 30                                            1.492980   0.665627
## age[60, 100)                                       0.265849   0.140993
## diag_1Circulatory                                  0.105830   0.105093
## diag_1Respiratory                                 -0.318237   0.120590
## diag_1Digestive                                   -0.064051   0.128546
## diag_1Injury and poisoning                        -0.004056   0.145035
## diag_1Musculoskeletal                             -0.721710   0.178061
## diag_1Genitourinary                               -0.277143   0.150977
## diag_1Neoplasms                                    0.157242   0.165829
## diag_1Other                                        0.028990   0.112182
## A1Cresulthigh_ch                                  -0.391866   0.140098
## A1Cresulthigh_noch                                -0.524933   0.216807
## A1CresultNormal                                    0.006444   0.150850
## dischargeOther:diag_1Circulatory                  -0.026805   0.111298
## dischargeOther:diag_1Respiratory                   0.098943   0.129727
## dischargeOther:diag_1Digestive                     0.024280   0.141330
## dischargeOther:diag_1Injury and poisoning          0.278475   0.148976
## dischargeOther:diag_1Musculoskeletal               0.423824   0.175323
## dischargeOther:diag_1Genitourinary                -0.185280   0.159566
## dischargeOther:diag_1Neoplasms                    -0.166814   0.178869
## dischargeOther:diag_1Other                         0.198855   0.119152
## dischargeOther:raceMissing                         0.285862   0.188412
## dischargeOther:raceOther                           0.496983   0.150673
## dischargeOther:raceCaucasian                       0.016840   0.071921
## dischargeOther:medical_specialtyFamily/GeneralPractice    0.318945   0.180204
## dischargeOther:medical_specialtyInternalMedicine   0.191219   0.164819
## dischargeOther:medical_specialtyMissing or Unknown 0.236041   0.154060
## dischargeOther:medical_specialtyOther              0.374703   0.165871
## dischargeOther:medical_specialtySurgery            0.723663   0.198158
## dischargeOther:time_in_hospital                   -0.027635   0.009256
```

```
## medical_specialtyFamily/GeneralPractice:time_in_hospital -0.061614   0.026133
## medical_specialtyInternalMedicine:time_in_hospital       -0.036593   0.023105
## medical_specialtyMissing or Unknown:time_in_hospital     -0.057010   0.021505
## medical_specialtyOther:time_in_hospital                  -0.051532   0.023610
## medical_specialtySurgery:time_in_hospital                -0.110316   0.029429
## medical_specialtyFamily/GeneralPractice:age< 30          -2.136860   0.844157
## medical_specialtyInternalMedicine:age< 30                -1.660124   0.732318
## medical_specialtyMissing or Unknown:age< 30              -1.108086   0.678110
## medical_specialtyOther:age< 30                           -2.059354   0.701992
## medical_specialtySurgery:age< 30                         -2.841808   1.216233
## medical_specialtyFamily/GeneralPractice:age[60, 100)      0.061838   0.180533
## medical_specialtyInternalMedicine:age[60, 100)           -0.015595   0.162112
## medical_specialtyMissing or Unknown:age[60, 100)         -0.096594   0.147866
## medical_specialtyOther:age[60, 100)                      -0.107182   0.159704
## medical_specialtySurgery:age[60, 100)                    -0.200098   0.196519
## time_in_hospital:diag_1Circulatory                       -0.034196   0.016936
## time_in_hospital:diag_1Respiratory                       -0.007534   0.019708
## time_in_hospital:diag_1Digestive                         -0.034019   0.021961
## time_in_hospital:diag_1Injury and poisoning              -0.042871   0.022531
## time_in_hospital:diag_1Musculoskeletal                    0.022708   0.027858
## time_in_hospital:diag_1Genitourinary                      0.041262   0.025003
## time_in_hospital:diag_1Neoplasms                         -0.047114   0.026554
## time_in_hospital:diag_1Other                             -0.057156   0.018053
## diag_1Circulatory:A1Cresulthigh_ch                        0.543173   0.169401
## diag_1Respiratory:A1Cresulthigh_ch                        0.323641   0.231354
## diag_1Digestive:A1Cresulthigh_ch                          0.509108   0.289201
## diag_1Injury and poisoning:A1Cresulthigh_ch              -0.152828   0.361694
## diag_1Musculoskeletal:A1Cresulthigh_ch                    0.789333   0.361859
## diag_1Genitourinary:A1Cresulthigh_ch                      0.439763   0.307627
## diag_1Neoplasms:A1Cresulthigh_ch                         -0.109459   0.542317
## diag_1Other:A1Cresulthigh_ch                              0.265904   0.204995
## diag_1Circulatory:A1Cresulthigh_noch                      0.516588   0.254539
## diag_1Respiratory:A1Cresulthigh_noch                      0.357910   0.335146
## diag_1Digestive:A1Cresulthigh_noch                        0.187666   0.427307
## diag_1Injury and poisoning:A1Cresulthigh_noch             0.276748   0.485056
## diag_1Musculoskeletal:A1Cresulthigh_noch                  0.793634   0.530661
## diag_1Genitourinary:A1Cresulthigh_noch                   -0.352285   0.632472
## diag_1Neoplasms:A1Cresulthigh_noch                        0.716933   0.653358
## diag_1Other:A1Cresulthigh_noch                            0.639614   0.294052
## diag_1Circulatory:A1CresultNormal                        -0.051316   0.169459
## diag_1Respiratory:A1CresultNormal                        -0.505827   0.212218
## diag_1Digestive:A1CresultNormal                          -0.073048   0.234012
## diag_1Injury and poisoning:A1CresultNormal               -0.596278   0.259305
## diag_1Musculoskeletal:A1CresultNormal                    -0.096769   0.282496
## diag_1Genitourinary:A1CresultNormal                       0.219060   0.246088
## diag_1Neoplasms:A1CresultNormal                           0.243043   0.306575
## diag_1Other:A1CresultNormal                              -0.070059   0.185184
##                                                          z value Pr(>|z|)
## (Intercept)                                              -18.014  < 2e-16 ***
## dischargeOther                                             1.497 0.134289
## raceMissing                                               -2.362 0.018175 *
## raceOther                                                 -2.817 0.004851 **
## raceCaucasian                                              0.329 0.742471
## sourceOther                                               -3.023 0.002506 **
```

```
## sourceReferral                                          -0.650 0.515680
## medical_specialtyFamily/GeneralPractice                  2.224 0.026125 *
## medical_specialtyInternalMedicine                        2.438 0.014776 *
## medical_specialtyMissing or Unknown                      2.813 0.004901 **
## medical_specialtyOther                                   1.762 0.078084 .
## medical_specialtySurgery                                 2.123 0.033728 *
## time_in_hospital                                         4.983 6.27e-07 ***
## age< 30                                                  2.243 0.024899 *
## age[60, 100)                                             1.886 0.059355 .
## diag_1Circulatory                                        1.007 0.313927
## diag_1Respiratory                                       -2.639 0.008315 **
## diag_1Digestive                                         -0.498 0.618292
## diag_1Injury and poisoning                              -0.028 0.977688
## diag_1Musculoskeletal                                   -4.053 5.05e-05 ***
## diag_1Genitourinary                                     -1.836 0.066407 .
## diag_1Neoplasms                                          0.948 0.343019
## diag_1Other                                              0.258 0.796087
## A1Cresulthigh_ch                                        -2.797 0.005157 **
## A1Cresulthigh_noch                                      -2.421 0.015469 *
## A1CresultNormal                                          0.043 0.965925
## dischargeOther:diag_1Circulatory                        -0.241 0.809679
## dischargeOther:diag_1Respiratory                         0.763 0.445641
## dischargeOther:diag_1Digestive                           0.172 0.863597
## dischargeOther:diag_1Injury and poisoning                1.869 0.061587 .
## dischargeOther:diag_1Musculoskeletal                     2.417 0.015633 *
## dischargeOther:diag_1Genitourinary                      -1.161 0.245583
## dischargeOther:diag_1Neoplasms                          -0.933 0.351025
## dischargeOther:diag_1Other                               1.669 0.095133 .
## dischargeOther:raceMissing                               1.517 0.129213
## dischargeOther:raceOther                                 3.298 0.000972 ***
## dischargeOther:raceCaucasian                             0.234 0.814869
## dischargeOther:medical_specialtyFamily/GeneralPractice   1.770 0.076743 .
## dischargeOther:medical_specialtyInternalMedicine         1.160 0.245975
## dischargeOther:medical_specialtyMissing or Unknown       1.532 0.125488
## dischargeOther:medical_specialtyOther                    2.259 0.023883 *
## dischargeOther:medical_specialtySurgery                  3.652 0.000260 ***
## dischargeOther:time_in_hospital                         -2.986 0.002830 **
## medical_specialtyFamily/GeneralPractice:time_in_hospital -2.358 0.018389 *
## medical_specialtyInternalMedicine:time_in_hospital      -1.584 0.113255
## medical_specialtyMissing or Unknown:time_in_hospital    -2.651 0.008024 **
## medical_specialtyOther:time_in_hospital                 -2.183 0.029065 *
## medical_specialtySurgery:time_in_hospital               -3.749 0.000178 ***
## medical_specialtyFamily/GeneralPractice:age< 30         -2.531 0.011362 *
## medical_specialtyInternalMedicine:age< 30               -2.267 0.023394 *
## medical_specialtyMissing or Unknown:age< 30             -1.634 0.102242
## medical_specialtyOther:age< 30                          -2.934 0.003351 **
## medical_specialtySurgery:age< 30                        -2.337 0.019462 *
## medical_specialtyFamily/GeneralPractice:age[60, 100)     0.343 0.731953
## medical_specialtyInternalMedicine:age[60, 100)          -0.096 0.923362
## medical_specialtyMissing or Unknown:age[60, 100)        -0.653 0.513593
## medical_specialtyOther:age[60, 100)                     -0.671 0.502139
## medical_specialtySurgery:age[60, 100)                   -1.018 0.308577
## time_in_hospital:diag_1Circulatory                      -2.019 0.043478 *
## time_in_hospital:diag_1Respiratory                      -0.382 0.702257
```

```
## time_in_hospital:diag_1Digestive                              -1.549 0.121367
## time_in_hospital:diag_1Injury and poisoning                   -1.903 0.057079 .
## time_in_hospital:diag_1Musculoskeletal                         0.815 0.415000
## time_in_hospital:diag_1Genitourinary                           1.650 0.098886 .
## time_in_hospital:diag_1Neoplasms                              -1.774 0.076010 .
## time_in_hospital:diag_1Other                                  -3.166 0.001545 **
## diag_1Circulatory:A1Cresulthigh_ch                             3.206 0.001344 **
## diag_1Respiratory:A1Cresulthigh_ch                             1.399 0.161844
## diag_1Digestive:A1Cresulthigh_ch                               1.760 0.078341 .
## diag_1Injury and poisoning:A1Cresulthigh_ch                   -0.423 0.672635
## diag_1Musculoskeletal:A1Cresulthigh_ch                         2.181 0.029159 *
## diag_1Genitourinary:A1Cresulthigh_ch                           1.430 0.152851
## diag_1Neoplasms:A1Cresulthigh_ch                              -0.202 0.840045
## diag_1Other:A1Cresulthigh_ch                                   1.297 0.194587
## diag_1Circulatory:A1Cresulthigh_noch                           2.030 0.042407 *
## diag_1Respiratory:A1Cresulthigh_noch                           1.068 0.285556
## diag_1Digestive:A1Cresulthigh_noch                             0.439 0.660529
## diag_1Injury and poisoning:A1Cresulthigh_noch                  0.571 0.568305
## diag_1Musculoskeletal:A1Cresulthigh_noch                       1.496 0.134769
## diag_1Genitourinary:A1Cresulthigh_noch                        -0.557 0.577529
## diag_1Neoplasms:A1Cresulthigh_noch                             1.097 0.272509
## diag_1Other:A1Cresulthigh_noch                                 2.175 0.029617 *
## diag_1Circulatory:A1CresultNormal                             -0.303 0.762024
## diag_1Respiratory:A1CresultNormal                             -2.384 0.017148 *
## diag_1Digestive:A1CresultNormal                               -0.312 0.754924
## diag_1Injury and poisoning:A1CresultNormal                    -2.300 0.021475 *
## diag_1Musculoskeletal:A1CresultNormal                         -0.343 0.731937
## diag_1Genitourinary:A1CresultNormal                            0.890 0.373376
## diag_1Neoplasms:A1CresultNormal                                0.793 0.427912
## diag_1Other:A1CresultNormal                                   -0.378 0.705192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 42244  on 69972  degrees of freedom
## Residual deviance: 41240  on 69883  degrees of freedom
## AIC: 41420
##
## Number of Fisher Scoring iterations: 6
```

Discrimination

To describe the discriminative ability of the model over different possible cutoffs, we can resort to the receiver operating characteristic (ROC) plot. The area under the ROC curve (AUC) is a popular indicator of how well the model performs with regards to discrimination.

2. Report AUC from ROC.

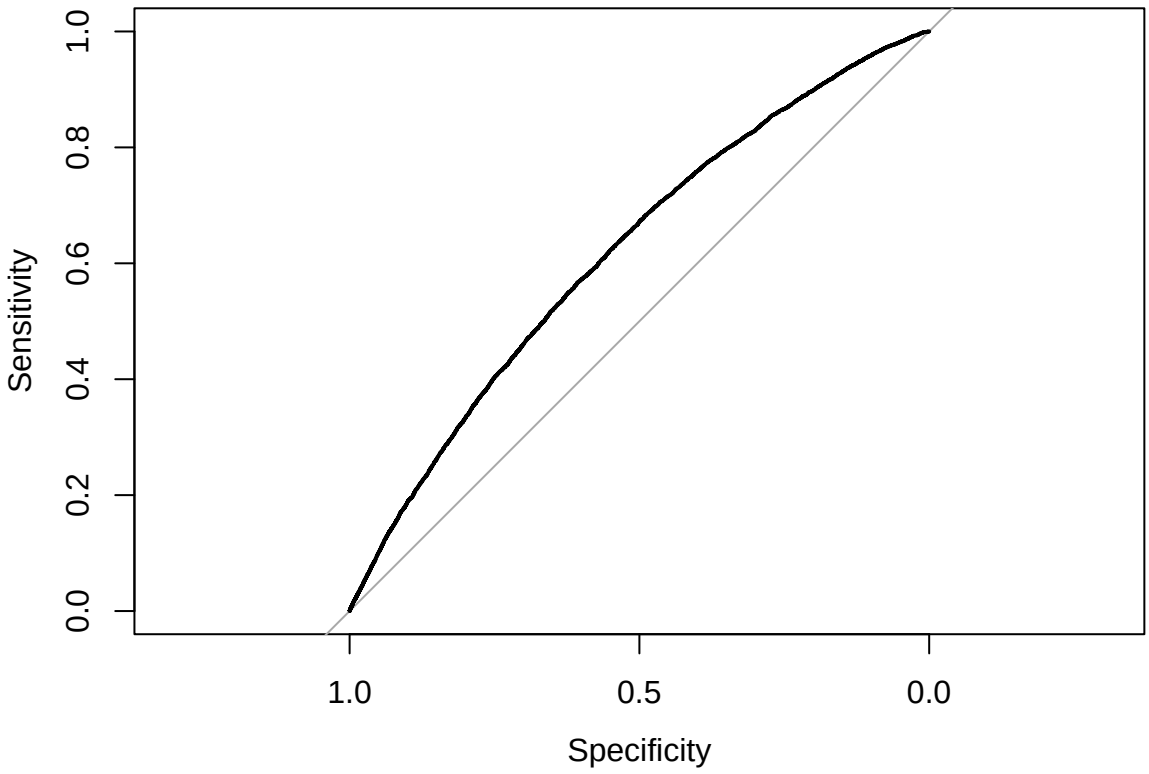
```
require(pROC)
dd.y2 <- diabetic.data$readmitted
prediction.y2 <- predict(lreg_fit, type = "response")
rocobject <- roc(dd.y2, prediction.y2)
```

```
## Setting levels: control = NO, case = YES

## Setting direction: controls < cases
```

```
rocobject
```

```
##
## Call:
## roc.default(response = dd.y2, predictor = prediction.y2)
##
## Data: prediction.y2 in 63696 controls (dd.y2 NO) < 6277 cases (dd.y2 YES).
## Area under the curve: 0.6189
```

```
plot(rocobject)
```



Validation

Previously, we considered measures of performance using the whole dataset, and predictions of the same observations that were used to build the model. For a more realistic assessment of the model's performance, the model should be validated, and there are a couple of options: split-sample validation and K-fold cross-validation (CV).

Cross-validation

3. Set up 10-fold cross-validation, fit logistic regression and obtain AUC from ROC from all the test datasets.

```
## Generalized Linear Model
##
## 69973 samples
##     8 predictor
##     2 classes: 'NO', 'YES'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 62977, 62975, 62975, 62976, 62976, 62975, ...
## Resampling results:
##
##   ROC        Sens  Spec
##   0.6093554  1     0
```

```
##          ROC Sens Spec Resample
## 1  0.5899997    1    0   Fold01
## 2  0.6214484    1    0   Fold02
## 3  0.6252887    1    0   Fold03
## 4  0.6196656    1    0   Fold04
## 5  0.5747153    1    0   Fold05
## 6  0.6185261    1    0   Fold06
## 7  0.6071611    1    0   Fold07
## 8  0.6155007    1    0   Fold08
## 9  0.6049925    1    0   Fold09
## 10 0.6162560    1    0   Fold10
```

Lasso, ridge or elastic net

4. Within 10-fold cross-validation, run the regularized regressions with the following parameter grids: `alpha = c(0,0.5,1), lambda = c(0.25, 0.75)`. Report the best alpha and lambda values the provides best AUC from ROC.

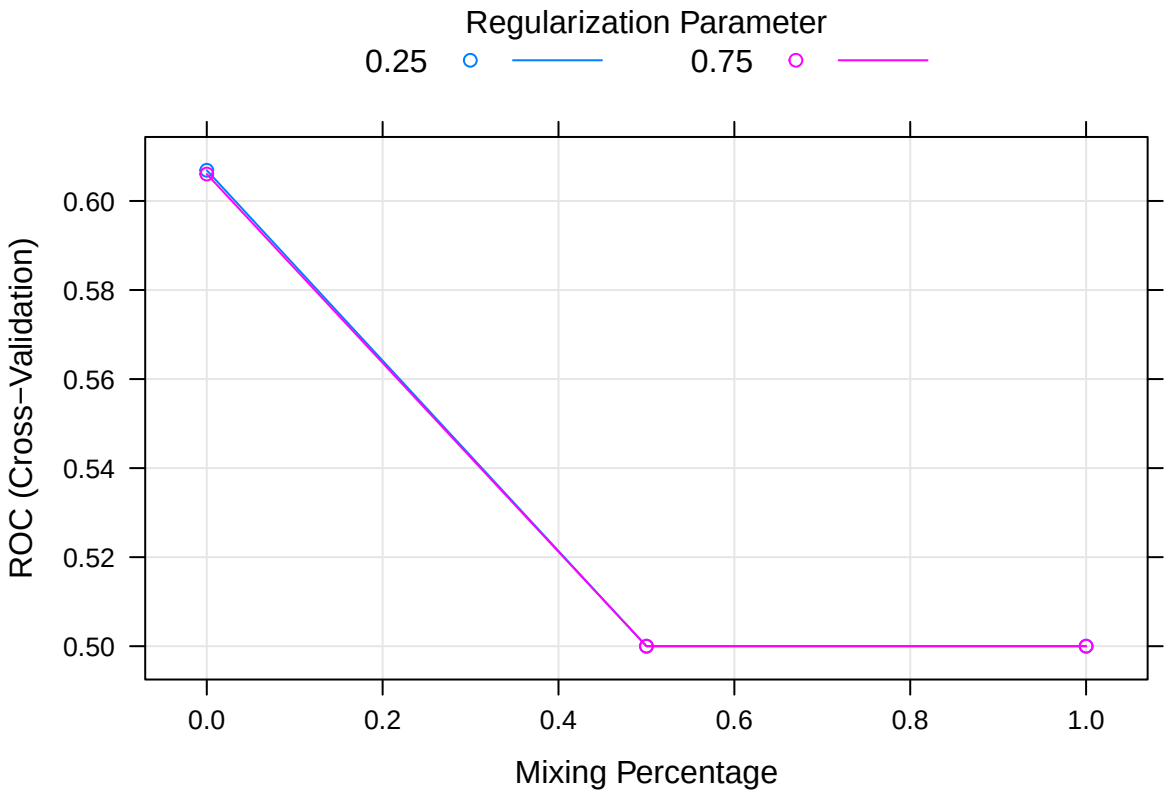
```
require(glmnet)



ctrl1<-trainControl(method = "cv", number = 10,
                    classProbs = TRUE,
                    summaryFunction = twoClassSummary)
fit.cv.bin1<-train(model.formula, trControl = ctrl1,
                data = diabetic.data, method = "glmnet",
                lambda= c(0.25, 0.75),
                tuneGrid = expand.grid(.alpha = c(0,0.5,1),
                                       .lambda = c(0.25, 0.75)),
                verbose = FALSE,
                metric="ROC")
fit.cv.bin1
```

```
## glmnet
##
## 69973 samples
##     8 predictor
##     2 classes: 'NO', 'YES'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 62976, 62976, 62976, 62976, 62975, 62976, ...
## Resampling results across tuning parameters:
##
##   alpha  lambda  ROC        Sens  Spec
##   0.0    0.25    0.6068955  1     0
##   0.0    0.75    0.6060432  1     0
##   0.5    0.25    0.5000000  1     0
##   0.5    0.75    0.5000000  1     0
##   1.0    0.25    0.5000000  1     0
##   1.0    0.75    0.5000000  1     0
##
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 0 and lambda = 0.25.
```

```
#best alpha value is 0, best lambda is 0.25, this is highest ROC which makes it a better model
```

5. Plot the AUC from ROCs for all combinations of parameter grids used in the previous analysis.

```
plot(fit.cv.bin1)
```



Decision Trees

In addition to regression methods, the data can be explored with decision trees (specification of interaction not necessary).

6. Within 10-fold cross-validation, run the regression trees.

```
require(caret)
set.seed(504)
model.formula0 <- as.formula("readmitted ~ discharge + race + source + medical_specialty + time_in_hospi
# your code here

ctrl<-trainControl(method = "cv", number = 10,
                   classProbs = TRUE,
                   summaryFunction = twoClassSummary)

fit.cv.bin1<-train(model.formula0, trControl = ctrl,
                data = diabetic.data, method = "rpart",
               metric="ROC")
fit.cv.bin1
```

```
## CART
##
## 69973 samples
```

```
##     8 predictor
##     2 classes: 'NO', 'YES'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 62977, 62975, 62975, 62976, 62976, 62975, ...
## Resampling results across tuning parameters:
##
##   cp            ROC        Sens       Spec
##   1.448289e-05  0.5841780  0.9991522  0.0006374506
##   1.770131e-05  0.5748267  0.9994819  0.0003187253
##   3.186235e-05  0.5633138  0.9998116  0.0000000000
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was cp = 1.448289e-05.
```

```
summary.res <- fit.cv.bin1$resample
summary.res
```

```
##          ROC      Sens        Spec Resample
## 1  0.6015380 0.9996860 0.000000000   Fold02
## 2  0.5503012 0.9984299 0.003189793   Fold01
## 3  0.5833110 0.9987441 0.001592357   Fold03
## 4  0.5848062 1.0000000 0.000000000   Fold06
## 5  0.5538480 0.9990581 0.000000000   Fold05
## 6  0.5938163 0.9993720 0.000000000   Fold04
## 7  0.5958440 0.9996860 0.000000000   Fold07
## 8  0.6013064 0.9985869 0.000000000   Fold10
## 9  0.5996910 0.9992151 0.000000000   Fold09
## 10 0.5773173 0.9987439 0.001592357   Fold08
```

Variable importance: Report the 5 most important predictor categories.

```
# your code here
caret::varImp(fit.cv.bin1, scale = FALSE)
```

```
## rpart variable importance
##
##   only 20 most important variables shown (out of 30)
##
##                                       Overall
## dischargeOther                         99.145
## time_in_hospital                       70.355
## age[60, 100)                           46.844
## diag_1Respiratory                      19.249
## diag_1Musculoskeletal                  16.106
## diag_1Circulatory                      12.166
## medical_specialtyInternalMedicine       9.803
## diag_1Injury and poisoning              8.648
## medical_specialtySurgery                8.204
## sourceReferral                          7.390
## medical_specialtyFamily/GeneralPractice 5.421
## A1CresultNormal                         4.655
```

```
## sourceOther                                4.536
## diag_1Other                                4.406
## A1Cresulthigh_noch                         3.815
## medical_specialtyMissing or Unknown        3.077
## raceOther                                  3.065
## diag_1Neoplasms                            3.026
## raceMissing                                2.966
## raceCaucasian                              2.352
```

Bagging

7. Within 10-fold cross-validation, run the bagging method.

```
set.seed(504)
require(caret)
ctrl<-trainControl(method = "cv", number = 10,
                   classProbs = TRUE,
                   summaryFunction = twoClassSummary)

fit.cv.bin2<-train(model.formula0, trControl = ctrl,
                data = diabetic.data, method = "bag",
                bagControl = bagControl(fit = ldaBag$fit,
                                        predict = ldaBag$pred,
                                        aggregate = ldaBag$aggregate),
                metric="ROC")
```

```
## Warning: executing %dopar% sequentially: no parallel backend registered
```

```
fit.cv.bin2
```

```
## Bagged Model
##
## 69973 samples
##     8 predictor
##     2 classes: 'NO', 'YES'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 62977, 62975, 62975, 62976, 62976, 62975, ...
## Resampling results:
##
##   ROC        Sens  Spec
##   0.6063624  1     0
##
## Tuning parameter 'vars' was held constant at a value of 25
```

Variable importance: Report the 5 most important predictor categories.

```
# your code here
caret::varImp(fit.cv.bin2, scale = FALSE)
```

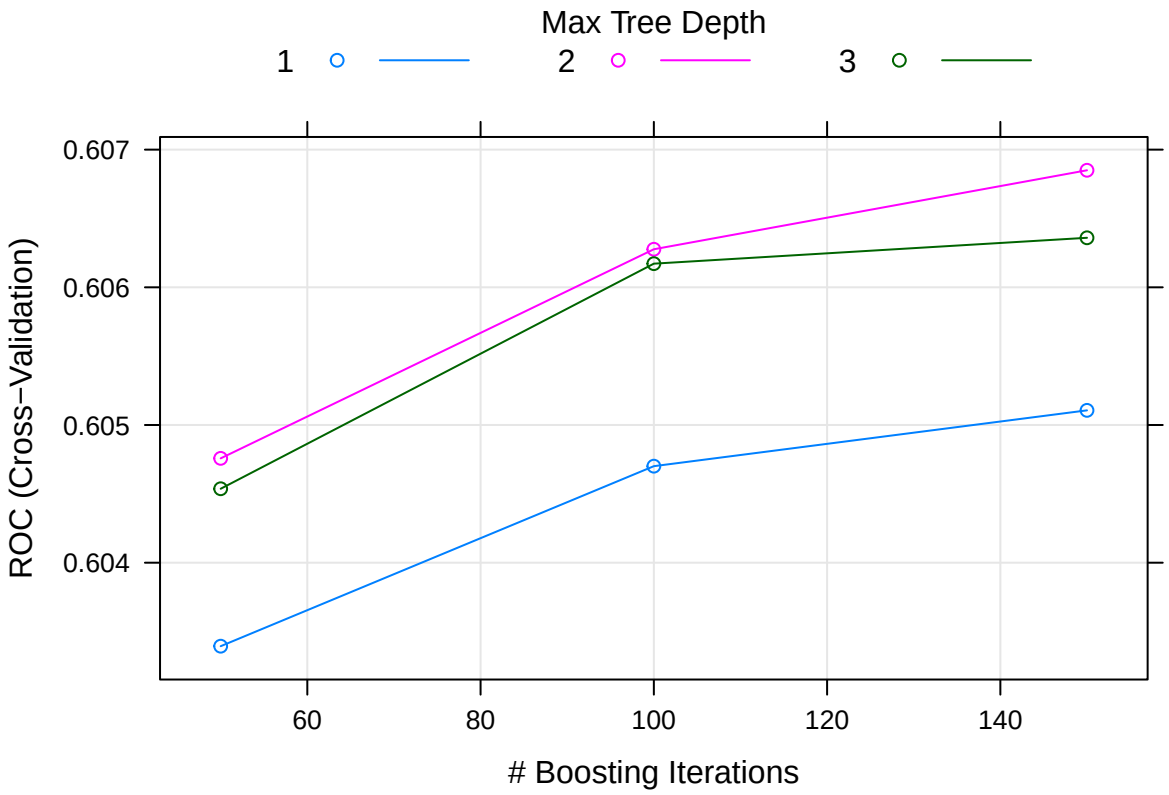
```
## ROC curve variable importance
##
##                   Importance
## discharge             0.5785
## time_in_hospital      0.5599
## age                   0.5379
## medical_specialty     0.5096
## source                0.5094
## race                  0.5088
## A1Cresult             0.5064
## diag_1                0.5055
```

```
#top 5 variables are discharge, time in hospital, age, medical specialty, source
```

Boosting

8. Within 10-fold cross-validation, run the boosting method.

```
# your code here
set.seed(504)
require(caret)
ctrl<-trainControl(method = "cv", number = 5,
                   classProbs = TRUE,
                   summaryFunction = twoClassSummary)

fit.cv.bin3<-train(model.formula0, trControl = ctrl,
               data = diabetic.data, method = "gbm",
               verbose = FALSE,
               metric="ROC")

plot(fit.cv.bin3)
```




```
fit.cv.bin3
```

I can't get fit.cv.bin3 to show up in the PDF: Stochastic Gradient Boosting

69973 samples 8 predictor 2 classes: 'NO', 'YES'

No pre-processing Resampling: Cross-Validated (5 fold) Summary of sample sizes: 55978, 55979, 55977, 55979, 55979 Resampling results across tuning parameters:

interaction.depth n.trees ROC Sens Spec 1 50 0.6033935 1 0
1 100 0.6047011 1 0
1 150 0.6051064 1 0
2 50 0.6047578 1 0
2 100 0.6062765 1 0
2 150 0.6068501 1 0
3 50 0.6045371 1 0
3 100 0.6061718 1 0
3 150 0.6063597 1 0

Tuning parameter 'shrinkage' was held constant at a value of 0.1 Tuning parameter 'n.minobsinnode' was held constant at a value of 10 ROC was used to select the optimal model using the largest value. The final values used for the model were n.trees = 150, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10.

Variable importance: Report the 5 most important predictor categories.

```
caret::varImp(fit.cv.bin3, scale = FALSE)
```

```
## gbm variable importance
##
##   only 20 most important variables shown (out of 25)
##
##                                        Overall
## dischargeOther                         136.215
## time_in_hospital                        49.763
## age[60, 100)                            18.392
## diag_1Respiratory                        9.382
## diag_1Circulatory                        9.375
## medical_specialtyInternalMedicine        7.253
## sourceOther                              6.406
## medical_specialtyFamily/GeneralPractice  4.904
## diag_1Musculoskeletal                    4.281
## raceOther                                4.185
## medical_specialtyOther                   3.950
## A1CresultNormal                          3.819
## medical_specialtySurgery                 3.773
## raceMissing                              3.380
## diag_1Other                              2.808
## A1Cresulthigh_ch                         2.735
## medical_specialtyMissing or Unknown      2.731
## diag_1Digestive                          2.697
## diag_1Genitourinary                      2.656
## A1Cresulthigh_noch                       2.582
```

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records", BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. https://doi.org/10.1155/2014/781670