P8106_MidtermProject

Yiwen Zhao

3/26/2022

Contents

```
library(caret)
library(doBy)
library(glmnet)
library(pROC)
library(pdp)
library(vip)
library(AppliedPredictiveModeling)
library(MASS)
library(klaR)
library(tidyverse)
library(corrplot)
library(earth)
```

Data Cleaning

```
data <- read.csv("Covid19_vacc_predict_handout.csv") %>%
  janitor::clean_names() %>%
  na.omit() %>%
  mutate(covid_vaccination = as.factor(covid_vaccination),
         sex_cd = as.factor(sex_cd),
         lang_spoken_cd = as.factor(lang_spoken_cd)) %>%
  select(id,cons_chmi,est_age,atlas_percapitainc,rwjf_uninsured_adults_pct,
         atlas_type_2015_mining_no,atlas_povertyallagespct,hum_region,
         atlas_hh65plusalonepct, sex_cd, lang_spoken_cd,atlas_pct_sbp15,
         rwjf_resident_seg_black_inx,cons_rxadhm,atlas_medhhinc,cons_lwcm07,
         atlas_low_education_2015_update,race_cd,covid_vaccination)

dat <- data[-c(1, 8, 10, 11)]

dim(dat)
```

```
## [1] 8308   15
```

```
summary(dat)
```

```
##    cons_chmi         est_age       atlas_percapitainc rwjf_uninsured_adults_pct
##  Min.   :  0.00   Min.   : 21.00   Min.   :10399      Min.   :0.02616
##  1st Qu.: 47.00   1st Qu.: 70.00   1st Qu.:23056      1st Qu.:0.08593
##  Median : 62.00   Median : 75.00   Median :26132      Median :0.13357
##  Mean   : 67.21   Mean   : 75.18   Mean   :26685      Mean   :0.13645
##  3rd Qu.: 79.00   3rd Qu.: 81.00   3rd Qu.:28949      3rd Qu.:0.17323
##  Max.   :255.00   Max.   :102.00   Max.   :66522      Max.   :0.43395
##  atlas_type_2015_mining_no atlas_povertyallagespct atlas_hh65plusalonepct
##  Min.   :0.00000           Min.   : 3.40           Min.   : 3.309
##  1st Qu.:0.00000           1st Qu.:11.60           1st Qu.: 9.626
##  Median :0.00000           Median :14.40           Median :10.878
##  Mean   :0.01577           Mean   :14.61           Mean   :10.993
##  3rd Qu.:0.00000           3rd Qu.:17.00           3rd Qu.:12.155
##  Max.   :1.00000           Max.   :45.20           Max.   :19.960
##  atlas_pct_sbp15 rwjf_resident_seg_black_inx  cons_rxadhm    atlas_medhhinc
##  Min.   :1.546   Min.   : 0.2584             Min.   :0.000   Min.   : 22045
##  1st Qu.:3.525   1st Qu.:40.3734             1st Qu.:1.000   1st Qu.: 45813
##  Median :4.110   Median :47.8798             Median :2.000   Median : 51864
##  Mean   :4.559   Mean   :48.1327             Mean   :1.921   Mean   : 53259
##  3rd Qu.:5.710   3rd Qu.:57.8018             3rd Qu.:2.000   3rd Qu.: 58742
##  Max.   :8.160   Max.   :89.6102             Max.   :9.000   Max.   :134609
##   cons_lwcm07      atlas_low_education_2015_update    race_cd
##  Min.   :0.03724   Min.   :0.00000                 Min.   :0.000
##  1st Qu.:0.18190   1st Qu.:0.00000                 1st Qu.:1.000
##  Median :0.22509   Median :0.00000                 Median :1.000
##  Mean   :0.23641   Mean   :0.06488                 Mean   :1.154
##  3rd Qu.:0.28204   3rd Qu.:0.00000                 3rd Qu.:1.000
##  Max.   :0.68722   Max.   :1.00000                 Max.   :6.000
##  covid_vaccination
##  no_vacc:6682
##  vacc   :1626
##
##
##
##
```

```
head(dat)
```

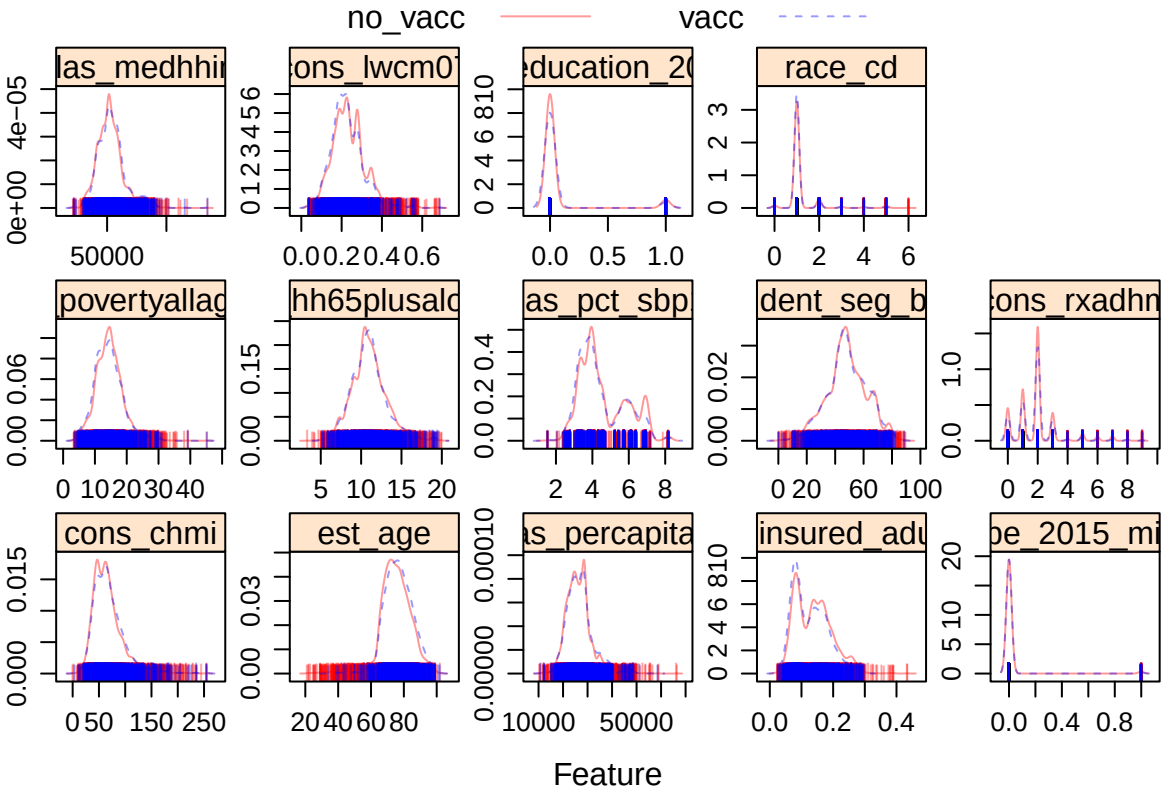
```
##   cons_chmi est_age atlas_percapitainc rwjf_uninsured_adults_pct
## 1        33      69              20228                0.14205338
## 2        53      81              30204                0.09244095
## 3        58      90              22569                0.09353293
## 4        81      77              27377                0.10517024
## 5        72      83              26461                0.08556130
## 6        45      94              38140                0.14642440
##   atlas_type_2015_mining_no atlas_povertyallagespct atlas_hh65plusalonepct
## 1                         0                    15.4              12.842051
## 2                         0                    11.6              11.628669
## 3                         0                    17.2              12.517076
## 4                         0                    17.2              10.404332
## 5                         0                    11.0              11.108695
## 6                         0                     4.6               8.342365
##   atlas_pct_sbp15 rwjf_resident_seg_black_inx cons_rxadhm atlas_medhhinc
## 1        4.539693                    45.89286           5          39631
```

```
## 2        3.206701                    49.13860           0          56439
## 3        3.798339                    67.39506           2          53006
## 4        4.109673                    48.53874           2          51960
## 5        4.024952                    61.37261           2          52736
## 6        4.109673                    41.27621           2          57324
##   cons_lwcm07 atlas_low_education_2015_update race_cd covid_vaccination
## 1     0.28204                               0       2              vacc
## 2     0.17038                               0       1           no_vacc
## 3     0.23359                               0       1           no_vacc
## 4     0.29048                               0       1           no_vacc
## 5     0.18632                               0       0           no_vacc
## 6     0.27305                               0       1           no_vacc
```

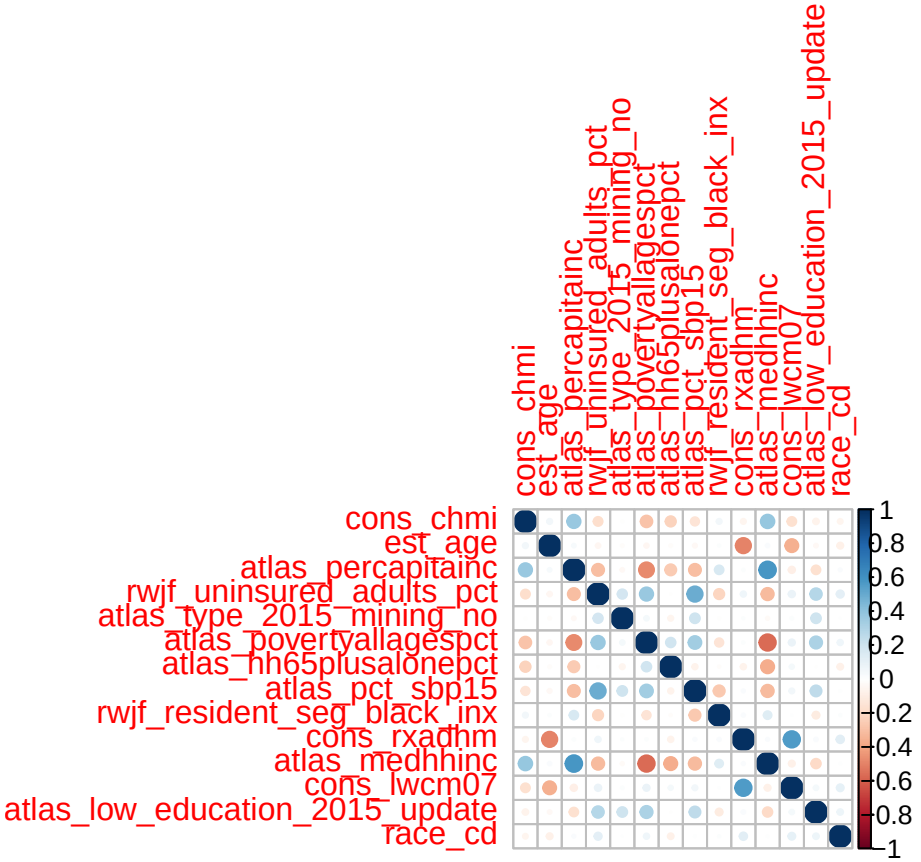
```
set.seed(1)
train <- createDataPartition(y = dat$covid_vaccination,
                             p = 0.7,
                             list = FALSE)
```

Visualization

```
# data visualization
theme1 <- transparentTheme(trans = .4)
trellis.par.set(theme1)

featurePlot(x = dat[, 1:14],
            y = dat$covid_vaccination,
            scales = list(x = list(relation = "free"),
                          y = list(relation = "free")),
            plot = "density", pch = "|",
            auto.key = list(columns = 2))
```




```
corrplot(cor(dat[,1:14]), method = "circle", type = "full")
```



Logistic Regression: GLM

```
# GLM
contrasts(dat$covid_vaccination)
```

```
##         vacc
## no_vacc    0
## vacc       1
```

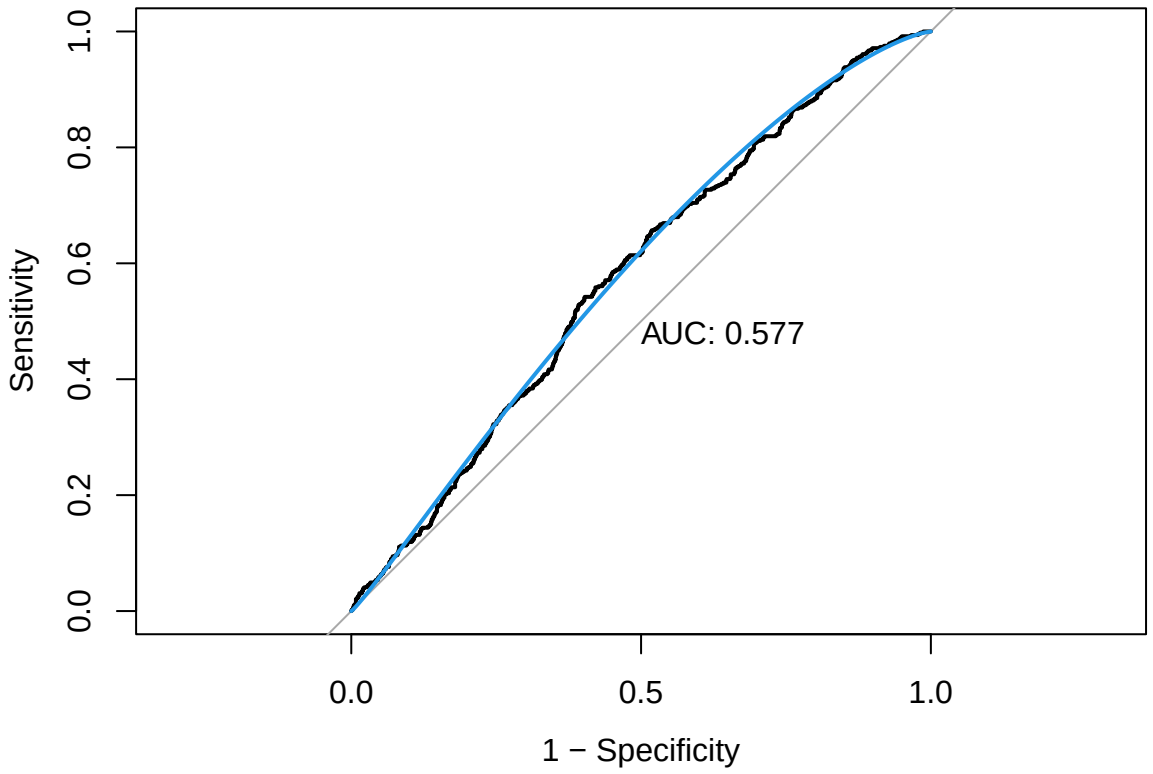
```
glm.fit <- glm(covid_vaccination ~ .,
               data = dat,
               subset = train,
               family = binomial(link = "logit"))

summary(glm.fit)
```

```
##
## Call:
## glm(formula = covid_vaccination ~ ., family = binomial(link = "logit"),
##     data = dat, subset = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9590  -0.6970  -0.6241  -0.5086   2.2682
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.989e+00  6.039e-01  -4.950 7.43e-07 ***
## cons_chmi                        1.598e-03  1.231e-03   1.299 0.193990
## est_age                          2.309e-02  4.115e-03   5.611 2.01e-08 ***
## atlas_percapitainc              -4.374e-06  7.714e-06  -0.567 0.570658
## rwjf_uninsured_adults_pct       -2.612e+00  7.337e-01  -3.560 0.000371 ***
## atlas_type_2015_mining_no       -2.751e-01  3.236e-01  -0.850 0.395266
## atlas_povertyallagespct          8.903e-03  9.839e-03   0.905 0.365550
## atlas_hh65plusalonepct           8.256e-03  1.745e-02   0.473 0.636129
## atlas_pct_sbp15                  3.860e-03  3.235e-02   0.119 0.905027
## rwjf_resident_seg_black_inx     -2.574e-04  2.494e-03  -0.103 0.917817
## cons_rxadhm                      2.888e-02  3.168e-02   0.912 0.362012
## atlas_medhhinc                   5.802e-06  4.050e-06   1.433 0.151956
## cons_lwcm07                     -1.324e+00  4.966e-01  -2.667 0.007655 **
## atlas_low_education_2015_update -1.951e-03  1.563e-01  -0.012 0.990037
## race_cd                         -8.384e-02  5.879e-02  -1.426 0.153861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5753.4  on 5816  degrees of freedom
## Residual deviance: 5659.3  on 5802  degrees of freedom
## AIC: 5689.3
##
## Number of Fisher Scoring iterations: 4
```

```
test.pred.prob <- predict(glm.fit, newdata = dat[-train,],
                          type = "response")
test.pred <- rep("no_vacc", length(test.pred.prob))

confusionMatrix(data = as.factor(test.pred),
                reference = dat$covid_vaccination[-train],
                positive = "vacc")                              #sensitivity is 0
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction no_vacc vacc
##    no_vacc    2004  487
##    vacc          0    0
##
##                Accuracy : 0.8045
##                  95% CI : (0.7884, 0.8199)
##     No Information Rate : 0.8045
##     P-Value [Acc > NIR] : 0.5121
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.0000
##             Specificity : 1.0000
##          Pos Pred Value :    NaN
##          Neg Pred Value : 0.8045
##              Prevalence : 0.1955
##          Detection Rate : 0.0000
##    Detection Prevalence : 0.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : vacc
##
```

```
roc.glm <- roc(dat$covid_vaccination[-train], test.pred.prob)
plot(roc.glm, legacy.axes = TRUE, print.auc = TRUE)
plot(smooth(roc.glm), col = 4, add = TRUE)
```



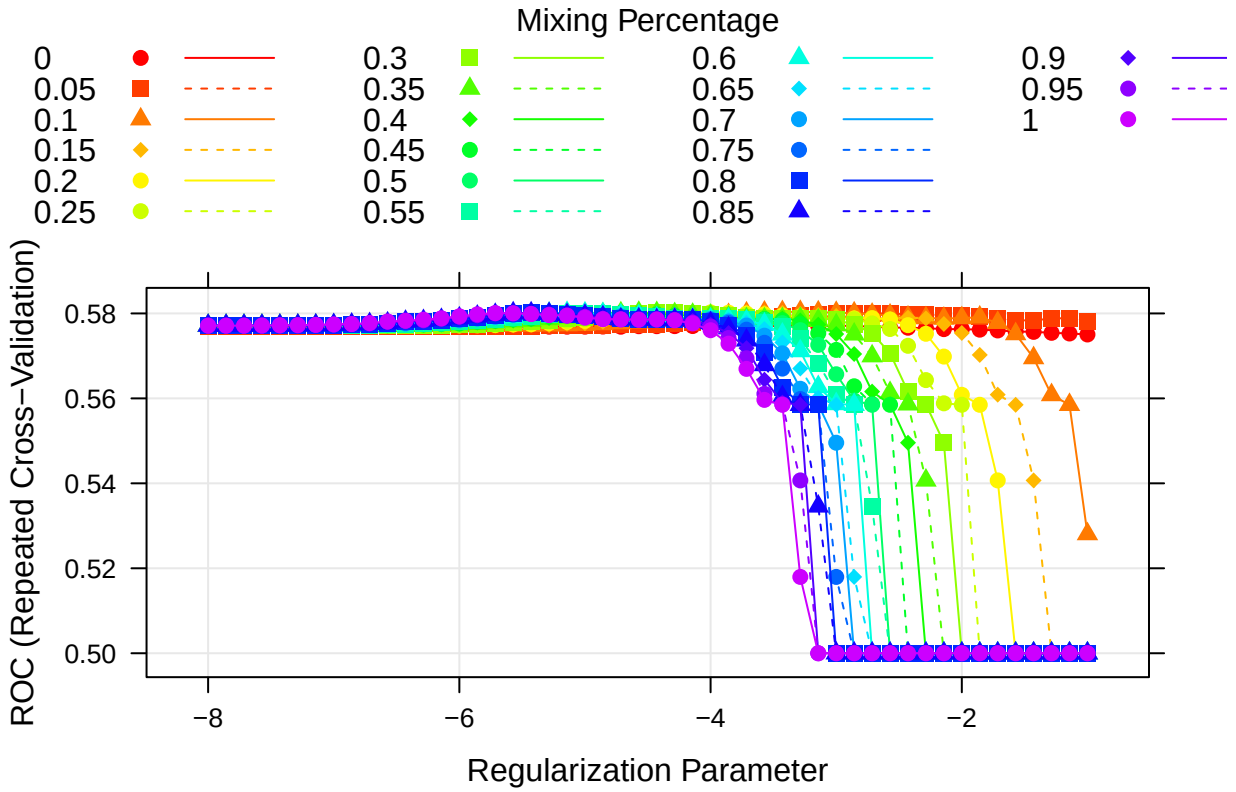

```
# caret
ctrl <- trainControl(method = "repeatedcv",
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE)
set.seed(1)
model.glm <- train(x = dat[train,1:7],
                   y = dat$covid_vaccination[train],
                   method = "glm",
                   metric = "ROC",
                   trControl = ctrl)
```

Penalized Logistic Regression

```
glmnGrid <- expand.grid(.alpha = seq(0, 1, length = 21),
                        .lambda = exp(seq(-8, -1, length = 50)))
set.seed(1)
model.glmn <- train(x = dat[train,1:14],
                    y = dat$covid_vaccination[train],
                    method = "glmnet",
                    tuneGrid = glmnGrid,
                    metric = "ROC",
                    trControl = ctrl)

model.glmn$bestTune
```

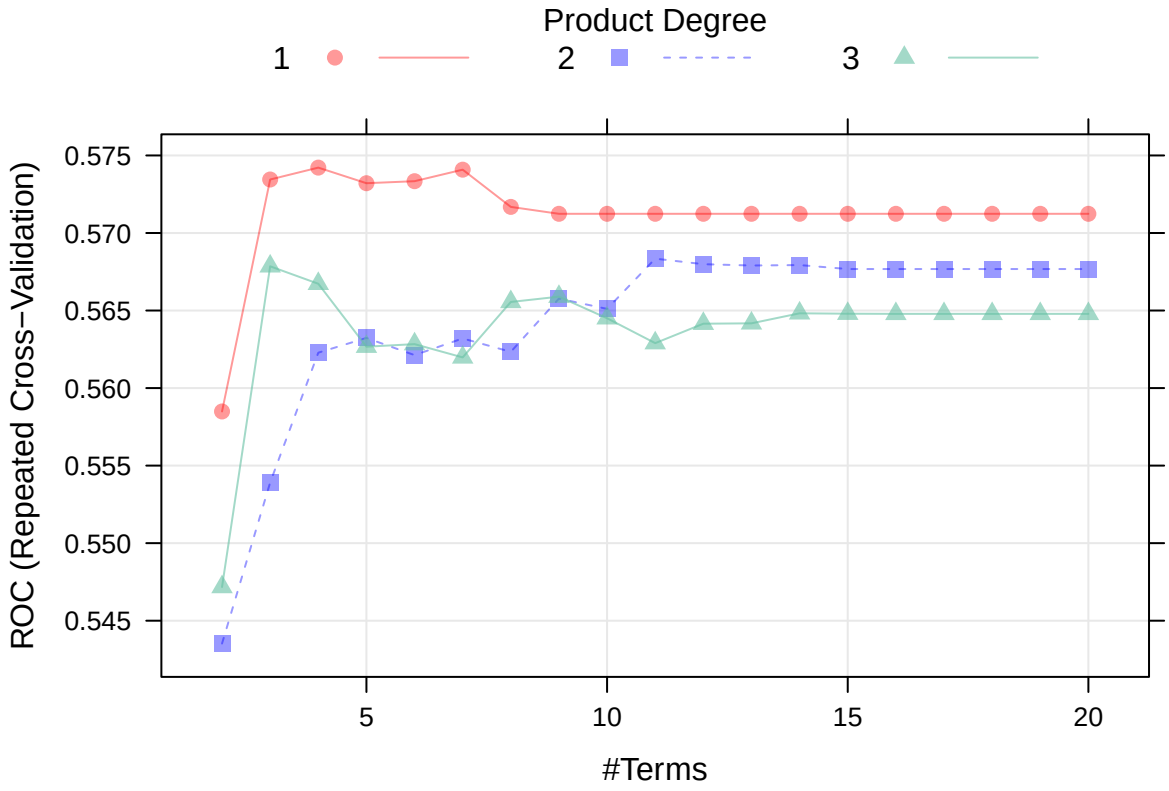
```
##     alpha     lambda
## 134   0.1 0.03741385
```

```
myCol <- rainbow(25)
myPar <- list(superpose.symbol = list(col = myCol),
              superpose.line = list(col = myCol))

plot(model.glmn, par.settings = myPar, xTrans = function(x) log(x))
```



MARS

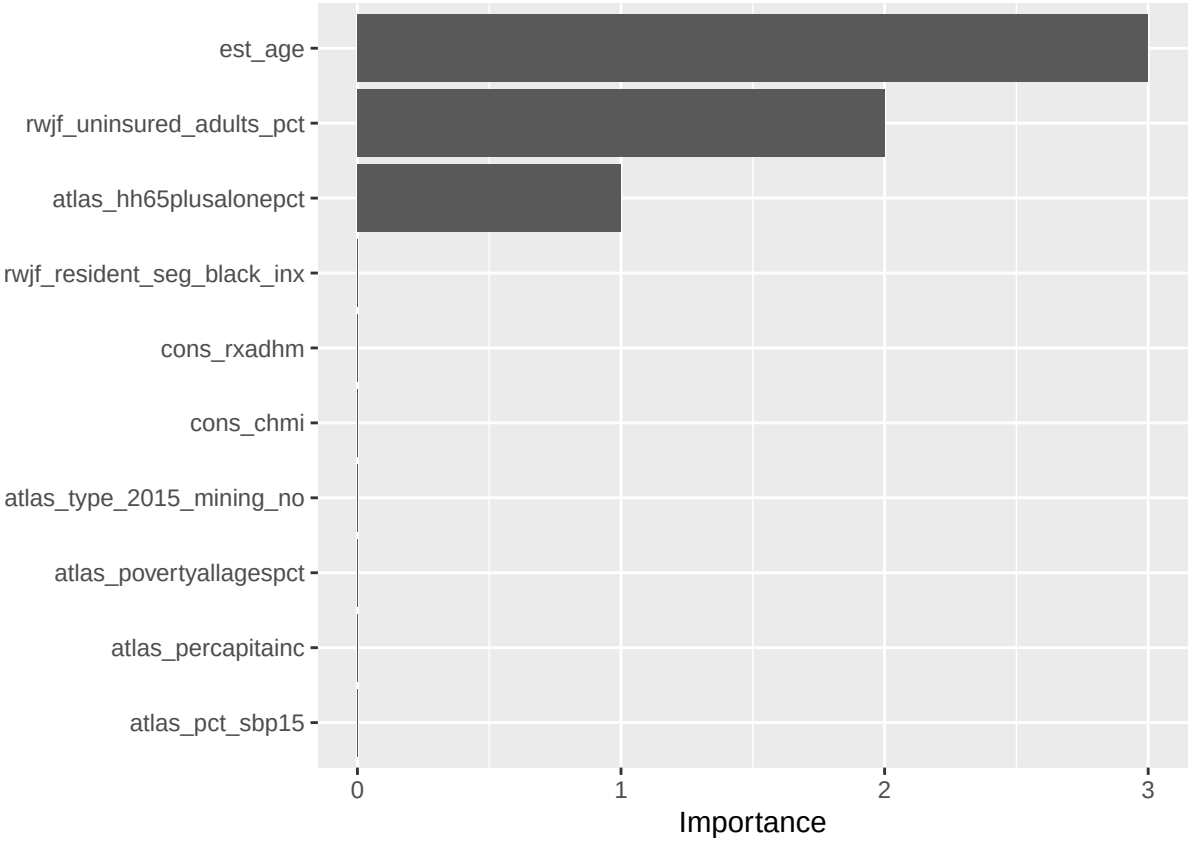
```
set.seed(1)
model.mars <- train(x = dat[train,1:14],
                    y = dat$covid_vaccination[train],
                    method = "earth",
                    tuneGrid = expand.grid(degree = 1:3,
                                           nprune = 2:20),
                    metric = "ROC",
                    trControl = ctrl)

plot(model.mars)
```




```
coef(model.mars$finalModel)
```

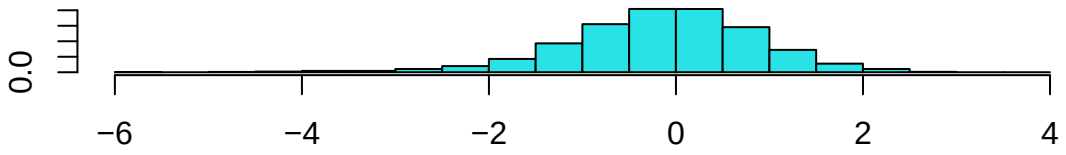
```
##                             (Intercept)                             h(99-est_age)
##                              -0.95401013                              -0.02500079
## h(0.121157-rwjf_uninsured_adults_pct)     h(atlas_hh65plusalonepct-18.2505)
##                               6.66928112                               1.61560562
```

```
vip(model.mars$finalModel)
```

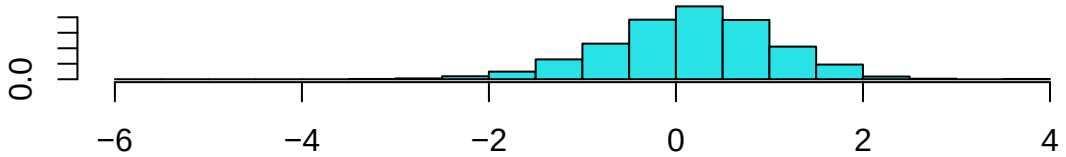


LDA

```
lda.fit <- lda(covid_vaccination~., data = dat,
               subset = train)
plot(lda.fit)
```



group no_vacc



group vacc

```
lda.fit$scaling
```

```
##                                      LD1
## cons_chmi                     5.368571e-03
## est_age                       6.991183e-02
## atlas_percapitainc           -1.348221e-05
## rwjf_uninsured_adults_pct    -8.160718e+00
## atlas_type_2015_mining_no    -6.545607e-01
## atlas_povertyallagespct       2.725220e-02
## atlas_hh65plusalonepct        2.930581e-02
## atlas_pct_sbp15               1.341964e-02
## rwjf_resident_seg_black_inx  -5.998705e-04
## cons_rxadhm                   1.126767e-01
## atlas_medhhinc                1.957556e-05
## cons_lwcm07                  -3.795710e+00
## atlas_low_education_2015_update  2.703274e-02
## race_cd                      -2.071219e-01
```

```
head(predict(lda.fit)$x)
```

```
##           LD1
## 2   0.7086198
## 4   0.3245458
## 5   1.3285887
## 7   1.1243922
## 8  -0.3859000
## 10 -0.1839894
```

```
mean(predict(lda.fit)$x)
```

```
## [1] -2.917938e-16
```

```
lda.pred <- predict(lda.fit, newdata = dat[-train,])
head(lda.pred$posterior)
```

```
##      no_vacc      vacc
## 1  0.8425982 0.1574018
## 3  0.7219217 0.2780783
## 6  0.7774544 0.2225456
## 9  0.7595194 0.2404806
## 11 0.7728520 0.2271480
## 12 0.7457247 0.2542753
```

```
# caret
set.seed(1)
model.lda <- train(x = dat[train,1:14],
                   y = dat$covid_vaccination[train],
                   method = "lda",
                   metric = "ROC",
                   trControl = ctrl)
```

QDA

```
qda.fit <- qda(covid_vaccination~., data = dat,
               subset = train)

qda.pred <- predict(qda.fit, newdata = dat[-train,])
head(qda.pred$posterior)
```

```
##      no_vacc      vacc
## 1  0.9268177 0.0731823
## 3  0.5531910 0.4468090
## 6  0.4631455 0.5368545
## 9  0.5385034 0.4614966
## 11 0.5819419 0.4180581
## 12 0.6528265 0.3471735
```

```
set.seed(1)
model.qda <- train(x = dat[train,1:14],
                   y = dat$covid_vaccination[train],
                   method = "qda",
                   metric = "ROC",
                   trControl = ctrl)
```

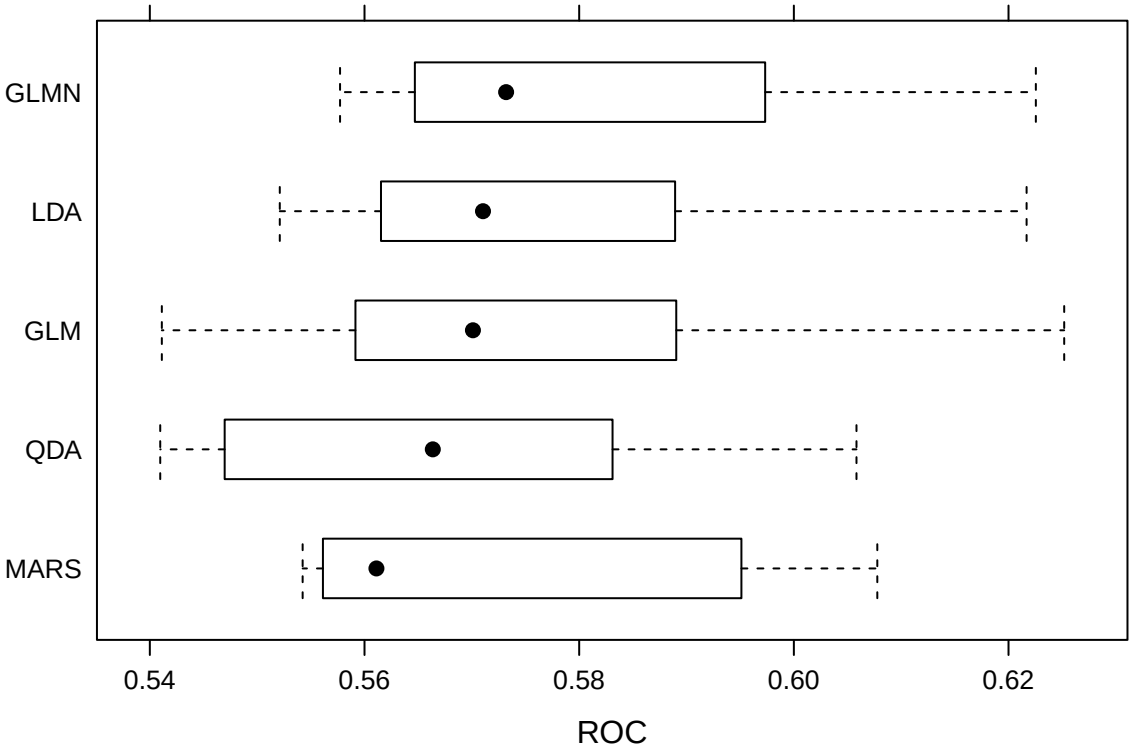
Resample

```
# resample
res <- resamples(list(MARS = model.mars,
                      GLM = model.glm,
                      GLMN = model.glmn,
```

```
                      LDA = model.lda,
                      QDA = model.qda))
summary(res)
```

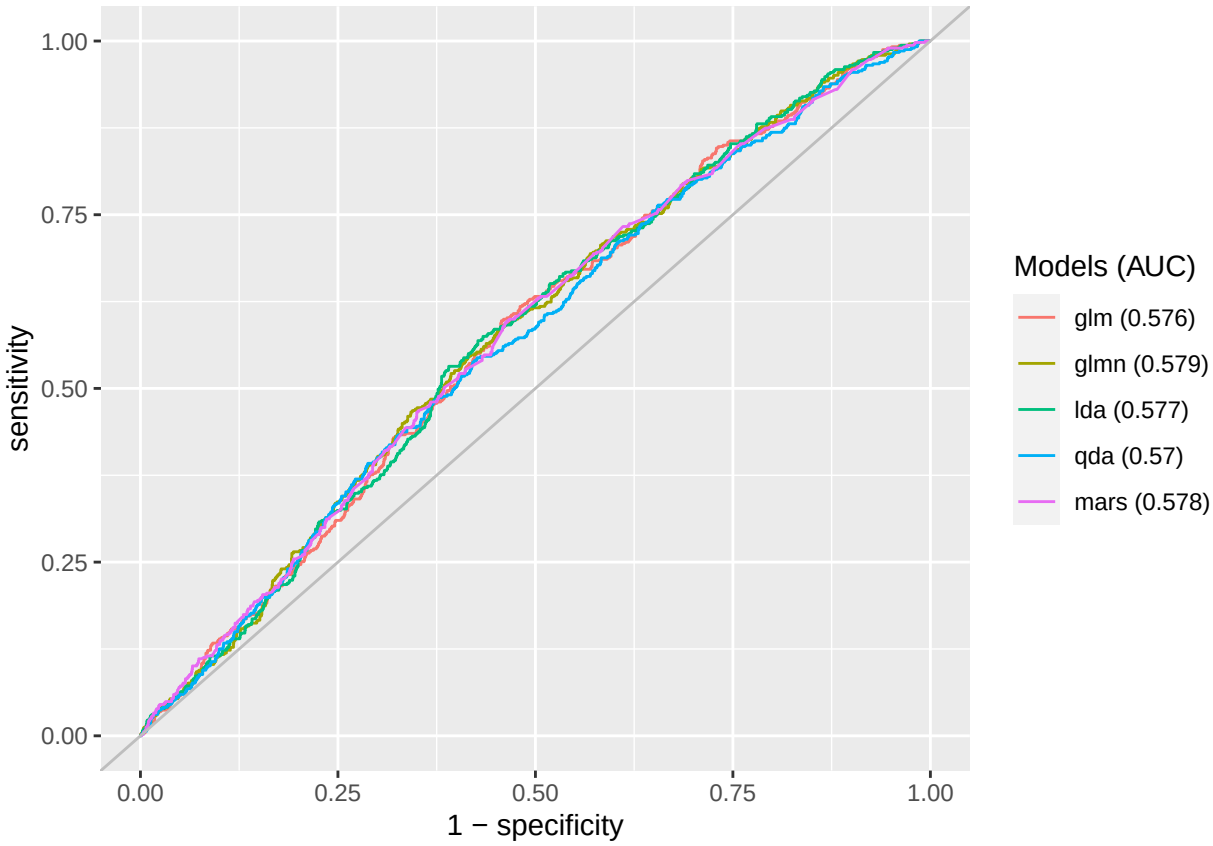
```
##
## Call:
## summary.resamples(object = res)
##
## Models: MARS, GLM, GLMN, LDA, QDA
## Number of resamples: 10
##
## ROC
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## MARS 0.5542341 0.5567294 0.5611111 0.5742158 0.5934874 0.6077748    0
## GLM  0.5411172 0.5596416 0.5701005 0.5764965 0.5883941 0.6251874    0
## GLMN 0.5577219 0.5648013 0.5731931 0.5803878 0.5951213 0.6225446    0
## LDA  0.5521056 0.5619611 0.5710376 0.5769682 0.5887174 0.6216824    0
## QDA  0.5409670 0.5500825 0.5663627 0.5672915 0.5810888 0.6058354    0
##
## Sens
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## MARS 0.9978632 1.0000000 1.0000000 0.9995726 1.0000000 1.0000000    0
## GLM  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
## GLMN 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
## LDA  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000    0
## QDA  0.9102564 0.9234876 0.9380342 0.9341546 0.9439103 0.9508547    0
##
## Spec
##            Min.    1st Qu.     Median        Mean   3rd Qu.       Max. NA's
## MARS 0.00000000 0.00000000 0.00000000 0.001754386 0.0000000 0.00877193    0
## GLM  0.00000000 0.00000000 0.00000000 0.000000000 0.0000000 0.00000000    0
## GLMN 0.00000000 0.00000000 0.00000000 0.000000000 0.0000000 0.00000000    0
## LDA  0.00000000 0.00000000 0.00000000 0.000000000 0.0000000 0.00000000    0
## QDA  0.03508772 0.06140351 0.07487191 0.078132278 0.0877193 0.14912281    0
```

```
bwplot(res, metric = "ROC")
```



ROC Curve

```
glm.pred <- predict(model.glm, newdata = dat[-train,], type = "prob")[,2]
glmn.pred <- predict(model.glmn, newdata = dat[-train,], type = "prob")[,2]
lda.pred <- predict(model.lda, newdata = dat[-train,], type = "prob")[,2]
qda.pred <- predict(model.qda, newdata = dat[-train,], type = "prob")[,2]
mars.pred <- predict(model.mars, newdata = dat[-train,], type = "prob")[,2]

roc.glm <- roc(dat$covid_vaccination[-train], glm.pred)
roc.glmn <- roc(dat$covid_vaccination[-train], glmn.pred)
roc.lda <- roc(dat$covid_vaccination[-train], lda.pred)
roc.qda <- roc(dat$covid_vaccination[-train], qda.pred)
roc.mars <- roc(dat$covid_vaccination[-train], mars.pred)

auc <- c(roc.glm$auc[1], roc.glmn$auc[1],
         roc.lda$auc[1], roc.qda$auc[1],
         roc.mars$auc[1])

modelNames <- c("glm","glmn","lda","qda", "mars")

ggroc(list(roc.glm, roc.glmn, roc.lda, roc.qda, roc.mars), legacy.axes = TRUE) +
  scale_color_discrete(labels = paste0(modelNames, " (", round(auc,3),")"),
                       name = "Models (AUC)") +
  geom_abline(intercept = 0, slope = 1, color = "grey")
```


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 (0.576)
glmn (0.579)
lda (0.577)
qda (0.57)
mars (0.578)