Model selection

Lu Zheng

2023-03-30

1 Load Required Libraries

```
library(boot)
library(pROC)
```

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

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(ROCR)
```

2 Load the data

```
data.credit = read.csv("Credit.csv")
# Transform categorical variables
data.credit$credit_risk = as.factor(data.credit$credit_risk)
data.credit$status  = as.factor(data.credit$status)
data.credit$savings = as.factor(data.credit$savings)
data.credit$property = as.ordered(data.credit$property)
data.credit$other_installment_plans = as.factor(data.credit$other_installment_plans)
```

3 Split the data into training set and testing set

```
set.seed(1006742107)

n = nrow(data.credit)
index = sample(n, round(0.75 * n), replace = FALSE)
traindata = data.credit[index, ]
testdata = data.credit[-index, ]
```

4 Main effect model

4.1 Training model

4.1.1 Forward method

```
step(glm(credit_risk ~ 1, family = binomial, data = traindata), scope =
       ~status + duration + savings + property + age +
       other_installment_plans, direction = "forward", test = "Chisq")
```

```
## Start:  AIC=921.66
## credit_risk ~ 1
##
##                            Df Deviance    AIC    LRT  Pr(>Chi)
## + status                    3   828.43 836.43 91.225 < 2.2e-16 ***
## + duration                  1   887.29 891.29 32.368 1.276e-08 ***
## + savings                   4   892.66 902.66 27.003 1.985e-05 ***
## + property                  3   903.76 911.76 15.899  0.001189 **
## + age                       1   909.12 913.12 10.537  0.001170 **
## + other_installment_plans   2   909.77 915.77  9.889  0.007123 **
## <none>                          919.66 921.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=836.43
## credit_risk ~ status
##
##                            Df Deviance    AIC     LRT  Pr(>Chi)
## + duration                  1   801.74 811.74 26.6915 2.387e-07 ***
## + property                  3   811.90 825.90 16.5377 0.0008796 ***
## + savings                   4   814.60 830.60 13.8362 0.0078365 **
## + other_installment_plans   2   820.32 832.32  8.1125 0.0173134 *
## + age                       1   823.40 833.40  5.0337 0.0248586 *
## <none>                          828.43 836.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=811.74
## credit_risk ~ status + duration
##
##                            Df Deviance    AIC     LRT Pr(>Chi)
## + savings                   4   787.09 805.09 14.6529 0.005478 **
## + other_installment_plans   2   794.74 808.74  7.0039 0.030138 *
## + age                       1   797.00 809.00  4.7411 0.029450 *
## + property                  3   794.81 810.81  6.9303 0.074154 .
## <none>                          801.74 811.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=805.09
## credit_risk ~ status + duration + savings
##
```

```
##                           Df Deviance    AIC    LRT Pr(>Chi)
## + other_installment_plans  2   779.60 801.60 7.4878  0.02366 *
## + age                      1   783.28 803.28 3.8072  0.05103 .
## + property                 3   780.56 804.56 6.5277  0.08858 .
## <none>                         787.09 805.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=801.6
## credit_risk ~ status + duration + savings + other_installment_plans
##
##            Df Deviance    AIC    LRT Pr(>Chi)
## + age       1   775.22 799.22 4.3864  0.03623 *
## <none>          779.60 801.60
## + property  3   774.53 802.53 5.0741  0.16645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=799.22
## credit_risk ~ status + duration + savings + other_installment_plans +
##     age
##
##            Df Deviance    AIC    LRT Pr(>Chi)
## + property  3   767.65 797.65 7.5621  0.05598 .
## <none>          775.22 799.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=797.65
## credit_risk ~ status + duration + savings + other_installment_plans +
##     age + property


##
## Call:  glm(formula = credit_risk ~ status + duration + savings + other_installment_plans +
##     age + property, family = binomial, data = traindata)
##
## Coefficients:
##              (Intercept)                   status2                   status3
##                  -0.72146                   0.45381                   0.85533
##                   status4                  duration                  savings2
##                   1.75141                  -0.03046                   0.25614
##                  savings3                  savings4                  savings5
##                   0.13951                   1.49757                   0.72922
## other_installment_plans2  other_installment_plans3                       age
##                   0.19640                   0.59292                   0.02236
##               property.L                property.Q                property.C
##                  -0.58463                  -0.16008                  -0.06641
##
## Degrees of Freedom: 749 Total (i.e. Null);  735 Residual
## Null Deviance:       919.7
## Residual Deviance: 767.7     AIC: 797.7
```

4.1.2 Backward method

```
step(glm(credit_risk ~status + duration + savings + property + age +
           other_installment_plans, family = binomial, data = traindata), test = "Chisq")
```

```
## Start:  AIC=797.65
## credit_risk ~ status + duration + savings + property + age +
##     other_installment_plans
##
##                            Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                          767.65 797.65
## - property                  3   775.22 799.22  7.562  0.055984 .
## - other_installment_plans   2   774.23 800.23  6.577  0.037312 *
## - age                       1   774.53 802.53  6.874  0.008744 **
## - savings                   4   781.43 803.43 13.779  0.008036 **
## - duration                  1   784.38 812.38 16.727 4.316e-05 ***
## - status                    3   832.72 856.72 65.065 4.857e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:  glm(formula = credit_risk ~ status + duration + savings + property +
##     age + other_installment_plans, family = binomial, data = traindata)
##
## Coefficients:
##              (Intercept)                  status2                  status3
##                  -0.72146                  0.45381                  0.85533
##                  status4                 duration                  savings2
##                   1.75141                 -0.03046                  0.25614
##                  savings3                  savings4                 savings5
##                   0.13951                  1.49757                  0.72922
##                property.L               property.Q               property.C
##                  -0.58463                 -0.16008                 -0.06641
##                        age other_installment_plans2 other_installment_plans3
##                   0.02236                  0.19640                  0.59292
##
## Degrees of Freedom: 749 Total (i.e. Null);  735 Residual
## Null Deviance:       919.7
## Residual Deviance: 767.7     AIC: 797.7
```

From above coding, we could find that both forward selection and backward elimination choose the model: glm(credit_risk ~status + duration + savings + property + age + other_installment_plans, family = binomial, data = traindata)

$$logit(\hat{\pi}) = -0.72 + 0.45 \cdot S_1 + 0.86 \cdot S_2 + 1.75 \cdot S_3 - 0.03 \cdot D + 0.26 \cdot SV_1 + 0.14 \cdot SV_2 + 1.50 SV_3 + 0.73 SV_4 - 0.58 \cdot P_L - 0.16 \cdot P_Q$$

$$-0.07 \cdot P_C + 0.02 \cdot A + 0.20 \cdot O_1 + 0.59 \cdot O_2$$

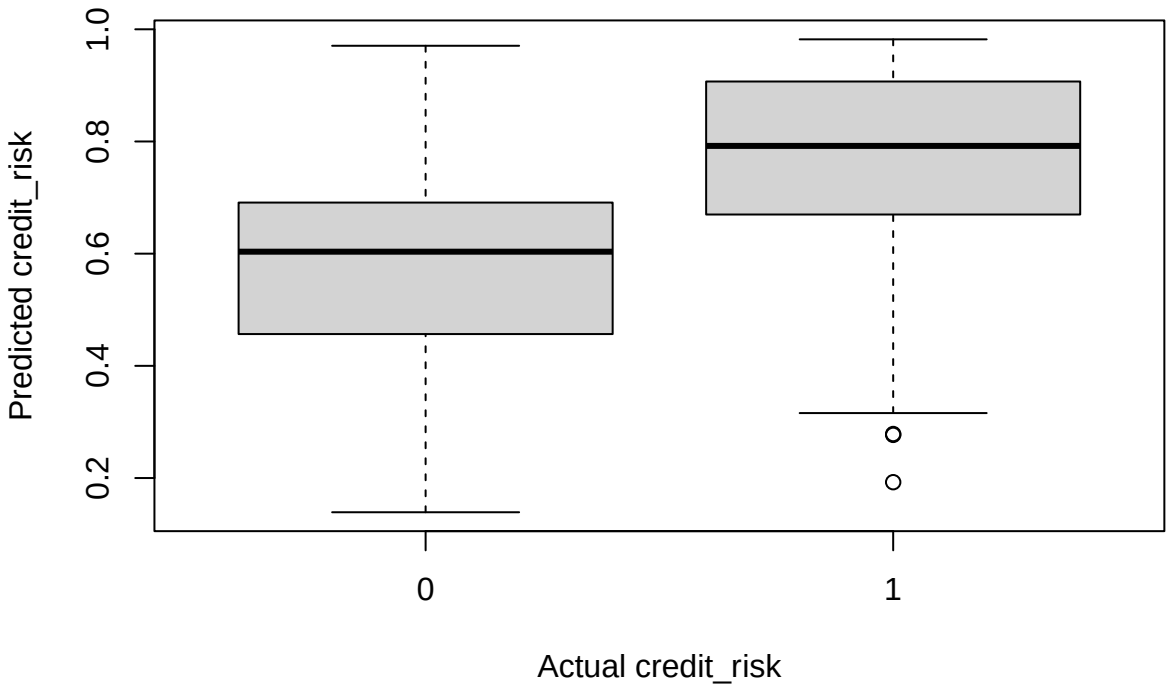
where * S_i's are dummy variables for status * D is duration * SV's are dummy variables for savings * P_i's are dummy variables for property * A is age * O_i's are dummy variables for other_installment_plans

```r
bestmodel.1 = glm(credit_risk ~status + duration + savings + property + age + other_installment_plans, f
summary(bestmodel.1)
```

```
##
## Call:
## glm(formula = credit_risk ~ status + duration + savings + property +
##     age + other_installment_plans, family = binomial, data = traindata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7517  -0.9595   0.4902   0.8265   1.8215
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.721459   0.436812  -1.652  0.09861 .
## status2                   0.453812   0.222048   2.044  0.04098 *
## status3                   0.855325   0.366951   2.331  0.01976 *
## status4                   1.751411   0.235658   7.432 1.07e-13 ***
## duration                 -0.030462   0.007502  -4.061 4.89e-05 ***
## savings2                  0.256136   0.302805   0.846  0.39762
## savings3                  0.139505   0.392170   0.356  0.72205
## savings4                  1.497575   0.657083   2.279  0.02266 *
## savings5                  0.729215   0.264897   2.753  0.00591 **
## property.L               -0.584630   0.213990  -2.732  0.00629 **
## property.Q               -0.160077   0.189526  -0.845  0.39832
## property.C               -0.066411   0.173897  -0.382  0.70254
## age                       0.022359   0.008708   2.568  0.01024 *
## other_installment_plans2  0.196405   0.440276   0.446  0.65553
## other_installment_plans3  0.592919   0.240369   2.467  0.01364 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 919.66  on 749  degrees of freedom
## Residual deviance: 767.65  on 735  degrees of freedom
## AIC: 797.65
##
## Number of Fisher Scoring iterations: 5
```

4.2 Testing model

```r
pred.1 = predict(bestmodel.1, newdata = testdata)
plot(testdata$credit_risk, inv.logit(pred.1), xlab = "Actual credit_risk", ylab = "Predicted credit_risk
```



From the plot, we find that the main effect model can describe the actual data fairly well.

5 Interaction model

5.1 Training model

5.1.1 Forward method

```
bestmodel.3 <- step(glm(credit_risk ~ 1, family = binomial, data = traindata), scope = ~status * durati
```

```
## Start:  AIC=921.66
## credit_risk ~ 1
##
##                            Df Deviance    AIC    LRT  Pr(>Chi)
## + status                    3   828.43 836.43 91.225 < 2.2e-16 ***
## + duration                  1   887.29 891.29 32.368 1.276e-08 ***
## + savings                   4   892.66 902.66 27.003 1.985e-05 ***
## + property                  3   903.76 911.76 15.899  0.001189 **
## + age                       1   909.12 913.12 10.537  0.001170 **
## + other_installment_plans   2   909.77 915.77  9.889  0.007123 **
## <none>                          919.66 921.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=836.43
## credit_risk ~ status
##
##                            Df Deviance    AIC     LRT  Pr(>Chi)
## + duration                  1   801.74 811.74 26.6915 2.387e-07 ***
```

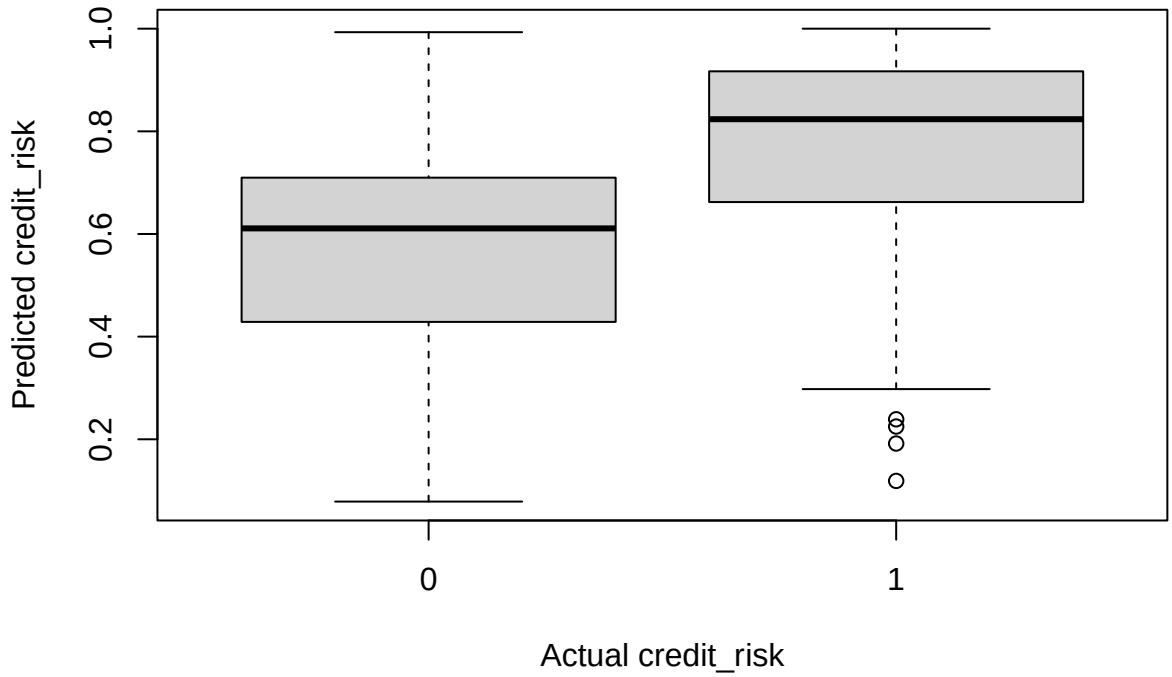
```
## + property                 3   811.90 825.90 16.5377 0.0008796 ***
## + savings                  4   814.60 830.60 13.8362 0.0078365 **
## + other_installment_plans  2   820.32 832.32  8.1125 0.0173134 *
## + age                      1   823.40 833.40  5.0337 0.0248586 *
## <none>                         828.43 836.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=811.74
## credit_risk ~ status + duration
##
##                           Df Deviance    AIC     LRT Pr(>Chi)
## + savings                  4   787.09 805.09 14.6529 0.005478 **
## + other_installment_plans  2   794.74 808.74  7.0039 0.030138 *
## + age                      1   797.00 809.00  4.7411 0.029450 *
## + property                 3   794.81 810.81  6.9303 0.074154 .
## <none>                         801.74 811.74
## + status:duration          3   797.97 813.97  3.7701 0.287383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=805.09
## credit_risk ~ status + duration + savings
##
##                           Df Deviance    AIC     LRT Pr(>Chi)
## + duration:savings         4   773.61 799.61 13.4777 0.009163 **
## + other_installment_plans  2   779.60 801.60  7.4878 0.023662 *
## + age                      1   783.28 803.28  3.8072 0.051032 .
## + property                 3   780.56 804.56  6.5277 0.088578 .
## <none>                         787.09 805.09
## + status:duration          3   783.47 807.47  3.6232 0.305134
## + status:savings          12   769.40 811.40 17.6913 0.125392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=799.61
## credit_risk ~ status + duration + savings + duration:savings
##
##                           Df Deviance    AIC     LRT Pr(>Chi)
## + other_installment_plans  2   765.10 795.10  8.5114  0.01418 *
## + age                      1   769.17 797.17  4.4427  0.03505 *
## + property                 3   766.99 798.99  6.6259  0.08483 .
## <none>                         773.61 799.61
## + status:duration          3   770.18 802.18  3.4347  0.32934
## + status:savings          12   760.49 810.49 13.1277  0.35983
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=795.1
## credit_risk ~ status + duration + savings + other_installment_plans +
##     duration:savings
##
##                                 Df Deviance    AIC     LRT Pr(>Chi)
## + age                            1   760.07 792.07  5.0305   0.0249 *
```

```
## <none>                                    765.10 795.10
## + property                            3   759.95 795.95  5.1531   0.1609
## + status:duration                     3   761.92 797.92  3.1767   0.3652
## + duration:other_installment_plans    2   765.05 799.05  0.0514   0.9746
## + savings:other_installment_plans     8   756.37 802.37  8.7308   0.3655
## + status:other_installment_plans      6   761.30 803.30  3.8027   0.7034
## + status:savings                     12   753.00 807.00 12.1014   0.4376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=792.07
## credit_risk ~ status + duration + savings + other_installment_plans +
##     age + duration:savings
##
##                                    Df Deviance    AIC     LRT Pr(>Chi)
## + age:other_installment_plans       2   750.02 786.02 10.0493 0.006574 **
## + property                          3   752.24 790.24  7.8305 0.049647 *
## <none>                                  760.07 792.07
## + duration:age                      1   759.61 793.61  0.4620 0.496669
## + status:duration                   3   756.45 794.45  3.6226 0.305204
## + duration:other_installment_plans  2   760.04 796.04  0.0335 0.983389
## + status:age                        3   758.23 796.23  1.8404 0.606175
## + savings:age                       4   757.98 797.98  2.0917 0.718903
## + savings:other_installment_plans   8   750.31 798.31  9.7585 0.282388
## + status:other_installment_plans    6   756.14 800.14  3.9290 0.686284
## + status:savings                   12   747.44 803.44 12.6261 0.396794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=786.02
## credit_risk ~ status + duration + savings + other_installment_plans +
##     age + duration:savings + other_installment_plans:age
##
##                                    Df Deviance    AIC     LRT Pr(>Chi)
## + property                          3   742.30 784.30  7.7188   0.0522 .
## <none>                                  750.02 786.02
## + duration:age                      1   749.76 787.76  0.2605   0.6098
## + status:duration                   3   745.89 787.89  4.1336   0.2474
## + status:age                        3   747.65 789.65  2.3749   0.4983
## + duration:other_installment_plans  2   749.97 789.97  0.0476   0.9765
## + savings:age                       4   747.70 791.70  2.3220   0.6768
## + status:other_installment_plans    6   745.83 793.83  4.1928   0.6506
## + savings:other_installment_plans   8   742.20 794.20  7.8179   0.4515
## + status:savings                   12   739.02 799.02 10.9992   0.5290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=784.3
## credit_risk ~ status + duration + savings + other_installment_plans +
##     age + property + duration:savings + other_installment_plans:age
##
##                                    Df Deviance    AIC     LRT Pr(>Chi)
## + status:property                   9   722.44 782.44 19.8671  0.01875 *
## <none>                                  742.30 784.30
```

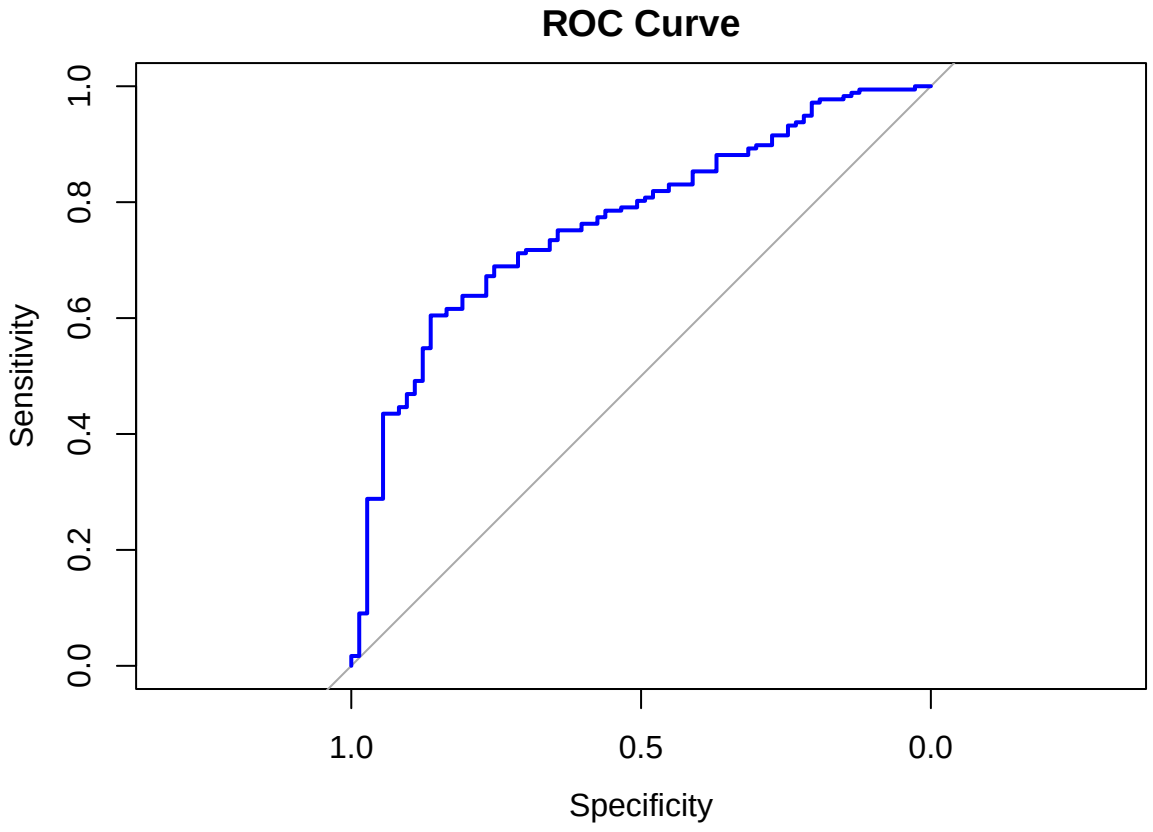
```
## + duration:age                       1   742.22 786.22  0.0787  0.77913
## + status:duration                    3   738.37 786.37  3.9359  0.26847
## + duration:property                  3   738.94 786.94  3.3634  0.33891
## + status:age                         3   740.05 788.05  2.2542  0.52135
## + duration:other_installment_plans   2   742.29 788.29  0.0141  0.99300
## + property:age                       3   741.47 789.47  0.8287  0.84258
## + savings:age                        4   739.79 789.79  2.5115  0.64258
## + property:other_installment_plans   6   737.35 791.35  4.9562  0.54944
## + status:other_installment_plans     6   738.23 792.23  4.0688  0.66737
## + savings:other_installment_plans    8   734.97 792.97  7.3372  0.50073
## + savings:property                  12   729.92 795.92 12.3800  0.41566
## + status:savings                    12   731.61 797.61 10.6915  0.55552
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=782.44
## credit_risk ~ status + duration + savings + other_installment_plans +
##     age + property + duration:savings + other_installment_plans:age +
##     status:property
##
##                                    Df Deviance    AIC     LRT Pr(>Chi)
## <none>                                  722.44 782.44
## + duration:age                      1   722.35 784.35  0.0832   0.7730
## + status:age                        3   718.69 784.69  3.7428   0.2906
## + duration:property                 3   718.92 784.92  3.5143   0.3189
## + status:duration                   3   720.15 786.15  2.2823   0.5159
## + duration:other_installment_plans  2   722.22 786.22  0.2146   0.8983
## + savings:age                       4   719.83 787.83  2.6071   0.6256
## + property:age                      3   721.88 787.88  0.5560   0.9064
## + property:other_installment_plans  6   717.17 789.17  5.2635   0.5105
## + savings:other_installment_plans   8   715.61 791.61  6.8292   0.5552
## + status:other_installment_plans    6   720.08 792.08  2.3606   0.8837
## + savings:property                 12   709.66 793.66 12.7724   0.3858
## + status:savings                   12   710.09 794.09 12.3503   0.4180
```

5.2 Testing model

```
pred.3 <- predict(bestmodel.3, newdata = testdata)
plot(testdata$credit_risk, inv.logit(pred.3), xlab = "Actual credit_risk", ylab = "Predicted credit_risk
```




```
roc(testdata$credit_risk~inv.logit(pred.3), plot=TRUE, main="ROC Curve", col="blue")
```




```
##
## Call:
## roc.formula(formula = testdata$credit_risk ~ inv.logit(pred.3),     plot = TRUE, main = "ROC Curve",
```

```
##
## Data: inv.logit(pred.3) in 73 controls (testdata$credit_risk 0) < 177 cases (testdata$credit_risk 1)
## Area under the curve: 0.7659
```

```
auc(testdata$credit_risk~inv.logit(pred.3))
```

```
## Area under the curve: 0.7659
```