Model selection

Lu Zheng

2023-03-30

1 Load Required Libraries

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

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

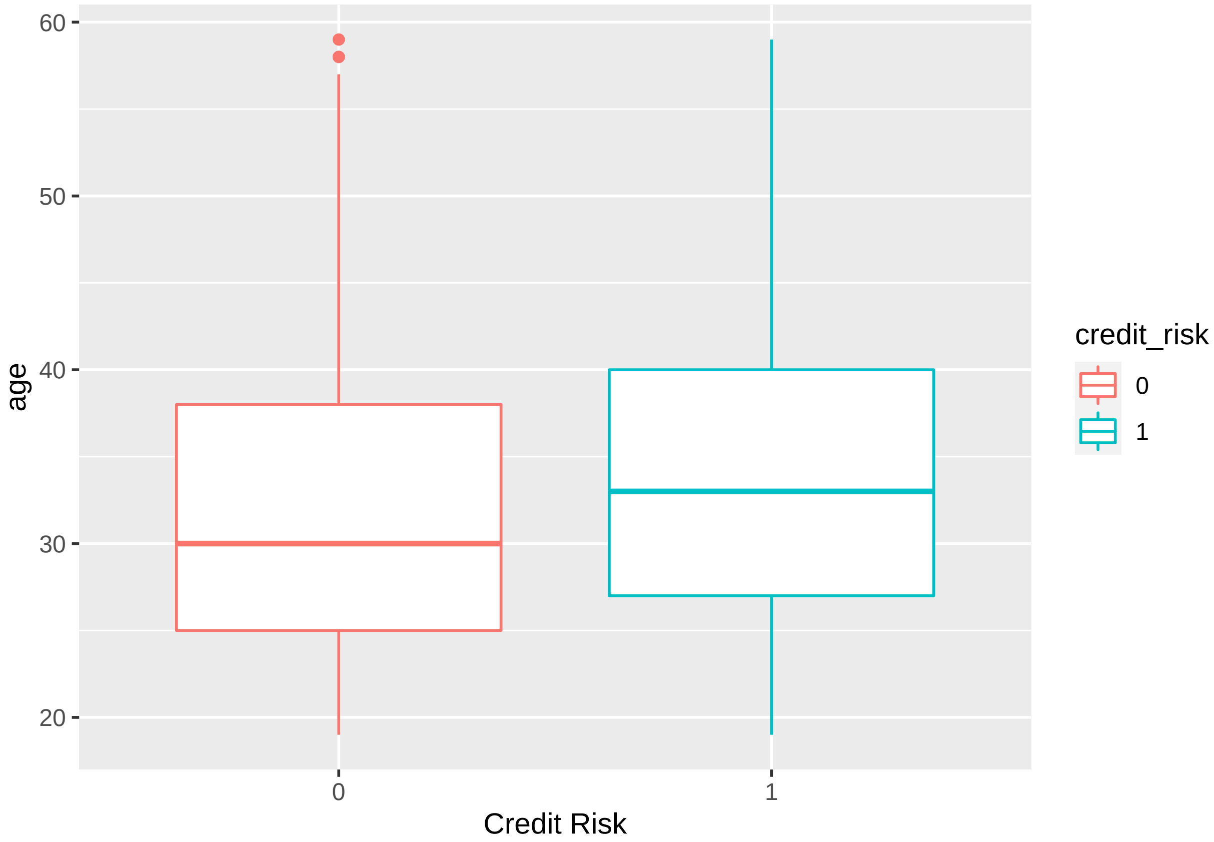
```
##
## Attaching package: 'pROC'
```

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

```r
library(ROCR)
library(ggplot2)
```

2 Load the data

```r
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)
# Remove outliers of age
data.credit = subset(data.credit, age < 60)
ggplot(data.credit, aes(x=as.factor(credit_risk), y=age, color=credit_risk)) +
      geom_boxplot() + xlab("Credit Risk")
```



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=890.33
## credit_risk ~ 1
##
##                           Df Deviance    AIC    LRT  Pr(>Chi)
## + status                   3   791.42 799.42 96.910 < 2.2e-16 ***
## + duration                 1   857.89 861.89 30.435 3.453e-08 ***
## + savings                  4   865.44 875.44 22.891 0.0001331 ***
## + property                 3   869.29 877.29 19.037 0.0002686 ***
## + age                      1   878.41 882.41  9.915 0.0016389 **
## + other_installment_plans  2   882.28 888.28  6.053 0.0484851 *
## <none>                         888.33 890.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=799.42
## credit_risk ~ status
##
##                           Df Deviance    AIC     LRT  Pr(>Chi)
## + duration                 1   765.81 775.81 25.6063 4.187e-07 ***
## + property                 3   772.48 786.48 18.9416 0.0002811 ***
## + age                      1   786.77 796.77  4.6539 0.0309837 *
## + savings                  4   781.03 797.03 10.3868 0.0343923 *
## + other_installment_plans  2   785.72 797.72  5.6974 0.0579184 .
## <none>                         791.42 799.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=775.81
## credit_risk ~ status + duration
##
##                           Df Deviance    AIC     LRT Pr(>Chi)
## + age                      1   760.69 772.69  5.1253  0.02358 *
## + property                 3   756.96 772.96  8.8566  0.03126 *
## + savings                  4   755.21 773.21 10.6035  0.03140 *
## + other_installment_plans  2   761.52 775.52  4.2925  0.11692
## <none>                         765.81 775.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=772.69
## credit_risk ~ status + duration + age
##
##                           Df Deviance    AIC     LRT Pr(>Chi)
## + property                 3   749.15 767.15 11.5332 0.009166 **
## + savings                  4   750.84 770.84  9.8479 0.043070 *
## + other_installment_plans  2   756.05 772.05  4.6367 0.098435 .
## <none>                         760.69 772.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=767.15
## credit_risk ~ status + duration + age + property
##
##                           Df Deviance    AIC     LRT Pr(>Chi)
## + savings                  4   738.78 764.78 10.3752  0.03456 *
## + other_installment_plans  2   744.65 766.65  4.5038  0.10520
## <none>                         749.15 767.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=764.78
## credit_risk ~ status + duration + age + property + savings
##
##                            Df Deviance    AIC    LRT Pr(>Chi)
## + other_installment_plans  2   733.74 763.74 5.0395  0.08048 .
## <none>                         738.78 764.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=763.74
## credit_risk ~ status + duration + age + property + savings +
##     other_installment_plans
```

```
##
## Call:  glm(formula = credit_risk ~ status + duration + age + property +
##     savings + other_installment_plans, family = binomial, data = traindata)
##
## Coefficients:
##              (Intercept)                   status2                   status3
##                 -1.02393                   0.53576                   0.95861
##                  status4                  duration                       age
##                  1.88266                  -0.02913                   0.03010
##               property.L                property.Q                property.C
##                 -0.73908                  -0.22543                  -0.11023
##                 savings2                  savings3                  savings4
##                  0.11049                   0.41649                   1.09690
##                 savings5  other_installment_plans2  other_installment_plans3
##                  0.66588                  -0.15812                   0.44864
##
## Degrees of Freedom: 711 Total (i.e. Null);  697 Residual
## Null Deviance:       888.3
## Residual Deviance: 733.7     AIC: 763.7
```

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=763.74
## credit_risk ~ status + duration + savings + property + age +
##     other_installment_plans
##
##                            Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                         733.74 763.74
## - other_installment_plans  2   738.78 764.78  5.040  0.080479 .
## - savings                  4   744.65 766.65 10.911  0.027583 *
## - property                 3   745.47 769.47 11.727  0.008381 **
## - age                       1   741.61 769.61  7.873  0.005018 **
## - duration                  1   748.03 776.03 14.287  0.000157 ***
## - status                    3   804.62 828.62 70.880 2.766e-15 ***
## ---
## 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
##                 -1.02393                   0.53576                   0.95861
##                  status4                  duration                  savings2
##                  1.88266                  -0.02913                   0.11049
##                 savings3                  savings4                  savings5
##                  0.41649                   1.09690                   0.66588
##               property.L                property.Q                property.C
##                 -0.73908                  -0.22543                  -0.11023
##                      age  other_installment_plans2  other_installment_plans3
##                  0.03010                  -0.15812                   0.44864
##
## Degrees of Freedom: 711 Total (i.e. Null);  697 Residual
## Null Deviance:       888.3
## Residual Deviance: 733.7     AIC: 763.7
```

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)

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

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

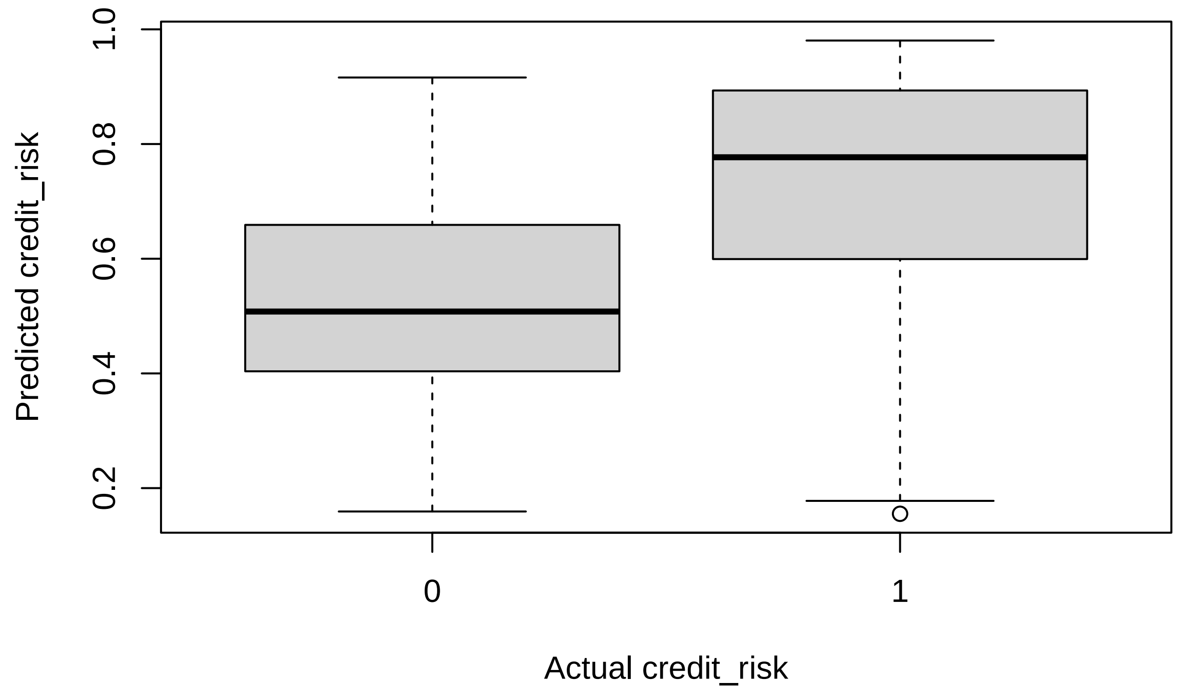
where * S_i's are dummy variables for status * D is duration * SV's are dummy variables for savings * P_i's are dummy variables for property * A is age * O_i's are dummy variables for other_installment_plans

```
bestmodel.1 = glm(credit_risk ~status + duration + savings + property + age + other_installment_plans, 
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
## -2.6697  -0.9490   0.4766   0.8229   1.6663
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.023927   0.499442  -2.050 0.040351 *
## status2                    0.535755   0.224788   2.383 0.017155 *
## status3                    0.958606   0.396762   2.416 0.015689 *
## status4                    1.882664   0.243751   7.724 1.13e-14 ***
## duration                  -0.029128   0.007773  -3.747 0.000179 ***
## savings2                   0.110489   0.296515   0.373 0.709429
## savings3                   0.416489   0.439685   0.947 0.343515
## savings4                   1.096905   0.534634   2.052 0.040199 *
## savings5                   0.665883   0.262515   2.537 0.011195 *
## property.L                -0.739080   0.218353  -3.385 0.000712 ***
## property.Q                -0.225429   0.193900  -1.163 0.244989
## property.C                -0.110234   0.178543  -0.617 0.536964
## age                        0.030105   0.010900   2.762 0.005745 **
## other_installment_plans2 -0.158117   0.439003  -0.360 0.718718
## other_installment_plans3  0.448638   0.253162   1.772 0.076372 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 888.33  on 711  degrees of freedom
## Residual deviance: 733.74  on 697  degrees of freedom
## AIC: 763.74
##
## Number of Fisher Scoring iterations: 4
```

4.2 Testing model

```
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

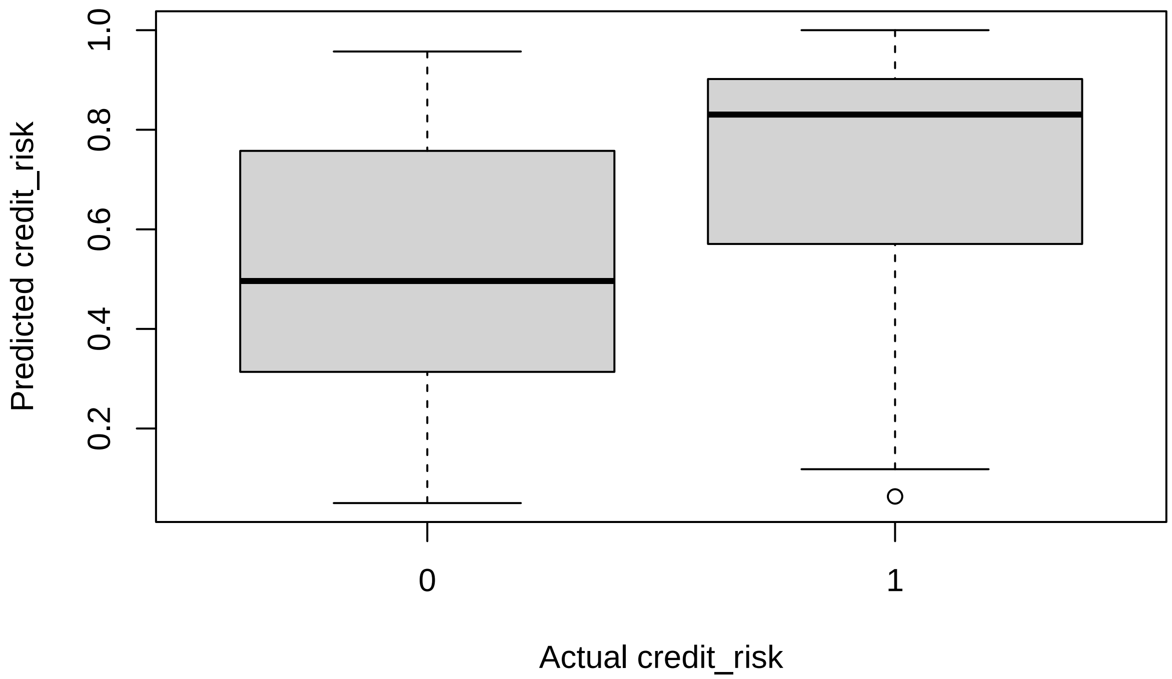
5.1.1 Forward method

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

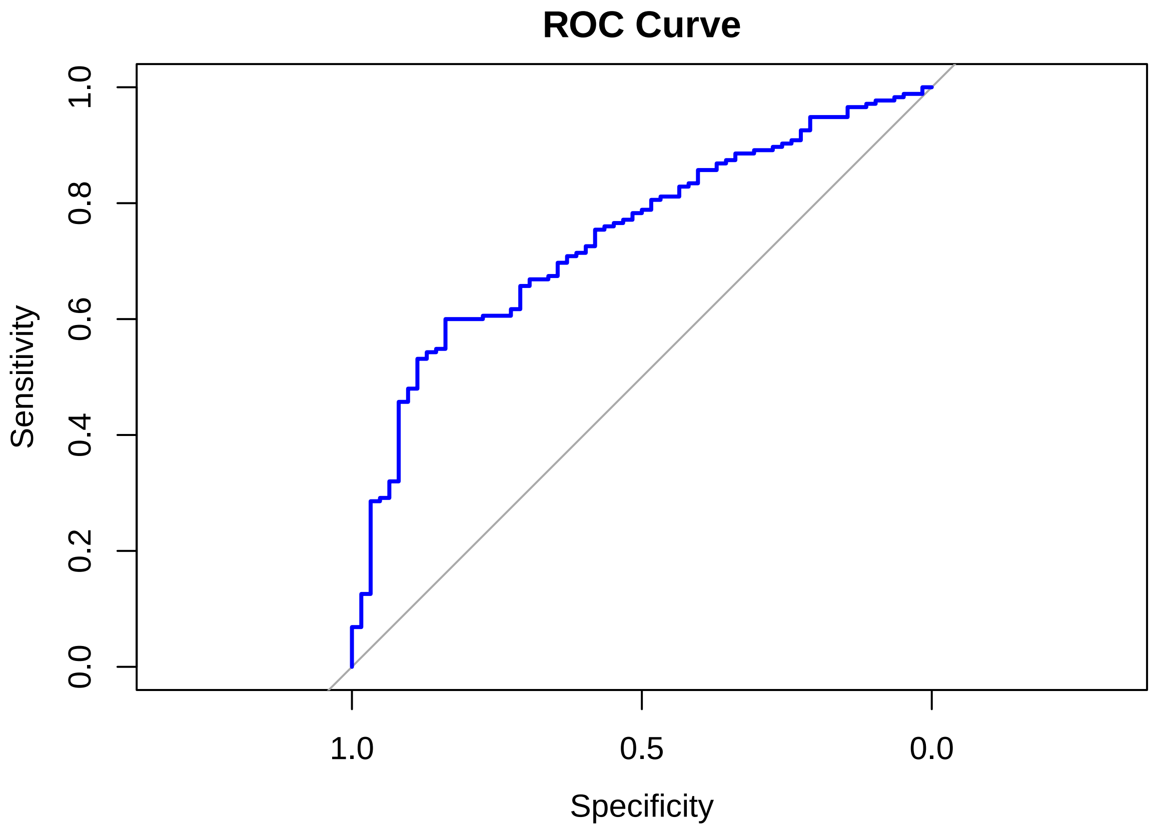
```
## Start:  AIC=890.33
## credit_risk ~ 1
##
##                            Df Deviance    AIC    LRT  Pr(>Chi)
## + status                    3   791.42 799.42 96.910 < 2.2e-16 ***
## + duration                  1   857.89 861.89 30.435 3.453e-08 ***
## + savings                   4   865.44 875.44 22.891 0.0001331 ***
## + property                  3   869.29 877.29 19.037 0.0002686 ***
## + age                       1   878.41 882.41  9.915 0.0016389 **
## + other_installment_plans   2   882.28 888.28  6.053 0.0484851 *
## <none>                          888.33 890.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=799.42
## credit_risk ~ status
##
##                            Df Deviance    AIC     LRT  Pr(>Chi)
## + duration                  1   765.81 775.81 25.6063 4.187e-07 ***
## + property                  3   772.48 786.48 18.9416 0.0002811 ***
## + age                       1   786.77 796.77  4.6539 0.0309837 *
## + savings                   4   781.03 797.03 10.3868 0.0343923 *
## + other_installment_plans   2   785.72 797.72  5.6974 0.0579184 .
## <none>                          791.42 799.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=775.81
## credit_risk ~ status + duration
##
##                            Df Deviance    AIC     LRT Pr(>Chi)
## + age                       1   760.69 772.69  5.1253  0.02358 *
## + property                  3   756.96 772.96  8.8566  0.03126 *
## + savings                   4   755.21 773.21 10.6035  0.03140 *
## + other_installment_plans   2   761.52 775.52  4.2925  0.11692
## <none>                          765.81 775.81
## + status:duration           3   764.31 780.31  1.4976  0.68282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=772.69
## credit_risk ~ status + duration + age
##
##                          Df Deviance    AIC     LRT Pr(>Chi)
## + property                3   749.15 767.15 11.5332 0.009166 **
## + savings                 4   750.84 770.84  9.8479 0.043070 *
## + other_installment_plans  2   756.05 772.05  4.6367 0.098435 .
## + duration:age            1   758.66 772.66  2.0254 0.154686
## <none>                        760.69 772.69
## + status:duration         3   758.99 776.99  1.6923 0.638638
## + status:age              3   760.35 778.35  0.3381 0.952717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=767.15
## credit_risk ~ status + duration + age + property
##
##                          Df Deviance    AIC     LRT Pr(>Chi)
## + savings                 4   738.78 764.78 10.3752  0.03456 *
## + status:property         9   729.78 765.78 19.3742  0.02219 *
## + property:age            3   742.28 766.28  6.8754  0.07598 .
## + other_installment_plans  2   744.65 766.65  4.5038  0.10520
## <none>                        749.15 767.15
## + duration:age            1   748.21 768.21  0.9457  0.33082
## + duration:property       3   746.02 770.02  3.1313  0.37183
## + status:duration         3   747.77 771.77  1.3876  0.70844
## + status:age              3   748.59 772.59  0.5606  0.90538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=764.78
## credit_risk ~ status + duration + age + property + savings
##
##                          Df Deviance    AIC     LRT Pr(>Chi)
## + status:property         9   717.23 761.23 21.5492  0.01042 *
## + other_installment_plans  2   733.74 763.74  5.0395  0.08048 .
## <none>                        738.78 764.78
## + property:age            3   732.97 764.97  5.8107  0.12119
## + duration:savings        4   731.61 765.61  7.1716  0.12709
## + duration:age            1   737.86 765.86  0.9235  0.33657
## + duration:property       3   735.30 767.30  3.4815  0.32317
## + savings:age             4   734.03 768.03  4.7521  0.31369
## + savings:property       12   718.13 768.13 20.6489  0.05576 .
## + status:duration         3   737.17 769.17  1.6124  0.65657
## + status:savings         12   719.39 769.39 19.3862  0.07963 .
## + status:age              3   738.46 770.46  0.3196  0.95630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=761.23
## credit_risk ~ status + duration + age + property + savings +
##     status:property
##
##                          Df Deviance    AIC     LRT Pr(>Chi)
## + other_installment_plans  2   711.79 759.79  5.4389  0.06591 .
## <none>                         717.23 761.23
## + property:age             3   711.41 761.41  5.8220  0.12060
## + duration:age             1   716.59 762.59  0.6388  0.42416
## + duration:property        3   712.91 762.91  4.3223  0.22869
## + savings:age              4   711.13 763.13  6.0988  0.19189
## + duration:savings         4   711.37 763.37  5.8631  0.20961
## + status:savings          12   695.67 763.67 21.5644  0.04270 *
## + savings:property        12   696.74 764.74 20.4884  0.05839 .
## + status:age               3   715.58 765.58  1.6512  0.64784
## + status:duration          3   716.22 766.22  1.0133  0.79805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=759.79
## credit_risk ~ status + duration + age + property + savings +
##     other_installment_plans + status:property
##
##                                     Df Deviance    AIC     LRT Pr(>Chi)
## + age:other_installment_plans       2   705.96 757.96  5.8322  0.05415 .
## + property:age                      3   705.77 759.77  6.0235  0.11047
## <none>                                  711.79 759.79
## + savings:age                       4   704.37 760.37  7.4217  0.11521
## + duration:age                      1   711.21 761.21  0.5770  0.44749
## + duration:savings                  4   705.42 761.42  6.3682  0.17329
## + duration:property                 3   708.06 762.06  3.7324  0.29185
## + status:savings                   12   690.68 762.68 21.1098  0.04880 *
## + savings:property                 12   691.21 763.21 20.5789  0.05690 .
## + duration:other_installment_plans  2   711.73 763.73  0.0623  0.96935
## + status:age                        3   710.28 764.28  1.5060  0.68088
## + status:duration                   3   710.76 764.76  1.0350  0.79279
## + status:other_installment_plans    6   706.74 766.74  5.0535  0.53697
## + property:other_installment_plans  6   706.87 766.87  4.9196  0.55417
## + savings:other_installment_plans   8   703.26 767.26  8.5339  0.38313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=757.96
## credit_risk ~ status + duration + age + property + savings +
##     other_installment_plans + status:property + age:other_installment_plans
##
##                                     Df Deviance    AIC     LRT Pr(>Chi)
## + property:age                      3   699.11 757.11  6.8521  0.07677 .
## <none>                                  705.96 757.96
## + savings:age                       4   698.03 758.03  7.9292  0.09421 .
## + duration:savings                  4   698.60 758.60  7.3581  0.11813
## + duration:age                      1   705.61 759.61  0.3456  0.55660
## + duration:property                 3   702.13 760.13  3.8252  0.28097
## + savings:property                 12   684.38 760.38 21.5788  0.04252 *
## + status:savings                   12   685.55 761.55 20.4087  0.05974 .
## + duration:other_installment_plans  2   705.76 761.76  0.1972  0.90611
## + status:age                        3   704.74 762.74  1.2230  0.74750
## + status:duration                   3   704.83 762.83  1.1241  0.77126
## + status:other_installment_plans    6   701.04 765.04  4.9176  0.55443
## + property:other_installment_plans  6   701.65 765.65  4.3134  0.63435
## + savings:other_installment_plans   8   699.50 767.50  6.4560  0.59629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=757.11
## credit_risk ~ status + duration + age + property + savings +
##     other_installment_plans + status:property + age:other_installment_plans +
##     age:property
##
##                                     Df Deviance    AIC     LRT Pr(>Chi)
## + savings:age                        4   690.90 756.90  8.2083  0.08424 .
## <none>                                   699.11 757.11
## + savings:property                  12   675.91 757.91 23.1937  0.02613 *
## + duration:age                       1   698.95 758.95  0.1584  0.69065
## + duration:savings                   4   693.08 759.08  6.0264  0.19718
## + duration:property                  3   695.63 759.63  3.4799  0.32338
## + duration:other_installment_plans  2   698.85 760.85  0.2552  0.88022
## + status:savings                    12   679.28 761.28 19.8272  0.07043 .
## + status:age                         3   697.69 761.69  1.4203  0.70077
## + status:duration                    3   697.81 761.81  1.2932  0.73076
## + status:other_installment_plans    6   693.89 763.89  5.2189  0.51606
## + property:other_installment_plans  6   694.01 764.01  5.0939  0.53183
## + savings:other_installment_plans   8   691.87 765.87  7.2319  0.51183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=756.9
## credit_risk ~ status + duration + age + property + savings +
##     other_installment_plans + status:property + age:other_installment_plans +
##     age:property + age:savings
##
##                                     Df Deviance    AIC     LRT Pr(>Chi)
## <none>                                   690.90 756.90
## + duration:savings                   4   684.38 758.38  6.5202  0.16352
## + duration:age                       1   690.86 758.86  0.0336  0.85464
## + savings:property                  12   668.98 758.98 21.9203  0.03842 *
## + duration:property                  3   687.87 759.87  3.0244  0.38788
## + duration:other_installment_plans  2   690.62 760.62  0.2797  0.86947
## + status:savings                    12   671.39 761.39 19.5032  0.07709 .
## + status:duration                    3   689.50 761.50  1.4024  0.70497
## + status:age                         3   689.77 761.77  1.1273  0.77048
## + property:other_installment_plans  6   685.49 763.49  5.4109  0.49229
## + status:other_installment_plans    6   685.59 763.59  5.3052  0.50531
## + savings:other_installment_plans   8   683.29 765.29  7.6109  0.47237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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$credit_risk ~ inv.logit(pred.3),     plot = TRUE, main = "ROC Curve",
##
## Data: inv.logit(pred.3) in 62 controls (testdata$credit_risk 0) < 175 cases (testdata$credit_risk 1)
## Area under the curve: 0.7415
```

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

```
## Area under the curve: 0.7415
```