Model Validation and Diagnostics

Zhiquan Cui

2023-04-03

Contents

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)
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

2 Model Building and Model Selection

2.1 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)
```

2.2 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, ]
```

2.3 Interaction model

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

3 Model Validation and Diagnostics

After choosing the final model, we will perform a model validation and diagnostics to examine the robustness of our model. Here, we run the final model with test data and check the resulting ROC, AUC, and perform Goodness of Fit Test.

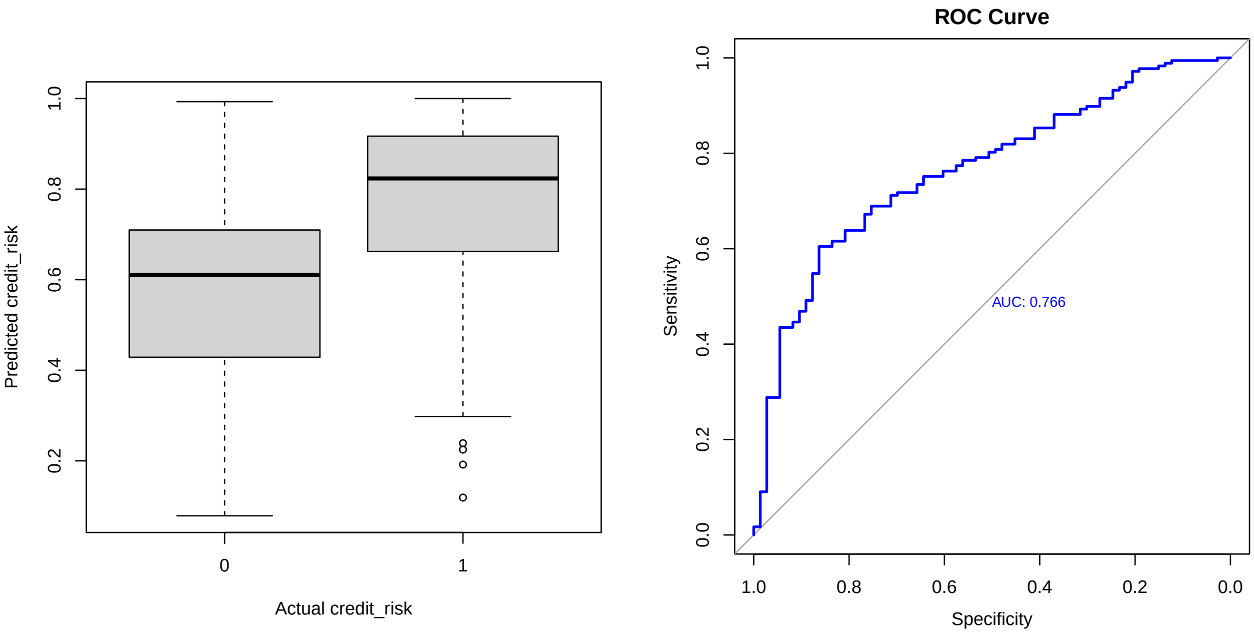
3.1 ROC Curve and AUC

```
pred.3 <- predict(bestmodel.3, newdata = testdata)
par(mfrow=c(2,2))
plot(testdata$credit_risk, inv.logit(pred.3), xlab = "Actual credit_risk", ylab = "Predicted credit_ris
roc(testdata$credit_risk~inv.logit(pred.3), plot=TRUE, main="ROC Curve", col="blue", print.auc=TRUE)
```

```
##
## 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
```



As we can see from the above results, the area under the ROC is 0.7659, which is fairly large. Also, the estimated probability of having good credit is lower when the actual credit risk is high compared to when the actual credit risk is low. Based on these two results, we can have some confidence on the robustness of the model.

3.2 Hosmer-Lemeshow Test

Now, we will perform Goodness of Fit Test. Since we are dealing with ungrouped data here, we will apply the Hosmer-Lemeshow Test.

```
saturated_model <- glm(credit_risk~status * duration * savings * property * age * other_installment_plan
                       family = binomial, data = traindata)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
hoslem.test(bestmodel.3$y, fitted(bestmodel.3), g=11)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  bestmodel.3$y, fitted(bestmodel.3)
## X-squared = 5.6322, df = 9, p-value = 0.7761
```

We can see that the p-value of the Hosmer-Lemeshow Test is 0.7761 which is much larger than the significance level $\alpha = 0.05$. Therefore, we fail to reject the null hypothesis and conclude that the selected model fits the data well.

3.3 Classification Table and Predictive Power

Next, we will analyse the predictive power of the selected model.

```
# Calculate the cutoff probability
n <- dim(testdata)[1]
prop <- sum(testdata$credit_risk == 1) / n
prop
```

```
## [1] 0.708
```

```
y <- (testdata$credit_risk == 1) * 1
predicted <- as.numeric(pred.3 > prop)
xtabs(~y + predicted)
```

```
##    predicted
## y     0   1
##   0  47  26
##   1  47 130
```

We can see that the cutoff probability is 0.708. This actually corresponds to the credit_risk odds ratio of 2.33 we got from Exploratory Data Analysis.

Based on the above classification table, we can calculate the followings:

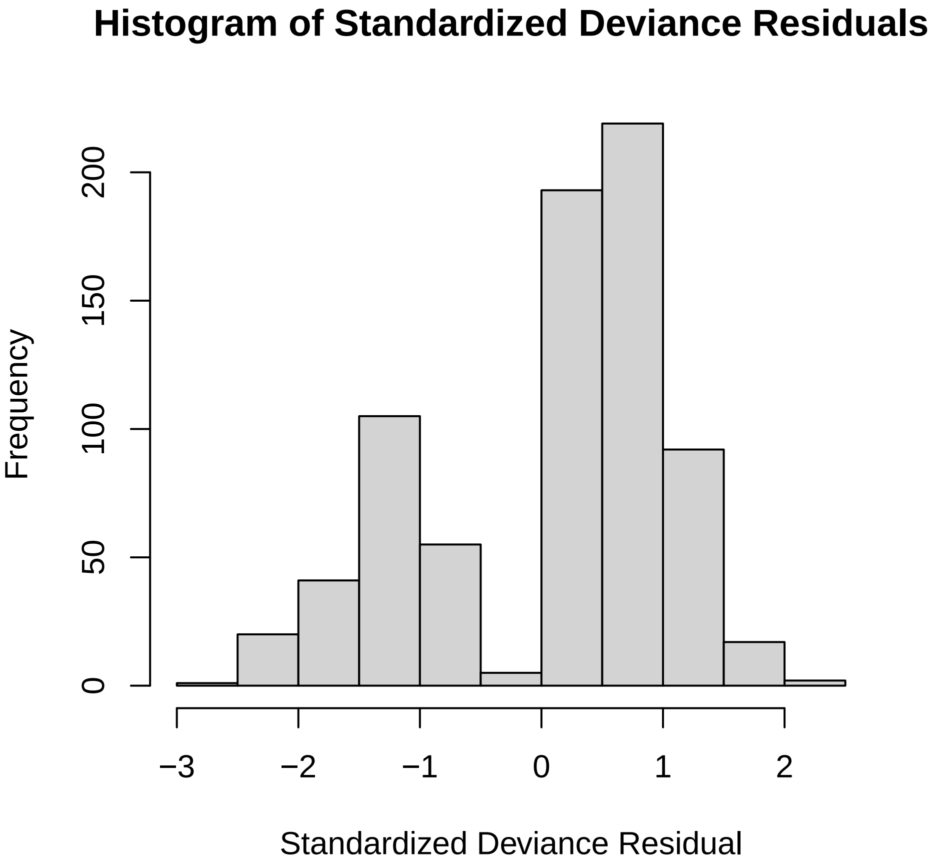
$$sensitivity = \frac{130}{130 + 47} = 0.7345$$

$$specificity = \frac{47}{47 + 26} = 0.6438$$

Since the model has a high sensitivity and specificity, we have strong confidence that the model fits the data well.

3.4 Residual Diagnostics

```r
hist(rstandard(bestmodel.3), main = "Histogram of Standardized Deviance Residuals",
     xlab = "Standardized Deviance Residual")
```



Based on the above residual histogram, we can see that there is no extreme value of residuals and the majority of the values is between -2 and 2. Therefore, we conclude that the selected model fits the data well.

In summary, we examined the ROC, AUC, Hosmer-Lemeshow Test and Predictive Power of the model. We found out the AUC is high, the p-value of Hosmer-Lemeshow Test is very large, the high sensitivity and high specificity of the model indicates a strong predictive power. All these clues show that the selected model is a robust model.