Predicting the Status of Credit

2023-04-04

Lu Zheng: Model Building, Model Selection

Yuxin Yao: Model Building, Model Selection

Zhiquan Cui 1005835857: Exploratory Data Analysis, Model Validation and Diagnostics

Contents

1 Load Required Libraries

```
library(dplyr)
library(insight)
library(knitr)
library(kableExtra)
library(ggplot2)
library(tidyverse)
library(corrplot)
library(patchwork)
library(rcompanion)
library(gridExtra)
library(boot)
library(pROC)
library(ROCR)
library(ResourceSelection)
```

2 Load Data & Inspect Variables

```
# Read the data
data <- read.csv("Credit.csv")
# Check the number of observations and number of variables
n <- nrow(data)
m <- ncol(data)
n
```

```
## [1] 1000
```

```
m
```

```
## [1] 21
```

```
# Check the data
kable(head(data[, 1:8]), format = "latex", align=rep("c", 8), booktabs=TRUE)
```

status	duration	credit_history	purpose	amount	savings	employment_duration	installment_rate
1	18	4	2	1049	1	2	4
1	9	4	0	2799	1	3	2
2	12	2	9	841	2	4	2
1	12	4	0	2122	1	3	3
1	12	4	0	2171	1	3	4
1	10	4	0	2241	1	2	1

```
kable(head(data[, 9:14]), format = "latex", align=rep("c", 6), booktabs=TRUE)
```

personal_status_sex	other_debtors	present_residence	property	age	other_installment_plans
2	1	4	2	21	3
3	1	2	1	36	3
2	1	4	1	23	3
3	1	2	1	39	3
3	1	4	2	38	1
3	1	3	1	48	3

```
kable(head(data[, 15:21]), format = "latex", align=rep("c", 7), booktabs=TRUE)
```

housing	number_credits	job	people_liable	telephone	foreign_worker	credit_risk
1	1	3	2	1	2	1
1	2	3	1	1	2	1
1	1	2	2	1	2	1
1	2	2	1	1	1	1
2	2	2	2	1	1	1
1	2	2	1	1	1	1

```
# Check invalid or missing values
anyNA(data)
```

```
## [1] FALSE
```

```
# Check the data type of each column
sapply(data, class)
```

```
##                  status                duration          credit_history
##               "integer"               "integer"               "integer"
##                 purpose                  amount                 savings
##               "integer"               "integer"               "integer"
##     employment_duration        installment_rate     personal_status_sex
##               "integer"               "integer"               "integer"
##           other_debtors       present_residence                property
##               "integer"               "integer"               "integer"
##                     age other_installment_plans                 housing
##               "integer"               "integer"               "integer"
##          number_credits                     job           people_liable
##               "integer"               "integer"               "integer"
##               telephone          foreign_worker             credit_risk
##               "integer"               "integer"               "integer"
```

As we can see from the above outputs, there is no NaN values so the data is clean. And all of the columns are of type integer. Some of them are quantitative variable while some of them are qualitative variables. Here is a summary of the variables:

- status: status of the debtor's checking account with the bank (categorical)
- duration: credit duration in months (quantitative)
- credit_history: history of compliance with previous or concurrent credit contracts (categorical)
- purpose: purpose for which the credit is needed (categorical)

- amount: credit amount in DM (quantitative; result of monotonic transformation; actual data and type of transformation unknown)
- savings: debtor's savings (categorical)
- employment_duration: duration of debtor's employment with current employer (ordinal; discretized quantitative)
- installment_rate: credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative)
- personal_status_sex: combined information on sex and marital status (categorical)
- other_debtors: is there another debtor or a guarantor for the credit? (categorial)
- present_residence: length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative)
- property: the debtor's most valuable property (ordinal)
- age: age in years (quantitative)
- other_installment_plans: installment plans from providers other than the credit-giving bank (categorical)
- housing: type of housing the debtor lives in (categorical)
- number_credits: number of credits including the current one the debtor has (or had) at the bank (ordinal; discretized quantitative)
- job: quality of debtor's job (ordinal)
- people_liable: number of persons who financially depend on the debtor (binary; discretized quantitative)
- telephone: is there a telephone landline registered on the debtor's name? (binary)
- foreign_ worker: is the debtor a foreign worker? (binary)
- credit_risk: has the credit contract been complied with (good) or not (bad)? (binary)

We can see that the **quantitative variables** include duration, amount and age, while **qualitative variables** include status, credit_history, purpose, savings, employment_duration, installment_rate, personal_status_sex, other_debtors, present_residence, property, other_installment_plans, housing, number_credits, job, people_liable, telephone, foreign_worker and credit_risk.

3 Univariate Data Analysis & Visualization

3.1 Histogram of Quantitative Variables

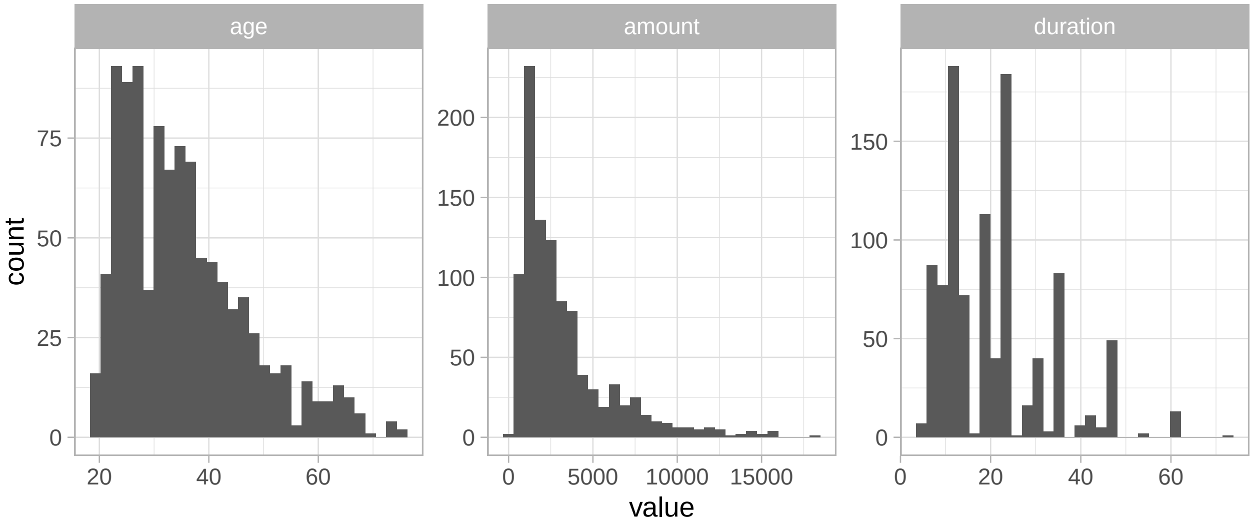
First we will perform univariate analysis on each of the variables and look at their distribution. Here is the summary statistics:

```
quant_vars <- c("duration", "amount", "age")
qual_vars <- c("status", "credit_history", "purpose", "savings", "employment_duration",
               "installment_rate", "personal_status_sex", "other_debtors", "present_residence",
               "property", "other_installment_plans", "housing", "number_credits", "job",
               "people_liable", "telephone", "foreign_worker", "credit_risk")
summary(data[, quant_vars])
```

```
##     duration        amount            age       
##  Min.   : 4.0   Min.   :  250   Min.   :19.00  
##  1st Qu.:12.0   1st Qu.: 1366   1st Qu.:27.00  
##  Median :18.0   Median : 2320   Median :33.00  
##  Mean   :20.9   Mean   : 3271   Mean   :35.54  
##  3rd Qu.:24.0   3rd Qu.: 3972   3rd Qu.:42.00  
##  Max.   :72.0   Max.   :18424   Max.   :75.00  
```

Next, let us check the histograms of the quantitative variables:

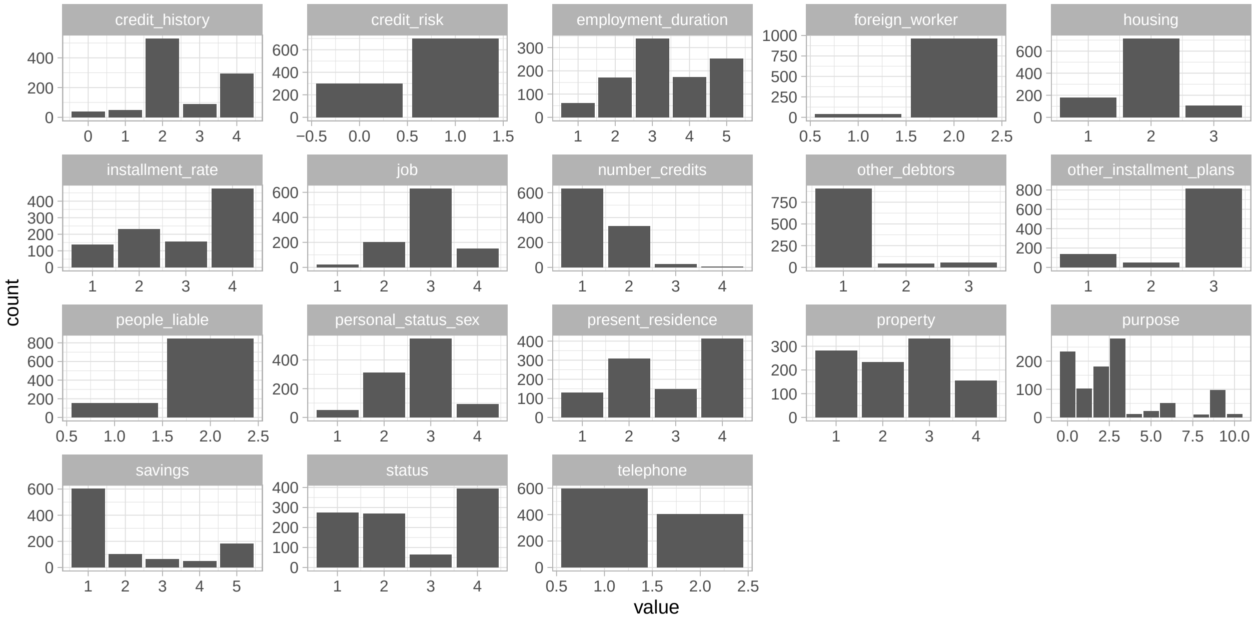
```
data[, quant_vars] %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram() +
    theme_light()
```



3.2 Barplot of Qualitative Variables

Then, let us check the barplots of qualitative variables:

```
data[, qual_vars] %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_bar() +
    theme_light()
```

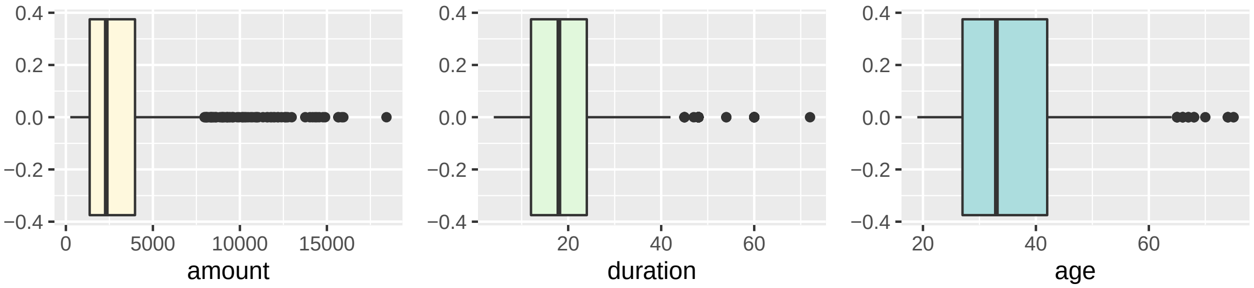


As we can see, the response variable credit risk is a binary variable while we have more than 2 predictors. This indicates that it is a good idea to use Multiple Logistic Regression as our model.

3.3 Boxplot of Quantitative Variables

After checking the histograms and barplots, we will check the boxplots of the quantitative variables. Here we will not check barplots for qualitative variables because it only makes sense to examine the median, first and third quartiles and maximum value for quantitative variables.

```r
g1 <- ggplot(data, aes(x = amount)) + geom_boxplot(fill="#FEF8DD")
g2 <- ggplot(data, aes(x = duration)) + geom_boxplot(fill="#E1F8DC")
g3 <- ggplot(data, aes(x = age)) + geom_boxplot(fill="#ACDDDE")
g1 + g2 + g3
```



From the above box plots, we can see that there are a few outliers for the variable amount. If we look at the histogram of variable amount, we can see that it is a right skewed distribution with a long right tail, which results in these outliers.

3.4 Sample Odds of Binary Variables

For binary variables people_liable, telephone, foreign_worker and credit_risk, we can calculate and interpret the sample odds:

```r
binary_var <- c("Statistics", "people_liable", "telephone", "foreign_worker", "credit_risk")
odds <- c("Sample Odds")
for (var in binary_var[2:5]) {
  if (var == "credit_risk") {y <- sum(data[, var] == 1)}
  else {y <- sum(data[, var] == 2)}
  n <- length(data[, var])
  odds <- append(odds, round(y / (n - y), 2))
}
kable(data.frame(t(odds)), col.names = binary_var, format = "latex") %>%
  kable_styling(position = "center", latex_options = "hold_position") %>% row_spec(0, bold = TRUE)
```

Statistics	**people_liable**	**telephone**	**foreign_worker**	**credit_risk**
Sample Odds	5.45	0.68	26.03	2.33

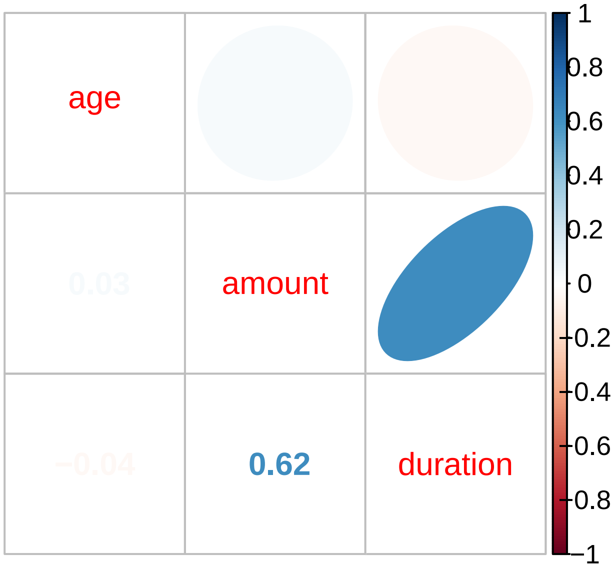
Based on our sample, the estimated probability of a person to have good credit is 2.33 times as likely as having a bad credit. Similarly, the estimated probability of a person to have a telephone landline registered on his/her name is 0.68 times as likely as not having such a telephone landline.

4 Multivariate Data Analysis & Visualization

4.1 Quantitative Variable

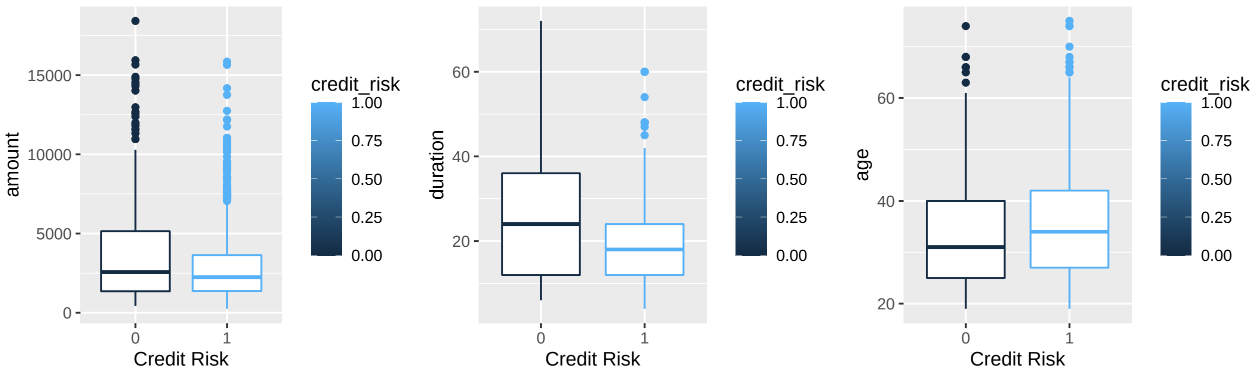
First, let us look at the correlation plots of the quantitative variables.

```r
corrplot.mixed(cor(data[quant_vars]), lower='number', upper='ellipse', order='AOE')
```



From the above correlation plot, we can see that the correlation coefficient between amount and duration is as high as 0.62, which indicates a strong positive correlation between the two variables. This also makes sense intuitively because the longer credit duration one has in months, he/she will have a higher chance to build up his/her credit and obtain a higher credit amount. Similarly, if one has a high credit amount, then he/she is more likely to have a long credit duration. In order to avoid multicollinearity, we will consider droping one of amount and duration in our model. However, before making a decision, we shall examine the side by side box plots.

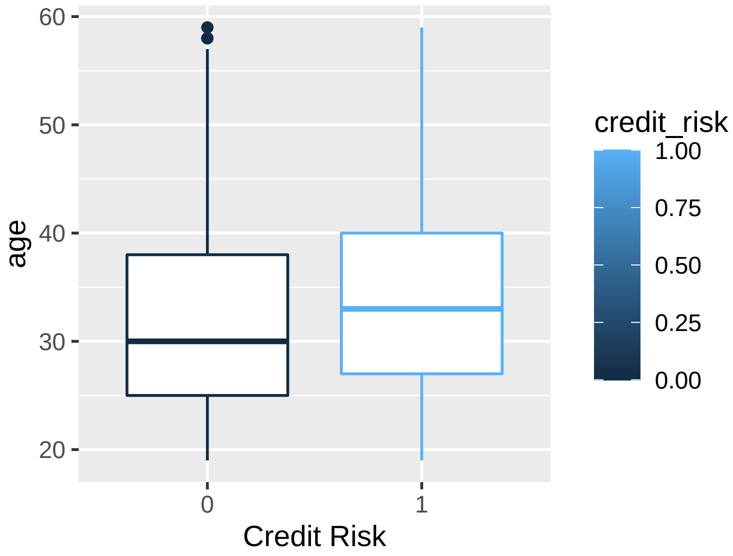
```
g1 <- ggplot(data, aes(x=as.factor(credit_risk), y=amount, color=credit_risk)) +
      geom_boxplot() + xlab("Credit Risk")
g2 <- ggplot(data, aes(x=as.factor(credit_risk), y=duration, color=credit_risk)) +
      geom_boxplot() + xlab("Credit Risk")
g3 <- ggplot(data, aes(x=as.factor(credit_risk), y=age, color=credit_risk)) +
      geom_boxplot() + xlab("Credit Risk")

grid.arrange(g1, g2, g3, nrow=1)
```



From the above side by side box plots, we can see that for variables duration and age, there are significant differences on the box plots between two levels of credit risks. This indicates a significant association between credit risk and these two variables. However, we don't see a significant difference between two credit risk levels for variable amount.

Therefore, we will drop the variable amount.

```
data <- subset(data, age < 60)
ggplot(data, aes(x=as.factor(credit_risk), y=age, color=credit_risk)) +
      geom_boxplot() + xlab("Credit Risk")
```



4.2 Qualitative Variables

After examining the quantitative variables, we will now look at the qualitative variables. Since they are not continuous and numeric data, we should not use the same methodology as above. Instead, we will use Pearson's Chi-sq Test of Indepence and Cramer's V designed for qualitative variables to examine the data.

```r
Pearson_chisq_test <- data.frame(matrix(0, ncol = length(qual_vars),
                                        nrow = length(qual_vars)), row.names = qual_vars)
colnames(Pearson_chisq_test) <- qual_vars
for (var in qual_vars) {
  for (var_2 in qual_vars) {
    test <- chisq.test(table(data[, var], data[, var_2]), simulate.p.value = TRUE)
    Pearson_chisq_test[var, var_2] <- test$p.value
  }
}
```

	status	credit_history	purpose	savings	employment_duration	installment_rate
status	0.0004998	0.0004998	0.0009995	0.0004998	0.0084958	0.4212894
credit_history	0.0004998	0.0004998	0.0004998	0.2603698	0.0054973	0.5512244
purpose	0.0004998	0.0004998	0.0004998	0.0754623	0.0084958	0.0004998
savings	0.0004998	0.2803598	0.0789605	0.0004998	0.0294853	0.5017491
employment_duration	0.0104948	0.0034983	0.0084958	0.0274863	0.0004998	0.0004998
installment_rate	0.4207896	0.5582209	0.0009995	0.5067466	0.0004998	0.0004998
personal_status_sex	0.0834583	0.0474763	0.0004998	0.5032484	0.0004998	0.0004998
other_debtors	0.0014993	0.0779610	0.0009995	0.0394803	0.2108946	0.9355322
present_residence	0.0019990	0.1154423	0.0054973	0.1434283	0.0004998	0.4682659
property	0.1229385	0.2043978	0.0004998	0.0619690	0.0004998	0.4667666
other_installment_plans	0.4227886	0.0004998	0.0044978	0.9995002	0.2723638	0.4137931
housing	0.0039980	0.0879560	0.0004998	0.8875562	0.0004998	0.1064468
number_credits	0.0159920	0.0004998	0.2883558	0.1469265	0.0049975	0.8285857
job	0.1724138	0.2863568	0.0004998	0.4872564	0.0004998	0.0369815
people_liable	0.0994503	0.0814593	0.0014993	0.8230885	0.0209895	0.1779110
telephone	0.0539730	0.2753623	0.0004998	0.0239880	0.0009995	0.5422289
foreign_worker	0.1714143	0.4567716	0.0084958	0.8860570	0.6221889	0.0069965
credit_risk	0.0004998	0.0004998	0.0004998	0.0004998	0.0009995	0.1214393

	personal_status_sex	other_debtors	present_residence	property	other_installment_plans	housing
status	0.0659670	0.0029985	0.0039980	0.1134433	0.4357821	0.0039980
credit_history	0.0589705	0.0779610	0.1154423	0.1959020	0.0004998	0.0819590
purpose	0.0004998	0.0004998	0.0064968	0.0004998	0.0029985	0.0004998
savings	0.4892554	0.0329835	0.1414293	0.0659670	0.9980010	0.8910545
employment_duration	0.0004998	0.2113943	0.0004998	0.0004998	0.2543728	0.0004998
installment_rate	0.0014993	0.9350325	0.4672664	0.4787606	0.4047976	0.1109445
personal_status_sex	0.0004998	0.6016992	0.0009995	0.0004998	0.3953023	0.0004998
other_debtors	0.6251874	0.0004998	0.5862069	0.0004998	0.2768616	0.1544228
present_residence	0.0004998	0.5962019	0.0004998	0.0004998	0.7506247	0.0004998
property	0.0004998	0.0004998	0.0004998	0.0004998	0.0764618	0.0004998
other_installment_plans	0.4067966	0.2768616	0.7541229	0.0759620	0.0004998	0.0024988
housing	0.0004998	0.1569215	0.0004998	0.0004998	0.0019990	0.0004998
number_credits	0.1189405	0.8945527	0.0059970	0.1159420	0.0164918	0.0169915
job	0.0094953	0.0714643	0.7831084	0.0004998	0.0389805	0.0044978
people_liable	0.0004998	0.4607696	0.1219390	0.0274863	0.1034483	0.0004998
telephone	0.0379810	0.0424788	0.0604698	0.0004998	0.1769115	0.0154923
foreign_worker	0.0904548	0.0024988	0.4557721	0.0004998	0.8870565	0.0429785
credit_risk	0.0159920	0.0194903	0.8485757	0.0004998	0.0029985	0.0004998

Based on the above table, we conclude that the following predictors are dependent to most of the predictors with $\alpha = 0.05$ according to Pearson's Chi-sq Test of Independence, and we consider dropping these predictors:

- job
- credit_history

	number_credits	job	people_liable	telephone	foreign_worker
status	0.0144928	0.1884058	0.0999500	0.0499750	0.1744128
credit_history	0.0004998	0.2738631	0.0684658	0.2718641	0.4487756
purpose	0.2863568	0.0004998	0.0019990	0.0004998	0.0074963
savings	0.1439280	0.4652674	0.8055972	0.0224888	0.9070465
employment_duration	0.0019990	0.0004998	0.0194903	0.0004998	0.6201899
installment_rate	0.8510745	0.0314843	0.1594203	0.5637181	0.0064968
personal_status_sex	0.1434283	0.0089955	0.0004998	0.0354823	0.0894553
other_debtors	0.8935532	0.0774613	0.4587706	0.0539730	0.0009995
present_residence	0.0059970	0.7956022	0.1224388	0.0614693	0.4687656
property	0.1114443	0.0004998	0.0264868	0.0004998	0.0004998
other_installment_plans	0.0144928	0.0339830	0.0994503	0.1854073	0.9015492
housing	0.0114943	0.0049975	0.0004998	0.0174913	0.0364818
number_credits	0.0004998	0.1869065	0.0084958	0.0134933	0.6771614
job	0.1744128	0.0004998	0.0004998	0.0004998	0.0489755
people_liable	0.0074963	0.0014993	0.0004998	0.6361819	0.0659670
telephone	0.0159920	0.0004998	0.6471764	0.0004998	0.0294853
foreign_worker	0.6941529	0.0494753	0.0679660	0.0334833	0.0004998
credit_risk	0.3273363	0.5002499	0.9310345	0.2728636	0.0144928

- purpose
- employment_duration
- housing
- people_liable

Also, we can see that the following predictors have very weak association with the response variable:

- installment_rate
- personal_status_sex
- other_debtors
- present_residence
- number_credits
- job
- people_liable
- telephone
- foreign_worker

To summarize, the variables we will use in model building are:

- status
- duration
- savings
- property
- age
- other_installment_plans

5 Model Building and Model Selection

5.1 Data Preparation

```r
# Transform categorical variables
data$credit_risk = as.factor(data$credit_risk)
data$status  = as.factor(data$status)
data$savings = as.factor(data$savings)
```

```r
# data$property = as.ordered(data$property)
data$other_installment_plans = as.factor(data$other_installment_plans)
```

Here we treat property as a quantitative variable as it is an ordinal variable. We will split the dataset into training set and testing set. Here, the split rate is set to be 0.75.

```r
set.seed(1006742107)

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

5.2 Main effect model

5.2.1 Forward method

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

5.2.2 Backward method

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

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)

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

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

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, 
summary(bestmodel.1)
```

5.2.3 Mannual Selection

From the above output, we can see that savings2 and Savings3, other_installment_plans2 seem to be insignificant. Therefore, we may consider combining level 1, 2 and 3 of savings. This also makes sense intuitively because people with no savings account, those with less than 100 DM in the savings account and those with between 100 DM and 500 DM might be treated similarly on the determination of their credit status. As for other_installment_plans, maybe only 'yes' or 'no' to the question makes a difference.

```
levels(traindata$savings) <- c(2, 2, 2, 4, 5)
levels(traindata$other_installment_plans) <- c(2, 2, 3)
bestmodel.1 = step(glm(credit_risk ~ 1, family = binomial, data = traindata), scope =
       ~status + duration + savings + property + age +
       other_installment_plans, direction = "forward", test = "Chisq")
```

```
summary(bestmodel.1)
```

```
##
## Call:
## glm(formula = credit_risk ~ status + duration + property + savings +
##     age + other_installment_plans, family = binomial, data = traindata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6294  -0.9783   0.4638   0.8147   1.6829
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.140437   0.469991  -0.299 0.765087
## status2                   0.567789   0.220175   2.579 0.009914 **
## status3                   0.997869   0.394480   2.530 0.011420 *
## status4                   1.939138   0.238316   8.137 4.06e-16 ***
## duration                 -0.029657   0.007716  -3.844 0.000121 ***
## property                 -0.283947   0.093488  -3.037 0.002387 **
## savings4                  1.035122   0.529097   1.956 0.050419 .
## savings5                  0.617737   0.255074   2.422 0.015444 *
## age                       0.025723   0.010407   2.472 0.013452 *
## other_installment_plans3  0.481794   0.220477   2.185 0.028872 *
## ---
## 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: 736.63  on 702  degrees of freedom
## AIC: 756.63
##
## Number of Fisher Scoring iterations: 4
```

Now the combined levels become significat predictors. The AIC of the model also dropped. Next, we will try to include some interaction terms and see whether we can obtain a better model.

5.3 Interaction model

5.3.1 StepAIC Algorithm

```
bestmodel.2 <- step(glm(credit_risk ~ 1, family = binomial, data = traindata),
                    scope = ~status * duration * savings * property * age *
                      other_installment_plans, direction = "both", test = "Chisq")
```

Check the selected interaction model.

```
summary(bestmodel.2)
```

```
##
## Call:
## glm(formula = credit_risk ~ status + duration + property + savings +
##     age + other_installment_plans + status:property + property:age +
##     duration:savings + status:savings + savings:age + age:other_installment_plans,
##     family = binomial, data = traindata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3222  -0.8325   0.4637   0.6904   2.2405
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -4.892e-01  1.319e+00  -0.371   0.7107
## status2                      1.897e-01  6.153e-01   0.308   0.7578
## status3                     -1.508e+00  9.944e-01  -1.517   0.1293
## status4                      7.603e-01  6.275e-01   1.212   0.2257
## duration                    -3.711e-02  9.063e-03  -4.095 4.23e-05 ***
## property                     2.551e-01  3.782e-01   0.674   0.5001
## savings4                     1.979e+01  6.049e+02   0.033   0.9739
## savings5                    -5.662e-01  1.142e+00  -0.496   0.6199
## age                          5.837e-02  3.634e-02   1.606   0.1082
## other_installment_plans3    -7.602e-01  9.618e-01  -0.790   0.4293
## status2:property             1.129e-01  2.325e-01   0.486   0.6272
## status3:property             1.119e+00  4.137e-01   2.706   0.0068 **
## status4:property             5.320e-01  2.394e-01   2.222   0.0263 *
## property:age                -2.364e-02  1.021e-02  -2.315   0.0206 *
## duration:savings4            3.007e-02  5.436e-02   0.553   0.5802
## duration:savings5            5.242e-02  2.301e-02   2.278   0.0227 *
## status2:savings4            -1.536e+01  6.049e+02  -0.025   0.9797
## status3:savings4            -6.964e-01  1.028e+03  -0.001   0.9995
## status4:savings4            -1.519e+01  6.049e+02  -0.025   0.9800
## status2:savings5             1.644e+00  7.192e-01   2.286   0.0222 *
## status3:savings5             5.125e-01  1.116e+00   0.459   0.6459
## status4:savings5             2.521e-01  6.752e-01   0.373   0.7089
## savings4:age                -1.322e-01  6.402e-02  -2.065   0.0389 *
## savings5:age                -2.006e-02  3.071e-02  -0.653   0.5135
## age:other_installment_plans3  4.030e-02  2.801e-02   1.439   0.1503
## ---
## 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: 694.67  on 687  degrees of freedom
## AIC: 744.67
##
## Number of Fisher Scoring iterations: 14
```

5.3.2 Manual Selection

Since the selected model doesn't give significant predictors, we will try adding the interaction term mannually.

```
fit1 <- glm(credit_risk~status * (duration + savings + property + age + other_installment_plans),
            family = binomial, data = traindata)
fit2 <- glm(credit_risk~status + savings * (duration + property + age) + other_installment_plans,
            family = binomial, data = traindata)
fit3 <- glm(credit_risk~status + property * (savings + duration + age + other_installment_plans),
            family = binomial, data = traindata)
fit4 <- glm(credit_risk~status + age * (savings + duration + property + other_installment_plans),
            family = binomial, data = traindata)
fit5 <- glm(credit_risk~status + other_installment_plans * (savings + duration + property + age),
            family = binomial, data = traindata)
fit6 <- glm(credit_risk~status + savings + duration + property + age + other_installment_plans,
            family = binomial, data = traindata)
bestmodel.1 <- fit2
```

Since the above model has a lower AIC and the majority of its predictors are significant, we decide to choose this as our final model.

5.4 Final Model

```
summary(bestmodel.1)
```

```
##
## Call:
## glm(formula = credit_risk ~ status + savings * (duration + property +
##     age) + other_installment_plans, family = binomial, data = traindata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4182  -0.8924   0.4444   0.7992   1.7706
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.232135   0.512469  -0.453  0.65057
## status2                  0.603676   0.223824   2.697  0.00699 **
## status3                  1.035245   0.400040   2.588  0.00966 **
## status4                  1.983709   0.243616   8.143 3.86e-16 ***
## savings4                 7.636093   3.747960   2.037  0.04161 *
## savings5                -0.129658   1.125252  -0.115  0.90827
## duration                 -0.039856   0.008843  -4.507 6.57e-06 ***
## property                 -0.291570   0.105396  -2.766  0.00567 **
## age                       0.033168   0.011777   2.816  0.00486 **
## other_installment_plans3  0.568023   0.226170   2.511  0.01202 *
## savings4:duration         0.063075   0.057578   1.095  0.27331
## savings5:duration         0.048758   0.021487   2.269  0.02326 *
## savings4:property        -1.237370   0.708441  -1.747  0.08071 .
## savings5:property         0.188123   0.259774   0.724  0.46896
## savings4:age             -0.125691   0.059822  -2.101  0.03563 *
## savings5:age             -0.026064   0.028452  -0.916  0.35963
## ---
## 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: 720.95  on 696  degrees of freedom
## AIC: 752.95
##
## Number of Fisher Scoring iterations: 6
```

6 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.

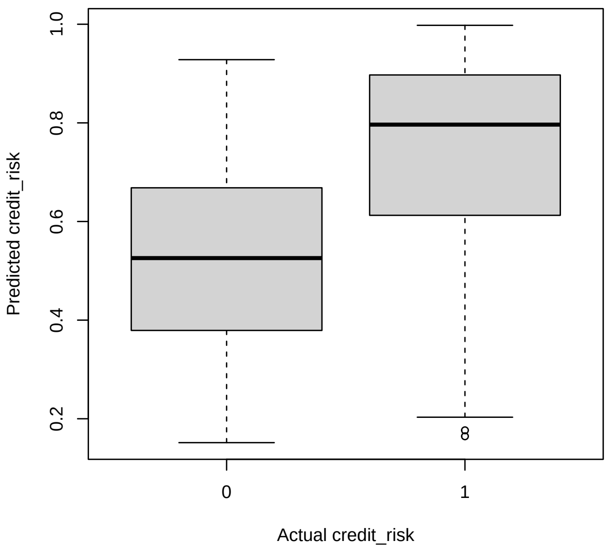
6.1 ROC Curve and AUC

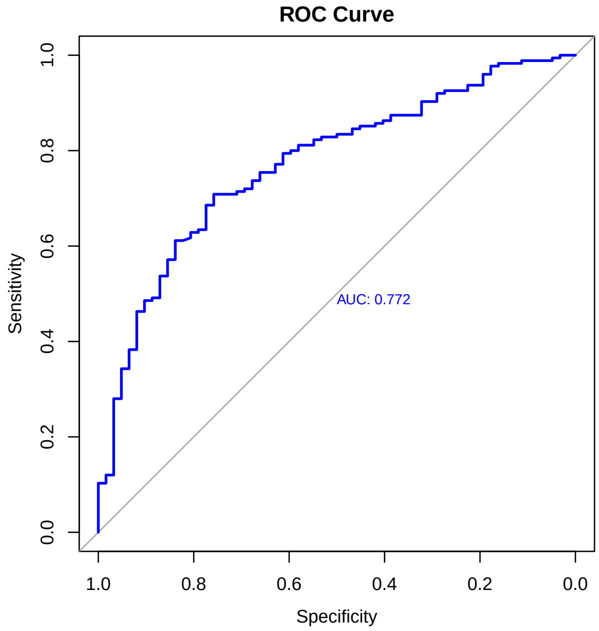
```
levels(testdata$savings) <- c(2, 2, 2, 4, 5)
levels(testdata$other_installment_plans) <- c(2, 2, 3)
pred.3 <- predict(bestmodel.1, newdata = testdata)
par(mfrow=c(2,2))
plot(testdata$credit_risk, inv.logit(pred.3), xlab = "Actual credit_risk", ylab = "Predicted credit_risk
roc(testdata$credit_risk~inv.logit(pred.3), plot=TRUE, main="ROC Curve", col="blue", print.auc=TRUE)
```

```
##
## Call:
## roc.formula(formula = testdata$credit_risk ~ inv.logit(pred.3),     plot = TRUE, main = "ROC Curve",
##
## Data: inv.logit(pred.3) in 62 controls (testdata$credit_risk 0) < 175 cases (testdata$credit_risk 1)
## Area under the curve: 0.7721
```

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

```
## Area under the curve: 0.7721
```





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.

6.2 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_pla
                       family = binomial, data = traindata)
```

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

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

```
hoslem.test(bestmodel.1$y, fitted(bestmodel.1), g=11)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  bestmodel.1$y, fitted(bestmodel.1)
## X-squared = 13.01, df = 9, p-value = 0.1622
```

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.

6.3 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] 0.7383966
```

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

```
##    predicted
## y     0   1
##   0  47  15
##   1  54 121
```

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:

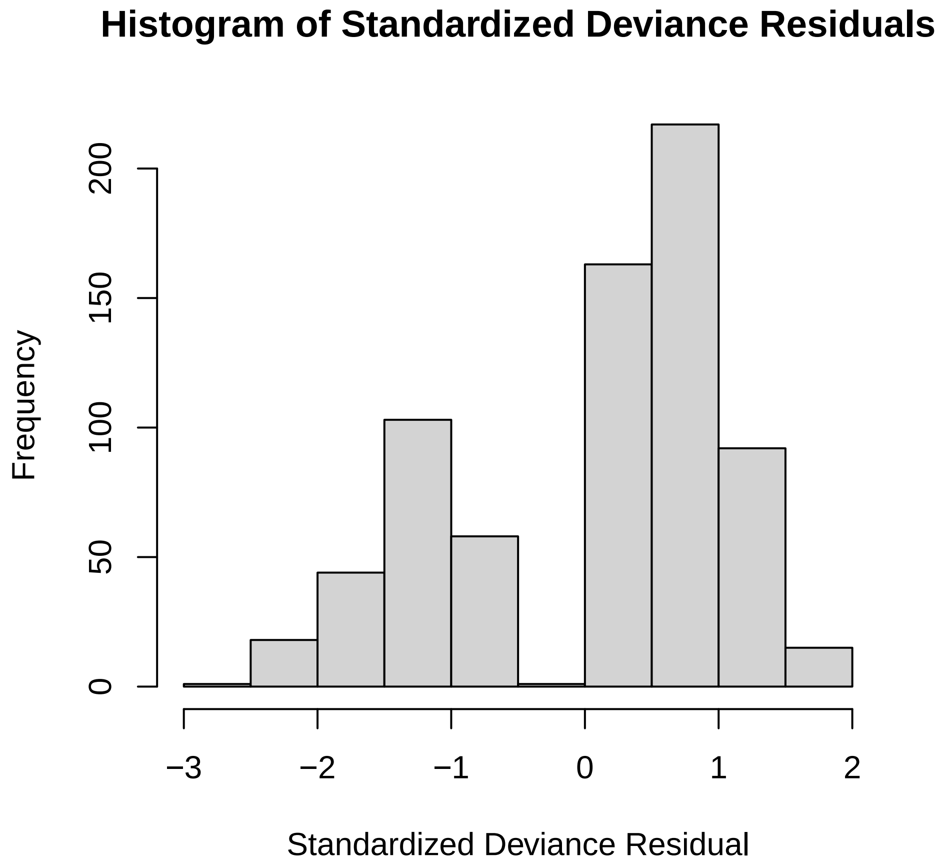
$$sensitivity = \frac{121}{121+54} = 0.6914$$

$$specificity = \frac{47}{47+15} = 0.7581$$

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

6.4 Residual Diagnostics

```
hist(rstandard(bestmodel.1), 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.

7 Discussion and Conclusion

8 References