Predicting the Status of Credit

2023-04-10

Lu Zheng 1006742107: Model Building, Model Selection

Yuxin Yao 1006917516: Model Building, Model Selection

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

Background and Significance

Predicting credit status involves assessing a person's ability to repay a loan based on factors such as credit score, income, debt-to-income ratio, employment history, and other financial indicators (Fontinelle, 2023). Financial institutions use predictive models and algorithms to make informed lending decisions and minimize the risk of default. Credit scores are a key factor in predicting credit status, and lenders also take into account other factors that can affect a borrower's ability to repay a loan. Overall, predicting credit status is a complex process that requires assessing a range of factors and weighing the potential risks and benefits of each loan. A study builts a linier discriminant function and shows that credit status is related to factors including checking amount and other installment plans (Aidi & Sari, 2012).

In this research, we aim to determine the main factors that influence one's credit status using a credit dataset containing 1000 customer profiles. From more than 15 predictors, we will select the most significant ones to build a multiple logistic regression model predicting the probability that a person has good credit status. Here is a description 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)

Exploratory Data Analysis

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

```
m <- ncol(data)
n
# Check invalid or missing values
anyNA(data)
# Check the data type of each column
sapply(data, class)
```

After performing the above basic data checks, we can see that 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.

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.

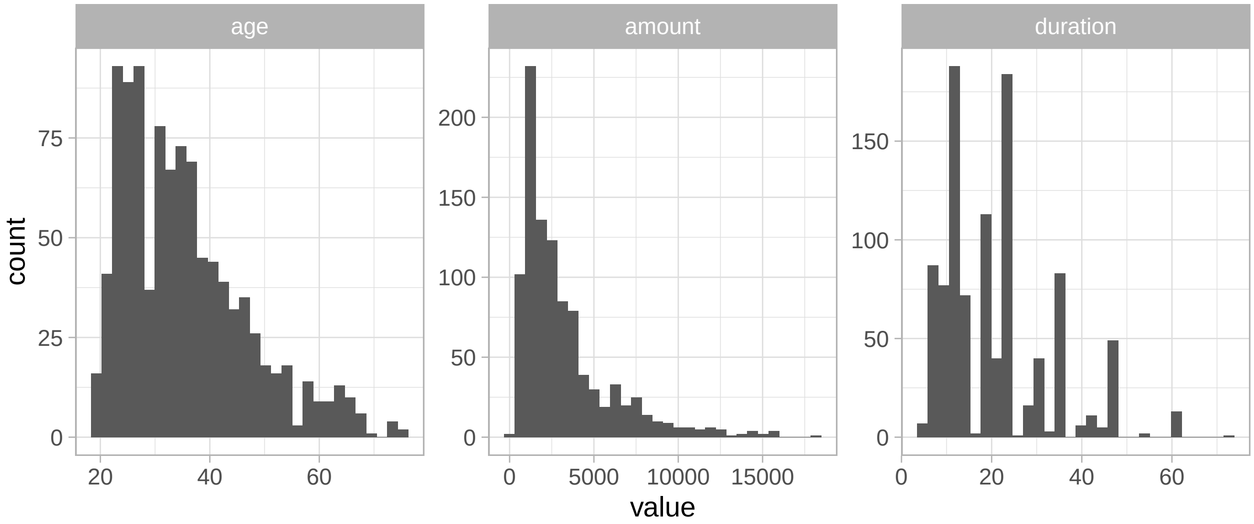
Univariate Data Analysis & Visualization

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:

```
##     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  
```

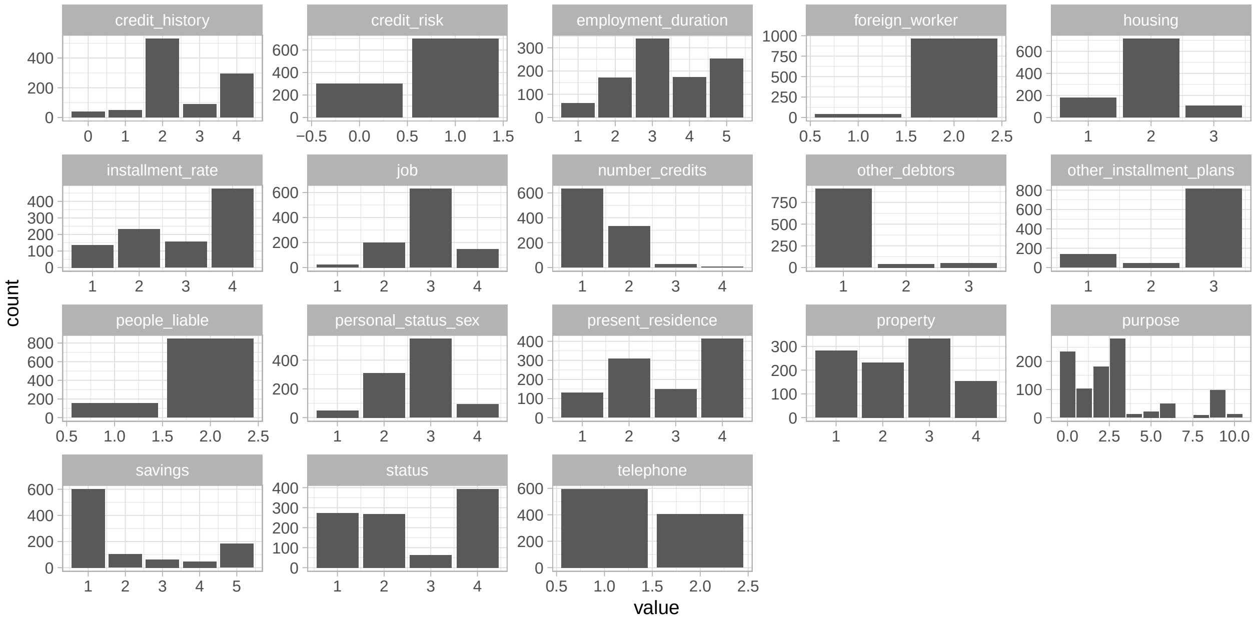
Next, let us check the histograms of the quantitative variables:



Barplot of Qualitative Variables

Then, let us check the barplots of qualitative variables:

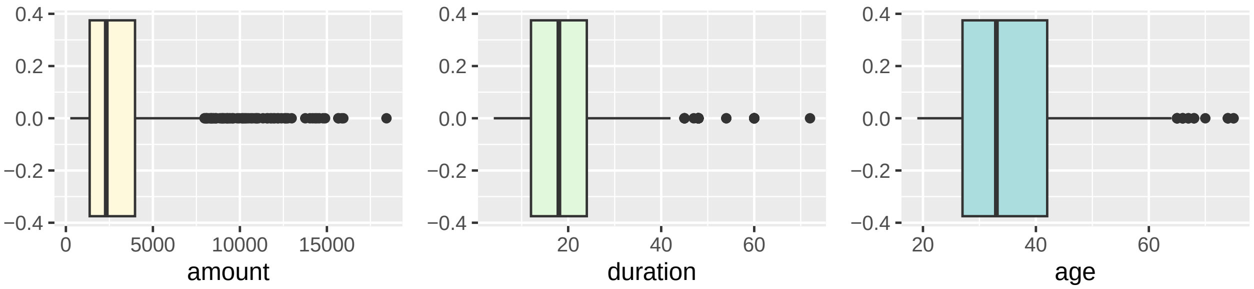
```
data[, qual_vars] %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_bar() +
    theme_light()
```



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 try Multiple Logistic Regression as our model.

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.



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.

Sample Odds of Binary Variables

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

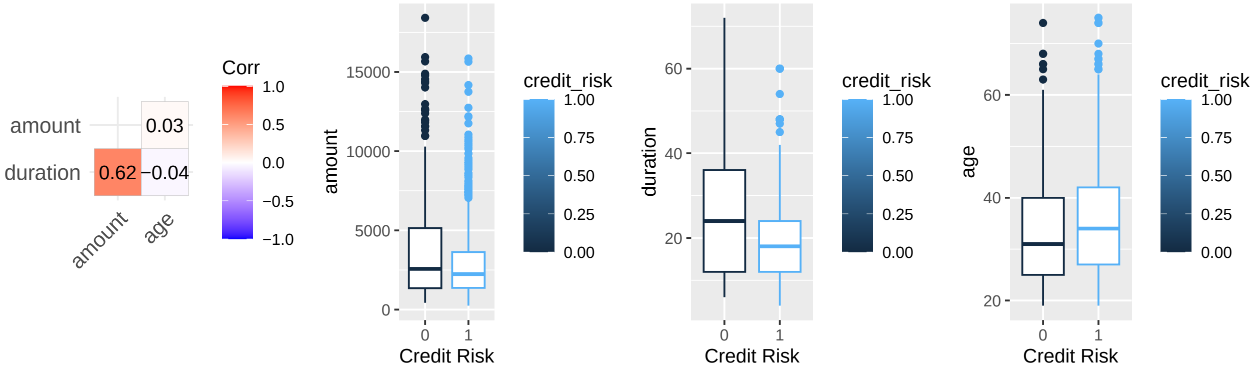
Statistics	people_liable	telephone	foreign_worker	credit_risk
Sample Odds	5.45	0.68	26.03	2.33

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.

Multivariate Data Analysis & Visualization

Quantitative Variable

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



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.

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.

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 group the data and then use Pearson's Chi-sq Test of Indepence designed for qualitative variables to examine the data.

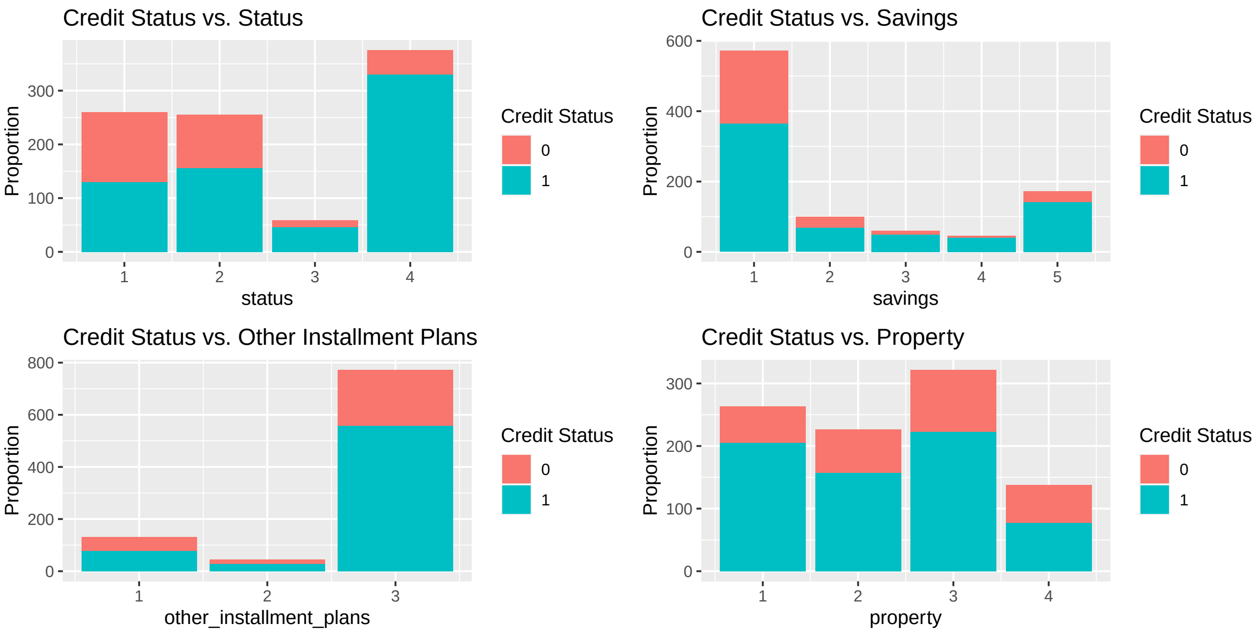
```
Pearson_chisq_test <- data.frame(matrix(0, ncol = length(qual_vars),
                                nrow = length(qual_vars)), row.names = qual_vars)
colnames(Pearson_chisq_test) <- qual_vars
for (var in qual_vars) {
  for (var_2 in qual_vars) {
    test <- chisq.test(table(data[, var], data[, var_2]),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

Based on the p-value table, we conclude that the following predictors are dependent to most of the predictors with $\alpha = 0.05$ because of their small p-values, and we consider dropping these predictors: job, credit_history, purpose, employment_duration, housing and people_liable.

Also, we can see that the following predictors have very weak association with the response variable because of a large p-value: installment_rate, personal_status_sex, other_debtors, present_residence, number_credits, job, people_liable, telephone, foreign_worker.

To summarize, the qualitative variables we will use in model building are: status, savings, property and other_installment_plans, while the quantitative variables are duration and age.

Now, let us look at their bar plots with the response variable to visualize the differences of responese variable between different levels.



Model Building and Model Selection

Data Preparation

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

```
set.seed(1006742107)
n = nrow(data)
index = sample(n, round(0.75 * n), replace = FALSE)
traindata = data[index, ]
testdata = data[-index, ]
```

Main effect model

Forward Selection & Backward Elimination

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

We find that both forward selection and backward elimination choose the same model which includes all of the given predictors.

Mannual Selection

However, from the above selected model, we see that savings2 and Savings3, other_installment_plans2 seem to be insignificant. We think the reason could be these levels are correlated to each other and results in multicollinearity. 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")
```

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.

Interaction model

Forward Selection & Backward Elimination

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

We find that the selected model by the algorithm gives lot of insignificant predictors. Therefore, we decide to perform mannual selection to eliminate the problems of multicollinearity.

Manual Selection & Final Model

Here, we try to add interaction term between variables one by one and remove those insignificant predictors.

```
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
summary(fit2)
```

```
## 
## 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
```

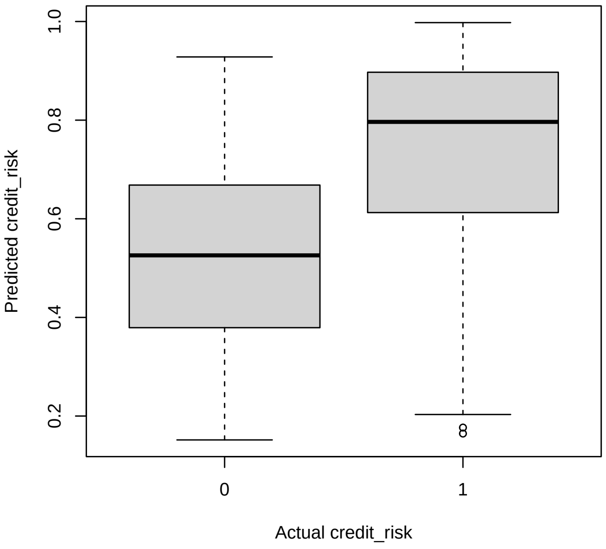
After comparing all the mannually constructed model, we decided to choose model fit2 as our final model because most of its predictors are significant and its AIC is lower than the selected main effect model.

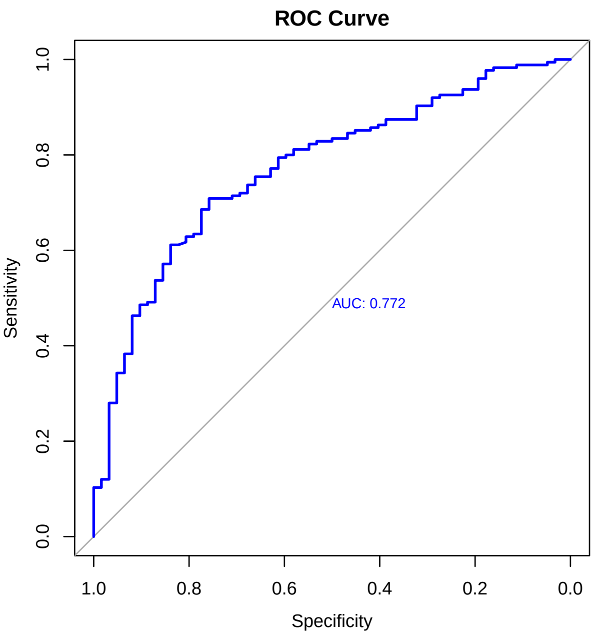
Model Validation and Diagnostics

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.

ROC Curve and AUC

```
##
## 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
```





As we can see from the above results, the area under the ROC is 0.772, 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.

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.

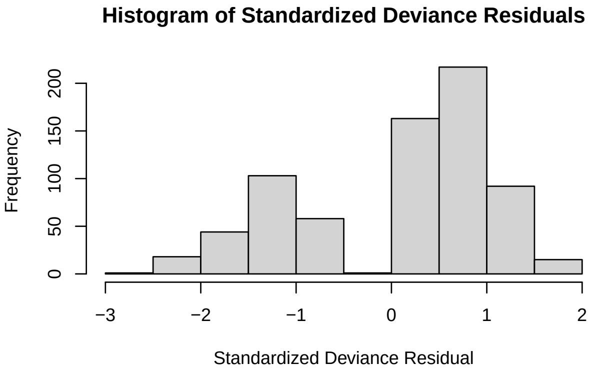
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  bestmodel.1$y, fitted(bestmodel.1)
## X-squared = 13.01, df = 9, p-value = 0.1622
```

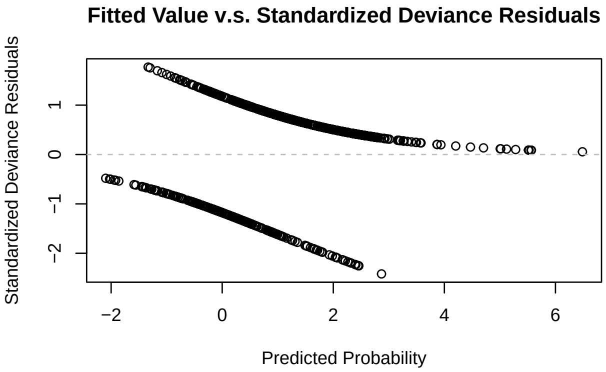
We can see that the p-value of the Hosmer-Lemeshow Test is 0.1622 which is much larger than the significance level $\alpha = 0.05$. Therefore, we fail to reject the null hypothesis and conclude that the selected model fits the data well.

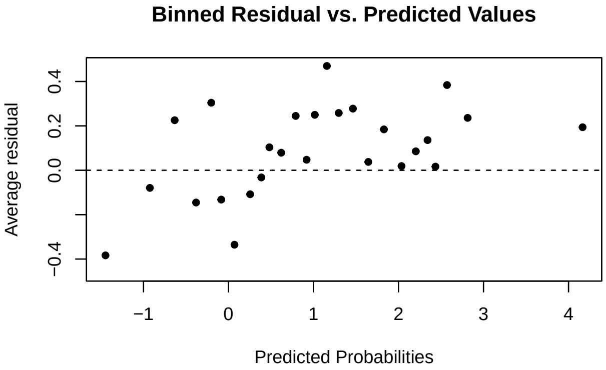
Predictive Power and Residual Diagnostics

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

We can see that the cutoff probability is 0.74. 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 sensitivity is 0.6914 and the specificity is 0.7581. Since the model has a high sensitivity and specificity, we have strong confidence that the model fits the data well.







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. More importantly, as we look at the binned residual plot, we can see that the mean residuals is near to 0.

In summary, we examined the ROC, AUC, Hosmer-Lemeshow Test, Predictive Power of the model and residual diagnostics. We found out the AUC is high, the p-value of Hosmer-Lemeshow Test is large, the high sensitivity and high specificity of the model indicates a strong predictive power. The mean residuals is near to 0 and near to normal. All these clues show that the selected model is a robust model.

Discussion and Conclusion

Based on this research, we successfully found several significant predictors for credit status, namely checking account status, savings amount, credit duration, the debtor's most valuable property, the debtor's age and debtor's stallment plans from providers other than the credit-giving bank. The most important determinants are duration and status. More specifically, whenever the credit duration increases by 1 month, the odds of having good credit status will be multiplied by a factor of $e^{-0.039856} = 0.96$, holding all else constant. Also, for people with checking account greater than or equal to 200 DM / salary for at least one year, their odds of having good credit status is $e^{1.983709} = 7.27$ times that for people with no checking account, holding all else constant. Such significant findings can help financial institutes to better assess a customer's credit status and issue credit limits.

However, there are also limitations in our model. For example, our model can only make predictions for people aged under 60 because we trimmed the data with age over 60 in model training. Also, the data we used in this research are collected from German, rather than from different regions or countries around the world. More than half of observations in our training set are with good credit status, which makes our model skewed in the prediction. Therefore, for future improvements, we could collect more data from seniors and from different regions to get more insight into the determinants of credit risk. Also, it is a good idea to have more data of bad credit status to increase the reliability of our model.

References

Fontinelle, A. (2023, April 5). The 5 biggest factors that affect your credit. Investopedia. Retrieved from https://www.investopedia.com/articles/pf/10/credit-score-factors.asp

Muhammad Nur Aidi and Resty Indah Sari. (2012) "Classification of debtor credit status and determination amount of credit risk by using linier discriminant function". AIP Conference Proceedings 1450, 280-285. Retrieved from https://doi.org/10.1063/1.4724155