1970 South German Credit Statistical Analysis

Group 15: Patrick Macadangdang, Meng He, Marco Wong, Chui Ling Mok

04/10/2022

Background/Summary

Data

Data Source: UCI Machine Learning Repository, "South German Credit" Dataset
Source: Ulrike Gromping, Beuth University of Applied Sciences Berlin

The dataset used for this research paper is the 1970 South German dataset which uses quantitative, categorical and ordinal variables to determine one's credit status. There was a total of 21 variables with the response variable being credit_risk. (ref 1)

Goal

The topic of interest for this analysis is to build a suitable model to predict the credit risk (status is good or bad) of a particular subject given the data for other explanatory variables. The analysis will select from the available quantitative and categorical predictor variables, and work towards a logistics regression model.

Purpose

The purpose of credit risk is to be used by creditors (ie. banks) to evaluate one's financial reliability and whether or not to give a line of credit. Whenever a loan is being made, there's always a risk that the borrower may not repay the lender. Even though calculating a borrower's credit risk may not 100% accurately determine rather they will return the loan or not, it can lessen the severity of a loss. Credit score can range from 300 to 850, where a higher score indicates more reliability and thus lesser chance the lender will suffer a lost from approving the loan request. For the purpose of this case study, R was used to build logistic regression models. (ref 2)

Note Most of the R codes used to generate this report is hidden in the PDF document. To find more information about the code used, please check the Rmarkdown file.

Variables

The "South German Credit" Dataset contains the following variables. To find out what each factor represents check reference 1 in the Apendix section (page 10).

Categorical	Quantitative
status (1:4)	duration
credit_history (0:4)	amount
purpose (0 to 10)	age
people_liable (1:2)	
savings (1:5)	
employment_duration (1:5)	
installment_rate (1:4)	
personal_status_sex (1:4)	
other_debtors (1:3)	
present_residence (1:4)	
property (1:4)	
other_installment_plans (1:3)	
housing (1:3)	
number_credits (1:4)	
job (1:4)	
telephone (1:2)	
foreign_worker (1:2)	
credit_risk (0,1)	

Summary of Procedures

The analysis will include a variety of statistical procedures, including the following:

- Frequency tables
- Testing for multicollinearity
- Multiple regression
- Logistics regression
- Forward selection, backward elimination, and comparing AIC values
- Testing goodness of fit for model
- ROC curves
- Residual inference

Visual Analysis

Just to get a general idea of the relationship between all the explantory variables with credit_risk, ggplot2 was used to visualize the distribution. The diagrams below show the histogram density graphs for all the variables against credit_risk

Looking at the histogram graphs above, there's a few things that stand out. For example, majority of the individuals that were sampled for this data were around 30-40 years old, has a credit duration of 20 months, had no other installments, and 0 to 2 people that are financially dependent of the debtor.

Feature Selection

There are 21 variables in this dataset. The first thing to do before building the model is to try and reduce the dimensionality of the data. Building a model with too many explanatory variables can lead to problems such as overfitting and creating very complex models that will take a long time to run. This is often refereed to as the curse of dimensionality.

The one in ten rule will be used to determine the maximum nunmber of parameters we should use in our model.

```
##   credit$credit_risk   n
## 1                  0 300
## 2                  1 700
```

The maximum number of predictors to use given this dataset would be $300/10 = 30$. This is even more than the number of variables available.

Forward Selection

Forward selection will be used to select significant variables.

```
model1 = glm(credit_risk ~ 1, data=credit)
model2 = formula(glm(credit_risk~.,data=credit))
model_fwd = step(model1, direction='forward',test="Chisq", scope=model2, trace=0)
summary(model_fwd)
```

```
##
## Call:
## glm(formula = credit_risk ~ status + duration + credit_history +
##     savings + other_debtors + installment_rate + personal_status_sex +
##     amount + employment_duration + other_installment_plans +
##     telephone + property + housing + number_credits + foreign_worker,
##     data = credit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0866  -0.3395   0.1146   0.2989   0.8373
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.818e-01  1.742e-01   1.618 0.105992
## status                   9.991e-02  1.077e-02   9.278  < 2e-16 ***
## duration                -4.348e-03  1.461e-03  -2.976 0.002994 **
## credit_history           6.520e-02  1.374e-02   4.746 2.38e-06 ***
## savings                  3.385e-02  8.429e-03   4.016 6.37e-05 ***
## other_debtors            5.828e-02  2.772e-02   2.102 0.035794 *
## installment_rate        -4.499e-02  1.287e-02  -3.496 0.000494 ***
## personal_status_sex      4.119e-02  1.848e-02   2.229 0.026046 *
## amount                  -1.487e-05  6.686e-06  -2.225 0.026325 *
## employment_duration      2.589e-02  1.101e-02   2.351 0.018927 *
## other_installment_plans  3.526e-02  1.873e-02   1.883 0.060055 .
## telephone                5.740e-02  2.769e-02   2.073 0.038412 *
```

```
## property               -3.360e-02  1.401e-02  -2.398 0.016652 *
## housing                 5.432e-02  2.615e-02   2.077 0.038060 *
## number_credits         -4.124e-02  2.504e-02  -1.647 0.099830 .
## foreign_worker         -1.072e-01  7.003e-02  -1.530 0.126332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1631397)
##
##     Null deviance: 210.00  on 999  degrees of freedom
## Residual deviance: 160.53  on 984  degrees of freedom
## AIC: 1042.6
##
## Number of Fisher Scoring iterations: 2
```

From the results above, it shows that the following variables are most significant to credit_risk.

- status
- duration
- credit_history
- savings
- installment_rate

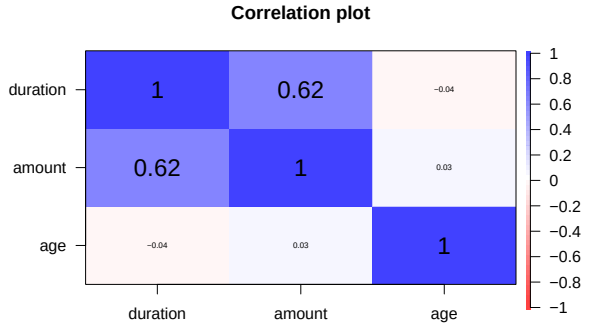
credit_risk is the reponse variable.

Multicolinearity

Multicollinearity needs to be checked between all the quantitative variables. This is important because multicollinearity can cause problems such as generating high variance of estimated coefficients which leads to inaccurate predictions.

In the South German Credit dataset, there are only 3 quantitative variables (duration, amount, age). Multicollinearity needs to be checked between the quantitative explanatory variables.

```
##              duration     amount         age
## duration  1.00000000 0.62498846 -0.03754986
## amount    0.62498846 1.00000000  0.03227268
## age      -0.03754986 0.03227268  1.00000000
```



According to the correlation matrix above, all the

quantitative variables seem to be independent to each other. The highest correlation is between amount and duration with a correlation of 0.62 which isn't high enough for it to be significant.

The final 5 significant explanatory variables that will be used for the rest of our analysis are the following:

- status (categorical)
- duration (quantitative)
- credit_history (categorical)
- savings (categorical)
- installment_rate (categorical)

Note: In the visual analysis section, we predicted that purpose, present_residence, job, and people_liable variable will not be significant according to just the graphs. Our prediction was correct.

Statistical Analysis

The analysis will being with using 3 different models to showcase the main effect, interaction, and no-predictor model.

```
status = as.factor(credit$status)
cred.hist = as.factor(credit$credit_history)
save = as.factor(credit$savings)
instal.rate = as.factor(credit$installment_rate)
y = as.numeric(credit$credit_risk>0)

# Main Effect Model
model.1 = glm(y~status + duration + cred.hist + save + instal.rate ,
              family=binomial,data=credit)

# Interaction Model
model.2 = glm(y~(status + duration + cred.hist  + save + instal.rate)^2 ,
              family=binomial,data=credit)

# No Predictors Model
model.3 = glm(y~1, family=binomial, data=credit)
```

Akaike Information Criterion Values

Between interaction, main effect and no predictor model, the AIC value will be calculated. The AIC value is an estimator of prediction error. A lower AIC value is better. The R code for AIC values give us the following:

```
## [1] 1026.502 1072.509 1223.729
```

AIC	Model
1026.502	Main Effect Model
1072.509	Interaction Model
1223.729	No Predictor Model

The model with the smallest AIC is the main effect model . Therefore, this is the preferred model.

Goodness of Fit (Hosmer-Lemeshow)

Since we are dealing with ungrouped data, we can use Hosmer-Lemeshow test for goodness-of-fit test. This partitioned our data into 10 equal sized groups and computed their chi-square statistic. For model 1, the p-value was 0.554 which exceeds 0.05. Therefore, we fail to reject and the current model fits the data well. The same follows for model 2 with a p-value of 0.1084 indicating a good fit.

```
hoslem.test(model.1$y, fitted(model.1),g=10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model.1$y, fitted(model.1)
## X-squared = 6.8399, df = 8, p-value = 0.554
```

```
hoslem.test(model.2$y, fitted(model.2),g=10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model.2$y, fitted(model.2)
## X-squared = 13.102, df = 8, p-value = 0.1084
```

Model Selection (Backward, Forward, Stepwise)

The backward, forward and stepwise function is used to select our model.

```
# 3 way interaction term
glm3 = glm(y~status*duration*cred.hist +
              status*duration*save+
              status*duration*instal.rate+
              duration * cred.hist * save +
              duration * cred.hist * instal.rate +
              cred.hist * save * instal.rate , family=binomial,data=credit)

aic1 = stepAIC(glm3, direction="backward", trace = 0)$aic
aic2 = stepAIC(model.3, direction="forward", scope=list(upper=glm3, lower=model.3), trace = 0)$aic
aic3 = stepAIC(model.3, direction="both", scope=list(upper=glm3, lower=model.3), trace = 0)$aic

cbind(aic1, aic2, aic3)
```

```
##          aic1     aic2     aic3
## [1,] 1044.954 1022.026 1022.026
```

Model Selection	AIC
Backward	1044.954
Forward	1022.459
Stepwise	1022.459

Forward and Stepwise model selection choose the same model, and its AIC value is smaller than the model selected by the backward elimination. Therefore, we will choose the model produced by forward selection as our model.

```
stepAIC(model.1, direction="both", scope=list(upper=glm3, lower=model.1), trace = 0)$call
```

```
## glm(formula = y ~ status + duration + cred.hist + save + instal.rate +
##     duration:cred.hist + duration:save, family = binomial, data = credit)
```

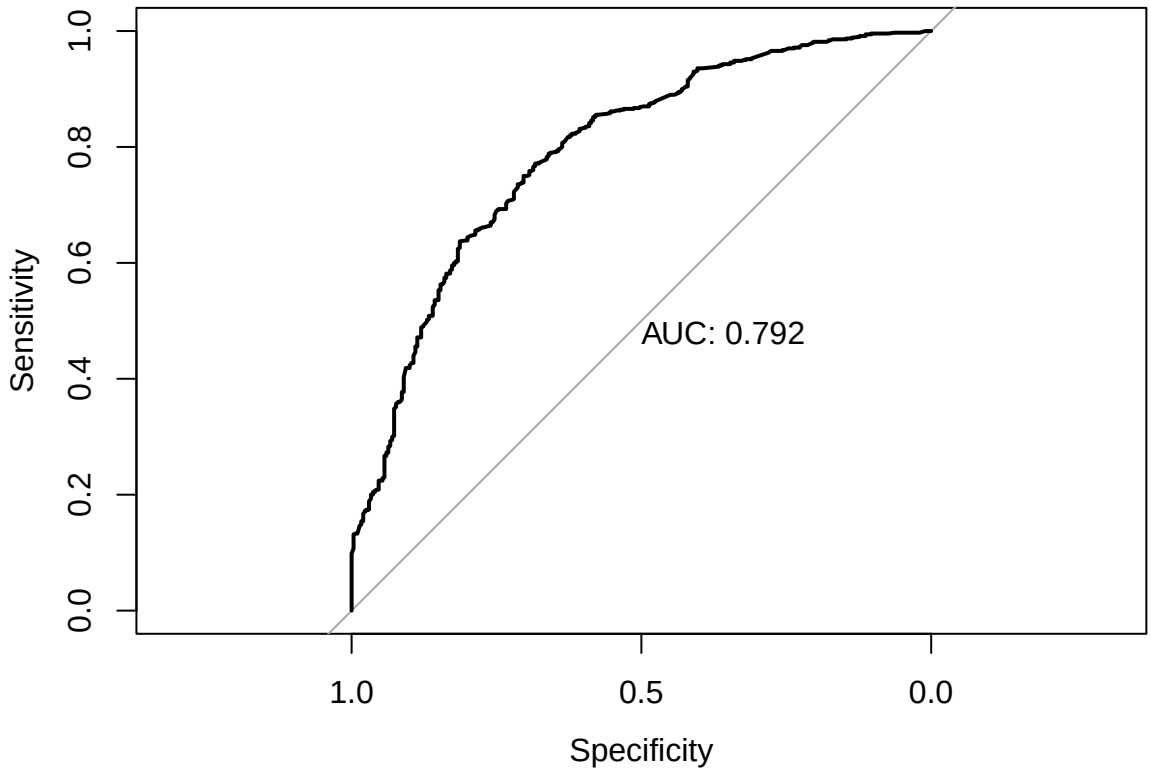
Therefore, out model is: $\text{logit}(\pi) = \alpha + \beta_1 status + \beta_2 duration + \beta_3 cred.hist + \beta_4 save + \beta_5 instal.rate + \beta_6 duration : cred.hist + \beta_7 duration : save$

ROC Curve

The area under a ROC curve is the concordance index x which estimates the probability that the predictions and outcomes are concordant. This means that the observation eith the larger y also has the larger $\hat{\pi}$. The larger the concordance index the better the model is. Note that a concordance of 0.5 corresponds to random guessing.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Our final model yielded a c value of 0.792 which is considered an acceptable value according to ScienceDirect (ref 3).

Conclusion

We have come up with a model to predict an individual's credit risk by analyzing multicollinearity, computing AIC values, goodness of fit, model selection and ROC curves. After our statistical analysis, we came up with a model that gives us a concordance index of 0.792 which is considered an acceptable value.

$\text{logit}(\pi) = \alpha + \beta_1 status + \beta_2 duration + \beta_3 cred.hist + \beta_4 save + \beta_5 instal.rate + \beta_6 duration : cred.hist + \beta_7 duration : save$

Status, duration, credit history, savings and installment rate is the significant explanatory variables used to predict credit_risk which is the respond variable.

Limitations

The dataset used for this report was from 1970 so it is a fairly old dataset. Many factors probably have changed by now that changes the prediction model. For example, the housing explanatory variable (1:for free, 2:rent , 3:own) might play a more significant role in terms of identifying credit risk now compared to 50 years ago. The model will probably still be a decent predictor for data in 2022 but might not be as accurate compared the data from 1970.

Apendix

Libraries Used for R Statistical Analysis

- rmarkdown
- tidyverse
- MASS
- psych
- tidyverse
- pROC
- ResourceSelection
- ggplot2

Reference

1. UCI Machine Learning Repository: Data Set. (1970). South German Credit. https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29/

2. What Is Credit Risk? (2022, March 15). Investopedia. https://www.investopedia.com/terms/c/creditrisk.asp#:%7E:text=Although%20it\T1\textquoterights%20impossible%20to%20know,reward%20for%20assuming%20credit%20risk

3. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. (2010, September 1). ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1556086415306043#:%7E:text=AREA%20UNDER%20THE%20ROC%20CURVE,-AUC%20is%20an&text=In%20general%2C%20an%20AUC%20of,than%200.9%20is%20considered%20outstanding

4. Schufa Credit Score Report in Germany. (2020b, January 2). Banks Germany. https://banks-germany.com/schufa-credit-score#:%7E:text=In%20Germany%2C%20the%20credit%20score,%2C%20or%20%E2%80%9CSchufa%20rating%E2%80%9D%2C